

Universidade de Brasília  
Instituto de Ciências Exatas  
Departamento de Estatística

Dissertação de Mestrado

Mensuração e regressão  
de variáveis latentes contínuas

por

Heitor Cova Gama

Orientadora: Prof<sup>a</sup>. Dr<sup>a</sup>. Cibele Queiroz da Silva

Julho de 2011

Heitor Cova Gama

**Mensuração e regressão  
de variáveis latentes contínuas**

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

**Universidade de Brasília  
Brasília, Julho de 2011**

*Dedico à Dona Clara,  
que a mim tanto dedicou.*

# Agradecimentos

Agradeço a toda minha família, especialmente meus avós, minha mãe e minha tia, que sempre me deram o apoio necessário para eu ir aonde precisava.

Ao Marcos, que me acolheu em tempos muito difíceis e que sabe como ninguém o quanto custou levar este projeto até o fim.

Ao Felipe, à Graziella e à Ana, pela amizade verdadeira em todos os momentos, bons ou não tão bons.

Ao Marcelo, ao Flávio, ao Narciso, que me proporcionaram momentos de serenidade na turbulência dos últimos meses.

A todos os colegas do Programa de *Trainees* do Sebrae Nacional, que me incentivaram a seguir em frente e me ajudaram a chegar ao fim, especialmente à Loreane, pelos conselhos, à Gabriela, pela compreensão, e ao Rafael, por viajar em meu lugar, para que eu estivesse em Brasília na data de entrega da Dissertação.

Aos colegas da graduação, especialmente ao André e ao Felipe, por serem amigos leais e terem tornado minhas primeiras lições de Estatística muito mais divertidas.

Aos colegas do Mestrado, pelos momentos de descontração e de estudo que compartilhamos.

À Professora Cibele, uma das maiores incentivadoras e apoiadoras do meu Mestrado.

Ao Professor Antonio Eduardo, por todo o suporte dado como Coordenador do Programa de Mestrado.

À Capes, pelo apoio financeiro concedido.

# Sumário

<b>Lista de Figuras</b>	<b>5</b>
<b>Lista de Tabelas</b>	<b>7</b>
<b>Resumo</b>	<b>8</b>
<b>Abstract</b>	<b>9</b>
<b>1 Introdução</b>	<b>10</b>
1.1 Objetivo da dissertação . . . . .	10
1.2 Estrutura da dissertação . . . . .	11
1.3 Terminologia e notação utilizadas . . . . .	11
<b>2 Representação e modelagem de variáveis latentes</b>	<b>16</b>
2.1 A onipresença das variáveis latentes . . . . .	16
2.2 Diagramas de caminho . . . . .	18
2.3 Regressão de variáveis latentes: atenuação . . . . .	21
2.4 Modelos de mensuração, de regressão e híbridos . . . . .	23
2.5 Atributos dos modelos em estudo . . . . .	24
2.5.1 Modelos de Regressão . . . . .	24
2.5.2 Modelos de Mensuração . . . . .	26
2.5.3 Modelos de Mensuração e Regressão ( <i>híbridos</i> ) . . . . .	27
<b>3 Mensuração de variáveis latentes contínuas</b>	<b>28</b>
3.1 Modelos de mensuração: definições preliminares . . . . .	28
3.2 Análise Fatorial (1904) . . . . .	30
3.2.1 Identificabilidade . . . . .	35

3.2.2	Interpretabilidade: rotações . . . . .	37
3.2.3	Componentes de variância . . . . .	38
3.2.4	Estimação dos parâmetros do modelo de Análise Fatorial . . . . .	40
3.2.5	Estimação dos escores dos fatores . . . . .	42
3.2.6	As duas modalidades de Análise Fatorial . . . . .	43
3.2.7	<i>Softwares</i> e exemplo de resolução . . . . .	44
3.2.8	Limitações . . . . .	47
3.3	Teoria de Resposta ao Item (1960) . . . . .	49
3.3.1	Modelos cumulativos unidimensionais . . . . .	52
3.3.2	Estimação dos parâmetros estruturais e incidentais . . . . .	53
3.3.3	<i>Softwares</i> e exemplo de resolução . . . . .	55
3.4	A Teoria de Resposta ao Item como um tipo de Análise Fatorial (1983) . . . . .	58
3.4.1	Duas abordagens para um mesmo problema . . . . .	58
3.4.2	Análise Fatorial ou TRI: uma questão de parametrização . . . . .	61
3.4.3	Semelhanças e dessemelhanças . . . . .	64
3.5	Apêndice . . . . .	65
3.5.1	Linhas de código e <i>output</i> do R: ajuste do modelo de Análise Fatorial aos dados da Seção 3.2 . . . . .	65
3.5.2	Linhas de código e <i>output</i> do R: ajuste do modelo logístico de 2 parâmetros, da Teoria de Resposta ao Item, aos dados da Seção 3.3 . . . . .	70
<b>4</b>	<b>Regressão de variáveis latentes contínuas</b> . . . . .	<b>72</b>
4.1	Modelos de Equações Estruturais (1973) . . . . .	73
4.1.1	Modelo de Equações Estruturais com Variáveis Observadas . . . . .	74
4.1.2	Análise Fatorial Confirmatória . . . . .	78
4.1.3	LISREL: Modelo de Equações Estruturais com Variáveis Latentes . . . . .	86
4.1.4	Estimação dos parâmetros estruturais e incidentais . . . . .	93
4.1.5	Limitações . . . . .	96
4.2	Regressão clássica entre escores de fatores (2001) . . . . .	97
4.2.1	Modelo de Regressão Latente . . . . .	98

4.2.2	Limitações . . . . .	102
4.3	Limitações das abordagens clássicas para a regressão de variáveis latentes contínuas . . . . .	102
4.4	Modelos Lineares Generalizados Mistos com Variáveis Latentes (1999)	103
4.4.1	Definição dos componentes do modelo . . . . .	104
4.4.2	Considerações sobre o modelo . . . . .	106
4.5	Apêndice . . . . .	108
4.5.1	Linhas de código e <i>output</i> do R: ajuste do modelo de Análise Fatorial Confirmatória aos dados da Seção 4.1.2 . . . . .	108
4.5.2	Linhas de código e <i>output</i> do R: ajuste do modelo de Equações Estruturais aos dados da Seção 4.1.3 . . . . .	109
<b>5</b>	<b>Visão geral dos modelos, recomendações práticas e trabalhos futuros</b>	<b>113</b>
5.1	Visão geral dos modelos apresentados . . . . .	113
5.1.1	Modelos de mensuração . . . . .	113
5.1.2	Modelos de regressão para variáveis latentes . . . . .	114
5.2	Recomendações para a mensuração de variáveis latentes . . . . .	118
5.3	Trabalhos futuros . . . . .	119

# Lista de Figuras

2.1	Diagrama de caminho de um modelo de mensuração para a variável latente $\xi_1$ . As variáveis latentes são representadas por círculos e as observáveis, representadas por quadrados. Setas indicam a relação linear de causalidade entre as variáveis e o coeficiente dessa relação é indicado sobre a seta. As pequenas setas representam erros de mensuração. . . . .	19
2.2	Diagrama de caminho da <i>equação estrutural</i> representada pela regressão entre as variáveis latentes $\eta_1$ e $\xi_1$ , cada uma aferida por seu <i>modelo de mensuração</i> específico. . . . .	20
2.3	Diagramas de caminho do modelo verdadeiro, em que os erros de mensuração estão presentes, e do modelo ajustado, em que são negligenciados. . . . .	22
2.4	Quadro de comparação entre alguns modelos de mensuração, de regressão e de híbridos. O valor da “complexidade” equivale à quantidade compreendida por cada modelo, dos atributos apresentados no quadro. . . . .	25
3.1	Modelo de Análise Fatorial com complexidade 1: seu diagrama de caminho representa o único fator latente ( $\xi_1$ ) comum a todas as seis variáveis observadas ( $\mathbf{x}$ ). . . . .	32
3.2	Modelo de Análise Fatorial com 2 fatores não-ortogonais, referente a notas obtidas em 6 avaliações aplicadas em 112 alunos (Bartholomew & Knott, 1990). . . . .	48
3.3	Ajuste de um modelo logístico de 2 parâmetros (ML2) para o conjunto de dados referente à atitude frente à legalização do aborto (McGrath & Waterton, 1986): Curvas Características dos Itens (CCI). . . . .	56

4.1	Matriz de covariâncias implicada pelo modelo de Análise Fatorial Confirmatória relativo às notas obtidas em uma bateria de seis avaliações educacionais aplicada a 112 alunos (Bartholomew & Knott, 1990). . . .	82
4.2	Diagrama de caminho exibindo estimativas dos parâmetros estimados para o modelo de Análise Fatorial Confirmatória relativo às notas obtidas em uma bateria de seis avaliações educacionais aplicada a 112 alunos (Bartholomew & Knott, 1990). A escala dos parâmetros é herdada das variáveis latentes, para as quais foi assumida média nula e variância unitária. . . . .	85
4.3	Modelo de Equações Estruturais para os dados de industrialização e democracia política em países em desenvolvimento (Bollen, 1989, caps. 2,7,8). . . . .	89
4.4	Modelo de Equações Estruturais para os dados de industrialização e democracia política em países em desenvolvimento: isolamento do modelo de mensuração. . . . .	91
4.5	Modelo de Equações Estruturais para os dados de industrialização e democracia política em países em desenvolvimento: isolamento do modelo estrutural. . . . .	91

# Lista de Tabelas

1.1	Terminologia própria de cada modelo para variáveis latentes. . . . .	12
1.2	Notação LISREL para a modelagem da regressão linear entre as variáveis latentes. . . . .	14
1.3	Notação LISREL para o modelo de mensuração associado às variáveis latentes endógenas ( $\eta$ ) e exógenas ( $\xi$ ). . . . .	15
3.1	Modelos de mensuração existentes, conforme tipos de variável. No presente trabalho, trataremos exclusivamente dos modelos para variáveis latentes contínuas. . . . .	30
3.2	Matriz de covariâncias de notas obtidas em 6 avaliações aplicadas em 112 alunos (Bartholomew & Knott, 1990). . . . .	31
3.3	Número mínimo de variáveis a serem observadas para possibilitar a identificabilidade de modelos com 1 a 10 fatores. . . . .	36
3.4	Extração da matriz e covariâncias exibida na Tabela 3.2. Utilizou-se a função <code>factanal()</code> do R para a extração de fatores. . . . .	46
3.5	Rotações da matriz de cargas apresentada na Tabela 3.4. As caselas vazias representam valores praticamente nulos. . . . .	47
3.6	Frequências de respostas dadas ao questionário de atitude em relação ao aborto (McGrath & Waterton, 1986). . . . .	51
3.7	Ajuste de um modelo logístico de 2 parâmetros (ML2) para o conjunto de dados referente à atitude frente à legalização do aborto (McGrath & Waterton, 1986): Estimativas dos parâmetros dos itens. . . . .	56

3.8	Ajuste de um modelo logístico de 2 parâmetros (ML2) para o conjunto de dados referente à atitude frente à legalização do aborto (McGrath & Waterton, 1986): Estimativas dos valores dos traços latentes, segundo o vetor de respostas . . . . .	57
3.9	Quatro parametrizações possíveis para identificabilidade de um modelo de Análise Fatorial Binária . . . . .	63
4.1	Parâmetros estimados para o modelo de Análise Fatorial Confirmatória relativo às notas obtidas em uma bateria de seis avaliações educacionais aplicada a 112 alunos (Bartholomew & Knott, 1990). . . . .	84
4.2	Parâmetros estimados e medidas de qualidade do ajuste, para o modelo de Equações Estruturais para os dados de industrialização e democracia política em países em desenvolvimento (Bollen, 1989, caps. 2,7,8) [ <i>continua na Tabela 4.3</i> ]. . . . .	94
4.3	Parâmetros estimados e medidas de qualidade do ajuste, para o modelo de Equações Estruturais para os dados de industrialização e democracia política em países em desenvolvimento (Bollen, 1989, caps. 2,7,8) [ <i>continuação da Tabela 4.2</i> ]. . . . .	95

# Resumo

Em geral, os diversos modelos de mensuração e de regressão de variáveis latentes surgiram independentemente uns dos outros. De fato, suas estruturas, seus métodos de estimação, a interpretação dos parâmetros estimados, a nomenclatura utilizada e as principais áreas de aplicação são por vezes bastante diferentes entre si. Este trabalho apresenta uma revisão bibliográfica dos principais modelos utilizados na mensuração e regressão de variáveis latentes contínuas, evidenciando suas conexões, normalmente muito pouco evidentes. Para esse fim, será considerada variável latente aquela que não é observada nem observável, passível de ser mensurada de forma aproximada, por meio de outras variáveis indicadoras observáveis.

**Palavras Chave:** *Variáveis latentes, Análise Fatorial, Teoria de Resposta ao Item (TRI), Equações Estruturais, LISREL, GLLAMM.*

# Abstract

In general, the many measurement and regression models involving latent variables have appeared independently of one another. In fact, their structures, methods of estimation, interpretation of parameters, taxonomy and main fields of application are often very distinct. This work presents a bibliographic review of the main models for the measurement and regression of latent continuous variables, eliciting their links, usually not self-evident. In this context, a latent variable will be considered to be that one which can be approximately measured, by means of manifest observable indicators.

**key words:** *Latent variables, Factor Analysis, Item Response Theory (IRT), Structural Equations, LISREL, GLLAMM.*

# Capítulo 1

## Introdução

### 1.1 Objetivo da dissertação

Em geral, os diversos modelos de mensuração e de regressão de variáveis latentes surgiram independentemente uns dos outros. De fato, suas estruturas, seus métodos de estimação, a interpretação dos parâmetros estimados, a nomenclatura utilizada e as principais áreas de aplicação são por vezes bastante diferentes entre si.

Este trabalho apresenta uma revisão bibliográfica dos principais modelos utilizados na mensuração e regressão de variáveis latentes contínuas, evidenciando suas conexões, normalmente muito pouco evidentes. Para esse fim, será considerada variável latente aquela que não é observada nem observável, passível de ser mensurada de forma aproximada, por meio de outras variáveis indicadoras observáveis.

Essa delimitação é importante devido à multiplicidade de interpretações que as variáveis latentes podem ter, como mencionado na Seção 2.1. Ao expandir a interpretação do conceito de *variável latente*, tem-se de entender o *erro de mensuração*, os *coeficientes de regressão aleatórios* e até mesmo *dados faltantes*, entre outros, como realizações de variáveis latentes. Caso todas as interpretações de variável latente fossem contempladas, uma série de outros modelos deveriam integrar o escopo desta dissertação, como a *Regressão com Erros nas Variáveis*, os *Modelos de Efeitos Aleatórios e Mistos* e os *Modelos para Dados Longitudinais*, por exemplo. Dentre esses, apenas a Regressão com Erros nas Variáveis será considerada próxima o suficiente dos modelos em que se deseja aprofundar a discussão, sendo por vezes mencionada,

mas para demarcar seu contraste em relação a esses modelos.

## 1.2 Estrutura da dissertação

Este trabalho explora inicialmente, em seu Capítulo 2, os principais modelos para variáveis latentes que serão abordados com maior profundidade nos capítulos seguintes, sua forma de representação gráfica e a dificuldade decorrente da regressão entre essas variáveis.

No Capítulo 3, são apresentados os modelos de mensuração mais difundidos para variáveis aleatórias: a Análise Fatorial e a Teoria de Resposta ao Item, trazendo também uma discussão sobre situações em que ambos são absolutamente equivalentes.

Ao tratar da regressão de variáveis latentes contínuas, o Capítulo 4 introduz o Modelo de Equações Estruturais (MEE), que possibilita a mensuração e regressão de variáveis latentes de forma simultânea. Em seguida, apresenta casos particulares em que a abordagem pode ser feita passo-a-passo, além do Modelo Linear Generalizado com Variáveis Latentes, que representa uma expansão do MEE.

Finalmente, o Capítulo 5 apresenta uma análise sobre os modelos apresentados, ressaltando suas conexões, bem como recomendações para situações práticas de pesquisa às quais essa análise pode trazer contribuições significativas.

## 1.3 Terminologia e notação utilizadas

No presente trabalho, serão abordados diversos modelos que tratam da mensuração e regressão de variáveis latentes contínuas. Desde o desenvolvimento dos primeiros modelos de mensuração (Spearman, 1904) até os mais recentes avanços nesse tipo de modelagem (Skrondal & Rabe-Hesketh, 2004), devido à independência observada no desenvolvimento das diversas abordagens, conceitos similares vieram sendo chamados por nomes que diferem de uma abordagem para outra. Para amenizar essa dissonância, a Tabela 1.1 apresenta as diversas nomenclaturas associadas a cada conceito.

Diante desse panorama, adotar-se-á a nomenclatura que parecer mais objetiva e geral, sempre que possível. Fazendo isso, pretende-se evidenciar as conexões entre os diversos modelos, chamando estruturas análogas pelo mesmo nome.

**Tabela 1.1:** Terminologia própria de cada modelo para variáveis latentes.

<b>Análise Fatorial</b>	<b>Teoria de Resposta ao Item</b>	<b>Equações Estruturais</b>
$x_i$ Variável observada ou manifesta $i$	$x_{ij}$ Resposta do indivíduo $j$ ao item $i$	$x_i$ (Variável) indicadora $i$ de $\zeta$ 's
$\zeta_i$ Fator $i$	$\zeta_i$ Traço latente $i$	$\zeta_i$ Variável latente exógena $i$
$\lambda_{ij}$ Carga do fator $j$ no item $i$	$\alpha_i$ Parâmetro de discriminação do item $i$	$\lambda_{ij}^x$ Coeficiente estrutural da variável $\zeta_j$ em $x_i$
$\delta_i$ Unicidade	$\beta_i$ Parâmetro de dificuldade do item $i$	$\delta_i$ Erro de mensuração da variável $\zeta_i$
	$c_i$ Parâmetro de acerto ao acaso do item $i$	
		$y_i$ (Variável) indicadora $i$ de $\eta$ 's
		$\eta_i$ Variável latente endógena $i$
		$\lambda_{ij}^y$ Coeficiente estrutural da variável $\eta_j$ em $y_i$
		$\epsilon_i$ Erro de mensuração da variável $\eta_i$
		$\zeta_i$ Erro estrutural de $\eta_i$

Além disso, tentou-se unificar a notação de todos os modelos com a mais correntemente adotada atualmente: a notação LISREL, apresentada nas Tabelas 1.2 e 1.3. Além da atualização, esse tipo de tradução também representa um ganho na percepção de semelhanças entre as estruturas dos diversos modelos, uma vez que elas estarão sendo sempre denotadas de forma consistente.

Em apenas dois modelos foram admitidos desvios a essa diretriz:

- No modelo GLLAMM, algumas notações serão levemente alteradas, em consonância com as alterações feitas pelos próprios autores à notação LISREL; e
- Nos modelo de TRI, a convenção largamente difundida, de denotar o parâmetro de dificuldade por  $\beta$  foi mantida, não se devendo confundir-lo com os homônimos da notação padrão da Tabela 1.2 ( $\mathbf{B} = \{\beta_{ij}\}$ ).

**Tabela 1.2:** Notação LISREL para a modelagem da regressão linear entre as variáveis latentes.

<i>Equação do modelo estrutural:</i>		
$\boldsymbol{\eta} = \mathbf{B} \boldsymbol{\eta} + \boldsymbol{\Gamma} \boldsymbol{\xi} + \boldsymbol{\zeta}$		
<i>Pressupostos:</i>		
$\mathbb{E}(\boldsymbol{\eta}) = \mathbf{0}$		
$\mathbb{E}(\boldsymbol{\xi}) = \mathbf{0}$		
$\mathbb{E}(\boldsymbol{\zeta}) = \mathbf{0}$		
$\boldsymbol{\zeta}$ é independente de $\boldsymbol{\xi}$		
$(\mathbf{I} - \mathbf{B})$ é não-singular		
Símbolo	Dimensões	Definição
<i>Variáveis</i>		
$\boldsymbol{\eta} = \{\eta_i\}$	$m \times 1$	Variáveis latentes endógenas
$\boldsymbol{\xi} = \{\xi_i\}$	$n \times 1$	Variáveis latentes exógenas
$\boldsymbol{\zeta} = \{\zeta_i\}$	$m \times 1$	Erros latentes
<i>Coefficientes</i>		
$\mathbf{B} = \{\beta_{ij}\}$	$m \times m$	Matriz de coeficientes das v. latentes endógenas
$\boldsymbol{\Gamma} = \{\gamma_{ij}\}$	$m \times n$	Matriz de coeficientes das v. latentes exógenas
<i>Matrizes de covariância</i>		
$\boldsymbol{\Phi} = \{\phi_{ij}\}$	$n \times n$	$\mathbb{E}(\boldsymbol{\xi}\boldsymbol{\xi}')$ : Matriz de covariância de $\boldsymbol{\xi}$
$\boldsymbol{\Psi} = \{\psi_{ij}\}$	$m \times m$	$\mathbb{E}(\boldsymbol{\zeta}\boldsymbol{\zeta}')$ : Matriz de covariância de $\boldsymbol{\zeta}$

**Tabela 1.3:** Notação LISREL para o modelo de mensuração associado às variáveis latentes endógenas ( $\boldsymbol{\eta}$ ) e exógenas ( $\boldsymbol{\xi}$ ).

<i>Equações do modelo de mensuração:</i>		
$\mathbf{x} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}$		
$\mathbf{y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon}$		
<i>Pressupostos:</i>		
$\mathbb{E}(\boldsymbol{\eta}) = \mathbf{0}, \mathbb{E}(\boldsymbol{\xi}) = \mathbf{0}, \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \mathbb{E}(\boldsymbol{\delta}) = \mathbf{0}$		
$\boldsymbol{\epsilon}$ não é correlacionado com $\boldsymbol{\eta}, \boldsymbol{\xi}$ e $\boldsymbol{\delta}$		
$\boldsymbol{\delta}$ não é correlacionado com $\boldsymbol{\xi}, \boldsymbol{\eta}$ e $\boldsymbol{\epsilon}$		
Símbolo	Dimensões	Definição
<i>Variáveis</i>		
$\mathbf{y} = \{y_i\}$	$p \times 1$	Variáveis indicadoras de $\boldsymbol{\eta}$
$\mathbf{x} = \{x_i\}$	$q \times 1$	Variáveis indicadoras de $\boldsymbol{\xi}$
$\boldsymbol{\epsilon} = \{\epsilon_i\}$	$p \times 1$	Erros de mensuração de $\mathbf{y}$
$\boldsymbol{\delta} = \{\delta_i\}$	$q \times 1$	Erros de mensuração de $\mathbf{x}$
<i>Coefficientes</i>		
$\boldsymbol{\Lambda}_y = \{\lambda_{ij}^y\}$	$p \times m$	Matriz de coeficientes relacionando $\mathbf{y}$ a $\boldsymbol{\eta}$
$\boldsymbol{\Lambda}_x = \{\lambda_{ij}^x\}$	$q \times n$	Matriz de coeficientes relacionando $\mathbf{x}$ a $\boldsymbol{\xi}$
<i>Matrizes de covariância</i>		
$\boldsymbol{\Theta}_\epsilon = \{\epsilon_{ij}\}$	$p \times p$	$\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')$ : Matriz de covariância de $\boldsymbol{\epsilon}$
$\boldsymbol{\Theta}_\delta = \{\delta_{ij}\}$	$q \times q$	$\mathbb{E}(\boldsymbol{\delta}\boldsymbol{\delta}')$ : Matriz de covariância de $\boldsymbol{\delta}$

## Capítulo 2

# Representação e modelagem de variáveis latentes

Embora tenham sido abordadas de formas bastante diferentes desde sua origem, alguns conceitos e formas de representação podem ser utilizados indistintamente nesses diversos contextos. Um exemplo de ferramenta transversal a diversos modelos para variáveis latentes é o *diagrama de caminho*, que evidencia as relações entre as variáveis observadas e latentes de um modelo.

Além disso, o problema da subestimação dos coeficientes do modelo de regressão ajustado a variáveis latentes é inerente a esse tipo de ajuste, independentemente da forma com que tais variáveis sejam mensuradas, salvo para casos extremamente particulares, apresentados na Seção 4.2.

Este Capítulo procura introduzir noções que perpassam múltiplas abordagens dadas às variáveis latentes, bem como apresentar de modo geral os diversos modelos que serão tratados nos Capítulos seguintes.

### 2.1 A onipresença das variáveis latentes

Estamos cercados de inúmeros fenômenos altamente complexos que, não raro, envolvem quantidades não observáveis diretamente. O estudo quantitativo de tais fenômenos requer que cada conceito não-mensurável seja desdobrado em termos de variáveis observadas que representem bem a variável *latente* de interesse. É evidente que tal

tradução acarreta a inserção de imprecisões na modelagem que se venha a realizar. Sendo assim, é preciso atentar para o fato de que a negligência da baixa confiabilidade das variáveis observadas escolhidas pode produzir estimações e inferências que não representam adequadamente o verdadeiro fenômeno de interesse, invalidando o estudo. Contudo, apesar da dificuldade de se pesquisar variáveis latentes, as ciências sociais e do comportamento lidam constantemente com diversas dessas variáveis e foram as precursoras desse tipo de modelagem (Clogg, 1992).

É o que ocorre em aferições de inteligência, por exemplo, utilizando-se testes por meio dos quais se espera mensurar, a partir do número de questões corretamente respondidas por um aluno, seu grau de proficiência em uma determinada disciplina. Nesse caso, a nota representa o valor correspondente à sua inteligência, medida na escala determinada pela prova aplicada, de modo que a aplicação da mesma prova em outro aluno permite a comparação de suas proficiências, ambas aferidas na mesma escala.

Evidentemente, esse tipo de técnica envolve um *erro de mensuração* que deve ser modelado e depende tanto do nível em que o respondente se situa na escala de medição do construto quanto das características dos itens do instrumento de coleta. Devido à imprecisão de suas medidas, variáveis latentes podem ser definidas como variáveis aleatórias, sendo suas realizações ocultas ao pesquisador (Skrondal & Rabe-Hesketh, 2004). Essa definição ampla permite que diversos fenômenos possam ser observados como ocorrências particulares de variáveis latentes, como por exemplo:

- variáveis contínuas não observadas, geradoras de respostas categóricas observadas (Pearson, 1901; Bartholomew, 1980);
- variáveis que tenham sido mensuradas com erro (Carroll *et al.*, 2006; Fuller, 1987);
- covariáveis não observadas dos indivíduos de uma população, que sejam responsáveis por heterogeneidades inicialmente desconsideradas no modelo proposto;
- dados faltantes (*missing data*) (Little & Rubin, 2002); e
- construtos hipotéticos não observáveis, estudados por meio de covariáveis observáveis (Spearman, 1904; Lord & Novick, 1968; Bollen, 1989).

Na presente dissertação, por motivos de objetividade e foco, concentrar-nos-emos no último caso, em que variáveis latentes são utilizadas como representação de construtos não observáveis diretamente. A necessidade de se aferir a magnitude de tais variáveis é recorrente e os modelos que descrevem a relação entre elas e suas covariáveis observáveis são convenientemente chamados de *modelos de mensuração*. Tais relações tornam-se mais facilmente apreensíveis quando representadas por *diagramas de caminho*.

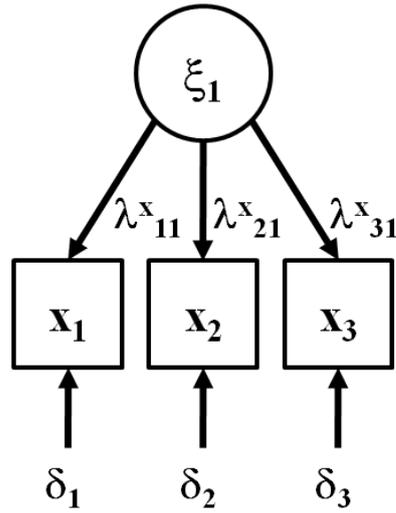
## 2.2 Diagramas de caminho

A representação usual dos modelos que envolvem variáveis latentes dá-se por meio do assim chamado *diagrama de caminho* (*path diagram*). Ele evidencia as relações existentes entre as variáveis latentes (representadas por círculos) e as observáveis (representadas por quadrados). Setas indicam a relação linear de causalidade entre as variáveis e o coeficiente dessa relação é indicado sobre a seta. Além disso, os erros de mensuração são representados por pequenas setas apontando na direção das variáveis observáveis em que ele incide. O diagrama da Figura 2.1 representa um *modelo de mensuração* em que, para cada elemento da amostra observada, a realização da variável latente  $\xi_1$  pode ser mensurada por meio das variáveis observáveis  $X_1$ ,  $X_2$  e  $X_3$ , que apresentam erros de mensuração  $\delta_1$ ,  $\delta_2$  e  $\delta_3$ , respectivamente. Esse modelo de mensuração também pode ser expresso pela equação

$$\mathbf{x} = \mathbf{\Lambda}_x \xi_1 + \boldsymbol{\delta}, \quad (2.1)$$

onde  $\mathbf{x} = [x_1, x_2, x_3]'$ ,  $\mathbf{\Lambda}_x = [\lambda_{11}^x, \lambda_{21}^x, \lambda_{31}^x]'$  e  $\boldsymbol{\delta} = [\delta_1, \delta_2, \delta_3]'$ .

Vale notar que o intercepto das relações lineares não é representado, nem nos diagramas, sequer nas equações, pelo fato de as variáveis em estudo serem preferencialmente padronizadas, caso em que o intercepto é nulo por definição. Tal escolha deve-se à facilidade de interpretação dos coeficientes estimados quando as variáveis em questão forem padronizadas, já que representarão a correlação entre as variáveis. Além disso, a padronização elimina problemas de estimação decorrentes de escalas com magnitudes muito diferentes e facilita a interpretação e estabilidade dos coeficientes estimados.



**Figura 2.1:** Diagrama de caminho de um modelo de mensuração para a variável latente  $\xi_1$ . As variáveis latentes são representadas por círculos e as observáveis, representadas por quadrados. Setas indicam a relação linear de causalidade entre as variáveis e o coeficiente dessa relação é indicado sobre a seta. As pequenas setas representam erros de mensuração.

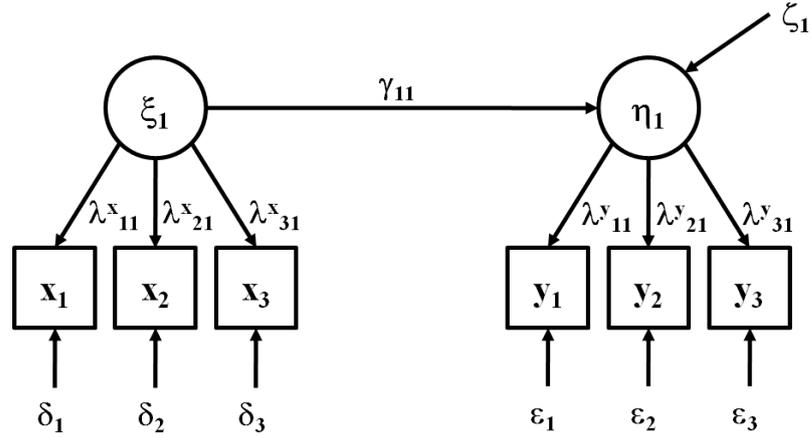
Além de modelos de mensuração, em que as relações de causalidade se dão entre as variáveis latentes e as observadas, os diagramas de caminho também podem ser utilizados na representação das relações entre variáveis latentes, chamadas de *equações estruturais*. A Figura 2.2 exibe o diagrama de caminho de um modelo com uma equação estrutural, relacionando as variáveis latentes  $\eta_1$  a  $\xi_1$ , cada uma delas com o seu respectivo modelo de mensuração, dado pelas variáveis de  $\mathbf{y}$ , no caso de  $\eta_1$ , e  $\mathbf{x}$ , no caso de  $\xi_1$ . O modelo da Figura 2.2 pode ser expresso pelas equações

$$\eta_1 = \gamma_{11} \xi_1 + \zeta_1 \quad (2.2)$$

$$\mathbf{x} = \mathbf{\Lambda}_x \xi_1 + \boldsymbol{\delta} \quad (2.3)$$

$$\mathbf{y} = \mathbf{\Lambda}_y \eta_1 + \boldsymbol{\epsilon} . \quad (2.4)$$

Na realidade, o modelo expresso pela Figura 2.2 contém aquele da Figura 2.1, o que é evidenciado pela identidade das equações (2.1) e (2.3). De fato, o modelo de mensuração (2.1) é utilizado na aferição da variável latente  $\xi_1$ , que compõe a regressão (2.2) como covariável. Como a variável latente  $\eta_1$  representa, não a covariável, mas a variável-resposta dessa mesma regressão (2.2), sua notação é diferente, sendo ela



**Figura 2.2:** Diagrama de caminho da *equação estrutural* representada pela regressão entre as variáveis latentes  $\eta_1$  e  $\xi_1$ , cada uma aferida por seu *modelo de mensuração* específico.

expressa pela letra grega  $\eta$  e as variáveis observáveis associadas pela letra  $y$ . Essa diferença de notação corresponde a uma distinção conceitual entre variáveis *endógenas* e *exógenas*. Em um diagrama de caminhos, todas as variáveis que forem modeladas como *variáveis-resposta* são consideradas endógenas e aquelas que não o forem em nenhuma equação, são consideradas exógenas. Em outras palavras, as variáveis exógenas não possuem suas “causas” explicitadas no modelo, ao passo que as “causas” das variáveis endógenas o são. Essa distinção desempenhará um papel importante na estimação e na identificabilidade dos modelos de Equações Estruturais, a serem discutidos no Capítulo 4.

Os usuais Estimadores de Mínimos Quadrados Ordinários, entretanto, são viesados para regressões entre variáveis latentes, sendo necessárias correções ou a utilização de outros modelos cuja definição contemple esse tipo de fenômeno, mais detalhado na seção seguinte, como os Modelos de Equações Estruturais e os Modelos Lineares Generalizados com Variáveis Latentes.

## 2.3 Regressão de variáveis latentes: atenuação

O Modelo de Regressão Linear Clássico é dado por

$$y_i = \mathbf{\Gamma}'\mathbf{x} + \zeta_i ,$$

com  $i = 1, 2, \dots, N$  onde o vetor  $\mathbf{x}$  é formado por valores considerados fixos e observados sem erro de mensuração. Além disso, assume-se que todas as covariáveis são independentes entre si e o único componente aleatório da equação é formado pelos erros  $\zeta_i \sim N(0, 1)$ , independentes entre si, bem como das covariáveis. Contudo, nos casos em que tais covariáveis forem medidas com erro ou consistirem em estimativas de quantidades não observáveis, como é o caso das variáveis latentes, é mais adequado considerá-las como variáveis aleatórias em vez de constantes predeterminadas. Negligenciar esse fato e utilizar o método de Mínimos Quadrados Ordinários para estimar os coeficientes  $\mathbf{\Gamma}$  faz com que os valores obtidos sejam subestimados, sendo esse efeito chamado correntemente de *atenuação*.

Para melhor ilustrar o fenômeno, tomemos como exemplo o caso em que se deseja estimar o coeficiente  $\gamma_{11}$  da relação linear entre uma variável-resposta  $y_1$  e uma única covariável  $x_1$ , ignorando o fato de ambas serem observadas com erro de mensuração. Nesse caso, o modelo verdadeiro é descrito pela *equação estrutural*

$$\eta_1 = \gamma_{11} \xi_1 + \zeta_1 \tag{2.5}$$

e, simultaneamente, pelo *modelo de mensuração*

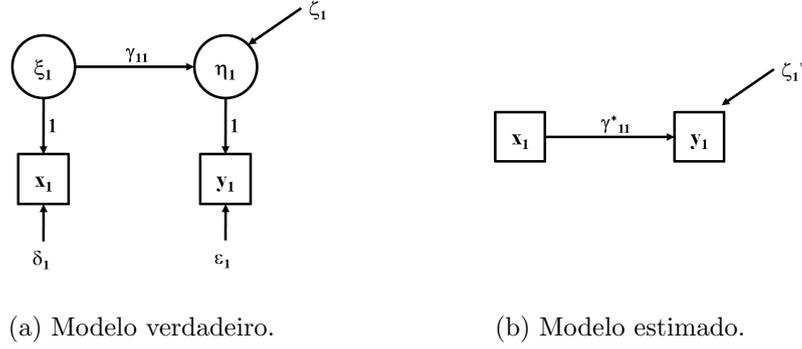
$$\mathbf{x} = \xi_1 + \boldsymbol{\delta} \tag{2.6}$$

$$\mathbf{y} = \eta_1 + \boldsymbol{\epsilon} , \tag{2.7}$$

em que se assumiu  $\mathbf{\Lambda}_x = \mathbf{\Lambda}_y = [1, 1, 1]$ , o que significa que as variáveis observadas  $\mathbf{x}$  e  $\mathbf{y}$  estão na mesma escala das respectivas variáveis latentes  $\xi_1$  e  $\eta_1$ , salvo pelos erros de mensuração  $\boldsymbol{\delta}$  e  $\boldsymbol{\epsilon}$ . Por outro lado, o modelo que negligencia o erro de mensuração é dado simplesmente por

$$y_1 = \gamma_{11}^* x_1 + \zeta_1^* , \tag{2.8}$$

em que o coeficiente linear  $\gamma_{11}^*$  é diferente daquele do modelo verdadeiro, dado por  $\gamma_{11}$ , bem como o erro estrutural  $\zeta_1^*$  difere de  $\zeta_1$  pelo fato de incorporar o erro de



**Figura 2.3:** Diagramas de caminho do modelo verdadeiro, em que os erros de mensuração estão presentes, e do modelo ajustado, em que são negligenciados.

mensuração, não especificado pelo modelo (2.8). A Figura 2.3 ilustra, por meio de seus diagramas de caminho, a diferença entre o modelo verdadeiro e o ajustado.

Evidentemente,  $\hat{\gamma}_{11}^*$  representa a estimativa ingênua do verdadeiro coeficiente  $\gamma_{11}$ . Sabemos que o Estimador de Mínimos Quadrados do coeficiente linear da equação (2.8) é dado por

$$\begin{aligned}
 \gamma_{11}^* &= \frac{\text{Cov}(X_1, Y_1)}{\mathbb{V}(X_1)} \\
 &= \frac{\text{Cov}(\xi_1 + \delta_1, \eta_1 + \epsilon_1)}{\mathbb{V}(X_1)} \\
 &= \frac{\text{Cov}(\xi_1, \eta_1)}{\mathbb{V}(X_1)}, \tag{2.9}
 \end{aligned}$$

pois os erros de mensuração de  $X_1$  e  $Y_1$ , respectivamente  $\delta_1$  e  $\epsilon_1$  são assumidos independentes, de modo que  $\text{Cov}(\delta_1, \epsilon_1) = 0$ . Como da equação (2.5) temos que

$$\begin{aligned}
 \text{Cov}(\xi_1, \eta_1) &= \text{Cov}(\xi_1, \gamma_{11} \xi_1 + \zeta_1) \\
 &= \text{Cov}(\xi_1, \gamma \xi_1) \\
 &= \gamma_{11} \mathbb{V}(\xi_1), \tag{2.10}
 \end{aligned}$$

substituindo (2.10) em (2.9) segue que

$$\gamma_{11}^* = \gamma_{11} \left[ \frac{\mathbb{V}(\xi_1)}{\mathbb{V}(X_1)} \right], \tag{2.11}$$

de onde obtemos

$$0 \leq \rho_{xx} = \frac{\mathbb{V}(\xi_1)}{\mathbb{V}(X_1)} \leq 1,$$

que representa o fator de *atenuação* do verdadeiro coeficiente  $\gamma_{11}$ , decorrente da negligência do erro de mensuração da variável observada  $x_1$ . Como  $\rho_{xx}$  indica a proporção da variância observada que pode ser atribuída à real variância da variável latente e não a outra fonte de variabilidade, considerada erro de mensuração, ele é comumente chamado de *confiabilidade*. Assim sendo, podemos afirmar que o coeficiente de uma regressão linear será estimado com um viés de atenuação inversamente proporcional à confiabilidade da covariável. O coeficiente da regressão ingênua, portanto, será igual ao da verdadeira regressão somente quando a confiabilidade for máxima ( $= 1$ ), isto é, na ausência de erro de mensuração. De fato, conforme já mencionado inicialmente, a ausência de erro de mensuração é um dos pressupostos para a utilização do Método de Mínimos Quadrados na estimação do coeficiente de uma Regressão Linear Simples. Finalmente, para o caso de mais covariáveis, obtemos o resultado análogo

$$\Gamma^{*'} = \Sigma_{xx'}^{-1} \Sigma_{x\xi'} \Gamma'$$

onde  $\Sigma_{ab'}$  denota a matriz de covariâncias populacional dos vetores aleatórios  $\mathbf{a}$  e  $\mathbf{b}$ .

## 2.4 Modelos de mensuração, de regressão e híbridos

Várias correções são possíveis para o fenômeno da atenuação, estando sua maioria no escopo dos Modelos de Regressão com Erros nas Variáveis (Fuller, 1987). Em tais modelos, há ao menos uma covariável mensurada com erro, seja por meio de um instrumento pouco preciso ou a partir de um modelo de mensuração. De uma forma ou de outra, o ajuste desses modelos pressupõe que a covariável que apresenta erro aleatório já foi mensurada.

Uma solução alternativa, que abarca e expande a da utilização de Modelos de Regressão com Erros nas Variáveis, é o ajuste de Modelos de Equações Estruturais, que unem tanto o problema de mensuração das variáveis latentes envolvidas quanto a regressão entre elas em uma só modelagem e, portanto, em um único processo de estimação. Esses modelos permitem o ajuste de modelos descritos por diagramas de caminho como o da Figura 2.2, que relaciona variáveis latentes, evidentemente mensuradas com erro, o que abarca inclusive a simples regressão representada pelo diagrama

da Figura 2.3a, eliminando o efeito da atenuação sobre os parâmetros estimados, uma vez que o modelo possibilita a especificação dos eventuais erros de mensuração.

Os Modelos de Equações Estruturais, portanto, lidam com o problema de mensuração e o de regressão simultaneamente, ao passo que outros modelos, ou tratam somente da mensuração, como a Análise Fatorial e a Teoria de Resposta ao Item, ou exclusivamente da regressão, a exemplo do Modelo de Regressão com Erros nas Variáveis. Na medida em que abarcam, como casos particulares, modelos de mensuração ou de regressão, além de combinarem os dois tipos de modelos em uma só classe, os Modelos de Equações Estruturais podem ser vistos como *híbridos*.

## 2.5 Atributos dos modelos em estudo

Ainda que frequentemente pouco evidente, os modelos de *mensuração*, os de *regressão* e os que podemos chamar de *híbridos*, por combinarem os dois tipos anteriores, apresentam conexões que podem se revelar úteis quando de sua aplicação. O quadro da Figura 2.4 apresenta sete atributos que podem ser incorporados ao Modelo de Regressão Linear Clássico, aumentando sua complexidade, no intuito de torná-lo compatível com fenômenos também mais complexos. A última coluna, intitulada “Complexidade”, apresenta a quantidade dentre os sete atributos elencados, que são compreendidas por cada modelo em questão. Vários dentre os modelos apresentados no quadro serão abordados na presente dissertação, que detalhará, oportunamente, as relações introduzidas na presente Seção.

### 2.5.1 Modelos de Regressão

#### Regressão com Erros nas Variáveis

Na Figura 2.4, dentre os Modelos de Regressão, vale notar que o *Modelo de Regressão com Erros nas Variáveis* consiste de uma adequação do modelo clássico, para comportar covariáveis que apresentem erro de mensuração. Conforme mencionado na Seção 2.3, o Modelo de Equações Estruturais Geral pode servir ao mesmo fim, mas por meio de uma abordagem via variáveis latentes. A diferença está no fato de que, no caso da Regressão com Erros nas Variáveis, as covariáveis mensuradas com erro representam

Classe de Modelos	Modelo	Erro de mensuração	Duas ou mais equações	Função de ligação	Variável-resposta com densidade da família exponencial	Possibilidade de uso de Indicadoras (respostas) categóricas	Covariáveis latentes	Possibilidade de regressão de variáveis latentes	Complexidade
<b>Modelos de Regressão</b>	Regressão Linear								0
	Regressão com Erros nas Variáveis	✓							1
	Equações Estruturais: Equações Simultâneas		✓						1
	Modelos Lineares Generalizados			✓	✓				3
<b>Modelos de Mensuração</b>	Equações Estruturais: Análise Fatorial Confirmatória		✓				✓		2
	Análise Fatorial Binária		✓			✓	✓		3
	Teoria de Resposta ao Item		✓	✓		✓	✓		4
<b>Modelos de Mensuração e Regressão</b>	Equações Estruturais: Modelo Geral (LISREL)		✓				✓	✓	3
	Modelos Lineares Generalizados com Variáveis Latentes		✓	✓	✓	✓	✓	✓	6

**Figura 2.4:** Quadro de comparação entre alguns modelos de mensuração, de regressão e de híbridos. O valor da “complexidade” equivale à quantidade compreendida por cada modelo, dos atributos apresentados no quadro.

dados efetivos que, mesmo com erros, serão utilizados no ajuste da regressão, em vez de serem representadas por meio de uma variável latente cujas mensurações serão obtidas por meio de outras variáveis observáveis correlacionadas a ela que, essas sim, serão utilizadas como dados para o ajuste simultâneo dos modelos de mensuração e regressão, conforme ilustrado pela Figura 2.2.

Como nos Modelos de Regressão com Erros nas Variáveis não ocorre a modelagem da mensuração das variáveis latentes, mas, ao contrário, utiliza medidas já observadas como dados para o ajuste, tais modelos não foram considerados pertencentes à classe dos Modelos de Mensuração.

### **Equações Simultâneas**

Partindo do Modelo de Regressão Linear Clássica, outra expansão possível é o ajuste de um sistema de equações de regressão lineares, em vez de apenas uma única equação. Tais modelos são frequentemente utilizados por econométristas, que costumam denominá-los *Equações Simultâneas*. Tais modelos representam uma dentre várias ocorrências particulares da classe mais geral compreendida pelos Modelos de Equações Estruturais (MEE), podendo também ser chamado de Modelo de Equações Estruturais com Variáveis Observadas (MEEVO).

## **2.5.2 Modelos de Mensuração**

### **Análise Fatorial Confirmatória**

No mesmo quadro, podemos observar que os MEE também estão presentes entre os modelos de mensuração, na forma de Análise Fatorial Confirmatória (AFC) e entre os modelos híbridos, de mensuração e regressão, na forma de LISREL. De fato, por conjugar o problema da mensuração com o da regressão, o LISREL comporta modelos exclusivamente de mensuração ou de regressão como casos particulares dessa classe de MEE.

No caso da Análise Fatorial Confirmatória, ela consiste de um modelo de mensuração para variáveis latentes contínuas, a partir de observações também contínuas.

## Teoria de Resposta ao Item

Uma especialização desse modelo pode ser feita por meio da utilização de funções de ligação que possibilitem a utilização de observações categorizadas para a mensuração de variáveis latentes contínuas. Na realidade, esse tipo de modelo de mensuração pertence à classe denominada Teoria de Resposta ao Item (TRI), que recebe esse nome pelo fato de seus modelos serem descritos em termos de uma função de resposta para os itens avaliados, de modo que o que é efetivamente modelado não é a resposta efetiva, isto é, a variável categórica observada, mas a probabilidade de cada categoria possível ser observada como resposta. Esses dois modelos de mensuração partem de abordagens tão distintas que parecem não ter conexão alguma. Contudo a proximidade entre eles, mostrada no quadro, é real e foi desvelada em primeira mão por Takane & de Leeuw (1987) (cf. Seção 3.4).

### 2.5.3 Modelos de Mensuração e Regressão (*híbridos*)

#### Modelo de Equações Estruturais Geral (LISREL)

Além da mensuração das variáveis latentes, pode ser interessante ajustar a regressão entre tais mensurações. Nesse contexto, os modelos de mensuração e regressão constituem importantes ferramentas de análise estatística e representam uma classe geral para todos os modelos citados previamente, além de outros que fogem ao escopo da presente dissertação, sendo o *híbrido* mais difundido o LISREL.

#### Modelos Lineares Generalizados com Variáveis Latentes (GLLMM)

Os Modelos Lineares Generalizados com Variáveis Latentes, chamados de GLLMM, representam uma expansão do LISREL. Evidentemente, tal generalidade vem acompanhada de uma maior complexidade, que pode tornar a aplicação de tais modelos proibitiva para o pesquisador de outras áreas de conhecimento. Entretanto, o conhecimento das relações entre os diversos modelos possibilita abordagens alternativas, de mesma eficácia e maior eficiência que a das abordagens mais complexas.

# Capítulo 3

## Mensuração de variáveis latentes contínuas

O primeiro desafio imposto pelas variáveis latentes ao pesquisador que deseja estudá-las é sua mensuração. Por não serem observáveis, sua aferição apoia-se em dados de variáveis observadas que estejam correlacionadas com elas. Neste Capítulo, são apresentadas algumas definições gerais concernentes aos modelos de mensuração, na Seção 3.1. Em seguida, serão apresentados dois modelos bastantes difundidos para a mensuração de variáveis contínuas: a Análise Fatorial, na Seção 3.2, e a Teoria de Resposta ao Item, na Seção 3.3. Conclui-se abordando a relação entre esses dois modelos, na Seção 3.4

### 3.1 Modelos de mensuração: definições preliminares

A fim de apurar o foco deste trabalho, cabe introduzir uma distinção importante. As variáveis latentes contínuas e os valores que assumem costumam ser denominados, respectivamente, *fatores* e *escores*. Tais escores, dependendo da forma com que os escores relacionam-se com as variáveis observadas, ainda podem ser de um dos dois tipos a seguir:

**Escores verdadeiros clássicos** As variáveis latentes não são diretamente mensuráveis por definição e as variáveis observadas representam *proxies*, i.e., indi-

cadores, que se aproximam da variável latente de real interesse por meio de relações traduzidas matematicamente em termos de *modelos de fator comum*, mais conhecidos como modelos de *Análise Fatorial*. Tais modelos podem ser unidimensionais ou multidimensionais, a depender do número de fatores. A utilização de indicadores não-contínuos dá origem aos modelos da *Teoria de Resposta ao Item* (TRI) (Harman, 1976; Lord & Novick, 1968); ou

**Escores verdadeiros platônicos**, quando existir a possibilidade de as variáveis latentes serem diretamente aferidas por um *padrão-ouro* de medida, sendo a opção por aferições com erro de mensuração feita por razões eminentemente práticas. A formulação matemática desse tipo de problema dá origem aos *modelos para erro de mensuração* (Carroll *et al.*, 2006; Fuller, 1987).

Nesta dissertação, focaremos nos modelos para escores verdadeiros clássicos. Como mencionado anteriormente, modelos para erro de mensuração, como a Regressão com Erros nas Variáveis serão mencionadas apenas a fim de contrastá-los com os modelos abordados.

Em termos matemáticos, a ideia básica que sustenta um modelo de mensuração para variáveis latentes consiste em encontrar, para um conjunto de variáveis observáveis diretamente  $X_1, \dots, X_q$ , um conjunto de variáveis latentes  $\xi_1, \dots, \xi_n$  (com  $n \ll q$ ) que contenha essencialmente a mesma informação sobre dependência entre as variáveis. Em geral, os modelos de mensuração de variáveis latentes equivalem a uma regressão a respeito de variáveis latentes, que segue a forma

$$\mathbb{E}(X_i | \boldsymbol{\xi}) = g(\kappa_{i0} + \kappa_{i1}\xi_1 + \dots + \kappa_{in}\xi_n), \quad i = 1, \dots, q, \quad (3.1)$$

onde  $g(\cdot)$  é uma função de ligação,  $\kappa_{i0}, \dots, \kappa_{in}$  são os coeficientes de regressão da  $i$ -ésima variável manifesta e  $X_i$  é independente de  $X_j$ , para  $i \neq j$ , dado  $\boldsymbol{\xi} = \{\xi_1, \dots, \xi_n\}$ .

A mesma equação dá origem ao modelo de Análise Fatorial e a alguns modelos da Teoria de Resposta ao Item, ainda que, sem a abstração proporcionada por essa equação, as definições dos dois modelos não pareçam ter relação.

Desse modo, o propósito de ambas as teorias é praticamente o mesmo. No entanto, as pequenas dessemelhanças estruturais entre as duas técnicas resultaram em desenvolvimentos teóricos bastante distintos, como se fará notar ao longo deste trabalho.

**Tabela 3.1:** Modelos de mensuração existentes, conforme tipos de variável. No presente trabalho, trataremos exclusivamente dos modelos para variáveis latentes contínuas.

Variável manifesta	Variável latente	
	Contínua	Categórica
Contínua	Análise Fatorial	Análise de Perfil Latente
Categórica	Modelo de Traço Latente (TRI)	Análise de Classe Latente

A diferença básica entre os modelos de mensuração de variáveis latentes situa-se no tipo de variável-resposta e no tipo das variáveis independentes das equações que definem o modelo. As diversas combinações dessas características dão origem aos modelos mostrados na Tabela 3.1.

Como o escopo do presente trabalho restringe-se à mensuração de escores verdadeiros clássicos e nesta Seção serão abordados os modelos de Análise de Fatorial e o de Teoria de Resposta ao Item. Na conclusão, apresentaremos a relação existente entre os dois tipos de modelos, apesar de sua aparente desconexão.

## 3.2 Análise Fatorial (1904)

A Análise Fatorial é, provavelmente, o primeiro modelo de mensuração de variáveis latentes de que se tem notícia. Sua origem é devida a Spearman (1904) e decorre da tentativa de provar matematicamente a existência de um fator latente geral, o assim chamado *fator G*, que seria responsável por uma parte significativa da variabilidade de todas as notas de uma mesma pessoa em diversas disciplinas. Em outras palavras, mesmo que uma pessoa tenha proficiências específicas, medidas por cada uma das provas, ela também possui uma proficiência geral, que explica boa parte das notas específicas.

Dando continuidade ao trabalho de Spearman, Thurstone (1931) ampliou a Análise Fatorial para comportar múltiplos fatores. O problema a ser enfrentado então era a deficiência dos procedimentos de estimação. Com o desenvolvimento de aplicações do método da máxima verossimilhança por Lawley (1940), foi aberto caminho para

**Tabela 3.2:** Matriz de covariâncias de notas obtidas em 6 avaliações aplicadas em 112 alunos (Bartholomew & Knott, 1990).

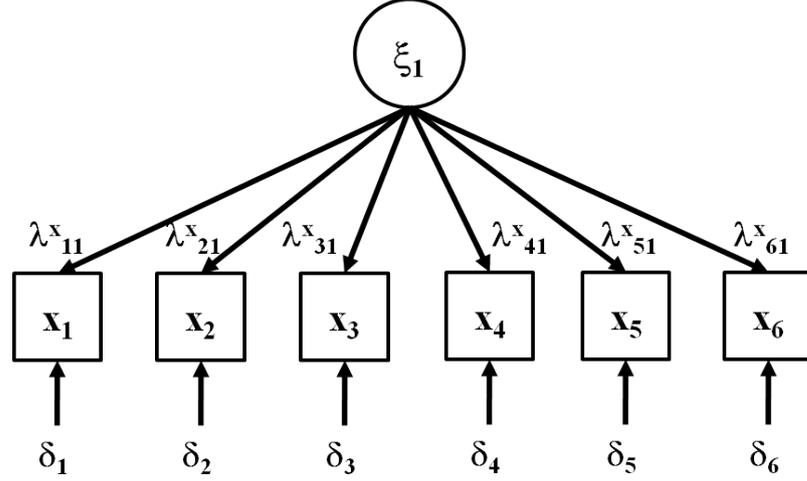
	general	picture	blocks	maze	reading	vocab
general	24.64	5.99	33.52	6.02	20.75	29.70
picture	5.99	6.70	18.14	1.78	4.94	7.20
blocks	33.52	18.14	149.83	19.42	31.43	50.75
maze	6.02	1.78	19.42	12.71	4.76	9.07
reading	20.75	4.94	31.43	4.76	52.60	66.76
vocab	29.70	7.20	50.75	9.07	66.76	135.29

o desenvolvimento de rotinas de estimação mais eficientes e confiáveis na década de 1960. Nesse ínterim, foram desenvolvidos outros modelos para variáveis latentes, como o Modelo de Classes Latentes (Lazarsfeld & Henry, 1968) e a Teoria de Resposta ao Item (TRI) (Lord & Novick, 1968). Somente nos anos 70 e 80 foram desenvolvidos os Modelos para Estrutura de Covariâncias, como o Modelo de Equações Estruturais, embora a Análise de Caminho, desenvolvida por Wright (1918), já constituísse um primeiro avanço nessa área, infelizmente subestimado na Sociologia até a década de 1960. Finalmente, um avanço importante na estimação desses modelos foi a criação do algoritmo EM (*Expectation Maximization*), por Dempster *et al.* (1977), possibilitando um tratamento mais adequado das variáveis latentes.

A Tabela 3.2, representa a matriz de covariâncias de um vetor cujas variáveis aleatórias correspondem às notas obtidas em uma bateria de seis avaliações educacionais aplicada a 112 alunos (Bartholomew & Knott, 1990). As habilidades avaliadas são: conhecimentos gerais (*general*), capacidade de completar figuras (*picture*), inteligência espacial (*blocks*), capacidade de resolução de labirintos (*maze*), leitura (*reading*) e vocabulário (*vocab*).

Segundo o modelo de Spearman (1904) para a inteligência, deveria ser possível encontrar um fator latente que fosse capaz de explicar, por si só, grande parte da variabilidade encontrada nas seis notas observadas entre os 112 alunos, conforme ilustrado no diagrama de caminho da Figura 3.1.

O modelo clássico de Análise Fatorial com um fator ( $\xi_1$ ) para as seis variáveis



**Figura 3.1:** Modelo de Análise Fatorial com complexidade 1: seu diagrama de caminho representa o único fator latente ( $\xi_1$ ) comum a todas as seis variáveis observadas ( $\mathbf{x}$ ).

observadas  $(x_1, x_2, \dots, x_6)$  é dado por

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \vdots \\ \lambda_{61} \end{bmatrix} \xi_1 + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_6 \end{bmatrix}. \quad (3.2)$$

onde  $\mathbb{E}(\boldsymbol{\delta}) = 0$ ,  $\delta_i$  e  $\delta_j$  são independentes  $\forall i \neq j \in \{1, 2, \dots, 6\}$ , bem como  $\xi_1$  e  $\boldsymbol{\delta}$  também são independentes. A modelagem considera, portanto, que existe um *fator geral* latente, comum a todas as variáveis, que é o responsável por parte significativa da variabilidade dessas variáveis.

Os valores de  $\boldsymbol{\lambda}$  são chamados de *cargas* dos fatores e indicam o peso que o fator tem na variável. No exemplo, as variáveis que possuem maiores cargas são as que possuem maior variabilidade em comum com o fator  $\xi_1$ . A análise comparativa das cargas evidencia as variáveis que indicam com maior confiabilidade o valor verdadeiro do fator (altas cargas) e podem revelar a ineficácia de alguma variável para esse fim (baixas cargas).

O caso em que um único fator é utilizado para modelar a variabilidade das variáveis observadas é chamado de *unidimensional*. Evidentemente, é possível que as

realizações do vetor aleatório do exemplo citado sejam melhor ajustadas se for considerada a existência de mais que um fator, caso em que a Análise Fatorial é chamada de *multidimensional*.

Para o caso de 2 fatores, por exemplo, o modelo é expresso pela equação

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \vdots & \vdots \\ \lambda_{61} & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_6 \end{bmatrix}. \quad (3.3)$$

onde  $\mathbb{E}(\boldsymbol{\delta}) = 0$ ,  $\delta_i$  e  $\delta_j$  são independentes  $\forall i \neq j \in \{1, 2, \dots, 6\}$ , bem como  $\xi$  e  $\boldsymbol{\delta}$  também são independentes. Nesse modelo bidimensional, é possível compreender a participação que cada fator tem uma variável observada, por meio da análise das cargas estimadas.

Para os dados da Tabela 3.2, os dois fatores poderiam representar, por exemplo, “proficiência linguística” e “proficiência matemática”, estando o primeiro fator mais fortemente associado às notas observadas nos testes de leitura e vocabulário, o segundo fator mais fortemente associado aos testes de completar figuras, noção espacial e labirintos, ao passo que ambos os fatores deveriam estar moderadamente associados às notas do teste de conhecimentos gerais.

Generalizando para  $n$  fatores e  $q$  variáveis observadas, obtemos o modelo

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{q1} & \lambda_{q2} & \cdots & \lambda_{qn} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_q \end{bmatrix}, \quad (3.4)$$

resumidamente denotado por

$$\mathbf{X} = \mathbf{\Lambda} \boldsymbol{\xi} + \boldsymbol{\delta},$$

onde  $\mathbb{E}(\boldsymbol{\delta}) = 0$ ,  $\delta_i$  e  $\delta_j$  são independentes  $\forall i \neq j \in \{1, 2, \dots, q\}$ , bem como  $\xi$  e  $\boldsymbol{\delta}$  também são independentes.

O modelo de Análise Fatorial, apresentado na equação (3.4), consiste de um sistema de equações lineares em que as variáveis observadas desempenham o papel de variável-reposta, ao passo que os fatores funcionam como covariáveis. Devido ao fato

de as covariáveis serem latentes, o problema de estimação não pode resumir-se ao de um sistema de regressões lineares.

De fato, a Análise Fatorial pertence à classe dos *Modelos de Estrutura de Covariância*. Essa classe é composta por modelos cujos parâmetros são quantidades desconhecidas que formam a matriz de covariâncias das variáveis observadas. No caso da Análise Fatorial, a matriz de covariância  $\Sigma$  implicada pelo modelo é dada por

$$\begin{aligned}\Sigma(\boldsymbol{\theta}) &= \Lambda \mathbb{E} (\boldsymbol{\xi}_1 \boldsymbol{\xi}_1') \Lambda' + \Theta_\delta \\ &= \Lambda \Phi \Lambda' + \Theta_\delta ,\end{aligned}\tag{3.5}$$

onde  $\boldsymbol{\theta}$  representa o vetor de parâmetros do modelo, composto por  $\Lambda$ ,  $\Phi$  e  $\Theta_\delta$ , devendo  $\Phi$  e  $\Theta_\delta$  ser positivas definidas, uma vez que representam matrizes de covariâncias. Além disso, requer-se que  $\Theta_\delta$  seja diagonal, dada a independência entre os elementos de  $\boldsymbol{\delta}$ . Os dados para estimação dos parâmetros de  $\Sigma(\boldsymbol{\theta})$  provêm da matriz de covariância amostral  $\mathbf{S}$ .

Contudo, o modelo (3.4) apresenta o que Anderson (1984, p.552) chamou de uma “indeterminação fundamental”. Sejam

$$\boldsymbol{\xi}^* = \mathbf{C}^{-1} \boldsymbol{\xi} \quad \text{e} \quad \Lambda^* = \Lambda \mathbf{C} ,\tag{3.6}$$

onde  $\mathbf{C}$  é uma matriz não-singular de ordem  $n \times n$ . Então, o modelo pode ser reescrito como

$$\mathbf{X} = \Lambda^* \boldsymbol{\xi}^* + \boldsymbol{\delta} ,\tag{3.7}$$

de onde obtemos

$$\text{Cov} (\boldsymbol{\xi}^*, \boldsymbol{\xi}^*) = \mathbf{C}^{-1} \Phi \mathbf{C} = \Phi^* .\tag{3.8}$$

Logo, existe uma transformação linear  $\mathbf{C}$  que, aplicada conforme descrito em (3.6), resulta em estimativas diferentes para os parâmetros de interesse do modelo.

É preciso fixar restrições que, mesmo que não garantam, ao menos possibilitem a identificabilidade do modelo. Nesse sentido, é comum que se adote uma restrição a respeito da matriz de covariâncias gerada pelo vetor de fatores de ordem  $n$ , para a qual se assume igualdade com a matriz identidade, isto é,

$$\Phi \equiv \text{Cov} (\boldsymbol{\xi}, \boldsymbol{\xi}) = \mathbf{I}_n .\tag{3.9}$$

Nesse caso, assume-se implicitamente que os fatores são ortogonais, uma vez que sua matriz de covariâncias ( $\Phi$ ) é diagonal.

Finalmente, uma restrição muito conveniente é a de que  $\Lambda' \Theta_\delta^{-1} \Lambda$  seja diagonal. Com essa restrição,  $\Lambda$  será determinado de forma única, desde que os elementos da diagonal de  $\Lambda' \Theta_\delta^{-1} \Lambda$  sejam diferentes entre si (Anderson, 1984, p.554).

### 3.2.1 Identificabilidade

O modelo de Análise Fatorial estará completamente identificado quando existir solução para o ajuste do modelo a partir dos dados observados e quando tal solução for única. Portanto, dados o número  $n$  de fatores que se assume para o modelo e a matriz de covariâncias  $\mathbf{S}$ , calculada a partir das  $q$  variáveis observadas, é preciso (Anderson, 1984, p.553):

1. Obter *alguma* trinca  $\{\Lambda, \Phi, \Theta_\delta\}$ , que satisfaça

$$\Sigma(\theta) = \Lambda \Phi \Lambda' + \Theta_\delta$$

2. Em caso de sucesso, verificar se a trinca é a *única* que satisfaz

$$\Sigma(\theta) = \Lambda \Phi \Lambda' + \Theta_\delta$$

Devido ao fato de qualquer trinca obtida poder ser transformada em uma equivalente, da forma  $\{\Lambda \mathbf{C}, \mathbf{C}^{-1} \Phi \mathbf{C}'^{-1}, \Theta_\delta\}$ , é preciso restringir o efeito da matriz  $\mathbf{C}$  por meio de  $n^2$  restrições, já que a matriz é de ordem  $n \times n$ .

Como as matrizes de covariâncias são simétricas, o número total de variâncias e covariâncias calculados na amostra e dispostos na matriz  $\mathbf{S}$  corresponde à quantidade de elementos abaixo ou acima da diagonal de  $\mathbf{S}$  (covariâncias), somada à quantidade de elementos na própria diagonal (variâncias), o que resulta em  $q(q+1)/2$  elementos distintos em uma matriz de ordem  $q$ . Além dessas estimativas para os parâmetros livres, devem ser contadas as  $n^2$  restrições já mencionadas.

Em relação ao número de parâmetros da trinca  $\{\Lambda, \Phi, \Theta_\delta\}$ , cada matriz possui, respectivamente,  $qn$ ,  $\frac{n(n+1)}{2}$  e  $q$  elementos. O número de elementos de  $\Lambda$  deve-se a

**Tabela 3.3:** Número mínimo de variáveis a serem observadas para possibilitar a identificabilidade de modelos com 1 a 10 fatores.

Nº de fatores	Nº mínimo de variáveis observadas	Nº de fatores	Nº mínimo de variáveis observadas
1	3	6	10
2	5	7	12
3	6	8	13
4	8	9	14
5	9	10	15

sua ordem, o de  $\Phi$  deve-se a sua simetria, por ser uma matriz de covariâncias, e o de  $\Theta_\delta$  deve-se a ela ser uma matriz diagonal.

Dessa forma, o valor

$$\frac{q(q+1)}{2} + n^2 - \left[ qn + \frac{n(n+1)}{2} + q \right] = \frac{1}{2} [(q-n)^2 - q - n] \quad (3.10)$$

representa o excesso do número de estimativas para os parâmetros livres, somado ao número de restrições, em relação ao número total de parâmetros da trinca  $\{\Lambda, \Phi, \Theta_\delta\}$ .

Caso esse excesso seja positivo, o modelo é denominado *sobre-identificado* e a solução de suas equações de ajuste, caso exista, será única, pois as informações observadas são mais numerosas que os parâmetros a estimar. Caso esse excesso seja negativo, o modelo é denominado *subidentificado* e não possui solução única, já que o número de informações disponíveis é insuficiente para que se selecione uma dentre as infinitas soluções possíveis para as equações do modelo. Por outro lado, caso o excesso seja nulo, temos que o número de informações obtidas com a amostra e as restrições predeterminadas corresponde exatamente ao número de parâmetros livres do modelo, de modo que é garantida a existência de um número finito de soluções para suas equações de ajuste.

A Tabela 3.3 apresenta o número mínimo de variáveis que devem ser observadas para que um modelo com um determinado número de fatores seja identificado ou sobre-identificado, calculado com base na expressão (3.10).

### 3.2.2 Interpretabilidade: rotações

Como evidenciou a equação (3.6), em um modelo de Análise Fatorial, a trinca de parâmetros  $\{\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{\Theta}_\delta\}$  pertence a um conjunto de infinitas trincas equivalentes no que diz respeito ao ajuste do modelo. As restrições de identificabilidade, normalmente feitas a respeito de  $\mathbf{\Phi}$  e de  $\mathbf{\Lambda}'\mathbf{\Theta}_\delta^{-1}\mathbf{\Lambda}$ , são escolhas relativamente arbitrárias e podem levar a uma solução que não possui interpretação prática. Entretanto, é possível escolher uma matriz  $\mathbf{C}$  conveniente, de modo que a transformação linear da solução, na forma da equação (3.6), resulte em um conjunto de parâmetros de maior interpretabilidade. Como tais transformações rotacionam os vetores paramétricos no hiperespaço em que se encontram, são mais comumente chamadas de *rotações*.

As rotações visam, portanto, à maximização da interpretabilidade do vetor de parâmetros estimados. Normalmente, considera-se que uma matriz de cargas  $\mathbf{\Lambda}$  apresenta estimativas interpretáveis quando cada variável observada possui carga mais expressiva em apenas um fator, sendo as cargas dos demais fatores pouco representativa. Se outra variável tiver carga elevada do mesmo fator, pode-se dizer que elas são variáveis mais afetadas pelo mesmo fator e compõem um conjunto de medidas observáveis para o fator latente a que estão associadas. Da mesma forma, é interessante que as demais variáveis estejam, cada uma, mais fortemente associadas a um fator específico, de modo que seja possível identificar grupos de variáveis mais associadas a um fator ou outro.

Na tentativa de maximizar a diferença entre o valor de uma das cargas de cada variável e suas demais cargas, várias rotações foram propostas. Basicamente, a um grupo de rotações *ortogonais* e um de *oblíquas*. Elas sempre impactam as matrizes  $\mathbf{\Lambda}$  e  $\mathbf{\Phi}$ , conforme mostra as equações (3.6), (3.7) e (3.8), mas as rotações ortogonais preservam a ortogonalidade da matriz  $\mathbf{\Phi}$ , de modo que os fatores permanecem independentes, conforme fixado na restrição de identificabilidade descrita pela equação (3.9). As rotações oblíquas, por sua vez, permitem que os fatores assumam posições não-ortogonais, deixando de ser independentes. As rotações oblíquas, portanto, só devem ser aplicadas quando não for postulada independência para os fatores.

## rotações ortogonais

Os critérios mais populares para se obter matrizes de rotação ortogonais são provavelmente o *quartimax* e o *varimax* (Krzanowski & Marriott, 1995). Cada método busca as cargas que maximizam uma determinada expressão, sendo a expressão maximizada pelo critério *quartimax* (Neuhaus & Wrigley, 1954) dada por

$$R_{\text{quartimax}} = \sum_{i=1}^q \sum_{j=1}^n (\lambda_{ij}^x)^4 - \frac{1}{qn} \left[ \sum_{i=1}^q \sum_{j=1}^n (\lambda_{ij}^x)^2 \right]^2 .$$

A expressão maximizada pelo critério *varimax*, por sua vez, é dada por

$$R_{\text{varimax}} = \frac{1}{q^2} \sum_{j=1}^n \left\{ n \sum_{i=1}^q \kappa_{ij}^4 - \left[ \sum_{i=1}^q \kappa_{ij}^2 \right]^2 \right\} ,$$

onde  $\kappa_{ij} = \lambda_{ij} / \left( \sum_{j=1}^n \lambda_{ij}^2 \right)^{1/2}$  (Kaiser, 1958).

Uma breve análise das expressões mostra que sua maximização procura justamente aumentar a variância entre as cargas, apresentando o critério *varimax* uma correção do valor das cargas que o torna frequentemente o mais eficaz (Marcoulides & Hershberger, 1997).

## rotações oblíquas

Quanto às rotações oblíquas, os critérios mais populares para se obter matrizes de rotação oblíquas são, provavelmente, o *promax* e o *oblimin*. O primeiro consiste na aplicação de uma rotação *varimax*, seguido da busca de uma rotação oblíqua que aproxime as cargas obtidas a uma determinada potência das cargas originais. O segundo, por sua vez, constitui um conjunto de rotações que visam à minimização da covariância entre os fatores oblíquos. (Harman, 1976, cap.14). Alternativamente, é possível buscar a solução que melhor se aproxima de uma especificação dada *a priori* utilizando-se análise de *procrustes* (Meredith, 1977).

### 3.2.3 Componentes de variância

Como consequência da restrição (3.9), obtemos fatores com variâncias unitárias e independentes entre si, que permitem o cálculo de uma medida conhecida como *comunalidade*. Ela expressa a parcela da variabilidade total de cada variável observada

que pode ser explicada pelos fatores comuns. Para a variável  $X_i$ , temos sua comunalidade expressa por

$$h_i = \sum_{j=1}^n (\lambda_{ij}^x)^2 . \quad (3.11)$$

Por exemplo, a comunalidade da variável  $X_2$  no modelo (3.3) é dada por

$$\mathbb{V} (\lambda_{21}^x \xi_1 + \lambda_{22}^x \xi_2) = (\lambda_{21}^x)^2 + (\lambda_{22}^x)^2 ,$$

em decorrência de  $\mathbb{V} (\xi_1) = \mathbb{V} (\xi_2) = 1$ .

Em contraposição à comunalidade, é comum denominar a variabilidade não capturada pelos fatores comuns e expressas em  $\delta$  como *unicidades*. De fato, para cada variável observada, essa componente representa a parcela da variância total que é “única” à variável, não sendo captada pelos fatores comuns. Sendo assim, um modelo de Análise Fatorial, será tanto melhor, quanto maiores forem as comunalidades de suas variáveis ou, em outras palavras, quanto menores forem suas unicidades (Anderson, 1984, Cap. 14).

É possível fazer com que as componentes de comunalidade e unicidade variem no intervalo  $[0, 1]$ , desde que se obtenha  $\mathbb{V} (X_i) = 1 \quad \forall i \in \{1, 2, \dots, q\}$ . Por ser comum a padronização de variáveis na aplicação da Análise Fatorial, esse requisito é atendido e obtemos

$$\mathbb{V} (X_i) = h_i + \delta_i = 1 ,$$

de onde segue que a unicidade pode ser expressa por sua fórmula mais difundida (Harman, 1976, Cap. 2):

$$\delta_i = 1 - h_i .$$

Finalmente, é possível decompor a unicidade como

$$\delta_i = \delta_i^s + \delta_i^e ,$$

com  $\delta_i^s$  e  $\delta_i^e$  denominados, respectivamente, *especificidade* e *erro de mensuração*. A especificidade é a parte do erro que pode ser atribuída a fatores que não participaram do modelo e que, se fossem incluídos, explicariam a total variabilidade das observações. O erro de mensuração, por sua vez, representa o erro atribuído à imprecisão decorrente do processo de aferição das observações. Contudo, a maior parte das aplicações

de Análise Fatorial não contemplam a desagregação da unicidade, por não serem separadamente identificáveis. A identificação pode ser obtida, entretanto, ao realizar observações replicadas do mesmo indivíduo amostrado (Skrondal & Rabe-Hesketh, 2004, p.67).

### 3.2.4 Estimação dos parâmetros do modelo de Análise Fatorial

Há pelo menos dois métodos de estimação mais utilizados no ajuste de modelos de Análise Fatorial: o método da Máxima Verossimilhança e o de Mínimos Quadrados.

A estimação pelo método da Máxima Verossimilhança implica na pressuposição de normalidade multivariada para as variáveis observadas e, conseqüentemente, para o vetor de fatores e para o vetor de unicidades. Desenvolvido mais tardiamente, o método da Máxima Verossimilhança apresenta uma exigência computacional muito maior que o de Mínimos Quadrados sendo, por outro lado, um método que apresenta propriedades interessantes, como eficiência e consistência assintóticas, invariância e a possibilidade de formulação de testes de razão de verossimilhança.

Para sua aplicação no ajuste do modelo (3.4), define-se a função de verossimilhança  $L(\cdot)$ , para uma amostra aleatória simples de tamanho  $N$  do vetor  $\mathbf{X}$ , como

$$L(\mathbf{\Lambda}, \mathbf{\Theta}_\delta) = (2\pi)^{-Nq/2} |\mathbf{\Sigma}|^{-N/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^N \mathbf{x}_i' \mathbf{\Sigma}^{-1} \mathbf{x}_i \right\},$$

com  $\mathbf{\Sigma}$ , conforme (3.5). A função de log-verossimilhança  $\ell(\cdot) = \log L(\cdot)$  é dada por

$$\ell(\mathbf{\Lambda}, \mathbf{\Theta}_\delta) = \frac{Nq}{2} \log(2\pi) - \frac{N}{2} \log |\mathbf{\Sigma}| - \frac{N}{2} \text{tr}(\mathbf{\Sigma}^{-1} \mathbf{S}),$$

onde  $\mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ . Maximizando  $\ell(\cdot)$  com respeito a  $\mathbf{\Lambda}$  e  $\mathbf{\Theta}_\delta$ , obtemos o sistema

$$\begin{cases} \left( \hat{\mathbf{\Theta}}_\delta^{-1/2} \mathbf{S} \hat{\mathbf{\Theta}}_\delta^{-1/2} \right) \left( \hat{\mathbf{\Theta}}_\delta^{-1/2} \hat{\mathbf{\Lambda}} \right) = \left( \hat{\mathbf{\Theta}}_\delta^{-1/2} \hat{\mathbf{\Lambda}} \right) \left( \mathbf{I} + \hat{\mathbf{\Lambda}}' \hat{\mathbf{\Theta}}_\delta^{-1} \hat{\mathbf{\Lambda}} \right) \\ \hat{\mathbf{\Theta}}_\delta = \text{diag} \left( \hat{\mathbf{\Lambda}} \hat{\mathbf{\Lambda}}' - \mathbf{S} \right) \end{cases}, \quad (3.12)$$

que pode ser resolvido iterativamente.

Essa abordagem dada por Lawley (1940), remonta à década de 1940. Como o procedimento iterativo em questão converge muito lentamente e as estimativas de

alguns elementos de  $\Theta_\delta$  podem tender a zero ou a estimativas negativas chamadas de *Heywood cases* (Heywood, 1931), a estimação por máxima verossimilhança no contexto da Análise Fatorial tornou-se pouco atrativa, salvo para problemas de pequena magnitude, até o trabalho de Jöreskog (1969), em que a segunda equação do sistema (3.12) é substituída pela maximização da log-verossimilhança, como uma função de  $\Theta_\delta$ , utilizando-se o algoritmo de Fletcher-Powell, baseado em uma aproximação de segunda ordem de  $\ell$ . Contudo, devido às limitações computacionais da época, até a década de 1970 a estimação dos modelos de Análise Fatorial eram eminentemente baseadas em métodos de Mínimos Quadrados, em vez dos de Máxima Verossimilhança.

Aplicados aos modelos de Análise Fatorial, os métodos de Mínimos Quadrados visam à obtenção de valores para os parâmetros por meio da minimização de uma soma de resíduos existentes entre os valores efetivamente observados da matriz de covariâncias  $\mathbf{S}$  e os valores estimados por meio de  $\Sigma(\theta)$ . Uma função de resíduos comumente utilizada é

$$R_{\text{MMQ}_1} = \text{tr} [(\mathbf{S} - \Sigma)^2] ,$$

cujas derivadas parciais, igualadas a zero, formam o sistema

$$\begin{cases} (\mathbf{S} - \hat{\Theta}_\delta) \hat{\Lambda}^x = \hat{\Lambda}^x (\hat{\Lambda}^{x'} \hat{\Lambda}^x) \\ \hat{\Theta}_\delta = \text{diag} (\mathbf{S} - \hat{\Lambda}^x \hat{\Lambda}^{x'}) \end{cases} . \quad (3.13)$$

O sistema (3.13) apresenta uma resolução computacionalmente menos intensiva que o do método da Máxima Verossimilhança. Além disso, suas equações sugerem um algoritmo iterativo intuitivo:

1. Toma-se uma estimativa inicial de  $\hat{\Theta}_\delta$ ;
2. Preenchem-se as colunas de  $\hat{\Lambda}^x$  com os autovetores de  $\mathbf{S} - \hat{\Theta}_\delta$ ;
3. Obtem-se uma atualização para  $\hat{\Theta}_\delta$  a partir da segunda equação do sistema;
4. Retorna-se a (1) até a convergência.

Esse método consagrado recebeu o nome de *fator principal* devido à sua semelhança com a decomposição realizada na análise de componentes principais, que constitui um caso particular desse esquema geral, obtido quando  $\Theta_\delta = \mathbf{0}$  e resolvido em apenas um passo.

Vale notar que foram propostas generalizações para a soma de resíduos, tais como

$$R_{\text{MMQ}_2} = \text{tr} [(\mathbf{S} - \mathbf{\Sigma}) \mathbf{\Theta}_\delta^{-1}]^2 \text{ e}$$

$$R_{\text{MMQ}_3} = \text{tr} [(\mathbf{S} - \mathbf{\Sigma}) \mathbf{S}^{-1}]^2 ,$$

de onde se obtêm sistemas similares que, resolvidos, fornecem estimativas para o modelo (Anderson, 1984).

### 3.2.5 Estimação dos escores dos fatores

Uma importante etapa da análise de um modelo de Análise Fatorial é a estimação dos escores dos fatores. A estimação tratada anteriormente refere-se aos *parâmetros estruturais* do modelo  $(\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{\Theta}_\delta)$ , presentes nas equações que relacionam os fatores latentes e as variáveis observadas. Contudo, uma vez que os valores dos próprios fatores são elementos desconhecidos, também podem considerados objetos de estimação, atuando como *parâmetros incidentais* ( $\xi$ ). Diferentemente dos parâmetros estruturais, os incidentais aumentam conforme a amostra aumente, já que dizem respeito a características dos indivíduos selecionados para a amostra. No caso da Análise Fatorial, cada fator é composto por um conjunto de escores latentes, um para cada indivíduo da amostra. Os fatores funcionam como covariáveis do modelo (3.4) e os escores equivalem às “observações” que seriam efetuadas a respeito dos fatores, caso não fossem latentes.

A estimação simultânea dos parâmetros estruturais e incidentais, entretanto, é impossível (Anderson, 1984, Seç. 14.4). Para estimar os escores dos fatores, portanto, é preciso já ter estimado os parâmetros estruturais. Além disso, para sua estimação, os fatores devem deixar de ser considerados vetores aleatórios normais multivariados. Em vez disso, são considerados fixos e abordados como parâmetros desconhecidos em uma função de verossimilhança que utilizará os parâmetros estruturais previamente estimados como valores fixos.

Duas abordagens para a estimação dos fatores são mais comumente utilizadas Krzanowski & Marriott (1995) e fornecem estimativas que são funções lineares das variáveis manifestas. O método de Bartlett (1937) consiste, basicamente, de uma abordagem de mínimos quadrados, que visa a minimizar a diferença entre os valores

preditos pelo modelo, a partir dos parâmetros estruturais previamente estimados, e as variáveis efetivamente observadas. Nessa abordagem, a quantidade a ser minimizada para o indivíduo  $i$  é

$$R_{\text{Bartlett}} = \left( \mathbf{x}_i - \hat{\Lambda}_x \boldsymbol{\xi}_i \right)' \hat{\Theta}_\delta^{-1} \left( \mathbf{x}_i - \hat{\Lambda}_x \boldsymbol{\xi}_i \right) ,$$

Um cálculo direto resulta no estimador

$$\hat{\boldsymbol{\xi}}_i^{\text{Bartlett}} = \left( \hat{\Lambda}'_x \hat{\Theta}_\delta^{-1} \hat{\Lambda}_x \right)^{-1} \hat{\Lambda}'_x \hat{\Theta}_\delta^{-1} \mathbf{x}_i . \quad (3.14)$$

Thomson (1951), por outro lado, assume normalidade multivariada para  $\mathbf{Y}$  e  $\boldsymbol{\delta}$ , aglutina as variáveis latentes e manifestas em um mesmo vetor e, utilizando propriedades da distribuição, obtém a esperança condicionada de  $\boldsymbol{\xi}$  dado  $\mathbf{X} = \mathbf{x}_i$ ,

$$\mathbb{E} (\boldsymbol{\xi} | \mathbf{X} = \mathbf{x}_i) = \Lambda'_x (\Lambda_x \Lambda'_x + \Theta_\delta)^{-1} \mathbf{x}_i ,$$

como em uma *regressão*, de onde temos o nome deste segundo tipo de estimativa para os fatores. Para o método da regressão, a estimativa óbvia é

$$\hat{\boldsymbol{\xi}}_i^{\text{Regressão}} = \hat{\Lambda}'_x \left( \hat{\Lambda}_x \hat{\Lambda}'_x + \hat{\Theta}_\delta \right)^{-1} \mathbf{x}_i . \quad (3.15)$$

### 3.2.6 As duas modalidades de Análise Fatorial

Há pelo menos dois usos consagrados para o modelo de análise fatorial. No âmbito das ciências sociais, por exemplo, é comum que não se conheça a estrutura latente por trás dos dados coletados, o que leva o pesquisador a valer-se de técnicas resumitivas, que evidenciem a relação entre as variáveis pesquisadas, de modo a agrupar as similares, eliminar as redundantes, contrastar as antagônicas etc.

Para tal fim, a Análise de Componentes Principais apresenta-se como uma técnica simples e eficaz, mas a Análise Fatorial inclui e expande a funcionalidade desse método, permitindo rotações das soluções obtidas, ampliando o poder explicativo e a compreensão das variáveis estudadas. Por outro lado, no âmbito das ciências do comportamento é comum que já se tenha formulado um construto teórico antes da coleta de dados e deseje-se validar o instrumento de mensuração utilizado para aferir o traço latente em estudo.

Os dois usos, no entanto, ainda que diverjam na finalidade, possuem formulações idênticas no que concerne ao interesse em elucidar as relações matemáticas existentes

entre variáveis observadas e latentes. Para discriminar um uso do outro, é costumeiro chamar de *Análise Fatorial Exploratória* a modalidade referente à redução da dimensionalidade dos dados levantados, ao passo que se costuma denominar de *Análise Fatorial Confirmatória* a modalidade devotada a avaliar uma estrutura latente descrita *a priori*.

Na Análise Fatorial Exploratória, a intenção do pesquisador reside na compreensão da estrutura latente dos dados observados, por meio da descoberta do número de fatores subjacentes e a forma com que interferem nas variáveis observadas. Na Análise Fatorial Confirmatória, o interesse do analista é confrontar modelos diferentes para o mesmo conjunto de dados e, não raro, estimar os escores dos fatores do modelo.

Ao passo que a Análise Fatorial Confirmatória requer a especificação prévia de detalhes do modelo fatorial, como parâmetros de escala, relações de causalidade entre variáveis latentes e observadas ou o agrupamento de variáveis correlacionadas, a pouca definição característica da Análise Fatorial Exploratória impõe o estabelecimento de diversas restrições arbitrárias aos parâmetros, a fim de possibilitar a identificabilidade do modelo, além de lançar mão de rotações das estimativas originais, a fim de aumentar sua interpretabilidade.

Neste Capítulo, apenas a Análise Fatorial Exploratória será abordada. A Análise Fatorial Confirmatória será oportunamente apresentada no Capítulo 4, onde serão abordados os modelos de Equações Estruturais, dos quais ela constitui caso particular.

### 3.2.7 *Softwares* e exemplo de resolução

Os insumos de uma Análise Fatorial Exploratória são as  $q$  variáveis observadas para cada elemento amostrado (ou sua matriz de covariâncias ou de correlações) e o número provável  $n$  de fatores do modelo, sendo desejável que  $n \ll q$ . Tendo esses elementos em mãos, a aplicação da Análise Fatorial Exploratória envolve basicamente cinco passos, a saber:

**Extração inicial dos fatores** Um primeiro ajuste é realizado, com o número de fatores postulado para o modelo. O número máximo de fatores pode ser utilizado quando o número provável de fatores for desconhecido.

**Determinação do número de fatores significativos a reter no modelo** Identificação

do número de fatores que acumula a maior parte da variabilidade do modelo.

**Rotação para uma solução final** Aplicação de rotações ortogonais ou oblíquas, conforme adequado, a fim de encontrar estimativas mais interpretáveis.

**Interpretação da solução rotacionada** Interpretação da solução obtida após as rotações.

**Estimação dos escores** Estimação dos escores, quando a intenção for a mensuração das variáveis latentes que os fatores representam. Essa estimação depende da disponibilidade das observações realizadas para cada elemento da amostra, sendo impossível executá-la a partir da matriz de covariância ou correlação isoladamente.

A Análise Fatorial Exploratória é uma importante técnica de análise multivariada e está implementada em diversos dos *softwares* estatísticos mais proeminentes da academia e do mercado. No SAS, o procedimento FACTOR realiza a estimação dos parâmetros e permite diversas rotações. No R ([R Development Core Team, 2010](#)), a função `factanal()` também permite o ajuste do modelo e as funções `varimax` e `promax`, disponíveis na biblioteca padrão `stat`, permitem a estimação da solução rotacionada. A biblioteca `GPARotation` ([Bernaards & I.Jennrich, 2005](#)) possui outras funções de rotação das soluções, como `oblimin` e `quartimax`.

Prosseguindo com o exemplo cujos dados figuram na Tabela 3.2, a Tabela 3.4 exhibe as cargas estimadas para o modelo ajustado considerando 1 ou 2 fatores. No modelo de 1 fator, as comunalidades somam 2,44, o que representa 41% da variância total de 6,00, referente à soma de variâncias das 6 variáveis observadas. O teste  $\chi^2$ , entretanto, apresentou um  $p$ -valor  $\approx 0$ , indicando um valor significativo para a distância entre os valores observados e os preditos pelo modelo ( $\chi_{\text{obs}}^2 = 75,18$ ). Estendendo o modelo para um de 2 fatores, obtemos uma soma de  $2,42 + 1,16 = 3,58$  para as comunalidades, o que representa 60% da variância total. Nesse caso, o valor observado para a estatística  $\chi^2$  foi de  $\chi_{\text{obs}}^2 = 6,11$ , apresentando um  $p$ -valor de 0,19, o que indica uma boa descrição dos dados pelo modelo, com apenas 2 fatores. Com relação aos outros elementos da trinca  $\{\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{\Theta}_\delta\}$ , os valores da diagonal de  $\mathbf{\Theta}_\delta$  para cada variável estão exibidos na saída do *software*, disponível no Apêndice deste

**Tabela 3.4:** Extração da matriz e covariâncias exibida na Tabela 3.2. Utilizou-se a função `factanal()` do R para a extração de fatores.

Variável	Cargas (1 fator)		Cargas (2 fatores)	
	Fator 1	Fator 1	Fator 1	Fator 2
general	0,68		0,65	0,35
picture	0,38		0,35	0,54
blocks	0,50		0,47	0,75
maze	0,30		0,25	0,41
reading	0,88		0,96	-0,14
vocab	0,85		0,82	
Comunalidades	2,44		2,42	1,16
Proporção da variância	41%		40%	19%
Variância acumulada	41%		40%	60%
$\chi^2$	75,18		6,11	
Graus de liberdade	9		4	
$p$ -valor	$1,46 \times 10^{-12}$		0,19	

Capítulo (Seção 3.5) e  $\Phi = \mathbf{I}_n$  devido à restrição para identificabilidade.

Tendo sido fixado o número de fatores do modelo e estimados os parâmetros livres, podemos rotacionar a solução encontrada para a matriz de cargas, a fim de obter uma solução mais interpretável. A Tabela 3.5 apresenta as cargas rotacionadas pelos métodos *promax*, *varimax* e *oblimin*. Dentre elas, a rotação *varimax* é a única ortogonal e apresenta fatores com cargas mais interpretáveis que a solução não rotacionada:

- parece haver uma influência parecida de ambos fatores na variável *general*;
- parece haver uma influência maior do Fator 1 nas variáveis *reading* e *maze*; e
- parece haver uma influência maior do Fator 2 nas demais variáveis.

As outras rotações apresentadas são oblíquas e salientam ainda mais essa interpretação. Cabe notar que a solução rotacionada pelo método *promax* faz com que o

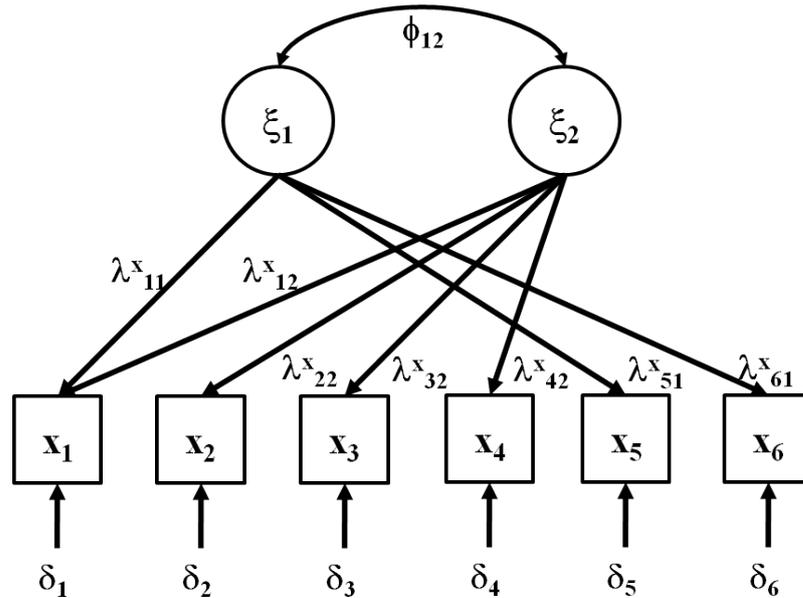
**Tabela 3.5:** Rotações da matriz de cargas apresentada na Tabela 3.4. As caselas vazias representam valores praticamente nulos.

Variável	Rotação promax		Rotação varimax		Rotação oblimin	
	Fator 1	Fator 2	Fator 1	Fator 2	Fator 1	Fator 2
general	0,36	0,47	0,50	0,54	0,39	0,48
picture		0,67	0,16	0,62		0,65
blocks		0,93	0,21	0,86		0,90
maze		0,51	0,11	0,47		0,49
reading	1,02		0,96	0,18	0,99	
vocab	0,81		0,79	0,23	0,79	
Comunalidades	1,85	1,81	1,86	1,72	1,76	1,69
Proporção da variância	0,31	0,30	0,31	0,29	0,29	0,28
Variância acumulada	0,31	0,61	0,31	0,60	0,29	0,57

modelo responda por 61% da variância total, sendo esta a melhor das três soluções rotacionadas apresentadas, segundo esse critério. Contudo, a carga de 1,02 do Fator 1 na variável *reading* representa um possível caso *Heywood*, em que a comunalidade de uma variável supera o valor unitário, fazendo com que a variância (unicidade) dessa variável precise assumir um valor negativo. Conferindo a saída do R, entretanto, verificamos que a unicidade calculada está conforme, com o valor positivo de 0,052. A Figura 3.2 apresenta o novo diagrama de caminho, do modelo ajustado, que pode ser comparado com o do modelo inicial, com apenas um fator (Figura 3.1).

### 3.2.8 Limitações

Conforme apresentado nesta Seção, a Análise Fatorial representa o mais tradicional modelo de mensuração conhecido no meio estatístico. No entanto, a questão da falta de identificabilidade do modelo desencadeou uma verdadeira celeuma entre vários estatísticos, devido à arbitrariedade, tanto das restrições impostas para possibilitar a identificabilidade, quanto das rotações aplicadas às cargas estimadas. Esse perigo, entretanto, só se apresenta na Análise Fatorial Exploratória e [Armstrong \(1967\)](#) apre-



**Figura 3.2:** Modelo de Análise Fatorial com 2 fatores não-ortogonais, referente a notas obtidas em 6 avaliações aplicadas em 112 alunos (Bartholomew & Knott, 1990).

sentou um exemplo em que os fatores latentes eram conhecidos e um modelo de Análise Fatorial Exploratória, ainda que bem ajustado, falhava sistematicamente em recuperar os valores conhecidos dos fatores. Grice (2001) também publicou uma revisão bibliográfica abrangente a respeito deste fenômeno, ao qual chama de *indeterminação dos fatores*, e alerta para a necessidade de se compreender a estrutura dos fatores, para que se possa validar e utilizar os escores estimados em análises posteriores. Por outro lado, na mesma linha de pesquisa, Beauducel (2007) analisa as matrizes de covariância produzidas pelos escores estimados e conclui que os métodos mais utilizados, entre eles os de Bartlett e o método da regressão, reproduzem a mesma matriz. De qualquer forma, a análise exploratória de fatores latentes, portanto, não deve preceder a teoria sobre eles e a Análise Fatorial Confirmatória representa um modelo mais consistente para a modelagem de variáveis latentes, ainda que tenha o custo de exigir uma especificação mais detalhada do modelo.

Outro aspecto que merece atenção é a utilização da Análise Fatorial na modelagem de variáveis latentes mensuradas por meio de observações categóricas, em vez de contínuas. Na verdade, uma escala de poucos níveis não pode ser considerada contínua

e há modelos mais adequados para lidar com esse tipo de observação, como a Teoria de Resposta ao Item ou alguma extensão da Análise Fatorial para dados categóricos (King, 1986, 682-683). Menos conhecidas, essas ferramentas acabam tendo seu espaço tomado pela Análise Fatorial clássica, para respostas contínuas linearmente relacionadas aos fatores, conduzindo a resultados eventualmente pouco satisfatórios e conclusões duvidosas, devido à negligência de um pressuposto elementar do modelo utilizado (Cooper, 1983).

Além disso, King (1986) alerta quanto a uma interpretação errônea, de que as variáveis latentes eliciadas pelos dados analisados constituiriam os efeitos das variáveis observadas, quando, na realidade, representam suas causas. Por exemplo, pode-se observar variáveis como raça ou gênero e identificar uma estrutura latente como *ideologia política* por trás delas. Não parece razoável, contudo, que uma ideologia política determine a raça ou o gênero de quem quer que seja, mas o contrário.

Finalmente, duas limitações decorrentes da definição do modelo são:

- a linearidade das relações entre variáveis latentes e observadas; e
- a continuidade das variáveis observadas.

Tais limitações, entretanto, já estão superadas por outros modelos como o de Análise Fatorial Não-linear (Yalcin & Amemiya, 2001) e os modelos de mensuração para variáveis latentes contínuas com observações categorizadas, como os da Teoria de Resposta ao Item, apresentados a seguir.

### 3.3 Teoria de Resposta ao Item (1960)

Como com a Análise Fatorial, a origem da Teoria de Resposta ao Item deu-se no contexto da psicometria. É possível, no entanto, retroceder a pelo menos duas localidades originárias distintas, uma nos Estados Unidos e outra na Europa.

Nos Estados Unidos, costuma-se atribuir o surgimento da TRI a Lord & Novick (1968). A obra destes autores contém quatro capítulos a respeito de TRI, escritos por Allan Birnbaum, e representa um marco na evolução dos métodos psicométricos por unificarem, com rigor, o tratamento estatístico da teoria sobre testes e por estarem os autores engajados na difusão dessa nova teoria, por meio da promoção de

eventos científicos e da aplicação da teoria na resolução de problemas reais. Coube a [Bock \(1972\)](#), auxiliado por diversos alunos seus, o desenvolvimento dos primeiros algoritmos efetivos para a estimação de parâmetros dos modelos da TRI<sup>Na</sup> Europa, diversos modelos da TRI surgiram com os estudos de [Rasch \(1960\)](#), que se interessou particularmente pelas propriedades científicas dos modelos de mensuração. Em Viena, [Fischer \(1973\)](#) estendeu o modelo para dados binários de [Rasch \(1960\)](#), de modo a incorporar considerações psicológicas nos parâmetros ([Embretson & Reise, 2000](#)).

Na TRI, diferentemente da Análise Fatorial, não se modela a estrutura de covariâncias dos dados observados, mas a probabilidade de ocorrência da resposta efetiva a cada item observado, em termos da variável latente não-observada. Além disso, em vez de *fatores*, as variáveis latentes costumam receber o nome de *traços latentes*. O histórico do surgimento da TRI evidencia o distanciamento do seu desenvolvimento em relação ao da Análise Fatorial e isso explica algumas das diferenças de nomenclatura e formato de concepção dessa que, embora outra teoria, trata-se de um modelo de mensuração, do mesmo modo que a Análise Fatorial.

A TRI pode ser usada, por exemplo, para modelar os dados da Tabela 3.6, que mostra a frequência das respostas dadas a uma pesquisa de atitude em relação ao aborto ([McGrath & Waterton, 1986](#)), cujo conjunto de dados está disponível na biblioteca `ltm` ([Rizopoulos, 2006](#)) do R, sob o nome `Abortion`. Os respondentes foram inquiridos se concordavam que a lei deveria permitir o aborto nas circunstâncias apresentadas em cada item:

**Item 1** A mulher decide por conta própria que não deseja ter a criança.

**Item 2** O casal concorda que não deseja ter a criança.

**Item 3** A mulher não é casada e não deseja casar com o pai da criança.

**Item 4** O casal não tem condições financeiras de ter mais um filho.

A última linha da Tabela 3.6 apresenta o total de respondentes que endossaram cada item. A partir dela, podemos concluir que o Item 1 é o que tem menor probabilidade de endosso, contanto com apenas 166 (44%) respondentes favoráveis, ao passo que o Item 4 é o de maior probabilidade, com 234 (62%) respostas afirmativas. Além disso, nota-se que os conjuntos de resposta mais frequentes foram: (1, 1, 1, 1), representando 131

**Tabela 3.6:** Frequências de respostas dadas ao questionário de atitude em relação ao aborto (McGrath & Waterton, 1986).

Resposta				Frequência
Item 1	Item 2	Item 3	Item 4	
0	0	0	0	103
0	0	0	1	13
0	0	1	0	10
0	0	1	1	21
0	1	0	0	9
0	1	0	1	6
0	1	1	0	7
0	1	1	1	44
1	0	0	0	1
1	0	0	1	0
1	0	1	0	0
1	0	1	1	6
1	1	0	0	3
1	1	0	1	3
1	1	1	0	12
1	1	1	1	141
166	225	241	234	379

(37%) das respostas;  $(0, 0, 0, 0)$ , representando 103 (27%) das respostas; e  $(0, 1, 1, 1)$ , representando 44 (12%) das respostas.

Daí conclui-se que quem é desfavorável a qualquer um dos itens, costuma ser desfavorável também aos demais. Já entre os que são favoráveis aos Itens 2, 3 ou 4, não raro são desfavoráveis ao Item 1, ainda que tendam a ser favoráveis a todos os demais. Dessa forma, podemos concluir que a maioria dos respondentes endossa ou todos ou nenhum dos itens, sendo o Item 1 polêmico. Esse tipo de análise das características dos itens é possibilitado de forma mais profunda pelos modelos da TRI. Inclusive, ela possibilita a estimação do valor do traço latente de cada respondente a partir do seu vetor de respostas.

O fato de os modelos da TRI constituírem uma alternativa à Análise Fatorial, nos casos em que as variáveis observadas não forem contínuas, é apenas uma das diferenças entre esses dois modelos. Essa distinção básica faz com que os modelos da TRI não sejam lineares e, em vez da distribuição normal atribuída às variáveis observadas no contexto da Análise Fatorial, as observações utilizadas na TRI seguem uma distribuição multinomial. Na TRI, portanto, modela-se a probabilidade de a variável-resposta assumir uma das categorias possíveis, dado o valor desconhecido da variável latente associado ao indivíduo amostrado. Neste trabalho, serão abordados os modelos mais tradicionais da TRI, que lidam com apenas um traço latente e são, por isso, chamados *unidimensionais*.

### 3.3.1 Modelos cumulativos unidimensionais

O modelo da TRI mais difundido para respostas dicotômicas é o de [Rasch \(1960\)](#), posteriormente aprimorado por [Wright & Stone \(1979\)](#) e [Fischer \(1995\)](#). Também chamado de modelo logístico de 1 parâmetro (ML1), ele define a probabilidade de endosso ao item  $i$  como

$$\Pr(x_{ij} = 1|\xi_j) = \frac{\exp(\xi_j - \beta_i)}{1 + \exp(\xi_j - \beta_i)}, \quad (3.16)$$

em que  $\beta_i$  é comumente chamado de *parâmetro de dificuldade* do item  $i$ . Quanto maior esse parâmetro, maior será o valor mínimo do traço latente  $\xi_j$  do respondente  $j$ , para que a probabilidade de endosso seja maior que a de não-endosso, isto é, para

que  $\Pr(x_{ij} = 1|\xi_j) > 0,5$ . Dessa afirmação, podemos deduzir uma das principais características positivas desse modelo: a escala do traço latente é a mesma do parâmetro de dificuldade do item.

No contexto de avaliações educacionais, por exemplo, esse modelo é frequentemente utilizado para a criação de bancos de itens calibrados, isto é, itens que têm seus parâmetros estimados. Conjuntos desses itens são posteriormente utilizados para compor avaliações que medirão a proficiência dos respondentes, na mesma escala de calibração dos itens. Por exemplo, um aluno com proficiência 1,5 terá maior probabilidade de acertar itens com dificuldade 1,5 ou menor, ao passo que estará mais propenso a errar os itens com dificuldades superiores a 1,5. A mesma propriedade é válida para o modelo logístico de 2 parâmetros (ML2), introduzido por [Birnbaum \(1968\)](#):

$$\Pr(x_{ij} = 1|\xi_j) = \frac{\exp[\alpha_i(\xi_j - \beta_i)]}{1 + \exp[\alpha_i(\xi_j - \beta_i)]}. \quad (3.17)$$

No ML2,  $\alpha_i$  é comumente chamado de *parâmetro de discriminação* do item  $i$ . Quanto maior esse parâmetro, melhor a estimativa dos traços latentes próximos a  $\beta_i$ . Os itens com baixa discriminação podem ser endossados por indivíduos que não possuem traço latente alto o suficiente para o endosso, assim como podem deixar de ser endossados por indivíduos com traço latente alto. A discriminação de um item, portanto, é uma medida da qualidade do item no que diz respeito ao seu poder discriminante dos respondentes que devem ou não endossá-lo.

Finalmente, o modelo logístico de 3 parâmetros (ML3) contém um parâmetro muito útil no contexto de avaliações educacionais: o *parâmetro de acerto ao acaso*, representado por  $c_i$ :

$$\Pr(x_{ij} = 1|\xi_j) = c_i + (1 - c_i) \frac{\exp \alpha_i(\xi_j - \beta_i)}{1 + \exp \alpha_i(\xi_j - \beta_i)} \quad (3.18)$$

Em itens de múltipla escolha, por exemplo, é comum que os alunos acertem a resposta a determinados itens, não por conhecerem o conteúdo, mas por assinalarem ao acaso a alternativa certa.

### 3.3.2 Estimação dos parâmetros estruturais e incidentais

Na TRI, assim como na Análise Fatorial, os parâmetros a se estimar dividem-se em *estruturais* e *incidentais*. Os parâmetros estruturais são os referentes às característi-

cas dos itens ( $\alpha_i$ ,  $\beta_i$  e  $c_i$ ). Os parâmetros incidentais, por sua vez, referem-se ao valor do traço latente ( $\xi_j$ ) de cada indivíduo da amostra. Na TRI, os parâmetros estruturais costumam ser estimados item-a-item e, em seguida, estimam-se os parâmetros incidentais.

A TRI também é afetada pela indeterminação da escala do traço latente. Na Seção 3.4, é feita uma discussão mais ampla a respeito das restrições que podem ser feitas para assegurar sua identificabilidade. No entanto, é comum fixar  $\mathbb{E}(\xi) = 0$  e  $\mathbb{V}(\xi) = 1$ , de modo a garantir a *identificabilidade* do modelo e a *ancoragem* da escala do traço latente.

Um pressuposto essencial para a aplicação do método da Máxima Verossimilhança no contexto da TRI é o de *independência condicional*. Esse pressuposto integra todos os modelos de mensuração para variáveis latentes e afirma que, dado o traço latente, as variáveis observadas causadas por ele são independentes. Em outras palavras, fixado o valor de  $\xi_j$ , as observações  $x_{i_1}|\xi_j$  e  $x_{i_2}|\xi_j$  são independentes  $\forall i_1 \neq i_2$ . Sob a hipótese de independência condicional, o cálculo da função de verossimilhança é direto.

Seja  $P_{ij} = \Pr(x_{ij} = 1|\xi_j)$  e  $Q_{ij} = \Pr(x_{ij} = 0|\xi_j)$ . Para o item  $i$ , podemos compor o vetor  $\mathbf{x}_i$  com as respostas dadas por cada um dos  $N$  indivíduos amostrados. Então, a função de verossimilhança para estimação dos parâmetros do item  $i$  é dada por

$$L(\theta_i|\mathbf{x}_i) = \prod_{j=1}^N P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} . \quad (3.19)$$

O procedimento de maximização da verossimilhança pode ser de três tipos no que diz respeito ao tratamento do traço latente (Baker & Kim, 2004):

**Máxima Verossimilhança Conjunta** A estimação é feita iterativamente, em cada ciclo fixando os valores dos traços latentes para estimar os parâmetros dos itens e, em seguida, estimando os valores dos traços latentes.

**Máxima Verossimilhança Condicional** A estimação é feita como na Máxima Verossimilhança Conjunta, mas utilizando-se estatísticas suficientes para substituir os valores dos traços latentes.

**Máxima Verossimilhança Marginal** A estimação é feita assumindo que o traço

latente segue uma distribuição  $N(0, 1)$ , de modo a fazer  $\xi_j$  desaparecer da função de verossimilhança, por meio de sua integração.

A Máxima Verossimilhança Marginal constitui o método mais utilizado na estimação dos parâmetros estruturais da TRI.

### 3.3.3 *Softwares* e exemplo de resolução

Poucos pacotes computacionais estatísticos oferecem comandos específicos para o ajuste de modelos da TRI (Zheng & Rabe-Hesketh, 2007), sendo o R uma exceção, já que conta com diversas bibliotecas especialmente elaboradas com essa finalidade, como `ltm` e `eRm` (Mair *et al.*, 2010). No SAS, é possível utilizar o procedimento NLMIXED para estimar modelos da TRI (cf. Seção 3.4 e Kamata & Bauer (2008, p.149-153)). Quanto ao STATA, há funções desenvolvidas por usuários, que podem ser instaladas para trabalhar mais diretamente com tais modelos. Além desses, há diversos programas exclusivamente dedicados à estimação de parâmetros de modelos da TRI, como o BILOG, para itens dicotômicos, o MULTILOG para itens politômicos e o PARSCALE para modelos com funcionamento diferencial do item (DIF) (Holland & Wainer, 1993).

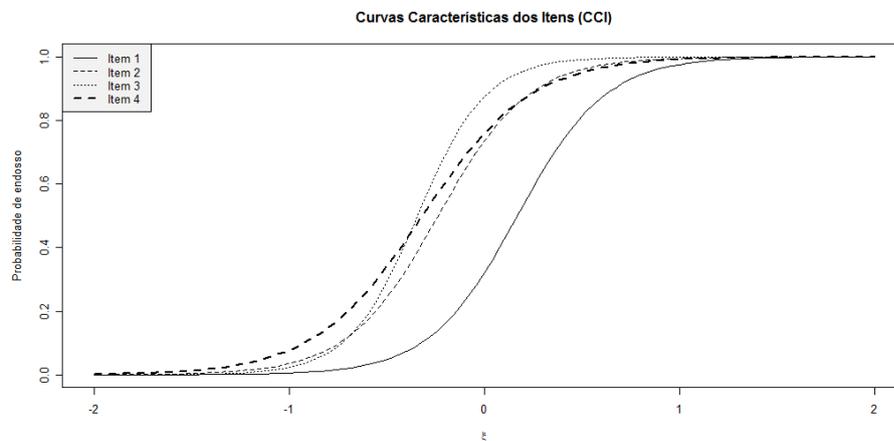
A Tabela 3.7 apresenta os parâmetros de um ML2 ajustado aos 4 itens do problema introduzido no início da Seção, referente à atitude frente à legalização do aborto. Cada item possui um parâmetro de dificuldade e um de discriminação que, substituídos na equação 3.17 determinam a curva característica de cada item, representadas conjuntamente no gráfico da Figura 3.3. As curvas do ML2 são deslocadas para a direita, quanto maior for o parâmetro de dificuldade, e têm sua inclinação acentuada, quanto maior for o parâmetro de discriminação.

A Tabela 3.8, por sua vez, apresenta os traços latentes estimados para cada padrão de resposta observado.

O Apêndice traz o código utilizado para a obtenção das estimativas utilizando a biblioteca `ltm` do R.

**Tabela 3.7:** Ajuste de um modelo logístico de 2 parâmetros (ML2) para o conjunto de dados referente à atitude frente à legalização do aborto (McGrath & Waterton, 1986): Estimativas dos parâmetros dos itens.

Item	Parâmetro	
	Dificuldade	Discriminação
	$(\beta_i)$	$(\alpha_i)$
1	0,17	4,453
2	-0,236	4,323
3	-0,343	5,664
4	-0,316	3,625



**Figura 3.3:** Ajuste de um modelo logístico de 2 parâmetros (ML2) para o conjunto de dados referente à atitude frente à legalização do aborto (McGrath & Waterton, 1986): Curvas Características dos Itens (CCI).

**Tabela 3.8:** Ajuste de um modelo logístico de 2 parâmetros (ML2) para o conjunto de dados referente à atitude frente à legalização do aborto (McGrath & Waterton, 1986): Estimativas dos valores dos traços latentes, segundo o vetor de respostas

Resposta				Valor do traço latente	Erro-padrão
Item 1	Item 2	Item 3	Item 4		
0	0	0	0	-0,898	0,463
0	0	0	1	-0,501	0,259
0	0	1	0	-0,378	0,236
0	0	1	1	-0,182	0,235
0	1	0	0	-0,456	0,249
0	1	0	1	-0,254	0,231
0	1	1	0	-0,143	0,239
0	1	1	1	0,094	0,277
1	0	0	0	-0,448	0,247
1	0	0	1	(não observado)	
1	0	1	0	(não observado)	
1	0	1	1	0,104	0,279
1	1	0	0	-0,210	0,233
1	1	0	1	0,006	0,260
1	1	1	0	0,161	0,292
1	1	1	1	0,660	0,524

## 3.4 A Teoria de Resposta ao Item como um tipo de Análise Fatorial (1983)

A Seção 3.2 apresentou a Análise Fatorial como precursora dos diversos modelos de mensuração conhecidos hoje. A Seção 3.3, por sua vez, apresentou a Teoria de Resposta ao Item como alternativa à Análise Fatorial, nos casos em que as variáveis observadas são categóricas. Mais especificamente, restringimo-nos ao caso em que as variáveis são dicotômicas e o problema de mensuração é unidimensional, isto é, envolve apenas um único traço latente (ou fator). A comparação das duas técnicas permite vislumbrar sua comunhão no que se refere à finalidade de mensurar quantidades não observáveis diretamente. Por outro lado, a modelagem feita em uma e na outra são tão díspares, que parece não haver conexão entre ambas, quando, na realidade, o modelo logístico de 2 parâmetros da TRI constitui um caso particular da Análise Fatorial clássica.

### 3.4.1 Duas abordagens para um mesmo problema

A extensão da Análise Fatorial, para que comporte respostas categóricas, é abordada de duas formas principais (Takane & de Leeuw, 1987). A primeira abordagem refere-se à chamada Análise Fatorial Binária, em que se observa a variável

$$x_i = \begin{cases} 1 & \text{se } x_i^* \geq \tau_i \\ 0 & \text{se } x_i^* < \tau_i \end{cases}, \quad i = 1, 2, \dots, q,$$

que corresponde à dicotomização de uma variável contínua subjacente:

$$x_i^* = \mu + \lambda_i \xi + \delta_i. \quad (3.20)$$

Esse tipo de tratamento foi originariamente chamado de *threshold model*, pois não retira o problema do contexto contínuo, apenas atribui a categorização da variável observada a um “threshold”, ou *limiar*, que particiona a variável contínua em dois intervalos distintos. Essa construção possibilita a substituição das indicadoras  $x_i$ , discretas, por suas análogas  $x_i^*$ , contínuas, acrescentando novos parâmetros ao modelo: os limiares  $\tau_i$ , um por variável para o caso de variáveis manifestas dicotômicas.

A função de verossimilhança para o modelo de limiar e uma amostra das observações  $j = 1, 2, \dots, N$  é da forma

$$L(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\tau}) = \prod_{j=1}^n \int_{a_{1j}}^{b_{1j}} \int_{a_{2j}}^{b_{2j}} \cdots \int_{a_{qj}}^{b_{qj}} \phi(\mathbf{y}_j) d\mathbf{y}_j,$$

ressaltando o caráter contínuo da abordagem, já que cada observação envolve a integração de uma densidade multivariada contínua ( $\phi(\mathbf{y}_j)$ ) em todas suas dimensões, na região correspondente às respostas dicotômicas observadas, que é definida pelos limites de integração  $a_{ij} = -\infty$  e  $b_{ij} = \tau_i$ , sempre que  $x_i = 0$ , ou  $a_{ij} = \tau_i$  e  $b_{ij} = \infty$ , quando  $x_i = 1$ .

A maximização de  $L(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\tau})$ , evidentemente, requer um investimento computacional elevado, de modo que soluções menos diretas foram formuladas por [Christofferson \(1975\)](#) e [Muthen \(1978\)](#), sendo bem acolhidas até o momento em que os métodos computacionalmente intensivos puderam ser mais comumente empregados.

Uma outra abordagem comumente dada às observações categorizadas de traços latentes contínuos é a modelagem por meio da *função de resposta*. Nela, busca-se a modelagem da probabilidade condicional de sucesso da variável  $X_i$ , em termos de um componente linear  $\nu$  dos fatores  $\xi_1, \xi_2, \dots, \xi_n$ , que pode ser dada na forma da equação [3.1](#), que aqui repetimos:

$$\mathbb{E}(X_i | \boldsymbol{\xi}) = g(\kappa_{i0} + \kappa_{i1}\xi_1 + \dots + \kappa_{in}\xi_n), \quad i = 1, \dots, q,$$

onde  $g(\cdot)$  é uma função de ligação,  $\kappa_{i0}, \dots, \kappa_{in}$  são os coeficientes de regressão da  $i$ -ésima variável manifesta e  $X_i$  é independente de  $X_j$ , para  $i \neq j$ , dado  $\boldsymbol{\xi} = \{\xi_1, \dots, \xi_n\}$ .

No modelo de Análise Fatorial, assume-se que os  $X_i$ 's são variáveis normais *i.i.d.* e  $g(\cdot)$  é a função de ligação *identidade*.

No contexto da TRI:

- trabalha-se mais comumente com apenas uma variável latente;
- $\mathbb{E}(X_i | \boldsymbol{\xi})$  expressa a probabilidade de endosso de uma das categorias de resposta do item  $i$ ; e
- a função  $g(\cdot)$  varia conforme o modelo adotado.

O modelo logístico de 2 parâmetros (ML2), por exemplo, resume-se ao caso linear generalizado, com função de ligação *logito* e componente linear

$$\nu_{ij} = \alpha_i(\xi_{ij} - \beta_i) ,$$

referente à observação  $j$  da variável  $i$ , com  $j = 1, 2, \dots, N$  e  $i = 1, 2, \dots, q$ . Como explicado no Capítulo 3.3, o parâmetro  $\beta_i$  refere-se à dificuldade do item  $i$  e  $\alpha_i$  ao seu poder discriminante. Modelos mais complexos, que fogem ao escopo linear generalizado, chegam a incluir 3 ou 4 parâmetros, normalmente relacionados à resposta contrária à esperada. No contexto dos testes educacionais, por exemplo, o modelo logístico pode conter um terceiro parâmetro, relacionado ao *acerto casual* do item (Birnbaum, 1968), e até mesmo um quarto parâmetro, relacionado ao *erro casual* do item (McDonald, 1967).

Por outro lado, há também a possibilidade de se trabalhar com um único parâmetro, utilizando o *Modelo de Rasch*, em que se restringe o parâmetro  $\alpha$  do ML2 fazendo-o igual a 1 e obtendo

$$\nu_{ij} = \xi_{ij} - \beta_i .$$

O modelo de Rasch teve grande relevância teórica, principalmente devido ao rigor metodológico de sua definição. Rasch desejou garantir que as medições realizadas com seu modelo tivessem propriedades idênticas às de qualquer medida física. Para tanto, formulou uma série de requisitos, atendidos pelo modelo que recebeu seu nome. A principal propriedade talvez seja a da *objetividade específica*, que requer que a comparação entre duas medidas independa de quaisquer outras medidas, isto é, a comparação depende exclusivamente de propriedades específicas dos objetos mensurados, sendo invariante quanto ao instrumento utilizado (Scheibleckner, 1977). Por exemplo, no âmbito das avaliações educacionais, é a objetividade específica que permite que as habilidades dos indivíduos e as dificuldades dos itens sejam comparáveis na mesma escala. Uma vantagem adicional dessa propriedade é que o método da máxima verossimilhança condicionada pode ser aplicado, pois o número de acertos de um indivíduo (seu score bruto) contém toda a informação relativa a sua habilidade, assim como o número de respostas corretas a um item contém toda a informação relativa a sua dificuldade, de modo que esses scores brutos representam estatísticas suficientes. Dessa forma, é possível fixar as habilidades dos indivíduos [os parâmetros dos itens] nos

valores dessas estatísticas, para se estimar os parâmetros dos itens [as habilidades dos indivíduos] correspondentes, dando início a um processo iterativo, conforme mostrado no Capítulo 3.3.

### 3.4.2 Análise Fatorial ou TRI: uma questão de parametrização

Sabendo que há duas abordagens para estender a Análise Fatorial tradicional, para que trate de respostas categorizadas, convém conjecturar se não existe uma conexão entre as duas abordagens. De fato, apesar de aparentemente distintas, uma abordagem consiste em uma parametrização da outra, pelo menos para o caso unidimensional (um único fator).

Para a função de resposta com função de ligação logito ou probito e componente linear escrito na forma tradicional, isto é,

$$\mathbb{E}(X_i | \xi) = \Pr(X_i = 1 | \xi) = g^{-1}(\nu) = F(\alpha_i \xi + \beta_i^*),$$

com  $g^{-1}(\cdot) = F(\cdot)$  representando a função de distribuição acumulada normal ou logística e  $\beta_i^* = -\alpha_i \beta_i$  uma reparametrização do parâmetro de dificuldade (cf. Seção 3.3), Takane & de Leeuw (1987) mostraram que sua correspondência com os parâmetros do modelo (3.20) é dada por

$$\alpha_i = \frac{\lambda_i}{q_i} \quad \text{e} \quad \beta_i^* = -\frac{\tau_i}{q_i}, \quad (3.21)$$

onde  $q_i = \sqrt{\mathbb{V}(\delta_i)}$  representa o desvio-padrão do erro de mensuração de (3.20), desde que a escala da Análise Fatorial Binária tenha sido fixada fazendo  $\mathbb{E}(\xi) = 0$  e  $\mathbb{V}(\xi) = 1$ .

A questão da fixação da escala, no entanto merece atenção. Apesar de a escolha de Takane & de Leeuw (1987) ser muito comum, há outras formas de se fixar a escala de um modelo de Análise Fatorial. Na verdade, cada opção de escala consiste de uma restrição para garantia de identificabilidade. Assim como a restrição feita na matriz de unicidades da Análise Fatorial do Capítulo 3.2 e a fixação da esperança e da variância do traço latente do Capítulo 3.3, é necessário escolher alguma restrição, tal que apenas uma das infinitas soluções possíveis para um modelo de mensuração seja obtida.

No caso da Análise Fatorial Clássica, ficou evidente que essa escolha é totalmente arbitrária, sendo possível até mesmo rotacionar a solução obtida até que se encontre aquela com interpretação mais conveniente.

No caso da Análise Fatorial Binária, há um complicador adicional: os parâmetros  $\tau_i$ , que definem os limiares das respostas categorizadas, não admitem unicidade se  $\mu$  for livre. Se assim fosse, poderíamos obter um conjunto de infinitas soluções equivalentes para os demais parâmetros, cada uma para um valor particular de  $\mu$ . Portanto, faremos  $\mu = 0$ .

Assim sendo, podemos enumerar as restrições mais utilizadas para tornar identificável um modelo de mensuração fatorial como sendo:

**Ancoragem** Realizada fixando em 1 a carga do fator sobre alguma das variáveis manifestas, fazendo com que a escala do fator seja dada na mesma escala da variável indicadora. No caso da Análise Fatorial Binária, também é necessário fixar o valor de algum dos limiares  $\tau_i$ . É comum que a primeira variável indicadora seja selecionada para sofrer as restrições, i.e.,  $\lambda_1 = 1$  e  $\tau_1 = 0$ .

**Padronização** Realizada fixando a locação e a escala da variável latente, normalmente em 0 e 1, obtendo  $\mathbb{E}(\xi) = 0$  e  $\mathbb{V}(\xi) = 1$ .

Além disso, a variabilidade das observações realizadas pode ser fixada:

**Marginalmente** Fazendo  $\mathbb{V}(X_i^*) = 1$ ; ou

**Condicionalmente** Fazendo  $\mathbb{V}(\delta_i) = 1$ .

Da combinação entre ancoragem / padronização e variância fixada marginalmente / condicionalmente, obtemos quatro parametrizações possíveis para a identificabilidade do modelo de Análise Fatorial Binária, resumidas na Tabela 3.9.

A partir das parametrizações apresentadas nessa tabela e das fórmulas da expressão (3.21), temos que a parametrização por padronização e com variabilidade condicional remete às igualdades  $\alpha_i = \lambda_i$  e  $\beta_i^* = -\tau_i$ , de onde temos que os parâmetros desse modelo coincidem com os da TRI, salvo pela inversão de sinal do parâmetro  $\beta_i^*$ .

Esse resultado já permite a tradução dos resultados de uma Análise Fatorial Binária dentro dessa parametrização para os parâmetros da TRI, observada a mudança

Variabilidade das observações	Tipo de restrição	
	Ancoragem	Padronização
Marginal	$\lambda_1 = 1, \tau_1 = 0, \mathbb{V}(X_i^*) = 1$	$\mathbb{E}(\xi) = 0, \mathbb{V}(\xi) = 1, \mathbb{V}(X_i^*) = 1$
Condicional	$\lambda_1 = 1, \tau_1 = 0, \mathbb{V}(\delta_i) = 1$	$\mathbb{E}(\xi) = 0, \mathbb{V}(\xi) = 1, \mathbb{V}(\delta_i) = 1$

**Tabela 3.9:** Quatro parametrizações possíveis para identificabilidade de um modelo de Análise Fatorial Binária

de sinal em  $\beta_i^*$  e atentando para o fato de que o parâmetro de dificuldade comumente utilizado é dado por  $\beta_i = -\alpha_i\beta_i^*$ .

Contudo, [Kamata & Bauer \(2008\)](#) estenderam a descoberta de [Takane & de Leeuw \(1987\)](#). Partindo do modelo (3.20) e da definição das variáveis observadas  $x_i$  em termos dos limiares  $\tau_i$ , executaram três transformações:

1. Dividiram todos os elementos por um desvio-padrão  $q_i$ ;
2. Recentralizaram o fator  $\xi$  subtraindo dele uma constante  $r$ ; e
3. Dividiram o fator transformado  $(\xi - r)$  por um desvio-padrão  $s$ , fazendo as alterações necessárias para garantir a igualdade da equação (3.20).

Dessa forma, obtiveram a transformação das equações do modelo, descritas por

$$\frac{x_i^* - \lambda_i r}{q_i} = \frac{\mu}{q_i} + \frac{\lambda_i s}{q_i} \left( \frac{\xi - r}{s} \right) + \frac{\delta_i}{q_i}.$$

e

$$x_i = \begin{cases} 1 & \text{se } \frac{x_i^* - \lambda_i r}{q_i} \geq \frac{\tau_i - \lambda_i r}{q_i} \\ 0 & \text{se } \frac{x_i^* - \lambda_i r}{q_i} < \frac{\tau_i - \lambda_i r}{q_i} \end{cases}, \quad i = 1, 2, \dots, q,$$

Essas expressões são genéricas o suficiente para permitir a conversão de uma parametrização em outra, bastando definir os valores de  $q_i$ ,  $r$  e  $s$ . Podemos, inclusive, atualizar o resultado (3.21), obtendo

$$\alpha_i = \frac{\lambda_i s}{q_i} \quad \text{e} \quad \beta_i^* = -\frac{\tau_i - \lambda_i r}{q_i}.$$

Se fizermos  $q_i = \sqrt{\mathbb{V}(\delta_i)}$ ,  $r = \mathbb{E}(\xi)$  e  $s = \sqrt{\mathbb{V}(\xi)}$ , obtemos as fórmulas de transformação para os parâmetros da TRI a partir de qualquer uma das quatro parametrizações da Tabela 3.9, bastando substituir nas fórmulas os valores correspondentes à

parametrização desejada:

$$\alpha_i = \frac{\lambda_i \sqrt{\mathbb{V}(\xi)}}{\sqrt{\mathbb{V}(\delta_i)}} \quad \text{e} \quad \beta_i^* = -\frac{\tau_i - \lambda_i \mathbb{E}(\xi)}{\sqrt{\mathbb{V}(\delta_i)}}.$$

É importante lembrar que o sinal de  $\beta_i^*$  fica invertido na transformação. Além disso, é interessante notar que embora a Análise Fatorial Binária exija a estimação de 3 parâmetros para cada item ( $\lambda_i$ ,  $\mathbb{V}(\delta_i)$  e  $\tau_i$ , sua transformação para os 2 parâmetros da TRI se dá pela conjugação da variância do erro ( $\mathbb{V}(\delta_i)$ ) nos demais parâmetros. Isso se dá porque, quando os dados são dicotômicos, a variância de cada variável manifesta não pode ser estimada devido à falta de informação relevante nos dados e, conseqüentemente,  $q_i$  pode assumir um valor arbitrário (Takane & de Leeuw, 1987). Portanto, o número efetivo de parâmetros nos dois modelos é idêntico.

Por fim, é digno de nota que os erros de mensuração ( $\delta_i$ ) da Análise Fatorial Binária usualmente seguem uma distribuição normal multivariada, mas podem ter sua distribuição definida como logística, a fim de conciliar seus resultados com os dos modelos logísticos da TRI. Alternativamente, pode-se utilizar a função de ligação probito na TRI, compatibilizando seus resultados com os da Análise Fatorial Binária de resíduos normais multivariados.

### 3.4.3 Semelhanças e dessemelhanças

Apesar de provado que a Análise Fatorial Binária e a TRI constituem formulações alternativas de um mesmo modelo, cabe notar que o uso de cada técnica costuma ser bastante diverso na prática.

Primeiramente, a TRI presta-se precipuamente à finalidade de mensuração em uma escala pré-definida, ao passo que a Análise Fatorial presta-se também à finalidade *exploratória*, auxiliando na busca de estruturas latentes. A indeterminação da escala é, inclusive, uma questão que deve ser analisada caso a caso nas aplicações da Análise Fatorial.

Em segundo lugar, a Análise Fatorial costuma tratar de problemas que apresentam uma estrutura latente multidimensional, enquanto a TRI é mais comumente utilizada em problemas unidimensionais, em que se deseja realizar medições de um único fator.

Em terceiro lugar, os parâmetros da TRI possuem interpretação fácil e direta, em

termos de características dos itens, ao passo que as cargas da Análise Fatorial Clássica não costumam ser interpretadas, mas utilizadas como indicativos do poder de influência de cada fator em cada variável observada. Inclusive, a capacidade de expansão dos modelos da TRI, para que contenham novos parâmetros com novos significados, é muito maior que a da Análise Fatorial. Exemplo disso são os tantos modelos utilizados em TRI, como aqueles para dados politômicos (nominais e ordinais), os modelos de desdobramento e os modelos com parâmetro de funcionamento diferencial do item (DIF), entre outros.

Finalmente, na TRI modela-se um questionário item a item, o que possibilita a criação de um banco de itens calibrados que podem ser combinados independentemente para formar um novo questionário para mensuração do traço latente, na mesma escala do primeiro. A Análise Fatorial não costuma ser aplicada dessa forma. O Exemplo da Seção 3.2 ilustra esse fato, pois nele cada observação modelada corresponde à nota de uma prova resolvida por um aluno da amostra, de modo que a carga e a variância estimadas para uma determinada prova só serão as mesmas caso uma prova idêntica seja aplicada em uma segunda amostra.

Tais diferenças provavelmente se devem ao contexto em que os modelos surgiram e aos problemas que procuraram resolver com sucesso. Decerto foi o uso, e não sua formulação matemática, que causou o tamanho diferenciamento entre modelos absolutamente iguais.

## 3.5 Apêndice

### 3.5.1 Linhas de código e *output* do R: ajuste do modelo de Análise Fatorial aos dados da Seção 3.2

```
.  
  
library(datasets)  
require(stats)  
#install.packages("GPArotation")  
library(GPArotation)
```

```

> ability.cov
$cov
      general picture blocks  maze reading  vocab
general 24.641  5.991 33.520  6.023 20.755 29.701
picture  5.991  6.700 18.137  1.782  4.936  7.204
blocks  33.520 18.137 149.831 19.424 31.430 50.753
maze     6.023  1.782 19.424 12.711  4.757  9.075
reading 20.755  4.936 31.430  4.757 52.604 66.762
vocab   29.701  7.204 50.753  9.075 66.762 135.292

$center
[1] 0 0 0 0 0 0

$n.obs
[1] 112

> AF.provas <- factanal(covmat=ability.cov , factors = 1)
> AF.provas

Call:
factanal(factors = 1, covmat = ability.cov)

Uniquenesses:
general picture blocks  maze reading  vocab
 0.535  0.853  0.748  0.910  0.232  0.280

Loadings:
      Factor1
general 0.682
picture 0.384
blocks  0.502
maze    0.300
reading 0.877
vocab   0.849

      Factor1
SS loadings  2.443
Proportion Var 0.407

```

Test of the hypothesis that 1 factor is sufficient.  
The chi square statistic is 75.18 on 9 degrees of freedom.  
The p-value is 1.46e-12

```
> update(AF.provas, factors=2 , rotation = "none")
```

Call:

```
factanal(factors = 2, covmat = ability.cov, rotation = "none")
```

Uniquenesses:

general	picture	blocks	maze	reading	vocab
0.455	0.589	0.218	0.769	0.052	0.334

Loadings:

	Factor1	Factor2
general	0.648	0.354
picture	0.347	0.538
blocks	0.471	0.748
maze	0.253	0.408
reading	0.964	-0.135
vocab	0.815	

	Factor1	Factor2
SS loadings	2.420	1.162
Proportion Var	0.403	0.194
Cumulative Var	0.403	0.597

Test of the hypothesis that 2 factors are sufficient.  
The chi square statistic is 6.11 on 4 degrees of freedom.  
The p-value is 0.191

```
> update(AF.provas, factors = 2 , rotation = "varimax")
```

Call:

```
factanal(factors = 2, covmat = ability.cov, rotation = "varimax")
```

Uniquenesses:

general	picture	blocks	maze	reading	vocab
0.455	0.589	0.218	0.769	0.052	0.334

Loadings:

	Factor1	Factor2
general	0.499	0.543
picture	0.156	0.622
blocks	0.206	0.860
maze	0.109	0.468
reading	0.956	0.182
vocab	0.785	0.225

	Factor1	Factor2
SS loadings	1.858	1.724
Proportion Var	0.310	0.287
Cumulative Var	0.310	0.597

Test of the hypothesis that 2 factors are sufficient.  
The chi square statistic is 6.11 on 4 degrees of freedom.  
The p-value is 0.191

```
> update(AF.provas, factors = 2 , rotation = "promax")
```

Call:

```
factanal(factors = 2, covmat = ability.cov, rotation = "promax")
```

Uniquenesses:

general	picture	blocks	maze	reading	vocab
0.455	0.589	0.218	0.769	0.052	0.334

Loadings:

	Factor1	Factor2
general	0.364	0.470
picture		0.671
blocks		0.932
maze		0.508
reading	1.023	
vocab	0.811	

	Factor1	Factor2
SS loadings	1.853	1.807
Proportion Var	0.309	0.301
Cumulative Var	0.309	0.610

Test of the hypothesis that 2 factors are sufficient.  
The chi square statistic is 6.11 on 4 degrees of freedom.  
The p-value is 0.191

```
> update(AF.provas, factors = 2 , rotation = "oblimin")
```

Call:

```
factanal(factors = 2, covmat = ability.cov, rotation = "oblimin")
```

Uniquenesses:

general	picture	blocks	maze	reading	vocab
0.455	0.589	0.218	0.769	0.052	0.334

Loadings:

	Factor1	Factor2
general	0.386	0.475
picture		0.646
blocks		0.896
maze		0.488
reading	0.990	
vocab	0.791	

	Factor1	Factor2
SS loadings	1.756	1.688
Proportion Var	0.293	0.281
Cumulative Var	0.293	0.574

Test of the hypothesis that 2 factors are sufficient.  
The chi square statistic is 6.11 on 4 degrees of freedom.  
The p-value is 0.191

### 3.5.2 Linhas de código e *output* do R: ajuste do modelo logístico de 2 parâmetros, da Teoria de Resposta ao Item, aos dados da Seção 3.3

.

```
> library(ltm)
> A <-ltm(Abortion ~ z1)
> A
```

Call:

```
ltm(formula = Abortion ~ z1)
```

Coefficients:

	Dffc1t	Dscrmn
Item 1	0.170	4.453
Item 2	-0.236	4.323
Item 3	-0.343	5.664
Item 4	-0.316	3.625

Log.Lik: -706.337

```
> factor.scores(A)
```

Call:

```
ltm(formula = Abortion ~ z1)
```

Scoring Method: Empirical Bayes

Factor-Scores for observed response patterns:

	Item 1	Item 2	Item 3	Item 4	Obs	Exp	z1	se.z1
1	0	0	0	0	103	103.412	-0.898	0.463
2	0	0	0	1	13	15.520	-0.501	0.259
3	0	0	1	0	10	11.148	-0.378	0.236
4	0	0	1	1	21	14.316	-0.182	0.235
5	0	1	0	0	9	8.914	-0.456	0.249
6	0	1	0	1	6	6.810	-0.254	0.231
7	0	1	1	0	7	12.053	-0.143	0.239

8	0	1	1	1	44	41.819	0.094	0.277
9	1	0	0	0	1	1.399	-0.448	0.247
10	1	0	1	1	6	7.193	0.104	0.279
11	1	1	0	0	3	0.896	-0.210	0.233
12	1	1	0	1	3	2.555	0.006	0.260
13	1	1	1	0	12	7.105	0.161	0.292
14	1	1	1	1	141	142.708	0.660	0.524

```

> # Gráfico das CCI
> CCI1 <- function (x) {1/(1+exp(-coef(A)[1,][2] * (x-coef(A)[1,][1])))}
> CCI2 <- function (x) {1/(1+exp(-coef(A)[2,][2] * (x-coef(A)[2,][1])))}
> CCI3 <- function (x) {1/(1+exp(-coef(A)[3,][2] * (x-coef(A)[3,][1])))}
> CCI4 <- function (x) {1/(1+exp(-coef(A)[4,][2] * (x-coef(A)[4,][1])))}
>
> par(cex=0.8)
> plot(CCI1,-2,2,main="Curvas Características dos Itens (CCI)" ,
+ xlab = expression(xi), ylab = "Probabilidade de endosso", lty = 1)
> curve(CCI2,-2,2,lty = 2, add=T)
> curve(CCI3,-2,2,lty = 3, add=T)
> curve(CCI4,-2,2,lwd=2,lty=2, add=T)
> legend("topleft",legend=paste("Item",1:4),lty=c(1,2,3,2),lwd=c(1,1,1,2),bg=grey(0.95))

```

## Capítulo 4

# Regressão de variáveis latentes contínuas

O Capítulo 3 apresentou diversos modelos de mensuração lineares. As variáveis latentes aferidas por meio de tais modelos evidentemente podem compor outros modelos, nos quais são utilizadas como covariáveis ou variáveis-resposta. O pressuposto de que as covariáveis representam efeitos fixos, no entanto, é essencial aos modelos clássicos de regressão e não é atendido pelas variáveis latentes, que representam efeitos essencialmente aleatórios. Além disso, se utilizadas como variáveis-resposta, o componente aleatório da equação de regressão será acrescido do erro de mensuração inerente a qualquer variável latente. Essas limitações impactam na estimação dos coeficientes de regressão, que tendem a ser subestimados, definindo o fenômeno chamado de *atenuação*, já detalhado na Introdução (Capítulo 2).

Na Seção 4.1 do presente Capítulo, o modelo de Equações Estruturais será apresentado, por constituir uma solução para o fenômeno da atenuação na regressão de variáveis aleatórias. De fato, essa, que na realidade é uma classe de modelos, apresenta outras ferramentas para a análise de variáveis latentes, como a Análise Fatorial Confirmatória, que é útil na validação do modelo de mensuração hipotetizado a partir dos resultados da Análise Fatorial Exploratória.

Além disso, o modelo de Equações Estruturais possibilita que se ajuste simultaneamente o modelo de regressão entre variáveis latentes e os modelos de mensuração dessas mesmas variáveis latentes, o que representa um ganho em relação ao ajuste

feito em etapas: primeiro dos parâmetros estruturais do modelo de mensuração, depois dos parâmetros incidentais do modelo de mensuração e, por último, da regressão entre os parâmetros incidentais estimados. Mesmo assim, Skrondal & Laake (2001) investigaram o processo de regressão de fatores em etapas e apresentaram casos específicos em que ela produz estimadores consistentes e livres do efeito da atenuação. Tais resultados são apresentados na Seção 4.2.

Finalmente, o Capítulo encerra-se com a Seção 4.3, que apresenta limitações do modelo de Equações Estruturais para a regressão de variáveis latentes.

## 4.1 Modelos de Equações Estruturais (1973)

Modelos de Equações Estruturais (MEE) são modelos compostos por múltiplas equações de regressão, nos quais a variável-resposta de uma equação pode figurar como covariável em outra equação do sistema. De fato, em um MEE duas variáveis podem afetar-se diretamente, indiretamente ou até mesmo reciprocamente. Além disso, podem conter variáveis latentes, mensuradas não somente por seus efeitos, como na Análise Fatorial ou na TRI, mas também por suas causas.

A equação geral de um MEE na notação LISREL é dada por

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \quad (4.1)$$

com  $\mathbb{E}(\boldsymbol{\eta}) = \mathbb{E}(\boldsymbol{\xi})\mathbb{E}(\boldsymbol{\zeta}) = \mathbf{0}$ ,  $(\mathbf{I} - \mathbf{B})$  não-singular e  $\boldsymbol{\zeta}$  independente de  $\boldsymbol{\eta}$  e de  $\boldsymbol{\xi}$ , conforme apresentado previamente na Tabela 1.2. Os MEE são usualmente representados por diagramas de caminho (Seção 2.2) que evidenciam graficamente as relações entre as variáveis do modelo.

A equação (4.1), que define a classe dos MEE é muito abrangente e origina subclasses inteiras de modelos, como:

- Os Modelos de Equações Estruturais com Variáveis Observadas (MEEVO);
- A Análise Fatorial Confirmatória (AFC); e
- Os Modelos Gerais de Equações Estruturais, aqui representados pelo LISREL<sup>1</sup> (Jöreskog, 1969).

---

<sup>1</sup>Acrônimo para *Linear Structural Relationships*

Dentre tantas especificações de modelo possíveis, uma importante questão dos MEE é precisamente a da identificabilidade. Nem todos os MEE formulados serão identificados, motivo que levou ao desenvolvimento de regras de identificabilidade, que são condições necessárias ou suficientes que, verificadas, garantem que o modelo é identificado. Evidentemente, há modelos de interesse científico cuja identificabilidade não é determinada por um conjunto dessas regras, devido ao fato de esses modelos atenderem plenamente às condições necessárias especificadas pelas regras e, simultaneamente, não atender às condições suficientes. Nesses casos, devem-se verificar outras regras, que representem condições necessárias e suficientes ou lançar mão de técnicas empíricas de análise da identificabilidade local que, entretanto, não garantem identificabilidade global (Bollen, 1989; Skrondal & Rabe-Hesketh, 2004).

Além dessa, outra questão relevante é a da estimação. Por se tratarem de modelos com múltiplas equações que podem contemplar variáveis latentes, a estimação pode lidar com as variáveis latentes como aleatórias ou fixas (cf. Seção 3.2.5) e, dependendo da relação existente entre as variáveis endógenas, o ajuste de todas as equações simultaneamente pode não se mostrar necessário, caso em que as equações podem ser estimadas uma a uma, por Mínimos Quadrados Ordinários ou por Mínimos Quadrados em Duas Etapas, por exemplo.

Como cada tipo de MEE apresenta peculiaridades relacionadas à forma de especificação, à identificabilidade e aos métodos de estimação, cada tipo será abordado individualmente a seguir.

#### 4.1.1 Modelo de Equações Estruturais com Variáveis Observadas

Embora os Modelos de Equações Estruturais com Variáveis Observadas (MEEVO) não tratem diretamente de variáveis latentes, os MEE com variáveis latentes herdam diversas propriedades dos MEEVO, sendo importante sua apresentação. Na notação LISREL, ele é descrito por

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{\Gamma}\mathbf{x} + \boldsymbol{\zeta} .$$

Por serem observadas, as covariáveis não possuem erro de mensuração, de onde segue que o modelo de mensuração é dado por

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\eta} \\ \mathbf{x} &= \boldsymbol{\xi}. \end{aligned} \quad (4.2)$$

Esse modelo, também conhecido pelo nome de Equações Simultâneas, foi profundamente estudado e utilizado por econométristas, principalmente no contexto dos modelos macroeconômicos, em que as variáveis estudadas possuem interações extremamente complexas entre si (Klein, 1950; Haavelmo, 1943).

### Matriz de covariâncias implicada pelo modelo

A matriz de covariâncias  $\Sigma(\boldsymbol{\theta})$  implicada pelo modelo é composta por três matrizes menores, a saber:

1. A matriz de covariâncias  $\Sigma_{yy}(\boldsymbol{\theta})$  implicada pelo vetor  $\mathbf{y}$ ,
2. A matriz de covariâncias  $\Sigma_{xy}(\boldsymbol{\theta})$  implicada pelos vetores  $\mathbf{x}$  e  $\mathbf{y}$ , e
3. A matriz de covariâncias  $\Sigma_{xx}(\boldsymbol{\theta})$  implicada pelo do vetor  $\mathbf{x}$ ,

organizadas na matriz-bloco

$$\Sigma(\boldsymbol{\theta}) = \begin{pmatrix} \Sigma_{yy}(\boldsymbol{\theta}) & \Sigma'_{xy}(\boldsymbol{\theta}) \\ \Sigma_{xy}(\boldsymbol{\theta}) & \Sigma_{xx}(\boldsymbol{\theta}) \end{pmatrix}. \quad (4.3)$$

Calculando, obtemos

$$\begin{aligned} \Sigma_{yy}(\boldsymbol{\theta}) &= \mathbb{E}(\mathbf{y}\mathbf{y}') \\ &= \mathbb{E} \left\{ (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}) [(\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta})]'\right\} \\ &= \mathbb{E} \left\{ (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta})(\mathbf{x}'\boldsymbol{\Gamma}' + \boldsymbol{\zeta}')(\mathbf{I} - \mathbf{B})^{-1'} \right\} \\ &= (\mathbf{I} - \mathbf{B})^{-1}[\mathbb{E}(\boldsymbol{\Gamma}\mathbf{x}\mathbf{x}'\boldsymbol{\Gamma}') + \mathbb{E}(\boldsymbol{\Gamma}\mathbf{x}\boldsymbol{\zeta}') + \mathbb{E}(\boldsymbol{\zeta}\mathbf{x}'\boldsymbol{\Gamma}') + \mathbb{E}(\boldsymbol{\zeta}\boldsymbol{\zeta}')] (\mathbf{I} - \mathbf{B})^{-1'} \\ &= (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})(\mathbf{I} - \mathbf{B})^{-1'}, \end{aligned} \quad (4.4)$$

$$\begin{aligned} \Sigma_{xx}(\boldsymbol{\theta}) &= \mathbb{E}(\mathbf{x}\mathbf{x}') \\ &= \boldsymbol{\Phi} \end{aligned} \quad (4.5)$$

e

$$\begin{aligned}
\Sigma_{xy}(\boldsymbol{\theta}) &= \mathbb{E}(\mathbf{xy}') \\
&= \mathbb{E} \left\{ \mathbf{x} [(\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta})]' \right\} \\
&= \boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I} - \mathbf{B})^{-1'}.
\end{aligned} \tag{4.6}$$

Os resultados (4.4), (4.5) e (4.6) permitem a obtenção da matriz de covariâncias (4.3) implicada pelo modelo:

$$\Sigma(\boldsymbol{\theta}) = \begin{pmatrix} (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})(\mathbf{I} - \mathbf{B})^{-1'} & (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I} - \mathbf{B})^{-1'} & \boldsymbol{\Phi} \end{pmatrix}. \tag{4.7}$$

### Identificabilidade

A questão da identificabilidade nos Modelos de Equações Estruturais (MEE) coloca-se devido a uma não-equivalência entre o número de parâmetros e o número de estimadores calculados a partir dos dados observados. Por modelarem a estrutura de covariância dos dados, qualquer MEE é dado por uma expressão do tipo

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}), \tag{4.8}$$

em que o lado esquerdo representa a matriz de covariâncias da população de  $\mathbf{X}$  e  $\mathbf{Y}$  e o lado direito representa cada elemento dessa matriz em termos dos parâmetros não-observados. Sendo assim, o problema de estimação de um modelo perfeitamente identificado consiste em resolver o sistema

$$\mathbf{S} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}), \tag{4.9}$$

em que  $\mathbf{S}$  é a matriz de covariâncias amostral. Como o número de elementos não-redundantes de  $(S)$  é de  $(p + q)(p + q + 1)/2$ , pois os elementos acima da diagonal principal são idênticos aos que estão abaixo, o sistema (4.9) possui  $(p + q)(p + q + 1)/2$  equações não-lineares (vide (4.7)), de modo que um número  $t$  de incógnitas (parâmetros) maior que o número de equações torna o sistema não-identificado.

Daí segue a primeira condição de identificabilidade, conhecida como Regra  $t$ , necessária embora não suficiente. Para que um modelo seja identificado, é necessário

que o número  $t$ , de parâmetros livres, atenda à restrição

$$t \leq \frac{(p+q)(p+q+1)}{2} .$$

Além da regra  $t$ , há duas condições suficientes comumente utilizadas na avaliação da identificabilidade de um MEEVO, que são baseadas na matriz  $\mathbf{B}$ . A *regra do  $\mathbf{B}$  nulo* assegura que o modelo será identificado se  $\mathbf{B} = \mathbf{0}$ . A prova baseia-se no fato de que  $\Sigma_{xx} = \Phi$  pela equação (4.7) e de que, substituindo  $\phi$  por  $\Sigma_{xx}$  nos demais quadrantes da matriz  $\Sigma(\theta)$ , obtemos expressões identificadas para  $\Gamma$  e  $\Psi$ , dado que  $\mathbf{B} = \mathbf{0}$  (Bollen, 1989, p.94).

Outra condição baseada na matriz  $\mathbf{B}$  é a *regra do modelo recursivo*. É considerado recursivo todo MEE em que:

1. Os erros latentes são independentes (ou pelo menos não-correlacionados), i.e.,  $\Psi$  é diagonal; e
2. As relações de causalidade entre as variáveis são unidirecionais e, portanto, sem reciprocidade direta ou indireta, i.e.,  $\mathbf{B}$  é triangular inferior.

A regra afirma que todo modelo recursivo é identificado e sua prova baseia-se no fato de que, como as variáveis-resposta do modelo não participam de outras equações como covariáveis, o erro latente  $\zeta_i$  de uma equação não sofre interferências do erro de outra, fazendo com que cada equação seja independente das demais. Essa característica possibilita até mesmo a estimação dos parâmetros equação a equação, pelo Método dos Mínimos Quadrados (Fox, 2006).

Finalmente, outras duas regras bastante difundidas estão relacionadas à matriz

$$\mathbf{C} = [(\mathbf{I} - \mathbf{B}) | -\Gamma] . \quad (4.10)$$

A *condição de ordem* é necessária para a identificabilidade e decorre de uma verificação de cada linha da matriz  $\mathbf{C}$ : para que o modelo seja identificado, é necessário que cada linha de  $\mathbf{C}$  contenha pelo menos  $p - 1$  zeros, sendo  $p$  o número de variáveis-resposta do MEEVO. A *condição de posto*, por sua vez, é necessária e suficiente para a identificação do modelo e pode ser verificada a partir de submatrizes  $\mathbf{C}_i$  de  $\mathbf{C}$ , que são formadas apenas pelas colunas de  $\mathbf{C}$  que possuem zeros na linha  $i$ , sendo as demais colunas excluídas. A condição é atendida se o posto de todas as submatrizes  $\mathbf{C}_i$  for

exatamente  $p - 1$ . É importante ressaltar que essa caracterização das condições de ordem e de posto pressupõem que todos os elementos de  $\Psi$  são parâmetros livres (Bollen, 1989, p.98-104).

### ***Softwares e exemplo de resolução***

Como o MEEVO constitui um caso particular dos Modelos de Equações Estruturais com Variáveis Latentes, todos os *softwares* apresentados na Seção 4.1.3 também prestam-se à estimação de parâmetros dos MEEVO.

Entretanto, alguns métodos como Mínimos Quadrados em 2 Passos podem ser utilizadas especificamente no ajuste de modelos de Equações Estruturais que não contenham variáveis latentes, como os MEEVO (Fox, 2006). No R, por exemplo, a biblioteca `sem` (Fox *et al.*, 2010) contém funções específicas para tal finalidade, como a `tsls()`.

### **4.1.2 Análise Fatorial Confirmatória**

Os modelos de Análise Fatorial Confirmatória AFC possuem estrutura similar à dos modelos de Análise Fatorial Exploratória AFE, mas uso bastante diverso. Se o propósito de se aplicar uma AFE é o de explorar características desconhecidas do modelo de mensuração que relaciona as variáveis observadas às latentes, o de uma AFC é testar um modelo de mensuração predeterminado e, eventualmente compará-lo com outros modelos, a fim de identificar o de melhor ajuste (cf. Seção 3.2.6). Entretanto, na AFC as associações entre cada variável latente e as variáveis observadas a ela associadas é predeterminado. Além disso, na AFC, a hipótese de que os fatores são ortogonais não é necessária como na AFE, podendo-se modelar as correlações entre diversos pares de fatores e fixar a ortogonalidade de outros pares, conforme parecer mais apropriado ao pesquisador. Quanto à escala dos escores dos fatores, pode-se fixar, para cada fator, uma de suas cargas com valor unitário, de modo que o fator herde a escala da variável que estiver associada a ele com essa carga (ancoragem) ou considerar que os fatores possuem esperança e variância iguais a 0 e 1, respectivamente (padronização) (cf. Seção 3.4.2).

## Matriz de covariâncias implicada pelo modelo

Matematicamente, a semelhança entre os modelos é evidenciada por meio da matriz implicada pelo modelo de AFC, que é idêntica à da AFE, isto é,

$$\Sigma(\theta) = \Lambda_x \Phi \Lambda'_x + \Theta_\delta . \quad (4.11)$$

Entretanto, no âmbito da AFC é possível optar por restringir ou manter livre quaisquer parâmetros do modelo, isto é, quaisquer elementos das matrizes  $\Lambda$ ,  $\Phi$  ou  $\Theta_\delta$ , desde que tais opções não tornem o modelo subidentificado.

## Identificabilidade

Assim como para os MEEVO, a confirmação da identificabilidade de um modelo proposto também é crucial no contexto da AFC. Uma primeira condição, necessária para identificabilidade, é a *regra t*, definida de forma análoga a sua homônima do contexto dos MEEVO. Para atender às equações (4.8) e (4.9), inerentes a qualquer Modelo de Equações Estruturais, dado que a AFC considera  $q$  variáveis observadas, temos que  $\mathbf{S}$  será composto por no máximo  $q(q+1)/2$  elementos distintos. Dessa forma, o número  $t$  de parâmetros livres do modelo de AFC não deve superar essa quantidade. A regra  $t$ , portanto, pode ser escrita como

$$t \leq \frac{q(q+1)}{2} .$$

Além da regra  $t$ , a *regra das três indicadoras* é muito difundida. Ela é atendida para modelos com pelo menos um fator, em que cada variável observada (indicadora) está associada a apenas um fator e os erros de mensuração são independentes. Sob tais condições, a regra garante a identificabilidade do modelo desde que cada fator conte com pelo menos *três indicadoras*. Caso se acrescente a condição de que pelo menos um par de fatores é correlacionado, isto é,  $\phi_{ij} \neq 0$  para ao menos um par  $i, j$  com  $i \neq j$ , a *regra das duas indicadoras* assegura que o modelo é identificado desde que cada fator conte com pelo menos *duas indicadoras* (Bollen, 1989, p.238-251).

## Softwares e exemplo de resolução

Como exemplo de aplicação, podemos tomar o modelo hipotetizado na Seção 3.2 para as notas obtidas em uma bateria de seis avaliações educacionais aplicada a 112

alunos (Bartholomew & Knott, 1990), exibido na Figura 3.2. Tal modelo descreve dois fatores correlacionados como os responsáveis por grande parte da variabilidade de 6 variáveis observadas, que referem-se às 6 notas obtidas por cada aluno na bateria de 6 avaliações. O número de 2 fatores foi hipotetizado a partir de uma Análise Fatorial Exploratória, assim como a associação da variável  $x_1$  a ambos os fatores, das variáveis  $x_5$  e  $x_6$  ao fator  $\xi_1$  e das demais ao fator  $\xi_2$ . Tais hipóteses dizem respeito ao modelo de mensuração dos fatores  $\xi_1$  e  $\xi_2$  a partir das variáveis observadas  $x_1$  a  $x_6$  e pode ser testado por meio de uma Análise Fatorial Confirmatória.

O modelo é definido como

$$\mathbf{x} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \quad (4.12)$$

onde as matrizes  $\mathbf{\Lambda}$ ,  $\boldsymbol{\Phi}$  ou  $\boldsymbol{\Theta}_\delta$  configuram-se da seguinte forma:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{11}^x & \lambda_{12}^x \\ 0 & \lambda_{22}^x \\ 0 & \lambda_{32}^x \\ 0 & \lambda_{42}^x \\ \lambda_{51}^x & 0 \\ \lambda_{61}^x & 0 \end{pmatrix}, \quad \boldsymbol{\Phi} = \begin{pmatrix} 1 & \\ \phi_{21} & 1 \end{pmatrix} \quad \text{e} \quad \boldsymbol{\Theta}_\delta = \begin{pmatrix} \delta_{11} & & & & & \\ 0 & \delta_{22} & & & & \\ 0 & 0 & \delta_{33} & & & \\ 0 & 0 & 0 & \delta_{44} & & \\ 0 & 0 & 0 & 0 & \delta_{55} & \\ 0 & 0 & 0 & 0 & 0 & \delta_{66} \end{pmatrix} \quad (4.13)$$

Portanto, os parâmetros a serem estimados são as 7 cargas ( $\lambda$ 's), a covariância entre os fatores ( $\phi_{21}$ ) e os 6 erros de mensuração ( $\delta$ 's), perfazendo um total de 14 parâmetros, o que atende à regra  $t$ , uma vez que

$$t = 14 < \frac{6(6+1)}{2} = 21.$$

Como essa condição é necessária, mas não suficiente para a identificação do modelo, é preciso verificar as demais regras. Entretanto, a verificação das demais regras são válidas apenas para modelos de complexidade 1, isto é, modelos em que cada variável observada está associada a apenas 1 dos fatores. No problema em questão, a variável  $x_1$  foi modelada como sendo influenciada pelos 2 fatores do modelo, uma vez que essa configuração para  $x_1$  era bastante evidente na Análise Fatorial Exploratória aplicada previamente e é, sobretudo, coerente com o problema modelado, já que tal variável

trata-se da nota em uma avaliação de conhecimentos gerais, e não específicos de um ou outro fator, como as demais provas.

Diante dessa impossibilidade de aplicar as regras de verificação de condições suficientes para a identificabilidade do modelo, é preciso verificá-la diretamente, a partir do confronto entre a matriz de covariâncias implicada pelo modelo  $\Sigma(\theta)$  e a matriz de covariâncias populacional  $\Sigma$ , que será aproximada pela amostral  $\mathbf{S} = \hat{\Sigma}$ . A Figura 4.1 exibe a matriz de covariâncias implicada pelo modelo, conforme a equação (4.11). Se chamarmos de  $\Sigma_{ij}$  o elemento da  $i$ -ésima linha e  $j$ -ésima coluna da matriz  $\Sigma$ , seguem da equação (4.8) expressões identificadas para os parâmetros do modelo.

Das equações

$$\lambda_{22}\lambda_{32} = \Sigma_{32} \quad , \quad \lambda_{32}\lambda_{42} = \Sigma_{13} \quad \text{e} \quad \lambda_{42}\lambda_{22} = \Sigma_{42}$$

obtemos expressões identificadas para  $\lambda_{22}$ ,  $\lambda_{32}$  e  $\lambda_{42}$ :

$$\lambda_{22}^2 = \frac{\Sigma_{32}\Sigma_{42}}{\Sigma_{43}} \quad , \quad \lambda_{32}^2 = \frac{\Sigma_{43}\Sigma_{32}}{\Sigma_{42}} \quad \text{e} \quad \lambda_{42}^2 = \frac{\Sigma_{42}\Sigma_{43}}{\Sigma_{32}} .$$

Das equações

$$\lambda_{51}\phi_{21} = \frac{\Sigma_{52}}{\lambda_{22}} \quad , \quad \lambda_{61}\phi_{21} = \frac{\Sigma_{62}}{\lambda_{22}} \quad \text{e} \quad \lambda_{51}\lambda_{61} = \Sigma_{65} ,$$

obtemos expressões identificadas para  $\phi_{21}$ ,  $\lambda_{51}$  e  $\lambda_{61}$ :

$$\phi_{21}^2 = \frac{\Sigma_{52}\Sigma_{62}\Sigma_{43}}{\Sigma_{65}\Sigma_{32}\Sigma_{42}} \quad , \quad \lambda_{51}^2 = \frac{\Sigma_{52}\Sigma_{65}}{\Sigma_{62}} \quad \text{e} \quad \lambda_{61}^2 = \frac{\Sigma_{65}\Sigma_{62}}{\Sigma_{52}} .$$

Finalmente, das equações

$$\lambda_{42}\phi_{21}\lambda_{11} + \lambda_{42}\lambda_{12} = \Sigma_{41} \quad \text{e} \quad \lambda_{51}\lambda_{11} + \lambda_{51}\phi_{21}\lambda_{12} = \Sigma_{51}$$

obtemos expressões identificadas para  $\lambda_{11}$  e  $\lambda_{12}$ :

$$\lambda_{11}^2 = \frac{\Sigma_{32}^2\Sigma_{42}\Sigma_{62}\Sigma_{65}(\Sigma_{51}^2\Sigma_{42} - \Sigma_{52}^2\Sigma_{41}^2\Sigma_{32}\Sigma_{65})}{(\Sigma_{65}\Sigma_{32}\Sigma_{42} - \Sigma_{52}\Sigma_{62}\Sigma_{43})^2} \quad \text{e} \quad \lambda_{12} = \frac{\Sigma_{41}\Sigma_{32}}{\Sigma_{42}\Sigma_{43}} - \lambda_{11} \left( \frac{\Sigma_{52}\Sigma_{62}\Sigma_{43}}{\Sigma_{65}\Sigma_{32}\Sigma_{42}} \right)^{1/2} .$$

Com tais parâmetros identificados, é fácil notar que os  $\delta_{ii}$  são facilmente obtidos após a substituição dos demais parâmetros pelas expressões apresentadas. Se tal abordagem para verificação da identificabilidade fosse inviável devido a uma quantidade excessiva de parâmetros, por exemplo, poderiam ser utilizados os testes empíricos mencionados na Seção 4.1.1, aplicáveis tanto aos MEEVO quanto à AFC, lembrando que verificam



apenas a identificabilidade *local* e não *global*, diferentemente do método algébrico aqui utilizado.

As expressões obtidas para cada parâmetro não são únicas neste problema, sendo tal modelo *sobre-identificado*. Para que o ajuste do modelo seja possível, entretanto, é suficiente garantir a existência de solução, ainda que não seja única. Utilizando a biblioteca `lavaan` (Rosseel, 2011) do R, foi possível estimar valores para os parâmetros do modelo proposto, por meio da função `cfa()`. As estimativas são mostradas na Tabela 4.1, onde também são apresentadas algumas medidas da qualidade do ajuste, calculadas com a função `fitMeasures()`, da mesma biblioteca. O código utilizado encontra-se no Apêndice deste Capítulo. A Figura 4.2, por sua vez, exhibe o diagrama de caminho do modelo ajustado, onde constam os parâmetros estimados. Cabe notar que a escala dos parâmetros é herdada das variáveis latentes, para as quais foi assumida média nula e variância unitária.

Uma breve análise dos parâmetros estimados mostra que todos são significativos, exceto possivelmente  $\delta_{55}$ , apresentando  $z = 1,454$  (vide Tabela 4.1). As medidas de qualidade do ajuste são adequadas, com  $\chi^2$  baixo, de  $p$ -valor 0,416, indicando um bom ajuste, além de CFI<sup>2</sup> máximo (1) e RMSEA de 0,013, bastante inferior aos limites máximos recomendados de 0,050 a 0,100. Verificamos que as variáveis  $x_3$  e  $x_6$  são as que apresentam maiores cargas, respectivamente dos fatores  $\xi_2$  e  $\xi_1$ . Por outro lado, são as variáveis que apresentam maiores erros de mensuração (32,964 e 38,678, respectivamente). As outras variáveis possuem cargas menos altas, mas também erros de mensuração mais aceitáveis. Finalmente, vale notar que a covariância entre os fatores foi estimada em 0,440, com erro-padrão de 0,094, indicando que esse parâmetro não deve ser nulo, isto é, esses fatores não devem ser ortogonais. Tal constatação corrobora aquela feita na Análise Fatorial Exploratória, quando as soluções rotacionadas oblíquas apresentaram melhores resultados que as soluções ortogonais.

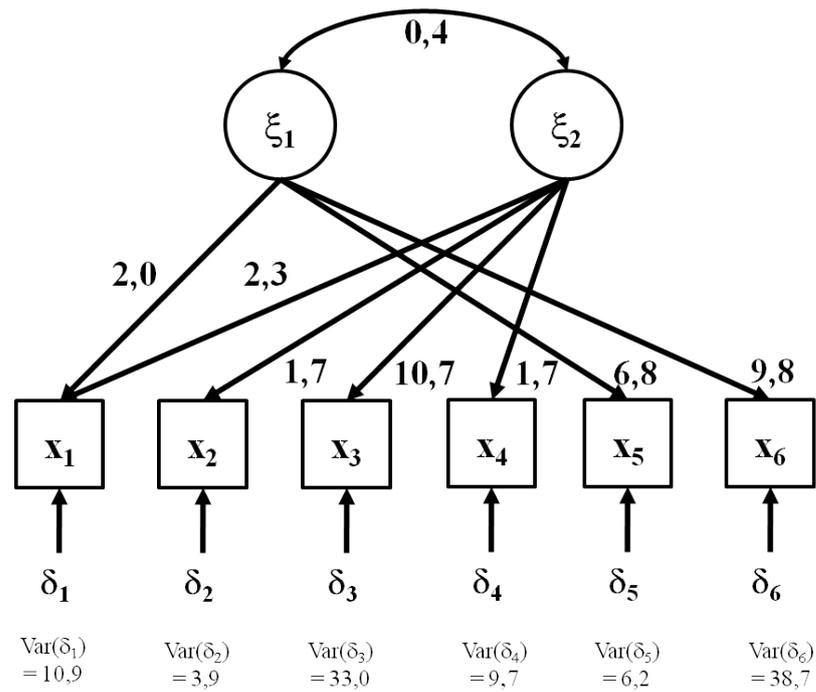
Como a AFC constitui um caso particular dos Modelos de Equações Estruturais com Variáveis Latentes, todos os *softwares* apresentados na Seção 4.1.3 também prestam-se à estimação de parâmetros dos MEEVO e das AFC.

---

<sup>2</sup>*Comparative Fit Index*

**Tabela 4.1:** Parâmetros estimados para o modelo de Análise Fatorial Confirmatória relativo às notas obtidas em uma bateria de seis avaliações educacionais aplicada a 112 alunos (Bartholomew & Knott, 1990).

Parâmetro	Estimativa	Erro-padrão	Z	Pr(>  z )
<b>Cargas das variáveis latentes:</b>				
<i>Inteligência verbal (<math>\xi_1</math>)</i>				
general ( $\lambda_{11}$ )	2,014	0,438	4,599	0,000
reading ( $\lambda_{51}$ )	6,774	0,598	11,332	0,000
vocab ( $\lambda_{61}$ )	9,768	0,986	9,905	0,000
<i>Inteligência espacial (<math>\xi_2</math>)</i>				
general ( $\lambda_{12}$ )	2,309	0,460	5,021	0,000
picture ( $\lambda_{22}$ )	1,659	0,243	6,814	0,000
blocks ( $\lambda_{32}$ )	10,748	1,122	9,584	0,000
maze ( $\lambda_{42}$ )	1,697	0,347	4,887	0,000
<b>Covariâncias:</b>				
$\phi_{21}$	0,440	0,094	4,698	0,000
<b>Variâncias:</b>				
general ( $\delta_{11}$ )	10,938	1,735	6,305	0,000
picture ( $\delta_{22}$ )	3,889	0,636	6,112	0,000
blocks ( $\delta_{33}$ )	32,964	15,043	2,191	0,028
maze ( $\delta_{44}$ )	9,718	1,388	7,000	0,000
reading ( $\delta_{55}$ )	6,247	4,295	1,454	0,146
vocab ( $\delta_{66}$ )	38,678	10,172	3,802	0,000
$\phi_{11}$	1,000			
$\phi_{22}$	1,000			
<b>Medidas de qualidade do ajuste:</b>				
$\chi^2$	7,125			
Graus de liberdade	7			
$p$ -valor da estatística $\chi^2$	0,416			
CFI	1,000			
RMSEA	0,013			



**Figura 4.2:** Diagrama de caminho exibindo estimativas dos parâmetros estimados para o modelo de Análise Fatorial Confirmatória relativo às notas obtidas em uma bateria de seis avaliações educacionais aplicada a 112 alunos (Bartholomew & Knott, 1990). A escala dos parâmetros é herdada das variáveis latentes, para as quais foi assumida média nula e variância unitária.

### 4.1.3 LISREL: Modelo de Equações Estruturais com Variáveis Latentes

O Modelo de Equações Estruturais com Variáveis Latentes é basicamente a união dos dois modelos previamente apresentados. De um lado, tem-se um modelo estrutural, formado por múltiplas equações lineares que associam entre si as variáveis em estudo, mas que, por outro lado, contempla a possibilidade de tais variáveis serem latentes e, nesse caso, terem um modelo de mensuração associado a cada uma delas, de modo que também a relação entre elas e as variáveis observadas seja especificada no modelo.

A equação geral de um LISREL é caracterizada por

$$\begin{aligned}\boldsymbol{\eta} &= \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \\ \mathbf{x} &= \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \\ \mathbf{y} &= \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\epsilon},\end{aligned}$$

sendo as equações (2.2), (2.3) e (2.4) um caso particular delas, em que a matriz  $\mathbf{B}$  é nula e há somente uma variável latente exógena ( $\xi_1$ ) e uma endógena ( $\eta_1$ ), conforme ilustrado no diagrama de caminho da Figura 2.2.

Um modelo mais complexo e que contempla todos os atributos da LISREL é apresentado em Bollen (1989, caps. 2,7,8), para dados de industrialização e democracia política. Esses dados de painel representam a relação entre a democracia política nos anos de 1960 e 1965 e a industrialização em 1960, em países em desenvolvimento. No estudo transnacional a respeito de democracia política, a industrialização é comumente vista como um fator que aumenta as chances de existência de um modelo político democrático. Ainda, o nível atual de democracia política costuma ser influenciado pelo nível anterior. Tais ideias sugerem que o modelo de variáveis latentes deva contemplar o nível de democracia em 1960 ( $\eta_1$ ) e a industrialização em 1960 ( $\xi_1$ ) como influenciadores da democracia em 1965 ( $\eta_2$ ). Além disso, contemplou-se um efeito da industrialização contemporânea influenciando na democracia de 1960.

Da descrição, segue que o modelo estrutural é dado por

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \beta_{21} & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \gamma_{11} \\ \gamma_{21} \end{pmatrix} \xi_1 + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}, \quad (4.14)$$

e assume-se que os erros latentes ( $\zeta$ 's) não são correlacionados entre si, de modo que  $\Psi$  é uma matriz diagonal. Quanto à matriz  $\Phi$ , ela é composta apenas pelo escalar  $\phi_{11}$ , devido à existência de apenas uma variável latente exógena no modelo.

Para o modelo de mensuração, adotaram-se quatro variáveis para a aferição da democracia, sendo elas medidas em 1960 para a variável latente  $\eta_1$  e em 1965 para a variável latente  $\eta_2$ . São elas:

$y_1$  e  $y_5$ : liberdade de imprensa;

$y_2$  e  $y_6$ : liberdade de grupos de oposição;

$y_3$  e  $y_7$ : justiça do processo eleitoral; e

$y_4$  e  $y_8$ : natureza eletiva e efetividade do corpo legislativo.

Quanto às restrições feitas, neste exemplo, cada fator herdará a escala da primeira variável observada associada a ele. No caso de  $\eta_1$  e  $\eta_2$ , portanto, será atribuída carga 1 às variáveis  $y_1$  e  $y_5$ , respectivamente. Além disso, como  $y_i$  e  $y_{i+4}$ ,  $i \in 1, 2, 3, 4$  representam a mesma variável em duas épocas distintas, assume-se que a carga de  $\eta_1$  em  $y_i$  é igual à carga de  $\eta_2$  em  $y_{i+4}$ . Quanto à matriz de covariâncias dos erros de mensuração, assume-se que  $y_i$  e  $y_{i+4}$ ,  $i \in 1, 2, 3, 4$  são possivelmente correlacionados, sendo esses parâmetros livres na matriz. Além desses, consideram-se correlacionados os pares  $(y_2, y_4)$  e  $(y_6, y_8)$ , pois o erro de mensuração das duas variáveis são advindos da mesmo fonte. Em suma, na matriz triangular inferior  $\Theta_\epsilon$ , são livres todos os elementos da diagonal principal (variâncias), além daqueles nas posições (5,1), (6,2), (7,3), (8,4), (4,2) e (8,6). A equação do modelo de mensuração para  $\boldsymbol{\eta}$  é, portanto:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda_{22} & 0 \\ \lambda_{33} & 0 \\ \lambda_{44} & 0 \\ 0 & 1 \\ 0 & \lambda_{22} \\ 0 & \lambda_{33} \\ 0 & \lambda_{44} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \end{pmatrix} \quad (4.15)$$

O modelo de mensuração da industrialização ( $\xi_1$ ) é mais simples, sendo composto por três variáveis observadas:

$x_1$ : produto interno bruto (PIB) *per capita*;

$x_2$ : consumo de energia *per capita*; e

$x_3$ : porcentagem da força de trabalho em ocupações industriais.

Utilizou-se o logaritmo das variáveis  $x_1$  e  $x_2$  e tomou-se o arco seno da raiz quadrada de  $x_3$ , a fim de que os valores tornassem-se mais próximos da distribuição normal. A escala de  $\xi_1$  é herdada de  $x_1$  fixando sua carga nessa variável em 1. Para a matriz de covariâncias dos erros de mensuração ( $\Theta_\delta$ ), assume-se que seja diagonal, já que não há suspeita de que as variáveis observadas sejam correlacionadas. Sendo assim, a equação do modelo de mensuração é dada por

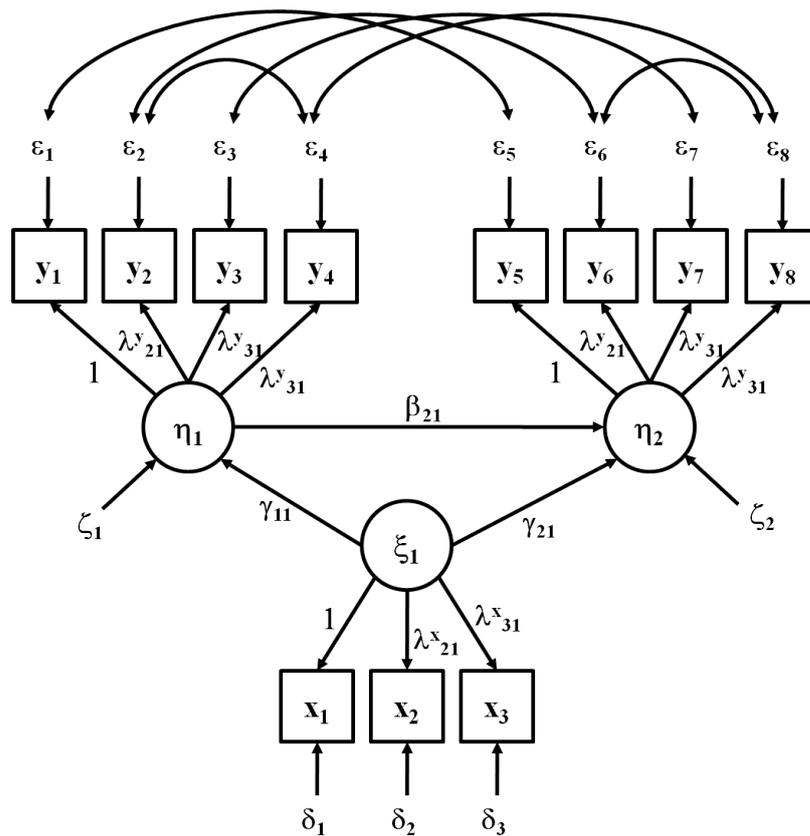
$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ \lambda_{21} \\ \lambda_{31} \end{pmatrix} \xi_1 + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix} \quad (4.16)$$

A Figura 4.3 exibe o diagrama de caminho do modelo descrito e, a seguir, será apresentada a matriz implicada pelos modelos LISREL e algumas regras de identificabilidade, para que, com os dados em mãos, seja possível dar início ao processo de estimação.

### Matriz de covariâncias implicada pelo modelo

A matriz de covariâncias implicada pelo modelo LISREL é similar à dos MEEVO, já que possuem estruturas similares, salvo pelo fato de que no LISREL as variáveis que participam da regressão são consideradas latentes e seus modelos de mensuração fazem parte da especificação do modelo, por meio das matrizes  $\Lambda_x$ ,  $\Lambda_y$ ,  $\Theta_\delta$  e  $\Theta_\epsilon$ . São precisamente essas matrizes que passam a figurar nas expressões que compõem a matriz de covariâncias do LISREL e a diferenciam daquela calculada para o MEEVO e apresentada na equação (4.7).

Usando a notação da Seção 4.1.1, obtemos



**Figura 4.3:** Modelo de Equações Estruturais para os dados de industrialização e democracia política em países em desenvolvimento (Bollen, 1989, caps. 2,7,8).

$$\begin{aligned}
\Sigma_{yy}(\boldsymbol{\theta}) &= \mathbb{E}(\mathbf{Y}\mathbf{Y}') \\
&= \mathbb{E}\left[(\Lambda_y\boldsymbol{\eta} + \boldsymbol{\epsilon})\left(\boldsymbol{\eta}'\Lambda_y' + \boldsymbol{\epsilon}'\right)\right] \\
&= \Lambda_y\mathbb{E}(\boldsymbol{\eta}\boldsymbol{\eta}')\Lambda_y' + \Theta_\epsilon \\
&= \Lambda_y(\mathbf{I} - \mathbf{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)\left[(\mathbf{I} - \mathbf{B})^{-1}\right]'\Lambda_y' + \Theta_\epsilon. \quad (4.17)
\end{aligned}$$

Para a covariância entre  $\mathbf{X}$  e  $\mathbf{Y}$ , obtemos

$$\begin{aligned}
\Sigma_{yx}(\boldsymbol{\theta}) &= \mathbb{E}(\mathbf{Y}\mathbf{X}') \\
&= \mathbb{E}\left[(\Lambda_y\boldsymbol{\eta} + \boldsymbol{\epsilon})(\boldsymbol{\xi}'\Lambda_x' + \boldsymbol{\delta}')\right] \\
&= \Lambda_y\mathbb{E}(\boldsymbol{\eta}\boldsymbol{\xi}')\Lambda_x' \\
&= \Lambda_y(\mathbf{I} - \mathbf{B})^{-1}\Gamma\Phi\Lambda_x'. \quad (4.18)
\end{aligned}$$

Analogamente às matrizes calculadas previamente para a AFE e a AFC e exibidas nas equações (3.5) e (4.11), obtemos

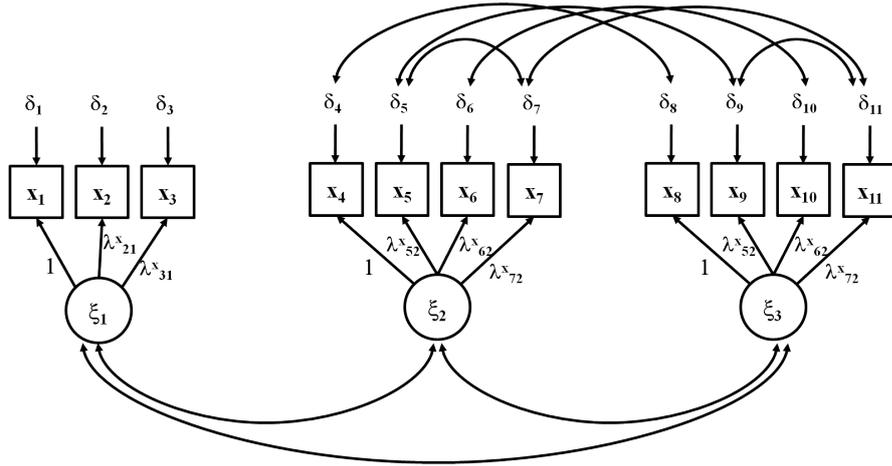
$$\begin{aligned}
\Sigma_{xx}(\boldsymbol{\theta}) &= \mathbb{E}(\mathbf{x}\mathbf{x}') \\
&= \Lambda_x\Phi\Lambda_x' + \Theta_\delta \quad (4.19)
\end{aligned}$$

A matriz de covariâncias implicada pelo modelo é obtida a partir dos resultados (4.17), (4.18) e (4.19):

$$\Sigma(\boldsymbol{\theta}) = \begin{pmatrix} \Lambda_y(\mathbf{I} - \mathbf{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)\left[(\mathbf{I} - \mathbf{B})^{-1}\right]'\Lambda_y' + \Theta_\epsilon & \Lambda_y(\mathbf{I} - \mathbf{B})^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'\left[(\mathbf{I} - \mathbf{B})^{-1}\right]'\Lambda_y' & \Lambda_x\Phi\Lambda_x' + \Theta_\delta \end{pmatrix}. \quad (4.20)$$

## Identificabilidade

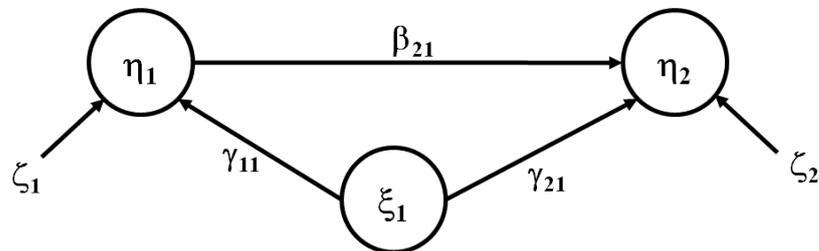
Nos modelos LISREL, a questão da identificabilidade é tão importante quanto no MEEVO e na AFC. Por se tratar de um caso geral constituído por uma espécie de “combinação” dos outros dois modelos, a identificabilidade pode ser verificada desmembrando o modelo em um MEEVO e uma AFC, para os quais a identificabilidade é verificada separadamente. Nesse caso, o LISREL será identificado na medida em que o MEEVO e a AFC derivados dele forem também identificados.



**Figura 4.4:** Modelo de Equações Estruturais para os dados de industrialização e democracia política em países em desenvolvimento: isolamento do modelo de mensuração.

Essa *regra dos dois passos* pode ser aplicada ao modelo das equações (4.14), (4.15) e (4.16). A Figura 4.5 exibe o resultado do isolamento das variáveis estruturais do modelo, ao passo que a Figura 4.4 exibe o modelo de mensuração isolado do modelo estrutural. Para a Figura 4.5, verificamos a identificabilidade aplicando a *regra do modelo recursivo*. Como a matriz  $\mathbf{B}$  do modelo estrutural é triangular inferior, segue que o modelo é identificado. Quanto ao modelo de mensuração, a verificação é mais trabalhosa.

As condições suficientes para verificação da identificabilidade dos modelos de men-



**Figura 4.5:** Modelo de Equações Estruturais para os dados de industrialização e democracia política em países em desenvolvimento: isolamento do modelo estrutural.

suração, apresentadas na Seção 4.1.2 exigem que  $\Theta_\delta$  seja diagonal, o que não é atendido no modelo hipotetizado para os dados de democracia política e industrialização, uma vez que os erros de mensuração são correlacionados. Nesse caso, é preciso verificar a identificabilidade algebricamente, como mostrado na Seção 4.1.2. Bollen (1989, p.251-253) mostra algebricamente que o modelo de mensuração da Figura 4.4 é sobre-identificado. Logo, da identificação do modelo estrutural e do modelo de mensuração, segue que o modelo LISREL composto por ambos também é identificado, de acordo com a *regra dos dois passos*.

Além da regra dos dois passos, a *regra* MIMIC assegura que todos os modelos do tipo MIMIC<sup>3</sup> (Goldberger, 1971; Robinson, 1974) são identificados. Tais modelos caracterizam-se por considerarem as variáveis observadas associadas à variável latente exógena, suas *causas* e não suas consequências, como nos modelos de mensuração tradicionais, a exemplo da AFC.

Em um modelo MIMIC, o modelo de mensuração para a variável  $\mathbf{x}$  é dado por

$$\mathbf{x} = \mathbf{I}\boldsymbol{\xi} ,$$

de modo que as variáveis observadas não possuam erro de mensuração e sejam equivalentes aos próprios  $\xi$ 's, devido às cargas unitárias atribuídas pela matriz identidade. Nesse caso, cada  $\xi_i$ , que nos modelos LISREL é por definição a causa de um ou mais  $\eta_j$ , é aferido sem erro de mensuração pela variável observada correspondente ( $x_i$ ), fazendo com que esta observação desempenhe o papel de causa direta de uma ou mais variáveis endógenas ( $\eta_j$ ).

Regras adicionais podem ser úteis na verificação da identificabilidade de um modelo e podem ser encontradas em Bollen & Davis (2009a,b); Muthen & Satorra (1995).

### ***Softwares e exemplo de resolução***

A estimação dos parâmetros de um modelo LISREL pode ser feita utilizando o R, por meio da biblioteca `sem` (Fox *et al.*, 2010) ou `lavaan` (Rosseel, 2011), por exemplo. Em ambas as bibliotecas, a função apropriada possui o mesmo nome, `sem()`, sendo a entrada de argumentos solicitada pela função da biblioteca `lavaan` mais amigável e o *output* mais completo (cf. Seção 4.5). No SAS, o procedimento CALIS pode ser

---

<sup>3</sup>Acrônimo para *Multiple Indicators Multiple Causes*.

utilizado para o mesmo fim. Outros *softwares* foram desenvolvidos exclusivamente com o propósito de ajustar modelos de equações estruturais com variáveis latentes, como o LISREL (SSI, Lincolnwood, IL – EUA), o AMOS (SPSS, Chicago, IL – EUA), o EQS (Multivariate Software, Encino, CA – EUA) e o Mplus (Muthén & Muthén, Los Angeles, CA – EUA). Esses *softwares* aceitam inclusive diagramas de caminho como forma de especificação do modelo.

As Tabelas 4.2 e 4.3 apresentam os parâmetros estimados pelo método da Máxima Verossimilhança para o modelo referente aos dados de industrialização e democracia em países em desenvolvimento. As estimativas coincidem com aquelas encontradas em Bollen (1989, p.335). Nelas, podemos constatar a qualidade do ajuste, por meio de um favorável teste  $\chi^2$  ( $p$ -valor 0,374), do índice CFI próximo de seu máximo (= 0,997) e do índice RMSEA inferior a 0,05 (= 0,028).

#### 4.1.4 Estimação dos parâmetros estruturais e incidentais

Os parâmetros estruturais dos MEE clássicos são normalmente estimados por um dentre três métodos mais difundidos: método da Máxima Verossimilhança, método dos Mínimos Quadrados ordinário ou método dos Mínimos Quadrados generalizado. Na abordagem por Mínimos Quadrados, a estimação baseia-se na obtenção de valores para os parâmetros que tornem a matriz  $\Sigma(\theta)$  o mais próxima possível de  $\hat{\Sigma} = \mathbf{S}$ . Tal finalidade é alcançada por meio da minimização de uma determinada função de resíduos.

Na abordagem por Máxima Verossimilhança, assume-se que o vetor de variáveis observadas ( $\mathbf{x}'|\mathbf{y}'$ ) segue uma distribuição normal multivariada e define-se a função de log-verossimilhança como

$$\log L(\theta) = \text{constante} - \frac{N}{2} \{ \log |\Sigma(\theta)| + \text{tr} [\mathbf{S}^* \Sigma^{-1}(\theta)] \} ,$$

onde  $\mathbf{S}^* = \frac{N-1}{N} \mathbf{S}$ , com as matrizes  $\Sigma(\theta)$  e  $\mathbf{S}$  positivas definidas e, portanto, não-singulares. A máxima verossimilhança será máxima quando a matriz de covariâncias estimada  $\Sigma(\hat{\theta})$  coincidir com a matriz de covariâncias amostral  $\mathbf{S}$ . Por isso é comum que no contexto dos MEE o método da Máxima Verossimilhança seja representado por uma função a ser minimizada, que assume seu valor mínimo, zero, quando  $\Sigma(\hat{\theta}) = \mathbf{S}$ .

**Tabela 4.2:** Parâmetros estimados e medidas de qualidade do ajuste, para o modelo de Equações Estruturais para os dados de industrialização e democracia política em países em desenvolvimento (Bollen, 1989, caps. 2,7,8) [continua na Tabela 4.3].

Parâmetro	Estimativa	Erro-padrão	Z	Pr(>  z )
<b>Cargas das variáveis latentes:</b>				
<i>Industrialização em 1960 (<math>\xi_1</math>):</i>				
$x_1$ ( $\lambda_{11}^x$ )	1,000			
$x_2$ ( $\lambda_{21}^x$ )	2,18	0,138	15,751	0,000
$x_3$ ( $\lambda_{31}^x$ )	1,818	0,152	11,971	0,000
<i>Democracia política em 1960 (<math>\eta_1</math>):</i>				
$y_1$ ( $\lambda_{11}^y$ )	1,000			
$y_2$ ( $\lambda_{21}^y$ )	1,191	0,139	8,551	0,000
$y_3$ ( $\lambda_{31}^y$ )	1,175	0,12	9,755	0,000
$y_4$ ( $\lambda_{41}^y$ )	1,251	0,117	10,712	0,000
<i>Democracia política em 1965 (<math>\eta_2</math>):</i>				
$y_5$ ( $\lambda_{11}^y$ )	1,000			
$y_6$ ( $\lambda_{21}^y$ )	1,191	0,139	8,551	0,000
$y_7$ ( $\lambda_{31}^y$ )	1,175	0,12	9,755	0,000
$y_8$ ( $\lambda_{41}^y$ )	1,251	0,117	10,712	0,000
<b>Coefficientes de regressão:</b>				
$\gamma_{11}$	1,471	0,392	3,75	0,000
$\gamma_{21}$	0,6	0,226	2,661	0,008
$\beta_{21}$	0,865	0,075	11,554	0,000

(continua)

**Tabela 4.3:** Parâmetros estimados e medidas de qualidade do ajuste, para o modelo de Equações Estruturais para os dados de industrialização e democracia política em países em desenvolvimento (Bollen, 1989, caps. 2,7,8) [continuação da Tabela 4.2].

Parâmetro	Estimativa	Erro-padrão	Z	Pr(>  z )
<b>Covariâncias:</b>				
$\epsilon_{51}$	0,583	0,356	1,637	0,102
$\epsilon_{62}$	2,183	0,737	2,96	0,003
$\epsilon_{73}$	0,712	0,611	1,165	0,244
$\epsilon_{84}$	0,363	0,444	0,817	0,414
$\epsilon_{42}$	1,44	0,689	2,092	0,036
$\epsilon_{86}$	1,372	0,577	2,378	0,017
<b>Variâncias:</b>				
$\delta_{11}$	0,081	0,019	4,182	0,000
$\delta_{22}$	0,12	0,07	1,729	0,084
$\delta_{33}$	0,467	0,09	5,177	0,000
$\epsilon_{11}$	1,855	0,433	4,279	0,000
$\epsilon_{22}$	7,581	1,366	5,549	0,000
$\epsilon_{33}$	4,956	0,956	5,182	0,000
$\epsilon_{44}$	3,224	0,723	4,458	0,000
$\epsilon_{55}$	2,313	0,479	4,831	0,000
$\epsilon_{66}$	4,968	0,921	5,393	0,000
$\epsilon_{77}$	3,56	0,71	5,018	0,000
$\epsilon_{88}$	3,308	0,704	4,701	0,000
$\phi_{11}$	0,449	0,087	5,175	0,000
$\psi_{11}$	3,875	0,866	4,477	0,000
$\psi_{22}$	0,164	0,227	0,725	0,469
<b>Medidas de qualidade do ajuste:</b>				
$\chi^2$	40,179			
Graus de liberdade	38			
p-valor	0,374			
CFI	0,997			
RMSEA	0,028			

As funções de estimação a serem minimizadas na aplicação do método da Máxima Verossimilhança é apresentada na equação (4.21). As equações (4.22) e (4.23) apresentam as funções referentes ao método de Mínimos Quadrados ordinário e generalizado, respectivamente.

$$F_{MV}(\boldsymbol{\theta}) = \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \text{tr} [\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})] \quad (4.21)$$

$$F_{MQO}(\boldsymbol{\theta}) = \frac{1}{2} \text{tr} \{[\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})]^2\} \quad (4.22)$$

$$F_{MQG}(\boldsymbol{\theta}) = \frac{1}{2} \text{tr} \{[\mathbf{I} - \boldsymbol{\Sigma}(\boldsymbol{\theta})\mathbf{S}^{-1}]^2\} \quad (4.23)$$

Vale lembrar que, para modelos perfeitamente identificados, a solução das equações de estimação é única e a função de resíduos vale zero, mas na maior parte dos casos, em que o modelo é sobre-identificado, o conjunto escolhido de parâmetros será determinado segundo o critério que cada abordagem de estimação apresenta, ora o de Mínimos Quadrados, ora o de Máxima Verossimilhança.

Os estimadores de Máxima Verossimilhança, no entanto, destacam-se por possuírem importantes propriedades. Embora possam ser viesados em amostras pequenas, são assintoticamente não-viesados. Além disso, possuem invariância de escala e são assintoticamente consistentes e eficientes, sendo a matriz de covariâncias assintótica do estimador de  $\boldsymbol{\theta}$  dada por

$$\left( \frac{2}{N-1} \right) \left[ \mathbb{E} \left( \frac{\partial^2 F_{MV}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right) \right]^{-1}.$$

Em relação aos parâmetros incidentais, isto é, aos escores dos fatores, os métodos de estimá-los são diversos. O mais popular é, provavelmente, o método da regressão, com estimador

$$\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\Phi}} \hat{\boldsymbol{\Lambda}}'_x \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{x}, \quad (4.24)$$

onde  $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}_x \boldsymbol{\Phi} \boldsymbol{\Lambda}'_x + \boldsymbol{\Theta}_\delta$  é o peso que multiplica  $\mathbf{x}$  é o estimador de Mínimos Quadrados dos coeficientes da regressão hipotética entre  $\boldsymbol{\xi}$  e  $\mathbf{x}$ .

### 4.1.5 Limitações

Embora os modelos de Equações Estruturais apresentem estimativas não-viesadas para os coeficientes de regressão entre escores de fatores, ele também apresenta algumas limitações:

1. Modelos que envolvem muitas variáveis observadas podem ser problemáticos, na medida em que a verificação algébrica da identificabilidade do modelo pode ser proibitiva, caso ele não se enquadre nas regras de identificabilidade preestabelecidas.
2. O modelo clássico considera que as variáveis observadas são contínuas, não comportando dados que contenham variáveis categorizadas.
3. Ampliando a advertência anterior, podemos mencionar que modelos de mensuração não-lineares ou lineares generalizados não são contemplados pelo modelo clássico.
4. O modelo clássico também não contempla relações estruturais não-lineares ou lineares generalizadas.

## 4.2 Regressão clássica entre escores de fatores (2001)

A Seção 4.1 mostrou como pode ser trabalhoso estimar os parâmetros de um modelo de regressão entre escores de fatores, principalmente se o número de variáveis for elevado ou se não se dispor de *softwares* específicos para o cálculo das estimativas. Daí a importância de se conhecer casos particulares em que os escores de fatores possam tomar parte em uma regressão com coeficientes estimados por técnicas simples, como o convencional método dos Mínimos Quadrados.

Skrondal & Laake (2001) obtiveram resultados interessantes que podem ser aplicados no ajuste de modelos de regressão entre fatores, que permite tratá-los como se fossem variáveis sem erro de mensuração e calcular estimadores assintoticamente consistentes por meio de técnicas simples.

A seguir, definir-se-á o modelo de Regressão Latente (RL), que subsidiará as abordagens

### 4.2.1 Modelo de Regressão Latente

O modelo de Regressão Latente (RL) é formado por um sistema de equações lineares, em que são fatores: as variáveis-resposta, as covariáveis ou ambas. Na notação LISREL, o modelo é descrito por

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \quad (4.25)$$

onde  $\boldsymbol{\eta}$  representa os escores dos fatores que comporão o sistema como variáveis-resposta (i.e., variáveis endógenas),  $\boldsymbol{\xi}$  representa os escores dos fatores que comporão o sistema como covariáveis (i.e., variáveis exógenas) e  $\boldsymbol{\zeta}$  é o vetor de erros aleatórios, sendo todos os valores dados em forma de desvios em relação à media. Temos, portanto, que o modelo de RL é um caso particular do LISREL, no qual

$$\mathbf{B} = \mathbf{0} \quad (4.26)$$

e, por se tratarem de variáveis latentes, os fatores  $\boldsymbol{\eta}$  e  $\boldsymbol{\xi}$  são aferidos segundo um modelo de mensuração descrito por

$$\mathbf{y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon} \quad (4.27)$$

$$\mathbf{x} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}. \quad (4.28)$$

Sendo assim, a matriz de covariâncias das variáveis observadas pode ser definida, a partir de (4.20) e (4.26), como

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} \Sigma_{yy}(\boldsymbol{\theta}) & \Sigma_{yx}(\boldsymbol{\theta}) \\ \Sigma'_{yx}(\boldsymbol{\theta}) & \Sigma_{xx}(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Lambda}_y(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})\boldsymbol{\Lambda}'_y + \boldsymbol{\Theta}_\epsilon & \boldsymbol{\Lambda}_y\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Lambda}'_x \\ \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'_y & \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}'_x + \boldsymbol{\Theta}_\delta \end{pmatrix}. \quad (4.29)$$

Além do fato de  $\mathbf{B} = \mathbf{0}$  e, portanto, as variáveis endógenas de uma equação não poderem participar de outras equações como covariáveis, há outra diferença crucial da aplicação do modelo LISREL em relação à do modelo de RL, que diz respeito ao propósito do modelo de mensuração. No LISREL, o modelo de mensuração presta-se a possibilitar a estimativa conjunta dos parâmetros do modelo estrutural e do modelo de mensuração simultaneamente, a partir das variáveis observadas. No modelo de RL, o modelo de mensuração também orienta a estimação dos escores dos fatores, mas essa estimação se dá em uma etapa preliminar para que, em uma segunda etapa, sejam utilizados como variáveis de uma regressão comum, a fim de estimar os parâmetros estruturais, que são os coeficientes da regressão ( $\boldsymbol{\Gamma}$ ).

Se definirmos os escores teóricos da realização  $i$ ,  $i = 1, 2, \dots, N$  dos fatores  $\xi$  e  $\eta$  como

$$\mathbf{F}_{\xi,i} = \mathbf{A}_\xi \mathbf{x}_i \quad \text{e} \quad \mathbf{F}_{\eta,i} = \mathbf{A}_\eta \mathbf{y}_i ,$$

temos que  $\mathbf{A}_\xi$  e  $\mathbf{A}_\eta$  são as matrizes que aplicam sobre o vetor observado,  $\mathbf{x}$  e  $\mathbf{y}$ , respectivamente, a transformação linear que resulta nas estimativas dos escores desses fatores. Para o método de Bartlett, temos

$$\mathbf{A}_\xi^B = (\Lambda'_x \Theta_\delta^{-1} \Lambda_x)^{-1} \Lambda'_x \Theta_\delta^{-1} \quad (4.30)$$

e

$$\mathbf{A}_\eta^B = (\Lambda'_y \Theta_\epsilon^{-1} \Lambda_y)^{-1} \Lambda'_y \Theta_\epsilon^{-1} . \quad (4.31)$$

Para o método da Regressão, temos (cf. Seção 3.2.5, eqs. (3.14) e (3.15))

$$\mathbf{A}_\xi^R = \Phi \Lambda'_x \Sigma_x^{-1} \quad (4.32)$$

e

$$\mathbf{A}_\eta^R = (\Gamma \Phi \Gamma' + \Psi) \Lambda'_y \Sigma_y^{-1} . \quad (4.33)$$

A partir dessas definições, é possível realizar o ajuste de uma regressão de escores por meio de uma das duas abordagens já mencionadas, que são detalhadamente apresentadas a seguir.

### Regressão de escores bloco a bloco (revisada)

A regressão de escores bloco a bloco revisada consiste de um algoritmo de 3 passos:

**Passo 1: Análise Fatorial Confirmatória bloco a bloco** Um modelo de AFC é ajustado às variáveis observadas  $\mathbf{x}$ , de modo a obter as estimativas consistentes  $\hat{\Lambda}_x$ ,  $\hat{\Theta}_\delta$  e  $\hat{\Phi}$ . Analogamente, são obtidas as estimativas consistentes  $\hat{\Lambda}_y$  e  $\hat{\Theta}_\epsilon$  aplicando a AFC às variáveis observadas  $\mathbf{y}$ .

**Passo 2: Estimação dos escores dos fatores** Por meio do método de Bartlett, obtêm-se os escores empíricos  $\tilde{\mathbf{F}}_{\eta,i}$  para os fatores  $\eta$  e, por meio do método da Regressão, os escores empíricos  $\tilde{\mathbf{F}}_{\xi,i}$  para os fatores  $\xi$ .

**Passo 3: Regressões com MMQ** Os parâmetros estruturais  $\mathbf{\Gamma}$  são estimados para cada equação do sistema (4.25), por meio do método dos Mínimos Quadrados (MMQ). O estimador pode ser expresso por

$$\hat{\mathbf{\Gamma}} = \mathbf{S}_{\tilde{F}_\eta^B \tilde{F}_\xi^R} \left( \mathbf{S}_{\tilde{F}_\xi^R} \right)^{-1},$$

onde  $\mathbf{S}_{\tilde{F}_\eta^B \tilde{F}_\xi^R}$  e  $\mathbf{S}_{\tilde{F}_\xi^R}$  representam as matrizes de covariâncias empíricas para os escores dos fatores subscritos.

Skrondal & Laake (2001) provaram que o estimador  $\hat{\mathbf{\Gamma}}$  é consistente desde que a estimação dos escores dos fatores ocorra pelos métodos descritos no Passo 2 do algoritmo. O epíteto “revisada” deve-se a esse detalhamento do algoritmo, inexistente na regressão de escores bloco a bloco difundida previamente ao estudo mencionado. Cabe mencionar que também provaram a consistência do estimador da matriz de covariâncias  $\mathbf{\Psi}$  e do coeficiente de determinação das variáveis-resposta.

Além dos resultados assintóticos, Skrondal & Laake (2001) fizeram estudos de simulação para amostras finitas, a respeito dos estimadores de  $\mathbf{\Gamma}$  e dos coeficientes de determinação obtidos segundo o algoritmo revisado. Foi evidenciada a ausência de viés quando comparados com os estimadores de Máxima Verossimilhança, além de se mostrarem quase tão eficientes quanto eles, que são conhecidos por serem assitoticamente eficientes para modelos corretamente especificados, como é o caso.

Exemplos de modelos particulares nos quais a regressão por Mínimos Quadrados pode ser aplicada por meio da abordagem bloco a bloco são o MIMIC, em que as variáveis explicativas não possuem erro de mensuração, e o *Modelo de Extensão do Fator*<sup>4</sup> (Tucker, 1971), em que se considera que as variáveis-resposta não possuem erro de mensuração.

### Regressão de escores fator a fator

Skrondal & Laake (2001) expandiram a regressão de escores para a modalidade fator a fator, que requer que cada variável observada esteja associada exclusivamente a um fator, formando grupos independentes. Tal abordagem dá-se também conforme um algoritmo de 3 passos:

---

<sup>4</sup>*Factor Extension Model*

**Passo 1: Análise Fatorial fator a fator** Para cada fator da regressão, um modelo de AFC, que neste caso coincide com um de AFE, é ajustado às variáveis observadas  $\mathbf{x}$  ou  $\mathbf{y}$  associadas a ele. É importante verificar a identificabilidade de cada Análise Fatorial conduzida, certificando-se de que cada uma seja composta por pelo menos 3 variáveis observadas. Seja  $\Sigma_x^\nu$  a matriz de covariâncias implicada pelo modelo de Análise Fatorial ajustado a um dos  $n$  fatores explicativos  $\nu$  e seu conjunto de variáveis observadas.

**Passo 2: Estimação dos escores dos fatores** Por meio do método de Bartlett, obtêm-se os escores empíricos  $\tilde{\mathbf{F}}_{\eta,i}$  para os fatores  $\boldsymbol{\eta}$  e, por meio do método da Regressão, os escores empíricos  $\tilde{\mathbf{F}}_{\xi,i}$  para os fatores  $\boldsymbol{\xi}$ . A independência dos fatores faz com que  $\boldsymbol{\Psi}$  seja diagonal e a matriz  $\mathbf{A}_\xi^R$  passe a ser definida como

$$\mathbf{A}_\xi^R = \text{diag}(\boldsymbol{\Phi})\boldsymbol{\Lambda}'_x(\Sigma_x^\bullet)^{-1},$$

onde

$$\Sigma_x^\bullet = \begin{pmatrix} \Sigma_x^1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma_x^2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma_x^n \end{pmatrix}.$$

A matriz  $\mathbf{A}_\eta^B$ , por outro lado, permanece idêntica à da regressão de escores bloco a bloco.

**Passo 3: Regressões com MMQ** Os parâmetros estruturais  $\boldsymbol{\Gamma}$  são estimados para cada equação do sistema (4.25), por meio do método dos Mínimos Quadrados (MMQ). O estimador pode ser expresso por

$$\hat{\boldsymbol{\Gamma}} = \mathbf{S}_{\tilde{\mathbf{F}}_\eta^B \tilde{\mathbf{F}}_\xi^R} \left( \mathbf{S}_{\tilde{\mathbf{F}}_\xi^R} \right)^{-1},$$

onde  $\mathbf{S}_{\tilde{\mathbf{F}}_\eta^B \tilde{\mathbf{F}}_\xi^R}$  e  $\mathbf{S}_{\tilde{\mathbf{F}}_\xi^R}$  representam as matrizes de covariâncias empíricas para os escores dos fatores subscritos.

A performance assintótica é semelhante à da regressão de escores bloco a bloco, mas exige-se que os fatores sejam independentes. Portanto, se cada variável observada estiver associadas exclusivamente a um único fator, temos que as regressões com um fator explicativo, bem como as de modelos MIMIC produzem estimadores consistentes por Mínimos Quadrados.

## 4.2.2 Limitações

Embora de fácil aplicação, o modelo de Regressão Latente apresenta as seguintes limitações:

1. As variáveis observadas não podem ser correlacionadas.
2. Os parâmetros dos modelos de mensuração utilizados na aferição dos escores que participam da regressão podem não ser identificados, ainda que os parâmetros estruturais o sejam.
3. A regressão de escores não produz as medidas convencionais de qualidade do ajuste, em que o modelo ajustado é comparado com a alternativa irrestrita, representada pela matriz de covariâncias amostral.
4. Versões corrigidas dos coeficientes de determinação requerem álgebra matricial elementar, enquanto que as dos erros-padrão e intervalos de confiança requerem métodos de reamostragem.
5. Transformações não-lineares dos escores prejudicam a estimação dos parâmetros da regressão e não devem ser realizadas.

De qualquer forma, nos casos em que pode ser aplicado, constitui uma útil simplificação para o modelo LISREL.

## 4.3 Limitações das abordagens clássicas para a regressão de variáveis latentes contínuas

Os modelos de Equações Estruturais constituem uma importante ferramenta na mensuração e regressão de variáveis latentes contínuas, uma vez que possibilitam a estimação simultânea, tanto dos parâmetros dos modelos de mensuração, quanto dos parâmetros da regressão.

Vale notar que a complexidade de tais modelos e a eventual falta de disponibilidade dos *softwares* necessários para o seu ajuste, pode fazer com que a aplicação de uma regressão clássica, ajustada pelo método dos Mínimos Quadrados seja ocasionalmente

mais plausível. De fato, em algumas situações essa abordagem produz estimadores consistentes, conforme demonstrado por Skrondal & Laake.

As limitações de ambas abordagens, no entanto, costuma residir no fato de o modelo clássico não comportar não-linearidades, tanto no modelo de mensuração quanto no estrutural. Abordagens mais modernas como o Modelo Linear Generalizado Misto com Variáveis Latentes, conhecido pelo acrônimo em inglês GLLAMM, possibilitam a resolução de problemas mais complexos, como um modelo de Equações Estruturais linear generalizado, para variáveis latentes mensuradas segundo um modelo da Teoria de Resposta ao Item.

Na Seção seguinte, apresentar-se-á o GLLAMM e suas potencialidades no contexto da mensuração e regressão de variáveis latentes contínuas.

## 4.4 Modelos Lineares Generalizados Mistos com Variáveis Latentes (1999)

Os Modelos Lineares Generalizados Mistos com Variáveis Latentes (GLLAMM)<sup>5</sup> (Rabe-Hesketh & Pickles, 1999) constituem uma importante expansão dos Modelos de Equações Estruturais convencionais, que congrega características marcantes dos Modelos Lineares Generalizados:

- Uma das funções de ligação dos Modelos Lineares Generalizados pode ser utilizada na definição das equações de regressão do modelo de mensuração; e
- A densidade das variáveis-resposta pode não ser normal, desde que pertença à família exponencial.

Exxtensões dos Modelos Lineares Generalizados também são contempladas, como superdispersão ou subdispersão, heteroscedasticidade, respostas politômicas ou ordinais e funções de ligação compostas (Skrondal & Rabe-Hesketh, 2004, Seção 2.3).

Além disso, o GLLAMM permite a definição de *clusters*, isto é, de *aglomerados* de variáveis, o que possibilita a modelagem de efeitos aleatórios não somente no nível do item ou do indivíduo amostrado, mas também de grupos de indivíduos, ou grupo de

---

<sup>5</sup>GLLAMM: *Generalized Linear Latent and Mixed Modeling*

grupos de indivíduos e assim sucessivamente. No contexto de avaliações educacionais em múltiplas escolas, por exemplo, é possível modelar não somente o efeito que a variável latente atribuída a cada aluno tem em sua resposta a um determinado item da avaliação, mas também o efeito que a turma em que ele está alocado exerce sobre as respostas a tal item, bem como o efeito que sua escola exerce.

De fato, mesmo o delineamento mais difundido, de itens e de indivíduos que respondem aos itens, já corresponde a um delineamento multinível, no qual as respostas a um conjunto de itens sempre estão associadas a um indivíduo. Ele, por sua vez, representa o aglomerado que agrupa tais respostas, de modo que cada resposta obtida é modelada em termos tanto das características do item, quanto das características do indivíduo respondente. Na realidade, portanto, o GLLAMM simplesmente amplia o número de níveis passíveis de ser modelados para além dos usuais dois níveis.

A seguir, será apresentada a definição do GLLAMM para modelos de mensuração de dois níveis, a fim de que se possa compará-la com a dos modelos apresentadas anteriormente.

#### 4.4.1 Definição dos componentes do modelo

##### Modelo de mensuração

O escopo do GLLAMM permite a modelagem de diversos tipos de variável-resposta, de forma análoga à dos Modelos Lineares Generalizados, bastando, para isso, que se especifique uma *função de ligação*  $g(\cdot)$  que torne verdadeira a equação para a resposta  $y_{ij}$ , do indivíduo  $j$  ao item  $i$ :

$$g \left[ \mathbb{E} (Y_{ij} | \mathbf{x}_j, \boldsymbol{\eta}_j) \right] = \nu_{ij} , \quad (4.34)$$

onde  $\boldsymbol{\eta}_j$  é o vetor composto pelas realizações de todas as variáveis latentes para o indivíduo  $j$ ,  $\mathbf{x}_j$  é um vetor de covariáveis observadas para o indivíduo  $j$  e a densidade de probabilidade da variável aleatória  $Y_{ij}$  pretence à família exponencial. Vale notar que a equação (4.34) representa a especificação para um GLLAMM com apenas dois níveis ( $i$  e  $j$ ), podendo esse número de níveis ser ampliado conforme necessário.

Os *componentes lineares*  $\nu_{ij}$ , para um modelo com  $M$  variáveis latentes, são defi-

nidos na equação

$$\boldsymbol{\nu}_j = \mathbf{X}_j \boldsymbol{\beta} + \sum_{m=1}^M \eta_{mj} \mathbf{Z}_{mj} \boldsymbol{\lambda}_m . \quad (4.35)$$

Diferentemente do LISREL, no GLLAMM, todas as variáveis latentes são chamadas de  $\eta$ , independentemente de sua exogeneidade ou endogeneidade. Além disso,  $\boldsymbol{\beta}$  é um vetor de *coeficientes atribuídos às covariáveis* representadas na matriz  $\mathbf{X}_j$ , não tendo relação alguma, com os elementos da matriz  $\mathbf{B}$  da equação estrutural, salvo pelo nome. O vetor  $\boldsymbol{\lambda}_m$  compõe-se de parâmetros da estrutura do modelo de mensuração, que relacionam cada item com a variável latente  $\eta_m$ . A matriz  $\mathbf{Z}_{mj}$  é uma *matriz de delineamento*, que associa as variáveis latentes apropriadas a cada componente linear  $\nu_{ij}$ , no nível da unidade  $j$

Como exemplo, a equação (4.36) exhibe a definição dos componentes lineares para o modelo representado no diagrama de caminho da Figura 2.2, com seis variáveis observadas e  $M = 2$  variáveis latentes:

$$\begin{pmatrix} \nu_{1j} \\ \nu_{2j} \\ \nu_{3j} \\ \nu_{4j} \\ \nu_{5j} \\ \nu_{6j} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} + \eta_{1j} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \end{pmatrix} + \eta_{2j} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_{42} \\ \lambda_{52} \\ \lambda_{62} \end{pmatrix} \quad (4.36)$$

onde as variáveis  $\xi_1$  e  $\eta_1$  passaram a ser denominadas, respectivamente, de  $\eta_1$  e  $\eta_2$ , tendo as cargas também mudado de nome, de  $(\lambda_{11}^x, \lambda_{21}^x, \lambda_{31}^x, \lambda_{12}^x, \lambda_{22}^x, \lambda_{32}^x)$  para  $(\lambda_{11}, \lambda_{21}, \lambda_{31}, \lambda_{42}, \lambda_{52}, \lambda_{62})$ .

No GLLAMM, os erros de mensuração ( $\Theta_\delta$  e  $\Theta_\epsilon$  do modelo LISREL) não são caracterizados diretamente e representam apenas a *variabilidade residual* das variáveis observadas, às quais é atribuída uma densidade de probabilidade da família exponencial.

A ampla gama de modelos que podem ser definidos a partir da especificação das matrizes de delineamento, das variáveis latentes e dos parâmetros do GLLAMM constitui sua principal característica. Nele, as variáveis latentes podem representar tanto efeitos aleatórios de um modelo Multinível quanto traços latentes de um modelo de Equações Estruturais. Além disso, as diversas densidades que podem ser atribuídas às

variáveis-resposta, bem como a possibilidade do uso de funções de ligação na definição dos componentes lineares permite a inclusão de modelos da TRI, modelos de Análise de Sobrevivência e diversos outros no escopo do GLLAMM. Até mesmo o modelo logístico de 3 parâmetros, da TRI, pode ser definido como um GLLAMM, utilizando *funções de ligação compostas* (Skrondal & Rabe-Hesketh, 2004, p.32).

### Modelo estrutural

O sistema de equações estruturais do GLLAMM é dado por

$$\boldsymbol{\eta}_j = \mathbf{B}\boldsymbol{\eta}_j + \boldsymbol{\Gamma}\mathbf{w}_j + \boldsymbol{\zeta}_j, \quad (4.37)$$

onde  $\boldsymbol{\eta}_j$  é o vetor das realizações associadas ao indivíduo  $j$ , de todas as  $M$  variáveis latentes do modelo, e  $\mathbf{B}$  é a matriz que contém os *coeficientes de regressão* das relações lineares entre tais variáveis. As *covariáveis* observadas para o indivíduo  $j$  são representadas por  $\mathbf{w}_j$ , sendo  $\boldsymbol{\Gamma}$  o vetor dos respectivos *coeficientes de regressão* associados às covariáveis. Finalmente,  $\boldsymbol{\zeta}_j$  é o vetor composto pelos  $M$  erros aleatórios observados em cada equação estrutural.

Para o exemplo descrito na Equação (4.36), o modelo estrutural define-se por

$$\begin{pmatrix} \eta_{1j} \\ \eta_{2j} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \beta_{21} & 0 \end{pmatrix} \begin{pmatrix} \eta_{1j} \\ \eta_{2j} \end{pmatrix} + \begin{pmatrix} \zeta_{1j} \\ \zeta_{2j} \end{pmatrix}$$

Como se nota, as diferenças mais marcantes do GLLAMM em relação ao LISREL encontram-se na sua definição do modelo de mensuração e não no modelo estrutural, que é praticamente idêntico.

## 4.4.2 Considerações sobre o modelo

### Identificabilidade

A identificabilidade dos GLLAMM pode ser verificada analítica ou empiricamente, de forma análoga à utilizada no contexto do LISREL. Skrondal & Rabe-Hesketh (2004, cap.5) apresentam métodos analíticos para a avaliação da identificabilidade, baseados no posto do jacobiano calculado para formas reduzidas do vetor de parâmetros do modelo. Em particular, modelos LISREL que são identificados também o são quando vertidos para a notação GLLAMM.

## Estimação

Quanto à estimação, várias abordagens são possíveis, inclusive bayesianas, sendo a mais utilizada aquela implementada no *software* que se tornou referência na estimação dos GLLAMM: o homônimo programa `gllamm` do pacote STATA. Nele, foi implementado o método da Máxima Verossimilhança com Informação Completa em combinação com o método da *Quadratura Adaptativa*, também implementado no procedimento NLMIXED do SAS, para realizar as integrações necessárias. Tal escolha deve-se ao método poder ser aplicado a qualquer modelo particular da enorme classe dos GLLAMM, além de possibilitar inferências baseadas na verossimilhança.

## Potencialidades e limitações

A opção por este método de estimação, no entanto, traz a inconveniência de sua eficiência computacional ser inferior à de diversos outros métodos menos gerais. O ajuste de um modelo logístico de 2 parâmetros da TRI, por exemplo, que leva não mais que alguns minutos ou segundos em *softwares* como o BILOG-MG ou utilizando-se o pacote `ltm` do R, pode levar horas no `gllamm`, especialmente se o número de variáveis latentes for elevado ou se o conjunto de dados compor-se de muitos itens ou muitos respondentes. Skrondal & Rabe-Hesketh (2004, cap.6) apresentam diversos métodos alternativos que podem ser computacionalmente mais viáveis em alguns casos.

De qualquer forma, o GLLAMM representa um importante avanço no âmbito dos modelos de regressão para variáveis latentes, na medida em que diversos modelos aparentemente desconexos constituem casos particulares dessa grande classe, como o Modelo de Regressão Linear Misto, os modelos de Equações Estruturais e diversos modelos da Teoria de Resposta ao Item. A possibilidade de modelar estruturas latentes complexas, a partir de observações não-contínuas pode compensar largamente o custo computacional comparativamente alto.

## 4.5 Apêndice

### 4.5.1 Linhas de código e *output* do R: ajuste do modelo de Análise Fatorial Confirmatória aos dados da Seção 4.1.2

```
> # install.packages("lavaan")
> library(lavaan)
> AFC.provas.modelo <- 'espacial =~ general + picture + blocks + maze
+                          linguagem =~ general + reading + vocab'
> AFC.provas.ajuste <- cfa(AFC.provas.modelo ,
+ sample.cov = ability.cov$cov , sample.nobs = 112 , std.lv = TRUE)
> summary(AFC.provas.ajuste)
```

Lavaan (0.4-7) converged normally after 110 iterations

Number of observations	112
Estimator	ML
Minimum Function Chi-square	7.125
Degrees of freedom	7
P-value	0.416

Parameter estimates:

Information	Expected
Standard Errors	Standard

	Estimate	Std.err	Z-value	P(> z )
Latent variables:				
espacial =~				
general	2.309	0.460	5.021	0.000
picture	1.659	0.243	6.814	0.000
blocks	10.748	1.122	9.584	0.000
maze	1.697	0.347	4.887	0.000
linguagem =~				
general	2.014	0.438	4.599	0.000
reading	6.774	0.598	11.332	0.000

```

      vocab          9.768    0.986    9.905    0.000

Covariances:
  espacial ~~
    linguagem      0.440    0.094    4.698    0.000

Variances:
  general         10.938    1.735    6.305    0.000
  picture          3.889    0.636    6.112    0.000
  blocks          32.964   15.043    2.191    0.028
  maze             9.718    1.388    7.000    0.000
  reading          6.247    4.295    1.454    0.146
  vocab           38.678   10.172    3.802    0.000
  espacial         1.000
  linguagem        1.000

```

```
> fitMeasures(AFC.provas.ajuste)
```

	chisq	df	pvalue	baseline.chisq	baseline.df
	7.125	7.000	0.416	277.864	15.000
	baseline.pvalue	cfi	tli	logl	unrestricted.logl
	0.000	1.000	0.999	-2020.752	-2017.190
	npar	aic	bic	ntotal	bic2
	14.000	4069.504	4107.563	112.000	4063.318
	rmsea	rmsea.ci.lower	rmsea.ci.upper	rmsea.pvalue	srmr
	0.013	0.000	0.117	0.598	0.031

## 4.5.2 Linhas de código e *output* do R: ajuste do modelo de Equações Estruturais aos dados da Seção 4.1.3

.

```

> library(lavaan)
> LISREL.modelo <- '
+ # Modelos de mensuração
+   ind60 =~ x1 + x2 + x3
+   dem60 =~ y1 + y2 + y3 + y4
+   dem65 =~ y5 + equal("dem60=~y2")*y6
+           + equal("dem60=~y3")*y7

```

```

+           + equal("dem60=~y4")*y8
+ # Modelo estrutural
+   dem60 ~ ind60
+   dem65 ~ ind60 + dem60
+ # Erros de mensuração
+   y1 ~~ y5
+   y2 ~~ y4 + y6
+   y3 ~~ y7
+   y4 ~~ y8
+   y6 ~~ y8'
+
> LISREL.ajuste <- sem(model, data=PoliticalDemocracy)
> summary(LISREL.ajuste, fit.measures=TRUE)
Lavaan (0.4-7) converged normally after 97 iterations

```

Number of observations	75
Estimator	ML
Minimum Function Chi-square	40.179
Degrees of freedom	38
P-value	0.374

Chi-square test baseline model:

Minimum Function Chi-square	730.654
Degrees of freedom	55
P-value	0.000

Full model versus baseline model:

Comparative Fit Index (CFI)	0.997
Tucker-Lewis Index (TLI)	0.995

Loglikelihood and Information Criteria:

Loglikelihood user model (H0)	-1548.818
Loglikelihood unrestricted model (H1)	-1528.728

Number of free parameters	28
Akaike (AIC)	3153.636
Bayesian (BIC)	3218.526
Sample-size adjusted Bayesian (BIC)	3130.277

Root Mean Square Error of Approximation:

RMSEA	0.028
90 Percent Confidence Interval	0.000 0.087
P-value RMSEA <= 0.05	0.665

Standardized Root Mean Square Residual:

SRMR	0.056
------	-------

Parameter estimates:

Information	Expected
Standard Errors	Standard

	Estimate	Std.err	Z-value	P(> z )
Latent variables:				
ind60 =~				
x1	1.000			
x2	2.180	0.138	15.751	0.000
x3	1.818	0.152	11.971	0.000
dem60 =~				
y1	1.000			
y2	1.191	0.139	8.551	0.000
y3	1.175	0.120	9.755	0.000
y4	1.251	0.117	10.712	0.000
dem65 =~				
y5	1.000			
y6	1.191	0.139	8.551	0.000
y7	1.175	0.120	9.755	0.000
y8	1.251	0.117	10.712	0.000

Regressions:

dem60 ~				
ind60	1.471	0.392	3.750	0.000
dem65 ~				
ind60	0.600	0.226	2.661	0.008
dem60	0.865	0.075	11.554	0.000

Covariances:

y1 ~~				
y5	0.583	0.356	1.637	0.102
y2 ~~				
y4	1.440	0.689	2.092	0.036
y6	2.183	0.737	2.960	0.003
y3 ~~				
y7	0.712	0.611	1.165	0.244
y4 ~~				
y8	0.363	0.444	0.817	0.414
y6 ~~				
y8	1.372	0.577	2.378	0.017

Variances:

x1	0.081	0.019	4.182	0.000
x2	0.120	0.070	1.729	0.084
x3	0.467	0.090	5.177	0.000
y1	1.855	0.433	4.279	0.000
y2	7.581	1.366	5.549	0.000
y3	4.956	0.956	5.182	0.000
y4	3.224	0.723	4.458	0.000
y5	2.313	0.479	4.831	0.000
y6	4.968	0.921	5.393	0.000
y7	3.560	0.710	5.018	0.000
y8	3.308	0.704	4.701	0.000
ind60	0.449	0.087	5.175	0.000
dem60	3.875	0.866	4.477	0.000
dem65	0.164	0.227	0.725	0.469

# Capítulo 5

## Visão geral dos modelos, recomendações práticas e trabalhos futuros

### 5.1 Visão geral dos modelos apresentados

#### 5.1.1 Modelos de mensuração

Tendo em perspectiva o conjunto de modelos apresentados nos Capítulos 3 e 4, para mensuração e regressão de variáveis latentes contínuas, observamos que as primeiras investidas no assunto diziam respeito exclusivamente à questão de sua *mensuração*. Nesse contexto, a Análise Fatorial Exploratória (AFE) e a Teoria de Resposta ao Item (TRI) possuem destaque, como principais representantes dos modelos de mensuração para variáveis observadas contínuas e discretas, respectivamente.

Cabe, ainda, uma outra diferenciação crucial entre ambas: ao passo que a abordagem exploratória da AFE proporciona uma compreensão maior das possíveis estruturas latentes para os dados observados, a TRI surge com um objetivo muito mais específico, que é o de mensurar uma variável latente pressuposta como principal geradora das observações efetuadas. Daí a proximidade maior da TRI com a modalidade Confirmatória da Análise Fatorial, e não com sua vertente Exploratória, dado que a Análise Fatorial Confirmatória (AFC) também pressupõe uma estrutura latente para os dados observados, que pode, inclusive, ser mais complexa que a estrutura unidimensional

normalmente adotada nas modelagens via TRI.

Essa observação é ainda mais necessária na medida em que a AFC é normalmente apresentada como um subproduto do Modelo de Equações Estruturais Geral, o LISREL, deixando a impressão de que se trata de um modelo para *regressão* de variáveis latentes, quando, na realidade, trata-se estritamente de um modelo de *mensuração*, mais próximo da TRI do que a própria AFE, precursora dos dois modelos.

Nesse sentido, é importante matizar a posição da AFE como modelo de mensuração, uma vez que, diante do quadro geral composto pelos modelos de mensuração e regressão de variáveis latentes hodierno, a AFE coloca-se mais como técnica exploratória do que como técnica de mensuração, propriamente dita.

A principal implicação desta constatação é, possivelmente, a de que os escores dos fatores, normalmente estimados no âmbito da AFE, não representam mensurações confiáveis, uma vez que são subprodutos de uma metodologia exploratória (Bollen, 1989; Skrondal & Rabe-Hesketh, 2004). O método mais adequado para a estimação dos escores dos fatores, portanto, é aquela que utiliza os parâmetros estruturais do modelo de Análise Fatorial *Confirmatória* para compor as transformações lineares (4.30), (4.31), (4.32) ou (4.33).

Ainda em relação aos modelos de mensuração, a relação, descrita por Takane & de Leeuw (1987) e ampliada por Kamata & Bauer (2008), entre a Análise Fatorial Binária e o Modelo Logístico de 2 Parâmetros, da Teoria de Resposta ao Item, pode ser muito útil, tanto no aproveitamento de conceitos e relações de uma teoria pela outra, quanto pela implementação computacional da TRI. Dado que poucos pacotes computacionais de estatística trazem implementados algoritmos de estimação para a TRI, é possível utilizar os métodos de estimação para Análise Fatorial, traduzindo os resultados em termos dos parâmetros da TRI. O ganho é mútuo, na medida em que o ajuste do modelo da TRI torna-se possível e os parâmetros do modelo de Análise Fatorial tornam-se mais interpretáveis.

### 5.1.2 Modelos de regressão para variáveis latentes

Um segundo momento que marca a evolução do estudo das variáveis latentes é a discussão sobre os problemas envolvidos na regressão dessas variáveis. Uma vez que as

aferições realizadas a respeito de uma variável latente representam *estimativas* do escore verdadeiro, sua utilização como variável de um modelo de regressão convencional não atende ao pressuposto de que as covariáveis representam efeitos fixos. A consequência dessa negligência é o fenômeno conhecido como *atenuação*, caracterizado pela subestimação dos coeficientes da regressão. Sendo assim, dois tipos de solução foram apresentados, tendo um deles sido focado.

### Mensuração e regressão simultâneas

A solução focada foi a da utilização dos Modelos de Equações Estruturais Gerais (LISREL), que possibilitam a estimação conjunta dos parâmetros do *modelo de mensuração* e do *modelo de múltiplas regressões*. A vantagem desta abordagem é que, ao conjugar a *Análise Fatorial Confirmatória* (AFC), como modelo de mensuração, e do modelo de *Equações Simultâneas*, como modelo de múltiplas regressões, o erro de mensuração das covariáveis é modelado, eliminando a atenuação dos coeficientes das regressões.

Além de tratar da atenuação, o LISREL possui a vantagem de estar implementado em diversos pacotes estatísticos e em *software* voltados especificamente à sua aplicação, para os mais diversos tipos de usuário.

Como o Capítulo 3 apresenta, a Teoria de Resposta ao Item (TRI) ganhou um espaço muito importante nas pesquisas que envolvem a mensuração de variáveis latentes contínuas, por possibilitarem sua aferição a partir de *variáveis observadas categorizadas* em vez de contínuas. Os modelos da TRI, no entanto, não podem compor um modelo LISREL clássico, que somente admite *variáveis observadas contínuas*, compondo um modelo de mensuração no molde de uma AFC, conforme mencionado anteriormente.

Ademais, os modelos de mensuração para observações categorizadas podem ser de dois tipos: *função de resposta* ou *resposta latente contínua*, também chamado de *modelo de limiar* (Seção 3.4). Diversos autores (Muthen, 1978; Bollen, 1989; Kamata & Bauer, 2008; Skrondal & Rabe-Hesketh, 2004) mostram como observações categorizadas podem ser modeladas a partir da hipótese de que uma *resposta latente contínua* é responsável por gerá-las. Essa formulação permite que a estrutura convencional dos

modelos de Análise Fatorial tradicionais seja ajustada a dados categorizados.

Os modelos de TRI, por sua vez, seguem a formulação da *função de resposta*, de modo que seus modelos não são diretamente adaptáveis à estrutura da Análise Fatorial. Contudo, um importante resultado apresentado na Seção 3.4 possibilita que se converta os parâmetros estimados para um modelo de Análise Fatorial naqueles de um modelo logístico de 1 ou 2 parâmetros da TRI (Takane & de Leeuw, 1987).

Ainda assim, diversos outros modelos da TRI não parecem ter conexão com a AFC ou o LISREL. Entretanto, a classe de modelos chamada GLLAMM possibilita a regressão de variáveis latentes estimadas por um modelo logístico de 3 parâmetros, da TRI, por exemplo, e de muitos outros modelos de mensuração não-lineares.

Isso é possível devido à incorporação de funções de ligação às equações que definem os modelos de mensuração do GLLAMM, bem como a possibilidade de às variáveis-resposta ser associada qualquer densidade da família exponencial. Essas características, portanto, conferem aos modelos de mensuração do GLLAMM propriedade típicas dos Modelos Lineares Generalizados.

A extensão que o GLLAMM propicia ao LISREL, no entanto, é ainda maior, na medida em que sua definição de variável latente é muito mais flexível, possibilitando que ela represente o coeficiente aleatório de um modelo Multinível. Exemplo disso, é que no GLLAMM, a denominação usual da TRI de *itens* e *respondentes*, é considerada um delineamento multinível, no qual o respondente é tido como um *cluster* de itens. Evidentemente, conjuntos de respondentes podem compor grupos ainda mais gerais abrangentes, sem sair do escopo do GLLAMM.

Embora extremamente geral, uma forte limitação do GLLAMM é o seu custo computacional, já que um modelo simples da TRI, normalmente ajustado em alguns minutos, pode levar horas para ter seus parâmetros estimados com a implementação computacional do GLLAMM proposta por seus autores em STATA.

### **Abordagens passo-a-passo**

Além dos modelos de ajuste simultâneo do modelo de mensuração e do de regressão, algumas abordagens passo-a-passo são possíveis em determinados casos.

Uma primeira abordagem envolve variáveis mensuradas por observações contínuas.

Para efetuar a regressão, é necessário ajustar o modelo de Análise Fatorial Confirmatória relativo às variáveis-resposta e, da mesma forma, o modelo das covariáveis, também em bloco. Em seguida, são aferidos os escores para as variáveis latentes, utilizando o método de Bartlett, e para as covariáveis, utilizando o método da Regressão. Os escores obtidos dessa forma constituem dados para o ajuste da regressão, isento de atenuação.

Além desse resultado, Skrondal & Laake (2001) obtiveram um outro, menos geral, que pode ser aplicado fator-a-fator, desde que eles sejam independentes e que as variáveis associadas a um fator não sejam associadas a nenhum outro. Nessa abordagem, um modelo de AFC é ajustado para cada fator, sendo seus escores estimados de forma análoga à abordagem bloco-a-bloco, isto é, usando o método de Bartlett para os escores que fazem papel de variável-resposta e o método da Regressão para os escores que servem como covariáveis.

Esse importante resultado, que permite uma abordagem passo-a-passo da regressão de escores de fatores, utilizando um modelo de mensuração consagrado, como a AFC, é extremamente recomendável nos casos em que o ajuste de um LISREL seja proibitivo ou simplesmente desnecessário, uma vez que a abordagem passo-a-passo pode ser muito mais acessível.

Uma última abordagem passo-a-passo que pode ser mencionada é a da Regressão com Erros nas Variáveis. Na abordagem passo-a-passo, os escores dos fatores são aferidos antes de se efetuar a regressão e podem ser consideradas variáveis com erro de mensuração. A variância desses erros pode ser diferente para cada observação e é dada pelo erro-padrão da estimativa do escore. Com essa informação sobre o erro de mensuração, é possível ajustar uma regressão que a leve em conta, de forma a eliminar o efeito da atenuação.

Battaui *et al.* (2008) apresenta um interessante problema resolvido por essa abordagem, no qual as variáveis-resposta são observadas e a covariável latente é estimada por um modelo da TRI. Nessa resolução, utiliza-se o método SIMEX (Carroll *et al.*, 2006) para realizar a estimativa do coeficiente da regressão. Em seguida, essa estimativa é agregada aos dados anteriores, formando um novo modelo que se propõe melhorar a confiabilidade dos valores aferidos por TRI, utilizados inicialmente.

## 5.2 Recomendações para a mensuração de variáveis latentes

Paralelamente à apresentação dos diversos métodos de mensuração e regressão das variáveis latentes, foram apresentadas limitações desses modelos e erros comuns de pesquisadores que utilizam tais métodos. A seguir, as principais dentre essas advertências estão compiladas em forma de recomendações:

- Ao se utilizar um modelo de Análise Fatorial para aferir os escores de um fator, é preciso checar se os escores não são indeterminados.
- A modalidade Confirmatória da Análise Fatorial é um modelo de mensuração de fato, ao passo que a modalidade Exploratória é uma ferramenta para investigação da estrutura latentes dos dados observados. Portanto, modelos de mensuração devem ser ajustados preferencialmente pela AFC e não pela AFE.
- Se as variáveis observadas  $\mathbf{x}$  representarem *causas* em vez de *consequências* da variável latente por elas mensurada, a utilização do modelo de Análise Fatorial não é adequada. Nesses casos, o modelo MIMIC representa o modelo correto a ser utilizado, pois considera  $\mathbf{x}$  um vetor de causas da variável latente. Vem a somar aos benefícios do MIMIC o fato de ele ser um modelo de mensuração identificado, desde que os erros de mensuração das variáveis que compõem  $\mathbf{x}$  possam ser considerados nulos.
- O ajuste de modelos de Análise Fatorial a dados categorizados é inadequado, sendo os modelos da TRI os mais apropriados para esse tipo de situação. O ajuste de modelos de Análise Fatorial a dados na escala Likert, composta por números inteiros, de 1 a 7, pode resultar em modelos de mensuração com graves problemas de especificação (Muthén & Kaplan, 1985).
- Independentemente dos modelos utilizados para aferir os escores de variáveis latentes contínuas, é importante ter em mente que tais escores possuem erro de mensuração e não devem ser modelados por modelos clássicos de regressão, que consideram as covariáveis como fixas, em vez de aleatórias. Alternativas corretas podem ser representadas pelos modelos de Equações Estruturais e

- É recomendável optar sempre pelo modelo mais simples que pode ser aplicado a um problema, devido à facilidade de especificação, à eficiência computacional e à maior disponibilidade de *softwares* adequados a diversos tipos de usuário. A obtenção de estimativas de modelos mais difundidos também é mais facilmente comunicável.

### 5.3 Trabalhos futuros

Ainda que os métodos de mensuração e regressão de variáveis latentes contínuas estejam bastante avançados, muitas questões importantes ainda estão por resolver. O resumo de alguns temas que podem ser abordados em trabalhos futuros é dado a seguir:

- [Skrondal & Rabe-Hesketh \(2004\)](#) mencionam a ampliação do estudo da identificabilidade e equivalência como necessária. Novas regras de identificabilidade, ainda que restritas, podem ser desenvolvidas para classes específicas de modelos.
- O nível de desenvolvimento avançado do tratamento da multidimensionalidade nos modelos LISREL pode ser aproveitado para os modelos da TRI, nos quais a unidimensionalidade ainda predomina. Pode ser possível o desenvolvimento de testes de multidimensionalidade para TRI, a partir do escopo dos Modelos de Equações Estruturais.
- Testes para seleção de modelos e medidas de qualidade do ajuste podem ser adaptadas do contexto dos MEE ao da TRI, ainda que para alguns modelos específicos.
- Medidas de confiabilidade podem ser melhoradas com uma melhor modelagem do erro de mensuração. Exemplo desse fato são as recentes críticas ao  $\alpha$  de Crombach, superado pelos próprios MEE para aferição da confiabilidade de múltiplas observações para uma mesma variável latente.
- Implementações computacionalmente mais eficientes do modelo GLLAMM podem ser estudadas e implementadas em interfaces mais amigáveis, tendo em

vista o público multidisciplinar que costuma utilizar os métodos para análise de variáveis latentes.

- Abordagens integradas dos diversos modelos para variáveis latentes podem ser desenvolvidas, como tem sido recorrente recentemente, a exemplo das obras de [Loehlin \(2004\)](#), [Muthén \(2002\)](#) e [Skrondal & Rabe-Hesketh \(2004\)](#).

# Referências Bibliográficas

- ANDERSON, T. W. 1984. *An introduction do multivariate statistical analysis*. John Wiley & Sons.
- ARMSTRONG, J. S. 1967. Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. *The American Statistician*, **21**, 17–21.
- BAKER, F. B., & KIM, S-H. 2004. *Item Response Theory: parameter estimation techniques*. 2 edn. Nova Iorque: Marcel Dekker.
- BARTHOLOMEW, D. J. 1980. Factor analysis for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **42**(3), 293–321.
- BARTHOLOMEW, D. J., & KNOTT, M. 1990. *Latent Variable Analysis and Factor Analysis*. 2 edn. Arnold.
- BARTLETT, M. S. 1937. The statistical conception of mental factors. *British Journal of Psychology*, **28**, 97–104.
- BATTAUZ, M., BELLIO, R., & GORI, E. 2008. Reducing measurement error in student achievement estimation. *Psychometrika*, **73**, 289–302.
- BEAUDUCEL, ANDRÉ. 2007. In Spite of Indeterminacy Many Common Factor Score Estimates Yield an Identical Reproduced Covariance Matrix. *Psychometrika*, **72**, 437–441. 10.1007/s11336-005-1467-5.
- BERNAARDS, COEN A., & I.JENNRICH, ROBERT. 2005. Gradient Projection Algorithms and Software for Arbitrary Rotation Criteria in Factor Analysis. *Educational and Psychological Measurement*, **65**, 676–696.

- BIRNBAUM, A. 1968. *Statistical theories of mental test scores*. Reading: Addison-Wesley. Chap. Test scores, sufficient statistics, and the information structures of tests, pages 425–435.
- BOCK, R. D. 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29–51.
- BOLLEN, KENNETH A. 1989. *Structural Equations with latent variables*. New York: Wiley.
- BOLLEN, KENNETH A., & DAVIS, WALTER R. 2009a. Causal Indicator Models: Identification, Estimation, and Testing. *Structural Equation Modeling*, **16:3**, 498 – 522.
- BOLLEN, KENNETH A., & DAVIS, WALTER R. 2009b. Two Rules of Identification for Structural Equation Models. *Structural Equation Modeling*, **16:3**, 523–536.
- CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A., & CRAINICEANU, C. M. 2006. *Measurement error in nonlinear models: a modern perspective*. 2 edn. Boca Raton: Chapman & Hall/CRC.
- CHRISTOFFERSON, A. 1975. Factor analysis of dichotomized variables. *Psychometrika*, **40**, 5–32.
- CLOGG, C. C. 1992. The impact of sociological methodology on statistical methodology. *Statistical Science*, **7**, 183–196.
- COOPER, JOHN C. B. 1983. Factor Analysis: An Overview. *The American Statistician*, **37**(2), pp. 141–147.
- DEMPSTER, A. P., LAIRD, N. M., & RUBIN, D. B. 1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39** (1), 1–38.
- EMBRETSON, S. E., & REISE, S. P. 2000. *Item Response Theory for psychologists*. Lawrence Erlbaum Associates.

- FISCHER, G. 1995. *Rasch Models. Foundations, recent developments, and applications*. Nova Iorque: Springer. Chap. Derivations of the Rasch model, pages 15–38.
- FISCHER, G. H. 1973. Linear logistic test model as an instrument in educational research. *Acta Psychologica*, **37**, 359–374.
- FOX, J. 2006. Structural Equation Modeling with the sem Package in R. *Structural Equation Modeling*, **13**, 465–486.
- FOX, JOHN, KRAMER, ADAM, & FRIENDLY, MICHAEL. 2010. *sem: Structural Equation Models*. R package version 0.9-21.
- FULLER, W.A. 1987. *Measurement error models*. Nova Iorque: Wiley.
- GOLDBERGER, R. M. HAUSER & A. S. 1971. The treatment of unobservable variables in path analysis. *Sociological Methodology*, **3**, 81–117.
- GRICE, J.W. 2001. Computing and evaluating factor scores. *Psychological Methods*, **6**, 430–450.
- HAAVELMO, TRYGVE. 1943. The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, **11**, 1–12.
- HARMAN, H. H. 1976. *Modern factor analysis*. 3 edn. Chicago: University of Chicago Press.
- HEYWOOD. 1931. On finite sequences of real numbers. *Proceedings of the Royal Society, Series A*, **134**, 486–510.
- HOLLAND, P.W., & WAINER, H. (eds). 1993. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum associates.
- JÖRESKOG, K. G. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, **34**, 183–202.
- KAISER, H. F. 1958. The varimax criterion for analytical rotation in factor analysis. *Psychometrika*, **23**, 187–200.

- KAMATA, A., & BAUER, D. J. 2008. A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, **15**, 136–153.
- KING, GARY. 1986. How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science. *American Journal of Political Science*, **30**(3), pp. 666–687.
- KLEIN, L. 1950. *Economic fluctuations in the United States 1921-1941*. Wiley.
- KRZANOWSKI, W. J., & MARRIOTT, F. H. C. 1995. *Multivariate Analysis, Part 2: Classification, covariance structures and repeated measurements*. Kendall's Library of Statistics, vol. 2. Londres: Arnold.
- LAWLEY, D. N. 1940. The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh, Sec. A*, **60**, 64–82.
- LAZARFELD, P. F., & HENRY, N. W. 1968. *Latent Structure Analysis*. Boston: : Houghton Mifflin.
- LITTLE, R. J. A., & RUBIN, D. B. 2002. *Statistical analysis with missing data*. 2 edn. Nova Iorque: Wiley.
- LOEHLIN, J. C. 2004. *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*. 3 edn. Mahwah, NJ: Erlbaum.
- LORD, F. M., & NOVICK, M. R. 1968. *Statistical theories of mental test scores*. EUA: Addison-Wesley.
- MAIR, PATRICK, HATZINGER, REINHOLD, & MAIER, MARCO. 2010. *eRm: Extended Rasch Modeling*. R package version 0.13-0.
- MARCOULIDES, G. A., & HERSHBERGER, S. L. 1997. *Multivariate statistical methods: a first course*. Mahwah: Lawrence Erlbaum Associates.
- MCDONALD, R. P. 1967. Nonlinear factor analysis. *Psychometrika Monograph*, **15**.
- MCGRATH, K., & WATERTON, J. 1986. *British Social Attitudes, 1983-1986 Panel Survey*. Tech. rept. Social and Community Planning Research, Londres.

- MEREDITH, W. 1977. On weighted Procrustes and hyperplane fitting in factor analysis rotation. *Psychometrika*, **42**, 491–522.
- MUTHEN, B. 1978. Contributions to factor analysis of dichotomous variables. *Psychometrika*, **43**, 551–560.
- MUTHÉN, B., & KAPLAN, D. 1985. A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, **38**, 171–189.
- MUTHÉN, B. O. 2002. Beyond SEM: general latent variable modeling. *Behaviormetrika*, **29**, 81–117.
- MUTHEN, BENGT O., & SATORRA, ALBERT. 1995. Complex Sample Data in Structural Equation Modeling. *Sociological Methodology*, **25**, 267–316.
- NEUHAUS, J., & WRIGLEY, C. 1954. The quartimax method: an analytical approach to orthogonal simple structure. *British Journal of Statistical Psychology*, **7**, 81–91.
- PEARSON, K. 1901. Mathematical contributions to the theory of evolution VII. On the inheritance of characters not capable of exact quantitative measurement. *Philosophical Transactions of the Royal Society A*, **195**, 79–150.
- R DEVELOPMENT CORE TEAM. 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- RABE-HESKETH, S., & PICKLES, A. 1999. Generalised linear latent and mixed models. *Pages 332–339 of: Proceedings of the 14th International Workshop on Statistical Modeling*.
- RASCH, G. 1960. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Dinamarca: Paedagogiske Institut.
- RIZOPOULOS, DIMITRIS. 2006. ltm: An R Package for Latent Variable Modeling and Item Response Analysis. *Journal of Statistical Software*, **17**(5), 1–25.

- ROBINSON, P. M. 1974. Identification, estimation, and large sample theory for regressions containing unobservable variables. *International Economic Review*, **15**, 680–692.
- ROSSEEL, YVES. 2011. *lavaan: Latent Variable Analysis*. R package version 0.4-7.
- SCHEIBLECKNER, H. 1977. *Structural models of thinking and learning*. Bern: Huber. Chap. Psychological models based upon conditional inference, pages 185–202.
- SKRONDAL, A., & LAAKE, P. 2001. Regression among factor scores. *Psychometrika*, **66**, 563–576.
- SKRONDAL, ANDERS, & RABE-HESKETH, SOPHIA. 2004. *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall/CRC.
- SPEARMAN, CHARLES. 1904. "General intelligence," objectively determined and measured. *American Journal of Psychology*, **15**, 201–293.
- TAKANE, Y., & DE LEEUW, J. 1987. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, **52**, 393–408.
- THOMSON, G. H. 1951. *The factorial analysis of human ability*. London University Press.
- THURSTONE, L. L. 1931. Multiple factor analysis. *Psychological Review*, **38(5)**, 406–427.
- TUCKER, L. R. 1971. Relations of factor score estimates to their use. *Psychometrika*, **36**, 427–436.
- WRIGHT, B. D., & STONE, M. H. 1979. *Best test design. Rasch measurement*. Chicago: Mesa Press.
- WRIGHT, S. 1918. On the nature of size factors. *Genetics*, **3**, 387.
- YALCIN, I., & AMEMIYA, Y. 2001. Nonlinear factor analysis as a statistical method. *Statistic*, **16**, 275–294.

ZHENG, X., & RABE-HESKETH, S. 2007. Estimating parameters of dichotomous and ordinal item response models with gllamm. *Stata Journal*, **7**(3), 313–333(21).