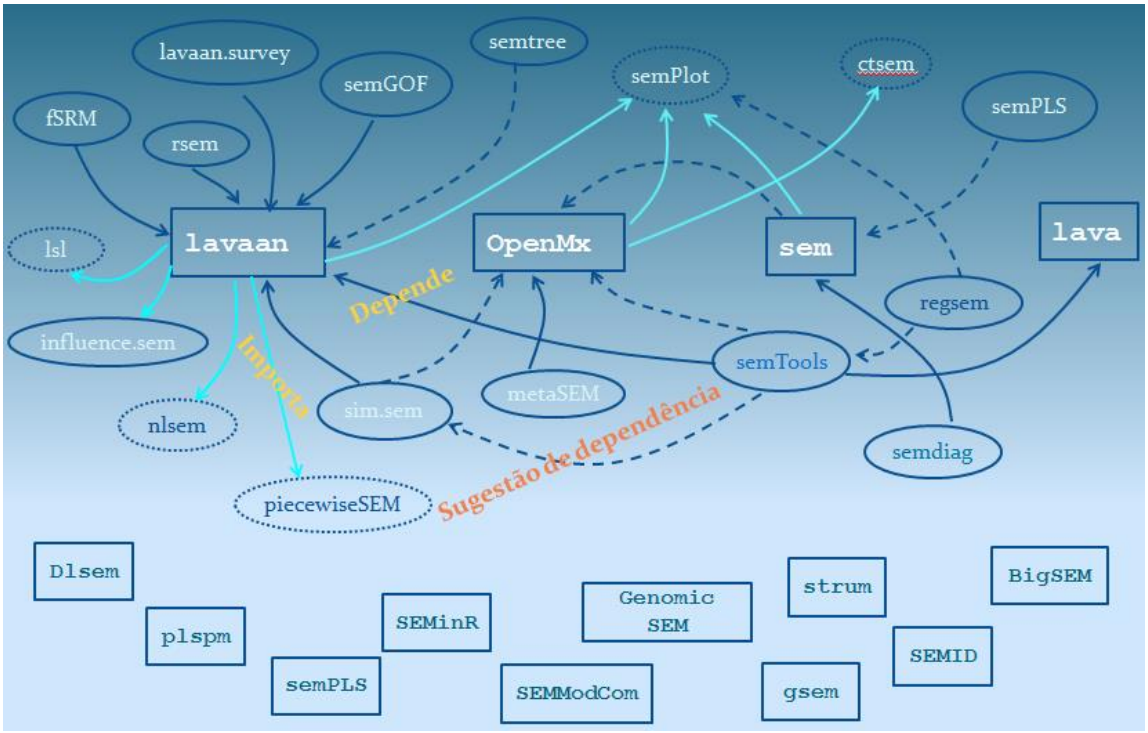
UNIVERSIDADE ABERTA

Modelos de Equações Estruturais – Métodos Computacionais



Maria da Conceição Dias Leal

Mestrado em Bioestatística e Biometria

Lisboa 2018

UNIVERSIDADE ABERTA

Modelos de Equações Estruturais – Métodos Computacionais

Maria da Conceição Dias Leal

Aluno nº 1002123

Mestrado em Bioestatística e Biometria

Dissertação orientada por

Orientadora

Prof.ª Doutora Teresa Paula Costa Azinheira Oliveira

Coorientador

Prof. Doutor Amílcar Manuel do Rosário Oliveira

Lisboa 2018

RESUMO

Resolver problemas complexos requer a capacidade, proporcionada pela Modelação de Equações Estruturais – SEM (*Structural Equation Modelling*), de examinar múltiplas influências e múltiplas respostas simultaneamente. Dada a sua flexibilidade e abrangência das aplicações, a SEM oferece um meio para desenvolver e avaliar ideias sobre relações multivariadas complexas, o que a torna capaz de responder a problemas e desafios quer das Ciências Sociais e Humanas quer das Ciências Naturais.

SEM "*is modeling hypotheses with structural equations*" (Grace, 2006). Esta é a definição que melhor se adequa à diversidade de abordagens que admite, seja na modelação, na análise ou ainda nas aplicações. Mais que uma metodologia, a SEM é uma coleção de técnicas estatísticas multivariadas que tem como objetivo principal avaliar em que grau um modelo teórico proposto é suportado pelos dados, o que a pode tornar um motor do conhecimento. De facto, na SEM a teoria é o motor da análise e os dados servem para testar a teoria, paradigma que rompe com a racionalidade estatística inferencial clássica, onde a análise dos dados precede a elaboração da teoria (Hair *et al.*, 2010).

Relativamente às técnicas multivariadas convencionais, a SEM tem duas vantagens que a tornam uma ferramenta capaz de lidar com problemas complexos e de gerar conhecimento em vários domínios: a capacidade de examinar simultaneamente múltiplas influências e múltiplas respostas e a capacidade de lidar com os erros de medição nos dados observados. Aliada a estas características acresce a facilidade em lidar com um elevado volume de dados e de diferentes tipos, de lidar com grupos múltiplos e com níveis múltiplos. Se se considerar ainda o facto de dispor de ferramentas para lidar com dados omissos, situação muito frequente quer nas Ciências Sociais e Humanas, quer na Ciências Naturais, melhor se percebe a importância da SEM na atualidade e a grande quantidade de artigos que ilustram a sua aplicação nas mais diversas áreas destas ciências.

O desenvolvimento computacional impulsionou a conceção de métodos estatísticos para melhorar a qualidade de produção científica e a automatização da recolha e armazenamento de dados, potenciando um aumento dramático da complexidade dos modelos e dos métodos. A SEM não foi exceção. Beneficiou, por um lado, com o

desenvolvimento de diversos *softwares* para a análise SEM, uns comerciais, como o AMOS, LISREL ou MPlus, a título de exemplo, e outros livres, disponíveis no *software* R, capazes de rivalizar com os comerciais. Por outro lado, por ser adequada para lidar com grandes volumes de dados, foi objeto de desenvolvimentos e aplicações cada vez mais complexas e mais abrangentes.

O objetivo do presente trabalho é o de fazer uma revisão da Modelação de Equações Estruturais, no que respeita a aplicações, fundamentos teóricos da SEM convencional (Analise Fatorial Confirmatória e Regressão Linear), com especial ênfase na análise SEM com dados omissos, tendo como motivação a exploração das potencialidades do *software* R, como recurso de livre acesso aos investigadores das diferentes áreas em que a SEM é especialmente útil. No âmbito das aplicações, as aplicações em Ciências da Vida e em Ciências Naturais foram o foco principal dado que, nas últimas décadas, estas sofreram grande expansão e a SEM tem contribuído para um maior amadurecimento de teorias e investigações. De facto, recorrer a formas de conectar modelos de equações estruturais com o processo científico é necessário se se quiser obter o máximo impacto de modelos e análises no processo de construção de conhecimento.

Palavras-chave: Modelação de Equações Estruturais, Dados Omissos, Pacotes SEM do R, Ciências Naturais.

ABSTRACT

Solving complex problems requires the capacity, provided by the SEM – *Structural Equation Modelling* – to analyse multiple influences and multiple responses simultaneously. Given its flexibility and the scope of its applications, SEM offers a means to develop and evaluate ideas about complex multivariate relations, which enables it to respond to problems and challenges both from Social and Human Sciences and Natural Sciences.

SEM "*is modeling hypotheses with structural equations*" (Grace, (2006). This is the definition that best suits the diversity of approaches it enables, whether considering modelling, analysis, or applications. More than just a methodology, SEM is a collection of multivariate statistical techniques the main objective of which is to evaluate the degree in which a proposed theoretical model is supported by data, making it a potential driving force for knowledge. In fact, within SEM, theory is the engine of analysis and data are used to test the theory, a paradigm that breaks away from the classical inferential statistical rationality, where data analyse precedes the elaboration of theory (Hair *et al.*, 2010).

In comparison to the conventional multivariate techniques, SEM offers two advantages that make it a useful tool both to deal with complex problems and to generate knowledge in several fields: the capacity to analyse simultaneously multiple influences and multiple responses, and the capacity to deal with the errors of measurement in the observed data. Together with these characteristics is the facility in dealing with a high volume of data, and of different types, and that of dealing with multiple groups and multiple levels. If one also considers the fact that it provides tools to deal with missing data, which frequently happens both in Social and Human Sciences and in Natural Sciences, then it's easy to understand the importance of SEM nowadays and the great amount of articles that illustrate its application in the most varied fields within these sciences.

The computational development pressed forward the conception of statistical methods to improve the quality of scientific production and the automation of data collection and storage, fostering a dramatic increase of the models and methods complexity. SEM wasn't an exception. On the one hand, it benefited with the development of several software's for

SEM analysis, ones that were commercial, like AMOS, LISREL or MPlus, and others that were free, available on R environment, capable of competing with the commercial ones. On the other hand, since it is adequate to deal with high data amounts, it was the object of more and more complex and wider developments and applications.

The aim of the present work is to review the Structural Equations Modelling in what concerns applications, theoretical grounds of conventional SEM (Confirmatory Factorial Analysis and Linear Regression), with special emphasis in the SEM analysis with missing data, so as to explore the potentialities of *software R* as a free access resource to researchers in the different areas where SEM is especially useful. As for applications, the ones in the fields of Life Sciences and Natural Sciences were the main focus since in the last decades they have undergone great expansion and SEM has been helping the process of maturing theories and investigations. In fact, one needs to reach for ways of connecting structural equations modelling to the scientific process if one wants to obtain the maximum impact of models and analysis in the knowing process.

Key Words: Structural Equation Modelling, R SEM Packages, Missing Data, Natural Sciences.

Aos meus pais, sempre

Às minhas amigas especiais…

AGRADECIMENTOS

A concretização de mais um sonho, de mais um projeto ou de mais um desafio nunca acontece se não for o resultado de um convergir de vontades.

A concretização do projeto aqui apresentado é o resultado disso mesmo. De uma recolha de vontades que se aliaram à minha!

À Professora Doutora Teresa Oliveira agradeço a vontade imensa de me fazer ir cada vez mais longe e o apoio nessa caminhada.

Ao Professor Doutor Amílcar Oliveira agradeço a vontade de ajudar, manifestada na disponibilidade com que sempre acolheu as minhas dúvidas.

A todos os Professores e Colegas que caminharam comigo agradeço a vontade de partilharem os seus conhecimentos e as suas experiências.

Às minhas amigas, que não preciso nomear, agradeço a vontade de partilharem o seu tempo e me apoiarem e estimularem em todas as etapas.

Aos meus pais agradeço a vontade de fazerem comigo, em todos os momentos e incondicionalmente, o caminho difícil e exigente, para mim e para eles, que culminou neste trabalho.

ÍNDICE

ÍNDICE DE QUADROS

ÍNDICE DE FIGURAS

SIMBOLOGIA E NOTAÇÕES

ADF	*(Asymptotic Distribution-Free)*
AVE	*(Average Variance Extrated)*
CFA	*(Confirmatory Factor Analysis)*
FIML	*(Full Information Maximum Likelihood)*
GLS	*(Generalized Least Square)*
GRN	*(Gene Regulatory Networks-s)*
GWAS	*(Genome-Wide Association Study (GWA study)*
LGM	*(Latent Growth Model)*
IM	*(Imputação Múltipla)*
LISREL	*(Linear Structural Relations)*
MAR	*(Missing At Random* e MNAR - *Missing Not At Random)*
MCAR	*(Missing Completely At Random)*
MCMC	*(Markov Chain Monte Carlo)*
MI	*Multiple Imputation*
ML	*(Maximum Likelihood)*
MV	*(Máxima Verosimilhança)*
PLS	*(Partial Least Square)*
QTL	*(Quantitative Traits Loci)*
SEM	*(Structural Equation Modelling)*

SNP	*(Single Nucleotide Polymorphism)*
ULS	*(Diagonally Weighted Least Squares)*
WLS	*(**W**eighted **L**east **S**quares)*
WLSMV	*(**W**eighted **L**east **S**quares for Mean and Variance)*
N	(Número de observações)
m	(Número de variáveis latentes endógenas)
n	(Número de variáveis latentes exógenas)
p	(Número de variáveis observadas endógenas)
q	(Número de variáveis observadas exógenas)
η	(Matriz de m variáveis latentes endógenas de ordem $m \times 1$ de elementos)
ξ	(Matriz de n variáveis latentes exógenas de ordem $n \times 1$)
ζ	(Matriz de erros estruturais de ordem $m \times 1$)
Y	(Matriz de p variáveis observadas endógenas de ordem $p \times 1$)
X	(Matriz de p variáveis observadas exógenas de ordem $q \times 1$)
ε	(Matriz de erros de medida de Y de ordem $p \times 1$)
δ	(Matriz de erros de medida de X de ordem $q \times 1$)
Λ_y	(Matriz de pesos fatoriais de η em Y de ordem $p \times m$)
Λ_x	(Matriz de pesos fatoriais de ξ em X de ordem $q \times n$)
B	(Matriz de coeficientes relacionando η com η, de ordem $m \times m$)
Γ	(Matriz de coeficientes relacionando ξ com η de ordem $m \times n$)
Φ	(Matriz de variâncias/covariâncias de ξ de ordem $n \times n$)

Ψ	(Matriz de variâncias/covariâncias de ζ de ordem $m \times m$)
Θ_{ε}	(Matriz de variâncias/covariâncias de ε de ordem $p \times p$)
Θ_{δ}	(Matriz de variâncias/covariâncias de δ de ordem $q \times q$)

INTRODUÇÃO

A Modelação de Equações Estruturais (SEM – *Structural Equation Modelling*) é uma coleção de técnicas estatísticas multivariadas, que na maioria dos casos, são usadas para formular, ajustar e testar uma grande variedade de modelos para dados contínuos que vão desde a análise fatorial exploratória e a análise fatorial confirmatória, a regressão linear multivariada, a análise de caminhos (*Path Analysis*), curvas de crescimento aleatório, e outros modelos longitudinais, a modelos de erros nas variáveis, modelos de mediações, entre outros. Um dos principais objetivos é o de determinar o grau em que um modelo teórico proposto é suportado pelos dados. Numa abordagem tradicional, a metodologia assenta em métodos baseados em covariâncias, e consiste num conjunto de procedimentos estatísticos muito geral que são largamente usados. Grace (2006) define SEM de uma forma muito geral – SEM "*is modeling hypotheses with structural equations*".

Nas ciências sociais a sua utilização predomina na Psicologia, mas a Educação, a Economia, a Ciência Política e a Sociologia são também campos de aplicação importantes. A Ecologia, a Biologia, a Genética, as Neurociências são áreas de aplicação mais recentes e direcionadas para a biometria.

No âmbito das Ciências Sociais e Humanas, a modelação consiste em quantificar de que forma variáveis observadas (indicadores ou variáveis manifestas) são indicativos indiretos de variáveis latentes (variáveis não-observadas que são construídas a partir dos indicadores), também conhecidas como constructos teóricos ou fatores (Hair *et al*., 2010). Na Ecologia, na Genética, na Biologia e noutras ciências da vida, a modelação SEM assume essencialmente a vertente de *path modelling (*numa perspetiva moderna que incorpora variáveis latentes na análise de caminhos – *path analysis)* para testar o ajustamento de um modelo "causal" hipotético que incorpora métodos de máxima verosimilhança e testes de ajustamento global do modelo. A modelação com variáveis latentes na perspetiva das aplicações às Ciências Sociais e Humanas também se aplica a problemas do âmbito destas ciências.

As relações entre variáveis podem ser descritas em termos de correlação, a qual indica o grau de linearidade entre duas variáveis, de covariância, que dá a medida de quanto duas variáveis variam juntas, e de regressão, que é a transformação da relação entre variáveis, observadas ou latentes, numa equação. Os modelos podem ter diversas "estruturas", isto é,

as relações entre as diferentes variáveis, traduzidas em diversas equações, podem ter diversas configurações espaciais, dependendo da teoria hipotética que se tem *a priori*, mas que pode sofrer modificações de acordo com os resultados obtidos por modelos hipotéticos concorrentes.

O interesse em SEM é essencialmente em construções teóricas que são interpretadas pelas variáveis latentes - variáveis que não podem ser diretamente observadas para todos os membros de uma dada amostra (Hoyle, 2012). Nas equações estruturais a teoria é o motor da análise e os dados servem para confirmar, ou não, a teoria. Este paradigma rompe com a racionalidade estatística inferencial clássica, onde a análise dos dados precede a elaboração da teoria (Hair *et al.*, 2010)

Byrne (2012) comparou o SEM com outras técnicas multivariadas e listou quatro características únicas da SEM:

1. SEM assume uma abordagem confirmatória para a análise de dados, especificando as relações entre as variáveis *a priori*.
2. SEM fornece estimativas explícitas de parâmetros de variância de erro. Outras técnicas multivariadas não são capazes de avaliar ou corrigir erros de medição.
3. Os procedimentos SEM incorporam variáveis não observadas e observadas, enquanto outras técnicas multivariadas baseiam-se apenas nas medidas observadas.
4. SEM é capaz de modelar relações multivariadas e estimar os efeitos diretos e indiretos das variáveis em estudo.

O modelo de equações estruturais é caraterizado por dois componentes básicos:

(i) um modelo de medida que especifica as relações entre variáveis observadas e variáveis latentes permitindo "medir" uma variável latente (cuja medida direta não se pode obter) através de um conjunto de variáveis observáveis. A análise fatorial confirmatória é frequentemente usada para testar esta componente do modelo. No modelo de medida, o investigador deve decidir que indicadores observados permitem definir os fatores latentes. A medida em que uma variável latente é definida com precisão depende de quão fortemente estão relacionados os indicadores observados, havendo uma falta de especificação do modelo nas relações hipotéticas entre as variáveis se um indicador selecionado para um fator estiver debilmente relacionado a outros indicadores selecionados para o mesmo fator. Assim, a análise fatorial confirmatória pode ser precedida de uma

análise fatorial exploratória para selecionar os fatores que melhor permitem "medir" a variável latente.

(ii) Um modelo estrutural que define as relações causais ou de associação entre as variáveis latentes, especificando se uma variável latente causa mudanças noutras variáveis latentes no modelo, direta ou indiretamente.

A popularidade crescente da SEM foi acompanhada ou foi impulsionada de/por mudanças na análise estatística de dados de pesquisa. A Lei de Moore (a complexidade dos circuitos de computador, ou seja, o poder de computação, duplica aproximadamente a cada 18-24 meses (Moore, 1965)) impulsionou parcialmente a exploração do desenvolvimento computacional na conceção de métodos estatísticos para melhorar a qualidade de produção científica, potenciando um aumento dramático da complexidade dos modelos e dos métodos. Por outro lado, também pela via do desenvolvimento computacional, os métodos de recolha de dados tornaram-se mais automatizados e o armazenamento de dados tornou-se acessível, levando a que os conjuntos de dados aumentassem drasticamente de tamanho. Como consequência, os projetos de pesquisa tornaram-se mais ambiciosos, na medida em que se tornou possível recolher muitas medidas de grandes amostras.

Assim, o desenvolvimento computacional levou ao surgimento de uma grande variedade de *softwares* comerciais que permitem implementar a análise SEM, nomeadamente Amos - SPSS (Byrne, 2012), Calis - SAS (PROC PROC CALIS, 2010), EQS (Bentler e Wu, 2005), LISREL (Jöreskog & Sörbom, 2009), Mplus (Muthén e Muthén, 2009), SEPath (SEPath, 2013). As opções referidas são todas comerciais, o que, para além do custo associado, limita a possibilidade de explorar novas ideias metodológicas, dado que os detalhes de muitos recursos normalmente permanecem ocultos ao utilizador. A implementação de ferramentas de análise SEM num ambiente *open-source* como R (R Core Team, 2017) resolve estes dois problemas, na medida em que é de utilização totalmente gratuita e permite aos utilizadores explorarem soluções que respondem aos seus problemas em particular. Atualmente, estão disponíveis diversas soluções, destacando-se os pacotes `sem` (Fox (2017)), `OpenMx` (Boker *et al.*, 2011), `lava` (Klaus, 2018) e `lavaan` (Rosseel, 2018). Muitos outros recursos estão disponíveis, ora para lidar com análises em contextos específicos ou com dados especiais, ora para suporte gráfico, ora com ferramentas para tratar etapas específicas da análise.

A Modelação de Equações Estruturais é regularmente aplicada em situações em que faltam dados em certas variáveis e não pode ser realisticamente assumido que a omissão dos dados ocorreu completamente ao acaso. Nestes casos os métodos comuns de análise SEM dão estimativas que são ineficientes e possuem grande viés amostral. Atualmente, nas ciências sociais e comportamentais, os dados omissos (*missing data*) são frequentemente tratados com métodos de imputação múltipla ou técnicas de máxima verossimilhança (FIML), mas há pesquisadores nestas ou noutras áreas que não adotaram estas metodologias na mesma medida e, com alguma frequência, utilizam técnicas de imputação tradicionais ou de análise de casos completos, o que pode comprometer o poder da análise e introduz enviesamento não intencional.

O objetivo deste trabalho é o de fazer uma revisão da análise de Modelos de Equações Estruturais, com especial ênfase na análise SEM com *missing data*, tendo como motivação a exploração das potencialidades do *software* R como recurso de livre acesso aos investigadores das diferentes áreas em que a SEM é especialmente útil.

Esta dissertação está estruturada em cinco capítulos além da presente introdução onde se descrevem os objetivos da dissertação e a estrutura da mesma.

No capítulo 1 é feita uma resenha histórica e no capítulo 2 é feita uma revisão de literatura no que respeita às aplicações, com especial enfase nas aplicações às ciências da vida.

O capítulo 3 contempla o enquadramento teórico da Modelação de Equações Estruturais. Explica as condições de identificação do modelo, a estimação dos parâmetros e a especificação, bem como a avaliação da qualidade do ajustamento do modelo através dos índices e outras medidas de ajustamento.

O capítulo 4 destina-se ao tratamento da análise de Modelos de Equações Estruturais com dados omissos. Além da caracterização dos mecanismos de *missig data*, são apresentadas as metodologias de análise de dados omissos no contexto da Modelação de Equações Estruturais. São ainda exploradas as ferramentas disponíveis em pacotes do *software* R para lidar com dados omissos.

No capítulo 5 é apresentada a panóplia de ferramentas disponíveis no *software* R para implementação da análise SEM.

Finalmente, no capítulo 6 são tecidas considerações finais.

CAPÍTULO 1

PERSPETIVA HISTÓRICA

PERSPETIVA HISTÓRICA

O desenvolvimento da Modelação de Equações Estruturais (SEM – ***S**tructural **E**quation **M**odelling*) fez-se essencialmente no seio da biometria, econometria, psicometria e sociometria, que têm como principais subdisciplinas a biologia, a economia, a psiquiatria e a sociologia.

A SEM tem a sua génese nos trabalhos seminais do geneticista Wright (1921, 1934), cujo objetivo principal foi o de estabelecer uma estrutura para aprender sobre processos causais. Os métodos estatísticos tradicionais não respondiam aos requisitos particulares exigidos na inferência de relações causa-efeito e a SEM, envolvendo uma sequência de etapas projetadas, permite tais inferências. Esta linha de desenvolvimento da SEM advém dos modelos de equações múltiplas, a partir dos quais Wright criou a análise de caminhos (*path analysis*) e a modelação gráfica para implementar a análise causal em sistemas biológicos. Seguiu-se a aplicação à econometria (Haavelmo, 1943) e às ciências sociais (Grace *et al*., 2012), o que abriu caminho à análise SEM, cuja implementação inicial se limitou à análise de matrizes de correlação.

Uma segunda linha de influência da SEM é o modelo analítico da análise fatorial desenvolvida pela psicologia e psicometria do início do século XX, graças a alguns trabalhos seminais (por exemplo, Pearson, 1901; Spearman, 1904; Thurstone,1931,1935, entre outros) e aperfeiçoada por outros psicometristas no decorrer do século XX.

De facto, no decorrer deste século vários psicometristas e estatísticos aperfeiçoaram o modelo da análise fatorial até ao desenvolvimento do modelo atual da SEM, especialmente devido ao desenvolvimento e aperfeiçoamento da capacidade de computação estatística por volta da década de 50.

Por esta génese híbrida, a SEM é apresentada por vários autores (p.ex. Ullman, 2007) como uma mistura de análise fatorial com regressão múltipla e análise de caminhos (*Path analysis*).

Após os primórdios do desenvolvimento da SEM, uma segunda geração da SEM surge dos trabalhos de Jöreskog (1973), Keesling (1972, *apud* Kaplan, 2009), e Wiley (1973, *apud* Kaplan, 2009), e que resultou numa abordagem analítica coerente que combina a análise fatorial e a modelação de equações simultâneas numa única metodologia – o modelo LISREL, descrita por Jöreskog (1973). O modelo LISREL contempla a

comparação entre a matriz das covariâncias implícitas no modelo e a matriz das covariâncias observadas e os métodos de máxima verosimilhança para a estimação, e permitiu um desenvolvimento muito significativo nas aplicações que envolvem variáveis latentes e variáveis observadas.

Até à publicação de alguns trabalhos na década de 90 (p.ex. Mitchell, 1992; Wootton 1992; Brown & Weis, 1995; Shipley, 1995; Pugesek & Tomer, 1996; Grace & Pugesek, 1997), que ilustram a utilidade potencial da SEM nas ciências naturais, em particular na ecologia e na biologia evolutiva, a sua utilização nestas ciências era muito incomum, resumindo-se quase aos trabalhos de Wright. Entretanto, verificou-se uma grande expansão no número e na variedade das aplicações da SEM no contexto das ciências naturais, acompanhada do crescimento do interesse dos cientistas por esta metodologia, nomeadamente com a exploração de ideias provenientes da inteligência artificial (Pearl, 2012) e da estatística *Bayesiana* (Scheines, Hoijtink & Boomsma, 1999; Lee, 2007a).

Pearl (2012) defendeu que a SEM é a linguagem natural para representar e estudar relações causais cuja explicitação é requisito para o desenvolvimento de sistemas inteligentes e desenvolveu uma teoria coerente que explica os requisitos para o raciocínio causal sob incerteza. Generalizou a SEM ao nível não paramétrico e aplicou à análise causal uma abordagem teórico-gráfica. Propôs ainda novos operadores matemáticos para apoiar a extração de interpretações causais de dados e sintetizou essas ideias num conjunto de requisitos para a SEM com requisitos de entrada definidos, que servem como um mecanismo de inferência (Pearl, 2012). Shipley (2000b, 20003, 2009) adotou algumas destas ideias mas as propostas de Pearl ainda não estão integradas na prática geral da SEM (Grace *et al*., 2010).

Uma das principais críticas ao modelo LISREL diz respeito aos pressupostos sobre a continuidade das variáveis observadas, as distribuições normais multivariadas e grandes tamanhos de amostra - necessários para capitalizar as propriedades assintóticas da estimação de máxima verossimilhança e dos testes associados, que raramente são verificados na prática. Em resposta a esta crítica, Browne (1984) fez uma contribuição marcante ao desenvolver um estimador com "distribuição assintótica livre" (ADF – *asymptotic distribution-free* ou WLS - ***W****eighted* ***L****east* ***S****quares*) e que foi crucial para o desenvolvimento de modelos SEM para variáveis dependentes dicotómicas, ordinais e limitadas (Müthen, 1984).

No contexto de algumas SEM básicas, muitos estudos (p. ex. Boomsma, 1982; Chou, Bentler & Satorra, 1991; Hu, Bentler & Kano, 1992; Hoogland & Boomsma, 1998) mostraram que as propriedades das estimativas de máxima verosimilhança não são robustas para amostras pequenas, resultante do facto de a matriz de covariância S da amostra ser assintoticamente normal, isto é, mesmo que os dados fornecidos sejam normais, a distribuição de S aproxima-se da normalidade apenas se o tamanho da amostra correspondente for grande. Pelo contrário, como apontado por muitos artigos importantes em análises *Bayesianas* de modelos de equações estruturais (Ansari & Jedidi, 2000; Ansari, Jedidi & Dube, 2002; Ansari, Jedidi & Jagpal, 2000; Scheines, Hoijtink & Boomsma, 1999; Lee & Song, 2014; Song & Lee, 2006; Muthén & Asparouhov, 2012; Merkle & Wang, 2016), os métodos *Bayesianos* baseados em amostragem dependem menos da teoria assintótica e, portanto, têm o potencial de produzir resultados confiáveis mesmo com amostras pequenas.

Há uma ligação natural da estimação pelo método da máxima verosimilhança com a estimação *bayesiana*, uma vez que as estimativas b*ayesianas* são simplesmente as probabilidades, ponderadas pelas probabilidades anteriores. A utilização da abordagem *Bayesiana* para a SEM permite incluir na inferência um conjunto mais amplo de fontes de informação prévia, útil para obter melhores resultados, especialmente quando o objetivo é prever observações futuras – a transição da análise retrospetiva para a análise prospetiva é facilitada. Permite ainda abrir o leque de especificações estatísticas que podem ser estimadas devido à flexibilidade dos procedimentos de Monte Carlo baseado em Cadeias de Markov (MCMC) (Gelman *et al*. 2014). As distribuições posteriores de parâmetros e de variáveis latentes podem ser estimadas usando um número suficientemente grande de observações que são simuladas a partir da distribuição posterior dos parâmetros desconhecidos, usando ferramentas eficientes em computação estatística, como os métodos MCMC. A abordagem *bayesiana* expande o leque de possíveis aplicações da informação (Kjaerulff & Madsen, 2008), o que é uma mais-valia muito significativa. Diversas generalizações da análise SEM padrão foram desenvolvidas nesta abordagem e incluem o desenvolvimento de modelos com covariáveis fixas, modelos não lineares, modelos multiníveis, modelos multiamostras, modelos de mistura, modelos com variáveis categóricas e contínuas misturadas, modelos com dados omissos e modelos com dados provenientes de uma família de distribuições exponenciais (Lee, 2007a).

Os desenvolvimentos metodológicos entretanto desencadeados têm conduzido a SEM ao que se pode designar por uma terceira geração de análise SEM, na qual a tradução da teoria (tradução da teoria em modelos de equações estruturais – Grace *et al.*, 2010), a inferência causal (Pearl, 2012) e a especificação estatística (Lee, 2007a) poderão ser integradas num processo de modelação, permitindo que a prática da SEM englobe uma metodologia científica mais completa (Grace *et al*., 2012).

A história do desenvolvimento da SEM reflete, em muitos aspetos, os desenvolvimentos em teoria estatística, metodologia e filosofia da ciência no século XX. Embora, ao longo dos anos, a SEM tenha sido objeto de desenvolvimentos metodológicos e aplicações, as últimas duas décadas presenciaram um grande aumento na implementação de SEM, tornando-se esta cada vez mais fácil, dado o desenvolvimento de programas computacionais muito eficientes que dispõem de vários métodos de estimação e que proporcionam informações adicionais necessárias para avaliar o ajustamento e a especificação dos modelos.

Karimi & Meyer (2014) conceberam o que designam por um pseudodiagrama de trajetórias com alguns desenvolvimentos em SEM e que teve a contribuição de Peter Bentler (Figura 1). O diagrama apresenta desenvolvimentos iniciais da SEM em azul, desenvolvimentos posteriores em amarelo e desenvolvimentos mais recentes, anteriores a 2014, em vermelho. Note-se que este diagrama contempla desenvolvimentos da SEM com interesse para a psicologia. Apesar da SEM *Bayesiana* não estar explicitamente referida no diagrama, a abordagem SEM, nesta perspetiva, surgiu no período abrangido pelo mesmo, como já referido.

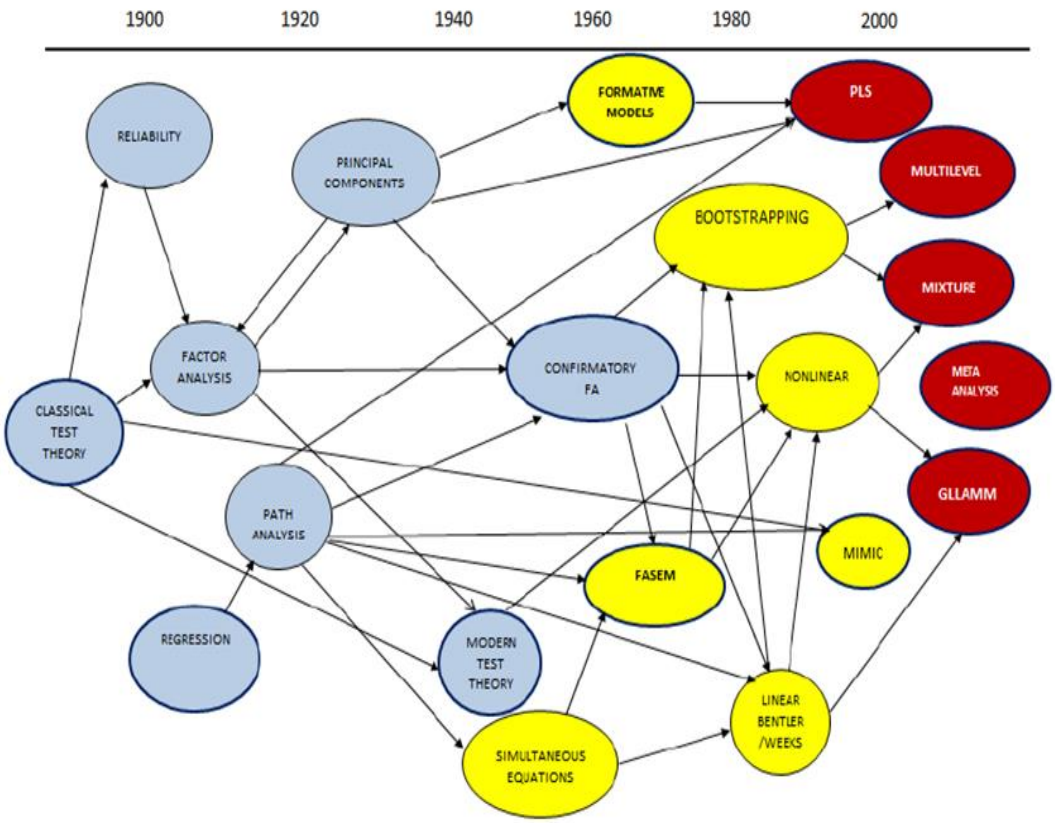


Figura 1: Pseudodiagrama de trajetórias com alguns desenvolvimentos em SEM (Karimi e Meyer, 2014). FASEM – *Factorial Analysis of SEM*; GLLAMM - *Generalized Linear Latent and Mixed Models*; MIMIC - *Multiple-Indicators Multiple-Causes Model*.

CAPÍTULO 2

APLICAÇÕES DE MODELOS DE EQUAÇÕES ESTRUTURAIS

APLICAÇÕES DE MODELOS DE EQUAÇÕES ESTRUTURAIS

Considerando a forma como a SEM se desenvolveu historicamente, esta metodologia pode considerar-se o lugar comum onde biometria, econometria, psicometria e sociometria se encontram. O desenvolvimento de ferramentas computacionais, tanto no que respeita à implementação da metodologia, como à melhoria das *interfaces* de utilizador, tornou prontamente acessível aos investigadores este poderoso e sofisticado método, e estes perceberam que a SEM está bem adaptado para responder a uma grande variedade de questões de pesquisa. Uma combinação de avanços metodológicos, da melhoria de *software* e da aplicabilidade a uma grande panóplia de problemas de interesse resultou numa ampla utilização da SEM, que por sua vez produziu avanços no conhecimento substantivo em diversas áreas (MacCallum & Austin, 2000).

A SEM é usada atualmente por um grande conjunto de investigadores nas mais diversas áreas, incluindo educação, *marketing*, psicologia, sociologia, administração, saúde, demografia, comportamento organizacional, biologia (em particular, da ecologia) e genética.

De facto, por fornecer um método direto para lidar com múltiplas relações simultaneamente com eficiência estatística, permitir avaliar as relações em âmbito geral e fornecer uma transição da análise exploratória para a análise confirmatória (Hair *et al*., 2010), a metodologia é amplamente aplicada nas ciências sociais e comportamentais, no *marketing*, nas ciências da informação, bem como nas ciências naturais, nas ciências da saúde e na bioestatística, aplicações que nestes últimos casos têm aumentado significativamente em quantidade e em importância. A capacidade que a metodologia tem de lidar com variáveis latentes, erros de medição e múltiplos indicadores, proporciona aos investigadores uma forma flexível e potente de examinar relações entre variáveis observadas e variáveis latentes, de testar similaridades entre grupos e diferenças entre variáveis latentes (Brown, 2006; Kline, 2011). A capacidade de corrigir erros de medição e o facto de ter menos pressupostos do que métodos clássicos (Little, 2013), fazem com que a metodologia SEM desempenhe um papel cada vez mais importante no desenvolvimento e na produção de conhecimento, na medida em que integra medição e teoria substantiva. Efetivamente, a SEM pode definir-se como um método para representar, estimar e testar

uma rede teórica de relações lineares entre variáveis, observadas ou latentes, nas mais diversas áreas de investigação.

As especificidades das diferentes áreas de investigação resultam em particularidades na abordagem SEM nas diferentes aplicações, facto que, por si só, ilustra a grande flexibilidade da metodologia. Podemos citar diferentes autores que se debruçaram sobre aspetos técnicos e metodológicos da SEM e sobre as suas aplicações em áreas específicas de investigação.

Shipley (2000a) explora, no contexto da Biologia, o estudo de relações de causa e efeito entre variáveis, num contexto causal multivariado, através de uma série de métodos estatísticos, com recurso ao R. Explica como testar hipóteses causais multivariadas usando equações estruturais e análise de caminhos, quando não é possível realizar experiências aleatórias (o que quase sempre acontece em Biologia), através de exemplos da ecologia fisiológica e desmistifica a utilização de hipóteses causais nesta metodologia, no contexto da Biologia.

Pugesek, Tomer e Von Eye (2003) têm como objetivo ajudar os biólogos a entenderem a diferença entre SEM e a análise de caminhos propriamente dita. Proporcionam a formulação básica do método, os detalhes técnicos sobre análise de dados, interpretação e relatórios, bem como inúmeros exemplos de projetos e aplicações de pesquisa que são adequados às necessidades e interesses da pesquisa biológica.

Skrondal e Rabe-Hesketh (2004) apresentam uma abordagem que unifica a modelação com variáveis latentes, incluindo modelos multinível, modelos longitudinais, modelos de itens de resposta, modelos de classe latente e modelos de equações estruturais. Explicam e comparam uma ampla gama de métodos de estimação e predição com abordagens específicas na bioestatística, na psicometria, na econometria e na estatística em geral. Incluem exemplos com diversos tipos de respostas não padrão, como dados ordinais, nominais, de contagem e de sobrevivência que ilustram a resolução de problemas concretos em áreas como medicina, economia e psicologia.

Kaplan (2009) apresenta uma visão geral de aspetos teóricos da SEM, incluindo muitos desenvolvimentos como a SEM multinível, SEM não normal, dados omissos, análise de classe latente, modelos mistos, discretos e contínuos, e curvas de crescimento latente. Utiliza e explora exemplos substantivos, extraídos de questões atuais no campo da

educação, guiados por uma estrutura teórica. Além do *software* Mplus o autor usa também o software R nas suas análises.

Grace (2006) apresenta a SEM como uma ferramenta que simultaneamente permite a análise de dados e proporciona uma forma diferente de fazer ciência, pois contempla métodos para desenvolver e avaliar teorias multivariadas. Na sua perspetiva, "compreender os sistemas requer a capacidade de examinar influências e respostas simultâneas" e a SEM oferece um meio de desenvolver e avaliar ideias sobre relações (multivariadas) complexas em sistemas naturais, em particular sistemas ecológicos, pois permite desenvolver *insights* mais profundos sobre as relações entre o padrão ecológico e o processo. Enfatiza exemplos de aplicações em Ecologia, pretendendo ilustrar as potencialidades da metodologia para ajudar à compreensão científica dos sistemas naturais, em contraponto com métodos científicos convencionais usados para a produção de conhecimento nas ciências naturais.

Lee (2007a) apresenta uma abordagem *Bayesiana* da SEM propondo generalizações da SEM com variáveis categóricas ordenadas, SEM com variáveis dicotômicas, SEMs não lineares, SEMs de dois níveis, SEMs *multisample*, SEMs de misturas, SEMs com *missing data* ignoráveis e/ou não-ignoráveis, SEMs com variáveis de famílias da distribuição exponencial e algumas das suas combinações. Na formulação de vários modelos SEM, e no desenvolvimento dos métodos *Bayesianos*, a enfase é colocada nas observações aleatórias individuais e não na matriz de covariância amostral em que assentam as abordagens tradicionais da SEM. A abordagem é exemplificada com dados reais relacionados com gestão, psicologia e sociologia.

Lee (2007b) apresenta uma vasta gama de modelos com variáveis latentes, predominando os modelos de equações estruturais, com diversas abordagens, e com exemplos ilustrativos usando conjuntos de dados reais de negócios, educação, medicina, saúde pública e sociologia.

Schumacker e Lomax (2004) apresentam diferentes tipos de modelos de equações estruturais, numa abordagem concetual orientada para aplicações. Abordam conceitos básicos, princípios e práticas necessárias para testar modelos teóricos e utilizam a SEM em exemplos significativos úteis em áreas como medicina, ciência política, sociologia, educação, psicologia, negócios e ciências biológicas.

Kline (2011) apresenta a SEM utilizando o mínimo de fórmulas e símbolos matemáticos possível, fazendo uma apresentação essencialmente conceitual. São apresentados muitos

exemplos de aplicação de SEM para investigar problemas em várias disciplinas, incluindo psicologia, educação, ciências da saúde, *marketing* e gestão.

Wang e Wang (2012) focam-se em aspetos conceituais e práticos da SEM e ilustram a aplicação dos modelos usando dados transversais e longitudinais extraídos de estudos de saúde pública.

Hoyle (2012) apresenta uma cobertura abrangente de SEM, começando com questões de fundo, continuando através de fundamentos estatísticos e etapas de implementação, passando para aplicações básicas e avançadas de SEM. São apresentadas aplicações da área das ciências sociais e comportamentais, nomeadamente na parte final do livro que contempla aplicações básicas cujo objetivo é o de ilustrar e discutir a aplicação da SEM em diversas abordagens e de diversos conceitos associados, e aplicações altamente especializadas, nomeadamente à genética e à neurociência com dados de neuroimagem.

Hair *et al.* (2014) explicam detalhadamente os fundamentos básicos da SEM usando os Mínimos Quadrados Parciais (PLS-SEM) como método de estimação complementar ao método mais amplamente aplicado, baseado em covariância (CB-SEM), e fornece diretrizes gerais para entender e avaliar os resultados da aplicação do método. O autor esclarece a natureza e o papel do PLS-SEM na investigação em ciências sociais, para tentar consciencializar os investigadores para uma ferramenta que poderá proporcionar novas e diferentes formas de desenvolver pesquisa.

2.1. Ciências Naturais

Apesar de a SEM ter as suas raízes na genética evolucionária (*path analysis* de Wright, 1921), a maioria das aplicações e dos desenvolvimentos ocorreram, como já referido, nas ciências sociais e humanas, no âmbito da econometria, psicometria e sociometria. No entanto, houve uma notável expansão no número e variedade de aplicações de SEM nas ciências naturais nos últimos anos (Shipley, 2000a; Pugesek, Tomer & Von Eye,2003; Grace, 2006).

As razões pelas quais a SEM pode ser usada nas ciências naturais prendem-se com: (1) ser fortemente orientada pela teoria, em oposição à hipótese nula, (2) a sua capacidade de representar hipóteses sobre redes causais, (3) os seus procedimentos para testar entre modelos concorrentes e (4) proporcionar uma estrutura para interpretação quando há um grande número de preditores e respostas com conexões causais complexas.

Chang (1981, *apud* Fan *et al*., 2016) e Maddox e Antonovics (1983, *apud* Fan *et al*., 2016) foram dos primeiros ecologistas a usarem SEM na pesquisa ecológica, esclarecendo as relações lógicas e metodológicas entre correlação e causalidade.

Fan *et al.* (2016) apresentam uma revisão das aplicações da SEM a problemas da ecologia. Esta revisão contempla aplicações básicas de SEM e resume as potenciais aplicações para modelos de SEM, incluindo os problemas e desafios que surgem na aplicação da modelação de equações estruturais nesta área científica.

Foram desenvolvidos estudos em diversas áreas da Ecologia. Podemos encontrar na literatura uma grande diversidade de aplicações na modelação em ecologia das populações animais e plantas e na ecologia das comunidades – comunidades de animais, comunidades de plantas, comunidades microbianas, bem como nas interações tróficas. Encontramos aplicações com modelação de relações macroecológicas e de processos em ecossistemas. Os problemas de sustentabilidade ambiental podem ser objeto de modelação SEM bem como um sem número de outras aplicações no âmbito das Ciências Naturais.

As aplicações da SEM verificam-se noutras áreas da Biologia para além das referidas, nomeadamente na modelação de processos evolutivos e na genética.

Enquanto a SEM tem sido mais comumente aplicada em **estudos observacionais** (p. ex. Iriondo, Albert & Escudero, 2003; Grace, 2008; Budtz-Jørgensen, 2010), tem havido inúmeras aplicações envolvendo **manipulações experimentais** (p. ex. Tonsor & Scheiner, 2007; Lamb & Cahill, 2008; Youngblood, Grace & McIver, 2009; Morrison, Morrison & McCutcheon, 2017). Até hoje, relativamente poucos estudos usaram **métodos *Bayesianos*** nas aplicações da SEM no contexto da Biologia e das Ciências Naturais (p. ex. Arhonditsis *et al.*, 2006; Grace, Harrison & Damschen, 2011; Gimenez, Anker-Nilssen & Grosbois, 2012).

2.1.1. Aplicações na Ecologia

Em investigação científica recorre-se com frequência a Modelação de Equações Estruturais, nomeadamente para testar e avaliar relações causais multivariadas nas ciências naturais.

Como é possível constatar com facilidade numa pesquisa *online*, existem múltiplas aplicações da SEM na literatura, nomeadamente na Ecologia e na Biologia. Grace *et al*. (2010), Eisenhauer *et al*. (2015) e Fan *et al*. (2016) fornecem uma revisão abrangente das aplicações de SEM em estudos ecológicos, em biologia e em ciências ambientais.

Segue-se a explicitação de exemplos da literatura, alguns referidos atrás, com enfoque na SEM e não nos conceitos específicos das áreas afetas.

Gimenez, Anker-Nilssen e Grosbois (2012), através da análise de caminhos, investigam fatores que impulsionam a variação no espaço e no tempo de parâmetros demográficos de populações naturais de animais e plantas, para avaliar a importância relativa de variáveis bióticas e abióticas na modelação da dinâmica de uma população; Detilleux *et al.* (2013) usam a SEM para quantificar o risco latente de infeção animal e estimar níveis diretos e indiretos de tolerância de animais infetados naturalmente por agentes patogénicos.

Arhonditsis *et al*. (2006) apresentam dois estudos de caso em que exploram estruturas ecológicas através de modelos de equações estruturais que descrevem a dinâmica de uma comunidade de fitoplâncton resultante da interação entre fatores físicos, químicos e biológicos. Aplicam a SEM numa abordagem *Bayesiana*, com o objetivo principal de discutir como a SEM pode ser combinada com a análise *Bayesiana* para auxiliar na gestão de recursos naturais. Cao *et al*. (2017) usam uma forma de sequenciamento de genes RNA para estudar as associações de comunidades bacterianas em termos de características ambientais, bem como mudanças ecológicas que influenciam a estrutura da comunidade bacteriana nos processos ecológicos, ao longo de um conjunto de variáveis, com base em dois modelos hipotéticos de equações estruturais, no mesmo tipo de ecossistema aquático.

Hodapp *et al*. (2015) aplicam um modelo SEM para monitorizar dados sobre comunidades de *fitoplâncton* marinho, incluindo dados sobre parâmetros ambientais, estrutura da comunidade e medidas de produtividade. A aplicação de um modelo SEM à análise de relações BEF (*Biodiversity−ecosystem functioning*) numa comunidade fitoplantónica natural mostra que os métodos multivariados não são apenas uma ferramenta adequada, mas também altamente recomendável quando se investiga essas complexas redes de interações.

Blüthgen *et al.* (2016) usam a SEM para testar hipóteses sobre a cadeia alimentar baseada no fitoplâncton e a sua fraca ligação com a cadeia alimentar baseada em partículas orgânicas /algas bentónicas, com dados do ecossistema do Great Salt Lake (Utah).

Li *et al*. (2017) usam uma abordagem de modelação de equações estruturais para quantificar a eficácia da restauração ecológica e os impactos de diferentes fatores socioeconómicos. Comber *et al*. (2017), na sequência do trabalho anterior, desenvolvem e

aplicam um modelo de equações estruturais baseado numa extensão metodológica da SEM a modelos de ponderação geográfica para estudar os fatores observados e latentes associados à restauração efetiva da paisagem. Exploram a heterogeneidade espacial no que respeita a processos e relacionamentos, desenvolvendo uma série de análises locais em vez de adotar uma abordagem de análise global.

Lam, Shirtliffe e May (2011) ilustram a aplicação da SEM, com uma gama diversa de modelos de variáveis latentes, a um conjunto dados agronómicos de vários locais e mostram que, regra geral, esta metodologia proporciona conhecimentos que uma análise univariada padrão não revela. Sugerem que esta metodologia seja usada pelos investigadores na área das plantas, para estudar processos ou mecanismos subjacentes às relações num grupo de variáveis intercorrelacionadas, nomeadamente para dados provenientes de estudos observacionais e experimentais desta área, onde não é possível um controlo experimental claro de múltiplas variáveis intercorrelacionadas. Lam e Maguire (2012) ilustram a aplicação da SEM na gestão florestal, com o objetivo de ajudar na investigação e compreensão de mecanismos causais entre componentes e processos estruturais que operam simultaneamente através de caminhos complexos e muitas vezes indiretos. Lam *et al.* (2014) propõem um método SEM espacialmente explícito, para o qual concebem um *package* de aplicação no R (`sesem`). Aplicam o método em três estudos, examinando as relações entre fatores ambientais, a estrutura da comunidade de plantas, a fixação de nitrogênio e a competição de plantas.

Joseph, Preston e Johnson (2016) combinam modelos de equações estruturais e modelos de ocupação (MacKenzie, Nichols & Lachman, 2002) para investigar influências complexas na ocorrência de espécies, abordagem que facilita uma representação mais mecanicista das ideias sobre as causas das distribuições de espécies no espaço e no tempo. De facto, a distribuição de espécies é direta e indiretamente influenciada por um conjunto diverso de fatores abióticos e bióticos e a SEM fornece uma estrutura que permite esclarecer, representar e avaliar hipóteses em ecologia que podem superar os métodos estatísticos tradicionais e os modelos de ocupação que não têm em conta a intercorrelação entre as covariáveis e assumem-nas independentes. Ilustram a metodologia num estudo de caso com uma amostra de anfíbios reprodutores de lagos para obter uma melhor compreensão dos fatores determinantes da composição da comunidade.

Laliberté, Zemunik e Turner (2014) usam a SEM para testar simultaneamente diversas teorias sobre os mecanismos que moldam a diversidade de plantas ao longo dos gradientes de recursos numa cronossequência de dunas de 2 milhões de anos. Os resultados da análise sugerem que a diversidade é determinada pela filtragem ambiental da flora regional e não pela competição por recursos, como teorias proeminentes enfatizam.

Tedersoo *et al.* (2014) usam a SEM para testar os efeitos diretos das variáveis climáticas sobre a riqueza de fungos e os seus grupos funcionais e os efeitos indiretos climáticos (via nutrientes do solo e vegetação), em conjunto com uma grande diversidade de ferramentas estatísticas.

Villarreal Ruiz *et al*. (2014) usam dados de áreas geográficas muito diversas e modelos de equações estruturais para determinar, via nutrientes do solo e vegetação, os efeitos diretos e indiretos do clima sobre a diversidade de fungos, química do solo e vegetação. Obtêm resultados para os padrões biogeográficos fúngicos consistentes com os paradigmas derivados de plantas e animais - os intervalos latitudinais das espécies aumentam em direção aos polos (regra de *Rapoport*) e a diversidade aumenta em direção ao equador.

Grace *et al.* (2016) usam a SEM para estudar como é que a produtividade do ecossistema e a riqueza de espécies estão interrelacionadas e quais são os mecanismos subjacentes que vinculam produtividade e riqueza. Desenvolvem um metamodelo de equações estruturais, com base na literatura sobre diversidade de produtividade, que assimila os constructos teóricos essenciais e as conexões hipotéticas numa rede de expectativas multivariadas.

Cubaynes *et al.* (2012) desenvolvem uma abordagem SEM combinando modelos de equações estruturais com modelos de captura-recaptura (CR-SEM) que permite a investigação de hipóteses concorrentes sobre variabilidade individual e ambiental, observadas em parâmetros demográficos. Usam a amostragem de Monte Carlo via Cadeias de Markov numa estrutura *bayesiana* para estimar parâmetros, selecionar modelos para avaliar hipóteses concorrentes sobre mecanismos causais e avaliar o ajustamento de modelos a dados usando verificações preditivas posteriores. Aplicam esta abordagem em dois estudos de caso com populações de aves selvagens.

He (2013) usa a SEM para testar quatro hipóteses multivariadas envolvendo história evolutiva, distribuição geográfica, diversidade genética e adequação – separar o efeito de

múltiplos fatores de interação, para revelar a força das interações diretas entre esses fatores, e explorar os mecanismos subjacentes aos processos ecológicos e evolutivos que moldam a distribuição geográfica, diversidade genética e adequação das espécies. Para tal usa dados comparativos sobre a história evolutiva, distribuição geográfica, extensão ecológica, diversidade genética e resistência à infeção por agentes patogénicos para espécies de uma planta predominantemente encontrada na Austrália.

Tortorec *et al*. (2013) usam a SEM para examinar as complexas associações hierárquicas entre a perda de *habitat* e a configuração espacial. Investigam os efeitos da fragmentação do *habitat* no desempenho de reprodução individual e mostram as vantagens de uma abordagem SEM como ferramenta para modelar associações ecológicas hierárquicas complexas, tornando possível encontrar associações indiretas que as abordagens estatísticas univariadas comumente usadas não seriam capazes de detetar.

Jing *et al*. (2015) ajustam um modelo de equações estruturais por *piecewise* para inferir sobre os efeitos diretos e indiretos do clima, das propriedades do solo (humidade do solo e pH do solo) e da biodiversidade acima e abaixo do solo na multifuncionalidade dos ecossistemas.

Delgado-Baquerizo *et al*. (2016) usam a SEM para testar se a relação entre a diversidade microbiana e multifuncionalidade é mantida quando se representam vários controladores de multifuncionalidade simultaneamente: efeitos diretos e indiretos do espaço, clima, pH do solo.

Trivedi *et al*. (2016) usam a SEM para identificar a importância relativa e os efeitos de genes funcionais *versus* fatores abióticos (C e pH totais) e a composição microbiana na função do solo (atividades enzimáticas).

Tallavaara, Eronen e Luoto (2017) usam modelação SEM para testar os efeitos potencialmente hierárquicos de produtividade primária líquida, biodiversidade e *stresse* ambiental patogénico na abundância global de caçadores-coletores. A análise revela que a produtividade primária líquida, a biodiversidade e o *stresse* por patogénicos ambientais interagem para impor limitações complexas e variadas na densidade populacional de caçadores-coletores em diferentes partes do mundo.

Bowker, Maestre e Escolar (2010) usam a SEM, recorrendo a indicadores de função do ecossistema relacionados com a hidrologia, a captura e a retenção de recursos do solo e a

ciclagem de nutrientes, aplicada a quatro conjuntos de dados observacionais para pesquisar sobre o papel da biodiversidade do solo na função dos ecossistemas.

You *et al.* (2014) usam a SEM para estabelecer e testar conexões hipotéticas entre os fatores bióticos e abióticos locais e atributos estruturais das comunidades microbianas do solo, bem como a ligação dos tipos de comunidade microbiana do solo com a função representada pelas atividades das enzimas extracelulares do solo.

Viswanath *et al.* (2015) desenvolvem modelos lineares multivariados para predizer a quantidade total de sólidos dissolvidos em termos de diferentes parâmetros físico-químicos de águas subterrâneas e aplicam a SEM para validar o modelo desenvolvido.

Ryberg (2017) usa a SEM para estudar fatores que influenciam a qualidade da água da bacia do Rio Vermelho. Para tal desenvolve potenciais modelos SEM para fatores que influenciam o peso total de fósforo na bacia do rio Vermelho, com base no conhecimento prévio da bacia e do ciclo de fósforo. Usa nos modelos uma variável latente representando práticas de gestão agrícola, que indica que as práticas agrícolas influenciam diretamente o peso anual total de fósforo no rio.

Prugh e Brashares (2012) usam a SEM para separar correlações e estimar a importância relativa dos efeitos de variáveis relacionadas com a engenharia dos ecossistemas (levada a cabo por organismos que "modulam a disponibilidade de recursos (que não são eles mesmos) para outras espécies, causando mudanças físicas de estado em materiais bióticos ou abióticos) e efeitos de variáveis não relacionadas com a engenharia dos ecossistemas na estrutura da comunidade, em comunidades coocorrentes num mesmo ecossistema. O estudo é aplicado para quantificar e particionar os efeitos da engenharia e da não-engenharia do rato canguru gigante (*Dipodomys ingens*) em espécies coocorrentes (em plantas, invertebrados e vertebrados) num ecossistema de pastagem semiárida.

Crouch e Mason-Gamer (2018) avaliam como a SEM ajuda a testar correlações entre variáveis que interagem em sistemas biológicos complexos, aplicando a metodologia num caso de estudo relacionado com a reprodução aviária.

Byun, de Blois e Brisson (2015) usam a SEM para testar diversas hipóteses com relações causais entre fatores que influenciam o sucesso de uma espécie durante o processo de invasão. Bowen *et al.* (2017) usam a SEM, com um modelo misto, para avaliar as

interações planta-micróbio, que desempenham papéis cruciais nos processos invasivos das espécies e que ilustram um estudo destas interações ao nível intraespecífico.

Dorresteijn *et al*. (2015) usam a *piecewise* SEM para estudar como a regulação dos ecossistemas pelos superpredadores é influenciada pelas atividades humanas.

Shao *et al.* (2015) usam a SEM para explorar as interações complexas entre múltiplos níveis tróficos nas teias alimentares do solo e ilustram o papel que esta metodologia pode desempenhar no entendimento das interações complexas e vias de energia em redes alimentares do solo num contexto multivariado, nos testes da teoria estabelecida e para propor novos testes experimentais.

Mora (2017) usa a SEM numa abordagem que integra sistemas de informações geográficas como uma forma de modelar a integridade ecológica como variável espacial latente. A SEM é usada para estabelecer um vínculo hipotético entre estrutura e função em ecossistemas, baseado na interação das variáveis espaciais usadas para definir vários conceitos.

Mortensen (2016) aplicam a SEM para avaliar hipóteses sobre padrões de interação direta e indireta do sistema consumidor-recurso e para usar variáveis latentes. Foi aplicada a um conjunto de dados de longo prazo de um ecossistema do Ártico Alto para analisar como as respostas fenológicas entre três níveis tróficos são acopladas a padrões de derretimento de neve e como as mudanças se podem espalhar através de interações consumidor-recurso.

Eisenhauer *et al*. (2015) fornecem alguns exemplos de como a SEM pode ser usada por ecologistas do solo para mudar o foco da descrição de padrões para o desenvolvimento do entendimento causal e para inspirar novos tipos de testes experimentais nesta área.

Capmourteres e Anand (2016) usam SEM para testar potenciais indicadores para uma variável latente que designam por "*Habitat Function*" e que representa a capacidade que um ecossistema tem de fornecer refúgio e *habitat* de reprodução a espécies selvagens de plantas e animais. Testam também hipóteses sobre a relação entre funções de *habitat* e componentes estruturais múlti-escala e procuram obter informações sobre a importância relativa de várias variáveis que podem influenciar funções de *habitat*.

Jiao *et al.* (2016) usam a SEM para medir o desempenho de urbanização sustentável, em termos económicos, sociais, ambientais e de recursos naturais.

Mardani *et al.*, (2016) apresentam uma revisão abrangente da aplicação da SEM em várias áreas de sustentabilidade ambiental, em artigos publicados entre 2005 e 2016.

2.1.2. Aplicações na Genética

A SEM é uma versão estendida da análise de caminhos de Wright (1921), com amplas aplicações à genética, pois oferece uma poderosa ferramenta para a modelação de redes em diversos contextos. Esta metodologia tem um número significativo de aplicações em redes biológicas, nomeadamente, na inferência de redes de fenótipo causal, em estudos de associação genómica ampla (GWAS - ***G****enome-****W****ide* ***A****ssociation* ***S****tudy*) e interações gene-ambiente, bem como para medir os efeitos de *loci* de características quantitativas (*Quantitative Traits Loci* - *QTLs*) em análises de ligação. Num complexo genótipo-fenótipo, envolvendo muitas características, uma determinada característica pode ser influenciada não apenas por fatores genéticos e sistemáticos, mas também por outras características (como covariáveis). Os *QTLs* podem não afetar diretamente o traço alvo, mas os efeitos podem ser mediados por traços a montante numa rede causal. Os efeitos indiretos podem, portanto, constituir uma proporção percebida de múltiplos efeitos dos genes, e esses conceitos aplicam-se a conjuntos de traços hereditários, organizados como redes, comuns em sistemas biológicos. Estas e muitas outras questões do âmbito da genética são objeto de estudo com recurso à SEM.

Neale e Cardon (1992) usam a SEM (LISREL) em estudos genéticos com gémeos, Posthuma *et al.* (2004) e van den Oord (2000) usam a SEM para identificar *QTLs* em estudos de associação em famílias nucleares de tamanho variável. Medland e Neale (2010) usam a SEM em estudos de associação alélica, com dados de famílias e sintetizam a metodologia para estudos genéticos com gémeos, assente na SEM, mais especificamente com o modelo LISREL. Morris, Elston e Stein (2010) propõem uma estrutura para a implementação da SEM em dados familiares que permite uma ampla variedade de modelos, podendo incluir componentes ambientais, poligénicas e outras componentes de variância genética.

Xiong, Li e Fang (2004) são os primeiros a aplicar a SEM à reconstrução de redes genéticas utilizando dados de expressão genética. No entanto, a sua aplicação foi limitada a

redes genéticas sem relações cíclicas, usando um SEM recursivo, que possui uma estrutura acíclica e erros não correlacionados e é equivalente a uma rede *bayesiana* gaussiana.

Liu, La Fuente e Hoeschele (2008) implementam a SEM para inferir uma rede reguladora de genes usando apenas traços de expressão (*etraits*). O método é avaliado num conjunto de dados simulados com estruturas de rede subjacentes conhecidas, e num conjunto de dados reais de levedura.

Rosa *et al.* (2011, 2016) apresentam algumas aplicações da SEM na inferência de redes fenotípicas causais, nomeadamente na reconstrução de redes fenotípicas em estudos genómicos genéticos, nos quais estão disponíveis tanto informações fenotípicas quanto moleculares, bem como no contexto da análise genética quantitativa clássica de múltiplas características fenotípicas, usando apenas informações fenotípicas e de *pedigree*.

Valente *et al.* (2013) discutem e investigam a vantagem do recurso à SEM de efeitos mistos, em comparação com os modelos *multitrait model* (MTM) *standard*, como ferramentas auxiliares na tomada de decisão em programas de melhoramento.

Cai, Bazerque e Giannakis (2013) usam uma abordagem SEM para integrar dados de expressão génica e *QTLs* de *cis*-expressão (*cis*-eQTL), para modelar redes reguladoras de genes, de acordo com evidências biológicas sobre genes reguladores ou regulados por pequeno número de genes.

Pepe e Grassi (2014) usam a SEM para investigar perfis de expressão genética, considerando, não apenas, genes desregulados, mas também as conexões entre os genes perturbados.

Tao, Sánchez e Mukherjee (2015) propõem uma estratégia de modelação SEM para examinar o impacto conjunto de genes e medidas de exposição múltipla nos resultados de saúde, medidos repetidamente ao longo do tempo.

Peñagaricano *et al.* (2015) usam a SEM com dados de vários fenótipos em suínos, para avaliar redes causais envolvendo variáveis latentes subjacentes a características fenotípicas complexas.

Song, Morris e Stein (2016) exploram diferentes técnicas SEM usando o *package* `strum` para analisar dados longitudinais multivariados e testar a associação de genótipos em características de pressão arterial.

Huisman *et al.* (2018) usam a SEM para encontrar variáveis latentes que explicam as alterações do volume cerebral num contexto de doença e que, por sua vez, são afetadas por variantes genéticas. Integram marcadores genéticos (polimorfismos de nucleótido único – SNP (*Single Nucleotide Polymorphism*)) e características de imagem e testam o modelo num ambiente de simulação aplicado num estudo de caso relacionado com neuroimagem da doença de Alzheimer.

Kim, Namkung e Park (2010) aplicam a SEM a dados de associação genética ampla (*Genome-wise association studies* – GWAS) para a modelação de relações complexas entre redes genéticas e características (*traits*) como fatores de risco (fatores de risco associados à obesidade na parte central do corpo). Defendem que a SEM permite alcançar uma melhor compreensão dos mecanismos biológicos, através da identificação de um maior número de genes e de *pathways* que estão associados a um conjunto de características e à relação entre eles.

Nuzhdin, Friesen e McIntyre (2012) propõem uma abordagem SEM para estudos GWAS que conectam estatisticamente genótipos a fenótipos, recorrendo a interações moleculares conhecidas, para estabelecer a ligação da função do gene ao fenótipo através de redes reguladoras de genes (*gene regulatory networks*-GRNs). Esta abordagem aproveita a miríade de polimorfismos em populações naturais para elucidar e quantificar as vias moleculares subjacentes à variação fenotípica. De acordo com os autores, a abordagem desenvolvida permite colocar num único modelo todos os efeitos da segregação de polimorfismos *cis*-reguladores e efeitos *trans*, permitindo assim que o fenótipo seja previsto.

Song *et al*. (2015,2016) desenvolvem um pacote para o R (`strum`) que permite implementar uma estrutura SEM para dados gerais de *pedigree* (Song *et al.* 2015) e cujas ferramentas permitem ajustar e simular uma ampla gama de modelos de equações estruturais com variáveis latentes e covariáveis. Recorrendo a este pacote, os autores usam dados de GWAS para testar a associação entre genótipo e características de pressão arterial e usam as covariáveis idade, sexo *e status* de fumador (Song *et al.* 2016).

Verhuls *et al*. (2017) apresentam um método para testar a associação de um PNU com múltiplos fenótipos ou um constructo latente, numa base genómica. Implementam este método num pacote do R – `GW-SEM`.

Grotzinger *et al*. (2018) desenvolvem um pacote para o R - *GenomicSEM* onde implementam um método multivariado (*Genomic* SEM) para analisar as arquiteturas genéticas conjuntas de características complexas, que sintetiza correlações genéticas e herdabilidades de SNPs, inferidos a partir de estatísticas sumárias de GWAS de características individuais de amostras com graus variados e desconhecidos de sobreposição, em contraste com os métodos que modelam a estrutura de covariância fenotípica usando dados brutos. A metodologia permite ao utilizador especificar e comparar uma gama de diferentes arquiteturas genéticas multivariadas, o que melhora as abordagens existentes para combinar informações através de características geneticamente correlacionadas que auxiliam na descoberta. Demonstram várias aplicações da SEM Genómica.

De los Campos, Gianola e Heringstad (2006) usam a SEM para estudar a relação entre a produção de leite e o *score* de células somáticas em cabras leiteiras.

Li *et al.* (2006) aplicam a SEM às características de tamanho corporal, adiposidade e geometria óssea para estudar como os efeitos de um *locus* genético podem ser decompostos ao longo de caminhos diretos e indiretos que podem ser mediados por interações com outras características.

Tsanousa *et al.* (2013) usam a SEM para investigar estatisticamente a via de sinalização do Recetor Toll-like (TLR - *Toll-like Receptor*) na Leucemia Linfocítica Crónica (LLC).

Cheung *et al.* (2013) usam a SEM para investigar as relações entre o metabolismo da glicose, componentes da síndrome metabólica e fosfatase alcalina específica do osso, bem como os efeitos diretos e indiretos de covariáveis como a hipertensão, colesterol HDL baixo, resistência à insulina, obesidade e inflamação podem ter na fosfatase alcalina específica do osso.

Nock *et al.* (2009) aplicam a SEM a dados de um estudo de controlo dos pólipos do colon para ilustrar matematicamente relações hierárquicas entre fatores genéticos e ambientais envolvidos em sistemas biológicos subjacentes que conduzem a doenças complexas. Tanaka *et al.* (2010) usam a SEM para estudar a estrutura de correlações do fenótipo metilador das ilhas CpG's, da metilação das ilhas CpG's e da instabilidade de microssatélites, em função do *status* de dois oncogenes (KRAS e BRAF) associados ao desenvolvimento do cancro colorretal. Mi *et al.* (2011) desenvolvem uma abordagem SEM

generalizada (em dois níveis) para identificar genes e interações gene-meio ambiente, no progresso da doença arterial coronária, tendo em conta a estrutura causal entre fatores de risco mediadores e a doença arterial coronária com os indivíduos no primeiro nível e com as famílias no segundo nível. Warrington *et al.* (2017) usam a SEM, em dados simulados e em dados reais de peso de bebés ao nascer, para estimar os efeitos maternos e fetais quando a informação fenotípica está presente para indivíduos em duas gerações e informações sobre genótipos estão disponíveis no indivíduo mais velho.

2.1.3. Aplicações na Agricultura

Smith *et al.* (2014) discutem a utilização da SEM como uma ferramenta de pensamento visual e metodologia estatística que pode melhorar o resultado de projetos de pesquisa integrada e transdisciplinar. A metodologia é apresentada como fornecendo um quadro comum entre disciplinas, facilitando o refinamento constante de hipóteses e métodos e promovendo a descoberta de novas questões e relacionamentos. Aplicam a SEM a uma investigação de campo multidisciplinar e em vários locais sobre a adaptação e mitigação de mudanças climáticas em agroecossistemas de culturas em linha, usando sistemas de gerenciamento zonal de precisão.

Brahim *et al.* (2011) desenvolvem dois modelos SEM para quantificar o carbono orgânico sob solos argilosos e arenosos em zonas semiáridas do Mediterrâneo, baseando-se em propriedades físicas e químicas do solo.

Ibrahim, Hatira e Gallali (2013) usam regressões lineares múltiplas para avaliar as relações entre o nitrogênio e as propriedades do solo e usam a SEM para investigar simultaneamente as interações entre os diferentes componentes das propriedades do solo e as suas relações com o nitrogênio.

Nazmi (2013) usa a SEM para estudar as interações simultâneas entre variáveis relacionadas com propriedades químicas do solo (variável latente), variáveis relacionadas com propriedades físicas (variável latente) e a produção de trigo (variável latente).

Crittenden e de Goede (2016) usam a SEM para modelar relações causais por meio de modelos que representam hipóteses dessas relações causais relacionadas com a qualidade biofísica do solo dos sistemas de cultivo sob sistemas de agricultura orgânicos e convencionais, correlacionando os dados físicos e biológicos.

Angelini, Heuvelink e Kempen (2017) usam a SEM para prever propriedades do solo em múltiplas camadas, considerando as inter-relações entre as propriedades e diferentes camadas do solo.

Zhang *et al.* (2017) usam a SEM para obter uma compreensão mecanicista da forma como a diversidade microbiana do solo altera a bioquímica do solo em ecossistemas de pastagem alpina, a partir da deteção dos efeitos diretos e indiretos entre variáveis.

Sharifzadeh *et al.* (2012) aplicam uma Teoria do Comportamento Planeado para prever a intenção comportamental e o comportamento real no uso voluntário de previsões de informações climáticas no apoio às decisões agrícolas e esta utilização é abordada, justificada e testada empiricamente com recurso à SEM.

Shadfar e Malekmohammadi (2013) usam a SEM para analisar as políticas do estado em relação ao desenvolvimento da produção de arroz nos principais países produtores de arroz.

Azadi *et al.* (2016) usam a SEM para identificar interações entre diferentes fatores (económicos, políticos, ambientais, biofísicos, institucionais e culturais) que condicionam a conversão de terras agrícolas no Irão.

Lamb, Shirtliffe e May (2011) apresentam aos cientistas de plantas os princípios e práticas da SEM, usando como exemplo um teste de campo agronómico. Demonstram a utilização da SEM com variáveis observadas e latentes usando uma experiência *multi-site* de campo, examinando os efeitos do tamanho da semente e da densidade de sementeira na densidade de plantas e no rendimento de aveia em Saskatchewan.

Zhang *et al.* (2014) usam a SEM para avaliar a importância relativa de múltiplos fatores que influenciam o rendimento e a produtividade na produção do linho, numa única análise abrangente, para permitir definir estratégias de melhoramento.

Mańkowski, Kozdój e Janaszek-Mańkowska (2016) usam a SEM para descrever e caracterizar as relações entre fatores produtivos e produtividade de grãos por planta duplo-haplóide de cevada, bem como a relação entre os componentes de produção e a duração de cada estágio de desenvolvimento da planta.

Cerda *et al.* (2017) aplicam o delineamento experimental e a SEM para quantificar as perdas de produtividade primária (incorridas no ano atual de produção) e as perdas de

rendimento secundárias (resultantes dos impactos negativos do ano anterior) na produção de café devido a pragas e doenças e para identificar os preditores mais importantes dos rendimentos e perdas de rendimento.

2.1.4. Aplicações nas Ciências Médicas, Epidemiologia e Saúde Pública

Beran e Violato (2010) discutem os possíveis contributos da SEM para a pesquisa médica e as ciências da saúde e proporcionam uma revisão da implementação em epidemiologia e pesquisa médica. Hays, Revicki e Coyne (2005), além de apresentarem dois exemplos de aplicação da SEM à pesquisa de resultados de saúde, discutem os prós e os contras sobre a aplicação de SEM na investigação neste contexto. Merchant *et al.* (2013) discutem a aplicação da SEM na pesquisa em reabilitação. Christ *et al.* (2014) discutem a abordagem SEM para modelação de dados de saúde ocular.

Chavance *et al.* (2010) aplicam a SEM a dados de um estudo longitudinal sobre a restrição da alimentação como fator de risco para o aumento de peso. Vilhena *et al.* (2014) usam a SEM para avaliar o impacto simultâneo de vários fatores psicossociais na qualidade de vida, em pacientes obesos portugueses e o papel mediador do estigma na relação entre afeto positivo/negativo e qualidade de vida.

Haber, Ahmed e Pekovic (2012) usam a SEM para estudar a força da associação entre a história familiar de cancro da mama e história familiar de outros cancros, com a perceção de risco de cancro da mama e a repetição da mamografia. Ma *et al.* (2013) usam a SEM para estudar e avaliar a relação entre o rastreamento do cancro da mama com fatores socioculturais de comportamento em relação à saúde, entre as mulheres chinesas que vivem nos EUA.

Green *et al.* (2012) usam a SEM para identificar fatores que diretamente predizem, moderem ou medeiam a obesidade em sobreviventes adultos a cancro na infância, para definir estratégias comportamentais que reduzam o risco de obesidade pós-terapia.

Rao *et al.* (2012) usam a SEM para estudar associações entre o estigma relacionado ao HIV, os sintomas depressivos e a adesão à medicação para HIV entre pessoas que vivem com o HIV em cuidados clínicos de rotina. Yoo-Jeong *et al.* (2016) usam a SEM para modelar o processo pelo qual os sintomas relacionados ao HIV estão relacionados com a

adesão à terapia antirretroviral, examinando o sintoma de depressão como mediador dessa relação.

Lee *et al.* (2017) usam a SEM para avaliar as relações entre as variáveis função física, fatores sociais, fatores psicológicos e a qualidade de vida dos doentes com fibromialgia pela primeira vez, no que respeita à saúde, bem como os efeitos dessas variáveis na qualidade de vida de doentes. Lee *et al.* (2016) usam a SEM para determinar as relações hipotéticas entre alfabetização em saúde, autoeficácia, atividades de autocuidado e a qualidade de vida, no que respeita à saúde, em pacientes com diabetes tipo 2.

Ahn (2017) usa a SEM para construir e testar um modelo hipotético incluindo fatores relacionados com competência cultural de enfermeiros que cuidam de pacientes estrangeiros. Lee e Yom. (2013) usam um planeamento transversal e a SEM para analisar as relações entre variáveis relacionadas com a reintegração social e no quotidiano de pessoas com queimaduras graves.

2.1.5. Aplicações nas Neurociências

Erickson *et al.* (2005) usam a SEM para avaliar as relações entre as regiões do cérebro envolvidas no controle da atenção. Esta análise específica complementa e acrescenta análises comuns de Modelos Lineares Generalizados de dados de neuroimagem que são tipicamente limitados a investigar a magnitude e a extensão da ativação neural.

Kim e Horwitz (2009), usando dados de fMRI (*functional Magnetic Resonance Image*) simulados a partir de um modelo de rede neurobiologicamente realista para investigar de que forma a SEM pode ser usada para examinar as diferenças na conectividade efetiva em distúrbios da conectividade cerebral.

Inmanl (2012) usa a SEM com dados de ressonâncias magnéticas feitas a sobreviventes de AVC em estado de repouso, para investigar a relação entre os *déficits* motores e a conectividade intrínseca efetiva entre as regiões cerebrais envolvidas no controle motor e na execução motora.

Kievit *et al*. (2012) exploram a SEM como uma metodologia que, com os devidos cuidados, é adequada para estudar a relação entre medidas comportamentais de inteligência geral e medidas neurológicas do cérebro, uma vez que possibilita a investigação de hipóteses conceptuais sobre a relação entre a inteligência e o cérebro.

Kievit *et al*. (2014), Nielsen e Wilms (2015) e Penke *et al.* (2012) usam a SEM para estudar relações entre variáveis globais do cérebro e diferenças individuais em medidas cognitivas associadas à idade, e Ritchie *et al*. (2015) usam a SEM para estudar a relação entre as competências cognitivas e o volume do cérebro.

Moreira *et al.* (2016) combinam dois procedimentos distintos de modelação SEM – autorregressivo e crescimento latente – para implementarem um modelo ATL (*Autoregressive Latent Trajectory*), para estudarem a aprendizagem de referência espacial através da análise do teste do Labirinto Aquático de Morris num planeamento experimental complexo envolvendo quatro fatores.

Kievit *et al.* (2017a) exploram uma classe de modelos SEM que designam por *Latent Growth Curve* e que apresentam como especialmente versátil e útil para os investigadores em neurociência cognitiva do desenvolvimento, uma vez que pode modelar mudanças no nível de construção, pode ser usada com um número relativamente pequeno de pontos de tempo e é especialmente poderosa para testar o comportamento cerebral. Ilustram a metodologia com dois estudos empíricos.

2.2. Aplicações nas Ciências Sociais

As aplicações da SEM começam predominantemente nas Ciências Sociais, nomeadamente na Psicometria e na Econometria, como anteriormente referido. Em particular, referem-se nesta trabalho alguns artigos relacionados com as aplicações da SEM na pesquisa em Psicologia e na Educação.

Tremblay e Gardner (1996) e Hershberger (2003) ilustram a enorme quantidade de artigos publicados com a SEM com aplicações à Psicologia e à Educação no período de 1984 a 2001.

MacCallum e Austin (2000) apresentam uma revisão de aplicações da SEM publicadas em revistas de pesquisa psicológica, centrada na diversidade de projetos de pesquisa e questões substantivas às quais a SEM pode ser aplicada de forma produtiva e nos vários problemas metodológicos e questões preocupantes que na sua perspetiva caracterizam parte dessa literatura.

Karimi e Meyer (2014) traçam a história da SEM no campo da psicologia e discutem os desenvolvimentos da metodologia bem como as diversas técnicas e a sua utilização nesta

área. Morrison, Morrison e McCutcheon (2017) fornecem uma breve visão geral da metodologia e proporcionam uma série de recomendações sobre as melhores práticas para testar modelos (antes e durante os testes) e relatar descobertas.

Merkle e Wang (2016) abordam a inferência com a SEM para extrair inferências de dados de psicologia experimental. Aplicam a SEM para reanalisar dados experimentais, comparando esta abordagem a métodos alternativos mais simples.

van de Schoot *et al.* (2017) propõem-se fornecer uma apresentação completa do papel que as estatísticas *Bayesianas* desempenham na pesquisa em psicologia, e em particular da SEM *Bayesiana*. Para tal, realizam uma revisão sistemática de artigos com estatística *Bayesiana* aplicada em estudos em Psicologia, publicados entre 1990 e 2015, nomeadamente de artigos em que é utilizada a SEM *Bayesiana* e discutem uma série de questões que surgem na sua aplicação.

Martens e Haase (2006) discutem três modelos de SEM, dois para análises longitudinais - projetos de painéis de desfasamento cruzado e modelos de curva de crescimento latente e um para testar modelos não-recursivos e ilustram a sua aplicação no contexto da psicologia do aconselhamento.

Ravens-Sieberer *et al.* (2009) usam a SEM para analisar a relação entre escola e satisfação com a vida, parcialmente mediada pela saúde emocional, seguindo um modelo que assume alguns pressupostos que relacionam as perceções escolares com o ajustamento escolar dos alunos e a satisfação com a vida mediada pela saúde emocional.

Schreibert *et al.* (2006), além de fornecerem uma introdução à SEM, apresentam exemplos de aplicação na Educação e, fazem uma revisão crítica de artigos publicados no *Journal of Educational Research* entre 1989 e 2004 sobre comportamento em diferentes níveis escolares.

Teo, Tsai e Yang (2013) fornecem uma introdução não técnica às várias facetas da SEM para investigadores em educação.

Khine (2013) proporciona uma série de exemplos e teorias internacionais para ilustrar as aplicações do SEM na pesquisa e na prática educacional.

Goldstein, Bonnet e Rocher (2007) usam a SEM multinível, aplicada a dados do PISA (2000, 2003, 2006), para explorar dados com uma estrutura complexa e comparar os

sistemas educativos de dois países, França e Inglaterra. Rabe-Hesketh, Skrondal e Zheng (2007) usam a SEM multinível, com dados do PISA 2000 dos EUA, para investigar a relação entre a variável latente '*excelência do professor'* ao nível da escola e a variável latente *'capacidade de leitu*ra' ao nível do aluno, cada uma medida por múltiplos indicadores ordinais. Babenko, Alves e Bahry (2012) usam a SEM com dados do PISA 2006 (EUA e Canadá), para avaliar as conexões entre os fatores de educação científica e conscientização dos alunos em relação à carreira na ciência e perspetivas de emprego. Sari (2015) usa a SEM em dados do PISA 2009, para determinar fatores que afetam as habilidades de leitura. Caro, Sandoval-Hernández e Lüdtke (2014) usam a SEM para avaliar constructos de capital cultural, económico e social, inspirados em teorias bem estabelecidas (Bourdieu & Passeron (1977), Bernstein (1975) e Coleman (1988)), em dados internacionais do PIRLS 2006 e PISA 2009.

Bulut, Delen e Kaya (2012) usam a SEM para criar variáveis latentes para tecnologia de leitura, atitude em relação à leitura e autorregulação e para estimar os coeficientes de caminho entre essas variáveis latentes e os desempenhos na leitura.

Phiakoksong, Niwattanakul e Angskun (2013) usam a SEM para explorar os principais fatores que afetam a qualidade do processo de ensino.

Afari (2013) usa a SEM para investigar os efeitos psicossociais do ambiente de sala de aula no aproveitamento de aulas de matemática pelos alunos e a autoeficácia acadêmica na aprendizagem de matemática nos Emirados Árabes Unidos. Borhan e Zakaria (2017) usam a SEM para determinar a relação entre crenças matemáticas e a atitude em relação às práticas de ensino de matemática com a matemática dos professores iniciantes na Malásia.

Hannula *et al.* (2014) usam a SEM com dados longitudinais nacionalmente representativos de resultados de aprendizagem de matemática na Finlândia, a fim de determinarem a direção da causalidade entre afeto e realização relacionada com a matemática.

A grande profusão do uso da SEM como metodologia de análise de dados também pode ser percebida pelo intenso debate que se dá em torno do tema. Desde 1993 está em funcionamento a SEMNET (*Structural Equation Modeling Network*) (http://www2.gsu.edu/~mkteer/semnet.html), lista de discussão multidisciplinar dedicada a debater os tópicos de interesse de investigadores e criadores dos modelos matemáticos de

SEM, contando, já em 2013, com mais de 3.000 integrantes de dezenas de países (Bollen e Pearl, 2013). Dentro dessa mesma perspetiva foi criado o periódico científico "*Structural Equation Modeling: A Multidisciplinary Journal*", que desde 1994 vem publicando artigos inéditos, empíricos ou teóricos que sejam relacionados com o tema de SEM.

CAPÍTULO 3

MODELOS DE EQUAÇÕES ESTRUTURAIS:

FUNDAMENTOS TEÓRICOS

MODELOS DE EQUAÇÕES ESTRUTURAIS: FUNDAMENTOS TEÓRICOS

3.1. Introdução

Os desenvolvimentos metodológicos da SEM têm dado flexibilidade crescente à metodologia e uma abrangência cada vez maior no que respeita a modelos e a aplicações.

Num sentido abrangente, os modelos SEM representam traduções de uma série de relações hipotéticas de causa e efeito entre variáveis numa hipótese composta – a função de distribuição de probabilidade conjunta das variáveis não é completamente especificada – referente a padrões de dependências estatísticas (Shipley, 2000a). O modelo SEM pode ser de vários tipos, nomeadamente análise de regressão, análise de caminhos, análise fatorial exploratória e análise fatorial confirmatória, análise fatorial de segunda ordem, modelos de estrutura de covariâncias, modelos de estrutura de correlações, equações econométricas simultâneas, modelos de curva de crescimento latente, para citar alguns.

A modelação SEM começa com a definição de um modelo teórico. O que deve guiar o investigador no desenvolvimento de um modelo teórico é a premissa de que a modelação de equações estruturais é baseada em relações causais, isto é, a mudança numa variável inevitavelmente acarretará mudança noutra variável. Convém salientar, no entanto, que nenhum método estatístico, por mais robusto que seja, é capaz de transformar dados transversais (correlacionais) em dados longitudinais (causais). Na interpretação dos dados transversais e do modelo SEM, deve-se trabalhar com a ideia de preditor *versus* consequência e não exatamente, causa *versus* efeito, como nas pesquisas longitudinais (Mueler, 1997).

O investigador deverá ter um conhecimento profundo do tema que investiga para determinar, no modelo, que variáveis são dependentes (consequência) e que variáveis são independentes (preditoras) (Hair *et al*., 2010). Esse cuidado assegurará que sejam respeitados os quatro critérios para o pressuposto de causalidade estabelecido pela SEM: (1) associação suficiente entre duas variáveis; (2) evidências anteriores de causa *versus* efeito; (3) falta de variáveis causais alternativas e (4) uma base teórica para a relação. Nem sempre é possível atender a todos os critérios, mas uma perspetiva teórica sólida permite fazer afirmações de causalidade. Além de possibilitar reconhecer as relações entre as variáveis para atender à causalidade, o conhecimento teórico aprofundado do tema permite que o pesquisador evite erros de especificação, que ocorre quando se omite uma variável

relevante ao modelo, o que causa uma avaliação errónea da importância das demais variáveis e por conseguinte, leva à falta de qualidade no ajustamento do modelo proposto.

O objetivo da análise SEM é determinar até que ponto um modelo teórico é suportado por dados de amostra. Se os dados da amostra suportarem o modelo teórico, então podem ser assumidos e testados modelos teóricos hipotéticos mais complexos, com base no modelo testado. Se os dados da amostra não suportam o modelo teórico, então o modelo original pode ser modificado e seguidamente testado, ou podem ser desenvolvidos e testados novos modelos teóricos. Assim, um dos principais objetivos da SEM consiste em testar modelos teóricos usando o método científico de teste de hipóteses, para se avançar na compreensão de relações complexas entre os constructos (Schumacker & Lomax, 2004).

As relações hipotéticas estabelecidas entre as variáveis são descritas por parâmetros que indicam a magnitude do efeito (direto ou indireto) que variáveis independentes têm sobre variáveis dependentes. A SEM oferece aos investigadores um método abrangente para quantificar e testar modelos teóricos, uma vez que permite traduzir relações hipotéticas em modelos matemáticos testáveis. De facto, proposta uma teoria, esta pode ser testada contra dados empíricos. O processo de testar um modelo teórico proposto é comumente referido como o aspeto "confirmatório" do SEM – CFA (*Confirmatory Factor Analysis*). Outro aspeto importante da SEM é o modo "exploratório". Este aspeto permite o desenvolvimento da teoria e, muitas vezes, envolve repetidas aplicações dos mesmos dados, de modo a explorar potenciais relações entre variáveis de interesse (Pugesek, Tomer & Von Eye, 2003).

Num modelo de equações estruturais podem intervir variáveis que podem ser medidas – variáveis observadas, manifestas ou indicadores, e variáveis que não podendo ser diretamente medidas, são operacionalizadas através das variáveis observadas – variáveis latentes ou constructos.

As variáveis, observadas ou latentes, quanto à influência que uma variável exerce sobre outras, podem ser variáveis exógenas (independentes ou preditoras) ou variáveis endógenas (dependentes). Uma variável exógena é uma variável que não é influenciada ou não sofre efeito de nenhuma outra variável no modelo. Usualmente, são causa de uma ou mais variáveis no modelo. Assume-se que estas variáveis são mensuradas sem erro. Uma variável endógena é uma variável que é influenciada por outras variáveis presentes no

modelo e é sempre acompanhada de um termo residual. É uma variável que é efeito de uma ou mais variáveis no modelo que que pode ser ela própria causa de outra variável endógena no modelo.

Os termos residuais ou de erro que, juntamente com o erro de mensuração, representam as causas omitidas agregadas das variáveis endógenas, podem ser associados a variáveis observadas ou a variáveis latentes, e são especificados como variáveis latentes, correspondendo a outra classe de variáveis em SEM (Kline, 2011).

Um termo residual na variável observada representa a variação não explicada pela variável latente que a variável correspondente deve medir, ou por qualquer outra variável latente, e não covaria com o termo residual de qualquer um dos outros indicadores. Parte dessa variação não explicada deve-se ao erro aleatório de medição ou à falta de fiabilidade (variação na variável que não é explicada por erro de medição e que resulta de imprecisões na representação do conceito teórico pelas variáveis observadas). Outra parte, designada variância específica – não é compartilhada com os demais indicadores da variável latente de interesse – é sistemática e não é relacionada com a variável latente subjacente ou com outra variável latente (Hoyle, 2012). A representação explícita do erro de medição e a utilização de variáveis latentes para o explicar são características especiais do SEM tornando a análise mais realista do que a de outras análises multivariadas que assumem que não há erros de medição.

O termo residual numa variável latente (*disturbance* ou *erro estrutural*) corresponde à variação nessa variável, não atribuível aos indicadores que a definem mas a todas as outras influências não observáveis, diferentes das influências nas outras variáveis latentes e que pode covariar com o termo residual das outras variáveis latentes.

Os erros são considerados variáveis latentes uma vez que a variação do erro deve ser estimada, considerando todo o modelo e os dados; assim, nesse sentido, a variação de erro não é diretamente observável nos dados brutos (Kline, 2011).

Os valores medidos das variáveis observadas constituem o conjunto de dados do investigador.

As variáveis observadas podem ser categóricas, ordinais ou contínuas. Tradicionalmente todas as variáveis latentes na SEM são contínuas, mas há desenvolvimentos recentes com variáveis latentes categóricas (Hoyle, (2012).

A capacidade de analisar variáveis observadas e variáveis latentes distingue a SEM de algumas técnicas estatísticas mais padronizadas, como a análise de variância (ANOVA) e a Regressão Múltipla, que analisam apenas as variáveis observadas. Na sua forma mais simples, a SEM reúne a análise de caminhos e a análise fatorial (confirmatória) num modelo onde estão contidas relações de dependência e é esta abordagem que vai ser caracterizada neste trabalho.

Na abordagem tradicional as relações entre indicadores e variáveis latentes e as relações entre variáveis latentes são avaliadas num único modelo que é definido por dois submodelos – submodelo de medida e submodelo estrutural, que são analisados simultaneamente (Hoyle, 2012). O submodelo estrutural define as relações causais ou de associação hipotéticas entre as variáveis latentes, especificando se uma variável latente causa mudanças noutras variáveis latentes no modelo, direta ou indiretamente. O submodelo de medida – define a forma como as variáveis observadas operacionalizam as variáveis latentes, e constitui-se como uma ligação entre o instrumento de medida (variáveis observadas) e os constructos teóricos em estudo. Na abordagem da SEM com os dois submodelos, o interesse incide na modelação das relações entre as variáveis, sendo analisadas apenas as covariâncias. Neste caso, assume-se que as médias são nulas. Em alguns casos, a hipótese de interesse exige a modelação de padrões de médias de variáveis observadas ou de variáveis latentes, sendo necessário ir além da modelação da estrutura de covariâncias que está subjacente à maioria das aplicações da SEM e considerar modelos que adicionam uma estrutura de médias à estrutura de covariâncias (Hoyle, 2012; Kline, 2011). Os dados de entrada para a análise de um modelo só com estrutura de covariâncias são covariâncias. Num modelo com uma estrutura de médias os dados de entrada são covariâncias e médias. A abordagem SEM para a análise de médias distingue-se de outros métodos multivariados pela capacidade de testar hipóteses sobre as médias das variáveis latentes (Kline, 2011).

Uma questão importante na análise SEM é a dimensão da amostra, sendo clara a necessidade de usar amostras de maior tamanho do que na maioria dos métodos estatísticos multivariados para obter parâmetros estáveis. Fazendo uma análise da literatura sobre os fundamentos teóricos da SEM, no que respeita a este tema, facilmente se constata a dificuldade em encontrar concordância quanto a definir regras gerais práticas, claras e facilmente aplicáveis. Tendo por referência o contexto a análise de regressão multivariada,

há autores que recomendam 5 observações por variável (Hill e Hill, 2009) para garantir variabilidade suficiente para estimar os parâmetros do modelo. No entanto, encontram-se outras recomendações no contexto da análise SEM: Hu e Benttler (1995) recomendam mais de 10 vezes o número de parâmetros livres do modelo, Kline (2011) recomenda que cada amostra contenha de 100-150 observações, Shumacker e Lomax (2004), recomendam de 250 a 500 observações. Estudos de simulação têm sido conduzidos (Wolf *et al*., 2013, Sideridis *et al.*, 2014) e têm revelado, pela diversidade de resultados para diferentes modelos e diferentes graus de complexidade, que a melhor estratégia para selecionar o tamanho da amostra consiste em recorrer à simulação Monte Carlo para decidir, caso a caso, qual o tamanho adequado (Muthén e Muthén, 2002). O pacote `simsem` do R é uma boa alternativa para implementar essa simulação (Jorgensen *et al*., 2018).

Na prática, a SEM é implementada, apesar da generalidade e flexibilidade, seguindo uma sequência de etapas expressas na Figura 3.1. A tracejado estão considerados aspetos que devem ser objeto de preocupação quando são implementadas as etapas principais representadas com linha contínua.

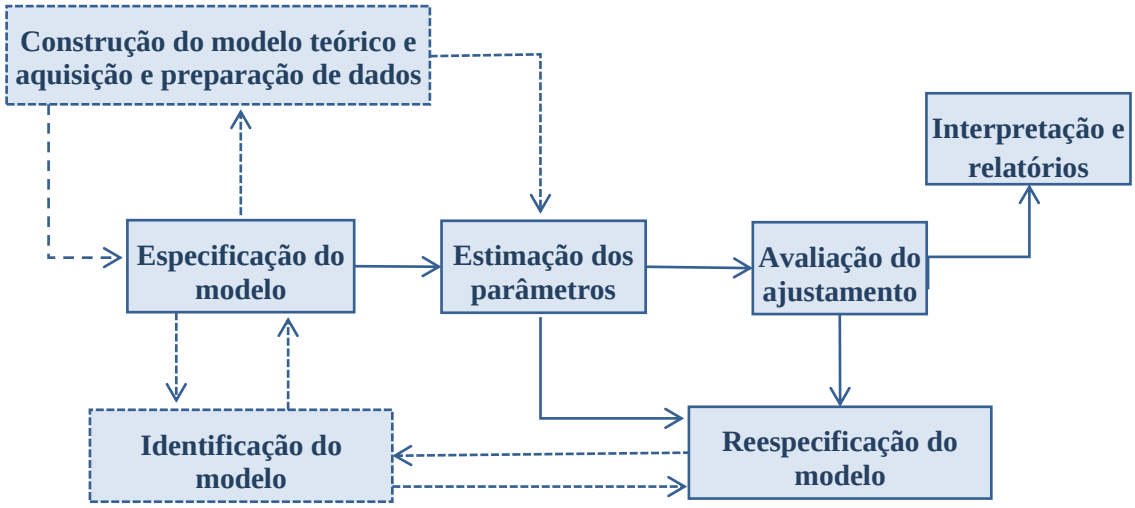


Figura 3.1: Etapas da implementação da SEM (adaptado de Hoyle, 2012).

A definição de um modelo SEM começa com a afirmação da teoria que explicita as relações hipotéticas entre um conjunto de variáveis estudadas (Marcoulides e Schumacker, 2009). As relações hipotéticas podem ter fundamentos teóricos ou resultar da pesquisa empírica do investigador, em particular, da obtenção de relações empíricas resultantes de

uma análise fatorial exploratória de dados de pesquisa, ou ainda de uma combinação dos dois.

3.2. Diagrama de caminhos (*Path Diagram*)

A complexidade dos modelos de equações estruturais justifica a utilização de um esquema visual para representar as relações hipotéticas assumidas. Wright (1921) propôs os designados diagramas de caminhos (*path diagram)* ou grafos orientados, diagramas estes que auxiliam, não apenas, na conceituação e comunicação de modelos teóricos, mas contribuem substancialmente para a criação do arquivo de entrada apropriado, necessário para testar e ajustar o modelo aos dados (Mulaik, 2009).

A título de exemplo, considere-se o modelo representado graficamente na Figura 3.2[1] (obtido com recurso ao pacote `semPLot` do R).

Neste modelo foram usados dados sobre industrialização e democracia em 75 países em vias de desenvolvimento.

O conjunto de dados utilizados é o conjunto de dados Industrialização e Democracia Política que é usado em todo o livro de Bollen (1989). Este conjunto de dados contém quatro medidas de democracia política, em dois momentos, 1960 e 1965, e três medidas de industrialização em 1960.

A base de dados contém 75 observações sobre as 11 variáveis identificadas no Quadro 3.1.

As variáveis y_1 a y_4 pretendem ser indicadores da variável latente Democracia Política em 1960 (D*60*); y_5 até y_8 são indicadores da variável latente Democracia Política em 1965 (D*65*); x_1 a x_3 são indicadores da variável latente Industrialização em 1960 (I*nd*).

Mais adiante será feita a descrição do modelo, nomeadamente as relações causais consideradas, as restrições impostas, não apenas nos pesos fatoriais, mas também nas variâncias/covariâncias dos resíduos, a identificação e a classificação das variáveis e o ajustamento do modelo.

[1] Bollen, K. A. (1989). Structural Equations with Latent Variables. New York. - capítulo 8 citado em "Fox, J. and Weisberg, S. (2012, *last revision*). Ver anexo.

Quadro 3.1: Codificação das variáveis observadas, usadas no modelo Industrialização e Democracia Política.

Variável	Descrição	Ano
y_1	Liberdade de imprensa	1960
y_2	Liberdade de oposição política	1960
y_3	Justiça das eleições	1960
y_4	Efetividade da legislatura eleita	1960
y_5	Liberdade de imprensa	1965
y_6	Liberdade de oposição política	1965
y_7	Justiça das eleições	1965
y_8	Efetividade da legislatura eleita	1965
x_1	PIB *per capita*	1960
x_2	Consumo de energia per capita	1960
x_3	Percentagem da força de trabalho na indústria	1960

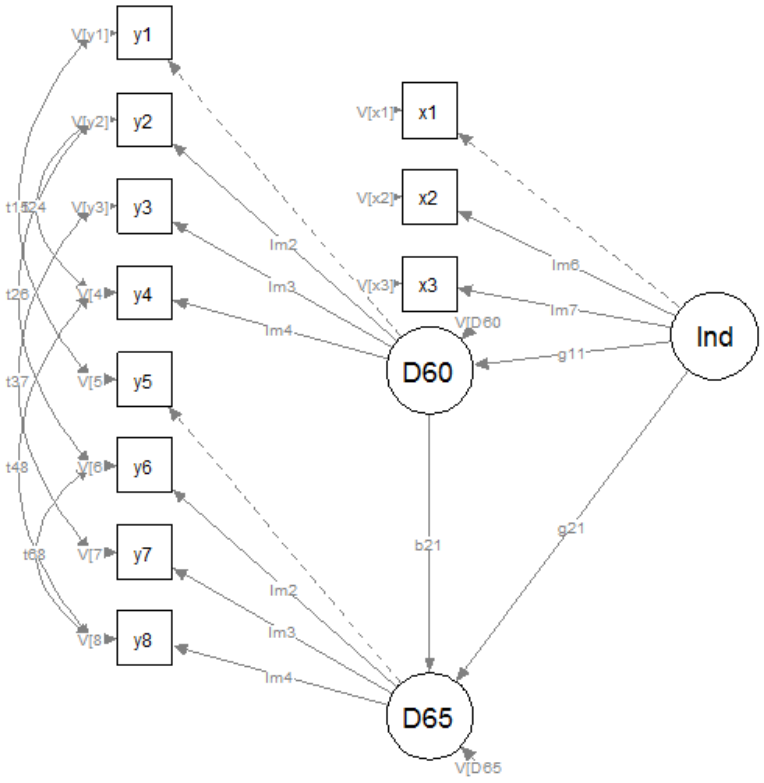


Figura 3.2: *Path Diagram* do modelo Industrialização e Democracia Política (Bollen, 1989)[2].

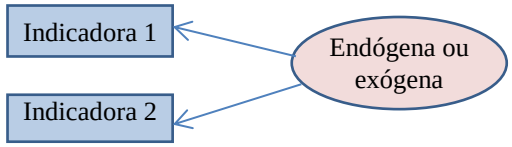
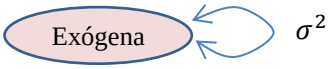
[2] Os coeficientes do modelo de medida são representados por λ_i (pesos fatoriais) para as variáveis indicadoras (no R foram representados por lm_i), os coeficientes do modelo estrutural são representados por γ_{ij} para as variáveis latentes exógenas ξ_j (g_{ij} no R) e β_{ij} para as variáveis latentes endógenas η_i (b_{ij} no R).

Neste tipo de diagrama as variáveis observadas são representadas por quadrados (ou retângulos) com os rótulos das variáveis escritos no seu interior. As variáveis latentes são representadas por círculos (ou elipses). Caminhos causais diretos são representados por setas unidirecionais que apontam da variável causal para a variável de efeito. As covariâncias entre pares de variáveis são identificadas por uma curva bidirecional. Em particular, a variância de uma variável é representada por um arco com dupla orientação para a variável respetiva. A covariância, geralmente, é considerada apenas entre variáveis exógenas. Estas, no diagrama, têm setas unidirecionais apontando apenas para outra variável e nenhuma apontando para elas, e representam entradas causais no sistema de variáveis. As variáveis endógenas têm setas apontando para elas e são variáveis dentro do sistema que são os efeitos de variáveis exógenas ou causas de outras variáveis endógenas dentro do sistema. Associado a cada caminho causal direto está um coeficiente estrutural, que representa o efeito causal direto da causa sobre a variável efeito. O efeito representa quanto a mudança de uma unidade na variável causal tem na variável de efeito ou proporcionalmente quanto da quantidade da variável causal é transferida para a variável de efeito. Quando não existe uma seta entre um par de variáveis, onde tal seta poderia existir, isso significa que não há conexão causal entre as variáveis e o coeficiente estrutural correspondente é zero. Assim, o que é deixado de fora de um diagrama de caminho é muito importante. Indicar que uma variável não é uma causa de outra variável é, muitas vezes, a maneira pela qual impomos restrições identificáveis e testáveis aos nossos modelos (Mulaik, 2009). No quadro 3.2 encontra-se uma síntese da simbologia usada para representar graficamente um modelo de equações estruturais.

O diagrama de caminhos (*path diagram*) traduz-se matematicamente por um conjunto de equações lineares que especifica as relações entre as variáveis e que formam conjuntos que constituem os submodelos de medida e estrutural.

As covariâncias dos erros de medida entre variáveis observadas são representadas por θ_{ij} (t_{ij} no R) e as variâncias das variáveis latentes exógenas por Φ (phi, no R). As variâncias e as covariâncias da variáveis observadas são representadas por σ_{ij} ([$V[x_{ii}]$ no R)

Quadro 3.2: Representação das relações entre variáveis (latentes e indicadores)

Tipo de relação	Representação
Entre uma variável latente (endógena ou exógena) e uma ou mais variáveis observadas	
Estrutural: dependência causal entre variáveis latentes	
Correlacional: variáveis latentes correlacionadas	
Variância de uma variável exógena	
Erro de medida e erro estrutural	

O modelo da Figura 3.2 pode ser representado analiticamente pelo seguinte conjunto de equações:

$$
\begin{aligned}
\eta_1 &= \gamma_{11}\xi_1 + \zeta_1 \\
\eta_2 &= \beta_{21}\eta_1 + \gamma_{21}\xi_1 + \zeta_2 \\
y_1 &= \eta_1 + \varepsilon_1 \\
y_2 &= \lambda_2\eta_1 + \varepsilon_2 \\
y_3 &= \lambda_3\eta_1 + \varepsilon_3 \\
y_4 &= \lambda_4\eta_1 + \varepsilon_4 \\
y_5 &= \eta_2 + \varepsilon_5 \\
y_6 &= \lambda_2\eta_2 + \varepsilon_6 \\
y_7 &= \lambda_3\eta_2 + \varepsilon_7 \\
y_8 &= \lambda_4\eta_2 + \varepsilon_8 \\
x_1 &= 1\xi_1 + \delta_1 \\
x_2 &= \lambda_6\xi_1 + \delta_2 \\
x_3 &= \lambda_7\xi_1 + \delta_3
\end{aligned}
\qquad \textbf{(1)}
$$

onde se incluem quatro medidas de democracia $[y]$ (liberdade de imprensa, liberdade da oposição política, eleições livres e cumprimento da legislatura eleita) em dois momentos

no tempo (1960 e 1965) e três medidas de industrialização em 1960 $[x]$ (PNB per capita, consumo de energia per capita e peso laboral na indústria, em percentagem). As variáveis η_1, η_2 são variáveis latentes endógenas e procuram exprimir a democracia política em 1960 e 1965 (representadas no diagrama por D60 e D65) e ξ_1 é uma variável latente exógena relacionada com a industrialização em 1960 (representada por Ind). As equações que definem η_1, η_2 constituem o modelo estrutural e as equações que definem as relações entre as variáveis indicadoras ($y_i, i = 1,2,3,4$ e $x_i, i = 1,2,3$) e as variáveis latentes η_1, η_2 e ξ_1 constituem o modelo de medida. Note-se que os pesos fatoriais dos indicadores y_1, y_5 e x_1 fixaram-se com o valor 1 (para definir as variáveis latentes correspondentes) e os pesos fatoriais dos pares de variáveis (y_2, y_6) , (y_3, y_7) e (y_4, y_8) foram forçados a serem iguais. Por outro lado, os erros associados a estes pares de variáveis, θ_{ij} são correlacionados.

O modelo com as estimativas dos parâmetros obtidas com recurso ao *package* `sem` do R encontra-se representado graficamente nas Figuras 3.3 e 3.4. Os diagramas foram obtidos nos packages `semPlot` e `sem`, respetivamente.

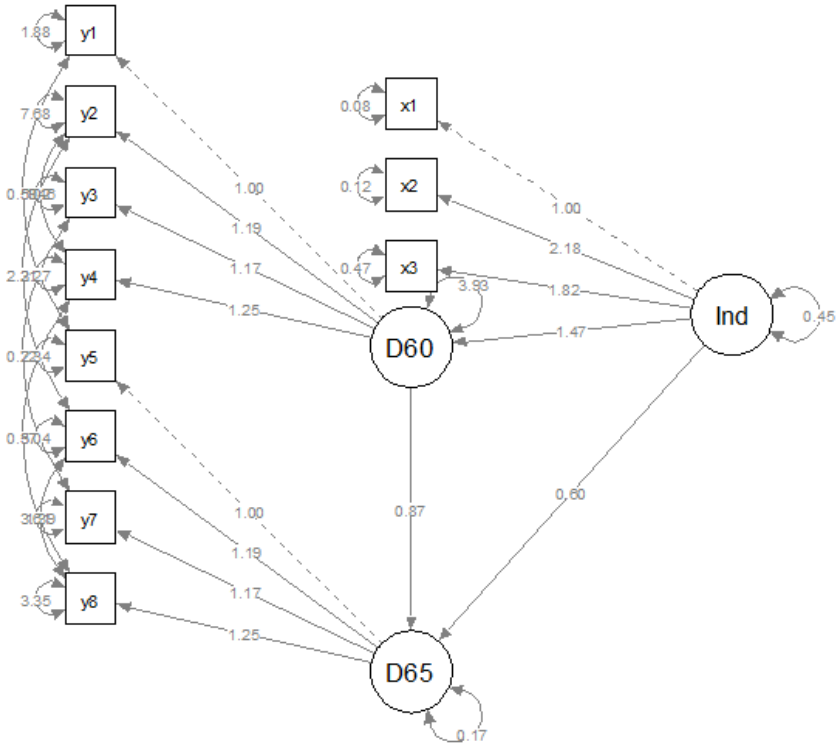


Figura 3.3: Representação gráfica do modelo SEM Industrialização e Democracia Política (Bollen, 1989) com as estimativas dos parâmetros, recorrendo à função ***semPaths***() do pacote ***semPlot***

Existem dois tipos de modelos estruturais, os recursivos (Figura 3.5a) e os não recursivos (Figura 3.5b). Os modelos recursivos possuem duas características básicas: os erros estruturais são não correlacionados e todos os efeitos causais são unidirecionais, isto é, nenhum par de variáveis endógenas é especificada como causa e efeito uma da outra. Estes modelos também podem ter erros estruturais correlacionados opcionais, mas apenas entre pares de variáveis endógenas sem efeitos diretos entre elas (Hoyle, 2012). Os modelos não-recursivos têm *loops* de *feedback* ou podem ter erros estruturais correlacionados entre pares de variáveis endógenas com efeitos diretos entre elas.

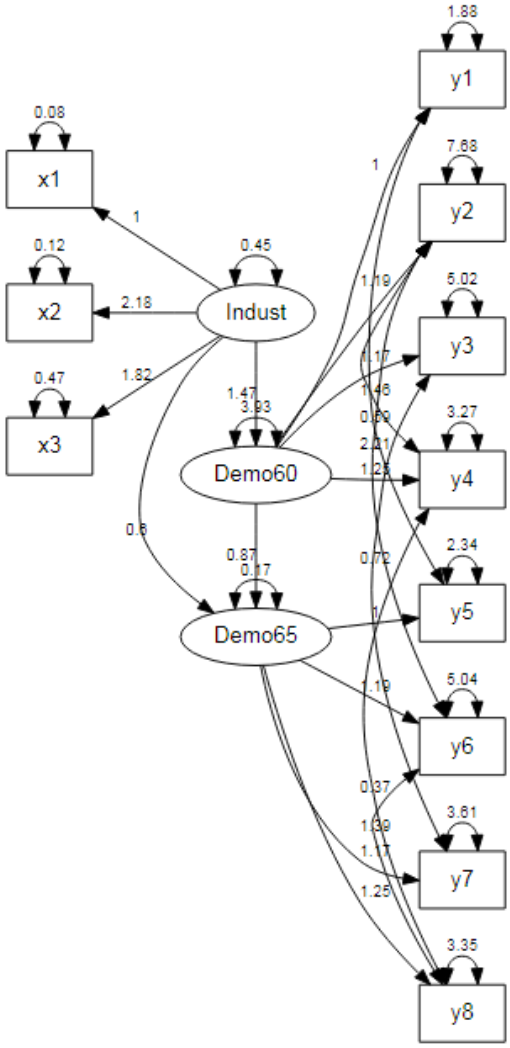


Figura 3.4: Representação gráfica do modelo SEM Industrialização e Democracia Política (Bollen, 1989) com as estimativas dos parâmetros, recorrendo à função ***pathDiagram***() do pacote ***semPlot***

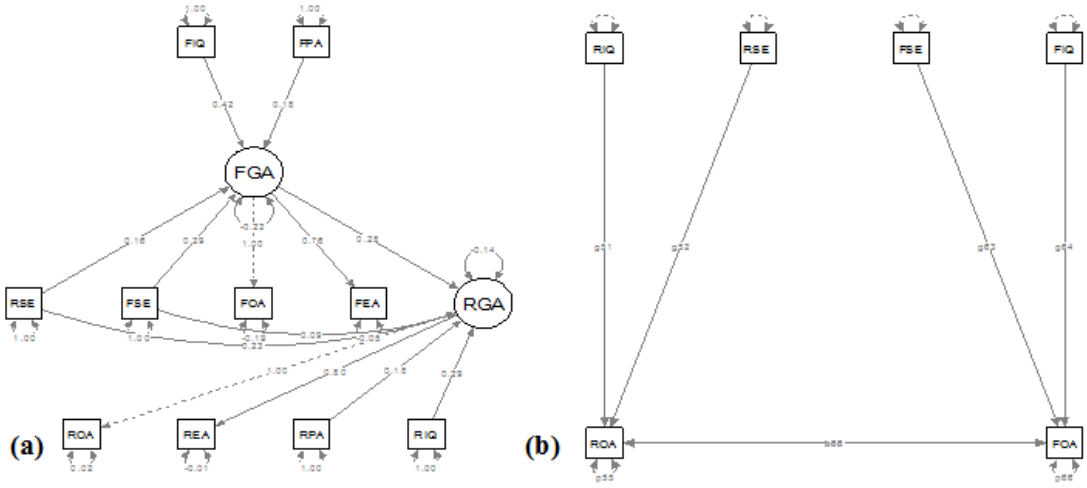


Figura 3.5: Modelos simulados com recurso ao pacote sem do software R: (a) Recursivo; (b) Não recursivo com um l*oop* de *feedback* direto ($\boldsymbol{ROA \rightarrow FOA}$ e $\boldsymbol{ROA \leftarrow FOA}$) .

3.3. Modelo matemático e pressupostos (especificação)

Tendo sido concebido o modelo teórico, é necessário proceder à sua especificação. A especificação do modelo consiste no desenho formal do mesmo que, *a priori*, reflete as hipóteses sobre o modelo de medida. Nesta fase, no contexto tradicional da SEM, deve ser tido em consideração que os fatores comuns latentes causam as variáveis observadas e que o comportamento destas resulta da manifestação dos fatores latentes; que a variância das variáveis observadas que não é explicada pelos fatores comuns latentes é explicada por fatores específicos latentes, nomeadamente os erros de medida ou resíduos; que os erros de medida são geralmente independentes embora possam estar correlacionados indicando uma fonte de variação comum dos itens, não explicada pelos fatores comuns presentes no modelo (Marôco, 2014). É nesta fase que se decide que variáveis observadas operacionalizam que constructos, quantas e quais variáveis observadas são incluídas/excluídas no modelo, que associações não causais devem ser omitas/incluídas e que erros devem ser correlacionados (Marôco, 2014).

O modelo de equações estruturais é o resultado da combinação de dois submodelos, o modelo de medida e o modelo estrutural, como referido anteriormente.

Quando a estrutura de covariância é analisada, o modelo geral de equações estruturais pode ser expresso por três equações básicas, escritas na forma matricial (Wang & Wang, 2012), na notação LISREL, por:

Modelo estrutural descrevendo as relações entre as variáveis latentes $\boldsymbol{\eta}$ (endógenas) e $\boldsymbol{\xi}$ (exógenas)

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta \qquad \textbf{(2)}$$

Modelo de medida descrevendo as variáveis de medida $\boldsymbol{Y}$ para as variáveis latentes endógenas $\boldsymbol{\eta}$:

$$Y = \Lambda_y \eta + \varepsilon \qquad \textbf{(3)}$$

Modelo de medida descrevendo as variáveis de medida $\boldsymbol{X}$ para as variáveis latentes exógenas $\boldsymbol{\xi}$:

$$X = \Lambda_x \xi + \delta \qquad \textbf{(4)}$$

onde

1) as matrizes de variáveis são:

   η é a matriz de m variáveis latentes endógenas de ordem $m \times 1$

   ξ é a matriz de n variáveis latentes exógenas de ordem $n \times 1$

   ζ é a matriz de erros estruturais de ordem $m \times 1$

   Y é a matriz de p variáveis observadas endógenas de ordem $p \times 1$

   X é a matriz de p variáveis observadas exógenas de ordem $q \times 1$

   ε é a matriz de erros de medida de Y de ordem $p \times 1$

   δ é a matriz de erros de medida de X de ordem $q \times 1$

2) as matrizes de coeficientes são:

   Λ_y é a matriz de pesos fatoriais de η em Y de ordem $p \times m$

   Λ_x é a matriz de pesos fatoriais de ξ em X de ordem $q \times n$

   B matriz de coeficientes relacionando η com η, de ordem $m \times m$

   Γ matriz de coeficientes relacionando ξ com η de ordem $m \times n$

Entretanto, é usual designar as matrizes de variâncias/covariâncias por:

Φ é a matriz de variâncias/covariâncias de ξ de ordem $n \times n$

Ψ é a matriz de variâncias/covariâncias de ζ de ordem $m \times m$

Θ_ε é a matriz de variâncias/covariâncias de ε de ordem $p \times p$

Θ_δ é a matriz de variâncias/covariâncias de δ de ordem $q \times q$

Assume-se que os termos residuais podem estar correlacionados entre si ($Cov(\varepsilon_i, \varepsilon_j)$ e $Cov(\delta_i, \delta_j)$ podem ser não nulos para algum par (i,j), $i \neq j$) mas não podem estar correlacionados entre submodelos ($Cov(\varepsilon_i, \delta) = 0, \forall i$ (Kline 2011, Marôco 2014) e são normalmente distribuídos:

$$\varepsilon \sim N_p(0, \Theta_\varepsilon),\ \delta \sim N_q(0, \Theta_\delta) \text{ e } \zeta \sim N(0, \Psi) \quad \textbf{(5)}$$

Além disso, uma variável dependente não é causa e efeito dela mesma, pelo que:

ε e η são independentes ($Cov(\varepsilon, \eta) = 0$)

δ e ξ são independentes ($Cov(\delta, \xi) = 0$);

ζ e ξ são independentes ($Cov(\zeta, \xi) = 0$);

ε, δ e ζ são mutuamente independentes

($Cov(\varepsilon, \delta) = 0,\ Cov(\varepsilon, \zeta) = 0), Cov(\zeta, \delta) = 0$).

Assume-se que as observações constituem amostras independentes. Sob estes pressupostos as variáveis observadas X e Y têm distribuição normal multivariada:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, \Sigma), \quad \textbf{(6)}$$

sendo Σ a matriz de covariância populacional dos indicadores.

A imposição destes pressupostos leva à seguinte expressão para a estrutura de covariância reproduzida (Σ) entre Y e X:

$$\Sigma = (YX)(YX)^T \quad \textbf{(7)}$$
$$= \begin{bmatrix} YY^T & YX^T \\ XY^T & XX^T \end{bmatrix}$$
$$= \begin{bmatrix} \Lambda_y(I - B)^{-1}(\Gamma\Phi\Gamma + \Psi)(I - B)^{-1}\Lambda_y^T + \Theta_\varepsilon & \Lambda_y(I - B)^{-1}\Gamma\Phi\Lambda_x^T \\ \Lambda_x\Phi\Gamma^T(I - B^T)^{-1}\Lambda_y^T & \Lambda_x\Phi\Lambda_X^T + \Theta_\delta \end{bmatrix}$$

Pode-se observar que Σ é função de oito matrizes de parâmetros a estimar, a saber: $\Lambda_y, \Lambda_x, B, \Gamma, \Phi, \Psi, \Theta_\varepsilon$, e Θ_δ.

A estimação dos parâmetros baseia-se na seleção dos valores dos parâmetros estruturais que reproduzem a matriz de covariância, uma vez que a questão empírica do SEM é a de

avaliar se o modelo proposto produz uma matriz de covariâncias que é consistente com a matriz covariâncias amostral.

Pode-se concluir ainda que os elementos da diagonal principal de B são nulos, que um elemento nulo nesta matriz representa a ausência de efeito de uma variável latente endógena noutra variável latente endógena. Além disso, a matriz $\mathrm{I} - \mathrm{B}$ tem que ser não singular para que exista $(\mathrm{I} - \mathrm{B})^{-1}$ e possa ser feita a estimação do modelo.

Além dos pressupostos sobre os erros e sobre B há outros pressupostos a ter em consideração. Refira-se que a suposição que o modelo definido pelo investigador está basicamente correto, antes que qualquer interpretação sobre causalidade possa ser feita, é o pressuposto mais abrangente de todos os que estão subjacentes à análise SEM (Kline, 2011). Este pressuposto implica que a relação entre as variáveis observadas e os seus constructos e entre um constructo e outro é linear e que deve haver uma relação de causa e efeito entre variáveis endógenas e exógenas (covariância não nula), e uma causa deve ocorrer antes do evento. Este pressuposto é relevante no modelo de medida. Entretanto, são assumidos mais alguns pressupostos que evitam que os resultados obtidos sejam comprometidos, podendo levar a conclusões incorretas (Kline, 2011, Hair *et al*., 2010), a saber:

a) A distribuição conjunta das variáveis endógenas deve ser normal multivariada para que possam ser usados os métodos mais comuns de estimação da SEM. Para garantir este pressuposto é necessário garantir que as distribuições univariadas são normais, que a distribuição conjunta de qualquer par de variáveis tem normalidade bivariada e que todos os gráficos bivariados são lineares e a distribuição dos resíduos é homocedástica, da mesma forma que deve ser evitada uma forte assimetria nos dados (Hair *et al*., 2010; Kline, 2011).

Existem testes estatísticos destinados a detetar a violação da normalidade multivariada, incluindo o teste de Mardia (Mardia, 1985), baseado em testes de assimetria e curtose e o teste de Cox-Small (Cox e Small, 1978), entre outros. Como a SEM deve ser aplicada a grandes amostras, e desvios leves da normalidade podem ser estatisticamente significativos em grandes amostras, os resultados destes testes ficam comprometidos caso não se verifique este pressuposto. Uma vez que a não normalidade multivariada é, em muitas situações, detetada através da análise à normalidade univariada, este pressuposto pode ser avaliado através dos índices de assimetria e de curtose, sendo casos extremos de desvio da

normalidade índices de assimetria acima de 3 e de curtose acima de 10. Os métodos gráficos, como o QQ-*Plot*, *boxplot*s ou histogramas, são alternativas, bem como a análise de resíduos.

Uma forma de lidar com a normalidade univariada – e, portanto, abordar a normalidade multivariada – consiste na implementação de transformações nos dados através de uma operação matemática, o que significa que os *scores* originais são convertidos em novos *scores* que terão distribuições mais próximas da normal. Também se pode recorrer a métodos de reamostragem, em particular por métodos *bootstrap* disponíveis.

b) A multicolinearidade extrema pode ocorrer se variáveis observadas, aparentemente separadas, realmente medirem a mesma coisa. Variáveis com um elevado grau de colinearidade não devem ser incluídas na mesma análise. Neste caso a matriz $I - \mathrm{B}$ pode não ser invertível por não ser definida positiva.

Para fazer o diagnóstico da multicolinearidade extrema, podem ser usados diversos métodos. Por exemplo, calcular o coeficiente de determinação (R^2) entre cada variável e todas as restantes variáveis observadas. Se para uma variável tomada como dependente, este valor for maior que 0.9, então suspeita-se da existência de extrema multicolinearidade. Um critério equivalente é a determinação da *Tolerance* $(1 - R^2)$ que se for inferior a 0.1 sugere a existência de extrema multicolinearidade. Se a estatística VIF $\left(\frac{1}{1-R^2}\right)$ for maior que 10, a variável é redundante.

c) A existência de *outliers* é um outro problema que pode comprometer os resultados pois estes podem afetar as covariâncias entre as variáveis e isto pode repercutir-se nas médias, desvios-padrão e covariâncias, comprometendo a qualidade de ajustamento do modelo (Schumacker e Lomax, 2004). A existência de um *outlie*r multivariado pode ser detetada se os valores de duas ou mais variáveis numa observação multivariada tiverem um *z score* superior a 3, ou se apresentar um padrão de *scores* atípico. Caso não haja *scores* individuais extremos, a sua deteção faz-se através do quadrado da Distância de Mahalanobis (D^2), que mede a distância de uma observação x_i à média de todas as observações de todas as variáveis $(\overline{x})$, que se designa por centróide.

$$D^2 = (x_i - \overline{x})^T \mathrm{S}^{-1} (x_i - \overline{x}) \qquad \textbf{(8)}$$

S é a matriz de covariâncias observadas na amostra.

Com amostras de grande dimensão, D^2 tem distribuição χ^2 e portanto é possível testar se um determinado dado provêm da mesma população dos restantes dados. No caso de se detetar um valor atípico pode, por exemplo, optar-se pela sua exclusão ou pela sua substituição por estatísticas robustas, nomeadamente pelo valor mais próximo que não é considerado *outlier*.

d) A existência de dados omissos também é um problema. A análise SEM deve ser feita com dados completos. Este tópico será abordado com maior profundidade no próximo capítulo.

Refiram-se outros pressupostos não tão relevantes mas que devem ser tidos em consideração. Cada variável latente deve ter 3 ou mais variáveis observadas e deve haver correlações medianas a fortes entre as variáveis observadas que operacionalizam o mesmo constructo (Hair *et al*., 2010; Marôco, 2014). Kline (2011) considera que se o modelo tiver duas ou mais variáveis latentes, este fica identificado se tiver pelo menos 2 indicadores por variável latente e cada variável latente estiver correlacionada com pelo menos uma outra variável latente. Na construção de alguns modelos teóricos há a omissão de variáveis preditoras fundamentais, conduzindo ao designado erro de especificação, erro este que compromete a avaliação da importância de outras variáveis. Porém o desejo de incluir todas as variáveis deve ser equilibrado com as limitações práticas da SEM. Mesmo não existindo um limite teórico para o número de variáveis a incluir no modelo, deve-se preservar a parcimónia e reconhecer os seus benefícios e os de modelos teoricamente concisos (Hair *et al*., 2010).

A SEM é mais sensível ao tamanho da amostra do que outras abordagens multivariadas (Hair *et al*., 2010). A maioria dos investigadores prefere um tamanho de amostra que corresponda a 10 a 20 casos por variável, devendo variar entre 100 e 500 dados, no mínimo, de acordo com os procedimentos de análise e características do modelo seguintes: (1) normalidade multivariada dos dados, (2) técnica de estimação, (3) complexidade do modelo, (4) quantidade de dados omissos e (5) erro médio variação entre os indicadores reflexivos (Hair *et al*.,2010).

3.4. Identificação do modelo

Para que um modelo seja estimável é necessário que seja atribuída pelo menos uma equação para estimar cada coeficiente e que o sistema não seja indeterminado, isto é, haja pelo menos tantos elementos na matriz de covariâncias de dados como parâmetros a estimar (Kline, 2011). Estamos aqui a discutir o problema da não identificação do modelo que geralmente ocorre quando a situação descrita não se verifica.

O problema de identificação do modelo diz respeito à existência de solução única na estimação dos parâmetros. Um modelo é dito identificado quando teoricamente é possível obter uma única estimativa para cada combinação dos parâmetros.

Num modelo com p variáveis observadas endógenas e q variáveis observadas exógenas, o número de elementos não redundantes da matriz de covariâncias é

$$\frac{(p+q)(p+q+1)}{2} \qquad \textbf{(9)}$$

E, sendo *t* o número de parâmetros a estimar, o número de graus de liberdade é

$$d_f = \frac{(p+q)(p+q+1)}{2} - t \qquad \textbf{(10)}$$

Para que um modelo seja identificado é necessário que $d_f \geq 0$ e que cada parâmetro seja univocamente estimado (Hair *et al*. (2010).

Quanto à identificação distinguem-se três tipos de modelos (Hair *et al*. 2010; Marôco 2014; Schumacker e Lomax 2004):

(i) o modelo sub-identificado ou indeterminado, quando o número de parâmetros a estimar é superior ao número de elementos não redundantes da matriz de covariância, $d_f < 0$;

(ii) o modelo é exatamente identificado, saturado ou determinado, quando o número de parâmetros a estimar é igual ao número de elementos não redundantes da matriz de covariância, $d_f = 0$;

(iii) o modelo é sobre-identificado ou sobressaturado quando o número de parâmetros a estimar é inferior ao número de elementos não redundantes da matriz de covariância, $d_f > 0$.

Um modelo teoricamente identificado ou sobre-identificado pode apresentar problemas de sub-idenficação empírica quando um parâmetro necessário para a identificação do modelo tem um valor próximo de zero. Esta situação pode ocorrer por diversos motivos, nomeadamente porque o processo iterativo de estimação do modelo elimina esse parâmetro ou há multicolinearidade entre variáveis o que conduz a instabilidade nas estimativas dos parâmetros associadas, podendo igualmente ocorrer a eliminação das variáveis observadas da análise (Marôco, 2014). A resolução deste problema passa, normalmente, por remover variáveis observadas colineares ou por aumentar a dimensão da amostra.

Embora alguns modelos possam precisar de reespecificação, muitas vezes os problemas de identificação surgem de erros comuns na especificação do modelo e dos dados de entrada. Alguns dos problemas mais comuns que levam a problemas de identificação incluem a especificação incorreta de uma variável observada, como por exemplo, não a ligar a qualquer constructo ou ligá-la a dois ou mais constructos, selecioná-la duas vezes no mesmo modelo, ou não criar um termo de erro para essa variável. O investigador deve inspecionar cuidadosamente a especificação do modelo.

Uma característica única na especificação das variáveis observadas para cada constructo é o processo de "definir a escala" de uma variável latente. Não "definir a escala" para cada variável latente origina um problema de identificação. Por não ser observada, uma variável latente não possui escala métrica e a definição da respetiva escala deve ser feita para variáveis latentes exógenas e endógenas. Definir uma escala para uma variável latente pode ser conseguido fixando pelo menos um dos pesos fatoriais em cada variável latente em um valor específico (normalmente 1 ou um valor conhecido previamente) ou fixar as variâncias das variáveis latentes exógenas (1 é um bom valor).

Segundo Hair *et al.* (2010) e Marôco (2014), para além do referido, problemas eventuais de identificação do modelo podem também ser corrigidos através da construção de um modelo com um número mínimo de coeficientes (reduzir o número de variáveis latentes, eliminar variáveis observadas multicolineares, fixar trajetórias em zero – eliminá-las), da fixação das variâncias de erros de medida, da fixação dos coeficientes conhecidos e da eliminação das variáveis problemáticas.

3.5. Estimação dos parâmetros do modelo

A matriz de covariância populacional das variáveis observadas y e x, Σ, pode ser expressa como função dos parâmetros livres de um modelo hipotético, ou seja, $\Sigma = \Sigma(\theta)$, onde $\Sigma(\theta)$ traduz a matriz de variância/covariância implícitada pelos parâmetros da população para o modelo hipotético.

A finalidade da estimação do modelo ou do ajustamento do modelo é encontrar um conjunto de parâmetros θ do modelo e produzir uma matriz de covariâncias $\Sigma(\theta)$ de modo que $[\Sigma - \Sigma(\theta)]$ possa ser minimizado. A discrepância entre Σ e $\Sigma(\theta)$ indica quão bem o modelo se ajusta aos dados.

Uma vez que Σ e $\Sigma(\theta)$ são desconhecidas, a estimação dos parâmetros do modelo consiste em minimizar a discrepância entre a matriz de covariâncias amostral observada S e a matriz de covariâncias estimadas a partir do modelo, $\Sigma(\hat{\theta})$, sendo $\hat{\theta}$ o vetor de parâmetros do modelo que reproduz o melhor possível a matriz de covariâncias observadas. Assim, a estimação do modelo tem como base as estimativas dos parâmetros (matrizes de covariâncias) que melhor reproduzem os dados observados. Os erros de ajustamento dizem respeito às variâncias e covariâncias entre cada uma das variáveis observadas e não aos valores individuais de cada observação.

O objetivo é encontrar a melhor estimativa de θ tal que (Marôco, 2014):

$$S = \Sigma(\hat{\theta}) = \begin{bmatrix} S_{YY} & S_{YX} \\ S_{XY} & S_{XX} \end{bmatrix} = \begin{bmatrix} \Sigma_{YY}(\hat{\theta}) & \Sigma_{YX}(\hat{\theta}) \\ \Sigma_{XY}(\hat{\theta}) & \Sigma_{XX}(\hat{\theta}) \end{bmatrix} \quad \textbf{(11)}$$

A discrepância é definida por uma função $F(S - \Sigma(\hat{\theta}))$ que deve ser escolhida e minimizada.

Uma das etapas da implementação da SEM é a escolha da função objetivo F, estimador dos parâmetros do modelo, e a implementação de métodos iterativos aplicados a este estimador usado para estimar os parâmetros.

Em SEM, os estimadores F mais frequentes são os de Máxima Verosimilhança (ML-*Maximum Likelihood*), o dos Mínimos Quadrados Generalizados (GLS-*Generalized Least Square*), o dos Mínimos Quadrados Ponderados (PLS), o dos Mínimos Quadrados Não Ponderados (ULS), o FIML (*Full Information Maximum Likelihood*) para dados omissos (Hair *et al*., 2010, Marôco, 2014, Schumacker e Lomax, 2004).

Os estimadores ML, GLS, FIML têm como pressupostos a normalidade multivariada dos dados e no caso do FIML os dados omissos devem ser pelo menos MAR (***M**issing **A**t **R**andom*). Os estimadores PLS e ULS são livres de pressupostos sobre a distribuição dos dados e dos erros.

Os estimadores ML, GLS e FIML podem ser definidos pelas expressões seguintes, sendo |... | a função determinante de uma matriz e *tr(...)* a função traço de uma matriz:

$$F_{ML} = ln\left|\Sigma\left(\hat{\theta}\right)\right| - ln|S| + tr\left[S\left(\Sigma\left(\hat{\theta}\right)\right)^{-1}\right] - (p+q) \quad \textbf{(12)}$$

$$F_{GLS} = \frac{1}{2}tr\left[\left(S - \Sigma\left(\hat{\theta}\right)W^{-1}\right)\right]^2 \quad \textbf{(13)}$$

$$F_{FIML} = k - \frac{1}{2}ln\left|\Sigma\left(\hat{\theta}\right)\right| - \frac{1}{2}\left(Y_{\mathrm{i}} - \mu_i(\hat{\theta})\right)\Sigma_{\mathrm{i}}^{-1}\left(Y_{\mathrm{i}} - \mu_i(\hat{\theta})\right) \quad \textbf{(14)}$$

onde $\Sigma\left(\hat{\theta}\right)$ é a matriz de covariâncias gerada pelo modelo, $\hat{\theta}$ o vetor de parâmetros do modelo, S é a matriz de covariâncias observada na amostra, W^{-1} é uma matriz de peso, definida positiva, para os resíduos, p e q os números de variáveis observadas endógenas e exógenas, respetivamente, y_i é o vetor de valores observados para o caso i, μ_i é o respetivo vetor de médias e Σ_i é a respetiva matriz de covariâncias.

S^{-1} funciona como a matriz de pesos dos resíduos W que, embora possa ser escolhida outra, é a que é usada nos pacotes SEM.

O estimador WLS, ou Distribuição Assimptótica Livre – ADF (*Asymptotic Distribution Free*) não exige, como referido, que as variáveis observadas apresentem distribuição normal multivariada. Este estimador é definido por

$$F_{ADF} = \left(s - \sigma(\hat{\theta})\right)^T W^{-1} \left(s - \sigma(\hat{\theta})\right) \sigma(\hat{\theta}) \quad \textbf{(15)}$$

sendo,

s^T é o vetor de elementos da matriz triangular inferior S incluindo a diagonal;

$\sigma(\hat{\theta})^T$ é o vetor de elementos da matriz triangular inferior $\Sigma(\hat{\theta})$ incluindo a diagonal;

W é uma matriz de distâncias de todas as observações às médias de todas as variáveis.

O estimador dos Mínimos Quadrados Não-ponderados (ULS) não exige suposições sobre a distribuição e não tem testes estatísticos associados, e as estimativas são dependentes de escala na medida - alterações na escala das variáveis observadas originam diferentes soluções ou conjuntos de estimativas, contrariamente aos restantes casos que são independentes da escala (Schumacker e Lomax, 2004). O estimador define-se por

$$F_{ULS} = \frac{1}{2} tr\left[S - \Sigma(\hat{\theta})\right]^2 \quad \textbf{(16)}$$

O estimador ML é cêntrico, consistente e eficiente para grandes amostras e as estimativas dos parâmetros são assintoticamente normalmente distribuídas. Além disso, é livre de escala e a função $F_{ML}(\hat{\theta})$ multiplicada por $(n - 1)$ aproxima-se de uma distribuição X^2 sob a hipótese de normalidade multivariada e tamanho de amostra grande, podendo o modelo X^2 ser usado para testar o ajustamento geral do modelo (Wang e Wang, 2012). A não verificação dos pressupostos compromete os resultados, nomeadamente os testes à significância do modelo. Para resolver os problemas associados à não normalidade dos dados, pode-se considerar transformações de variáveis não-normais que as fazem aproximar melhor à normalidade multivariada, remover os *outliers* dos dados, aplicar procedimentos de *bootstrap* para estimar as variâncias de estimativas de parâmetros para os testes de significância (Bollen e Stine, 1993; Efron e Tibshirani, 1993). Em alternativa podem ser usados estimadores alternativos, robustos à não normalidade como o estimador ADF ou o estimador ULS.

3.6. Avaliação da qualidade do ajustamento do modelo

A avaliação da qualidade do modelo tem como objetivo aferir sobre quão bem o modelo teórico é capaz de reproduzir a estrutura de covariância ou a estrutura correlacional das

variáveis observadas na amostra em estudo, sendo pouco consensual entre os investigadores. Ao longo dos tempos foram sendo desenvolvidas diferentes estratégias para avaliar a qualidade do modelo, e que se podem agrupar em (1) testes de ajustamento, (2) índices empíricos baseados nas funções de verosimilhança e (3) análise de resíduos e da significância dos parâmetros. A utilização de uma ou outra estratégia ou de mais que uma depende do se pretender testar o ajustamento global ou local e das características do modelo e dos dados, nomeadamente da dimensão da amostra, do modelo e da razão entre o número de variáveis latentes e de variáveis observadas, dos graus de liberdade, do método de estimação e do grau de especificação errada.

A extensão segundo a qual $\Sigma(\hat{\theta})$ se ajusta aos dados, isto é, difere de S, pode ser avaliada por diversos índices de ajustamento (Hair *et al*., 2010), no pressuposto da normalidade dos dados: teste χ^2, parâmetro de não centralidade (NCP), índice de qualidade do ajustamento (GFI), raiz do resíduo quadrático médio (RMSR), raiz do erro quadrático médio de aproximação (RMSEA), índice ajustado de qualidade do ajustamento (AGFI), índice de Tuker-Lewis (TLI) , índice de ajustamento não-ponderado (NNFI), índice de ajustamento ponderado (NFI), índice de ajustamento comparativo (CFI), índice de ajustamento incremental (IFI), índice de ajustamento relativo (RFI), Critério de Informação Akaike (AIC), entre outros. Estes e outros índices estão disponíveis nos pacotes do R. No caso do estimador PLS, porque não assume hipóteses sobre a distribuição das observações e dos erros do modelo adotado, os testes tradicionais baseados na teoria do χ^2 não são apropriados, sendo usados outros critérios: o Coeficiente de Determinação R^2 e a Variância Média Extraída (AVE) que se destinam a analisar a qualidade do ajustamento e a capacidade de previsão do modelo e *Jacknifing* ou *Bootstraping* que se destinam a testar a estabilidade das estimativas dos parâmetros.

Se matriz de variância/covariância estimada pelo modelo, $\Sigma(\hat{\theta})$, não é estatisticamente diferente da matriz de covariância dos dados observados, S, então diz-se que o modelo se ajusta bem os dados, e aceita-se a hipótese nula $H_0: S = \Sigma(\hat{\theta})$, ou diz-se que o modelo apoia a plausibilidade das relações postuladas entre as variáveis; caso contrário, o modelo não se ajusta aos dados e a hipótese nula deve ser rejeitada (Wang e Wang, 2012).

Assim, a avaliação do modelo é realizada através do teste χ^2 de ajustamento e também através de alguns índices que na sua maioria são utilizados em todos os *softwares* de SEM.

3.6.1. Teste do χ^2 de ajustamento

O teste do χ^2 de ajustamento testa a significância da função de discrepância

$$f = \mathrm{F}\left(\mathrm{S} - \Sigma\left(\hat{\theta}\right)\right) \quad \textbf{(17)}$$

minimizada durante o ajustamento do modelo. A estatística do teste do χ^2de Ajustamento é dada por:

$$\chi^2 = (n-1)f_{min} \underset{\sim}{a} \chi^2_{(g.l.)} \quad \textbf{(18)}$$

Como os resíduos, ou seja, os elementos de $\mathrm{S} - \Sigma\left(\hat{\theta}\right)$ devem ser próximos de zero para um bom ajustamento do modelo, interessa obter um valor χ^2 com graus de liberdade associados não significativo.

Este teste é altamente sensível ao tamanho da amostra, pelo que requer alguns cuidados no que respeita a este aspeto. Quanto maior o tamanho da amostra, maior a probabilidade de rejeitar o modelo, portanto, é mais provável que se rejeite a hipótese correta (erro tipo I). Para amostras pequenas, o teste tem probabilidades maiores de não rejeitar a hipótese de que o modelo se ajusta bem aos dados quando o ajustamento é mau (erro tipo II), para além de a função de ajustamento poder não seguir uma distribuição χ^2. Acresce que χ^2 é muito sensível à violação do pressuposto de normalidade multivariada e aumenta quando o número de variáveis num modelo aumenta.

Quando a normalidade multivariada não é válida, o teste pode ser corrigido usando a Correção de Satorra-Bentler, de forma a considerar a distribuição amostral não central da estatística do teste (Marôco (2014).

Pelo exposto, o resultado do teste χ^2 não deve ser uma razão por si só para rejeitar um modelo. Para abordar estas limitações vários índices de ajustamento de modelo foram propostos para o teste de ajustamento do modelo.

3.6.2. Índices de qualidade de ajustamento

Existe uma grande panóplia de índices de qualidade de ajustamento, dos quais se seguem alguns dos mais utilizados de acordo com Byrne (2012), Hair *et al*. (2010), Marôco (2014), Schermelleh-Engel Moosbrugger e Müller (2003) e Wang e Wang (2012).

I. Índices absolutos

Estes índices avaliam o ajustamento do modelo sem comparação com outros modelos.

- **CMIN/DF**

$$\text{CMIN/DF} = \frac{\chi^2}{\text{gl}}. \qquad \textbf{(19)}$$

$\text{CMIN}(\text{ estatística } \chi^2)$ é o valor mínimo de discrepância.

O ajustamento do modelo considera-se muito bom se o valor do índice for igual ou inferior a 1, bom se estiver entre 1 e 2, tolerável se variar entre 2 e 5 e mau se for superior a 5 (Portela, 2012).

- ***RMSR*** **(***Root Mean Square Residual***).**

$$RMSR = \sqrt{\frac{\sum_{i=1}^{p+q}\sum_{j=1}^{i}\left(s_{ij} - \sigma_{ij}(\hat{\theta})\right)^2}{\frac{(p+q)(p+q+1)}{2}}} \qquad \textbf{(20)}$$

onde s_{ij} e $\sigma_{ij}(\hat{\theta})$ são os elementos das matrizes de variância/covariância observada S e estimada pelo modelo $\Sigma(\hat{\theta})$ e p e q o número de variáveis observadas endógenas e exógenas, respetivamente.

Este índice resulta da raiz quadrada da média dos resíduos. O ajustamento será tanto melhor quanto menor for o valor de RMSR, sendo que um valor de zero revela um ajustamento perfeito.

- ***GFI*** **(***Goodness of Fit Index***)**.

$$GFI = 1 - \frac{\left(s - \sigma(\hat{\theta})\right)^T W^{-1}\left(s - \sigma(\hat{\theta})\right)}{s^T W\hat{}(-1)s} \qquad \textbf{(21)}$$

O numerador é o mínimo da discrepância generalizada depois do modelo ter sido ajustado.

Este índice explica a proporção da covariância entre as variáveis observadas. O ajustamento considera-se muito bom se for igual ou superior a 0.95, bom entre 0.9 e 0.95,

tolerável entre 0.8 e 0.9 e mau se for inferior a 0.8. No caso de ser 1, o ajustamento é perfeito.

II. Índices relativos (medidas de ajustamento incrementais**)**

Os índices relativos avaliam a qualidade do modelo relativamente a um modelo de independência (não há relações entre variáveis observadas – as covariâncias são nulas) ou a um modelo saturado (modelo com melhor ajustamento possível). Os mais comuns são os seguintes.

- ***NFI (****Normal Fit Index***)**

$$NFI = 1 - \frac{\chi^2}{\chi_b^2} \qquad \textbf{(22)}$$

Este índice avalia a percentagem de incremento na qualidade do ajustamento do modelo ajustado, relativamente ao modelo de independência total ou ao modelo basal. É pouco utilizado por ser pouco fiável em amostras pequenas e é tanto maior quanto maior for o número de parâmetros a estimar e maior for a dimensão da amostra. Considera-se que o ajustamento é muito bom se o valor do índice for igual ou superior a 0.95 (o modelo está a 95% do percurso entre o pior e o melhor modelos possíveis), bom entre 0.9 e 0.95, tolerável entre 0.8 e 0.9 e mau se for inferior a 0.8.

- ***CFI (****Comparative Fit Index****)***

$$CFI = 1 - \frac{máx(\chi^2 - gl, 0)}{máx(\chi_b^2 - gl_b, 0)} \qquad \textbf{(23)}$$

Este índice corrige a subestimação que, regra geral, ocorre com o NFI para amostras pequenas. É independente da dimensão da amostra mas em amostras pequenas diminui com o aumento do número de variáveis com correlações não muito fortes.

Compara o ajustamento do modelo em estudo, com gl graus de liberdade, com o ajustamento do modelo basal com gl_b graus de liberdade. O ajustamento considera-se muito bom se o valor do índice for igual ou superior a 0.95, bom entre 0,9 e 0.95, tolerável entre 0.8 e 0.9 e mau se for inferior a 0.8.

- ***RFI*** **(***Relative Fit Index***).**

$$RFI = 1 - \frac{\frac{\chi^2}{gl}}{\frac{\chi_b^2}{gl_b}} \quad \textbf{(24)}$$

Este índice avalia o ajustamento do modelo comparando o χ^2 normalizado pelos graus de liberdade, com o modelo basal também normalizado pelos graus de liberdade. É pouco utilizado por ser pouco fiável em amostras pequenas. Considera-se o ajustamento muito bom quanto mais próximo de 1 estiver o valor do índice e mau se for inferior a 0.9.

- ***TLI*** **(***Tucker-Lewis Index***)** ou ***NNFI*** **(** *Bentler-Bonett non-normed fit index***)**

$$TLI = \frac{\frac{\chi_b^2}{gl_b} - \frac{\chi^2}{gl}}{\frac{\chi_b^2}{gl_b} - 1} \quad \textbf{(25)}$$

Neste índice há uma combinação de uma medida de parcimónia com um índice comparativo entre o modelo em estudo e o modelo basal, na medida em que é moderadamente corrigido pela parcimónia do modelo (Wang e Wang, 2012). O ajustamento considera-se muito bom se o valor do índice for igual ou superior a 0.95, bom entre 0.9 e 0.95, tolerável entre 0.8 e 0.9 e mau se for inferior a 0,8. É de notar que este índice, normalmente, varia entre 0 e 1 mas não está limitado a este intervalo. Por exemplo, se o modelo especificado tiver muito poucos graus de liberdade e as correlações entre as variáveis observadas forem baixas, este índice pode ser negativo.

III. Índices de parcimónia

As medidas de parcimónia relacionam o ajustamento do modelo com o número de coeficientes estimados necessários para atingir o nível de adequação pretendido. O objetivo básico é diagnosticar se o ajustamento do modelo foi atingido pelo sobreajustamento (*overfitting*) dos dados com o uso de muitos coeficientes.

Os índices de parcimónia obtêm-se a partir da correção dos índices relativos com um fator de penalização associado à complexidade do modelo (Marôco, 2014), estando, por isso, relacionados com os índices relativos descriminados acima. Estes índices penalizam os índices relativos por um fator de complexidade estimado como $\frac{gl}{gl_b}$.

- ***PCFI (****Parsimony Comparative Fit Index****).***

$$PCFI = \text{CFI} \times \frac{gl}{gl_b} \quad \textbf{(26)}$$

Penaliza o índice CFI.

- ***PGFI (****Parsimony Goodness of Fit Index****).***

$$PGFI = \text{GFI} \times \frac{gl}{gl_b} \quad \textbf{(27)}$$

Penaliza o índice GFI.

- ***PGFI (****Parsimony Normal Fit Index****).***

$$PNFI = \text{NFI} \times \frac{gl}{gl_b} \quad \textbf{(28)}$$

Penaliza o índice NFI.

Em todos estes índices considera-se o ajustamento do modelo muito bom se o seu valor for igual ou superior a 0,8, bom se estiver entre 0.6 e 0.8 e mau se for inferior a 0.6.

IV. Índices de Discrepância Populacional

Os índices de discrepância populacional comparam o ajustamento do modelo obtido com as médias e variâncias amostrais com o ajustamento do modelo que seria obtido com as médias e variâncias da população.

- ***NCP*** *(Parâmetro da não-centralidade)*

$$NPC = máx(\chi^2 - gl, 0) \quad \textbf{(29)}$$

Reflete o grau de desajustamento do modelo proposto à estrutura de variância-covariância observada. O ajustamento será tanto melhor quanto menor for o valor de NCP, sendo que um valor de zero revela um ajustamento perfeito.

- ***RMSEA*** *(Root Mean Square Error of Approximation).*

$$RMSEA = \sqrt{\frac{F_o}{gl}} \quad \textbf{(30)}$$

sendo F_o a estatística que corresponde ao mínimo relativo do índice NCP, isto é

$$F_o = máx\left[\frac{(\chi^2 - gl)}{n-1}\ , 0\right] = \frac{NCP}{n-1} \quad \textbf{(31)}$$

Segundo Hair *et al*. (2010), RMSEA é um índice absoluto que tenta corrigir a tendência da estatística χ^2 em rejeitar modelos com amostras de grandes dimensões. Tem tendência a favorecer modelos mais complexos. Este índice de ajustamento de modelo permite determinar um intervalo de confiança em torno de seu valor, para além da estimativa pontual. Este intervalo de confiança do RMSEA é assimétrico em torno da estimativa pontual e varia em $[0, +\infty[$ (Wang e Wang, 2012). É um índice que tende a ser sobrestimado para amostras pequenas e para modelos com poucos graus de liberdade. O ajustamento considera-se muito bom se a estimativa pontual do RMSEA for igual ou inferior a 0.05, bom entre 0.05 e 0.08, medíocre entre 0.08 e 0.10 e inaceitável se for superior a 0.10. Numa estimação intervalar, o teste de hipóteses mais comum é o que considera as hipóteses

$$H_o: RMSEA \leq 0.05 \quad \textit{Versus} \quad H_1: RMSEA > 0.05 \quad \textbf{(32)}$$

e que é rejeitada caso

$$p - value = 1 - \Phi(\chi^2\ |0.05^2 \times (n-1) \times g.l., g.l.) \quad \textbf{(33)}$$

obtido com distribuição não central χ^2 seja inferior ao nível de significância previamente fixado. Note-se que os valores referidos para a estimação pontual podem ser usados com a mesma interpretação comparando-os com os limites inferior e superior do intervalo de confiança. Por exemplo, se o limite superior do intervalo for inferior a 0.1, o ajustamento é bom.

V. Índices baseados na Teoria da Informação

Estes índices também são índices relativos de ajustamento mas num sentido diferente dos anteriormente referido, uma vez que com estas estatísticas comparam-se modelos alternativos e a comparação não é feita apenas de um modelo estipulado para o modelo basal. São úteis para comparar modelos não aninhados.

Cada um destes índices reflete a extensão em que as matrizes de covariâncias observadas e previstas diferem umas das outras mas têm um termo que penaliza o modelo em função da sua complexidade e, como referido, são adequados para a comparação de vários modelos alternativos que se ajustem igualmente aos dados. Cada uma dessas estatísticas cria uma medida composta de má qualidade de ajustamento e complexidade, formando uma soma ponderada das duas. Modelos complexos e mal ajustados obtêm pontuações altas.

O melhor modelo é aquele que apresentar os valores menores num ou em mais destes índices, sendo os mais habituais os seguintes:

- ***AIC*** *(Akaike Information Criterion)*

$$AIC = \chi^2 + 2t \quad \textbf{(34)}$$

t é o número de parâmetros livres a estimar no modelo.

Este índice favorece modelos muito complexos em pequenas amostras pelo facto de não ter em conta o efeito da dimensão da amostra na seleção do modelo. Para resolver este problema podem ser usadas estratégias que reduzam o peso do tamanho da amostra na seleção do modelo e melhorar o desempenho do índice – pode usar-se por exemplo variantes *bootstrap* do critério AIC.

- ***BIC*** *(Bayes Information Criterion)*

$$BIC = \chi^2 + t\, Ln(n) \quad \textbf{(35)}$$

n é o número de elementos da amostra. Este índice penaliza mais os modelos complexos e com amostras maiores do que o AIC.

Este critério tem como objetivo selecionar o modelo que mais provavelmente gerou os dados no "sentido *bayesiano*".

- ***ECVI*** *(Expected Cross-Validation Index)*

$$ECVI = \frac{AIC}{n-1} \qquad \textbf{(36)}$$

Este índice pode ser interpretado como a discrepância média nas matrizes de covariância ajustadas entre duas amostras de igual dimensão, em todas as combinações possíveis de duas amostras da mesma população. Reflete o ajustamento teórico do modelo em outras amostras semelhantes à que foi usada para ajustar o modelo, mas a partir de uma única amostra. Estudos mostram que este índice, obtido com uma única amostra, produz resultados de validação cruzada semelhantes aos obtidos com duas amostras independentes.

Este índice é especialmente adequado para comparar modelos não aninhados. O melhor modelo alternativo com melhor ajustamento é o que tem menor valor de ECVI.

Em síntese:

Quadro 3.3: Quadro resumo dos valores de referência dos índices de ajustamento do modelo (adaptado de Portela, 2012).

	Qualidade / **Índice**	**Muito Bom**	**Bom**	**Tolerável**	**Mau**
Índices Absolutos	**CMIM/DF**	≤ 1	$]1,2]$ $]1,3]$	$]2,5]$ $]3,5]$	> 5
	RMSR	Tanto melhor quanto mais próximo de 0			
	GFI	≥ 0.95	$[0.9, 0.95[$	$[0.8, 0.9[$	< 0.8
Índices Relativos	**NFI** **CFI**	≥ 0.95	$[0.9, 0.95[$	$[0.8, 0.9[$	< 0.8
	RFI	Tanto melhor quanto mais próximo de 1			
	TLI	≥ 0.95	$[0.9, 0.95[$	$[0.8, 0.9[$	< 0.8
Índices de Parcimónia	**PCFI** **PGFI** **PNFI**	≥ 0.8	$[0.6, 0.8[$		< 0.6
Índices de Discrepância Populacional	**NCP**	Tanto melhor quanto mais próximo de 0			
	RMSEA	≤ 0.05	$]0.05, \ldots$	$]0.08, \ldots$	> 0.10

Os índices apresentados no quadro 3.3. correspondem apenas a uma amostra da panóplia de índices que têm sido desenvolvidos para estudar o ajustamento global do modelo, em média, em alternativa ao teste χ^2. Um modelo com excelentes índices de

ajustamento não significa necessariamente que seja um modelo correto. Por um lado, há outras componentes do modelo que devem ser consideradas para a avaliação do modelo – as estimativas de coeficientes devem ser interpretáveis, os R^2 das equações devem ser aceitáveis, não deve haver soluções impróprias (variância negativa, correlação menor que -1 ou maior que 1). Por outro lado, pode haver muitos modelos que se ajustam igualmente bem aos dados, a julgar pelos índices de ajustamento do modelo, devendo ser assumido o mais parcimonioso. A avaliação do modelo também tem que ter em conta a teoria e os resultados empíricos, não devendo ser aceite se não fizer sentido substantivo, mesmo que, estatisticamente, se ajuste muito bem aos dados.

3.6.3. Ajustamento local do modelo

Para além da avaliação do ajustamento global do modelo, é necessário estudar o ajustamento local, uma vez que o modelo pode apresentar um mau ajustamento local, mesmo que o ajustamento global seja bom, uma vez que as medidas de ajustamento global são medidas de ajustamento global médio aos dados. Esta situação pode ocorrer porque um ou mais parâmetros do modelo não são significativos ou há uma reduzida fiabilidade de um ou mais indicadores.

A avaliação do ajustamento local pode ser feita através da análise de resíduos, da significância dos parâmetros ou da fiabilidade individual das variáveis observadas (Marôco, 2014).

I. Avaliação dos resíduos estandardizados

Os resíduos estandardizados são estimados pela expressão

$$r_{ij} = \frac{e_{ij}}{\hat{\sigma}_{\varepsilon_{ij}}} \qquad \textbf{(37)}$$

sendo

- e_{ij} o elemento da linha i e da coluna j da matriz $E = S - \Sigma(\hat{\theta})$, dos Mínimos Quadrados Ponderados e
- $\hat{\sigma}_{\varepsilon_{ij}}$ a estimativa do desvio-padrão de e_{ij} estimada por

$$\hat{\sigma}_{\varepsilon_{ij}} = \sqrt{\frac{\hat{\sigma}_{ii}^2\hat{\sigma}_{jj}^2 + \hat{\sigma}_{ij}^2}{n}} \qquad \textbf{(38)}$$

em que $\hat{\sigma}_{ii}^2, \hat{\sigma}_{jj}^2$ e $\hat{\sigma}_{ij}^2$ são os elementos de $\Sigma(\hat{\theta})$.

Para amostras de grande dimensão, $r_{ij} \overset{a}{\sim} N(0,1)$ e portanto $|r_{ij}| > 2$ indicam, com 95% de confiança, valores muito díspares – *outliers*, o que é indicador de problemas de ajustamento local.

II. Avaliação dos erros-padrão assintóticos dos parâmetros do modelo e a sua significância

A significância dos parâmetros γ_{ij} (coeficientes das variáveis observadas exógenas no modelo estrutural) do modelo pode avaliar-se por um teste Z onde as hipóteses são

$$H_o: \gamma_{ij} = 0 \quad \text{Versus} \quad H_1: \gamma_{ij} \neq 0 \tag{39}$$

A estatística do teste é

$$\boldsymbol{Z} = \frac{\hat{\gamma}_{ij}}{\hat{\sigma}_{\gamma_{ij}}} \overset{a}{\sim} N(0,1) \tag{40}$$

onde $\hat{\sigma}_{\gamma_{ij}}$ é a estimativa do erro-padrão assintótico do parâmetro γ_{ij}, estimada pelo elemento correspondente da matriz assintótica de covariância de θ estimada (ACOV), no pressuposto de normalidade multivariada.

Sendo $LL(\hat{\theta})$ a função de log- verosimilhança, então esta pode ser estimada por

$$\boldsymbol{ACOV}(\hat{\theta}) = \left(-E\left(\frac{\partial^2 LL(\theta)}{\partial\theta\, \partial\theta^T}\right)\right)^{-1} \tag{41}$$

Rejeita-se H_0 , ao nível de significância α, para um $p - value < \alpha$, no teste Z.

III. Avaliação da fiabilidade individual das variáveis observadas

A fiabilidade de uma variável observada endógena é estimada pela fração da variância dessa variável que é explicada pela variável latente, conceito idêntico ao coeficiente de determinação da regressão linear. Este valor é igual ou aproximadamente igual ao quadrado do peso fatorial dessa variável, $R_j^2 \cong \lambda_{ij}^2$, valor que é especialmente apropriado para avaliar a relevância das variáveis observadas no modelo de medida. Geralmente,

valores de R^2 inferiores a 0.25 indicam possíveis problemas de ajustamento local com essa variável.

Em síntese, a utilização do teste χ^2 ou dos índices de ajustamento não garante, por si só, a validade do modelo, uma vez que assenta no pressuposto que o ajustamento do modelo é perfeito, o que é irrealista. Por outro lado, neste teste, para amostras grandes aumenta a probabilidade de erro tipo I e para amostras pequenas aumenta a probabilidade de erro tipo II. Da mesma forma, e uma vez que a maioria dos índices de ajustamento não tem distribuição amostral conhecida estes também apresentam problemas. De facto, a decisão sobre a adequação do ajustamento do modelo com base nos valores dos índices é feita com base ou na observação empírica ou em estudos simulados que dependem da especificação do modelo, da dimensão da amostra e dos graus de liberdade, e não em resultados da inferência estatística. Há ainda a salientar, como referido anteriormente, que o ajustamento "médio" global pode não ter correspondência num bom ajustamento local dos parâmetros do modelo.

A avaliação do modelo deve, por tudo isto, resultar da aplicação simultânea de várias medidas de ajustamento global e local que, com resultados coerentes, podem dar ao investigador confiança de que o modelo reproduz convenientemente a estrutura de relações efetivamente existente entre as variáveis.

Além da avaliação do modelo SEM como um todo, pode-se ainda avaliar globalmente a qualidade do ajustamento de cada um dos submodelos.

A qualidade do ajustamento local relativa à globalidade do modelo de medida pode avaliar-se por (Marôco, 2014):

$$R^2 = 1 - \frac{\left|\widehat{\Theta}_{\delta}\right|}{\left|\sum_{xx} \hat{\theta}\right|} \qquad \textbf{(42)}$$

sendo $\left|\widehat{\Theta}_{\delta}\right|$ o determinante da matriz de variância-covariância dos erros do modelo de medida e x e $\left|\sum_{xx} \hat{\theta}\right|$ o determinante da matriz de covariância estimada pelo modelo de medida de x.

Já a qualidade do ajustamento do modelo estrutural pode medir-se globalmente pela fração de variância total das variáveis latentes explicada pelo modelo estrutural – coeficiente de determinação (Marôco, 2014), definida por

$$R^2 = 1 - \frac{\left|\hat{\Psi}\right|}{\left|\sum_{\eta\eta} \hat{\theta}\right|} \qquad \textbf{(43)}$$

sendo $\left|\hat{\Psi}\right|$ o determinante da matriz de variância-covariância dos erros das variáveis latentes e $\left|\sum_{\eta\eta} \hat{\theta}\right|$ o determinante da matriz de covariância das variáveis latentes η.

Note-se que a fração de variância de uma variável latente endógena η_i, explicada pelas suas variáveis preditoras, é dada por

$$R^2_{\eta_i} = 1 - \frac{Var(\zeta_i)}{Var(\eta_i)} \qquad \textbf{(44)}$$

Os valores R^2, seja para o modelo de medida, seja para o modelo estrutural, devem ser superiores ou iguais a 0.25 para se poder considerar que o modelo de medida e/ou o modelo estrutural explicam uma percentagem considerável da variância da(s) variável(eis) observada(s)/latente(s) endógena(s).

3.7. Validade e fiabilidade de um modelo de medida com um constructo reflexivo

Um constructo é reflexivo quando as variáveis latentes se manifestam ou refletem nas variáveis observadas, que estão positivamente correlacionadas. O conjunto de variáveis observadas são as manifestações da variável latente. Assume-se que a ação causal flui da variável latente para os indicadores. O modelo considerado na secção 3.1. (Figura 3.6) é um modelo reflexivo.

A validade de um instrumento de medida é a capacidade que este tem de medir ou operacionalizar, efetivamente, a variável latente que realmente se quer medir.

A fiabilidade de um modelo de medida é a extensão de quão fiável é o referido modelo na medição das variáveis latentes pretendidas, isto é, se a medida é consistente e reprodutível (Marôco, 2014).

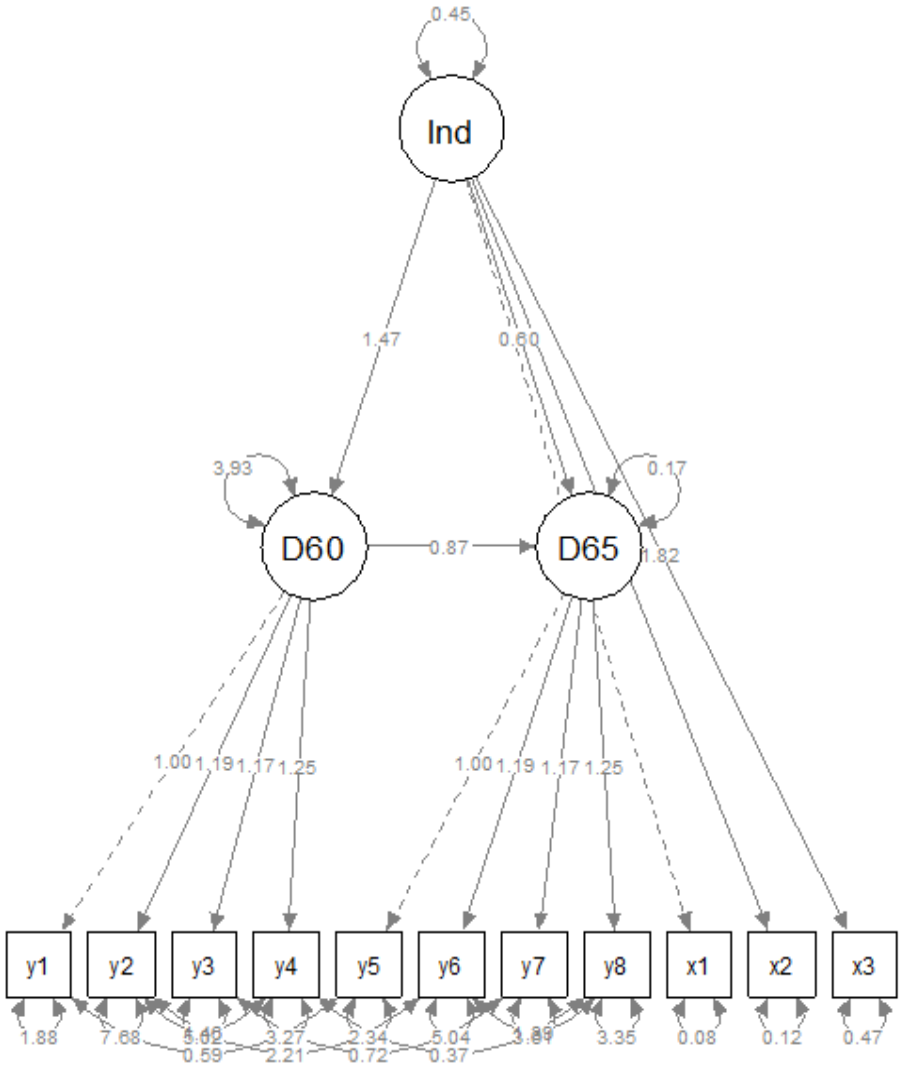


Figura 3.6: Modelo SEM Industrialização e Democracia Política (Bollen, 1989) – modelo reflexivo.

Para proceder à validação do modelo de medida com constructos reflexivos, devem ser verificadas a validade convergente, a validade fatorial, a validade discriminante, a fiabilidade interna.

3.7.1. Validade convergente

A validação convergente avalia o grau em que duas medidas do mesmo constructo, que teoricamente devem estar correlacionadas, estão efetivamente correlacionadas. É indicada pela evidência de que diferentes indicadores de constructos teoricamente semelhantes ou sobrepostos são fortemente correlacionadas.

A validade convergente demonstra-se quando todas as variáveis observadas de constructos teoricamente semelhantes ou sobrepostos num modelo de medida são estatisticamente significativas, isto é, apresentam correlações positivas e elevadas (Marôco, 2014). A validade convergente também se verifica quando a Variância Extraída Média (AVE – ***A**verage **V**ariance **E**xtrated*) for superior a 0.50 ou os pesos fatoriais são superiores a 0.7 (Hair *et al*., 2010).

Para um determinado constructo j, com k indicadores, a AVE é dada por (Marôco, 2014):

$$\widehat{AVE}_j = \frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{k} \lambda_i^2 + \sum_{i=1}^{k} \varepsilon_{ij}} \quad \textbf{(45)}$$

sendo λ_i o peso fatorial estandardizado do indicador i e $\varepsilon_{ij} \approx 1 - \lambda_{ij}^2$ o erro do k-ésimo indicador do constructo j.

3.7.2. Validade fatorial

A validade fatorial ocorre quando os indicadores são reflexo da variável latente que se pretende medir.

É usual assumir que se os pesos fatoriais, λ_{ij}, de todos os indicadores forem superiores ou iguais a 0.50, o constructo apresenta validade fatorial. Idealmente devem ser superiores a 0.70 (Hair *et al*., 2014).

3.7.3. Validade discriminante

A validação discriminante verifica o grau em que um constructo sob estudo é verdadeiramente diferente dos demais. A validade discriminante ocorre quando o modelo de medida não tem variáveis observadas redundantes. Além disso, a raiz quadrada das estimativas AVE, para quaisquer duas variáveis latentes, devem ser maiores que a correlação entre as respetivas variáveis latentes, na lógica de que uma variável latente deve explicar melhor a variância de seu próprio indicador do que a variância de outras variáveis latentes, para fornecer evidências de validade discriminante (Hair *et al*, 2010).

3.7.4. Fiabilidade interna

A fiabilidade interna mede a consistência interna entre os valores medidos dos indicadores de um constructo. Para medida de fiabilidade interna são utilizados o Alfa de Cronbach e a Fiabilidade Compósita. Os indicadores Alfa de Cronbach e Fiabilidade Compósita devem ser preferencialmente maiores ou iguais a 0.70 para indicar a fiabilidade do constructo (Hair *et al*., 2010). O Alfa de Cronbach tem uma relação positiva com o número de indicadores – o aumento do número de indicadores, mesmo com o mesmo grau de intercorrelação, aumenta o valor desta medida –, pelo que, para um grande número de indicadores é necessário dispor de outros critérios, nomeadamente a Fiabilidade Compósita que é uma estimativa de confiança menos enviesada do que o Alpha de Chonbach. O seu valor também deve ser superior ou igual a 0.7 para validar a fiabilidade interna do constructo.

As expressões do Alfa de Cronbach (Kline, 2011) e da Fiabilidade Compósita para um constructo j são dadas (Marôco, 2014), respetivamente, por:

$$\alpha_{C_j} = \frac{n\,\overline{r}_{ij}}{1 + (n-1)\overline{r}_{ij}} \qquad \textbf{(46)}$$

onde n é o número de indicadores e $\overline{r}_{ij}$ é a correlação média de Pearson entre todos os pares de indicadores.

$$FC_j = \frac{\left(\sum_{i=1}^{k} \lambda_{ij}\right)^2}{\left(\sum_{i=1}^{k} \lambda_{ij}\right)^2 + \sum_{i=1}^{k} \varepsilon_{ij}} \qquad \textbf{(47)}$$

sendo λ_{ij} o peso fatorial estandardizado do indicador i e $\varepsilon_{ij} \approx 1 - \lambda_{ij}^2$ o erro do k-ésimo indicador do constructo..

Segundo Kline (2011), os coeficientes de fiabilidade em torno de 0.90 são considerados "excelentes", os valores em torno de 0.80 são "muito bons" e os valores em torno de 0.70 são "adequados". Assim, para valores superiores ou iguais a 0.7 a fiabilidade do constructo é apropriada, embora possam ser aceitáveis valores inferiores (Hair *et al*., 2014).

3.8. Reespecificação do modelo

A variação da qualidade do ajustamento do modelo quando são comparados dois modelos pode ser medida por um teste de razão de verosimilhança com a estatística

$$LR = (n-1)(f_{ML_r} - f_{ML_u}) \quad \textbf{(48)}$$

sendo f_{ML_r} a função de discrepância ML do modelo restrito e f_{ML_u} a função de discrepância com um parâmetro livre (Marôco, 2014). A estatística LR tem distribuição χ^2 com graus de liberdade calculados pela diferença dos graus de liberdade dos dois modelos. O teste avalia a hipótese nula (H_0) que postula que a especificação das variações de pesos fatoriais, das variâncias e covariâncias de fatores e de erros de medida, para o modelo em estudo, são válidas. A estatística de teste de razão de verosimilhança (χ^2), testa simultaneamente a extensão em que esta especificação é verdadeira. O valor de probabilidade associado a χ^2 representa a probabilidade de obter um valor χ^2 que exceda o valor de χ^2 quando a H_0 for verdadeira. Assim, quanto maior a probabilidade associada a χ^2, mais próximo fica o ajuste entre o modelo hipotético (sob H_0) e o ajuste perfeito (Byrne, 2012).

Se o modelo apresenta problemas de ajustamento aos dados, é necessário reespecificá-lo ou modificá-lo. Este processo pode ser feito com o auxílio de um número reduzido de transformações, que podem passar pela exclusão, adição ou alteração de parâmetros do modelo, de forma a melhorar significativamente o ajustamento aos dados. Neste processo podem ser incluídas ou excluídas variáveis, observadas ou latentes, pode considerar-se novas relações ainda não especificadas ou excluir outras, correlacionar erros de medida, ou ainda considerar efeitos mediadores (interações).

Os programas de *software* permitem o cálculo de índices de modificação para estimar a redução ou aumento da estatística χ^2 do modelo se uma das ações anteriormente referidas for implementada para reespecificar o modelo, quando se assumem erros de especificação resultantes do mau ajustamento do modelo aos dados.

Os erros de especificação do modelo podem ser externos ou internos. Quando variáveis irrelevantes são incluídas no modelo ou variáveis substantivamente importantes foram deixadas de fora, ocorre erro externo de especificação e a resolução do problema só pode

ocorrer por meio da reespecificação do modelo com base numa teoria mais relevante. Quando são incluídos caminhos não importantes entre variáveis ou são omitidos caminhos importantes, ocorrem erros internos de especificação que podem ser diagnosticados e solucionados usando estatísticas de Wald (estima a alteração do ajuste por meio da eliminação de parâmetros do modelo - aumento previsto no qui-quadrado se um parâmetro estimado anteriormente fosse fixado em algum valor conhecido, por exemplo, zero) e a estatística do Multiplicador de Lagrange (testa e compara o incremento de ajustamento, a partir da inserção de parâmetros específicos no modelo - redução estimada no qui-quadrado se um parâmetro previamente fixado fosse agora estimado) (Mueller e Hancock, 2008; Ullman, 2007).

Depois de um modelo ser especificado novamente, o processo descrito recomeça e esta tarefa pode repetir-se até que se obtenha um modelo que apresenta um bom ajustamento aos dados.

Importa não esquecer que toda e qualquer modificação no modelo só deve ser feita com base na teoria, a não ser que alguma conclusão empírica suporte, fortemente, a definição de novas hipóteses que questionem a teoria existente (Byrne, 2012). O uso cego de índices de modificação pode levar os investigadores a adotar modelos que se desviam de objetivos substantivos originais, sendo imperativo considerar apenas a alteração de parâmetros que têm uma interpretação substantiva clara (Raykov e Marcoulides, 2006).

Além disso, é necessário ter em conta que não existe nenhum modelo que se adeque perfeitamente à realidade. Um modelo modificado deverá passar por uma validação cruzada – com dados diferentes daqueles usados para estimar o modelo anterior – antes de ser aceite. De facto, na tentativa de encontrar o "melhor" modelo corre-se o risco de ter um modelo que, ao ser objeto de grandes modificações para se adequar à amostra em estudo, este não se adeque a outras amostras ou ao universo populacional. Qualquer estratégia de geração de modelos deve estar sujeita a condições - deve-se reconhecer que o modelo resultante é em parte orientado por dados, as modificações devem ser substantivamente significativas e o modelo modificado deve ser avaliado, ajustando-o a uma amostra independente (MacCallum e Austin 2000).

CAPÍTULO 4

MODELOS DE EQUAÇÕES ESTRUTURAIS COM DADOS OMISSOS

MODELOS DE EQUAÇÕES ESTRUTURAIS COM DADOS OMISSOS

4.1. Introdução

Um problema comum em SEM prende-se com a existência frequente de dados omissos (*missing data*), problema transversal a todos os tipos de análises de dados.

Os métodos mais comuns de implementação da modelação SEM com dados incompletos são aqueles em que se utiliza o estimador de Máxima Verosimilhança com matrizes de covariâncias estimadas com dados completos, após proceder à correção dos dados, com recurso a metodologias para resolver o problema dos dados omissos que se baseiam, tradicionalmente, em procedimentos *ad hoc* e que não possuem suporte teórico.

Destes procedimentos podemos citar métodos de exclusão, como *listwise deletion e pairwise deletion*, ou métodos de substituição única, como imputação pela média *(Mean imputation)* ou imputação por regressão (*Regression imputation)*, entre outros. Tal como acontece em outros métodos estatísticos, os dados omissos geralmente criam grandes problemas para a estimativa de Modelos de Equações Estruturais. A utilização dos métodos convencionais referidos para lidar com este problema, em muitas situações, conduz a estimativas enviesadas. Métodos de Máxima Verosimilhança adequados para lidar com dados omissos, como o método de Máxima Verosimilhança de Informação Completa (*Full-Information Maximum Likelihood* – FIML), e métodos de substituição múltipla de dados omissos, como a Imputação Múltipla (*Multiple Imputation* – MI), possuem muito melhores propriedades estatísticas do que os métodos convencionais, sob pressupostos consideravelmente mais fracos, uma combinação rara nos métodos estatísticos.

A implementação destes métodos em *softwares* de modelação SEM veio facilitar a sua utilização. No entanto, o facto de esta implementação ocorrer em *softwares* comerciais limitou a sua proliferação. O *software* R veio colmatar este problema com pacotes como `sem` (Fox *et al*., 2017), `lavaan` (Rosseel, 2018), `lavaan.survey` (Oberski,2014), específicos para a implementação da SEM e que incorporam mecanismos para lidar com dados omissos, ou o `Amelia II` (Honaker, King & Blackwell, 2011) para imputação múltipla, entre outros.

Neste capítulo, será feito o enquadramento teórico da SEM com dados omissos, em particular será feita uma revisão dos mecanismos que conduzem a dados omissos e dos procedimentos para lidar com estes dados no contexto da SEM. Será dada especial atenção aos recursos disponíveis no *software* R para lidar com os dados omissos neste contexto.

4.2. Dados omissos – Enquadramento teórico

O modelo de equações estruturais é definido por dois submodelos. O submodelo estrutural, que define as relações causais ou de associação hipotéticas entre os constructos não observados (variáveis latentes), especificando se uma variável latente causa mudanças noutras variáveis latentes no modelo, direta ou indiretamente. O submodelo de medida, que relaciona as variáveis observadas e as variáveis não observadas, constitui-se como uma ligação entre o instrumento de medida (variáveis observadas) e os constructos teóricos em estudo.

O modelo pode ser representado por um diagrama (*path diagram*) que se traduz matematicamente num conjunto de equações que especifica as relações entre as variáveis e que constituem os submodelos de medida e estrutural.

Os objetivos da análise de equações estruturais são, como em qualquer análise de dados:

(i) Obter estimativas não enviesadas dos parâmetros;
(ii) Obter uma boa avaliação da variabilidade em torno destas estimativas (estimar erros padrão ou intervalos de confiança);
(iii) Maximizar o poder estatístico no processo.

A análise de equações estruturais com dados omissos assenta nos mesmos objetivos, mas estes ficam comprometidos com a omissão de dados numa ou mais variáveis. Daí que a forma como o investigador lida com esta situação, que é a mais comum na análise de dados reais, seja de extrema importância.

Antes de mais, importa referir que os mecanismos que originam a omissão de dados influenciam as metodologias ao dispor do investigador para lidar com este problema. O mecanismo gerador da omissão de dados deve ser tomado como pressuposto para a aplicação de técnicas de análise com dados omissos.

Rubin (1976) definiu três mecanismos que causam dados omissos: omissão de dados completamente ao acaso (MCAR – *Missing Completely At Random)*, omissão de dados ao

acaso (MAR - *Missing At Random)* e omissão de dados não ao acaso (MNAR - *Missing Not At Random).* Estes mecanismos podem ser entendidos como explicações probabilísticas dos dados omissos. Operacionalmente, a omissão numa variável X é definida como uma variável indicadora r que assume o valor 1 se X é observada e 0 se X estiver em falta. Portanto, r não é, por si só, uma probabilidade. É uma variável aleatória que tem uma distribuição de probabilidade definida pela probabilidade de assumir o valor 0 ou 1, condicionada pelo valor de alguma outra variável. No caso de a omissão ser resultante de um mecanismo MCAR, a probabilidade de a observação estar omissa numa variável X é independente de qualquer variável Y que tenha sido observada ou que esteja omissa: $Pr(r|(y_{obs}, y_{miss})) = Pr(r)$. No caso do mecanismo MAR a probabilidade da omissão de uma observação depende apenas dos valores dos dados observados: $Pr(r|(y_{obs}, y_{miss})) = Pr(r|y_{obs})$. Se a omissão resulta de um mecanismo MNAR, a probabilidade de omissão de uma observação depende dos valores das variáveis não observadas: $Pr(r|(y_{obs}, y_{miss})) = Pr(r|y_{miss})$. Neste último caso a observação omissa é dita informativa.

Por outro lado, como referido, estes mecanismos representam pressupostos que determinam as condições em que um dado método de análise, adequado para lidar com os dados omissos, proporcionará um desempenho ótimo (Enders, 2010).

Consideremos uma matriz A de dados completos, com M linhas, as quais representam as M unidades amostrais, N colunas, as quais representam as N variáveis, e com linhas $a_i = (a_{i1}, \ldots, a_{iN})$, onde a_{ij} é o valor da variável j para a unidade i. Pode-se dividir A em dois conjuntos, o conjunto dos dados observados e o conjunto dos dados omissos:

$$A = \{A_{obs}, A_{miss}\} \quad \textbf{(49)}$$

onde A_{obs} contém os dados observados (não-omissos) e A_{miss} contém os dados omissos.

Para cada matriz de dados A, existe um identificador de dados omissos, uma matriz R, com a mesma dimensão de A, onde $r_{ij} = 1$, se a_{ij} é observado, e $r_{ij} = 0$, caso contrário.

O mecanismo de dados omissos é caraterizado pela distribuição condicional de R dado A, $P(R \mid A)$, a qual pode ser de três tipos, como referido: MCAR, MAR ou NMAR (Little & Rubin, 2002).

4.2.1. Dados Omissos Completamente ao Acaso (MCAR – *Missing Completely At Random*)

Os dados omissos não dependem dos valores de A, isto é:

$$P(R \mid A) = P(R) \quad \textbf{(50)}$$

Um mecanismo MCAR corresponde a uma situação em que os dados observados da variável com dados omissos foram gerados por um processo aleatório. Os valores dos dados observados constituem uma amostra aleatória representativa do conjunto de dados hipoteticamente completo (Peugh & Enders, 2004; Graham, 2012). Os valores dos dados omissos numa variável são como uma amostra aleatória simples dos dados dessa variável e portanto a distribuição dos valores omissos é a mesma que a dos valores observados (Zhang, 2003; Graham, 2012). Mesmo que o evento responsável pela omissão de dados não seja totalmente aleatório, o mecanismo é MCAR se os valores omissos da variável particular X não estão correlacionados com outras variáveis no conjunto de dados, nem com os valores subjacentes da própria variável X (Peugh & Enders, 2004). A falta de dados numa variável X não está relacionada nem com os valores observados, nem com os valores em falta nessa variável, ou com qualquer valor das restantes variáveis.

A grande vantagem deste mecanismo é que a causa que levou aos dados omissos não precisa fazer parte da análise para controlar a influência destes nos resultados da pesquisa (Peugh e Enders, 2004). Por este facto, embora o pressuposto MCAR seja forte, há situações em que é razoável, especialmente quando os dados estão em falta devido a decisões do processo de recolha (Allison, 2003). É de notar que, apesar das vantagens associadas ao facto dos dados serem MCAR, nem todos os métodos capazes de lidar com dados omissos produzirão bons resultados com este tipo de dados.

4.2.2. Dados Omissos ao Acaso (MAR – *Missing at Random*)

Os dados omissos não dependem dos valores de A_{miss} mas apenas dos valores de A_{obs}, isto é:

$$P(R \mid A) = P(R|A_{obs}) \quad \textbf{(51)}$$

Neste caso, os dados omissos são causados por alguma variável observada, disponível para análise e correlacionada com a variável que possui dados omissos (Peugh & Enders, 2004). Ainda neste caso, existência de dados omissos numa variável X não tem relação

com os valores que a variável assume nas outras observações, podendo, ou não, ter relação com as restantes variáveis. A ocorrência de um mecanismo MAR significa que existe uma relação sistemática entre uma ou mais variáveis medidas e a probabilidade de omissão de dados (Enders, 2010). Os valores omissos para uma variável são como uma amostra aleatória simples dos dados para essa variável, dentro de subgrupos definidos pelos valores observados, e a distribuição de valores omissos é a mesma que a distribuição dos valores observados dentro de cada subgrupo (Zhang, 2003).

O mecanismo de omissão de dados MAR tem um problema prático, uma vez que não há como confirmar que a probabilidade de dados omissos numa variável é apenas função de outras variáveis medidas. Este facto compromete os resultados da aplicação das metodologias mais comumente usadas para lidar com dados omissos: a estimação por máxima verosimilhança e a imputação múltipla, que assumem um mecanismo MAR na origem dos dados omissos (Enders, 2010).

A(s) variável(eis) que é (são) causa da omissão de dados deve(m) ser incluída(s) na análise para controlar todas as influências causadas por ela(s). Se alguma causa da omissão, mesmo que disponível, não for incluída no modelo de análise de dados omissos, então, por definição, a omissão é MNAR, e haverá um viés de estimativa associado com os dados omissos (Peugh & Enders, 2004; Graham, 2012).

4.2.3. Dados Omissos Não ao Acaso (MNAR – *Missing Not At Random*)

Quando a distribuição de R depende dos dados omissos contidos na matriz A (A_{miss}), podendo também depender dos dados observados (A_{obs}), tem-se:

$$P(R \mid A) \neq P(R|A_{obs}) \qquad \textbf{(52)}$$

Neste caso, o mecanismo da omissão de dados não é MAR mas MNAR. Num mecanismo MNAR a causa da omissão de dados numa variável é ela própria, podendo ser causa dos seus próprios valores, observados ou omissos.

Como, neste caso, a causa da omissão não pode ser incluída no modelo, este tipo de mecanismo produz um viés de estimativa associado com os dados omissos.

A teoria desenvolvida por Rubin (1976) sobre dados omissos envolve dois conjuntos de parâmetros: os parâmetros que abordam as questões de pesquisa substantivas (aqueles que o investigador estimaria sem dados omissos) e os parâmetros que descrevem a probabilidade de perda de dados. No entanto, geralmente, não há como determinar ou estimar os parâmetros que descrevem a propensão para os dados omissos. Ao contrário do que acontece no mecanismo MAR, onde se podem estimar os dados omissos a partir das variáveis completas, no caso de dados omissos do tipo MNAR tal estimativa não pode ser obtida, já que existe uma relação entre as omissões de dados e os valores que existiriam nessas omissões.

Rubin (1976) clarificou condições que garantem estimativas precisas dos parâmetros substantivos sem conhecer os parâmetros da distribuição de dados omissos e que dependem das técnicas usadas para analisar os dados. Análises cujas técnicas dependem de uma distribuição amostral, como é o caso das *listwise deletion e pairwise deletion*, são válidas apenas quando os dados são MCAR. Análises baseadas na verosimilhança, como a estimação por máxima verosimilhança ou a imputação múltipla, dispensam informação sobre os parâmetros da distribuição dos dados omissos se estes forem MAR ou MCAR e por isso, neste caso, são ditos *ignoráveis* no sentido em que são considerados mais fáceis de lidar, dado que os seus efeitos nos modelos estão disponíveis para o analista. No entanto, as análises baseadas na verosimilhança produzem viés com dados MNAR. As metodologias desenvolvidas para análise de dados MNAR exigem pressupostos muito restritivos que limitam a sua utilidade prática (Enders, 2010).

O mecanismo MCAR não deve ter grande impacto na estimação dos parâmetros substantivos, pois os dados omissos ocorrem de forma completamente aleatória. Quando o mecanismo de omissão é MAR, existe um processo sistemático subjacente à falta de dados que pode ser modelado através dos dados observados. Ao contrário, para o mecanismo MNAR, dito *não-ignorável*, não existe nenhuma informação dentro do conjunto de dados que permita modelar e compreender a maneira como os dados omissos aconteceram. Este facto, leva a que o efeito deste mecanismo seja desconhecido e potencialmente perigoso, devendo ser modelado para que sejam obtidas boas estimativas dos parâmetros de interesse.

4.3. Diagnóstico

Diagnosticar o mecanismo gerador de dados omissos ajuda o investigador a entender a natureza dos dados omissos e o seu potencial impacto nos resultados dos estudos e nas interpretações destes.

Na maioria das situações de dados omissos, não se consegue obter os dados perdidos. Uma primeira abordagem será a de examinar padrões nos dados para ter uma ideia de qual será o mecanismo mais provável. Traçar padrões de dados omissos pode revelar padrões inesperados que não foram detetados durante as etapas de recolha de dados. No entanto, padrões de dados omissos, sozinhos, não informam sobre quais são os mecanismos que estão subjacentes a essa omissão.

Em princípio, é possível verificar se um conjunto de dados é MCAR, uma vez que é o único mecanismo de dados omissos que produz proposições testáveis.

Testar se uma coleção inteira de variáveis é consistente com MCAR, por um lado, pode ser demasiado fastidioso pois poderá exigir um elevado número de testes e, por outro, provavelmente não é útil. De facto, é altamente improvável que todas as variáveis com dados omissos de um conjunto de dados sejam MCAR e algumas terão dados omissos de forma sistemática. Acresce o facto de que encontrar evidências da consistência, ou não, dos dados MCAR não altera a recomendação para usar a máxima verosimilhança ou a imputação múltipla nas análises, por exigirem apenas o pressuposto MAR que é menos rigoroso que MCAR. No entanto, identificar variáveis individuais que não são MCAR é potencialmente útil porque pode haver uma relação entre essas variáveis e a probabilidade de omissão (Enders, 2010). A incorporação de causas de omissão no tratamento de dados omissos é recomendada, pois pode mitigar o viés e melhorar a possibilidade de os dados satisfazerem a suposição de MAR (Schafer & Graham, 2002).

MCAR exige que os dados observados constituam uma amostra aleatória simples do conjunto de dados hipoteticamente completo, o que implica que os casos com dados omissos pertençam à mesma população e, portanto, compartilham o mesmo vetor de médias e a mesma matriz de covariância dos casos com dados completos - homogeneidade de médias e covariâncias. Separar os casos completos e os casos com dados omissos numa variável e aplicar testes para avaliar a igualdade de médias e/ou de variâncias entre grupos

definidos pelos valores observados de outra variável permite recolher evidências de que os dados são MCAR, ou não, relativamente a essa variável.

Foram propostos diversos métodos para testar se os dados omissos são MCAR (Little, 1988; Chen & Little, 1999; Kim & Bentler, 2002; Enders, 2010; Jamshidian & Jalal, 2014; Li, 2013).

A título de exemplo, considere-se testes t, independentes, aplicados a grupos de dados omissos (ou observados) com padrões comuns, para cada variável com dados omissos ou entre o grupo com dados completos e aquele com dados omissos. A aplicação destes testes permite avaliar se a omissão em cada variável está relacionada com os valores observados de outras variáveis. Se todos os t-testes forem não significativos, então pode-se assumir que os dados omissos nesse conjunto de dados são MCAR; se não, são MAR ou MNAR nas variáveis com testes significativos. No entanto, à medida que o tamanho da matriz de dados cresce, a avaliação de múltiplos testes t torna-se fastidiosa e, no que respeita ao desempenho, podem resultar erros de Tipo I. Para além do grande número de estatísticas t, estes testes não têm em consideração as correlações entre as variáveis, sendo possível que um indicador de dados omissos produza diferenças de médias em várias variáveis, mesmo que exista apenas uma única causa para os dados omissos. Seguindo a mesma lógica pode ser testada a igualdade de variâncias e covariâncias entre o grupo de dados completos e o grupo com dados omissão, em relação a cada uma das variáveis (teste de Levene, teste de Bartlet ou teste F).

Também se pode recorrer ao teste de Little (1988), extensão multivariada da abordagem de teste t, baseado num quociente de verosimilhança. Este teste é destinado a avaliar simultaneamente diferenças de médias entre vários subgrupos de casos que compartilham o mesmo padrão de dados omissos (Enders, 2010). A rejeição da hipótese de igualdade de médias entre os grupos indica que os dados não são MCAR. Kim e Bentler (2002) apresentam dois testes propostos por Little para testar homogeneidade das matrizes de covariâncias.

O teste de Little para a igualdade de médias tem alguns problemas que importa considerar. O teste não identifica as variáveis específicas que violam o MCAR e pressupõe que os padrões de dados omissos compartilham uma matriz de covariância comum. Estudos de simulação sugerem que o teste de Little é pouco potente especialmente quando

o número de variáveis que violam MCAR é pequeno, a relação entre os dados omissos é fraca ou os dados são MNAR (Enders, 2010).

As comparações de médias não fornecem um teste conclusivo da coerência com MCAR porque os mecanismos MAR e MNAR podem produzir subgrupos de dados omissos com as mesmas médias.

Kim e Bentler (2002) propuseram um teste semelhante ao de Little mas baseado no raciocínio dos Mínimos Quadrados. Este teste tem menos pressupostos que o de Little, não exigindo que o número de observações seja, no mínimo, igual o número de variáveis, nem a normalidade dos dados.

Jamshidian e Jalal (2010) propuseram um teste à homocedasticidade para dados que respeitam o pressuposto da normalidade e um teste não paramétrico para dados que não respeitam este pressuposto, baseados em métodos adequados para dados completos e na imputação de dados.

O *software* R dispõe de pacotes que disponibilizam o teste de Little: `BaylorEdPsych` (Beaujean e Beaujean, 2018) e o pacote `MissMech` (Jamshidian, Jalal & Jansen, 2014). Este último pacote dispõe de uma função para testar a homocedasticidade, a normalidade multivariada e MCAR de dados omissos, seguindo a metodologia proposta por Jamshidian e Jalal (2010).

Se o mecanismo não for MCAR, é necessário saber se o mecanismo que criou os dados omissos é relacionado com as informações conhecidas, mas não existe nenhum método formal para esse efeito. Assim, se os dados não são MCAR não há testes estatísticos que permitam distinguir dados MAR ou MNAR, uma vez que o que distingue a distribuição de probabilidades dos dados MAR e MNAR são os dados omissos e não se tem como saber os valores que tomariam. Também não há técnicas visuais que auxiliem nesta distinção. Uma das fontes de distinção pode ser o conhecimento teórico que se tem do problema.

Porém, segundo Schafer (1997), existem algumas situações, nomeadamente quando as omissões são planeadas, em que o pressuposto de que o mecanismo é ignorável (a(s) causa(s) da omissão é (são) incorporada(s) na análise) é bastante plausível e as simplificações analíticas que resultam dessa hipótese são altamente benéficas.

Assim, se:

- Algumas informações são recolhidas de todos os objetos da base de dados, e outras informações adicionais são recolhidas apenas de um subgrupo da amostra original, sendo que esse subgrupo é selecionado devido a alguma informação recolhida na amostra toda;
- Os investigadores podem substituir os objetos incompletos por outros completos, com as mesmas características;
- Em testes controlados aleatoriamente, em que o número de objetos, nas diferentes intervenções, é não equilibrado devido a causas inesperadas e não devido a um processo sistemático;
- As informações são recolhidas de uma amostra e, posteriormente, informações adicionais são recolhidas de um subgrupo selecionado aleatoriamente, ou selecionado baseado nas informações recolhidas previamente;
- o estudo é longitudinal e mede-se uma subamostra em cada ponto de tempo.

o mecanismo pode ser considerado pelo menos MAR. Caso contrário, deve ser considerado MNAR.

Depois de avaliados os dados, para assumir qual o mecanismo que levou à existência de dados omissos, passa-se à seleção da metodologia de análise SEM com dados omissos.

4.4. Metodologias de análise de dados omissos no contexto da Modelação de Equações Estruturais

Ir-se-á analisar as abordagens tradicionais de exclusão de dados omissos, como *listwise deletion e pairwise deletion*, ou métodos de substituição como *mean imputation* ou *regression imputation* e abordagens baseadas na Máxima Verosimilhança e na Imputação Múltipla. Estas duas últimas abordagens, sob pressupostos idênticos, produzem estimativas consistentes, assintoticamente eficientes e assintoticamente normais.

4.4.1. *Listwise Deletion* e *Pairwise Deletion*

Estes métodos são métodos de exclusão de dados e que eram tradicionalmente (até 1987) usados por se basearem em algoritmos de análise de dados completos, únicos algoritmos disponíveis nos programas de implementação de SEM.

No método *listwise deletion* ou *complete–case analysis* são eliminados todos os casos com qualquer dado em falta.

Esta metodologia é simples e geral e tem duas propriedades estatísticas importantes. Primeiro, se os dados forem MCAR, a exclusão dos casos com dados omissos não introduzirá nenhum enviesamento nas estimativas dos parâmetros: a subamostra com dados completos é efetivamente uma amostra aleatória simples da amostra original e portanto não é introduzido qualquer viés. Segundo (e pelo mesmo motivo), as estimativas de erro padrão sob *listwise deletion* devem ser estimativas aproximadamente imparciais dos verdadeiros erros padrão. Por outro lado, se os dados não são MCAR a *listwise deletion* pode conduzir a estimativas enviesadas dos parâmetros. Para além do enviesamento, há ainda a considerar que esta metodologia pode levar a uma grande redução do número de elementos da amostra na análise, especialmente se o modelo tem muitas variáveis e cada variável tem pelo menos uma observação com um dado omisso, o que compromete a inferência sobre os parâmetros. Mesmo que os erros padrão produzidos sejam estimativas aproximadamente não enviesadas dos verdadeiros erros padrão, esses erros padrão podem ser substancialmente maiores que os que se obtêm com métodos que preservaram mais dados disponíveis, o que resulta em intervalos de confiança mais amplos e em testes de hipóteses com menor poder do que os produzidos por métodos mais eficientes (Allison, 2003; Peugh & Enders, 2004).

Em algumas situações não MCAR é possível reduzir os enviesamentos resultantes da exclusão de casos com a aplicação de pesos. Após a remoção dos casos incompletos, os casos completos restantes são ponderados de modo que sua distribuição se assemelhe mais à da amostra completa ou da população em relação a variáveis auxiliares. Os pesos são obtidos das probabilidades de resposta, que devem ser estimadas a partir dos dados (por exemplo, por uma regressão logística ou *probit*). A ponderação pode eliminar o viés devido à resposta diferencial relacionada com as variáveis usadas para modelar as probabilidades de resposta, mas não pode corrigir o viés relacionado com as variáveis que não são utilizadas ou não são medidas (Schafer & Graham, 2002). Apesar de ser uma técnica não paramétrica, por não exigir um modelo para a distribuição dos valores omissos na população, exige algum modelo para a probabilidade de resposta, o que a pode tornar pouco atrativa, na medida em que pode ser necessário calcular um conjunto diferente de pesos para cada variável.

O método *pairwise deletion* tenta usar todos os dados disponíveis, não descartando casos análise da análise. É frequentemente descrito no contexto de uma matriz de covariâncias (ou correlações). No âmbito da análise SEM, o método é usado na matriz de covariâncias, em que cada elemento é a variância ou covariância estimada com todos os dados completos disponíveis para cada variável ou par de variáveis. Se os dados forem MCAR, a *pairwise deletion* produz estimativas de parâmetros consistentes e, portanto, aproximadamente não enviesados. Mas a metodologia tem limitações importantes. Por um lado, a matriz de covariâncias, com pares eliminados, pode não ser definida positiva, o que implica que os parâmetros para muitos modelos lineares não possam ser estimados. Por outro lado, e talvez o mais importante, as estimativas de erros padrão obtidas sob *pairwise deletion* não são estimativas consistentes dos verdadeiros erros padrão, o que põe em causa a validade dos intervalos de confiança e dos testes de hipóteses (Allison, 2003). A estimativa de erro padrão requer especificar o tamanho da amostra, e não há nenhuma maneira óbvia de o fazer com a *pairwise deletion*. Esta metodologia não é recomendada para a análise SEM (Peugh & Enders, 2004).

4.4.2. Imputação de dados - métodos de substituição

Uma abordagem possível para a imputação de dados na análise de dados omissos consiste em fazer algumas suposições razoáveis para os valores dos dados omissos, usar esses valores para os substituir e proceder a uma análise convencional dos dados completos (os dados reais mais os dados imputados).

Na literatura têm sido propostas várias formas de proceder à imputação de dados, as quais se enquadram num de dois tipos: simples ou múltipla. Na imputação simples cada valor ausente é substituído por um único valor imputado. Na imputação múltipla substitui-se cada valor em falta por mais de um valor imputado (m valores), resultando em m conjuntos de dados completos que são analisados por um método convencional para dados completos e posteriormente combinados de forma simples e apropriada, resultando assim num conjunto de dados completos. Os conjuntos de dados assim completados são analisados e usados para estimar um valor plausível que representa a incerteza sobre o valor a ser imputado (Little & Rubin, 2002).

Na imputação simples pode ser usada, por exemplo, (1) a imputação, em todos os valores omissos da variável com dados omissos, do valor médio ou do valor mediano dessa variável, obtidos com os dados observados, opção que produz estimativas enviesadas de

muitos parâmetros; (2) podem ser imputados os valores previstos (média condicional) para os dados omissos, obtidos com uma equação de regressão dos dados observados ou por interpolação, opções que, sob o pressuposto MCAR, produzem estimativas aproximadamente imparciais. No entanto, verifica-se uma tendência geral para a imputação da média condicional produzir subestimações de variâncias e superestimar as correlações (Allison, 2003).

Todos os métodos convencionais de imputação levam a subestimar os erros padrão, por presumirem que todos os dados são reais, mesmo que se possa evitar o viés nas estimativas dos parâmetros (Little & Rubin, 2002). Se alguns dados são imputados, o processo de imputação introduz variabilidade amostral adicional que não é devidamente contabilizada. Por outro lado, não se deve esquecer que a variância, geralmente, é subestimada com a utilização destes métodos.

Outras metodologias mais recentes como a FIML (*Full-Information Maximum Likelihood*) ou a Imputação Múltipla (IM) estão disponíveis em programas de implementação da SEM, em particular em pacotes do *software* R como o `sem` (Fox *et al*., 2017), `lavaan` (Rosseel *et al*., 2018), `Amelia II`, (Honaker, King & Blackwell, 2011), `mice` (Buuren & Groothuis-Oudshoorn, 2011), entre outros. Estas metodologias, sob determinados pressupostos, nomeadamente a normalidade multivariada e MAR, produzem estimativas de parâmetros que possuem propriedades, para grandes amostras, como consistência, eficiência assintótica e normalidade assintótica.

4.4.3. *Full-Information Maximum Likelihood* (FIML)

O objetivo básico da estimação por Máxima Verosimilhança (MV) é identificar os valores dos parâmetros populacionais mais prováveis de terem produzido uma determinada amostra de dados. Na FIML o ajustamento dos dados a um determinado conjunto de valores de parâmetros é avaliado por um valor de log-verosimilhança que quantifica a probabilidade relativa de uma amostra particular, no pressuposto que os dados provêm de uma população normal multivariada. Calcula-se uma função de verosimilhança para cada caso, usando apenas as variáveis que são observadas para o caso i.

A função de log-verosimilhança para o caso i é dada por:

$$l_i(\theta|Y) = k_i - \frac{1}{2} ln|\Sigma_i| - \frac{1}{2}(y_i - \mu_i)\Sigma_i^{-1}(y_i - \mu_i) \quad \textbf{(53)}$$

em que k_i é uma constante (é um fator de escala que depende do número de dados não omissos para o caso i e pode ser ignorado durante a estimação), y_i é o vetor de valores observados para o caso i, μ_i é o respetivo vetor de médias e Σ_i é a respetiva matriz de covariâncias. É importante notar que a derivação desta equação depende explicitamente da hipótese de normalidade multivariada. Embora as estimativas de parâmetros tendam a ser precisas quando os dados não são normais, os erros padrão serão muito baixos, resultando em taxas de erro de Tipo I elevadas (Enders, 2010).

O aspeto importante da equação (53) reside no facto de o vetor de dados observados não precisar ser completo - o tamanho e o conteúdo das matrizes de parâmetros (μ_i e Σ_i) são ajustados, de modo que $l_i(\theta|Y)$ é calculada usando apenas as variáveis e os parâmetros para os quais o caso tem dados completos. As linhas e colunas correspondentes a dados omissos são removidos e o ajustamento dos dados brutos aos parâmetros é baseado apenas nos dados observados.

Adicionando as log-verosimilhança para todos os N elementos da amostra, obtém-se um valor de log-verosimilhança que quantifica a probabilidade relativa de os dados serem provenientes de uma população normalmente distribuída com um vetor de médias μ e matriz de covariância Σ, particulares.

$$l(\theta|Y) = \sum_{i=1}^{N} l_i(\theta|Y) \qquad \textbf{(54)}$$

Em cada iteração do processo de estimação, os valores de μ e Σ são ajustados na tentativa de identificar o conjunto de valores com a maior log-verosimilhança (ou seja, maior probabilidade de produzir os dados da amostra) (Peugh & Enders, 2004).

Não é óbvio, mas a inclusão de casos com dados parciais contribui para a estimação de todos os parâmetros. De facto, embora os valores em falta não sejam imputados durante o processo de estimação, os dados parciais implicam valores prováveis para os dados omissos através das correlações entre as variáveis. O algoritmo FIML não imputa valores omissos, mas o "empréstimo" de informações da parte observada dos dados é conceitualmente análogo à substituição de pontos Y de dados ausentes pela expectativa condicional de Y, dado X (Enders & Bandalos, 2001).

A abordagem FIML, também referida como a Máxima Verosimilhança dos dados brutos, usa os dados brutos, caso a caso, e maximiza a função de Máxima Verosimilhança, caso a caso, usando toda a informação disponível para cada caso. A combinação das estimativas de todos os casos produz uma estimativa geral da função de MV. Os *softwares* que usam esta metodologia proporcionam excelentes estimativas de parâmetros (MV) para o modelo em estudo e também fornecem erros padrão razoáveis, num único passo (Grahm, 2012).

A estimação FIML é implementada recorrendo aos algoritmos convencionais de otimização para a estimação dos parâmetros. Uma das formas possíveis de implementação da FIML corresponde à utilização do algoritmo EM (*Expectation- Maximization*) para estimação dos parâmetros. O algoritmo EM usualmente é implementado em duas etapas para produzir estimativas MV da matriz de variâncias-covariâncias e do vetor de médias: a etapa E (*Expectation*) em que, em cada iteração, os dados omissos são substituídos por melhores valores preditos por modelos de regressão estimados com estatísticas suficientes (média, variâncias e covariâncias) estimadas ou assumidas na iteração anterior e a etapa M (*Maximization*), em que os parâmetros são re-estimados por Máxima Verosimilhança com os dados completos resultantes da imputação. O processo repete-se até que haja convergência. Para obter as estimativas de FIML usando o algoritmo EM, na etapa E, os valores omissos são substituídos por valores previstos, assumindo parâmetros conhecidos do modelo. Na etapa M, os parâmetros são re-estimados por ML assumindo dados completos.

O algoritmo EM também pode ser usado na FIML em duas etapas. Na primeira etapa obtém-se estimativas ML da média e das variâncias e covariâncias, com recurso ao algoritmo EM. Na segunda etapa, essas estimativas são usadas para obter estimativas dos pesos fatoriais, das variações de erro e assim por diante. Neste caso pode-se obter erros padrão demasiado pequenos (Hirose, 2013; Graham & Coffman, 2012).

Quando o algoritmo EM é usado com a FIML numa única etapa, as estimativas obtidas são os pesos fatoriais, variâncias de erro e assim por diante. Não envolve a obtenção de uma estimativa ML de médias, variâncias e covariâncias via EM e a entrada dessas estatísticas num programa de *software* SEM para estimar os parâmetros do modelo via ML, nem envolve a realização de Imputação Múltipla seguida pela adaptação do SEM a cada

um dos conjuntos de dados completos e pela combinação dos resultados (Graham & Coffman, 2012).

A metodologia FIML produz erros padrão e teste qui-quadrado (*likelihood ratio*) corretos quando os dados são MAR e com distribuição normal multivariada. As estimativas de erro padrão para FIML com o algoritmo EM situam-se entre as estimativas otimistas e pessimistas produzidas pelo método EM em duas etapas (Allison, 2003).

Dadas as vantagens, a metodologia FIML parece ser o melhor método para lidar com dados omissos, para a maioria das aplicações SEM. No entanto, não é de descurar a exigência dos pressupostos sobre os dados omissos – devem ser MAR e os dados devem ter distribuição normal multivariada. Para além de ser difícil que os dados empíricos sejam normalmente distribuídos e dificilmente se possa garantir que sejam MAR, acresce a grande dificuldade em testar este último pressuposto. Os erros padrão não são produzidos como subproduto da estimativa de parâmetros quando se utiliza a FIML com o algoritmo EM, o que resulta na única desvantagem da metodologia FIML implementada com este algoritmo de otimização. A forma mais comum para estimar erros padrão com estimativas EM é usar procedimentos de *bootstrap*. Uma vantagem muito importante da utilização do algoritmo EM com *bootstrapping* é que esta é uma boa abordagem quando os dados não são normalmente distribuídos (Graham, 2012).

4.4.4. Imputação Múltipla (MI – *Multiple Imputation*)

Apesar de predominar o recurso a FIML na análise SEM, a Imputação Múltipla é uma alternativa plausível, até porque, sob os mesmos pressupostos, produz resultados com as mesmas propriedades estatísticas para amostras grandes: consistentes, assintoticamente eficientes e assintoticamente normais.

Apesar de a Imputação Múltipla (IM) ser mais flexível do que a FIML, na primeira é introduzida variação aleatória no processo de imputação, o que leva a que não seja produzido o mesmo resultado cada vez que se usa a metodologia para o mesmo conjunto de dados. Além disso, há um grande número de maneiras de implementar a IM, sendo difícil escolher a melhor abordagem para cada aplicação.

A metodologia assenta basicamente em duas etapas: considerar apenas os dados completos e estimar com estes dados um modelo de regressão para a variável com dados omissos em função das restantes variáveis. De seguida, usar o valor predito pelo modelo

para substituir o dado em falta. A análise SEM será feita sobre os dados completos após a imputação de todos os dados omissos. No entanto, surgem alguns problemas: superestimar a correlação resultante da imputação determinista, mas que pode ser resolvido introduzindo uma correção na variância estimada; estimativas dos parâmetros não totalmente eficientes, resultado da variação aleatória adicionada e estimativas convencionais de erro padrão muito baixas por não ter em conta a variabilidade adicional (Allison, 2003). A solução para estes dois últimos problemas consiste em gerar M conjuntos de dados completos, cada um com dados imputados ligeiramente diferentes (5 conjuntos é suficiente segundo Allison (2003) mas Graham, Olchowski e Gilreath (2007) sugerem que são necessárias mais imputações – por exemplo, M = 20 a 40 ou mais, para obter o poder estatístico de procedimentos de ML equivalente). A análise SEM é implementada e os parâmetros e os erros padrão destes M conjuntos de dados completos são estimados e são depois combinados num único conjunto de estimativas de parâmetros e num erro padrão. Para os parâmetros considera-se a média dos seus valores nos M conjuntos de dados e o erro padrão pode resultar da média dos quadrados das estimativas dos erros padrão dos M conjuntos de dados completos, da variância das estimativas dos parâmetros nas M replicações, entre outros (Allison, 2003).

O método descrito é bastante bom, mas ainda produz intervalos de confiança que são um pouco grandes demais e testes de hipóteses cujas taxas de erros verdadeiros são um pouco maiores do que as taxas de erros nominais. O motivo é que, ao gerar os valores imputados, são utilizados os valores estimados dos coeficientes de regressão, em vez dos valores reais da população, mas os erros padrão estimados não refletem esse facto. Uma solução poderá estar em usar coeficientes de regressão simulados aleatoriamente, cada um com uma variável de distribuição normal, tomando a média igual à estimativa do parâmetro e a variância igual ao quadrado da estimativa do seu erro padrão (Allison, 2003).

Para padrões de dados omissos mais complexos, que tornam quase impraticável o método descrito, pode-se recorrer ao algoritmo de Monte Carlo via Cadeias de Markov (MCMC).

Para resolver outros potenciais problemas da Imputação Múltipla e aproximar os dados imputados dos dados reais, pode-se recorrer a transformação das variáveis, impor limites aos dados imputados, etc.

O método de imputação múltipla aqui descrito, baseado na regressão múltipla, é apenas um dos mais comumente usados. Mas há muitas outras alternativas que estão disponíveis em diversos pacotes do *software* R: `mice` (van Buuren *et al*., 2015), `mi` (Sue *et al*., 2011), `Amelia II` (Honaker *et al*., 2011), `mitools` (Lumley, Lumley & RODBC, 2015).

4.4.5. Variáveis auxiliares

Para incrementar as propriedades estatísticas das estimativas obtidas em SEM, quer pela Máxima Verosimilhança quer pela Imputação Múltipla, muitas vezes é desejável incluir no processo de imputação ou de modelação variáveis que não são significativas para o modelo, isto é, não seriam consideradas ou incluídas no modelo final se os dados estivessem completos, mas podem *(a)* ser uma causa potencial ou estar correlacionadas com a omissão de dados, ou *(b)* ser correlacionadas com a variável que contém dados omissos - são designadas por variáveis auxiliares.

As variáveis auxiliares a incorporar numa análise SEM devem ser *(i)* correlacionadas entre si; *(ii)* correlacionadas com todas as variáveis preditoras observadas; e (*iii*) correlacionadas com os resíduos de qualquer variável de resposta observada.

É importante notar que estas regras não se aplicam a variáveis latentes. Em nenhuma situação, uma variável auxiliar deve ser correlacionada com uma variável latente (Enders *et al.*, 2006).

No caso da Imputação Múltipla, a inclusão de variáveis auxiliares é mais fácil que na FIML, uma vez que são simplesmente adicionadas como variáveis preditoras durante a fase de imputação e podem ser ignoradas durante todas as análises subsequentes (os valores preenchidos já estão condicionados às variáveis extras). Neste caso, é importante que as variáveis auxiliares escolhidas sejam relacionadas com a variável que está a ser imputada e potencialmente relacionadas com omissão de dados nessa variável.

A inclusão de variáveis auxiliares na imputação ou na modelação pode tornar o pressuposto MAR mais plausível para além de poder melhorar a precisão dos resultados obtidos a partir de uma análise de dados omissos.

4.5. *Software* R – Ferramentas para lidar com os dados omissos na SEM

O *software* R, por defeito, na maioria dos modelos de regressão exclui os dados omissos automaticamente, atribuindo o valor `na.omit` à função `na.action`. Para que os dados omissos sejam considerados, o valor da função deve ser `na.action=na.pass`.

Entretanto, o *software* R tem um conjunto significativo de ferramentas para lidar com os dados omissos especificamente para a análise SEM. As estratégias principais para lidar com os dados omissos consistem na estimação por FIML (*Full Information Maximum Likelihood*) e, mais geralmente, na incorporação de métodos de Imputação Múltipla disponíveis num grande leque de pacotes e com recurso a diversos métodos.

Começando pelos pacotes de Imputação Múltipla refiram-se o `Amelia II` (Honaker, King & Blackwell, 2011), o `mice` (Buuren & Groothuis-Oudshoorn, 2011), o `mi` (Gelman *et al*., 2015), entre outros, em particular os pacotes `pan` (Zhao & Schafer, 2018) e o `jomo` (Quartagno & Carpenter, 2018) que permitem a Imputação Múltipla para a modelação conjunta multinível. O pacote `mitml` (Grund, Robitzsch & Luedtke, 2018) fornece um conjunto de ferramentas para imputação múltipla de dados omissos na modelação multinível e inclui uma *interface* amigável aos pacotes `pan` e `jomo`, além de diversas funções para visualização, gestão e análise de conjuntos de dados com imputação múltipla. Estes últimos pacotes serão particularmente úteis nas análises SEM em Educação. No que respeita a pacotes que lidam diretamente com a SEM com dados omissos, refira-se o `sem` (Fox *et al.*, 2017), o `lavaan` (Rosseel *et al*., 2018) e o `OpenMx` (Boker *et al*., 2011). O pacote `sem` permite recorrer a *listwise deletion* caso se defina o argumento `na.action=na.omit`, à estimação por FIML e à Imputação Múltipla, com recurso à função `miSem()` que usa a função `mi()` do pacote `mi` para gerar múltiplas imputações de dados omissos, ajustando o modelo especificado a cada conjunto completo de dados. O pacote `lavaan` permite definir o método para lidar com dados omissos através do argumento `missing="listwise"` (para dados MCAR) ou `missing="fiml"` (para dados pelo menos MAR), além de permitir a Imputação Múltipla com recurso à função `runMI()` do pacote `semTools` que implementa numa única etapa a imputação múltipla, através dos pacotes `Amélia` ou `mice`, e ajusta um modelo SEM. O pacote `OpenMx` usa a estimação por FIML sempre que são usados os dados brutos.

O R oferece funcionalidades adicionais, tais como testes de normalidade multivariada, testes à existência de *outliers* ou MCAR (pacotes `mvnormtest` (Jarek & Jarek, 2009), `MVN` (Korkmaz, Goksuluk & Zararsiz, 2014), `BaylorEdPsych` (Beaujean & Beaujean, 2018), `MissMech` (Jamshidian, Jalal & Jansen, 2014), imputações múltiplas (pacotes já referidos, `mvnmle` (Gross & Bates, 2018), `mitools` (Lumley, Lumley & RODBC, 2015), `Hmisc` (Harrell & Harrell, 2018), etc), *bootstrapping* (pacote `boot` (Canty & Ripley), 2017), também incorporado nos pacotes `sem` e `lavaan` e funções de SEM adicionais, como estatísticas Qui-quadrado corrigidas de Satorra-Bentler, aproximação da raiz quadrada do erro quadrático médio (*root mean squared error* - RMSEA) ou *bootstrap Bollen-Stine* (pacote `sem.additions`) que são úteis para os investigadores SEM. Estas opções são integradas num ambiente de trabalho unificado, evitando a importação e exportação de dados e permitindo que o investigador crie *scripts* completos de análise de dados usando um único idioma.

Ainda no que respeita à Imputação Múltipla, Buuren (2012) proporciona uma lista de ferramentas disponíveis no R, que atualiza em http://www.stefvanbuuren.nl/mi/Software.html, lista esta que ilustra a grande variedade de possibilidades nesta área.

De seguida faz-se referência explícita a alguns pacotes de análise SEM que dispõem de ferramentas para implementar a metodologia na presença de dados omisso e a pacotes de Imputação Múltipla e ilustram-se algumas ferramentas.

a) Pacotes `sem, lavaan, lavaan.survey e OpenMx`

Na função `sem()`, do pacote `sem`, a existência de dados omissos é processada da seguinte forma: o parâmetro `na.action` está predefinido como `na.omit` e nestas condições só serão processados os registos sem dados omissos, ou seja, a forma de lidar com os dados omissos é a *listwise diletion*. Definindo o parâmetro `na.action` como `na.pass`, fica ativo o estimador FIML (também possível de selecionar através do parâmetro `objective`). Este estimador, na ausência de dados omissos, produz os mesmos resultados que o estimador ML (predefinido quando `na.action= na.omit`). No entanto, é possível proceder à Imputação Múltipla com recurso à função `miSem()`,

do pacote `mi`, que estima o modelo SEM por Imputação Múltipla via Cadeias de Markov. Efetua iterações sobre os padrões específicos de cada variável quando em presença de dados omissos, tendo em vista a sua imputação.

O pacote `lavaan`, por defeito, usa a estimação por Máxima Verosimilhança, mas para dados omissos dispõe de duas opções: `"MLF"` para a estimação por Máxima Verosimilhança com erros padrão baseados nas derivadas de primeira ordem e testes estatísticos convencionais e `"MLR"` para estimação pela Máxima Verosimilhança com erros padrão robustos (Huber-White) (Freedman, 2006) e uma estatística de teste que é (assintoticamente) igual à estatística de teste Bentler Yuan (Bentler & Yuan, 1999).

Para as diversas formas de implementação das estimativas de Máxima Verosimilhança o `lavaan`, por defeito, baseia a análise na chamada matriz de covariância amostral enviesada, onde os elementos são divididos por N em vez de $N - 1$. Isso é feito internamente e não deve ser feito pelo utilizador. Além disso, a estatística do qui-quadrado é calculada multiplicando o valor mínimo da função pelo fator N (e não por $N - 1$). Este procedimento é semelhante ao usado no *software* Mplus. Se o investigador preferir usar uma matriz de covariâncias não enviesada terá que especificar essa opção em `likelihood="wishart"` e a estatística Qui-quadrado é calculada usando o fator $N - 1$. No caso da existência de dados omissos, o comportamento padrão é a aplicação da metodologia `listwise deletion` (que só pode ser usada no caso dos dados serem MCAR). Se o mecanismo gerador de dados omissos for, pelo menos, MAR, o pacote fornece a estimativa FIML que deve ser ativada com uma das opções `missing=c("ml", "direct","fiml")` quando é chamada a função de ajustamento (`sem()`, `cfa()` ou `growth()`ou `lavaan()`).

Os erros padrão são, por defeito, baseados na matriz de informação esperada, exceto quando há dados omissos e é usada a FIML, sendo, neste caso usada a matriz de informação observada que proporciona melhores resultados (Savalei, 2010).

O pacote `lavaan.survey` usa um modelo ajustado com recurso ao pacote `lavaan` e um *design* de pesquisa complexo, e devolve uma análise SEM em que o *design* complexo de amostragem é tido em conta. As estimativas dos parâmetros do modelo SEM e os erros padrão são baseados no *design* amostral. O pacote ajusta o modelo SEM, incluindo análise

fatorial, modelos de regressão multivariada com variáveis latentes e muitos outros modelos com variáveis latentes enquanto corrige estimativas, erros padrão e medidas de ajustamento, derivadas do qui-quadrado para um projeto de amostragem complexo. Incorpora *clusters*, estratificação, pesos amostrais e correções de população finita na análise SEM. Para lidar com dados omissos o pacote não incorpora nenhuma função específica mas o autor sugere a utilização da Imputação Múltipla, recorrendo aos pacotes `mice` e `mitools`, como ilustrado no artigo de introdução do pacote (Oberski, 2014), embora seja adequado qualquer pacote que implemente a Imputação Múltipla.

O pacote `OpenMx` usa o estimador FIML sempre que são usados os dados brutos e uma função de Máxima Verosimilhança para o ajustamento (função ML). Portanto, o utilizador não precisa fazer mais do que usar os dados brutos para ativar o estimador FIML, tendo assim a garantia que este será o estimador utilizado caso existam dados omissos. Quando não há dados omissos, são produzidos os mesmos resultados que com um estimador de Máxima Verosimilhança. Quando os padrões de omissão não são condicionados às variáveis resposta, o FIML considera a omissão e produz estimativas imparciais. A Imputação Múltipla não está incorporada no pacote.

b) Amelia II, mice e mi.

De acordo com Honaker, King e Blackwell (2011), `Amelia II` é um pacote de extensão ao *software* R que tem por objetivo auxiliar na imputação de dados, em conjuntos com dados omissos. Este pacote utiliza a técnica de *bootstrap* em conjunto com estimação por Máxima Verosimilhança para imputar valores em conjuntos de dados e assim produzir múltiplas saídas para a análise de dados. Baseia-se no algoritmo EM (*Expectiation Maximization*). Permite a imputação de dados longitudinais.

Este pacote possui um ambiente visual, que facilita o trabalho de imputação dos dados, bastando utilizar a função `AmeliaView()` para aceder ao ambiente visual. Pode simplesmente ser utilizada a função `amelia()` em ambiente de linha de comando.

No pacote `mice`, a imputação múltipla tem por base a Especificação Totalmente Condicional (FCS – *Fully Conditional Specification*), possuindo cada variável o seu próprio modelo de imputação. Os modelos de imputação incorporados são adequados para dados contínuos (*Predictive mean matching* (PMM) para dados não normais (Vink *et al.*,

2014) e modelo normal para dados normais), para dados binários (regressão logística), para dados categóricos não ordenados (regressão logística politómica) e para dados categóricos ordenais (*odds* proporcionais). O `mice` também pode imputar dados contínuos de dois níveis (modelo normal, `pan` (pacote), variáveis de segundo nível). Para manter a consistência entre as variáveis pode ser usada a imputação passiva (o utilizador pode especificar, em qualquer ponto no algoritmo de amostragem `mice Gibbs`, uma função para os dados imputados.). Estão disponíveis vários gráficos que permitem inspecionar a qualidade das imputações.

O pacote `mi` proporciona funções para manipulação de dados e para imputar dados omissos com recurso ao algoritmo EM, numa abordagem aproximadamente Bayesiana. O algoritmo gera valores imputados iterativamente da distribuição condicional para cada variável, dados os valores observados e imputados das outras variáveis nos dados. Dispõe de representações gráficas para visualizar padrões de dados omissos, para diagnosticar os modelos usados para gerar as imputações e para avaliar a convergência. Dispõe ainda de funções para analisar conjuntos de dados imputados com o grau apropriado de incerteza amostral. Os dados podem consistir em variáveis contínuas, semicontínuas, binárias, categóricas e/ou de contagem.

Exemplo de aplicação

Portugal participa regularmente no estudo internacional PISA onde se pretende avaliar literacia de jovens de 15 anos e a sua capacidade para enfrentar os desafios que a transição para a vida adulta lhes coloca.

O exemplo que se apresenta assenta em dados do PISA 2012 de Portugal (http://www.oecd.org/pisa/data/), tendo sido neste ciclo a literacia em Matemática o domínio principal de competência em avaliação.

Note-se que o objetivo deste exemplo é o de ilustrar algumas ferramentas e procedimentos disponíveis no R para tratar os dados omissos na SEM e não o de proceder à interpretação dos resultados da análise SEM do conjunto de dados.

Foram consideradas apenas as variáveis que dizem respeito à competência geral em Matemática e que são designadas por valores plausíveis (em vez de se estimar o *score* global de cada aluno com as respostas correspondentes às perguntas relativas às três

competências do PISA, é estimado um conjunto de valores possíveis (*plausible values*) e a respetiva probabilidade e são imputados valores para cada aluno em função desses valores plausíveis) e as variáveis que dizem respeito à eficácia na Matemática (questão QT37, que está subdividida em oito questões, cada com uma escala de quatro pontos: *muito confiante, confiante, não muito confiante, nada confiante*) e ao nível do autoconceito a Matemática (questão Q42, com cinco subquestões, com uma escala de quatro pontos: *discordo totalmente, discordo, concordo, concordo totalmente*). Foram ainda consideradas as variáveis: índice socioeconómico e cultural do aluno, tipo de escola (geral, pré-vocacional (CEF), profissional) e sexo. A base de dados tem uma quantidade significativa de observações com dados omissos que foram codificados com NA (68,47%).

As variáveis utilizadas foram codificadas da seguinte forma:

PV1MATH: "Valores plausíveis" para a capacidade matemática geral da criança (imputação 1).
PV2MATH: "Valores plausíveis" para a capacidade matemática geral da criança (imputação 2).
PV3MATH: "Valores plausíveis" para a capacidade matemática geral da criança (imputação 3).
PV4MATH: "Valores plausíveis" para a capacidade matemática geral da criança (imputação 4).
PV5MATH: "Valores plausíveis" para a capacidade matemática geral da criança (imputação 5).
ST37Q01: Sentiu-se confiante a realizar a tarefa: "horário" 1 (muito) - 4 (não em todos).
ST37Q02: Sentiu-se confiante a realizar a tarefa: "desconto" (1-4).
ST37Q03: Sentiu-se confiante a realizar a tarefa: "área" (1-4).
ST37Q04: Sentiu-se confiante a realizar a tarefa: "gráficos" (1-4).
ST37Q05 Sentiu-se confiante a realizar a tarefa: "linear" (1-4).
ST37Q06: Sentiu-se confiante a realizar a tarefa: "distância" (1-4).
ST37Q07: Sentiu-se confiante a realizar a tarefa: "quadrática" (1-4).
ST37Q08: Sentiu-se confiante a realizar a tarefa: "taxa" (1-4).
ST42Q02: "Eu não sou bom em Matemática" 1 (concordo totalmente) - 4 (discordo totalmente).
ST42Q04: "Recebo boas notas em Matemática" (1-4).
ST42Q06: "Eu aprendo Matemática rapidamente" (1-4).
ST42Q07: "Sempre acreditei que a Matemática é uma das minhas melhores matérias" (1-4).
ST42Q09: "Na minha aula de Matemática, entendo até o trabalho mais difícil" (1-4).
ESCS: Índice de Status Socioeconômico e Cultural.
sexo: (1 = feminino, 2 = masculino).
School.type: Tipo de escola - nível de dificuldade de estudos secundários (1-3).

```
> head(prt.pisa.12.2)

sexo ST37Q01 ST37Q02 ST37Q03 ST37Q04 ST37Q05 ST37Q06 ST37Q07 ST37Q08 ST42Q02
1    1       1       1       3       1       1       4       2       2       3
2    2       1       1       2       1       1       2       1       2       3
3    1       2       2       3       2       1       3       2       3      NA
4    2       1       2       3       2       1       2       2       1       1
5    2      NA      NA      NA      NA      NA      NA      NA      NA       3
6    1       2       2       3       2       1       3      NA       3       2
  ST42Q04 ST42Q06 ST42Q07 ST42Q09  ESCS  PV1MATH  PV2MATH  PV3MATH  PV4MATH
1       2       2      NA       3 -1.26 416.1941 432.5518 391.2681 427.0992
2       2       2       1       2  0.56 381.8430 363.9274 434.8107 404.4322
3      NA      NA      NA      NA -0.01 379.5062 316.4122 352.2433 335.8857
4       4       4       4       1 -2.13 338.3004 335.1846 337.5214 337.5214
5       2       2       2       2 -0.59 586.7814 593.7919 614.0442 586.0025
6       3       3       3       3  0.67 417.2067 376.7020 401.6280 375.1441
   PV5MATH school.type
1 403.7311           1
2 387.2955           1
3 387.2955           1
4 365.5632           1
5 594.5708           1
6 380.5967           1
```

O R dispõe de diversas ferramentas para análise dos padrões de dados omissos.

Com recurso aos pacotes `dplyr` (Wickham *et al.*, 2018) e `ggploty2` (Wickham, 2018) obtém-se uma representação gráfica da percentagem de omissão de cada variável (Figura 4.1).

```
> library("dplyr")

> propmiss <- function(dataframe) {
+   m <- sapply(dataframe, function(x) {
+     data.frame(
+       nmiss=sum(is.na(x)),
+       n=length(x),
+       propmiss=sum(is.na(x))/length(x)
+     )
+   })
+   d <- data.frame(t(m))
+   d <- sapply(d, unlist)
+   d <- as.data.frame(d)
+   d$variable <- row.names(d)
+   row.names(d) <- NULL
+   d <- cbind(d[ncol(d)],d[-ncol(d)])
+   return(d[order(d$propmiss), ]) }
>
> miss_vars<-propmiss(prt.pisa.12.2)
> miss_vars_mean<-mean(miss_vars$propmiss)
> miss_vars_ges<- miss_vars  %>% arrange(desc(propmiss))
> library(ggplot2)

> plot1<-ggplot(miss_vars_ges,aes(x=reorder(variable,propmiss),y=propmiss*100)) +
+   geom_point(size=3) +
+   coord_flip() +
+   theme_bw() + xlab("") +ylab("Missingness por variável") +
+   theme(panel.grid.major.x=element_blank(),
+         panel.grid.minor.x=element_blank(),
+         panel.grid.major.y=element_line(colour="grey60",linetype="dashed")) +
+   ggtitle("Percentagem
+ de missingness")
> plot1
```

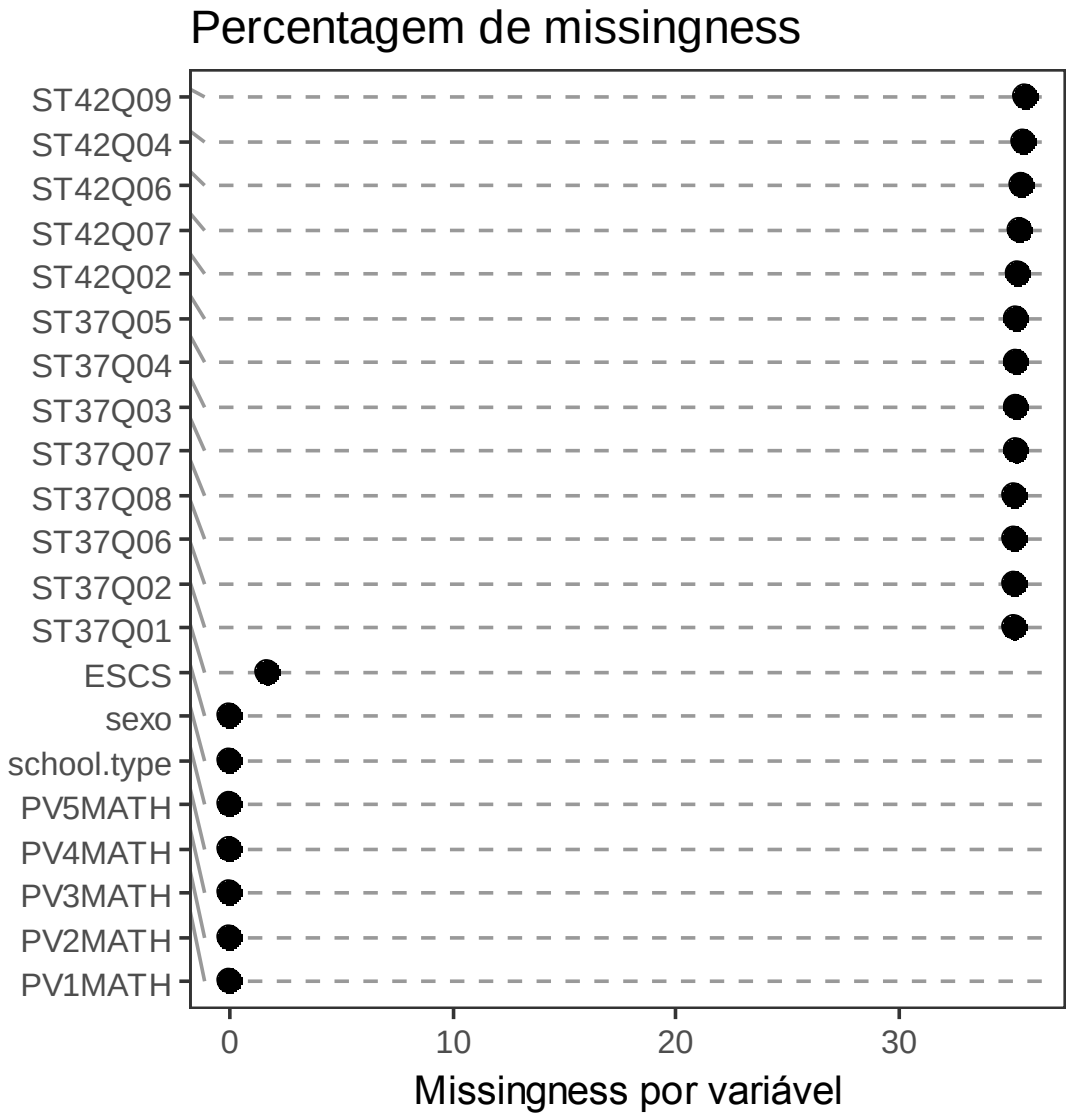


Figura 4.1: Representação gráfica da percentagem de omissão por variável.

O pacote `lavaan`, através das funções `lavTech()` (TRUE – dado não omisso; FALSE – dado omisso) ou `lavInspect() ()` (1 – dado não omisso; 0 – dado omisso) identifica os padrões de dados omissos numa matriz cujas linhas representam os padrões identificados.

```
> lavTech(model1,what="patterns")
     PV1MAT PV2MAT PV3MAT PV4MAT PV5MAT ST37Q01 ST37Q02 ST37Q03 ST37Q04 ST37Q05 ST37Q06 ST37Q07 ST37Q08 ST42Q02 ST42Q04 ST42Q06
ST42Q07 ST42Q09   ESCS   sexo   schl.t
 [1,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
 [2,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [3,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
 [4,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
 [5,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
 [6,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
 [7,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
 [8,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [9,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[10,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[11,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[12,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
[13,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[14,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[15,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[16,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[17,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[18,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[19,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
[20,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[21,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[22,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[23,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
[24,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
[25,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
[26,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
[27,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
[28,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
[29,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[30,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
[31,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[32,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[33,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[34,] TRUE TRUE TRUE TRUE TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[35,] TRUE TRUE TRUE TRUE TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[36,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[37,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[38,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[39,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[40,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[41,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
[42,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[43,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE
[44,] TRUE TRUE TRUE TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
```

O pacote `mice` proporciona uma matriz semelhante mas os resultados são algo diferentes do pacote `lavaan`: identifica 56 padrões e não 44 como acontece com o pacote `lavaan`. Além de uma matriz de 0's e 1's, o pacote proporciona a representação gráfica da matriz (Figura 4.2). Na margem esquerda está contabilizado o número de observações correspondente a cada padrão e na margem direita está contabilizado o número de dados omissos por padrão. Na margem inferior está contabilizado o número de dados omissos por variável.

```
> md.pattern(prt.pisa.12.2)
```

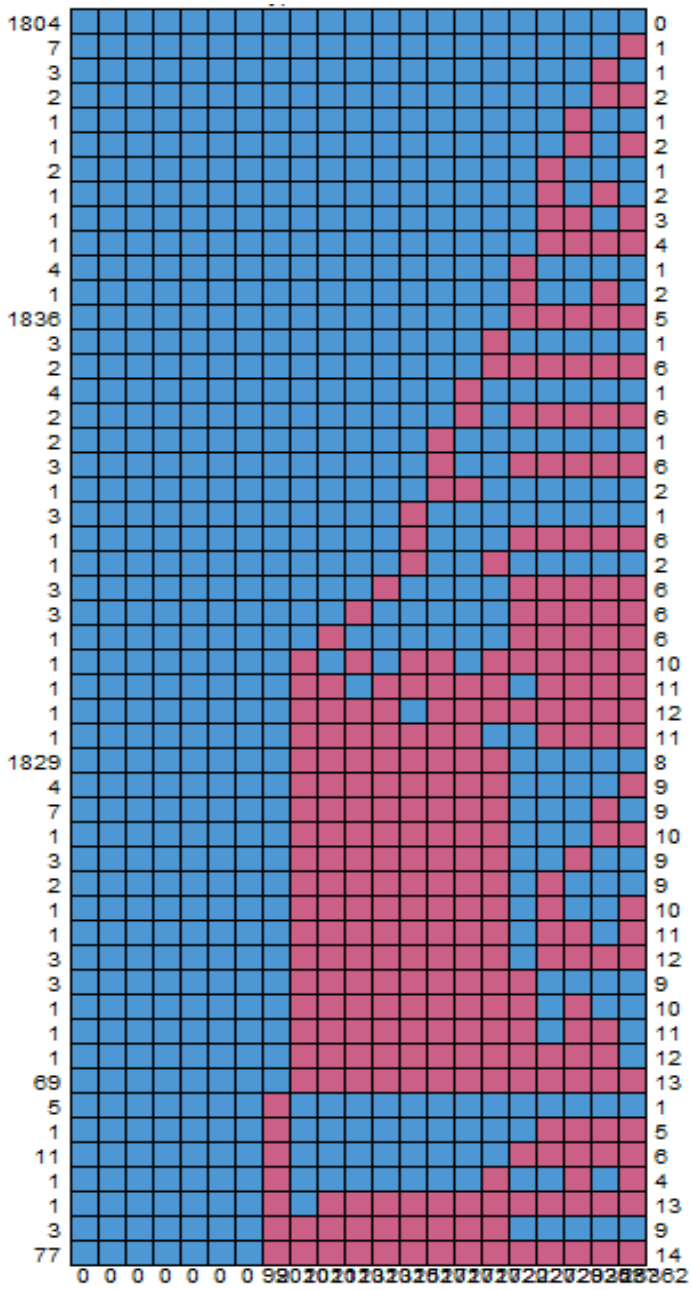


Figura 4.2: Representação gráfica da matriz de padrões obtida com recurso à função **`md.pattern()`**.do pacote **`mice`**.

O pacote `Amelia II` proporciona um mapa representativo dos dados omissos por observação (linha) permitindo perceber o padrão de omissão de dados por simples observação. Note-se, no entanto, que as observações não são agrupadas por padrões idênticos.

```
library(Amelia II)

missmap1(prt.pisa.12.2)
```

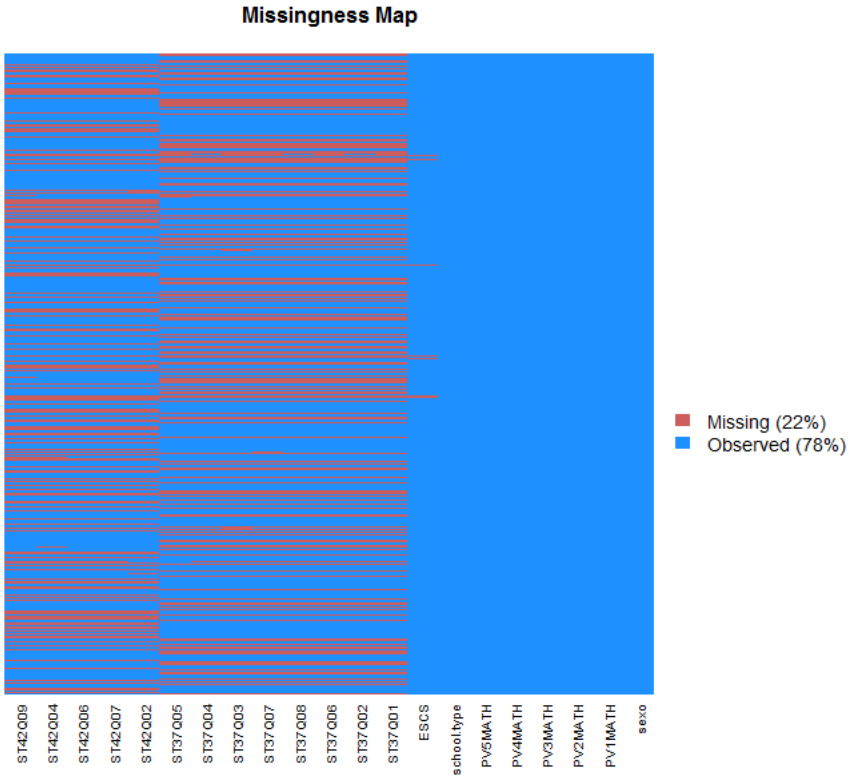


Figura 4.3: Mapa de omissão de dados, por observação obtido com recurso à função `missmap1()` do pacote `Amelia II`.

Com este mapa é fácil perceber que se se considerar a *listwise diletion* o número de dados completos será numa percentagem muito reduzida relativamente ao número total de dados. Parte significativa da omissão de dados ocorreu pelo planeamento experimental usado. Parece plausível que os dados sejam assumidos como MAR, uma vez que parece haver um padrão determinista de omissão: as observações com dados omissos nas subquestões da questão QT37 não têm dados omissos nas subquestões da questão QT42. O padrão da matriz sugere que a presença de dados omissos não depende dos valores omissos, nem das variáveis com dados omissos, mas depende de valores observados sobre outras variáveis da matriz de dados.

Os pacotes `mvnormtest` (Jarek & Jarek, 2009), `MVN` (Korkmaz, Goksuluk & Zararsiz,2018), `psych` (Revelle, 2018), `BaylorEdPsych` (Beaujean & Beaujean, 2018), `MissMech` (Jamshidian, Jalal & Jansen, 2014) proporcionam testes à normalidade

multivariada, à homogeneidade de variâncias, ao pressuposto MCAR. No entanto, tendo em conta que a base de dados com os dados omissos tem 5772 observações e as funções têm como limite 5000 observações, alguns dos testes não puderam ser aplicados. A função `mshapiro.test()` do pacote `mvnormtest` foi alterada para permitir a aplicação do teste (foi aumentado o limite máximo de observações admitido). No entanto, tal não foi possível no pacote `MVN`.

Assim, os resultados da aplicação de alguns testes são os seguintes:

```
Library(mvnormtest)
> mshapiro.test <-
+  function (x) {
+    if (!is.matrix(x)) {x <- as.matrix(x)}
+    x <- x[complete.cases(x),]
+    x <- t(x)
+    n <- ncol(x)
+       if (n<3 || n>500000) {stop("sample size must be between 3 and 5000")}
+    rng <- range(x)
+    rng <- rng[2]-rng[1]
+    if (rng==0) {stop("all `x[]' are identical")}
+    Us <- apply(x,1,mean)
+    R <- x-Us
+    M.1 <- solve(R%*%t(R),tol=1e-50)
+    Rmax <- diag(t(R)%*%M.1%*%R)
+    C <- M.1%*%R[,which.max(Rmax)]
+    Z <- t(C)%*%x
+    result <- shapiro.test(Z)
+    result$method <- "Multivariate Shapiro-Wilk normality test"
+    result$data.name <- paste("(",paste(rownames(x),collapse=","),")",sep="")
+    return(result)
+  }
> mshapiro.test(prt.pisa.12.2)

Multivariate Shapiro-Wilk normality test

data:
(sexo,ST37Q01,ST37Q02,ST37Q03,ST37Q04,ST37Q05,ST37Q06,ST37Q07,ST37Q08,ST4
2Q02,ST42Q04,ST42Q06,ST42Q07,ST42Q09,ESCS,PV1MATH,PV2MATH,PV3MATH,PV4MATH
,PV5MATH,school.type)

W = 0.97895, p-value = 1.236e-15
```

```
library(BaylorEdPsych)
> library(BaylorEdPsych)
> test_mcar<-LittleMCAR(prt.pisa.12.2)
Loading required package: mvnmle
this could take a while>
> # print p-value of mcar-test
> print(test_mcar$p.value)
[1] 8.881784e-16
```

```
> library("MissMech")
> out<-TestMCARNormality(prt.pisa.12.2)
```

```
> print(out)
Call:
TestMCARNormality(data = prt.pisa.12.2)

Number of Patterns:  8

Total number of cases used in the analysis:  5640

 Pattern(s) used:
         sexo   ST37Q01   ST37Q02   ST37Q03   ST37Q04   ST37Q05   ST37Q06
group.1     1         1         1         1         1         1         1
group.2     1         1         1         1         1         1         1
group.3     1        NA        NA        NA        NA        NA        NA
group.4     1        NA        NA        NA        NA        NA        NA
group.5     1        NA        NA        NA        NA        NA        NA
group.6     1         1         1         1         1         1         1
group.7     1         1         1         1         1         1         1
group.8     1        NA        NA        NA        NA        NA        NA
         ST37Q07   ST37Q08   ST42Q02   ST42Q04   ST42Q06   ST42Q07   ST42Q09
group.1        1         1         1         1         1         1         1
group.2        1         1        NA        NA        NA        NA        NA
group.3       NA        NA         1         1         1         1         1
group.4       NA        NA        NA        NA        NA        NA        NA
group.5       NA        NA        NA        NA        NA        NA        NA
group.6        1         1         1         1         1         1        NA
group.7        1         1        NA        NA        NA        NA        NA
group.8       NA        NA         1        NA         1         1         1
         ESCS   PV1MATH   PV2MATH   PV3MATH   PV4MATH   PV5MATH   school.type
group.1     1         1         1         1         1         1             1
group.2     1         1         1         1         1         1             1
group.3     1         1         1         1         1         1             1
group.4    NA         1         1         1         1         1             1
group.5     1         1         1         1         1         1             1
group.6     1         1         1         1         1         1             1
group.7    NA         1         1         1         1         1             1
group.8     1         1         1         1         1         1             1
         Number of cases
group.1             1804
group.2             1836
group.3             1829
group.4               77
group.5               69
group.6                7
group.7               11
group.8                7

    Test of normality and Homoscedasticity:
  ------------------------------------------

Hawkins Test:

    P-value for the Hawkins test of normality and homoscedasticity:  3.130527e-
263

    Either the test of multivariate normality or homoscedasticity (or
both) is rejected.
    Provided that normality can be assumed, the hypothesis of MCAR is
    rejected at 0.05 significance level.

Non-Parametric Test:

    P-value for the non-parametric test of homoscedasticity:
0.0001926078

    Hypothesis of MCAR is rejected at 0.05 significance level.
```

```
    The multivariate normality test is inconclusive.
```

Como se pode observar e como era de esperar a hipótese dos dados serem MCAR é rejeitada, bem como a normalidade multivariada e a homocedasticidade. No entanto, os testes são sensíveis a amostras grandes e portanto vai-se assumir a normalidade multivariada e o pressuposto MAR.

Estudou-se um modelo simplificado apresentado por Oberski (2014).

```
> library(semPlot)
>    semPaths(model1, intercept = FALSE)
```

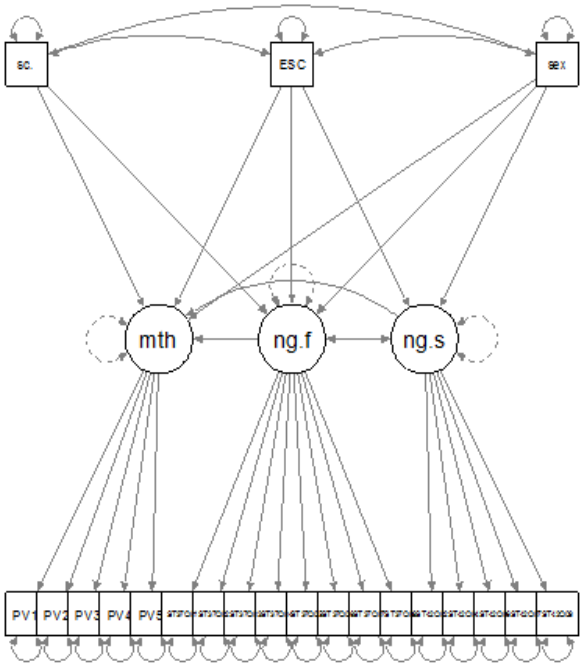


Figura 4.4: Diagrama de caminhos do modelo.

As variáveis `PV1MATH,PV2MATH,PV3MATH,PV4MATH,PV5MATH` foram estandardizadas, uma vez que tomam valores muito diferentes dos valores das restantes variáveis.

Ajustando o modelo com o pacote `lavaan` e na presença de dados omissos, obtém-se o resultado seguinte:

```
>library(lavaan)
> pisa.prt.12 <- "
+ math = ~ PV1MATH + PV2MATH + PV3MATH + PV4MATH+PV5MATH
+ neg.efficacy = ~ ST37Q01 + ST37Q02 + ST37Q03 + ST37Q04 +
+   ST37Q05 + ST37Q06 + ST37Q07 + ST37Q08
+ neg.selfconcept = ~ ST42Q02 + ST42Q04 + ST42Q06 + ST42Q07 + ST42Q09
```

```
+
+    neg.selfconcept ~ neg.efficacy + ESCS + sexo
+   neg.efficacy ~ neg.selfconcept + school.type + ESCS + sexo
+   math ~ neg.selfconcept + neg.efficacy + school.type + ESCS + sexo
+   sexo~~ESCS
+   ESCS~~school.type
+   sexo~~school.type
+   "
>xx=rapply(prt.pisa.12.2[,16:20],scale,c("numeric","integer"),how="replace")
prt.pisa.12.2[,16:20]<-xx
> model1<- sem(pisa.prt.12, data = prt.pisa.12.2, auto.var = TRUE, std.lv = TRUE,
+   int.ov.free = TRUE, missing="fiml" )
> summary(model1, estimates = TRUE, fit.measures = TRUE)
lavaan 0.6-2 ended normally after 128 iterations

  Optimization method                           NLMINB
  Number of free parameters                         75

  Number of observations                          5722
  Number of missing patterns                        51

  Estimator                                         ML
  Model Fit Test Statistic                    3003.209
  Degrees of freedom                               177
  P-value (Chi-square)                           0.000

Model test baseline model:

  Minimum Function Test Statistic           82764.527
  Degrees of freedom                              210
  P-value                                       0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.966
  Tucker-Lewis Index (TLI)                       0.959

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -75829.964
  Loglikelihood unrestricted model (H1)     -74328.360

  Number of free parameters                         75
  Akaike (AIC)                               151809.929
  Bayesian (BIC)                             152308.834
  Sample-size adjusted Bayesian (BIC)        152070.506

Root Mean Square Error of Approximation:

  RMSEA                                          0.053
  90 Percent Confidence Interval          0.051  0.054
  P-value RMSEA <= 0.05                          0.002

Standardized Root Mean Square Residual:

  SRMR                                           0.042

Parameter Estimates:

  Information                                 Observed
  Observed information based on                Hessian
  Standard Errors                             Standard

Latent Variables:
```

```
                   Estimate  Std.Err  z-value  P(>|z|)
  math =~
    PV1MATH           0.642    0.008   79.339    0.000
    PV2MATH           0.641    0.008   79.150    0.000
    PV3MATH           0.642    0.008   79.320    0.000
    PV4MATH           0.642    0.008   79.366    0.000
    PV5MATH           0.641    0.008   79.170    0.000
  neg.efficacy =~
    ST37Q01           0.365    0.009   39.764    0.000
    ST37Q02           0.398    0.010   41.597    0.000
    ST37Q03           0.443    0.011   41.992    0.000
    ST37Q04           0.311    0.009   36.104    0.000
    ST37Q05           0.390    0.010   38.914    0.000
    ST37Q06           0.386    0.010   36.991    0.000
    ST37Q07           0.416    0.011   37.188    0.000
    ST37Q08           0.342    0.010   35.104    0.000
  neg.selfconcept =~
    ST42Q02           0.687    0.068   10.146    0.000
    ST42Q04          -0.646    0.064  -10.136    0.000
    ST42Q06          -0.679    0.067  -10.202    0.000
    ST42Q07          -0.774    0.076  -10.191    0.000
    ST42Q09          -0.596    0.059  -10.134    0.000

Regressions:
                    Estimate  Std.Err  z-value  P(>|z|)
  neg.selfconcept ~
    neg.efficacy       0.039    0.115    0.334    0.739
    ESCS               0.228    0.038    5.951    0.000
    sexo               0.292    0.033    8.789    0.000
  neg.efficacy ~
    neg.selfconcpt    -0.839    0.200   -4.203    0.000
    school.type        0.262    0.038    6.945    0.000
    ESCS              -0.284    0.033   -8.587    0.000
    sexo              -0.001    0.053   -0.010    0.992
  math ~
    neg.selfconcpt     0.113    0.032    3.550    0.000
    neg.efficacy      -0.627    0.025  -24.598    0.000
    school.type       -0.171    0.029   -5.883    0.000
    ESCS               0.250    0.015   16.775    0.000
    sexo               0.031    0.031    0.981    0.327

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  ESCS ~~
    sexo              0.007    0.008    0.846    0.398
    school.type      -0.123    0.009  -14.043    0.000
  sexo ~~
    school.type       0.024    0.004    6.489    0.000

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)
   .PV1MATH           0.232    0.042    5.471    0.000
   .PV2MATH           0.231    0.042    5.470    0.000
   .PV3MATH           0.232    0.042    5.471    0.000
   .PV4MATH           0.232    0.042    5.471    0.000
   .PV5MATH           0.231    0.042    5.470    0.000
   .ST37Q01           1.642    0.029   56.629    0.000
   .ST37Q02           1.645    0.031   52.404    0.000
   .ST37Q03           1.891    0.035   54.096    0.000
   .ST37Q04           1.628    0.025   64.552    0.000
   .ST37Q05           1.534    0.031   49.135    0.000
   .ST37Q06           1.966    0.031   63.060    0.000
   .ST37Q07           1.814    0.034   53.944    0.000
   .ST37Q08           1.962    0.028   70.454    0.000
   .ST42Q02           2.239    0.042   53.072    0.000
```

```
   .ST42Q04           2.724    0.039   69.035    0.000
   .ST42Q06           2.753    0.041   66.919    0.000
   .ST42Q07           3.092    0.047   65.699    0.000
   .ST42Q09           2.921    0.037   79.582    0.000
    ESCS             -0.494    0.016  -31.703    0.000
    sexo              1.501    0.007  227.144    0.000
    school.type       1.221    0.007  167.629    0.000
   .math              0.000
   .neg.efficacy      0.000
   .neg.selfconcpt    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .PV1MATH           0.066    0.002   42.414    0.000
   .PV2MATH           0.070    0.002   43.007    0.000
   .PV3MATH           0.067    0.002   42.541    0.000
   .PV4MATH           0.066    0.002   42.316    0.000
   .PV5MATH           0.069    0.002   42.986    0.000
   .ST37Q01           0.231    0.006   37.364    0.000
   .ST37Q02           0.216    0.006   36.011    0.000
   .ST37Q03           0.273    0.008   36.298    0.000
   .ST37Q04           0.251    0.006   39.418    0.000
   .ST37Q05           0.306    0.008   37.493    0.000
   .ST37Q06           0.367    0.009   39.243    0.000
   .ST37Q07           0.428    0.011   38.388    0.000
   .ST37Q08           0.342    0.009   39.816    0.000
   .ST42Q02           0.362    0.010   36.180    0.000
   .ST42Q04           0.262    0.008   34.776    0.000
   .ST42Q06           0.197    0.006   31.182    0.000
   .ST42Q07           0.310    0.009   33.330    0.000
   .ST42Q09           0.287    0.008   36.728    0.000
    ESCS              1.369    0.026   52.868    0.000
    sexo              0.250    0.005   53.488    0.000
    school.type       0.304    0.006   53.488    0.000
   .math              1.000
   .neg.efficacy      1.000
   .neg.selfconcpt    1.000
```

Na Figura 4.5 encontra-se representado o diagrama de caminhos obtido.

Considerando o ajustamento com a *listwise deletion*, o *output* é o seguinte:

```
> model2<- sem(pisa.prt.12, data = prt.pisa.12.2, missing="listwise",estimator = "MLM" )
```

```
> summary(model2, estimates = FALSE, fit.measures = TRUE)
lavaan 0.6-2 ended normally after 78 iterations

  Optimization method                           NLMINB
  Number of free parameters                         54

                                                  Used       Total
  Number of observations                          1804        5722

  Estimator                                         ML      Robust
  Model Fit Test Statistic                    1611.046    1445.191
  Degrees of freedom                           177         177
  P-value (Chi-square)                          0.000       0.000
  Scaling correction factor                                 1.115
    for the Satorra-Bentler correction

Model test baseline model:
```

```
  Minimum Function Test Statistic             31332.464   29110.958
  Degrees of freedom                                210         210
  P-value                                         0.000       0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                     0.954       0.956
  Tucker-Lewis Index (TLI)                        0.945       0.948

  Robust Comparative Fit Index (CFI)                          0.955
  Robust Tucker-Lewis Index (TLI)                             0.946

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -31040.544  -31040.544
  Loglikelihood unrestricted model (H1)      -30235.021  -30235.021

  Number of free parameters                          54          54
  Akaike (AIC)                                62189.088   62189.088
  Bayesian (BIC)                              62485.967   62485.967
  Sample-size adjusted Bayesian (BIC)         62314.412   62314.412

Root Mean Square Error of Approximation:

  RMSEA                                           0.067       0.063
  90 Percent Confidence Interval          0.064   0.070       0.060   0.066
  P-value RMSEA <= 0.05                           0.000       0.000

  Robust RMSEA                                                0.067
  90 Percent Confidence Interval                              0.063   0.070

Standardized Root Mean Square Residual:
  SRMR                                            0.045       0.045
```

Como se pode ver, o número de observações usadas reduz de 5772 para 1804.

As estimativas dos parâmetros do modelo são as seguintes:

```
Parameter Estimates:

  Information                                 Expected
  Information saturated (h1) model          Structured
  Standard Errors                           Robust.sem

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  math =~
    PV1MATH           1.000
    PV2MATH           0.998    0.010  102.507    0.000
    PV3MATH           0.996    0.009  106.011    0.000
    PV4MATH           1.003    0.010  105.177    0.000
    PV5MATH           0.993    0.009  107.309    0.000
  neg.efficacy =~
    ST37Q01           1.000
    ST37Q02           1.104    0.032   34.231    0.000
    ST37Q03           1.249    0.039   31.836    0.000
    ST37Q04           0.837    0.031   26.670    0.000
    ST37Q05           1.105    0.043   25.547    0.000
    ST37Q06           1.094    0.043   25.692    0.000
    ST37Q07           1.204    0.048   25.086    0.000
    ST37Q08           1.001    0.037   26.953    0.000
  neg.selfconcept =~
```

```
    ST42Q02             1.000
    ST42Q04            -0.973    0.032  -30.558    0.000
    ST42Q06            -1.059    0.034  -31.543    0.000
    ST42Q07            -1.158    0.038  -30.648    0.000
    ST42Q09            -0.899    0.035  -25.925    0.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  neg.selfconcept ~
    neg.efficacy       0.137    0.358    0.383    0.702
    ESCS               0.145    0.062    2.358    0.018
    sexo               0.258    0.050    5.196    0.000
  neg.efficacy ~
    neg.selfconcpt    -0.458    0.106   -4.334    0.000
    school.type        0.090    0.018    4.922    0.000
    ESCS              -0.101    0.016   -6.385    0.000
    sexo              -0.003    0.032   -0.101    0.920
  math ~
    neg.selfconcpt     0.110    0.038    2.893    0.004
    neg.efficacy      -1.086    0.071  -15.211    0.000
    school.type       -0.069    0.031   -2.201    0.028
    ESCS               0.172    0.016   10.703    0.000
    sexo               0.051    0.033    1.565    0.118

Covariances:
                  Estimate  Std.Err  z-value  P(>|z|)
  ESCS ~~
    sexo             0.003    0.014    0.217    0.828
    school.type     -0.140    0.014  -10.260    0.000
  sexo ~~
    school.type       0.021    0.007    3.277    0.001

Variances:
                  Estimate  Std.Err  z-value  P(>|z|)
   .PV1MATH           0.062    0.003   22.597    0.000
   .PV2MATH           0.068    0.003   24.714    0.000
   .PV3MATH           0.066    0.003   23.561    0.000
   .PV4MATH           0.065    0.003   23.590    0.000
   .PV5MATH           0.066    0.003   23.200    0.000
   .ST37Q01           0.227    0.011   20.674    0.000
   .ST37Q02           0.219    0.011   19.094    0.000
   .ST37Q03           0.259    0.012   21.980    0.000
   .ST37Q04           0.243    0.010   23.475    0.000
   .ST37Q05           0.278    0.013   20.866    0.000
   .ST37Q06           0.363    0.014   25.810    0.000
   .ST37Q07           0.392    0.016   24.098    0.000
   .ST37Q08           0.330    0.013   25.295    0.000
   .ST42Q02           0.370    0.022   16.496    0.000
   .ST42Q04           0.246    0.012   20.177    0.000
   .ST42Q06           0.167    0.009   17.890    0.000
   .ST42Q07           0.305    0.016   18.633    0.000
   .ST42Q09           0.290    0.016   17.789    0.000
    ESCS              1.347    0.035   38.296    0.000
    sexo              0.250    0.000 9577.754    0.000
    school.type       0.309    0.019   16.668    0.000
   .math              0.426    0.019   22.398    0.000
   .neg.efficacy      0.127    0.010   12.605    0.000
   .neg.selfconcpt    0.450    0.140    3.203    0.001
```

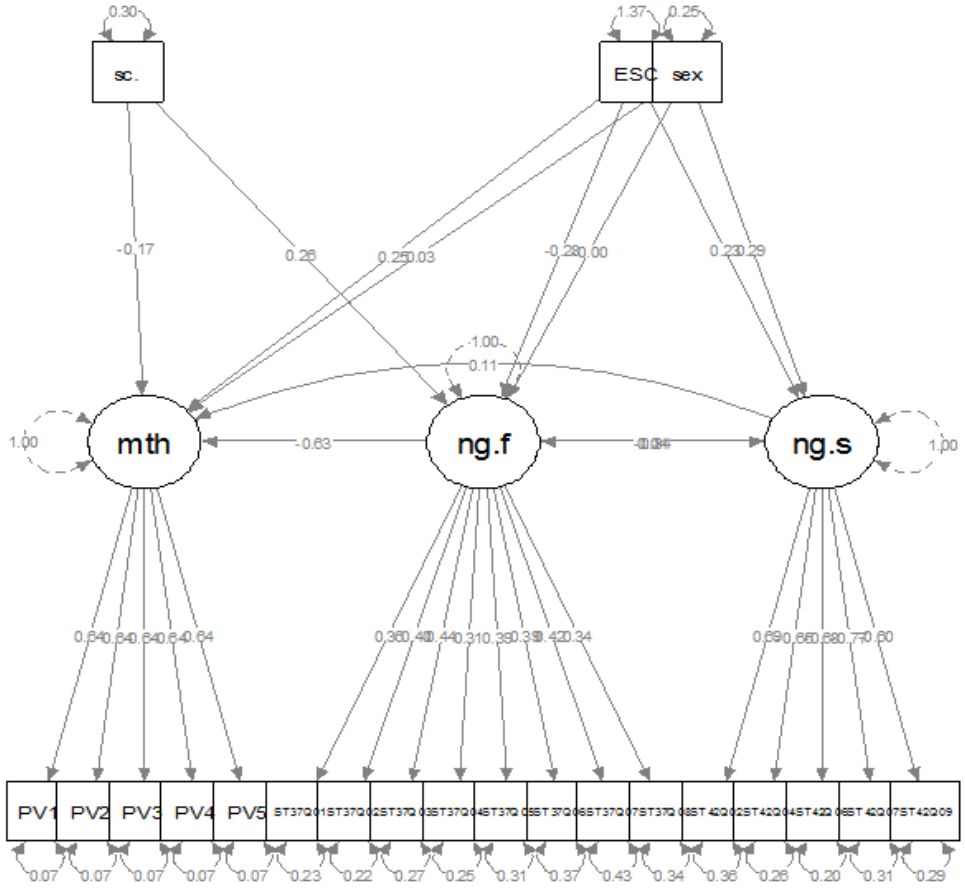


Figura 4.5: Diagrama de caminhos do modelo ajustado

De seguida será considerada a imputação múltipla. Vai-se assumir que os dados são MAR. O tamanho da amostra, sendo elevado, ajuda a legitimar a assunção deste pressuposto.

Também neste tópico o R dispõe de uma grande quantidade de ferramentas, seja para implementar a imputação e o ajustamento do modelo em simultâneo, seja para realizar a tarefa em duas etapas: primeiro os dados são imputados e de seguida é ajustado o modelo com os dados completos resultantes da imputação.

O pacote `sem` dispõe da função `simSem()` que permite realizar a tarefa numa única etapa. Os pacotes `lavaan`, `lavaan.survey` e `OpenMx` não têm qualquer função que permita a imputação múltipla e o ajustamento do modelo numa única etapa. Lidam com os dados omissos usando o estimador adequado (FIML). O `lavaan` dispõe ainda de uma opção designada por *"two.stage*" que consiste numa primeira etapa em que as estimativas do vetor de médias, $\bar{X}_n$, e da matriz de covariâncias, S_n, são obtidas através do algoritmo

EM baseado em uma suposição de normalidade multivariada e numa segunda etapa, em que se procede à análise como no caso dos dados completos, usando $\bar{X}_n$ e S_n estimadas como o vetor de médias amostral e a matriz de covariâncias amostral (Rosseel, 2018). Para usar a imputação múltipla será necessário proceder primeiro à imputação de dados e de seguida proceder ao ajustamento do modelo com os dados entretanto completados.

No entanto, o pacote `semTools` permite realizar o ajustamento do modelo sobre um conjunto de bases de dados completos por imputação múltipla, sem que seja efetuado esse conjunto de imputações separadamente.

Os *outputs* seguintes exemplificam a imputação de dados nos pacotes `mi`, `mice` e `Amelia II`.

A imputação com o pacote `mi` é iniciada com a criação de uma *data.frame* usando a função `missing_data.frame()` que devolve um objeto semelhante a uma *data.frame*, mas que é personalizada para que as variáveis com dados omissos sejam modeladas para imputação múltipla. Esta *data.frame* contém a lista de variáveis, a respetiva classificação, o número de dados omissos por variável, o método que será usado para a imputação por variável, bem como a função que relaciona cada variável com dados omissos com as restantes variáveis. De seguida, a função `mi()` é usada para executar um algoritmo iterativo onde cada variável com dados omissos é modelada (usando o modelo estabelecido) como função de todos as outras variáveis. Caso não seja definido o número de Cadeias de Markov a usar e o número de iterações para cada cadeia, serão executadas 4 cadeias e 30 iterações.

```
> miss.mi<-missing_data.frame(prt.pisa.12.2)
NOTE: In the following pairs of variables, the missingness pattern of the first
is a subset of the second.
 Please verify whether they are in fact logically distinct variables.
     [,1]      [,2]
[1,] "ST37Q01" "ST37Q03"
> show(miss.mi)
Object of class missing_data.frame with 5722 observations on 21 variables

There are 51 missing data patterns

Append '@patterns' to this missing_data.frame to access the corresponding pattern
for every observation or perhaps use table()

                            type missing method  model
```

```
sexo                      binary       0   <NA>   <NA>
ST37Q01     ordered-categorical    2010    ppd ologit
ST37Q02     ordered-categorical    2011    ppd ologit
ST37Q03     ordered-categorical    2017    ppd ologit
ST37Q04     ordered-categorical    2017    ppd ologit
ST37Q05     ordered-categorical    2017    ppd ologit
ST37Q06     ordered-categorical    2013    ppd ologit
ST37Q07     ordered-categorical    2015    ppd ologit
ST37Q08     ordered-categorical    2013    ppd ologit
ST42Q02     ordered-categorical    2022    ppd ologit
ST42Q04     ordered-categorical    2035    ppd ologit
ST42Q06     ordered-categorical    2029    ppd ologit
ST42Q07     ordered-categorical    2027    ppd ologit
ST42Q09     ordered-categorical    2037    ppd ologit
ESCS                 continuous      99    ppd linear
PV1MATH              continuous       0   <NA>   <NA>
PV2MATH              continuous       0   <NA>   <NA>
PV3MATH              continuous       0   <NA>   <NA>
PV4MATH              continuous       0   <NA>   <NA>
PV5MATH              continuous       0   <NA>   <NA>
school.type ordered-categorical       0   <NA>   <NA>

                 family     link transformation
sexo               <NA>     <NA>           <NA>
ST37Q01     multinomial    logit           <NA>
ST37Q02     multinomial    logit           <NA>
ST37Q03     multinomial    logit           <NA>
ST37Q04     multinomial    logit           <NA>
ST37Q05     multinomial    logit           <NA>
ST37Q06     multinomial    logit           <NA>
ST37Q07     multinomial    logit           <NA>
ST37Q08     multinomial    logit           <NA>
ST42Q02     multinomial    logit           <NA>
ST42Q04     multinomial    logit           <NA>
ST42Q06     multinomial    logit           <NA>
ST42Q07     multinomial    logit           <NA>
ST42Q09     multinomial    logit           <NA>
ESCS           gaussian identity    standardize
PV1MATH            <NA>     <NA>    standardize
PV2MATH            <NA>     <NA>    standardize
PV3MATH            <NA>     <NA>    standardize
PV4MATH            <NA>     <NA>    standardize
PV5MATH            <NA>     <NA>    standardize
school.type        <NA>     <NA>           <NA>
> imputations<-mi(miss.mi)
> show(imputations)
Object of class mi with 4 chains, each with 30 iterations.
Each chain is the evolution of an object of missing_data.frame class with 5722
observations on 21 variables.
> summary(imputations)
$sexo
$sexo$is_missing
[1] "all values observed"

$sexo$observed
   1    2
2853 2869
```

```
$ST37Q01
$ST37Q01$crosstab
    observed imputed
  1     6328    2196
  2     6880    3883
  3     1428    1558
  4      212     403

• • • •

$ESCS
$ESCS$is_missing
missing
FALSE  TRUE
 5623    99

$ESCS$imputed
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.43771 -0.57769 -0.26149 -0.25429  0.06409  1.06469

$ESCS$observed
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.45048 -0.38436 -0.08465  0.00000  0.33923  1.36253

$PV1MATH
$PV1MATH$is_missing
[1] "all values observed"

$PV1MATH$observed
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.7899392 -0.3555382  0.0007553  0.0000000  0.3609884  1.6139769


……
> if(!exists("imputations", env = .GlobalEnv)) {
+   imputations <- mi:::imputations # cached from example("mi-package")
+ }
> data.frames <- complete(imputations, 3)
> lapply(data.frames, summary)
$`chain:1`
 sexo     ST37Q01  ST37Q02  ST37Q03  ST37Q04  ST37Q05  ST37Q06  ST37Q07
 1:2853   1:2096   1:2169   1:1610   1:2315   1:2935   1:1440   1:1935
 2:2869   2:2646   2:2430   2:2247   2:2787   2:1821   2:2362   2:2062
          3: 825   3: 973   3:1583   3: 558   3: 712   3:1627   3:1189
          4: 155   4: 150   4: 282   4:  62   4: 254   4: 293   4: 536

 ST37Q08  ST42Q02  ST42Q04  ST42Q06  ST42Q07  ST42Q09       ESCS
 1:1358   1:1271   1: 726   1: 584   1: 707   1: 413   Min.   :-3.8700
 2:2743   2:1933   2:2504   2:2484   2:1329   2:1998   1st Qu.:-1.3900
 3:1410   3:1945   3:1947   3:2081   3:2195   3:2518   Median :-0.6800
 4: 211   4: 573   4: 545   4: 573   4:1491   4: 793   Mean   :-0.4927
                                                      3rd Qu.: 0.3000
                                                      Max.   : 2.7000
    PV1MATH             PV2MATH             PV3MATH             PV4MATH
 Min.   :-3.579878   Min.   :-3.614958   Min.   :-3.47223   Min.   :-3.16974
 1st Qu.:-0.711076   1st Qu.:-0.706824   1st Qu.:-0.71065   1st Qu.:-0.71289
 Median : 0.001511   Median : 0.005445   Median : 0.01429   Median : 0.00886
 Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.00000   Mean   : 0.00000
 3rd Qu.: 0.721977   3rd Qu.: 0.710459   3rd Qu.: 0.71727   3rd Qu.: 0.70672
 Max.   : 3.227954   Max.   : 3.156665   Max.   : 3.19667   Max.   : 3.44085
    PV5MATH          school.type missing_ST37Q01 missing_ST37Q02
 Min.   :-3.680236   1:4833      Mode :logical   Mode :logical
 1st Qu.:-0.709414   2: 513      FALSE:3712      FALSE:3711
 Median : 0.005409   3: 376      TRUE :2010      TRUE :2011
 Mean   : 0.000000
 3rd Qu.: 0.712519
 Max.   : 3.129869
```

```
 missing_ST37Q03 missing_ST37Q04 missing_ST37Q05 missing_ST37Q06
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:3705      FALSE:3705      FALSE:3705      FALSE:3709
 TRUE :2017      TRUE :2017      TRUE :2017      TRUE :2013

missing_ST37Q07 missing_ST37Q08 missing_ST42Q02 missing_ST42Q04
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:3707      FALSE:3709      FALSE:3700      FALSE:3687
 TRUE :2015      TRUE :2013      TRUE :2022      TRUE :2035

missing_ST42Q06 missing_ST42Q07 missing_ST42Q09 missing_ESCS
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:3693      FALSE:3695      FALSE:3685      FALSE:5623
 TRUE :2029      TRUE :2027      TRUE :2037      TRUE :99
```

```
> plot(imputations)
```

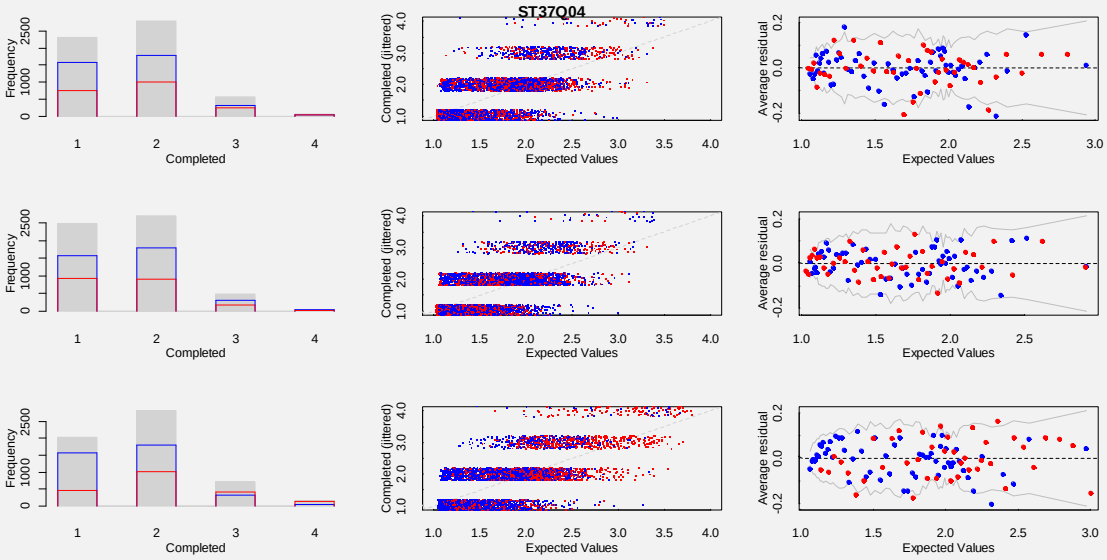


Figura 4.7. Representação dos resultadis da imputação múltipla no pacote `mi`– variável ST37Q04

Usando o pacote `Amelia II`, com a função `amelia()` obtém-se o seguinte *output*,

```
> miss.Amelia<-amelia(prt.pisa.12.2)
> miss.Amelia

Amelia output with 5 imputed datasets.
Return code:  1
Message:  Normal EM convergence.

Chain Lengths:
--------------
Imputation 1:  12
Imputation 2:  12
Imputation 3:  13
Imputation 4:  12
Imputation 5:  12
```

Para efetuar o diagnóstico dos dados imputados e analisar a plausibilidade dos dados imputados em relação aos dados observados, o pacote dispõe de diversas ferramentas, nomeadamente gráficos que permitem comparar as funções de densidade dos dados observados e dos dados imputados. As ferramentas gráficas de diagnóstico auxiliam, por exemplo, na verificação da plausibilidade do pressuposto MAR, uma vez que este não pode ser testado a partir dos dados (Buuren & Groothuis-Oudshoorn, 2011).

```
> plot(miss.Amelia, which.vars = 13:16)
```

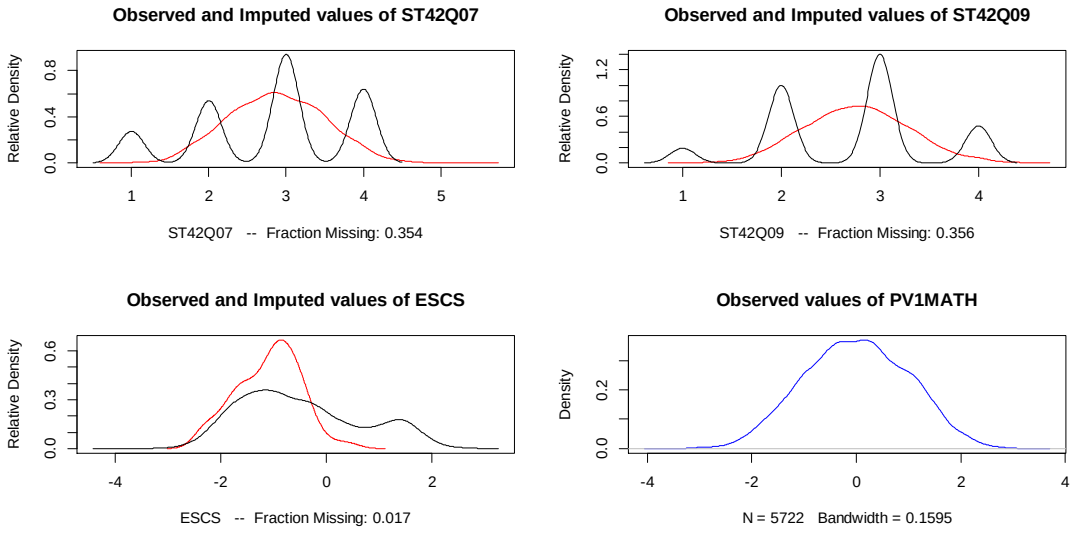


Figura 4.6: Função de densidade dos dados observados e dos dados imputados para quatro das variáveis.

O pacote `mice`, usando a função `mice()`, devolve o seguinte *output*

```
miss.mice <- mice(prt.pisa.12.2, m=5,method = "pmm", seed = 500)
```

```
> miss.mice
Class: mids
Number of multiple imputations:  5
Imputation methods:
       sexo     ST37Q01     ST37Q02     ST37Q03     ST37Q04     ST37Q05
         ""       "pmm"       "pmm"       "pmm"       "pmm"       "pmm"
    ST37Q06     ST37Q07     ST37Q08     ST42Q02     ST42Q04     ST42Q06
      "pmm"       "pmm"       "pmm"       "pmm"       "pmm"       "pmm"
    ST42Q07     ST42Q09       ESCS     PV1MATH     PV2MATH     PV3MATH
      "pmm"       "pmm"       "pmm"          ""          ""          ""
    PV4MATH     PV5MATH school.type
          ""          ""          ""
PredictorMatrix:
$chainMean
, , Chain 1

                   1          2         3         4         5
sexo             NaN        NaN       NaN       NaN       NaN
ST37Q01     1.710448  1.7004975  1.691045  1.704478  1.676119
ST37Q02     1.676280  1.6911984  1.691198  1.694679  1.674789
ST37Q03     2.002479  2.0099157  1.967774  1.990084  1.984135
```

```
ST37Q04      1.662866  1.6440258  1.656916  1.694100  1.683193
ST37Q05      1.528508  1.5404065  1.546852  1.575607  1.597918
ST37Q06      2.073522  2.0452062  2.042722  2.045206  2.046200
ST37Q07      1.852605  1.8565757  1.849628  1.883871  1.933499
ST37Q08      2.045206  1.9955291  2.040238  2.030303  1.993045
ST42Q02      2.478239  2.4668645  2.473294  2.439664  2.445104
ST42Q04      2.487961  2.4864865  2.520885  2.549877  2.548894
ST42Q06      2.497782  2.5194677  2.523411  2.561853  2.553475
ST42Q07      2.842625  2.8307844  2.834238  2.888505  2.882585
ST42Q09      2.706431  2.7349043  2.757977  2.751105  2.748159
ESCS        -0.990000 -0.9966667 -1.215354 -1.082121 -1.171313
PV1MATH           NaN        NaN       NaN       NaN       NaN
PV2MATH           NaN        NaN       NaN       NaN       NaN
PV3MATH           NaN        NaN       NaN       NaN       NaN
PV4MATH           NaN        NaN       NaN       NaN       NaN
PV5MATH           NaN        NaN       NaN       NaN       NaN
school.type       NaN        NaN       NaN       NaN       NaN

, , Chain 2
```

Para efetuar o diagnóstico dos dados imputados, o `mice` dispõe de diversas ferramentas gráficas. A título de exemplo, comparem-se as funções de densidade dos dados observados e dos dados imputados, representadas nos gráficos da figura 4.8, obtidos com o *script* seguinte:

```
> densityplot(miss.mice,scales = list(x = list(relation = "free"), cex=0.6),par.strip.text = list(cex = 0.6), par.settings = simpleTheme(col.line = rep(mdc(1:2))) )
```

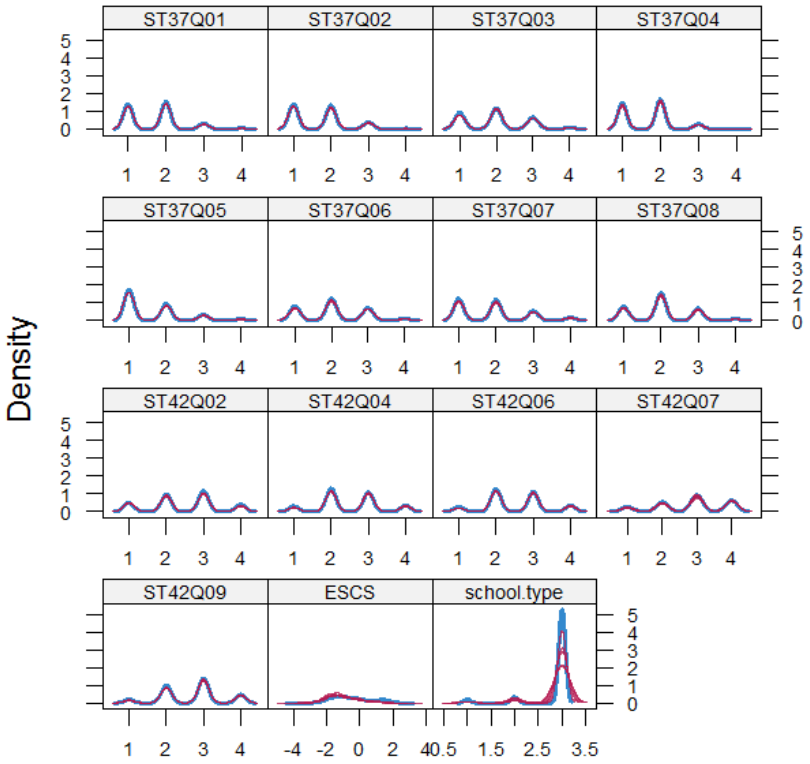


Figura 4.8: Funções de densidade dos dados observados e dos dados imputados, por variável.

Após efetuar um determinado número de imputações é possível ajustar um modelo a cada base de dados completos, e obter por fim estimativas dos parâmetros do modelo, combinando as estimativas dos diferentes modelos ajustados às diferentes bases de dados imputados.

A função `runMi()`, em geral, e as funções `sem.mi()`, `cfa.mi()` e `growth.mi()`, em particular, do pacote `semTools` permitem realizar as duas etapas (imputação e ajustamento do modelo) numa única etapa. Dispõem de duas opções para a imputação múltipla: o pacote `mice` e o pacote `Amelia II`.

Assim, considerando o modelo ajustado com a função `lavaan()` do pacote `semTools`, com cinco bases de dados completos por imputação múltipla, recorrendo ao pacote `Amelia II`, obtém-se o seguinte *output*,

```
# usando a função sem.mi
model4.1<-sem.mi(pisa.prt.12,data=pisa.12.s.2,m=5,miPackage="Amelia",seed=5000)
# usando a função runMI com a função sem()
>  model4.2<-runMI(pisa.prt.12,data=pisa.12.s.2,  m=5,fun  =  "sem",  miPackage  =
"Amelia", seed = 5000)
# Criando as bases de dados por imputação múltipla e de seguida ajustando o
modelo com recurso à função sem.mi
> set.seed(5000)
> pisa.amelia <- amelia(pisa.12.s.2, m = 5)
-- Imputation 1 --
  1  2  3  4  5  6  7  8  9 10 11 12

-- Imputation 2 --
  1  2  3  4  5  6  7  8  9 10 11 12

-- Imputation 3 --
  1  2  3  4  5  6  7  8  9 10 11 12 13

-- Imputation 4 --
  1  2  3  4  5  6  7  8  9 10 11 12

-- Imputation 5 --
  1  2  3  4  5  6  7  8  9 10 11 12 13 14

> imps <- pisa.amelia$imputations
> model4.3 <- sem.mi(pisa.prt.12, data = imps)
> summary(model4.1)
lavaan.mi object based on 5 imputed data sets.
See class?lavaan.mi help page for available methods.

Convergence information:
The model converged on 5 imputed data sets

Rubin's (1987) rules were used to pool point and SE estimates across 5 imputed
data sets, and to calculate degrees of freedom for each parameter's t test and
CI.

Parameter Estimates:

  Information                                 Expected
  Information saturated (h1) model          Structured
```

```
  Standard Errors                           Standard

Latent Variables:
                   Estimate  Std.Err  t-value       df  P(>|t|)
  math =~
    PV1MATH           1.000
    PV2MATH           0.998    0.006  163.410      Inf    0.000
    PV3MATH           1.000    0.006  165.249      Inf    0.000
    PV4MATH           1.000    0.006  166.012      Inf    0.000
    PV5MATH           0.998    0.006  163.472      Inf    0.000
  neg.efficacy =~
    ST37Q01           1.000
    ST37Q02           1.091    0.024   46.202   57.227    0.000
    ST37Q03           1.219    0.026   46.193   12.970    0.000
    ST37Q04           0.849    0.022   39.271   21.959    0.000
    ST37Q05           1.089    0.025   42.890   16.737    0.000
    ST37Q06           1.062    0.027   40.069   25.803    0.000
    ST37Q07           1.175    0.029   40.887   13.738    0.000
    ST37Q08           0.933    0.025   37.914   15.957    0.000
  neg.selfconcept =~
    ST42Q02           1.000
    ST42Q04          -0.929    0.019  -50.013  115.252    0.000
    ST42Q06          -0.985    0.018  -54.075   98.635    0.000
    ST42Q07          -1.118    0.022  -51.957   74.768    0.000
    ST42Q09          -0.861    0.018  -47.571  395.116    0.000

Regressions:
                   Estimate  Std.Err  t-value       df  P(>|t|)
  neg.selfconcept ~
    neg.efficacy      0.092    0.239    0.385  137.807    0.700
    ESCS              0.159    0.044    3.653  214.090    0.000
    sexo              0.207    0.030    6.910   17.838    0.000
  neg.efficacy ~
    neg.selfconcpt   -0.435    0.068   -6.412  132.033    0.000
    school.type       0.094    0.013    7.532 6871.826    0.000
    ESCS             -0.106    0.011   -9.265  569.435    0.000
    sexo              0.000    0.019    0.019   16.650    0.985
  math ~
    neg.selfconcpt    0.109    0.023    4.745   50.348    0.000
    neg.efficacy     -1.116    0.038  -29.182  617.384    0.000
    school.type      -0.112    0.020   -5.506  807.414    0.000
    ESCS              0.157    0.010   15.176   71.331    0.000
    sexo              0.018    0.022    0.843 1790.160    0.400

Covariances:
                   Estimate  Std.Err  t-value       df  P(>|t|)
  ESCS ~~
    sexo              0.006    0.009    0.694      Inf    0.488
    school.type      -0.122    0.010  -11.731 7979.653    0.000
  sexo ~~
    school.type       0.024    0.004    5.415      Inf    0.000

Variances:
                   Estimate  Std.Err  t-value       df  P(>|t|)
   .PV1MATH           0.066    0.002   35.433      Inf    0.000
   .PV2MATH           0.070    0.002   35.927      Inf    0.000
   .PV3MATH           0.067    0.002   35.522      Inf    0.000
   .PV4MATH           0.066    0.002   35.347      Inf    0.000
   .PV5MATH           0.070    0.002   35.914      Inf    0.000
   .ST37Q01           0.234    0.006   39.423   18.609    0.000
   .ST37Q02           0.219    0.006   37.992    9.488    0.000
   .ST37Q03           0.273    0.007   37.998  687.885    0.000
   .ST37Q04           0.253    0.006   41.175   90.777    0.000
   .ST37Q05           0.300    0.008   39.836  231.656    0.000
   .ST37Q06           0.369    0.009   40.919   23.024    0.000
```

```
   .ST37Q07           0.419    0.010   40.639   73.731    0.000
   .ST37Q08           0.343    0.008   41.556   44.373    0.000
   .ST42Q02           0.361    0.010   37.982    9.400    0.000
   .ST42Q04           0.263    0.007   36.723    7.096    0.000
   .ST42Q06           0.193    0.006   32.459   29.250    0.000
   .ST42Q07           0.315    0.009   35.030   13.702    0.000
   .ST42Q09           0.282    0.007   38.318   43.571    0.000
    ESCS              1.364    0.031   44.636      Inf    0.000
    sexo              0.250    0.006   44.636      Inf    0.000
    school.type       0.304    0.007   44.636      Inf    0.000
   .math              0.409    0.011   37.702   42.575    0.000
   .neg.efficacy      0.134    0.006   22.462   17.102    0.000
   .neg.selfconcpt    0.483    0.100    4.853  105.889    0.000
```

O pacote `lavaan` disponibiliza a função `fitMeasures()` que permite obter uma grande bateria de testes de ajustamento do modelo.

```
> fitMeasures(model4.1, "all")
anova() provides more control over options for pooling chi-squared before
calculating fit indices from multiple imputations. See the class?lavaan.mi help
page for details.

            chisq                df            pvalue              npar
         2264.212           177.000             0.000            54.000
           ntotal              logl unrestricted.logl               aic
         5722.000       -100099.571        -97560.172        200307.143
             bic              bic2    baseline.chisq       baseline.df
       200666.355        200494.759         46506.787           210.000
  baseline.pvalue               cfi               rni              nnfi
           0.000             0.955             0.955             0.947
             tli               rfi               nfi              pnfi
           0.947             0.942             0.951             0.802
             ifi               mfi              rmsea    rmsea.ci.lower
           0.955             0.833             0.045             0.044
   rmsea.ci.upper      rmsea.pvalue          gammaHat       adjGammaHat
           0.047             1.000             0.966             0.956
             rmr       srmr_bollen      srmr_bentler
           0.031             0.044             0.044
Warning messages:
1: In lavaan(model = PT, data = d, slotOptions = lavoptions, group = group) :
  lavaan WARNING: the optimizer warns that a solution has NOT been found!
2: In lavaan(model = PT, data = d, slotOptions = lavoptions, group = group) :
  lavaan WARNING: the optimizer warns that a solution has NOT been found!
```

É possível ajustar o modelo em duas etapas, recorrendo, por exemplo ao pacote `mice` e ao `lavaan.survey`, usando o seguinte *script*:

```
> mice.imp <- NULL
> for(i in 1:5) mice.imp[[i]] <- complete(pisa_imp, action=i, inc=FALSE)
> mice.imp2<-lapply(seq(pisa_imp$m),function(im) complete(pisa_imp,im))
> mice.imp2<-mitools::imputationList(mice.imp2)
#survey-criar objeto
> svy.df_imp<-survey::svydesign(id=~1,weights=~1,data=mice.imp2)
> lavaan_fit_pisa<-sem(pisa.prt.12, meanstructure = FALSE)
> model5<-lavaan.survey(lavaan_fit_pisa, svy.df_imp)
> summary(model5)
```

```
lavaan 0.6-2 ended normally after 112 iterations

  Optimization method                           NLMINB
  Number of free parameters                         75

  Number of observations                          5722

  Estimator                                         ML      Robust
  Model Fit Test Statistic                    4882.096    4547.036
  Degrees of freedom                           177         177
  P-value (Chi-square)                          0.000       0.000
  Scaling correction factor                                 1.074
    for the Satorra-Bentler correction

Parameter Estimates:

  Information                                 Expected
  Information saturated (h1) model          Structured
  Standard Errors                           Robust.sem

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  math =~
    PV1MATH           1.000
    PV2MATH           0.998    0.006  180.611    0.000
.....
  neg.efficacy =~
    ST37Q01           1.000
    ST37Q02           1.093    0.018   60.652    0.000
.....
  neg.selfconcept =~
    ST42Q02           1.000
    ST42Q04          -0.938    0.016  -57.510    0.000
.....

Regressions:
                 Estimate  Std.Err  z-value  P(>|z|)
  neg.selfconcept ~
    neg.efficacy    0.082    0.211    0.389    0.697
    ESCS            0.150    0.036    4.211    0.000
    sexo            0.200    0.025    8.170    0.000
  neg.efficacy ~
    neg.selfconcpt -0.390    0.071   -5.479    0.000
    school.type     0.085    0.011    7.619    0.000
    ESCS           -0.104    0.011   -9.219    0.000
    sexo           -0.008    0.018   -0.453    0.651
  math ~
    neg.selfconcpt  0.193    0.018   10.440    0.000
    neg.efficacy   -1.041    0.032  -32.295    0.000
    school.type    -0.128    0.018   -7.258    0.000
    ESCS            0.170    0.009   19.307    0.000
    sexo            0.013    0.018    0.685    0.493

Covariances:
                 Estimate  Std.Err  z-value  P(>|z|)
  ESCS ~~
```

```
     sexo              0.006    0.008     0.790    0.429
     school.type      -0.122    0.008   -15.857    0.000
  sexo ~~
     school.type       0.024    0.004     6.538    0.000

Intercepts:
                   Estimate  Std.Err  z-value   P(>|z|)
   .PV1MATH           0.231    0.043     5.376    0.000
……
Variances:
                   Estimate  Std.Err  z-value   P(>|z|)
   .PV1MATH           0.066    0.002    40.105    0.000
……
```

4.6. Conclusão

Os métodos convencionais para lidar com dados omissos deixam muito a desejar. Podem produzir estimativas enviesadas de parâmetros, dos seus erros padrão ou ambos, pois geralmente fazem um uso ineficiente dos dados. Apenas para os dados MCAR produzem boas estimativas, mecanismo de omissão de dados que em dados reais é praticamente inexistente. Em contrapartida, os métodos Máxima Verosimilhança para lidar com dados omissos, FIML, têm propriedades estatísticas quase ótimas sob a suposição MAR, o que permite que a omissão dependa apenas dos dados observados. Como o pressuposto da normalidade multivariada é comum na análise SEM, esta é a configuração mais fácil para implementar métodos Máxima Verosimilhança para dados omissos.

A Imputação Múltipla possui propriedades estatísticas quase tão boas quanto as da FIML, sendo mais atrativa, na medida em pode ser facilmente implementada com qualquer tipo de modelo ou método de estimação. A principal desvantagem é que, ao contrário de Máxima Verosimilhança, não produz um resultado determinado, pela incerteza que introduz na imputação. A grande diversidade de abordagens e algoritmos também pode ser um obstáculo pela confusão e incerteza que pode induzir no investigador sobre a melhor maneira de a implementar.

Vários programas de SEM já implementam a estimação FIML para dados omissos e a Imputação Múltipla e como se viu, o *software* R não é exceção, com a grande vantagem de ser livre e permitir a investigadores, alunos e professores disporem de uma gama de ferramentas suficientemente abrangente e completa para lidar com o problema dos dados omissos no contexto da SEM.

CAPÍTULO 5

PACOTES DO R PARA ANÁLISE DE EQUAÇÕES ESTRUTURAIS

PACOTES DO R PARA ANÁLISE DE EQUAÇÕES ESTRUTURAIS

5.1. Introdução

A popularidade crescente da SEM foi acompanhada e impulsionada de/por mudanças na análise estatística de dados de pesquisa. A conceção de métodos estatísticos, resultamte do desenvolvimento computracional impulsionado parcialmente pela Lei de Moore[3] permitiu a melhoraria da qualidade de sua produção científica, resultante de um aumento dramático da complexidade dos modelos e dos métodos. Também pela via do desenvolvimento computacional, os métodos de recolha de dados tornaram-se mais automatizados e o armazenamento de dados tornou-se acessível, levando a que o tamanho dos conjuntos de dados aumentasse drasticamente. Em consequência destes desenvolvimentos. Os projetos de pesquisa tornaram-se mais ambiciosos, na medida em que se tornou possível recolher um grande número de medidas de grandes amostras.

O desenvolvimento computacional levou ao surgimento de uma grande variedade de *softwares* comerciais que permitem implementar a análise SEM. Opções comerciais, como o AMOS - SPSS (Byrne, 2012), Calis - SAS (PROC PROC CALIS, 2010), EQS (Bentler e Wu, 2005), LISREL (Jöreskog e Sörbom, 2009), Mplus (Muthén e Muthén, 2009), SEPath (SEPath, 2013), tendo associado um custo, também limitam a possibilidade de explorar novas ideias metodológicas, uma vez que os detalhes de muitos recursos, normalmente, permanecem ocultos ao utilizador. O R (R Development Core Team, 2017), sendo umsoftware em ambiente *open-source*, possibilita a implementação de ferramentas de análise SEM que resolvem estes dois problemas, uma vez que são de utilização totalmente gratuita e permitem aos utilizadores explorarem soluções que respondem aos seus problemas particulares. Atualmente, estão disponíveis diversas soluções, destacando-se os pacotes `sem` (Fox, 2017), `OpenMx` (Boker *et al*, 2011), `lava` (Klaus, 2013, 2018), `lavaan` (Rosseel, 2018). O R proporciona muitos outros recursos, quer para lidar com análises em contextos específicos e análises com dados especiais, quer para suporte gráfico ou ferramentas para tratar etapas específicas da análise.

[3] A complexidade dos circuitos de computador, ou seja, o poder de computação, duplica aproximadamente a cada 18-24 meses (Moore, 1965)).

No presente capítulo é vertido o resultado da pesquisa de pacotes do R para Análise de Equações Estruturais, bem como uma breve exposição das potencialidades de cada um. São ainda incluídos alguns exemplos que ilustram estas análises ou as potencialidades dos recursos referidos.

5.2. Pacotes do R para Equações Estruturais

Feita uma pesquisa aturada das ferramentas disponíveis no R para implementar a metodologia ou para implementar etapas ou procedimentos associados, apresenta-se uma breve resenha dos pacotes que têm como foco a implementação da SEM, ou outros que, sendo mais abrangentes, têm como componente a implementação da SEM, ou ainda, pacotes que dispõem de ferramentas que suportam modelos criados por pacotes para a SEM e que permitem incrementar potencialidades desses pacotes ou colmatar lacunas que estes têm.

5.2.1. PACOTE `sem` (Fox *et al.*, 2017)

Este pacote proporciona ferramentas para a ajustar equações estruturais em modelos de variáveis observadas, pelo método dos Mínimos Quadrados em duas etapas (2LS) através da função `tsls()`, e para ajustar modelos com equações estruturais lineares gerais (com variáveis observadas e latentes) usando a abordagem RAM. Usa estimadores de Máxima Verosimilhança (ML – *Maximum Likelihood*) e de Máxima Verosimilhança de Informação Completa (FIML- *Full Information Maximum Likelihood*), assumindo a distribuição normal multivariada. Embora este último estimador forneça as mesmas estimativas que o estimador ML quando não há dados omissos, é substancialmente mais rápido, assim como o estimador GLS (*Generalized Least Square*), também disponível, pois usam códigos compilados. Usando o pacote `polycor` (Fox, 2016) é possível estimar a matriz de correlações policóricas para variáveis endógenas ordinais e usá-la para ajustar modelos SEM com este tipo de variáveis. A função `path.diagram()` cria a descrição do Diagrama de Caminhos de um modelo SEM ou objeto `SEMspecification`, a ser processado pelo programa de desenho de gráficos *dot*, que pode ser chamado automaticamente se estiver instalado ou, com recurso automático ao pacote `DiagrammeR` proporciona uma representação do Diagrama de Caminhos em HTML.

5.2.2. PACOTE `lavaan` (Rosseel *et al.*, 2018)

Este pacote fornece uma coleção de ferramentas que permite explorar, estimar e compreender uma grande variedade de modelos estatísticos multivariados com variáveis latentes, a saber, Análise Fatorial Confirmatória, Análise de Caminhos, Equações Estruturais, modelos longitudinais (curvas de crescimento), modelos multinível, Itens de Resposta e modelos com *missing data*, usando as funções de ajustamento `sem()`, `cfa()`, `growth(`. A função `lavaan()` permite implementar os modelos anteriores chamando a função respetiva. Para especificar o modelo de forma compacta dispõe-se da sintaxe do modelo `lavaan`. O ajustamento é feito com as funções já referidas, com recurso a estimadores de Máxima Verosimilhança (ML – *Maximum Likelihood*), Mínimos Quadrados Generalizados (GLS – *Generalized Least Squares*), Mínimos Quadrados Ponderados (WLS – *Weight Least Squares*), Mínimos Quadrados Ponderados na Diagonal (ULS – *Diagonally Weighted Least Squares*). É implementada ainda uma série de estimadores robustos, quer com o estimador ML quer com os estimadores WLS e ULS, que fornecem erros padrão robustos e uma estatística de teste corrigida (p.ex., com correções Satorra-Bentler), apenas para dados completos ou para dados completos e dados omissos. O estimador ML para dados omissos é o estimador FIML que é ativado com a opção `missing="ML"`. Caso contrário, é usado o método `listwise deletion`. O pacote é totalmente compatível com as estruturas de médias e multigrupos e os *outputs* apresentam soluções padronizadas, medidas de ajustamento, índices de modificação e outras informações. O modelo multinível pode ser implementado mas com algumas limitações. Dispõe da função `fitMeasures()` para calcular várias medidas da qualidade do ajustamento global, a função `bootstrapLavaan()` para aplicar o *bootstrap* a qualquer estatística de um objeto `lavaan` ou vetor de estatísticas. É possível ainda simular dados de um objeto `lavaan`, entre outras potencialidades. Quando as variáveis observadas são ordinais, a função `ordered()` aplicada à *data.frame* com os dados para análise ou o argumento `ordered` associado à(s) variável(eis) binária(s) ou ordinal(ais) numa função de ajustamento, faz com que o pacote implemente automaticamente o estimador WLSMV (mínimos quadrados ponderados ajustados para média e variância): são usados os mínimos quadrados ponderados na diagonal (DWLS) para estimar os parâmetros do modelo, mas é usada a matriz de pesos completa para calcular erros padrão robustos e um teste estatístico ajustado por média e variância.

O pacote **`blavaan`** (Merkle e Rosseel, 2015) fornece um conjunto flexível de ferramentas para estimar modelos *Bayesianos* de equações estruturais, usando a sintaxe do `lavaan` e as mesmas funcionalidades. É possível estimar versões Bayesianas de modelos SEM clássicos com a sintaxe `lavaan` e obter medidas Bayesianas de ajustamento de última geração associadas aos modelos.

5.2.3. PACOTE `OpenMx` (Neale *et al*., 2016)

Este pacote permite a estimativa de uma grande variedade de modelos estatísticos multivariados avançados. Consiste numa biblioteca de funções e otimizadores que permitem definir um modelo SEM de forma rápida e flexível e estimar os parâmetros a partir de dados observados.

Os modelos podem ser especificados diretamente através de álgebra matricial (através das matrizes de covariâncias e de estruturas de médias das variáveis latentes e observadas – Método de matrizes) ou através da formulação RAM ou LISREL. Para ajustar o modelo inclui estimadores FIML, ML e WLS. Os modelos disponíveis são dos mais diversos. Alguns dos mais populares que estão em uso atual incluem: análise fatorial confirmatória, autorregressão multivariada com atrasos cruzados, curvas de crescimento latente, modelos de mediação latente, de ecossistemas multivariados, modelos multigrupos com restrições, modelos epidemiológicos genéticos e genéticos comportamentais, modelos multivariados ordinais com estimativa de limiar, modelos de mistura de fatores, equações diferenciais latentes, modelos de classe latente.

5.2.4. PACOTE `lava` (Klaus *et al*., 2018)

É um pacote adequado para especificar e estimar modelos de variáveis latentes lineares, em particular, os modelos de equações estruturais. Abrange a análise clássica da estrutura de covariâncias. O pacote `lava.tobit` generaliza a estrutura para variáveis censuradas e dicotómicas através da formulação de uma função de ligação *probit*.

Subjacente à implementação está uma filosofia que consiste em separar a especificação do modelo dos dados reais, o que leva a uma forma dinâmica e fácil de modelar estruturas hierárquicas complexas. São implementados vários recursos avançados, incluindo erros padrão robustos para dados correlacionados em *cluster*, análises em multigrupos, restrições não-lineares de parâmetros, inferência com dados omissos, estimativas de máxima verosimilhança com observações censuradas e binárias e estimadores de variáveis

instrumentais. Além disso, contempla rotinas de simulação que abrange uma ampla gama de modelos de equações estruturais generalizadas não-lineares.

5.2.5. PACOTE `nlsem` (Umbach *et al*., 2017)

O pacote `nlsem` permite ajustar modelos de mistura de equações estruturais não-lineares para variável latente endógena, recorrendo ao algoritmo EM (*Expectation-Optimization*). São implementadas três abordagens diferentes: (i) LMS (Equações Estruturais Moderadas Latentes) e *QuasiMaximum Likelihood* (QML) que permitem interação bidirecional e termos quadráticos no modelo estrutural. Dado o carater não linear não pode ser assumida a normalidade multivariada das variáveis latentes, sendo aproximada por uma mistura de distribuições normais nos modelos LMS (Klein *et al*., 2007) e por um produto de densidade normal e densidade condicionalmente normal da função densidade do vetor indicador conjunto, no caso da QML (Klein e Moosbrugger, 2000); (ii) STEMM (Modelos de Mistura de Equações Estruturais) que usa misturas para modelar classes latentes e pode lidar com heterogeneidade na amostra ou com não-linearidade e não-normalidade das variáveis latentes e dos seus indicadores (Jedidi, Jagpal, & DeSarbo, 1997) e (iii) NSEMM (Modelos de Mistura de Equações Estruturais Não-Lineares), que mistura as duas abordagens anteriores permitindo a modelação de termos quadráticos e de interação bem como das classes latentes (Kelava, Nagengast & Brandt, 2014).

A especificação do modelo é feita usando a função `specify.sem()`. O *output* é um objeto `singleClass`, `semm` ou `nsemm`, dependendo da existência de interações e do número de classes na especificação do modelo. Para além do estimador EM, o modelo `singleClass` também pode ser ajustado com a função `qml()`.

O pacote **`plotSEMM`** (Kok *et al*., 2017) é adequado para representar as relações não-lineares entre variáveis latentes dos modelos de mistura de equações estruturais, permitindo investigar interações de variáveis latentes não-lineares em modelos de regressão latente, na medida em que permite visualizar relações não-lineares potenciais entre um preditor latente e os resultados.

5.2.6. PACOTE `lavaan.survey` (Oberski, 2014)

O pacote `lavaan.survey` é especialmente útil para lidar com amostras que não são independentes e identicamente distribuídas, permitindo modelar equações estruturais baseadas em *designs* amostrais complexos. Alavancado nos códigos do pacote `lavaan` e do pacote `survey`, este pacote permite a análise SEM com dados estratificados, *clusters*, dados de amostras ponderadas, correções de populações finitas, bem como com dados de *designs* complexos resultantes de imputação múltipla com recurso aos pacotes `mice`, `mi`, e `Amelia II` ou a programas externos.

Permite o ajustamento de modelos de equações estruturais, nomeadamente análise fatorial, modelos de regressão multivariada com variáveis latentes e muitos outros modelos de variáveis latentes, corrigindo estimativas de parâmetros, erros padrão e medidas de ajustamento derivadas do qui-quadrado, para *designs* amostrais complexos.

Ainda não é possível analisar dados categóricos no `lavaan.survey`, sendo esta uma limitação do pacote.

5.2.7. PACOTES `semPLS`, `plspm` e `SEMinR`

O pacote `semPLS` (Monecke e Leisch, 2012) proporciona uma modelação baseada nas variâncias, que é uma alternativa à SEM baseada na matriz de covariâncias e que é especialmente adequada para situações em que os dados não são normalmente distribuídos. Ajusta modelos de equações estruturais usando Mínimos Quadrados Parciais (PLS – *Partial Least Square*), com exigências mínimas em relação às escalas de medição, tamanho de amostra e distribuições residuais. A implementação dos PLS, com ênfase na SEM, é feita no pacote `plspm` (Sanchez, 2013). A função `plspm.fit()` deste pacote devolve uma lista, incluindo todos os parâmetros estimados e quase todas as estatísticas associadas aos modelos de caminho PLS: o modelo externo (modelo de medida), o modelo interno (modelo estrutural), as variáveis latentes padronizadas, pesos externos, pesos fatoriais, matrizes de coeficientes de caminho, R^2, correlações externas, modelo interno resumido, efeitos totais, unidimensionalidade, resultados de *boot-boot, bootstrap* (se for selecionada esta opção) e a matriz de dados. Um método gráfico cria uma representação gráfica dos modelos através da função `plot.plspm()`: do modelo de medida e do modelo estrutural, incluindo os parâmetros estimados. Para o tratamento da heterogeneidade

observada é fornecido como pacote complementar o `pathmox` (Sanchez e Aluja 2012), com uma abordagem de segmentação de árvores da modelação de caminhos PLS.

A implementação da SEM via PLS, em relação à SEM baseada nas covariâncias, tem a vantagem de não requerer suposições sobre a distribuição dos dados, ser robusta para pequenas amostras, para situações em que o número de variáveis é maior do que o número de indivíduos e em que existem modelos complexos e na presença de *missing data*. Hair *et al.* (2010) apresentam um conjunto de regras de ouro que, a seu ver, devem orientar o investigador na escolha do método a usar em cada caso: SEM baseada nas covariâncias ou SEM implementada via PLS.

`SEMinR` (Ray e Danks, 2018) cria e estima modelos de equações estruturais usando os Mínimos Quadrados Parciais na Modelação de Caminhos (PLS *Path Modelling* - PLS-PM). Usa a estimativa de variância baseada em PLS para modelar construções de fatores compostos e comuns.

5.2.8. PACOTE `metaSEM` (Cheung, 2015)

O pacote fornece uma coleção de funções para realizar meta-análises univariadas e multivariadas através de uma abordagem implementada com recurso ao pacote `OpenMx`. Implementa também a abordagem SEM de duas etapas (TSSEM – *Two Stage* SEM) (Cheung e Chang, 2005) para conduzir a modelação de equações estruturais meta-analíticas (MASEM), em matrizes de correlação ou covariância, para efeitos fixos e efeitos aleatórios. Para implementar a TSSEM é necessário dispor das matrizes de correlações ou de covariâncias e dos tamanhos das amostras e podem existir variáveis em falta nos diferentes estudos. Numa primeira etapa pode ser usada a Análise Fatorial Confirmatória para testar a homogeneidade das matrizes de correlação em todos os estudos. A matriz de correlação agrupada e a sua matriz de covariância assintótica podem ser obtidas nesta fase de análise. Se a homogeneidade das matrizes de correlação não for rejeitada, pode-se prosseguir para a segunda etapa. Se a hipótese de homogeneidade for rejeitada, os moderadores categóricos potenciais podem ser usados para classificar os estudos em subgrupos homogêneos. Na segunda etapa, a matriz de correlação agrupada e a sua matriz de covariância assintótica são utilizadas como entradas para o método de estimação ADF (*Asymptotically distribution-free*) do modelo. O tamanho total da amostra de todos os estudos é usado como o tamanho da amostra para ajustar o modelo.

5.2.9. PACOTE `fSRM` (Stas, Schönbrodt & Loeys, 2016)

Este pacote implementa quase automaticamente análises de SRM (*Social Relations Model*) via SEM, bastante complexas e introduz novas possibilidades para avaliar as diferenças entre médias e entre variâncias de SRM, tanto dentro como entre grupos de famílias. Usando dados familiares sobre processos negativos, são formulados diferentes tipos de questões de pesquisa e são apresentadas as análises correspondentes.

5.2.10. Pacotes adequados para análise de dados com características especiais

`Dlsem` (Magrini, 2018) ajusta modelos de atraso distribuído SEM (*distributed lag* SEM), com formas de atraso limitado, que são modelos para dados de séries temporais, em que equações de regressão são usadas para prever valores atuais de uma variável dependente com base nos valores atuais de uma variável explicativa e os valores de atraso (período passado) desta variável explicativa.

`strum` (Song *et al*., 2015) é um pacote que dispõe de ferramenta adequada à análise genética. Permite modelar a associação genética, a análise de ligações, efeitos poligénicos, o ambiente compartilhado, e averiguação combinada com a análise fatorial confirmatória e SEM geral. Fornece ainda uma ferramenta conveniente para visualização de modelos e integra ferramentas para simular dados de *pedigree*.

`Genomic SEM` (Grotzinger *et al*., 2018) é um pacote que assenta num novo método para modelar a arquitetura genética multivariada de constelações de características e incorporar a estrutura de covariância genética na descoberta multivariada de GWAS. Utilizando métodos SEM, modela formalmente a estrutura de covariância genética das estatísticas de resumo do GWAS a partir de amostras de graus de sobreposição variáveis e potencialmente desconhecidas. Permite ainda que o utilizador especifique e compare uma gama de diferentes arquiteturas genéticas multivariadas, o que melhora as abordagens existentes para combinar informações através de características geneticamente correlacionadas para auxiliar na descoberta.

`GW-SEM` (Verhulst, Maes & Neale, 2017) é um pacote que fornece funções específicas para estimar quatro modelos SEM comuns: um modelo de um fator, um modelo de

resíduos de um fator, um modelo de dois fatores e um modelo de crescimento latente (LGM).

ctsem (Driver, Oud & Voelkle , 2017) ajusta SEM de tempo contínuo utilizando equações diferenciais estocásticas lineares. Ao interagir com o `OpenMx`, o `ctsem` combina a especificação flexível de modelos de equações estruturais com oportunidades de recolha de dados aprimoradas e estimativa melhorada de modelos de tempo contínuo.

gSEM (Ma *et al*., 2016) fornece uma análise estatística de equações estruturais generalizadas, semi-supervisionadas num quadro de dados de observações coincidentes de variáveis contínuas múltiplas.

piecewiseSEM (Lefcheck, 2016) implementa a modelação de equações estruturais por partes.

rsem (Yuan e Zhang, 2012) implementa um procedimento robusto para estimar médias e matriz de covariâncias de múltiplas variáveis com dados omissos, usando o peso de Huber e, em seguida, estimar um modelo de equação estruturais usando o pacote `lavaan` ou o *software* EQS. Implementa modelos com variáveis auxiliares.

regsem (Jacobucci, 2017) implementa a Regularização em SEM. Incorpora várias formas de estimativa de verosimilhança penalizada numa ampla gama de modelos de equações estruturais. O `regsem` é particularmente útil para modelos de equações estruturais que têm um pequeno rácio entre o número de parâmetros e o tamanho da amostra, uma vez que a adição de penalidades pode reduzir a complexidade, reduzindo assim o viés das estimativas dos parâmetros.

lsl (Huang, 2017) é um pacote concebido para a realização de métodos de aprendizagem de estrutura latente. O modelo de equações estruturais, através de verosimilhança penalizada, pode ser implementado usando a classe de referência.

semtree (Brandmaier *et al.*, 2013) implementa o particionamento recursivo (SEM *Trees* e SEM *Forests*), o que corresponde à implementação dos modelos árvores de decisão à SEM. As SEM *Trees* dividem hierarquicamente os dados empíricos em grupos homogéneos que compartilham padrões de dados semelhantes em relação a um SEM, ou

seja, constroem estruturas de árvores que separam um conjunto de dados de forma recursiva em subconjuntos com estimativas de parâmetros significativamente diferentes, selecionando recursivamente preditores ótimos dessas diferenças. As SEM *Forests* são conjuntos de SEM *Trees* criados numa amostra aleatória dos dados originais. Ao agregar-se as SEM *Trees* numa SEM *Forest*, obtêm-se medidas de importância variável que são mais robustas do que medidas de árvores únicas.

Sesem (Lamb *et al.*, 2016) implementa um método simples da SEM espacialmente explícito com base na análise de matrizes de variâncias/covariâncias calculadas numa faixa de distâncias de desfasamento. Este método fornece gráficos prontamente interpretados da mudança nos coeficientes de caminho em função da escala.

5.2.11. Pacotes adequados para implementação de rotinas/etapas específicas da *SEM*

BigSEM (Chen e Zhang, 2016) constrói grandes sistemas de equações estruturais usando uma abordagem de Mínimos Quadrados Penalizados de duas etapas (2SPLS - *Two-stage Penalized Least Squares*).

SEMModComp (Levy, 2010) realiza testes de razão de verosimilhança para modelação de estruturas de médias e covariâncias em SEM.

semdiag (Zhang e Yuan, 2012) implementa diagnósticos a *outliers* e alavancagem; **influence.SEM** (Pastore e Altoe', 2018) dispõe de um conjunto de ferramentas para avaliar várias medidas de influência de casos para SEM; **semGOF** (Bertossi, 2012) proporciona um conjunto de catorze índices de qualidade de ajustamento para SEM.

SEMID (Barber, Drton & Weihs, 2017) fornece rotinas baseadas na representação gráfica dos modelos de equações estruturais por um diagrama de trajeto/gráfico misto, para verificar a identificação ou não - identificação de modelos de equações estruturais lineares.

RAMpath (Zhang *et al.*, 2015) executa análises regulares SEM através do pacote **lavaan**, mas possui recursos exclusivos. Assim, pode gerar diagramas de caminhos de acordo com um determinado modelo, pode exibir regras de rastreamento de caminhos através de diagramas de caminhos e decompor os efeitos totais nos respetivos efeitos

diretos e indiretos, bem como decompor a variância e a covariância em pontes individuais. Permite ainda ajustar modelos de sistemas dinâmicos automaticamente, com base em *scores* de mudanças latentes e gerar gráficos de campos vetoriais com base nos resultados obtidos de um sistema dinâmico bivariado.

semPlot (Epskamp, 2017) produz Diagramas de Caminhos e análise visual, para resultados de vários pacotes SEM (`lavaan`, `sem`, `OpennMx`, MPlus, LISREL, Onyx). O modelo pode ser especificado através da sintaxe do `lavaan`, da função `lisrelModel()` (usando matrizes do modelo LISREL estendido) e `ramModel()` (usando matrizes do modo RAM, explicitado mais à frente).

simsem (Jorgensen *et al*., 2018) fornece funcionalidades de simulação abrangente para a modelação SEM. Permite gerar dados com base num modelo especificado pelo utilizador, análises de dados gerados e armazenamento e processamento de resultados de simulação. A geração de dados e modelos de análise podem ser especificados usando a sintaxe `lavaan`, mais fácil para os utilizadores familiarizados com este pacote, ou as especificações do modelo `OpenMx`, para gerar dados com base em valores iniciais, ou num conjunto de matrizes.

FIAR (Roelstraete e Rosseel, 2011) é um pacote que permite executar algumas das técnicas mais populares e recentes para o estudo da integração funcional em redes cerebrais, nomeadamente Modelos de Equações Estruturais Autorregressivas (ARSEM). O pacote contém a função `ARsem()` que é um invólucro em torno da função `sem`() do pacote `lavaan`. A função toma como primeiro argumento o modelo de conectividade de uma análise clássica SEM. O modelo é especificado como um vetor que assume o valor 1 quando se assume uma conexão entre duas regiões e 0 caso contrário. As colunas representam as regiões "de" e as linhas as regiões "para". Os dados devem conter apenas as séries temporais (linhas) das regiões (colunas) no modelo. É especificada na função um argumento que define a ordem autorregressiva (AR) do modelo de conectividade.

MplusAutomation (Hallquist e Wiley, 2018) é um pacote que procura otimizar e agilizar a utilização do Mplus (Muthén e Muthén, 2009) para projetos complexos, como estudos de simulação de Monte Carlo ou a comparação de muitos modelos. Em particular,

o `MplusAutomation` fornece rotinas para: (i) criar e gerenciar a sintaxe para grupos de modelos relacionados; (ii) automatizar a estimativa de muitos modelos; e (iii) fornecer ferramentas para extrair e comparar estatísticas de ajustamento de modelos, estimativas de parâmetros e saídas de modelos auxiliares, a partir de quatro rotinas básicas que suportam esses objetivos: `createModels`, `runModels`, `readModels` e `compareModels`.

`REQS` (Mair, Wu & Wu, 2011) é um *interface* entre o ambiente R e o *software* EQS para SEM. O pacote consiste em três funções principais: `run.eqs()`, `call.eqs()` e `read.eqs()`. A função `run.eqs()` chama um arquivo de *script* EQS, executa a estimativa EQS e, finalmente, importa os resultados como objetos R. Estas ações podem ser executadas separadamente: a função `call.eqs()` chama e executa o *script* EQS, enquanto a função `read.eqs()` importa saídas existentes de EQS como objetos para R.

`xxM` (Mehta, 2013) é um pacote que permite a implementação da SEM multinível (ML-SEM), com estruturas de dados dependentes complexas e permite estimar modelos com qualquer número de níveis, com variáveis observadas e latentes em todos os níveis.

Este pacote está disponível apenas no *website* https://xxm.times.uh.edu/get-started/.

A estrutura de especificação do modelo por matrizes é projetada para ter recursos adicionais que os pacotes de `sem`, `lavaan` ou `OpenMx` não possuem (por exemplo, parâmetros aleatórios, não-especificação de modelo sofisticada, distribuição de fatores não normais e dados com covariáveis fixas).

Os dados podem ser gerados num formato e a análise pode ser implementada noutro formato.

`semTools` (Jorgensen, Pornprasertmanit & Schoemann, 2018) é um pacote que tem como objetivo reunir num único pacote um conjunto de funções úteis para a modelação SEM. Está projetado para ser suportado pelos utilizadores de SEM que são encorajados a enviar funções e ideias para funções adicionais. Inclui funções que estendem as potencialidades dos pacotes `lavaan` e `OpenMx`, mas dispõe também de outras funções que não estão ligadas a nenhum pacote específico. Tem funções para trabalhar com dados omissos, com a função `auxiliary()` que permite adicionar facilmente variáveis auxiliares ao estimador

FIML no `lavaan`, a função `runMI()` que, a partir de um conjunto de dados incompletos, faz imputação usando por exemplo os pacotes `Amelia`, `mice` ou dados fornecidos pelo utilizador, ajusta o modelo para dados imputados com o `lavaan` e agrupa os dados ou ainda a função `bsBootMiss()` que permite implementar o *bootstrap* com dados omissos. Tem funções para avaliação do modelo que incluem a determinação de índices adicionais, não incluídos no `lavaan` ou no `OpenMix` ou para implementar a análise de poder seja a um modelo seja a um par de modelos aninhados. Dispõe de funções para a medição da Invariância, para investigar a existência de interação entre variáveis latentes, para implementar a Análise Fatorial Exploratória e outras mais.

Cada formato para a especificação do modelo nos diferentes pacotes (`lavaan` ou `OpenMx`) tem um conjunto de características específicas, havendo depois um conjunto de características comuns a todos os formatos.

Assim, enumeram-se características em que é mais vantajoso usar um dos formatos:

a) *formato lavaan*:

- Gerar dados com base em parâmetros padronizados
- Entrada baseada em sintaxe para geração de dados e análise de dados
- Criar variáveis categóricas ordenadas endógenas

b) formato OpenMx:

- Criar variáveis categóricas ordenadas endógenas
- Simular dados com base em variáveis de definição
- Acomodar o modelo de mistura
- Gerar dados com base em parâmetros padronizados
- Gerar dados com parâmetros aleatórios
- Gerar dados com falta de modelo
- Controlar a ordem para 1) encontrar parâmetros não especificados (por exemplo, encontrar variâncias residuais quando as variações totais são especificadas), 2) impor restrições de igualdade/não-linear e 3) impor de falta de especificação do modelo
- Método sequencial para geração de dados (gerar dados no nível de fator e usá-los para criar dados de indicadores)

- Distribuição não-normal de fator e distribuição de erro normal
- Criar dados com base em covariáveis exógenas
- Implementar o Bollen-Stine *bootstrap*
- É um pouco mais rápido

No que respeita às características comuns aos formatos dos dois pacotes enumeram-se as seguintes:

- Processamento paralelo
- Distribuição não-normal do indicador (por cópula ou método de Vale e Maurelli)
- Impor *missing data* (MCAR, MAR ou *missing data* planeados)
- Simulação com diferentes amostras ou percentagens perdidas em repetições
- Restrições não-lineares e parâmetros definidos
- Gerar dados da saída de `lavaan`
- Imputação múltipla
- Modelação de variáveis auxiliares
- Análise de potência à significância das estimativas de parâmetros e da análise de poder na rejeição de maus modelos usando ajustamento de modelo absoluto, comparação de modelos aninhados ou comparação de modelos não aninhados, precisão na estimação de parâmetros, taxa de cobertura de intervalos de confiança.
- Transformar dados gerados e extrair saídas adicionais
- Executar uma simulação com base num conjunto de dados de população ou numa lista de conjuntos de dados de amostra
- Executar uma simulação até obter o número especificado de replicações convergentes
- Os utilizadores podem escrever uma função que devolve um vetor de estimativas de parâmetros, erros padrão, índices de ajustamento e *status* de convergência e usar a função na análise de dados gerados, que serão salvos automaticamente no resultado da simulação.

5.3. Alguns exemplos de modelação *SEM* com pacotes do R

5.3.1. Modelo RAM

No capítulo 3 foi apresentada a formulação matemática do modelo SEM na notação LISREL. Embora esta seja a notação mais comumente usada, existem várias maneiras equivalentes de representar modelos gerais de equações estruturais, nomeadamente a formulação em Modelo de Ação Reticular (RAM) (McArdle e McDonald, 1984; Fox *et al*., 2012), que tem uma conexão direta com o diagrama de caminhos subjacente e também cobre explicitamente modelos de análise de caminhos. Esta é a formulação usada, por exemplo, no pacote `sem` (Fox *et al*., 2017).

Seja U o vetor estocástico incluindo as variáveis latentes η e todas as variáveis observadas $Z = (Y_1, \dots, Y_p, X_1, \dots, X_q)$

$$U = (Z_1, Z_2, \dots, Z_{p+q}, \eta_1, \dots, \eta_l) \quad \textbf{(55)}$$

A formulação RAM do modelo fica

$$U = \upsilon_\theta + A_\theta U + \epsilon \quad \textbf{(56)}$$

onde υ_θ é o vetor de interceptos e ϵ é o termo de resíduos, que se assume ter distribuição $N_{p+q}(0, P_\theta)$, sendo $P_\theta = Var(\epsilon)$.

Assim, o modelo é completamente especificado pelos vetores $\upsilon_\theta, P_\theta$ e A_θ sendo as matrizes, geralmente, escassas e A_θ tem zeros na diagonal principal. Em termos gráficos, a matriz A_θ representa os caminhos assimétricos e P_θ representa os caminhos simétricos.

De forma a simplificar as equações estruturais pode-se considerar as variáveis centradas na média, os interceptos tornam-se zero e o modelo toma a forma

$$U = A_\theta U + \epsilon \quad \textbf{(57)}$$

A chave para estimar o modelo é a conexão entre as covariâncias das variáveis observadas, que podem ser estimadas diretamente a partir de dados de amostra, e os parâmetros em A_θ e P_θ. Seja r o número de variáveis em U e, sem perda de generalidade, as b primeiras variáveis sejam as variáveis observadas no modelo. Seja $J_{r\times r}$ a matriz de seleção para escolher as variáveis observadas:

$$J = \begin{bmatrix} I_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \textbf{(58)}$$

onde I_b é a matriz identidade de ordem b e **0** é a matriz de zeros de ordem apropriada.

O modelo implica as seguintes covariâncias entre as variáveis observadas:

$$C(\theta) = E(JUU^TJ^T) = J(I_r - A_\theta)^{-1}P_\theta(I_r - A_\theta)^{-1T}J^T \quad \textbf{(59)}$$

Seja S a matriz de covariâncias das variáveis observadas, calculadas diretamente da amostra.

O estimador ML que permite ajustar o modelo aos dados - isto é, estimar os parâmetros livres em A_θ e P_θ que tornam o S o mais próximo possível das covariâncias implícitas no modelo C, é dada, sob os pressupostos de que os erros e as variáveis latentes têm distribuição normal multivariada, por

$$F_{ML} = ln\left|\mathrm{C}(\hat{\theta})\right| - ln|S| + tr\left[S\left(\mathrm{C}(\hat{\theta})\right)^{-1}\right] - b \quad \textbf{(60)}$$

5.3.2. Especificação do modelo de acordo com a sintaxe específica de pacotes `SEM` do *software* R. Estimação do modelo

A etapa em que o modelo SEM é especificado é fundamental.

Com base num referencial, regra geral, teórico, é necessário definir um conjunto de relações entre as variáveis, observadas ou latentes, endógenas ou exógenas, que podem ser representadas através de um diagrama de caminhos ou de um conjunto de equações, sendo possível expressar cada uma das representações na outra forma. Assim, o investigador deve identificar cada um dos tipos de variáveis, bem como as relações entre elas, de forma a definir o modelo de medida e o modelo estrutural. Tem ainda que definir a estrutura de variância/covariâncias e de médias, se for do seu interesse.

Os pacotes mais relevantes do R que implementam a SEM, `sem`, `lavaan`, `OpenMx` têm formas diferentes de especificação do modelo.

Atualmente o pacote `sem` tem a formulação RAM e pode ser especificado através da função `specifyModel()`, através da função `specifyEquations`() usando as respetivas equações, ou ainda através da função `cfa()` para especificar um modelo de Análise Fatorial Confirmatória.

O `lavaan` dispõe de uma sintaxe específica que irá ser sintetizada e ilustrada mais adiante.

O `OpenMx` dispõe de dois métodos de especificação do modelo, o método da Análise de Caminhos (na formulação RAM ou na formulação LISREL) e o método das matrizes que exemplificaremos.

I. Especificação e estimação do modelo no pacote `sem` (Fox *et al*., 2017)

Usando a função `specifyModel()`, cada linha tem três entradas separadas por vírgulas.

- Na primeira entrada é definida a relação entre as variáveis: uma seta unidirecional indica um coeficiente de regressão e corresponde a uma seta unidirecional no diagrama de caminhos com a mesma orientação; uma seta bidirecional representa uma variância ou covariância e corresponde a uma seta bidirecional no diagrama de caminhos.
- A segunda entrada fornece o nome (arbitrário) de um parâmetro livre a ser estimado. Se o nome for NA (*missing*) significa que deve ser fixado um valor específico para o parâmetro. Se se atribuir o mesmo nome em duas ou mais linhas, esta ação estabelece uma restrição de igualdade entre os parâmetros correspondentes.
- A terceira entrada em cada linha atribui um valor a um parâmetro que se fixou ou define um valor inicial para um parâmetro livre; no último caso, inserir NA ou omitir faz com que a função de estimação (`sem` ou `cfa`) calcule o valor inicial.

Considere-se o modelo Industrialização e Democracia Política[4] especificado no diagrama de caminhos da Figura 5.1.

[4] Bollen, K. A. (1989). Structural Equations with Latent Variables. New York. - capítulo 8 citado em "Fox, J. and Weisberg, S. (2012, *last revision*). Descrição em anexo.

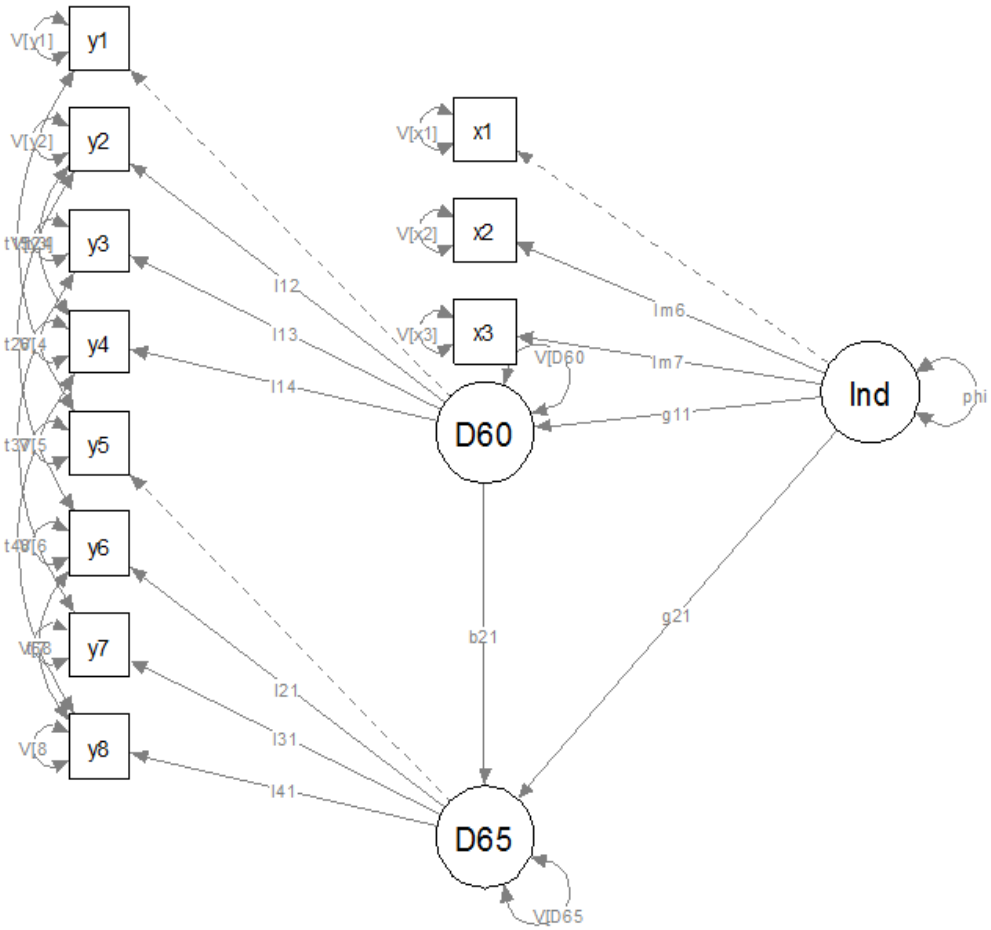


Figura 5.1: Modelo Industrialização e Democracia Política (Bollen, 1989).

O modelo especificado através do função `specifyModel()` fica:

```
> model.bollen.2<-specifyModel()
1:    Demo60 -> y1,       NA, 1
2:   Demo60 -> y2,        lambda12
3:    Demo60 -> y3,       lambda13
4:    Demo60 -> y4,       lambda14
5:   Demo65 -> y5 ,       NA, 1
6:    Demo65 -> y6,       lambda21
7:    Demo65 -> y7,       lambda31
8:    Demo65 -> y8,       lambda41
9:    Indust -> x1,       NA, 1
10:   Indust -> x2 ,      lam6
11:   Indust -> x3,       lam7
12: y1 <-> y5,            theta15
13:   y2 <-> y4,          theta24
14:   y2 <-> y6,          theta26
15:   y3 <-> y7,          theta37
16:   y4 <-> y8 ,         theta48
17:   y6 <-> y8 ,         theta68
```

```
18:  Indust -> Demo60,  gamma11
19:  Indust -> Demo65,  gamma21
20:  Demo60 -> Demo65,  beta21
21:  Indust <-> Indust, phi
22:
Read 21 records
NOTE: adding 3 variances to the model
```

Traduzindo estas relações num sistema de equações, o modelo pode ser especificado usando a função `specifyEquations()`:

```
>  model.bollen <- specifyEquations()
1:  y1 = 1*Demo60
2:  y2 = lam2*Demo60
3:  y3 = lam3*Demo60
4:  y4 = lam4*Demo60
5:  y5 = 1*Demo65
6:  y6 = lam2*Demo65
7:  y7 = lam3*Demo65
8:  y8 = lam4*Demo65
9:  x1 = 1*Indust
10:  x2 = lam6*Indust
11:  x3 = lam7*Indust
12:  c(y1, y5) = theta15
13:  c(y2, y4) = theta24
14:  c(y2, y6) = theta26
15:  c(y3, y7) = theta37
16:  c(y4, y8) = theta48
17:  c(y6, y8) = theta68
18:  Demo60 = gamma11*Indust
19:  Demo65 = gamma21*Indust + beta21*Demo60
20:  v(Indust) = phi
21:
Read 20 items
NOTE: adding 13 variances to the model
```

Estas equações são convertidas no formato RAM pela função. Note-se que não é incluído qualquer termo de erro nas equações. As variâncias dos erros são especificadas por meio dos argumentos `covs=` ou `v()=parâmetro` ou são adicionados automaticamente ao modelo quando está definida a opção `endog.variances=TRUE` (opção definida por defeito). É de referir ainda que em $covs = c("x_1", x_2")$ são estimadas apenas as variâncias de x_1 e de x_2, enquanto $covs = c("x_1, x_2")$ estima as variâncias e a covariância. Fixar um valor inicial para um parâmetro livre de x_1 corresponde a especificar esse valor do seguinte modo: `par(2)*` x_1, se se pretender que o valor inicial seja 2.

Estas duas funções são as mais comuns que o pacote `sem` dispõe para especificar o modelo (Fox *et al.*, 2017).

A função `sem()` estima modelos de equações estruturais gerais (com variáveis observadas e variáveis latentes): Análise Fatorial Confirmatória (CFA – *Confirmatory Factor Analysis*) usando a especificação do modelo com a função `cfa()` – mais simples (Fox, 2017), a CFA combinada com regressões nas variáveis latentes e modelos multigupo. Os dados podem ser fornecidos na forma de matriz de variâncias/covariâncias (das variáveis observadas) simétrica ou triangular, ou pode ser, também, uma matriz de momentos bruta (isto é, não corrigida – a soma dos quadrados e os produtos divididos por N e não por $N-1$). Em qualquer destes dois últimos casos tem que ser fornecido o número de dados. Os dados podem ser fornecidos na forma de uma *data-frame*.

A função `sem()` dispõe de estimadores de MV para dados completos e para dados omissos (`objectiveML` - Máxima Verosimilhança de Informação Completa normal multivariada, `objectiveFIML` - Máxima Verosimilhança de Informação Completa normal multivariada com dados omissos e `msemObjectiveML` - FIML normal multivariada multigrupo) e o estimador dos Mínimos Quadrados Generalizados (`objectiveGLS` - *Generalized Least Squares*).

Foi ajustado o modelo Industrialização e Democracia Política e obtido o *output* seguinte:

```
> sem.bollen <- sem(model.bollen, data=Bollen)
> summary(sem.bollen)

 Model Chisquare =  37.61688   Df =  35 Pr(>Chisq) = 0.3502626
 AIC =  99.61688
 BIC =  -113.4952

 Normalized Residuals
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.703373 -0.346526 -0.008777 -0.033954  0.181890  1.058388

 R-square for Endogenous Variables
Demo60     y1     y2     y3     y4 Demo65     y5     y6     y7     y8     x1
0.1996 0.7232 0.5143 0.5218 0.7152 0.9610 0.6529 0.5565 0.6784 0.6853 0.8461
    x2     x3
0.9468 0.7606

 Parameter Estimates
          Estimate   Std Error  z value    Pr(>|z|)
lambda12  1.25674622 0.18366822  6.8424804 7.783355e-12 y2 <--- Demo60
lambda13  1.05771642 0.15240266  6.9402753 3.913366e-12 y3 <--- Demo60
lambda14  1.26478659 0.14598269  8.6639489 4.556977e-18 y4 <--- Demo60
lambda21  1.18569692 0.16994727  6.9768519 3.018675e-12 y6 <--- Demo65
lambda31  1.27951250 0.16097853  7.9483427 1.890231e-15 y7 <--- Demo65
```

```
lambda41  1.26594759 0.15917601  7.9531307 1.818562e-15 y8 <--- Demo65
lam6      2.18036756 0.13944198 15.6363789 4.114703e-55 x2 <--- Indust
lam7      1.81851100 0.15298139 11.8871384 1.380559e-32 x3 <--- Indust
theta15   0.63209941 0.36560795  1.7288995 8.382708e-02 y5 <--> y1
theta24   1.33085641 0.71626050  1.8580620 6.316021e-02 y4 <--> y2
theta26   2.18195145 0.74869952  2.9143220 3.564621e-03 y6 <--> y2
theta37   0.80570567 0.62005813  1.2994034 1.938055e-01 y7 <--> y3
theta48   0.35293105 0.45123456  0.7821454 4.341291e-01 y8 <--> y4
theta68   1.37449259 0.57984417  2.3704517 1.776636e-02 y8 <--> y6
gamma11    1.48299988  0.40183636   3.6905567  2.237638e-04 Demo60 <--- Indust
gamma21    0.57233671  0.22280402   2.5687899  1.020543e-02 Demo65 <--- Indust
beta21     0.83734437  0.09901324   8.4568926  2.745965e-17 Demo65 <--- Demo60
phi        0.45449734  0.08845494   5.1381793  2.774131e-07 Indust <--> Indust
V[Demo60]  4.00949199  0.93992012   4.2657795  1.992054e-05 Demo60 <--> Demo60
V[y1]     1.91695469 0.45346274  4.2273698 2.364389e-05 y1 <--> y1
V[y2]     7.47250322 1.40183600  5.3305117 9.793642e-08 y2 <--> y2
V[y3]     5.13594770 0.97108480  5.2888766 1.230699e-07 y3 <--> y3
V[y4]     3.19043996 0.75380947  4.2324222 2.311879e-05 y4 <--> y4
V[Demo65]  0.17481220  0.21917326   0.7975982  4.251037e-01 Demo65 <--> Demo65
V[y5]     2.38274086 0.49000599  4.8626770 1.158087e-06 y5 <--> y5
V[y6]     5.02090675 0.93284034  5.3823860 7.350494e-08 y6 <--> y6
V[y7]     3.47774495 0.72734175  4.7814455 1.740392e-06 y7 <--> y7
V[y8]     3.29806077 0.70873188  4.6534675 3.263991e-06 y8 <--> y8
V[x1]     0.08265139 0.01988614  4.1562313 3.235403e-05 x1 <--> x1
V[x2]     0.12142542 0.07113866  1.7068836 8.784368e-02 x2 <--> x2
V[x3]     0.47300948 0.09199017  5.1419571 2.718913e-07 x3 <--> x3

 Iterations =  211
```

Note-se que o *output*, para além das estimativas dos parâmetros, dos R^2 e das variâncias, proporciona pouca informação sobre a qualidade do ajustamento do modelo aos dados. Proporciona apenas o teste Qui-Quadrado, e os índices de ajustamento AIC e BIC. Esta situação deve-se a não terem sido especificados os índices pretendidos na função `sem()`, quando se ajustou o modelo.

Indicando como opção todos os índices de ajustamento disponíveis, as primeiras linhas do *output* são as seguintes:

```
> sem.bollen <- sem(model.bollen, covBollen, 75, options(fit.indices = c("GFI",
"AGFI", "RMSEA", "NFI", "NNFI",
+ "CFI", "RNI", "IFI", "SRMR", "AIC", "AICc", "BIC", "CAIC")))

>  summary(sem.bollen)

Model Chisquare =  37.61688   Df =  35 Pr(>Chisq) = 0.3502626
 Goodness-of-fit index =  0.922671
 Adjusted goodness-of-fit index =  0.8541796
 RMSEA index =  0.03178646   90% CI: (NA, 0.09142272)
 Bentler-Bonett NFI =  0.9478204
 Tucker-Lewis NNFI =  0.9938246
 Bentler CFI =  0.9960702
 Bentler RNI =  0.9960702
 Bollen IFI =  0.9961848
 SRMR =  0.04441754
 AIC =  99.61688
 AICc =  83.75642
 BIC =  -113.4952
 CAIC =  -148.4952
```

Observando os resultados pode concluir-se que o ajustamento do modelo aos dados é bom, uma vez que o teste Qui-Quadrado é não significativo, ao nível de significância 5%, e que os valores dos índices de ajustamento indiciam, de acordo com o Quadro 3.3, um bom nível de ajustamento.

II. Especificação e estimação do modelo no pacote `lavaan` (Rosseel, 2018)

A sintaxe `lavaan` especifica o modelo através de grupos de equações que podem contemplar: (1) um conjunto de equações que especificam o modelo de medida e que têm a forma do modelo de regressão linear simples mas com o símbolo $=\sim$ a ligar a variável latente aos indicadores que a "medem"; (2) um grupo de equações de regressão entre variáveis latentes sendo, como usualmente, o símbolo ~ a ligar a variável dependente às variáveis preditoras; (3) um grupo de condições que especificam as correlações residuais do tipo $variável \sim\sim variável$. Se as variáveis forem iguais, fica definida a respetiva variância, se forem diferentes fica definida a covariância entre as variáveis. Pode ainda haver um outro conjunto de equações a definir as restrições aos parâmetros. No Quadro 5.1 encontra-se a síntese da sintaxe do pacote `lavaan`. O painel superior do quadro contém os quatro tipos de fórmula que podem ser usados para especificar um modelo na sintaxe `lavaan` do modelo. O painel inferior contém operadores adicionais que são permitidos na sintaxe `lavaan` do modelo.

Quadro 5.1: Síntaxe do modelo SEM no pacote `lavaan`

Tipo de fórmula	Operador	Significado
Variável latente	= ~	É manifestada por
Regressão	~	É dependente de (por regressão)
(Resíduo) (co)variância	~ ~	É correlacionada com
Intercepto	~ 1	Intercepto
Parâmetro definido	: =	É definido como
Restrição de igualdade	= =	É igual a
Restrições de desigualdade	< (>)	É menor (maior) do que

Na especificação do modelo podem ser tidas em conta diversas considerações sobre os parâmetros.

Assim, os parâmetros das equações podem ser livres ou sujeitos a restrições e a sintaxe contempla algumas regras:

1) Por defeito, os pesos do primeiro indicador nas equações do modelo de medida valem 1, para que a variável latente correspondente seja criada e a respetiva escala fique definida. Se quisermos alterar esta condição temos que pré-multiplicar o indicador por NA. Neste caso, será a variância da variável latente que valerá 1 para que esta seja criada.
2) Podemos usar modificadores que vão obrigar a que os parâmetros sejam sujeitos a determinadas condições no ajustamento do modelo, através de um mecanismo de pré-multiplicação com o operador *. Qualquer variável do segundo membro que não tenha pré-multiplicador é livre e o respetivo parâmetro será etiquetado com o par de variáveis a que diz respeito e que estão ligadas pela relação estabelecida entre elas nas equações (= ~, ~ ou ~ ~), com exceção dos primeiros indicadores das equações do modelo de medida, que têm peso 1.

Portanto, temos modificadores numéricos (fixam o valor do parâmetro), a função `equal()` que obriga a que as estimativas do parâmetros sujeitas à restrição sejam iguais, a função `start()` que define os valores iniciais dos parâmetros a quem está associado e que, por defeito, são gerados automaticamente pelo `lavaan`, as etiquetas que permitem introduzir equações com restrições não lineares ou com desigualdades. Podemos ainda obrigar que pares de fatores sejam ortogonais usando a pré-multiplicação por zero numa equação de covariâncias.

Por defeito, o `lavaan` fixa os interceptos como zero. Se pretendermos incluir uma estrutura de médias, incluímos um conjunto de equações de regressão para o intercepto das médias não nulas que pretendemos incluir, com equações do tipo: $variável \sim 1$ ou $variável \sim valor\ fixo * 1$, se se pretender fixar o seu valor.

Se pretendermos incluir todos os interceptos, basta usar o argumento `meanstructure=TRUE` na função de ajustamento.

O pacote `lavaan` dispõe de três funções de ajustamento, `sem()`,`cfa()` e `growth()`, que implementam os modelos de Equações Estruturais, Análise Fatorial

Confirmatória e Curvas de Crescimento Latente, como os nomes sugerem. A função `lavaan()`, usada para ajustar um modelo geral com variáveis latentes, fornece um invólucro para estes três modelos que podem ser chamadas diretamente ou através do argumento `model.type=" "`. Esta função, por defeito, ajusta um modelo `sem`. Os dados são fornecidos na forma de *data.frame* ou na forma de matriz de variâncias/covariâncias (`sample.cov=`) acompanhada do número de dados (`sample.nobs=`). Se algumas variáveis forem declaradas como fatores ordenados, o `lavaan` irá tratá-las como variáveis ordinais.

A função `lavaan()`, por defeito, usa o estimador ML para dados contínuos. No entanto, dispõe de diversos estimadores alternativos para dados completos e para dados omissos, a referir na seção 5.2.2.

Usando ainda os dados de Industrialização e Democracia Política, mas com condições muito diferentes das consideradas na especificação do modelo para o pacote `sem`, apresentam-se diversas situações que ilustram as possibilidades de especificação no `lavaan`.

Assim, no modelo seguinte estão ilustradas diversas considerações sobre os parâmetros: fixar o valor dos parâmetros (o peso de x_2 no fator ind60 é fixa $= 0.15$); impor que mais que um indicador tenham pesos fatoriais iguais (pesos dos indicadores y_2 e y_4 no fator dem60 são iguais); o peso do primeiro indicador num fator não seja 1, o que acontece por defeito (o peso de y_1 no fator dem60 é livre); fatores são ortogonais (dem60 e dem65 têm covariância zero); etiquetar parâmetros (na equação do fator dem65, os indicadores y_7 e y_8 têm os coeficientes etiquetados com a_1 e a_2); definir restrições com igualdade ou desigualdade sobre os parâmetros (estão definidas duas restrições envolvendo a_1 e a_2); definir o valor inicial para o parâmetro que, por defeito, é gerado automaticamente pelo `lavaan` (a estimação do modelo começa com o valor inicial 0.35 para o peso de x_3 no fator ind60).

```
> library(lavaan)
This is lavaan 0.6-1
lavaan is BETA software! Please report any bugs.

Attaching package: 'lavaan'

The following objects are masked from 'package:sem':

    cfa, sem
```

```
  > modelo <- '
  + # Modelo de medida (o peso do indicador x2 para o fator ind60 está fixo em
0.35,
  + # o peso dos indicadores y2 e y4 são iguais para o fator dem60;
  + # O peso do indicador y1 ficou livre com a pré-multiplicação por NA, uma vez
que
  + # vamos fixar a variância do fator dem60.
  + ind60 =~ x1 + 0.15*x2 + start(0.35)*x3
  + dem60 =~ NA*y1 + y2 + y3 + equal("dem60=~y2")*y4
  + dem65 =~ y5 + y6 + a1*y7 + a2*y8
  + # Regressões
  + dem60 ~ ind60
  + dem65 ~ ind60 + dem60
  + # (Co( Variâncias residuais
  + y1 ~~ y5
  + y2 ~~ y4 + y6
  + y3 ~~ y7
  + y4 ~~ y8
  + y6 ~~ y8
  + # Fatores ortogonais: demo60 e demo65
  + dem60~~0*dem65
  + # Fixar a variância do fator demo60
  + dem60~~1*dem60
  +   #Restrições nos parâmetros a1 e a2
  +   a1==a2*a2+1
  +   a2<0.515 '
```

Ajustando o modelo, obtemos o seguinte *output*:

```
>  fit <- sem(modelo,meanstructure=TRUE, data = PoliticalDemocracy)

> summary(fit, fit.measures=TRUE )

lavaan (0.6-1) converged normally after 152 iterations

  Number of observations                            75

  Estimator                                         ML
  Model Fit Test Statistic                     202.603
  Degrees of freedom                            38
  P-value (Chi-square)                           0.000

Model test baseline model:

  Minimum Function Test Statistic              730.654
  Degrees of freedom                            55
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.756
  Tucker-Lewis Index (TLI)                       0.647

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -1630.030
  Loglikelihood unrestricted model (H1)      -1528.728

  Number of free parameters                         39
  Akaike (AIC)                                3338.060
  Bayesian (BIC)                              3428.442
  Sample-size adjusted Bayesian (BIC)         3305.524
```

```
Root Mean Square Error of Approximation:

  RMSEA                                          0.240
  90 Percent Confidence Interval          0.208  0.273
  P-value RMSEA <= 0.05                          0.000

Standardized Root Mean Square Residual:

  SRMR                                           0.222

Parameter Estimates:

  Information                                 Expected
  Information saturated (h1) model          Structured
  Standard Errors                             Standard

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  ind60 =~
    x1                1.000
    x2                0.150
    x3                1.543    0.244    6.315    0.000
  dem60 =~
    y1                2.026    0.232    8.744    0.000
    y2      (.p5.)    2.289    0.277    8.249    0.000
    y3                2.168    0.306    7.082    0.000
    y4      (d60=)    2.289    0.277    8.249    0.000
  dem65 =~
    y5                1.000
    y6                0.777    0.108    7.172    0.000
    y7        (a1)    1.265
    y8        (a2)    0.515

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  dem60 ~
    ind60             0.714    0.216    3.304    0.001
  dem65 ~
    ind60             0.554    0.247    2.245    0.025
    dem60             1.780    0.213    8.355    0.000

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
 .y1 ~~
   .y5                0.322    0.358    0.901    0.368
 .y2 ~~
   .y4                1.476    0.728    2.026    0.043
   .y6                2.193    0.753    2.911    0.004
 .y3 ~~
   .y7                0.721    0.642    1.123    0.261
 .y4 ~~
   .y8                0.930    0.514    1.809    0.070
 .y6 ~~
   .y8                2.615    0.713    3.665    0.000
 .dem60 ~~
   .dem65             0.000

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)
   .x1                5.054    0.084   60.099    0.000
   .x2                4.792    0.163   29.469    0.000
   .x3                3.558    0.161   22.066    0.000
   .y1                5.465    0.303   18.053    0.000
   .y2                4.256    0.438    9.727    0.000
```

```
   .y3                6.563    0.379   17.309    0.000
   .y4                4.453    0.369   12.076    0.000
   .y5                5.136    0.308   16.697    0.000
   .y6                2.978    0.343    8.685    0.000
   .y7                6.196    0.398   15.552    0.000
   .y8                4.043    0.310   13.024    0.000
    ind60             0.000
   .dem60             0.000
   .dem65             0.000

Variances:
                  Estimate  Std.Err  z-value  P(>|z|)
   .dem60            1.000
   .x1               0.004    0.070    0.050    0.960
   .x2               1.971    0.322    6.123    0.000
   .x3               0.695    0.202    3.448    0.001
   .y1               1.665    0.454    3.664    0.000
   .y2               7.715    1.389    5.553    0.000
   .y3               4.821    0.938    5.139    0.000
   .y4               3.550    0.748    4.745    0.000
   .y5               1.986    0.466    4.258    0.000
   .y6               5.735    0.966    5.940    0.000
   .y7               3.722    0.806    4.618    0.000
   .y8               5.873    0.971    6.046    0.000
    ind60            0.527    0.111    4.732    0.000
   .dem65            0.188    0.300    0.626    0.531

Constraints:
                                             |Slack|
    a1 - (a2*a2+1)                             0.000
    0.515 - (a2)                               0.000
```

O modelo foi ajustado com a função `sem()`. Note-se que se fosse usada a função `lavaan()` seria este mesmo modelo a ser ajustado aos dados. Pelo *output* podemos verificar que o pacote dispõe de quantidade interessante de medidas de ajustamento. Caso não se forçasse a função `summary()` a apresentar os valores dos índices de ajustamento, nenhum seria apresentado. Apenas se disporia do *p-value* do teste χ^2. Observe-se ainda que ao impor o peso do primeiro indicador com um fator diferente de 1, a respetica variância vale 1.

É de referir que o objetivo deste exemplo não foi o de obter um bom modelo, mas sim o de ilustrar, por um lado a sintaxe `lavaan` para especificação do modelo e, por outro, a informação que é possível extrair do modelo. Foi usado o estimador ML pois é este que é aplicado por defeito quando todos os dados são contínuos. É necessário definir o estimador `(estimator=" ")` e o teste (`test=" "`) e ainda `missing=" "` para lidar com dados omissos. Comos seria de esperar o teste χ^2 resultou muito significativo e os índices têm valores que de acordo com o Quadro 3.3 revelam um mau ajustamento do modelo aos dados.

Na presença de dados omissos, por defeito, o `lavaan` recorre à *listwise deletion*. Se o mecanismo de omissão de dados for MCAR ou MAR, o pacote `lavaan` fornecerá uma estimativa de máxima verossimilhança FIML, que deverá ser ativada usando o argumento `missing="ML"` na função de ajustamento.

III. Especificação do modelo no pacote `OpenMx` (Neal *et al*., 2016)

No OpenMx o modelo é estimado em três etapas independentes. Numa primeira etapa são simulados dados a partir do modelo e são calculadas expectativas do modelo, em particular, a matriz de covariâncias e as médias, usando uma função da classe `mxExpectation`, nomeadamente `mxExpectationNormal()`, `mxExpectationRAM()` e `mxExpectationLISREL()`.Numa segunda etapa, a matriz de covariâncias (e/ou as médias) é (são) comparada(s) aos dados usando uma função de ajustamento, geralmente através de um cálculo de probabilidade, para determinar quão bem os dados se ajustam ao modelo. As funções da classe `mxfit` incluem a função `mxFitFunctionML()`e a função `mxFitFunctionWLS()`. A função `mxFitFunctionWLS()` calcula os mínimos quadrados ponderados das diferenças entre os dados e as expectativas implícitas no modelo para os dados, com base nos parâmetros livres e na função de expectativa utilizada. A função `mxFitFunctionML()` calcula $-2\,log\,(verosimilhança)$ de cada dado, tomados os valores presentes dos parâmetros livres e a função de expectativa selecionada para o modelo. Devolve a $\sum -2\,log\,(verosimilhança)$ das observações. Finalmente, é usado um algoritmo de otimização para encontrar o conjunto de parâmetros do modelo que minimizam o desajustamento do modelo aos dados. A função `mxRun()`implementa o processo de otimização dos parâmetros livres em objetos MxModel com base numa função de expectativa e numa função de ajustamento. Os objetos MxModel incluídos na função `mxRun()`devem incluir uma função apropriada de expectativa e de ajustamento.

Estas três etapas não são totalmente independentes uma vez que nem todas as expectativas podem ser operadas por todas as funções de ajustamento e por todos os algoritmos de otimização disponíveis

A especificação do modelo que vai ser usada na função de expectativas pode ter três tipos de formulação, a formulação RAM, a formulação LISREL e a formulação com

álgebra matricial. Nas duas primeiras formulações podem ser usadas matrizes ou caminhos (função `mxPath()` ou `mxMatrix()`).

A formulação RAM consiste em:

- Definir um vetor para cada variável latente com os indicadores que a quantificam. Esta ação facilita a definição da função `mxPath()` correspondente. Pode optar-se por fazer esta ação na própria função `mxPath()`;

- Criar um vetor com todas as variáveis observadas;

- Fixar o peso fatorial de um indicador em 1 para cada variável latente para a criar e a dimensionar na mesma unidade desse indicador (equivale ao primeiro indicador para cada fator no `lavaan`, que tem peso 1), ou fixar a variância de cada variável latente em 1 para criar a variável latente e dimensioná-la na unidade estandardizada. Neste caso, as covariâncias entre as variáveis latentes são interpretadas como as correlações. Também se pode optar pela combinação das duas ações para diferentes variáveis latentes.

Esta ação faz com que o sistema de equações que define o modelo tenha tantas equações como variáveis e o modelo seja identificado.

- Definir um vetor com as variáveis latentes.

- Definir o modelo `mxModel` começando por especificar as variáveis latentes e as variáveis observadas; usar funções `mxPath()` para: especificar os pesos fatoriais livres, dimensionar as variáveis latentes, especificar as variâncias das variáveis observadas, especificar as variâncias das variáveis latentes, especificar as covariâncias das variáveis latentes, especificar a estrutura de médias, caso se pretenda avaliar a estrutura de médias. Com a função `mxData()` anexar os dados ao modelo (em *data.frame* ou em matriz de covariâncias observadase o número de observações, como acontece no `lavaan`).

Para ajustar o modelo recorre-se à função `mxRun()`.

Mais uma vez consideremos o modelo Indústria e Democracia Política.

```
> library(OpenMx)
> observadas<-names(Bollen)
> latentes<-c("Dem60", "Dem65","Ind")
> model.B<-mxModel(model="Industria e Democracia", type="RAM",
+ manifestVars=observadas,
```

```
+ latentVars=latentes,
+ mxPath(from="Dem60", to=c("y1","y2","y3","y4"), arrows=1,
+       free=TRUE, values=1, labels=c("l1","l2","l3","l4") ),
+ mxPath(from="Dem65", to=c("y5","y6","y7","y8"), arrows=1,
+       free=TRUE, values=c(1,1,1,1), labels=c("l1","l2","l3","l4") ),
+ mxPath(from="Ind", to=c("x1","x2","x3"), arrows=1,
+       free=TRUE, values=c(1,1,1), labels=c("l5","l6","l7") ),
+ mxPath(from="Ind", to=c("Dem60","Dem65"), arrows=1,
+       free=TRUE, values=c(1,1), labels=c("g11","g21") ),
+ mxPath(from="Dem60", to=c("Dem65"),
+       arrows=1, labels=c("b21"), free=TRUE,values=0.8 ),
+ mxPath(from="y1", to=c("y5"),
+       arrows=1, labels=c("t15"),free=TRUE,values=0.8 ),
+ mxPath(from="y2", to=c("y4"),
+       arrows=1, labels=c("t24"),free=TRUE,values=0.8 ),
+ mxPath(from="y3", to=c("y7"),
+       arrows=1, labels=c("t37"),free=TRUE,values=0.8 ),
+ mxPath(from="y4", to=c("y8"),
+       arrows=1, labels=c("t48"),free=TRUE,values=0.8 ),
+ mxPath(from="y6", to=c("y8") ,
+       arrows=1, labels=c("t68"),free=TRUE,values=0.8),
+ mxPath(from=latentes,arrows=2,free=FALSE,
+ labels=c("phi1","phi2","phi3"),values=c(1,1,1)),
+
mxPath(from=observadas,arrows=2,labels=c("d1","d2","d3","d4","d5","d6","d7","d8",
"e1","e2","e3"),values=0.8),
+ #mxPath(from = "one", to = observadas), # introdução de uma estrutura de médias
+  mxData(covBollen, type="cov",numObs = 75) )
> model.B1<-mxRun(model.B)
> summary(model.B1)

   Summary of Industria e Democracia

   free parameters:
      name matrix   row   col   Estimate  Std.Error A
   1   t15      A    y5    y1 0.19345996 0.10651820
   2   t24      A    y4    y2 0.26817001 0.08604755
   3   t37      A    y7    y3 0.23949328 0.08652924
   4   t48      A    y8    y4 0.15739107 0.09591510
   5   t68      A    y8    y6 0.16740501 0.10701965
   6    l1      A    y1 Dem60 1.39405570 0.18262383
   7    l2      A    y2 Dem60 1.86742018 0.22608689
   8    l3      A    y3 Dem60 1.66073947 0.20304255
   9    l4      A    y4 Dem60 1.43655673 0.26090361
   10  b21      A Dem65 Dem60 0.74661907 0.14437760
   11   l5      A    x1   Ind 0.67190266 0.06462291
   12   l6      A    x2   Ind 1.45530167 0.12816623
   13   l7      A    x3   Ind 1.21775541 0.12827244
   14  g11      A Dem60   Ind 0.70909878 0.16914139
```

```
15  g21      A Dem65   Ind 0.21660832 0.17299843
16   d1      S    y1    y1 2.14735164 0.58040097
17   d2      S    y2    y2 8.27960984 1.59831671
18   d3      S    y3    y3 4.87532145 0.99454560
19   d4      S    y4    y4 2.84015692 0.68487014
20   d5      S    y5    y5 2.33423065 0.50969430
21   d6      S    y6    y6 4.73731965 1.08155744
22   d7      S    y7    y7 2.91502635 0.64556720
23   d8      S    y8    y8 2.50454946 0.54661889
24   e1      S    x1    x1 0.07853316 0.01949056
25   e2      S    x2    x2 0.13377447 0.07009006
26   e3      S    x3    x3 0.46674776 0.08860487

Model Statistics:
               |  Parameters  |  Degrees of Freedom  |  Fit (-2lnL units)
       Model:             26                    40              1628.350
   Saturated:             66                     0              1541.282
Independence:             11                    55              2271.936
Number of observations/statistics: 75/66

chi-square:  χ² ( df=40 ) = 87.06838,  p = 2.416137e-05
Information Criteria:
      |  df Penalty  |  Parameters Penalty  |  Sample-Size Adjusted
AIC:       7.068383              139.0684                168.3184
BIC:     -85.631141              199.3231                117.3779
CFI: 0.9303366
TLI: 0.9042128   (also known as NNFI)
RMSEA:  0.1252576  [95% CI (0.08165346, 0.1677037)]
Prob(RMSEA <= 0.05): 0.0008683751
timestamp: 2018-07-31 17:09:47
Wall clock time: 0.301018 secs
optimizer:  CSOLNP
OpenMx version number: 2.9.9
```

Note-se que as variâncias das variáveis latentes foram fixadas em 1 e que, portanto, as covariâncias entre as variáveis latentes devem ser interpretadas como as respetivas correlações.

O *output* fornece todos os parâmetros do modelo, erros padrão, o resultado do teste χ^2e alguns índices de ajustamento, nomeadamente os índices baseados na Teoria da Informação. O pacote não calcula muitos dos índices referidos na secção 3.6.2. dados os problemas a que estão sujeitos e que foram analisados nessa secção.

Usando o pacote `semPlot` obtém-se o Diagrama de Caminhos do modelo ajustado.

```
>  semPaths(model.B1, intercept = FALSE, whatLabel = "est",
+           residuals = TRUE, exoCov = FALSE, rotation=4, layout="tree2")
```

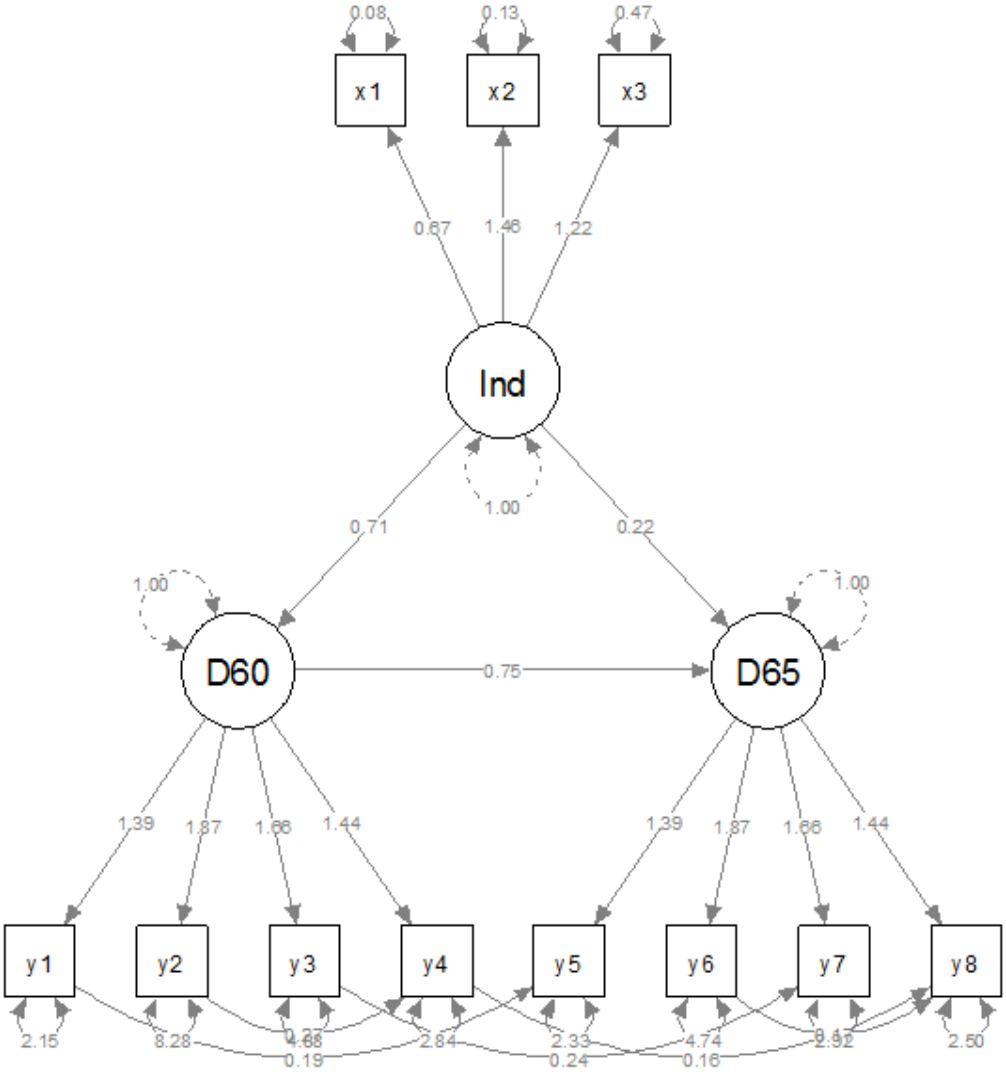


Figura 5.2: *Path Diagram* do modelo `model.B1`

O **método das matrizes** consiste em especificar uma estrutura de matrizes de covariância, mediante álgebra matricial.

Na Análise Fatorial Confirmatória a matriz R de covariâncias esperadas é dada por $R = ALA^T$, em que A é a matriz de pesos fatoriais, L é a matriz das intercorrelações (simétrica) e U é a matriz diagonal das variâncias das variáveis observadas. No modelo SEM geral, a matriz de covariância esperada é dada por

$$Cov = F(I - A)^{-1}S\,(I - A)^{-1}F^T \qquad \textbf{(61)}$$

sendo

- A a matriz assimétrica que define os caminhos assimétricos do modelo,

- S a matriz simétrica que define os erros e

- F, designada por matriz filtro, a matriz que define que variáveis são observadas e que variáveis são latentes.

A implementação do método consiste em:

- Definir um vetor com os nomes dos indicadores;

- Definir o modelo `mxModel` começando por:

- especificar a matriz de pesos fatoriais, incluindo os seus valores iniciais e os elementos que são livres;
- especificar a matriz de intercorrelação de fatores;
- especificar a matriz de variâncias dos fatores;
- especificar álgebra que resulta nas covariâncias esperadas do modelo;
- especificar um modelo para as médias, fixadas em zero.

Escolher o estimador entre as funções `mxExpectationNormal()` ou `mxExpecationRAM()` e a função a minimizar `MxFitFunctionML()`.

- Anexar os dados ao modelo;

- Ajustar o modelo com a função `mxRun()`.

Ilustremos estas ações ajustando o modelo com duas variáveis latentes e cinco variáveis observadas para definir cada fator (Boker *et al.*, 2011).

```
> data(TwoFactor)

> # ler os nomes das variáveis observadas a partir da dataframe
> observadas <- names(Towfactor)
> # especificar o modelo e armazená-lo em "Twofactor.m"
>  Twofactor.m <- mxModel("Two Factor",
+  # especificar a matriz dos pesos fatoriais, incluindo os valores
iniciais e
+ # quais elementos que são livres e os que são fixos
+              mxMatrix("Full",              nrow=10,              ncol=2,
values=c(1,rep(0.2,4),rep(0,10),1,rep(0.2,4)),
+               free=c(FALSE,rep(TRUE,4),rep(FALSE,10),FALSE,rep(TRUE,4)),
name="A"),
+ # especificar a matriz de intercorrelação dos fatores
+ mxMatrix("Symm", nrow=2, ncol=2, values=.8, free=TRUE, name="L"),
```

```
+ # especificar a matriz de variâncias exclusivas das variáveis latentes
+ mxMatrix("Diag", nrow=10, ncol=10, values=1, free=TRUE, name="U"),
+ # especificar a álgebra que resulta nas expectativas do modelo
+  mxAlgebra(A  %*%  L  %*%  t(A)  +  U,  dimnames  =  list(indicators,
indicators), name="R" ),
+ # especificar um modelo para os médias fixadas em zero
+    mxMatrix("Full",    nrow=1,    ncol=10,    values=0,    free=FALSE,
dimnames=list(NULL, indicators), name="M" ),
+ # escolha a função objetivo
+  mxExpectationNormal(covariance = "R")
+ # Anexar os dados ao modelo
+ mxData(Twofactor, type="raw") )
> #executar o modelo
> factorModelOut.m <- mxRun(factorModel)
Running Two Factor with 21 parameters
> # imprimir um resumo dos resultados
> summary(factorModelOut.m)
Summary of Two Factor

free parameters:
                  name matrix row col   Estimate  Std.Error A
1    Two Factor.A[2,1]      A   2   1 0.88922458 0.01738650
2    Two Factor.A[3,1]      A   3   1 0.78571245 0.01762168
3    Two Factor.A[4,1]      A   4   1 0.69146921 0.01601991
4    Two Factor.A[5,1]      A   5   1 0.59995290 0.01659923
5    Two Factor.A[7,2]      A   7   2 0.88220886 0.01991042
6    Two Factor.A[8,2]      A   8   2 0.78364326 0.01925239
7    Two Factor.A[9,2]      A   9   2 0.67946769 0.01849522
8   Two Factor.A[10,2]      A  10   2 0.61498704 0.02009256
9    Two Factor.L[1,1]      L   1   1 2.64704929 0.19029913
10   Two Factor.L[1,2]      L   1   2 1.10656919 0.11698068
11   Two Factor.L[2,2]      L   2   2 1.90453140 0.14070107
12   Two Factor.U[1,1]      U   1   1 0.37671872 0.02788014
13   Two Factor.U[2,2]      U   2   2 0.08472853 0.01020803
14   Two Factor.U[3,3]      U   3   3 0.16839981 0.01312125
15   Two Factor.U[4,4]      U   4   4 0.15131163 0.01143024
16   Two Factor.U[5,5]      U   5   5 0.22118984 0.01522579
17   Two Factor.U[6,6]      U   6   6 0.33483849 0.02520463
18   Two Factor.U[7,7]      U   7   7 0.09539094 0.01045257
19   Two Factor.U[8,8]      U   8   8 0.13445685 0.01121354
20   Two Factor.U[9,9]      U   9   9 0.15947383 0.01181517
21 Two Factor.U[10,10]      U  10  10 0.24607273 0.01686752

Model Statistics:
               |  Parameters  |  Degrees of Freedom  |  Fit (-2lnL units)
       Model:             21                  4979              9242.452
   Saturated:             65                  4935                    NA
Independence:             20                  4980                    NA
Number of observations/statistics: 500/5000

Information Criteria:
      |  df Penalty  |  Parameters Penalty  |  Sample-Size Adjusted
AIC:      -715.5477              9284.452                 9286.385
BIC:    -21700.0814              9372.959                 9306.304
To get additional fit indices, see help(mxRefModels)
timestamp: 2018-08-03 17:16:04
Wall clock time: 0.1990111 secs
optimizer:  CSOLNP
OpenMx version number: 2.9.9
```

Observe-se que o *output* apresenta um número reduzido de índices, situação que será analisada no próximo exemplo.

Representemos graficamente o Diagrama de Caminhos (Figura 5.3)

```
>    semPaths(factorModelOut.c,  intercept  =  FALSE,  whatLabel  =  "est",
color="lightblue" ,
+          residuals = TRUE, exoCov = FALSE,  layout="tree")
```

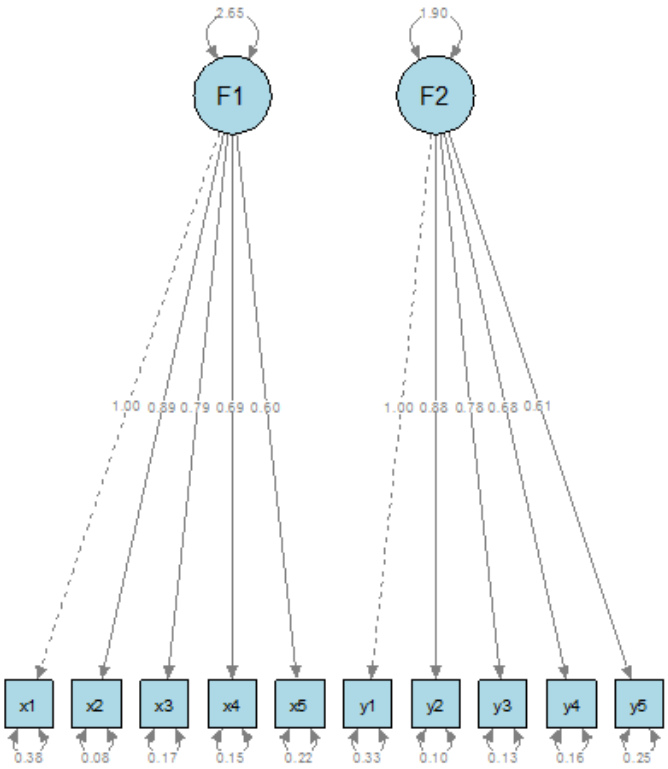


Figura 5.3: *Path Diagram* do modelo `factorModelOut.c`[5]

Embora o método de Análise de Caminhos (especificação do modelo com a função `mxPath()`) possa ser o mais adequado para alguns modelos, muitas vezes é mais fácil ou é necessário usar matrizes para especificar um modelo. Muitos dos modelos avançados disponíveis no OpenMx não possuem um Diagrama de Caminhos equivalente e, portanto, a álgebra da covariância deve ser especificada por meio de matrizes. Além disso, a

[5]. Para obter esta representação foi necessário especificar o modelo com recurso a caminhos (ajustados com a função `mxPath()`), uma vez que o pacote `semPlot` só ajusta modelos MxModel se todos os submodelos foram da classe MxRAMModel (Figura 5.3)

especificação da matriz é frequentemente mais compacta do que a especificação de todos os caminhos.

A formulação LISREL na forma matricial é feita especificando o modelo com as matrizes do modelo LISREL (expressão 61) e com as funções `mxExpectationLISREL()`.

Outro exemplo[6]:

Um modelo de medição para o humor e inteligência baseou-se em três perguntas de pesquisa, pedindo aos sujeitos que avaliassem quanto gostavam de *The Simpsons*, *Family Guy* e *American Dad*. O modelo contém ainda um caminho estrutural conectando inteligência ao humor.

Fazendo a especificação do modelo usando caminhos:

```
> library(foreign)
>    intell<-read.spss("http://www.methodsconsultants.com/data/intelligence.sav",
to.data.frame=TRUE)
>   names(intell)
[1] "reading"   "writing"   "math"      "analytic"  "simpsons"  "familyguy"
[7] "amerdad"
> observadas<-names(intell)
> latentes<-c("intelligence","humor")
> Intel<-mxModel("Inteligência", type="RAM",
+  manifestVars=observadas,
+  latentVars=latentes,
+  mxPath(from="humor", to=c("simpsons","familyguy","amerdad"),
+  free=c(FALSE,TRUE,TRUE), values=c(1,1,1), labels=c("l1","l2","l3") ),
+  mxPath(from="intelligence", to=c("reading","writing","math","analytic"),
+  free=c(TRUE,TRUE,TRUE,TRUE), values=c(1,1,1,1),
+  labels=c("l4","l5","l6","l7")),
+  mxPath(from="intelligence", to="humor",labels="g1"),
+  mxPath(from=observadas,arrows=2,labels=c("d1","d2","d3","d4","e1","e2","e3")),
+  mxPath(from=latentes, arrows=2,free=c(FALSE,TRUE),values=c(1,1)),
+  #mxPath(from = "one", to = observadas),
+  mxData(cov(intell),type="cov",numObs=100))
```

[6] Dados em https://www.methodsconsultants.com/tutorial/structural-equation-models-using-the-sem-package-in-r/

```
>  Intel.path<-mxRun(Intel)
Running Inteligência with 15 parameters
>   summary(Intel.path)
Summary of Inteligência

free parameters:
                  name matrix       row          col  Estimate  Std.Error A
1                   l4      A   reading intelligence 0.9379670 0.08316332
2                   l5      A   writing intelligence 0.8384168 0.07694728
3                   l6      A      math intelligence 0.9082987 0.07708918
4                   l7      A  analytic intelligence 0.8593180 0.07684932
5                   g1      A     humor intelligence 0.2404915 0.07983138
6                   l2      A familyguy        humor 1.0150917 0.09031129
7                   l3      A   amerdad        humor 1.0995753 0.08514730
8                   d1      S   reading      reading 0.2259886 0.04389873
9                   d2      S   writing      writing 0.2193856 0.03949217
10                  d3      S      math         math 0.1603442 0.03494912
11                  d4      S  analytic     analytic 0.2011758 0.03749557
12                  e1      S  simpsons     simpsons 0.1158015 0.03134843
13                  e2      S familyguy    familyguy 0.2599027 0.04627608
14                  e3      S   amerdad      amerdad 0.1503261 0.03858122
15 Inteligência.S[9,9]      S     humor        humor 0.5091682 0.08970626

Model Statistics:
               |  Parameters  |  Degrees of Freedom  |  Fit (-2lnL units)
       Model:             15                    13              44.10016
   Saturated:             28                     0              30.31894
Independence:              7                    21             621.27394
Number of observations/statistics: 100/28
chi-square:  χ² ( df=13 ) = 13.78122,  p = 0.3894301
Information Criteria:
      |  df Penalty  |  Parameters Penalty  |  Sample-Size Adjusted
AIC:      -12.21878              43.78122                       NA
BIC:      -46.08599              82.85877                   35.485
CFI: 0.9986293
TLI: 0.9977858   (also known as NNFI)
RMSEA:  0.02451404  [95% CI (0, 0.1145892)]
Prob(RMSEA <= 0.05): 0.6136771
optimizer:  CSOLNP
```

Fazendo a especificação do modelo usando matrizes:

```
> Intel.m <- mxModel("Inteligência",
+       mxMatrix("Full", 9, 9,
+
free=c(rep(FALSE,63),rep(TRUE,4),rep(FALSE,4),TRUE,rep(FALSE,5),rep(TRUE,2),rep(F
ALSE,2)),
+
values=c(rep(0:0,63),rep(1:1,4),rep(0:0,4),1,rep(0:0,4),rep(1:1,3),rep(0:0,2)),na
me="A",
+                  dimnames=list(c(observadas,latentes),c(observadas, latentes)),
byrow=FALSE),
+        mxMatrix("Full", 9, 9,
+                              free=c(TRUE,  rep(FALSE,9),TRUE,  rep(FALSE,9),TRUE,
rep(FALSE,9),TRUE,
+                         rep(FALSE,9),TRUE, rep(FALSE,9),TRUE, rep(FALSE,9),TRUE,
rep(FALSE,19),TRUE),
+             values=c(1,rep(0:0,9),1,rep(0:0,9),1,rep(0:0,9),1,rep(0:0,9),1,
+                      rep(0:0,9),1,rep(0:0,9),1,rep(0:0,9),1,rep(0:0,9),1),
+                               dimnames=list(c(observadas,latentes),c(observadas,
latentes)), name="S",byrow=FALSE),
+               mxMatrix("Full",  7,  9,  values=c(1,  rep(0:0,7),1,  rep(0:0,7),1,
rep(0:0,7),1,
+                    rep(0:0,7),1, rep(0:0,7),1, rep(0:0,7),1,rep(0:0,14)),
+             name="F", dimnames=list(observadas,c(observadas,latentes))),
+                     mxMatrix("Full",  1,  9,  name="M",  values=c(rep(0:0,9)),
free=c(rep(FALSE,9)),byrow=FALSE),
+        mxExpectationNormal("A", "S", "F",
+              dimnames =c(observadas,latentes)),
+              mxFitFunctionML(),
+                    #mxMatrix("Full",  nrow=1,  ncol=7,  values=0,  free=FALSE,
dimnames=list(NULL, NULL), name="M" ),
+       # mxData(cov(intell), type="cov",numObs=100))

+       mxData(observed=intell, type="raw"))
>    intelFit.m<-mxRun(Intel.m)


> summary(intelFit.m)
Summary of Inteligência

free parameters:
                  name matrix       row          col  Estimate  Std.Error A
1  Inteligência.A[1,8]      A   reading intelligence 0.9449307 0.08351610
2  Inteligência.A[2,8]      A   writing intelligence 0.8417415 0.07707880
3  Inteligência.A[3,8]      A      math intelligence 0.9075950 0.07727951
4  Inteligência.A[4,8]      A  analytic intelligence 0.8667470 0.07748796
5  Inteligência.A[9,8]      A     humor intelligence 0.2385208 0.07980906
6  Inteligência.A[6,9]      A familyguy        humor 1.0152192 0.09077342
7  Inteligência.A[7,9]      A   amerdad        humor 1.1010749 0.08542098
8  Inteligência.S[1,1]      S   reading      reading 0.2252645 0.04399622
9  Inteligência.S[2,2]      S   writing      writing 0.2186573 0.03945540
10 Inteligência.S[3,3]      S      math         math 0.1638218 0.03529624
11 Inteligência.S[4,4]      S  analytic     analytic 0.2042539 0.03810556
12 Inteligência.S[5,5]      S  simpsons     simpsons 0.1169587 0.03149209
13 Inteligência.S[6,6]      S familyguy    familyguy 0.2625547 0.04666014
14 Inteligência.S[7,7]      S   amerdad      amerdad 0.1494548 0.03866183
15 Inteligência.S[9,9]      S     humor        humor 0.5093009 0.08989999

Model Statistics:
               |  Parameters  |  Degrees of Freedom  |  Fit (-2lnL units)
       Model:             15                   685              1334.805
   Saturated:             35                   665                    NA
Independence:             14                   686                    NA
Number of observations/statistics: 100/700
```

```
Information Criteria:
      |  df Penalty  |  Parameters Penalty  |  Sample-Size Adjusted
AIC:       -35.19474              1364.805                 1370.520
BIC:     -1819.73632              1403.883                 1356.509

optimizer:  CSOLNP
```

Note-se que, neste caso, contrariamente ao que aconteceu no *output* anterior, são fornecidos apenas dois índices de ajustamento para além dos que são baseados na Teoria da Informação. A função `mxRefModels()` proporciona mais informação a este respeito.

```
> summary(intelFit.m, refModels=mxRefModels(intelFit.m, run = TRUE))

Running Saturated Inteligência with 35 parameters
Running Independence Inteligência with 14 parameters
Summary of Inteligência
…
Model Statistics:
               |  Parameters  |  Degrees of Freedom  |  Fit (-2lnL units)
       Model:             15                   685              1334.805
   Saturated:             35                   665              1316.798
Independence:             14                   686              1907.753
Number of observations/statistics: 100/700

chi-square:  χ² ( df=20 ) = 18.00761,  p = 0.5869072
Information Criteria:
      |  df Penalty  |  Parameters Penalty  |  Sample-Size Adjusted
AIC:       -35.19474              1364.805                 1370.520
BIC:     -1819.73632              1403.883                 1356.509
CFI: 1.003496
TLI: 1.00367   (also known as NNFI)
RMSEA: 0 *(Non-centrality parameter is negative)  [95% CI (0, 0.08665643)]
Prob(RMSEA <= 0.05): 0.8150079
```

Podemos observar que os efeitos positivos diretos da inteligência e do humor sobre as variáveis observadas que as mediram (pesos fatoriais) podem ser considerados adequados, uma vez que o ajustamento pode ser considerado bom, tendo em conta que o teste Qui-Quadrado é não significativo ao nível de significância de 5% (*p-value>0.05*) e os índices de avaliação da qualidade do ajustamento têm valores que, de acordo com o Quadro 3.3, permitem tomar o ajustamento como bom ou muito bom. Verifica-se ainda que existe associação entre a inteligência e o humor, confirmando o que é expectável.

Representemos graficamente o Diagrama de Caminhos com as estimativas dos coeficientes do modelo (Figura 5.4).

```
>semPaths(Intel.path, intercept = FALSE, whatLabel = "est", color="lightblue" ,
+residuals = TRUE, exoCov = FALSE, rotation=4, layout="tree")
```

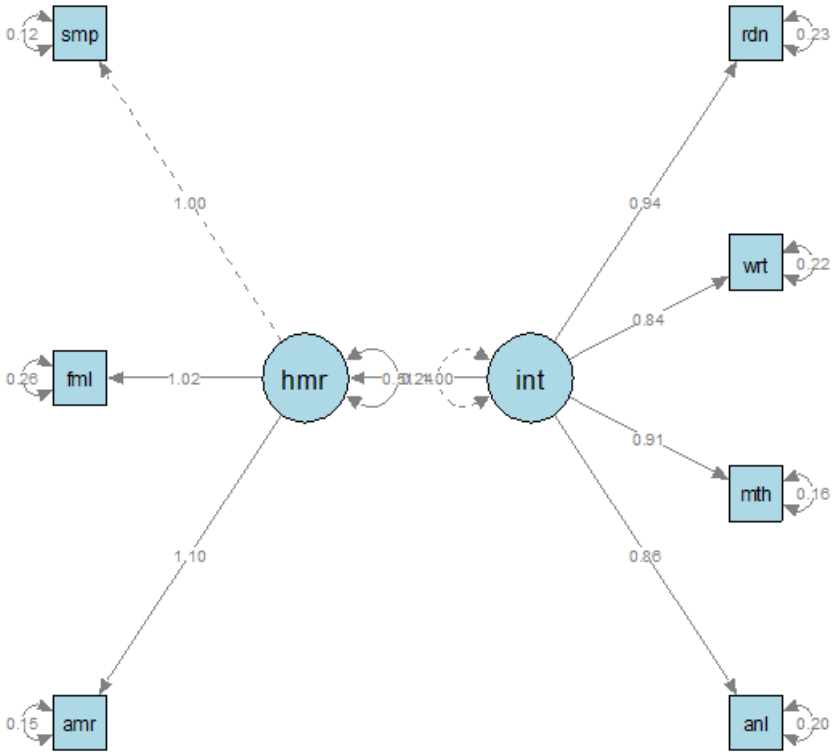


Figura 5.4: *Path Diagram* do modelo `inteFit.m`

5.3. Conclusões

O R dispõe de uma grande panóplia de ferramentas ao serviço da SEM. No presente capítulo, exploramos apenas as ferramentas diretamente relacionadas com especificação e o ajustamento do modelo. Muito mais pode ser feito com recurso ao R, nomeadamente a exploração e caracterização dos dados, como ilustrado, em parte, no capítulo 4 com a verificação de pressupostos, em particular a existência de dados omissos e respetivos padrões, a normalidade multivariada e a homogeneidade de variâncias e com as diferentes formas de implementar a metodologia na presença de dados omissos.

O tema da SEM é de tal forma abrangente e as ferramentas disponíveis são de tal forma completas que é impossível, num trabalho deste âmbito, explorar devidamente todos os aspetos. Muitos ficam por explorar e por exemplificar para além dos enunciados. Refira-se em especial os pacotes adequados para estudos de simulação como o `simsem`, que tem imensas potencialidades e o `lavaan.survey` que possibilita o desenvolvimento de estudos em projetos com planeamentos amostrais complexos, nomeadamente na área da Educação.

CAPÍTULO 6

CONSIDERAÇÕES FINAIS E PERSPETIVAS DE INVESTIGAÇÃO FUTURA

CAPÍTULO 6

CONSIDERAÇÕES FINAIS E PERSPETIVAS DE INVESTIGAÇÃO FUTURA

CONSIDERAÇÕES FINAIS E PERSPETIVAS DE INVESTIGAÇÃO FUTURA

Desde a introdução do método de análise de trajetórias por Wright (1918) que a SEM, como abordagem estatística, ganhou impulso e progrediu de forma constante nos últimos anos.

Nas últimas décadas, o poder computacional e estatístico cresceu e vários programas de *software* foram desenvolvidos e provaram que a abordagem SEM é uma opção viável para a investigação de redes complexas de relacionamentos, mas também para representar conceitos teóricos usando variáveis latentes, além de permitir lidar com um elevado volume de dados.

O desenvolvimento da SEM iniciou-se, no princípio do século XX, por duas vias. Uma, pela via da biometria, com a Análise de Caminhos (*Path Analysis*) e a modelação gráfica para implementar a análise causal de sistemas biológicos e que posteriormente se aplicou à Econometria e às Ciências Sociais, e outra através da Análise Fatorial que surgiu no âmbito da Psicologia e da Psicometria. Jöreskog (1973), entre outros, combinou estas duas vias, a análise fatorial e a modelação de equações simultâneas, numa abordagem analítica coerente – o modelo LISREL. Como referido a SEM tem-se desenvolvido numa abordagem que integra a tradução da teoria (Grace *et al*., 2010), a inferência causal (Pearl, 2012) e a especificação estatística (Lee, 2007) num único processo de modelação, levando a que, segundo Grace *et a.l* (2012), a prática da SEM englobe uma metodologia científica mais completa.

O desenvolvimento computacional impulsionou a conceção de métodos estatísticos para melhorar a qualidade da produção científica e a automatização da recolha e armazenamento de dados, potenciando um aumento dramático da complexidade dos modelos e dos métodos. A SEM não foi exceção. Beneficiou, por um lado, com o desenvolvimento de diversos *softwar*es para a análise SEM, uns comerciais e outros livres, destes últimos a maioria disponível no *software* R e capazes de rivalizar com os comerciais. Por outro lado, por ser adequada para lidar com grandes volumes de dados, a metodologia foi objeto de desenvolvimentos e aplicações com complexidade crescente e numa vasta gama de domínios do conhecimento.

No presente trabalho dá-se conta de uma grande diversidade de aplicações da metodologia SEM, com especial ênfase nas Ciências Naturais e Ciências da Vida, tendo como objetivo ilustrar a importância que a metodologia readquiriu na biometria, área que esteve na sua génese, após ter passado por um longo período em que os desenvolvimentos metodológicos e as aplicações tiveram o foco principal nas Ciências Sociais e Humanas e na Psicometria.

Os desenvolvimentos metodológicos da SEM foram impulsionados quer pela diversidade de áreas de aplicação quer pelo desenvolvimento da capacidade computacional. A diversidade de áreas de aplicação, por um lado coloca desafios para a especificação adequada de ideias teóricas em modelos de equações estruturais, e por outro potencia a pesquisa de ferramentas estatísticas e de modelos capazes de lidar com diversos tipos de dados amostrais e diversos tipos de planeamentos experimentais. Computadores eficientes no armazenamento e processamento de grandes conjuntos de informações disponíveis hoje pelas mais diversas vias, grande profusão de programas de análise de dados e bibliografia disponível, seja com aspetos teóricos e metodológicos da SEM, seja com estratégias e aplicações específicas ou ainda com descrição de *softwares* para aplicação da SEM, tornaram a SEM acessível a um maior número de investigadores e a uma maior diversidade de áreas de investigação.

A metodologia SEM tem tal vastidão que se estende a um nível de flexibilidade maior que o dos Modelos Lineares Generalizados (Kline, 2011), que se apresenta como um caso particular da SEM, e contempla ainda modelos não lineares, modelos hierárquicos (multinível) e abordagem Bayesiana, para além de outras especificidades dependentes da área de aplicação.

Não é possível entender a aplicação da SEM em abordagens mais avançadas, sem conhecer os conceitos chave e a formulação básica da metodologia, bem como os detalhes técnicos sobre a análise de dados, interpretação e relatórios. Assim, foi feita uma revisão da SEM clássica, a SEM baseada na estrutura de covariâncias, que se tentou fosse o mais completa possível, tendo em conta as limitações deste tipo de trabalho e a dificuldade em sintetizar todos os detalhes, os cuidados e as questões práticas que há que ter na implementação da metodologia.

Os dados reais dificilmente são completos. Por um ou outro motivo há dados omissos no conjunto de dados em análise. Uma das questões aqui tratada com mais cuidado foi a omissão de dados e a forma como o *software* lida com essa situação.

Considerando que a implementação da metodologia depende da utilização de *software*, uma vez que requer grandes amostras, esta é uma questão incontornável. Durante muito tempo, o *software* LISREL foi o único disponível para implementar a SEM. Entretanto, deu-se a proliferação de *softwares* de implementação mais simples e que não exigem quase conhecimentos técnicos sobre a metodologia. No entanto, com exceção dos *softwares* em ambiente R, do Onyx e do JASP, todos os restantes são comerciais.

Dada a flexibilidade do ambiente R que permite o desenvolvimento integrado de cálculos estatísticos e gráficos e a enorme diversidade de ferramentas disponíveis para diferentes etapas da SEM, neste trabalho apresentou-se uma resenha dos pacotes que, de alguma forma, permitem implementar a SEM nas suas diferentes abordagens, de acordo com diferentes contextos, bem como implementar tarefas associadas, nomeadamente a verificação de pressupostos, as representações gráficas e o tratamento de dados omissos.

A SEM foi implementada com alguns exemplos e com diferentes pacotes.

Os resultados do programa PISA revelaram recentemente que em Portugal se tem feito um caminho muito positivo, nomeadamente nos resultados em Matemática. Para aprofundar o exemplo usado no capítulo 4, no tratamento de dados omissos, uma possível exploração futura consistirá na realização de estudos comparativos neste âmbito entre os resultados de 2003, 2012 e 2015, com recurso à metodologia SEM multinível e num estudo longitudinal. Poderá ser uma tarefa muito complexa, dada a também complexa estrutura das bases de dados, mas será certamente uma tarefa muito desafiante.

No âmbito dos processos de autoavaliação das escolas, a SEM oferece um sem número de possibilidades, seja no que se refere à avaliação da satisfação com serviços (administrativos e educativos), seja no que se refere ao sucesso escolar. Esta será também uma área de exploração futura.

O *software* R constituir-se-á num auxiliar precioso numa tarefa desta natureza.

A exploração da metodologia e do software R em áreas da saúde é um objetivo futuro, tendo em vista o desenvolvimento de competências noutras áreas para além da Educação, em conformidade com o percurso formativo que vem sendo desenvolvido.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Afari, E. (2013). The effects of psychosocial learning environment on students' attitudes towards mathematics. In Application of structural equation modeling in educational research and practice (pp. 91-114). SensePublishers, Rotterdam.
2. Ahn, J. W. (2017). Structural Equation Modeling of Cultural Competence of Nurses Caring for Foreign Patients. Asian nursing research, 11(1), 65-73.
3. Allison, P. D. (2003). Missing data techniques for structural equation modeling. Journal of abnormal psychology, 112(4), 545.
4. Angelini, M. E., Heuvelink, G. B. M., & Kempen, B. (2017). Multivariate mapping of soil with structural equation modelling. European Journal of Soil Science.
5. Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. Psychometrika, 65(4), 475-496.
6. Ansari, A., & Jedidi, K. (2001). Bayesian structural equation models for multilevel data. In New developments and techniques in structural equation modeling (pp. 149-178). Psychology Press.
7. Ansari, A., Jedidi, K., & Jagpal, S. (2000). A hierarchical Bayesian methodology for treating heterogeneity in structural equation models. Marketing Science, 19(4), 328-347.
8. Arhonditsis, G. B., Stow, C. A., Steinberg, L. J., Kenney, M. A., Lathrop, R. C., McBride, S. J., & Reckhow, K. H. (2006). Exploring ecological patterns with structural equation modeling and Bayesian analysis. Ecological Modelling, 192(3-4), 385-409.
9. Azadi, H., Barati, A. A., Rafiaani, P., Raufirad, V., Zarafshani, K., Mamoorian, M., ... & Lebailly, P. (2016). Agricultural Land Conversion Drivers in Northeast Iran: Application of Structural Equation Model. Applied Spatial Analysis and Policy, 9(4), 591-609.
10. Babenko, O., Alves, C. B., & Bahry, L. M. (2012). Using Structural Equation Modeling to Investigate Students' Career Awareness in Science. CJNSE/RCJCÉ, 4(1).
11. Barber, R. F., Drton, M., Weihs, L. (2017). SEMID: Identifiability of Linear Structural Equation Models. R package version 0.3.1
12. Beaujean, A. A., & Beaujean, M. A. A. (2018). Package 'BaylorEdPsych'. Version 0.5
13. Bentler, P. M., & Wu, E. J. (2005). EQS 6.1 for Windows. Encino, CA: Multivariate Software INC.
14. Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. Multivariate behavioral research, 34(2), 181-197.
15. Beran, T. N., & Violato, C. (2010). Structural equation modeling in medical research: a primer. BMC research notes, 3(1), 267.
16. Bernstein, B. B. (1975). Class, codes and control. Towards a theory of educational transmissions (Vol. 3). London: Routledge.
17. Bertossi, E. (2012). semGOF: Goodness-of-fit indexes for structural equation models. R package version 0.2-0
18. Blüthgen, N., Simons, N. K., Jung, K., Prati, D., Renner, S. C., Boch, S., ... & Tschapka, M. (2016). Land use imperils plant and animal community stability through changes in asynchrony rather than diversity. Nature communications, 7, 10697.
19. Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., ... & Mehta, P. (2011). OpenMx: an open source extended structural equation modeling framework. Psychometrika, 76(2), 306-317.

20. Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In Handbook of causal analysis for social research (pp. 301-328). Springer, Dordrecht.
21. Bollen, KA.; Stine, RA. (1993).Bootstrapping goodness-of-fit measures in structural equation models. In: Bollen, KA.; Long, JS., editors. Testing structural equation models. Sage; Newbury Park, CA: p. 111-135.
22. Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog, 2(3), 139-168.
23. Borhan, N., & Zakaria, E. (2017, May). Structural equation modeling assessing relationship between mathematics beliefs, teachers' attitudes and teaching practices among novice teachers in Malaysia. In AIP Conference Proceedings (Vol. 1847, No. 1, p. 030012). AIP Publishing.
24. Bourdieu, P., & Passeron, J.-C. (1977). Reproduction in education, society and culture. Theory, Culture & Society. London: Sage Publications, Inc.
25. Bowen, J. L., Kearns, P. J., Byrnes, J. E., Wigginton, S., Allen, W. J., Greenwood, M., ... & Meyerson, L. A. (2017). Lineage overwhelms environmental conditions in determining rhizosphere bacterial community structure in a cosmopolitan invasive plant. Nature communications, 8(1), 433.
26. Bowker, M. A., Maestre, F. T., & Escolar, C. (2010). Biological crusts as a model system for examining the biodiversity–ecosystem function relationship in soils. Soil Biology and Biochemistry, 42(3), 405-417.
27. Brahim, N., Blavet, D., Gallali, T., & Bernoux, M. (2011). Application of structural equation modeling for assessing relationships between organic carbon and soil properties in semiarid Mediterranean region. International Journal of Environmental Science & Technology, 8(2), 305-320.
28. Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. Psychological methods, 18(1), 71.
29. Brown, D. G., & Weis, A. E. (1995). Direct and indirect effects of prior grazing of goldenrod upon the performance of a leaf beetle. Ecology, 76(2), 426-436.
30. Brown, T. A. (2006). Confirmatory factor analysis for applied research. Guilford Publications.
31. Browne, M. W.(1984). Asymptotically distribution-free methods for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 37(1), 62-83.
32. Bulut, O., Delen, E., & Kaya, F. (2012). An SEM Model Based on PISA 2009 in Turkey: How Does the Use of Technology and Self-regulation Activities Predict Reading Scores?. Procedia-Social and Behavioral Sciences, 64, 564-573.
33. Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of statistical software, 45(3).
34. Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2015). Package 'mice'. Computer software]. Retrieved from: http://cran. r-project. org/web/packages/mice/mice. pdf.
35. Byrne, B. M.(2012).Structural Equation Modeling with Mplus Basic Concepts, Applications, and Programming. Taylor & Francis Ltd
36. Byun, C., de Blois, S., & Brisson, J. (2015). Interactions between abiotic constraint, propagule pressure, and biotic resistance regulate plant invasion. Oecologia, 178(1), 285-296.
37. Cai, X., Bazerque, J. A., & Giannakis, G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. PLoS computational biology, 9(5), e1003068.

38. Canty, A. & Ripley, B. (2017). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-20.
39. Cao, X., Wang, J., Liao, J., Gao, Z., Jiang, D., Sun, J., ... & Luan, S. (2017). Bacterioplankton community responses to key environmental variables in plateau freshwater lake ecosystems: A structural equation modeling and change point analysis. Science of the Total Environment, 580, 457-467.
40. Capmourteres, V., & Anand, M. (2016). Assessing ecological integrity: A multi-scale structural and functional approach using Structural Equation Modeling. Ecological indicators, 71, 258-269.
41. Caro, D. H., Sandoval-Hernández, A., & Lüdtke, O. (2014). Cultural, social, and economic capital constructs in international assessments: An evaluation using exploratory structural equation modeling. School Effectiveness and School Improvement, 25(3), 433-450.
42. Cerda, R., Avelino, J., Gary, C., Tixier, P., Lechevallier, E., & Allinne, C. (2017). Primary and secondary yield losses caused by pests and diseases: Assessment and modeling in coffee. PloS one, 12(1), e0169133.
43. Chang WY (1981) Path analysis and factors affecting primary productivity. J Freshwater Ecol 1(1):113–120
44. Chavance, M., Escolano, S., Romon, M., Basdevant, A., de Lauzon-Guillain, B., & Charles, M. A. (2010). Latent variables and structural equation models for longitudinal relationships: an illustration in nutritional epidemiology. BMC medical research methodology, 10(1), 37.
45. Chen, C. and Zhang, D. (2016) BigSEM: Constructing Large Systems of Structural Equations. R package version 0.2
46. Chen, H., & Little, R. (1999). A test of missing completely at random for generalised estimating equations with missing data. Biometrika, 86 (1), 1.
47. Cheung, C. L., Tan, K. C., Lam, K. S., & Cheung, B. M. (2013). The relationship between glucose metabolism, metabolic syndrome, and bone-specific alkaline phosphatase: a structural equation modeling approach. The Journal of Clinical Endocrinology & Metabolism, 98(9), 3856-3863.
48. Cheung, M. W. L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. Frontiers in Psychology, 5, 1521.
49. Cheung, M. W. L., & Chan, W. (2005). Meta-analytic structural equation modeling: a two-stage approach. Psychological methods, 10(1), 40.
50. Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a Monte Carlo study. British Journal of Mathematical and Statistical Psychology, 44(2), 347-357.
51. Christ, S. L., Lee, D. J., Lam, B. L., & Zheng, D. D. (2014). Structural equation modeling: A framework for ocular and other medical sciences research. Ophthalmic epidemiology, 21(1), 1-13.
52. Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. American Journal of Sociology, 94, S95–S120. doi: 10.1086/228943.
53. Comber, A., Li, T., Lü, Y., Fu, B., & Harris, P. (2017). Geographically Weighted Structural Equation Models: spatial variation in the drivers of environmental restoration effectiveness. In Societal Geo-Innovation. 20th AGILE Conference Proceedings.
54. Cox, D. R. and Small, N. J. H. (1978). Testing multivariate normality. Biometrika 65, 263{272.

55. Crittenden, S. J., & de Goede, R. G. M. (2016). Integrating soil physical and biological properties in contrasting tillage systems in organic and conventional farming. European Journal of Soil Biology, 77, 26-33.
56. Crouch NMA, Mason-Gamer RJ (2018) Structural equation modeling as a tool to investigate correlates of extra-pair paternity in birds. PLoS ONE 13(2): e0193365. https://doi.org/10.1371/journal.pone.0193365
57. Cubaynes, S., Doutrelant, C., Grégoire, A., Perret, P., Faivre, B., & Gimenez, O. (2012). Testing hypotheses in evolutionary ecology with imperfect detection: capture–recapture structural equation modeling. Ecology, 93(2), 248-255.
58. De los Campos, G., Gianola, D., & Heringstad, B. (2006). A structural equation model for describing relationships between somatic cell score and milk yield in first-lactation dairy cows. Journal of dairy science, 89(11), 4445-4455.
59. Delgado-Baquerizo, M., Maestre, F. T., Reich, P. B., Jeffries, T. C., Gaitan, J. J., Encinar, D., ... & Singh, B. K. (2016). Microbial diversity drives multifunctionality in terrestrial ecosystems. Nature communications, 7, 10541.
60. Detilleux, J., Theron, L., Duprez, J. N., Reding, E., Humblet, M. F., Planchon, V., ... & Hanzen, C. (2013). Structural equation models to estimate risk of infection and tolerance to bovine mastitis. Genetics Selection Evolution, 45(1), 6.
61. Dorresteijn, I., Schultner, J., Nimmo, D. G., Fischer, J., Hanspach, J., Kuemmerle, T., ... & Ritchie, E. G. (2015). Incorporating anthropogenic effects into trophic ecology: predator–prey interactions in a human-dominated landscape. Proc. R. Soc. B, 282 (1814), 20151602.
62. Driver, C. C., Oud, J. H., & Voelkle, M. C. (2017). Continuous time structural equation modeling with R pacote ctsem. Journal of Statistical Software, 77(5).
63. Efron, Bradley, Tibshirani, R.J. (1993). An Introduction to the Bootstrap Chapman and Hall/CRC
64. Eisenhauer, N., Bowker, M. A., Grace, J. B., & Powell, J. R. (2015). From patterns to causal understanding: structural equation modeling (SEM) in soil ecology. Pedobiologia, 58(2), 65-72.
65. Enders, C. K. (2010). Applied missing data analysis. Guilford press.
66. Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Structural equation modeling, 8(3), 430-457.
67. Epskamp, Sacha (2017). semPlot: Path Diagrams and Visual Analysis of Various SEM Packages' Output. An R package version 1.1
68. Erickson, K. I., Ho, M. H. R., Colcombe, S. J., & Kramer, A. F. (2005). A structural equation modeling analysis of attentional control: an event-related fMRI study. Cognitive Brain Research, 22(3), 349-357.
69. Fan, Y., Chen, J., Shirkey, G., John, R., Wu, S. R., Park, H., & Shao, C. (2016). Applications of structural equation modeling (SEM) in ecological studies: an updated review. Ecological Processes, 5(1), 19.
70. Fox, J. (2016). polycor: Polychoric and Polyserial Correlations. R package version 0.7-9.
71. Fox, J. and Weisberg, S. (2012, last revision). Structural Equation Modeling in R with the sem Package. An Appendix to An R Companion to Applied Regression, Second Edition.
72. Fox, J., Nie, Z., Byrnes, J., Culbertson, M., DebRoy, S., Friendly, M., ... & Fox, M. J. (2017). Package 'sem'.

73. Freedman, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors". The American Statistician, 60(4), 299-302.
74. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D.B., Vehtari, A., & Rubin, D. B. (2014). Bayesian data analysis. Third EduitionChapman and Hall/CRC.
75. Gelman, A., Hill, J., Su, Y. S., Yajima, M., Pittau, M., Goodrich, B., ... & Goodrich, M. B. (2015). Package 'mi'.
76. Gimenez, O., Anker-Nilssen, T., & Grosbois, V. (2012). Exploring causal pathways in demographic parameter variation: path analysis of mark–recapture data. Methods in Ecology and Evolution, 3(2), 427-432.
77. Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance.
78. Grace, J. B. (2006). Structural Equation Modeling and Natural Systems. Cambridge University Press, Cambridge, UK.
79. Grace, J. B. (2008). Structural Equation Modeling for Observational Studies. The Journal of Wildlife Management, 14-22.
80. Grace, J. B., & Pugesek, B. H. (1997). A structural equation model of plant species richness and its application to a coastal wetland. The American Naturalist, 149(3), 436-460.
81. Grace, J. B., Anderson, T. M., Olff, H., & Scheiner, S. M. (2010). On the specification of structural equation models for ecological systems. Ecological Monographs, 80(1), 67-87.
82. Grace, J. B., Anderson, T. M., Seabloom, E. W., Borer, E. T., Adler, P. B., Harpole, W. S., ... & Bakker, J. D. (2016). Integrative modelling reveals mechanisms linking productivity and plant species richness. Nature, 529 (7586), 390.
83. Grace, J. B., Harrison, S., & Damschen, E. I. (2011). Local richness along gradients in the Siskiyou herb flora: RH Whittaker revisited. Ecology, 92(1), 108-120.
84. Grace, J. B., Schoolmaster, D. R., Guntenspergen, G. R., Little, A. M., Mitchell, B. R., Miller, K. M., & Schweiger, E. W. (2012). Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere, 3(8), 1-44
85. Graham, J. W. (2012). Analysis of missing data. In Missing data(pp. 47-69). Springer, New York, NY.
86. Graham, J. W., & Coffman, D. L. (2012). Structural equation modeling with missing data. In R. H. Hoyle (Ed.), Handbook of structural equation modeling (pp. 277-295). New York, NY, US: Guilford Press.
87. Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How Many Imputations are Really Needed? Some Practical Clarifi cations of Multiple Imputation Theory. Prevention Science, 8, 206–213.
88. Green, D. M., Cox, C. L., Zhu, L., Krull, K. R., Srivastava, D. K., Stovall, M., ... & Meacham, L. R. (2012). Risk factors for obesity in adult survivors of childhood cancer: a report from the Childhood Cancer Survivor Study. Journal of Clinical Oncology, 30(3), 246.
89. Gross, K., Bates, D.. (2018). mvnmle: ML Estimation for Multivariate Normal Data with Missing Values. R package version 0.1-11.1
90. Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., ... & Harden, K. P. (2018). Genomic SEM Provides Insights into the Multivariate Genetic Architecture of Complex Traits. bioRxiv, 305029. Package dispoível em https://github.com/MichelNivard/GenomicSEM/wiki
91. Grund, S., Robitzsch, A., Luedtke, O. (2018). mitml: Tools for Multiple Imputation in Multilevel Modeling. An RE package. Version 0.3-6

92. Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. Econometrica, Journal of the Econometric Society, 1-12.
93. Haber, G., Ahmed, N. U., & Pekovic, V. (2012). Family history of cancer and its association with breast cancer risk perception and repeat mammography. American journal of public health,102(12), 2322-2329.
94. Hair, J. F. Jr, Back, W.C, Babin, B.J., Anderson, R. E. (2010). Multivariate data analysis. Pearson. 7th Edition.
95. Hair, J.F., Hult, G.T.M., Ringle, C.M. and Sarstedt, M. (2014), A Primer on Partial Least Squares Structural Equation Modeling, 2nd Ed. Sage, Thousand Oaks, CA.
96. Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in M plus. Structural Equation Modeling: A Multidisciplinary Journal, 1-18
97. Hannula, M. S., Bofah, E., Tuohilampi, L., & Metsämuuronen, J. (2014). A longitudinal analysis of the relationship between mathematics-related affect and achievement in Finland. In Proceedings of the Joint Meeting of PME (Vol. 38, pp. 249-256).
98. Harrell Jr, F. E., & Harrell Jr, M. F. E. (2018). Package 'Hmisc'. R Foundation for Statistical Computing.
99. Hays, R. D., Revicki, D., & Coyne, K. S. (2005). Application of structural equation modeling to health outcomes research. Evaluation & the Health Professions, 28(3), 295-309.
100. He, T. (2013). Structural equation modelling analysis of evolutionary and ecological patterns in Australian Banksia. Population ecology, 55(3), 461-467.
101. Hershberger, S. L. (2003). The growth of structural equation modeling: 1994-2001. Structural Equation Modeling, 10(1), 35-46.
102. Hill, M. M. e A. Hill (2009). Investigação por questionário. 2ª Ed.Lisboa, Edicões SÍLABO,
103. Hirose, K., Kim, S., Kano, Y., Imada, M., Yoshida, M., & Matsuo, M. (2013). Full information maximum likelihood estimation in factor analysis with a lot of missing values. arXiv preprint arXiv:1312.5458.
104. Hodapp, D., Meier, S., Muijsers, F., Badewien, T. H., & Hillebrand, H. (2015). Structural equation modeling approach to the diversity-productivity relationship of Wadden Sea phytoplankton. Marine Ecology Progress Series, 523, 31-40.
105. Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. Journal of statistical software, 45(7), 1-47.
106. Hoyle, R. H. (Ed.). (2012). Handbook of structural equation modeling. Guilford Press. www. handbookofsem.com
107. Hu, L. T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted?. Psychological bulletin, 112(2), 351.
108. Huang, Po-Hsien (2017). lsl: Latent Structure Learning. R package version 0.5.6
109. Huisman, S. M., Mahfouz, A., Batmanghelich, N. K., Lelieveldt, B. P., & Reinders, M. J. (2018). A structural equation model for imaging genetics using spatial transcriptomics. bioRxiv, 253443.
110. Ibrahim, H., Hatira, A., & Gallali, T. (2013). Relationship between nitrogen and soil properties: Using multiple linear regressions and structural equation modeling. Int. J. Res. Appl. Sci, 2, 1-7.
111. Inman, C. S., James, G. A., Hamann, S., Rajendra, J. K., Pagnoni, G., and Butler, A. J. (2012). Altered resting-state effective connectivity of fronto-parietal motor control systems on the

primary motor network following stroke. Neuroimage 59, 227–237. doi: 10.1016/j.neuroimage.2011.07.083
112. Iriondo, J. M., Albert, M. J., & Escudero, A. (2003). Structural equation modelling: an alternative for assessing causal relationships in threatened plant populations. Biological Conservation, 113(3), 367-377.
113. Jacobucci, R. (2017). regsem: Regularized structural equation modeling. arXiv preprint arXiv:1703.08489.
114. Jamshidian, M., & Jalal, S. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. Psychometrika, 75(4), 649-674.
115. Jamshidian, M., Jalal, S. J., & Jansen, C. (2014). MissMech: an R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR).
116. Jarek, S., & Jarek, M. S. (2009). Package 'mvnormtest'.
117. Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). STEMM: A general finite mixture structural equation model. Journal of Classification, 14(1), 23-50.
118. Jiao, L., Shen, L., Shuai, C., & He, B. (2016). A novel approach for assessing the performance of sustainable urbanization based on structural equation modeling: a China case study. Sustainability, 8(9), 910.
119. Jing, X., Sanders, N. J., Shi, Y., Chu, H., Classen, A. T., Zhao, K., ... & He, J. S. (2015). The links between ecosystem multifunctionality and above-and belowground biodiversity are mediated by climate. Nature Communications, 6, ncomms9159.
120. Jöreskog, K. G. (1973). "A general method for estimating a linear structural equation system.". In Structural equation models in the social sciences Edited by: Goldberger, A. S. and Duncan, O. D. 85–112. New York, NY: Academic..
121. Jöreskog, KG.; Sörbom, D. (2009).LISREL. Scientific Software International; Chicago.
122. Jorgensen, T. D., Pornprasertmanit, S., Miller, P., Schoemann, A., Quick, C. (2018). simsem: SIMulated Structural Equation Modeling. R package version 0.5-14
123. Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. (2018).semTools: Useful Tools for Structural Equation Modeling. R package version 0.5-0
124. Joseph, M. B., Preston, D. L., & Johnson, P. T. (2016). Integrating occupancy models and structural equation models to understand species occurrence. Ecology, 97(3), 765-775.
125. Kaplan, D. (2009). Structural equation modeling: Foundations and extensions. Thousand Oaks, CA: Sage Publications, Inc..
126. Karimi, L., & Meyer, D. (2014). Structural equation modeling in psychology: the history, development and current challenges. International Journal of Psychological Studies, 6(4), 123.
127. Keesling, J. W. (1972). Maximum likelihood approaches to causal analysis. Ph. D. dissertation. Department of Education: University of Chicago.
128. Kelava, A., Nagengast, B., & Brandt, H. (2014). A nonlinear structural equation mixture modeling approach for non-normally distributed latent predictor variables. Structural Equation Modeling, 21, 468-481. doi:http://dx.doi.org/10.1080/10705511.2014.915379
129. Khine, M. S. (2013). Structural Equation Modeling Approaches in Educational Research and Practice. In Application of structural equation modeling in educational research and practice (pp. 279-283). Sense Publishers, Rotterdam.
130. Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A. L., de Mooij, S. M., Moutoussis, M., ... & Lindenberger, U. (2017a). Developmental cognitive neuroscience using Latent Change Score models: A tutorial and applications. Developmental cognitive neuroscience.

131. Kievit, R. A., Davis, S. W., Mitchell, D. J., Taylor, J. R., Duncan, J., Tyler, L. K., ... & Dalgleish, T. (2014). Distinct aspects of frontal lobe structure mediate age-related differences in fluid intelligence and multitasking. Nature communications, 5, 5658.
132. Kievit, R. A., Van Rooijen, H., Wicherts, J. M., Waldorp, L. J., Kan, K. J., Scholte, H. S., & Borsboom, D. (2012). Intelligence and the brain: A model-based approach. Cognitive neuroscience, 3(2), 89-97.
133. Kim J, Namkung J, Lee S, Park T(2010). Application of structural equation models to genome-wide association analysis. Genomics & Informatics, 8:150–158.
134. Kim, J., & Horwitz, B. (2009). How well does structural equation modeling reveal abnormal brain anatomical connections? An fMRI simulation study. Neuroimage, 45(4), 1190-1198.
135. Kim, K., & Bentler, P. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. Psychometrika, 67 (4), 609–623.
136. Kjaerulff, U. B. and Madsen, A. L. (2008). Bayesian networks and influence diagrams. Springer Science+ Business Media, 200:114.
137. Klaus K. H., Brice, Ozenne, Gerds, T.. (2018). Lava: Latent Variable Models . R package version 1.6.2
138. Klein, A. & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. Psychometrika, 65 (4), 457-474.
139. Klein, A. G., & Muthén, B. O. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. Multivariate Behavioral Research, 42(4), 647-673.
140. Kline, R. B. (2011). Principles and practice of structural equation modeling. New York, NY: Guilford.
141. Kok, B. , Pek, J. , Sterba, S. , Bauer, D. Chalmers, P. (2017). plotSEMM: Graphing Nonlinear Relations Among Latent Variables from Structural Equation Mixture Models. R package version 2.4
142. Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. The R Journal, 6(2), 151-162.
143. Laliberté, E., Zemunik, G., & Turner, B. L. (2014). Environmental filtering explains variation in plant diversity along resource gradients. Science, 345(6204), 1602-1605.
144. Lam, T. Y., & Maguire, D. A. (2012). Structural equation modeling: theory and applications in forest management. International Journal of Forestry Research, 2012.
145. Lamb, E. G., Mengersen, K., Stewart, K. J., Attanayake U., and Siciliano, S. D. (2016) sesem: Spatially Explicit Structural Equation Modeling. R package version 1.0.2
146. Lamb, E. G., Shirtliffe, S. J., & May, W. E. (2011). Structural equation modeling in the plant sciences: An example using yield components in oat. Canadian Journal of Plant Science, 91(4), 603-619.
147. Lamb, E.G. & Cahill, J.F. (2008) When competition does not matter: grassland diversity and community composition. American Naturalist, 171, 777–787.
148. Lee, E. H., Lee, Y. W., & Moon, S. H. (2016). A structural equation model linking health literacy to self-efficacy, self-care activities, and health-related quality of life in patients with type 2 diabetes. Asian nursing research, 10(1), 82-87.
149. Lee, G. S., & Yom, Y. H. (2013). Structural equation modeling on life-world integration in people with severe burns. Asian nursing research, 7(3), 112-119.

150. Lee, J. W., Lee, K. E., Park, D. J., Kim, S. H., Nah, S. S., Lee, J. H., ... & Lee, H. S. (2017). Determinants of quality of life in patients with fibromyalgia: A structural equation modeling approach. PloS one, 12(2), e0171186.
151. Lee, S. Y. (2007a). Structural equation modeling: A Bayesian approach (Vol. 711). John Wiley & Sons.
152. Lee, S. Y. (2007b). Handbook of latent variable and related models (Vol. 1). Elsevier.
153. Lee, S. Y., & Song, X. Y. (2003). Bayesian analysis of structural equation models with dichotomous variables. Statistics in medicine, 22(19), 3073-3088.
154. Lee, S. Y., & Song, X. Y. (2014). Bayesian structural equation model. Wiley
155. Lefcheck, J. S. (2016). piecewiseSEM: piecewise structural equation modelling in R for ecology, evolution, and systematics. Methods in Ecology and Evolution, 7(5), 573-579. doi:10.1111/2041-210X.12512
156. Levy, R. (2010). SEMModComp: An R Package for Calculating Likelihood Ratio Tests for Mean and Covariance Structure Models. Applied Psychological Measurement, 34(5), 370-371.
157. Li, C. (2013). Little's test of missing completely at random. The Stata Journal, 13(4), 795-809.
158. Li, R., Tsaih, S. W., Shockley, K., Stylianou, I. M., Wergedal, J., Paigen, B., & Churchill, G. A. (2006). Structural model analysis of multiple quantitative traits. PLoS genetics, 2(7), e114.
159. Li, T., Lü, Y., Fu, B., Comber, A. J., Harris, P., & Wu, L. (2017). Gauging policy-driven large-scale vegetation restoration programmes under a changing environment: Their effectiveness and socio-economic relationships. Science of the Total Environment, 607, 911-919.
160. Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), 1198-1202.
161. Little, R. J. A., Rubin, D. B. (2002).Statistical analysis with missing data, 2. ed. Hoboken: John Wiley & Sons,.
162. Little, T. D. (2013). Longitudinal structural equation modeling. Guilford press.
163. Liu, B., de La Fuente, A., & Hoeschele, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. Genetics, 178(3), 1763-1776.
164. Lumley, T., Lumley, M. T., & RODBC, S. (2015). Package 'mitools'.
165. Ma, G. X., Fang, C., Wang, M. Q., Shive, S. E., & Ma, X. S. (2013). Pathways of breast cancer screening among Chinese American women. Journal of community medicine & health education, 3(209).
166. Ma, J., Wheeler, N., Xu, Y., Du, W., Gok, A., Sun, J. (2016). gSEM: Semi-Supervised Generalized Structural Equation Modeling. R package version 0.4.3.4
167. MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. Annual review of psychology, 51(1), 201-226.
168. MacKenzie, D., J. Nichols, and G. Lachman. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255.
169. Maddox GD, Antonovics J (1983) Experimental ecological genetics in Plantago: a structural equation approach to fitness components in P. aristata and P. patagonica. Ecology 64(5):1092–1099
170. Magrini, Alessandro. (2018). dlsem: Distributed-Lag Linear Structural Equation Models. R package version 2.3
171. Mair, P., Wu, E., & Wu, M. E. (2011). R Package 'REQS'.
172. Mańkowski, D. R., Kozdój, J., & Janaszek-Mańkowska, M. (2016). Structural Equation Model as a Tool to Assess the Relationship Between Grain Yield Per Plant and Yield Components in

Doubled Haploid Spring Barley Lines (Hordeum vulgare L.). Plant Breeding and Seed Science, 73(1), 63-77.
173. Marcoulides, G.A.e Schumacker, R.E..(2009). New Developments and Techniques an Structural Equation Modeling. Psychology Press. USA
174. Mardani, A., Streimikiene, D., Zavadskas, E. K., Cavalaro, F., Nilashi, M., Jusoh, A., & Zare, H. (2017). Application of Structural Equation Modeling (SEM) to Solve Environmental Sustainability Problems: A Comprehensive Review and Meta-Analysis. Sustainability, 9(10), 1814.
175. Mardia, K.V. (1985). "Mardia's Test of Multinormality," in S. Kotz and N.L. Johnson, eds., Encyclopedia of Statistical Sciences, vol. 5 (NY: Wiley), pp. 217-221.
176. Marôco, J. (2014). Análise de equações estruturais: Fundamentos teóricos, software & aplicações. ReportNumber, Lda.
177. Martens, M. P., & Haase, R. F. (2006). Advanced applications of structural equation modeling in counseling psychology research. The Counseling Psychologist, 34(6), 878-911.
178. Medland, S. E., & Neale, M. C. (2010). An integrated phenomic approach to multivariate allelic association. European Journal of Human Genetics, 18(2), 233.
179. Mehta, P. D. (2013). n-level structural equation modeling. In Y. Petscher, C. Schatschneider & D. L. Compton (Eds.), Applied quantitative analysis in the social sciences (pp. 329-362). New York: Routledge. http://xxm.times.uh.edu/
180. Merchant, W. R., Li, J., Karpinski, A. C., & Rumrill Jr, P. D. (2013). A conceptual overview of structural equation modeling (SEM) in rehabilitation research. Work, 45(3), 407-415.
181. Merkle, E. C., & Rosseel, Y. (2015). blavaan: Bayesian structural equation models via parameter expansion. arXiv preprint arXiv:1511.05604.
182. Merkle, E. C., & Wang, T. (2016). Bayesian latent variable models for the analysis of experimental psychology data. Psychonomic bulletin & review, 1-15.
183. Mi, X., Eskridge, K. M., George, V., & Wang, D. (2011). Structural equation modeling of gene–environment interactions in coronary heart disease. Annals of human genetics, 75(2), 255-265.
184. Monecke, A., & Leisch, F. (2012). semPLS: structural equation modeling using partial least squares.
185. Moore, Gordon. (1965) "Cramming More Components onto Integrated Circuits," Electronics Magazine Vol. 38, No. 8 (April 19, 1965).
186. Mora, F. (2017). A structural equation modeling approach for formalizing and evaluating ecological integrity in terrestrial ecosystems. Ecological Informatics, 41, 74-90.
187. Moreira, P. S., Sotiropoulos, I., Silva, J., Takashima, A., Sousa, N., Leite-Almeida, H., & Costa, P. S. (2016). The Advantages of Structural Equation Modeling to Address the Complexity of Spatial Reference Learning. Frontiers in behavioral neuroscience, 10, 18.
188. Morris, N. J., Elston, R. C., & Stein, C. M. (2010). A framework for structural equation models in general pedigrees. Human heredity, 70(4), 278-286.
189. Morrison, J. A. (2017). Effects of white-tailed deer and invasive plants on the herb layer of suburban forests. AoB Plants, 9(6), plx058.
190. Morrison, T. G., Morrison, M. A., & McCutcheon, J. M. (2017). Best Practice Recommendations for Using Structural Equation Modelling in Psychological Research. Psychology, 8(09), 1326.

191. Mortensen, L. O., Schmidt, N. M., Høye, T. T., Damgaard, C., & Forchhammer, M. C. (2016). Analysis of trophic interactions reveals highly plastic response to climate change in a tri-trophic High-Arctic ecosystem. Polar Biology, 39(8), 1467-1478.
192. Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. Best practices in quantitative methods, 488508.
193. Mulaik, S. A. (2009). Linear causal modeling with structural equations. Chapman and Hall/CRC.
194. Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. Psychological Methods, 17, 313–335.
195. Muthén, B., & Muthén, B. O. (2009). Mplus:Statistical analysis with latent variables. New York, NY: Wiley.
196. Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. Structural equation modeling, 9(4), 599-620.
197. Nazmi, L. (2013). Modeling for relationships between soil properties and yield components of wheat using multiple linear regression and structural equation modeling. Biol, 7(2), 235-242.
198. Neale, M. C., Hunter, M. D., Pritkin, J., Zahery, M., Brick, T. R., Kirkpatrick, R. M., ... & Boker, S. M. (2016). OpenMx 2.0: Extended Structural Equation and Statistical Modeling. Psychometrika, 81(2), 535. <doi:10.1007/s11336-014-9435-8>
199. Neale, M., & Cardon, L. (1992). Methodology for Genetic Studies of Twins and Families (Vol. 67). Springer Science & Business Media.
200. Nielsen, S., & Wilms, L. I. (2015). Cognitive aging on latent constructs for visual processing capacity: a novel structural equation modeling framework with causal assumptions based on a theory of visual attention. Frontiers in psychology, 5, 1596.
201. Nock, N. L., Li, L., & Elston, R. C. (2009, June). Modeling genetic and environmental factors in biological systems using structural equation modeling: an application to energy balance. In Bioinformatics, 2009. OCCBIO'09. Ohio Collaborative Conference on (pp. 3-8). IEEE.
202. Nuzhdin, S. V., Friesen, M. L., & McIntyre, L. M. (2012). Genotype–phenotype mapping in a post-GWAS world. Trends in Genetics, 28(9), 421-426.
203. Oberski, D. (2014). lavaan. survey: An R package for complex survey analysis of structural equation models. Journal of Statistical Software, 57(1), 1-27. URL http://www.jstatsoft.org/v57/i01/.
204. Pastore, M. & Altoe',G (2018).influence.SEM: Case Influence in Structural Equation Models. R package version 2.2
205. Pearl, J. 2012. The causal foundations of structural equation modeling (Vol. 370). CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE.
206. Pearson, K. (1901). Mathematical contributions to the theory of evolution. VII: On the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society of London, Series A, 195, 1–47.
207. Peñagaricano, F., Valente, B. D., Steibel, J. P., Bates, R. O., Ernst, C. W., Khatib, H., & Rosa, G. J. M. (2015). Searching for causal networks involving latent variables in complex traits: Application to growth, carcass, and meat quality traits in pigs. Journal of animal science, 93(10), 4617-4623.
208. Penke, L., Maniega, S. M., Bastin, M. E., Hernández, M. V., Murray, C., Royle, N. A., ... & Deary, I. J. (2012). Brain white matter tract integrity as a neural foundation for general intelligence. Molecular psychiatry, 17(10), 1026.

209. Pepe, D., & Grassi, M. (2014). Investigating perturbed pathway modules from gene expression data via structural equation models. BMC bioinformatics, 15(1), 132.
210. Peugh, J. L., & Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. Review of Educational Research, 74(4), 525-556.
211. Phiakoksong, S., Niwattanakul, S., & Angskun, T. (2013). An application of structural equation modeling for developing good teaching characteristics ontology. Informatics in Education, 12(2), 253-272.
212. Portela, D. M. P. (2012). Contributo das técnicas de análise fatorial para o estudo do programa" ocupação científica de jovens nas férias" (Doctoral dissertation).
213. Posthuma, D., De Geus, E. J. C., Boomsma, D. I., & Neale, M. C. (2004). Combined linkage and association tests in mx. Behavior Genetics, 34(2), 179-196.
214. PROC CALIS. SAS Institute, Inc.; Cary, NC: 2010.
215. Prugh, L. R., & Brashares, J. S. (2012). Partitioning the effects of an ecosystem engineer: kangaroo rats control community structure via multiple pathways. Journal of Animal Ecology, 81(3), 667-678.
216. Pugesek, B. H., & Tomer, A. (1996). The Bumpus house sparrow data: a reanalysis using structural equation models. Evolutionary Ecology, 10(4), 387-404.
217. Pugesek, B. H., Tomer, A., & Von Eye, A. (Eds.). (2003). Structural equation modeling: applications in ecological and evolutionary biology. Cambridge University Press.
218. Quartagno M, Carpenter J (2018). JOMO: A Package For Multilevel Joint Modelling Multiple Imputation. https://CRAN.R-project.org/package=jomo.
219. R Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
220. Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). Multilevel structural equation modeling. Handbook of latent variable and related models, 209-227.
221. Rao, D., Feldman, B. J., Fredericksen, R. J., Crane, P. K., Simoni, J. M., Kitahata, M. M., & Crane, H. M. (2012). A structural equation model of HIV-related stigma, depressive symptoms, and medication adherence. AIDS and Behavior, 16(3), 711-716.
222. Ravens-Sieberer, U., Freeman, J., Kokonyei, G., Thomas, C. A., & Erhart, M. (2009). School as a determinant for health outcomes–a structural equation model analysis. Health Education, 109(4), 342-356.
223. Ray, Soumya & Danks, Nicholas. (2018). SEMinR. https://cran.rstudio.com/web/packages/seminr/vignettes/SEMinR.html
224. Raykov, T., & Marcoulides, G. A. (2006). A first course in structural equation modeling. 2nd Edition.Routledge.
225. Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.4.
226. Ritchie, S. J., Booth, T., Hernández, M. D. C. V., Corley, J., Maniega, S. M., Gow, A. J., ... & Bastin, M. E. (2015). Beyond a bigger brain: Multivariable structural brain imaging and intelligence. Intelligence, 51, 47-56.
227. Roelstraete, B., & Rosseel, Y. (2011). FIAR: an R package for analyzing functional integration in the brain. Journal of Statistical Software, 44(13), 1-32.

228. Rosa, G. J., Felipe, V. P., & Peñagaricano, F. (2016). Applications of graphical models in quantitative genetics and genomics. In Systems Biology in Animal Production and Health, Vol. 1 (pp. 95-116). Springer, Cham.
229. Rosa, G. J., Valente, B. D., de los Campos, G., Wu, X. L., Gianola, D., & Silva, M. A. (2011). Inferring causal phenotype networks using structural equation models. Genetics Selection Evolution, 43(1), 6.
230. Rosseel, Y. (2018). The lavaan tutorial. Department of Data Analysis: Ghent University.
231. Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.
232. Ryberg, K. R. (2017). Structural Equation Model of Total Phosphorus Loads in the Red River of the North Basin, USA and Canada. Journal of environmental quality, 46(5), 1072-1080.
233. Sanchez G, Aluja T (2012). pathmox: Segmentation Trees in Partial Least Squares Path.
234. Sanchez, G. (2013). PLS path modeling with R. Berkeley: Trowchez Editions,
235. Sari, A. A. (2015). Using structural equation modeling to investigate students' reading comprehension abilities. İlköğretim Online, 14(2).
236. Savalei, V. (2010). Expected vs. observed information in SEM with incomplete normal and nonnormal data. Psychological Methods, 15, 352-367.
237. Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological methods, 7(2), 147.
238. Schafer, J.L. (1997), Analysis of Incomplete Multivariate Data, New York: Chapman and Hall.
239. Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. Psychometrika, 64(1), 37-52.
240. Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. Methods of psychological research online, 8(2), 23-74.
241. Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. The Journal of educational research, 99(6), 323-338.
242. Schumacker, R. E., & Lomax, R. G. (2004). A beginner's guide to structural equation modeling (2nd ed.). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
243. Shadfar, S., & Malekmohammadi, I. (2013). Application of Structural Equation Modeling (SEM) in restructuring state intervention strategies toward paddy production development. International Journal of Academic Research in Business and Social Sciences, 3(12), 576.
244. Shao, Y., Bao, W., Chen, D., Eisenhauer, N., Zhang, W., Pang, X., ... & Fu, S. (2015). Using structural equation modeling to test established theory and develop novel hypotheses for the structuring forces in soil food webs. Pedobiologia-Journal of Soil Ecology, 4(58), 137-145.
245. Sharifzadeh, M., Zamani, G. H., Khalili, D., & Karami, E. (2012). Agricultural climate information use: an application of the planned behaviour theory. Journal of Agricultural Science and Technology, 14(3), 479-492.
246. Shipley, B. (1995). Structured interspecific determinants of specific leaf area in 34 species of herbaceous angiosperms. Functional Ecology, 312-319.
247. Shipley, B. (2000a). Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations, and Causal Inference in R.. Cambridge, UK: Cambridge University Press, 317 pages.
248. Shipley, B. 2000b. A new inferential test for path models based on directed acyclic graphs. Structural Equation Modeling 7:206–218.

249. Shipley, B. 2003. Testing recursive path models with correlated errors using d-separation. Structural Equation Modeling 10:214–221.
250. Shipley, B. 2009. Confirmatory path analysis in a generalized multilevel context. Ecology 90:363–368.
251. Sideridis, G., Simos, P., Papanicolaou, A., & Fletcher, J. (2014). Using Structural Equation Modeling to Assess Functional Connectivity in the Brain Power and Sample Size Considerations. Educational and Psychological Measurement, dpi: 10.1177/0013164414525397
252. Skrondal, A.and Rabe-Hesketh., S. (2004). Generalized Latent Variable Modeling: Multilevel,Longitudinal, and Structural Equation Models.Boca Raton, FL: Chapman & Hall/CRC, 2004.
253. Smith, R., Davis, A. S., Jordan, N. R., Atwood, L. W., Daly, A. B., Grandy, A. S., ... & Kane, D. (2014). Structural equation modeling facilitates transdisciplinary research on agriculture and climate change. Crop Science, 54(2), 475-483.
254. Song, X. Y., & Lee, S. Y. (2006). Bayesian analysis of structural equation models with nonlinear covariates and latent variables. Multivariate Behavioral Research, 41(3), 337-365.
255. Song, X. Y., & Lee, S. Y. (2012). Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences. Chichester, England: Wiley.
256. Song, Y. E., Morris, N. J., & Stein, C. M. (2016, October). Structural equation modeling with latent variables for longitudinal blood pressure traits using general pedigrees. In BMC proceedings (Vol. 10, No. 7, p. 55). BioMed Central.
257. Song, Y. E., Stein, C. M., & Morris, N. J. (2015). strum: an R package for stru ctural m odeling of latent variables for general pedigrees. BMC genetics, 16(1), 35.
258. Spearman, C. (1904). "General intelligence", objectively determined and measured. American Journal of Psychology, 15, 201–292.
259. Stas, L., Schönbrodt, F. D., & Loeys, T. (2016). fSRM: Social Relations Analyses with Roles ("Family SRM"). R package version 0.6.4
260. StatSoft, Inc. (2013). (Electronic Version): Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/. http://documentation.statsoft.com/STATISTICAHelp.aspx?path=sepath/Indices/SEPATHAnalysis_HIndex
261. Tallavaara, M., Eronen, J. T., & Luoto, M. (2017). Productivity, biodiversity, and pathogens influence the global hunter-gatherer population density. Proceedings of the National Academy of Sciences, 201715638.
262. Tanaka, N., Huttenhower, C., Nosho, K., Baba, Y., Shima, K., Quackenbush, J., ... & Ogino, S. (2010). Novel application of structural equation modeling to correlation structure analysis of CpG island methylation in colorectal cancer. The American journal of pathology, 177(6), 2731-2740.
263. Tao, Y., Sánchez, B. N., & Mukherjee, B. (2015). Latent variable models for gene–environment interactions in longitudinal studies with multiple correlated exposures. Statistics in medicine, 34(7), 1227-1241.
264. Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N. S., Wijesundera, R., ... & Smith, M. E. (2014). Main content area Global diversity and geography of soil fungi. Science, 346(6213), 1256688-1256688.

265. Teo, T., Tsai, L. T., & Yang, C. C. (2013). Applying Structural Equation Modeling (SEM) in Educational Research. In Application of structural equation modeling in educational research and practice (pp. 3-21). SensePublishers.
266. Thurstone, L. L. (1931). Multiple factor analysis. Psychological Review, XXXVIII, 406–427.
267. Thurstone, L. L. (1935). The Vectors of the Mind. Chicago: Chicago University Press.
268. Tonsor, S. J., & Scheiner, S. M. (2007). Plastic trait integration across a CO2 gradient in Arabidopsis thaliana. The American Naturalist, 169(5), E119-E140.
269. Tortorec, E., Helle, S., Käyhkö, N., Suorsa, P., Huhta, E., & Hakkarainen, H. (2013). Habitat fragmentation and reproductive success: a structural equation modelling approach. Journal of Animal Ecology, 82(5), 1087-1097.
270. Tremblay, P. F., & Gardner, R. C. (1996). On the growth of structural equation modeling in psychological journals. Structural Equation Modeling: A Multidisciplinary Journal, 3(2), 93-104.
271. Trivedi, P., Delgado-Baquerizo, M., Trivedi, C., Hu, H., Anderson, I. C., Jeffries, T. C., ... & Singh, B. K. (2016). Microbial regulation of the soil carbon cycle: evidence from gene–enzyme relationships. The ISME journal, 10(11), 2593.
272. Tsanousa, A., Angelis, L., Ntoufa, S., Papakonstantinou, N., & Stamatopoulos, K. (2013, August). A Structural Equation Modeling Approach of the Toll-Like Receptor Signaling Pathway in Chronic Lymphocytic Leukemia. In Database and Expert Systems Applications (DEXA), 2013 24th International Workshop on (pp. 71-75). IEEE.
273. Ullman, J. B. (2007). Structural Equation Modeling. Em B. G. Tabachnick & L. S. Fidell (Orgs.), Using multivariate statistics (5ª ed.). Boston: Pearson Education.
274. Umbach, N., Naumann, K., Brandt, H., & Kelava, A. (2017). Fitting nonlinear structural equation models in R with package nlsem. Journal of Statistical Software, 77(7), 1-20.
275. Valente, B. D., Rosa, G. J., Gianola, D., Wu, X. L., & Weigel, K. (2013). Is structural equation modeling advantageous for the genetic improvement of multiple traits?. Genetics, 194(3), 561 572.
276. van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. Psychological Methods, 22(2), 217.
277. van den Oord, E. J. (2000). Framework for identifying quantitative trait loci in association studies using structural equation modeling. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 18(4), 341-359.
278. Verhulst, B., Maes, H. H., & Neale, M. C. (2017). GW-SEM: A Statistical Package to Conduct Genome-Wide Structural Equation Modeling. Behavior genetics, 47(3), 345-359.
279. Vilhena, E., Pais-Ribeiro, J., Silva, I., Cardoso, H., & Mendonça, D. (2014). Predictors of quality of life in Portuguese obese patients: a structural equation modeling application. Journal of obesity, 2014.
280. Villarreal Ruiz, L., Vasco-Palacios, A. M., Thu, P. Q., Suija, A., Smith, M. E., Sharp, C., ... & Ratkowsky, D. (2014). Fungal biogeography. Global diversity and geography of soil fungi. Science (New York, NY), 346, 1256688.
281. Vink, G., Frank, L. E., Pannekoek, J., & Van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. Statistica Neerlandica, 68(1), 61-90.
282. Viswanath, N. C., Kumar, P. D., Ammad, K. K., & Kumari, E. U. (2015). Ground water quality and multivariate statistical methods. Environmental Processes, 2(2), 347-360.

283. Wang, J., Wang, X. (2012). Structural equation modeling : applications using Mplus. UK: John Wiley & Sons Ltd
284. Warrington, N., Freathy, R., Neale, M. C., & Evans, D. M. (2017). Using Structural Equation Modeling to Jointly Estimate Maternal and Foetal Effects on Birthweight in the UK Biobank. bioRxiv, 160044.
285. Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger & O. D. Duncan (eds.), Structural Equation Models in the Social Sciences. New York: Seminar Press A.S.
286. Wickham, H., François, R.,. Henry, L., & Müller, K. (2018). R package `dplyr`.
287. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
288. Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. Educational and psychological measurement, 73(6), 913-934.
289. Wootton, J. T. (1992). Indirect effects, prey susceptibility, and habitat selection: impacts of birds on limpets and algae. Ecology, 73(3), 981-991.
290. Wright, S. (1921). Correlation and causation. Journal of Agricultural Research, XX, 557–215.
291. Wright, S. (1934). The method of path coefficients. Annals of Mathematical Statistics, V, 161–215.
292. Xiong, M., J. Li and X. Fang. (2004). Identification of genetic networks. Genetics 166: 1037–1052.
293. Yoo-Jeong, M., Waldrop-Valverde, D., McCoy, K., & Ownby, R. L. (2016). A Structural Equation Model of HIV-related Symptoms, Depressive Symptoms, and Medication Adherence. Journal of HIV and AIDS, 2(3).
294. You, Y., Wang, J., Huang, X., Tang, Z., Liu, S., & Sun, O. J. (2014). Relating microbial community structure to functioning in forest soil organic carbon transformation and turnover. Ecology and evolution, 4(5), 633-647.
295. Youngblood, A., Grace, J. B., & McIver, J. D. (2009). Delayed conifer mortality after fuel reduction treatments: interactive effects of fuel, fire intensity, and bark beetles. Ecological Applications, 19(2), 321-337.
296. Yuan, K. H., & Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. Psychometrika, 77(4), 803-826.
297. Zhang, P. (2003). Multiple imputation: theory and method. International Statistical Review, 71(3), 581-592.
298. Zhang, T., Lamb, E. G., Soto-Cerda, B., Duguid, S., Cloutier, S., Rowland, G. & Booker, H. M. (2014). Structural equation modeling of the Canadian flax (Linum usitatissimum L.) core collection for multiple phenotypic traits. Canadian journal of plant science, 94(8), 1325-1332.
299. Zhang, Y., Dong, S., Gao, Q., Liu, S., Ganjurjav, H., Wang, X., ... & Wu, X. (2017). Soil bacterial and fungal diversity differently correlated with soil biochemistry in alpine grassland ecosystems in response to environmental changes. Scientific Reports, 7.
300. Zhang, Z., Hamagami, F., Grimm, K. J., & McArdle, J. J. (2015). Using R Package RAMpath for tracing SEM path diagrams and conducting complex longitudinal data analysis. Structural Equation Modeling: A Multidisciplinary Journal, 22(1), 132-147.
301. Zhang, Z., Yuan K. (2012).semdiag: Structural equation modeling diagnostics. R package version 0.1.2~
302. Zhao, Jing Hua, Schafer, Joseph L. (2018). pan: Multiple imputation for multivariate panel or clustered dataR package version 1.6.