

ANÁLISE FATORIAL

Daniel Abud Seabra Matos e
Erica Castilho Rodrigues

COLEÇÃO

Metodologias
de Pesquisa

Análise fatorial

Enap Escola Nacional de Administração Pública

Presidente

Diogo Godinho Ramos Costa

Diretoria de Seleção e Formação de Carreiras

Diana Magalhães de Souza Coutinho

Diretor de Educação Continuada

Paulo Marques

Diretor de Inovação e Gestão do Conhecimento

Guilherme Alberto Almeida de Almeida

Diretor de Pesquisa e Pós-Graduação

Fernando de Barros Filgueiras

Diretora de Gestão Interna

Camile Sahb Mesquita

Editor: Fernando de Barros Filgueiras. *Revisão:* Luiz Augusto Barros de Matos e Renata Fernandes Mourão. *Projeto gráfico e editoração eletrônica:* Ana Carla Gualberto Cardoso.

Análise fatorial

*Daniel Abud Seabra Matos e
Erica Castilho Rodrigues*

Brasília – DF
Enap
2019

Ficha catalográfica elaborada pela equipe da Biblioteca Graciliano Ramos da Enap

M4336a Matos, Daniel Abud Seabra

Análise fatorial / Daniel Abud Seabra Matos; Erica Castilho Rodrigues. -- Brasília: Enap, 2019.

74 p. : il. –

Inclui bibliografia.

ISBN: 978-85-256-0118-6

1. Análise Fatorial. 2. Análise Estatística. 3. Mensuração.
4. Ciências Sociais. 5. Políticas Sociais. I. Título. II. Rodrigues,
Erica Castilho.

CDU 519.237.7

Bibliotecária: Tatiane de Oliveira Dias – CRB1/2230

As opiniões emitidas nesta publicação são de exclusiva e inteira responsabilidade do(s) autor(es), não exprimindo, necessariamente, o ponto de vista da Escola Nacional de Administração Pública (Enap). É permitida a reprodução deste texto e dos dados nele contidos, desde que citada a fonte. Reproduções para fins comerciais são proibidas.

Enap Fundação Escola Nacional de Administração Pública

SAIS – Área 2-A

70610-900 – Brasília, DF

Telefones: (61) 2020 3096 / 2020 3102 – Fax: (61) 2020 3178

Sítio: www.enap.gov.br

SUMÁRIO

Introdução: pontos principais do livro	7
Análise fatorial	9
Níveis de mensuração das variáveis e diferentes tipos de correlação.....	11
Fatores, suas diferentes formas de representação e cargas fatoriais	15
Escore fatoriais.....	22
Método da regressão	24
Comunalidade, análise fatorial e análise de componentes principais.....	26
Autovalores, diagrama de inclinação e porcentagem de variância acumulada	28
Extração de fatores.....	32
Rotação de fatores	34
A escolha do método de rotação dos fatores.....	37
Planejamento e etapas da análise fatorial exploratória	39
<i>Etapas 1 – Verificação da adequação da base de dados.....</i>	<i>39</i>
<i>Etapas 2 – Número de fatores, extração, rotação e interpretação</i>	<i>44</i>
Exemplos de pesquisa	46
<i>Exemplo 1 – Concepções de avaliação</i>	<i>46</i>

<i>Passo 1 – Tamanho da amostra.....</i>	<i>48</i>
<i>Passo 2 – Nível de mensuração</i>	<i>48</i>
<i>Passo 3 – Matriz de correlações</i>	<i>49</i>
<i>Passo 4 – Teste de Bartlett (BTS).....</i>	<i>50</i>
<i>Passo 5 – Teste de Kaiser-Meyer-Olkin (KMO)</i>	<i>51</i>
<i>Passo 6 – Determinação do número de fatores.....</i>	<i>52</i>
<i>Passo 7 – Extração das cargas fatoriais</i>	<i>54</i>
<i>Passo 8 – Rotação dos fatores.....</i>	<i>55</i>
<i>Síntese dos resultados</i>	<i>59</i>
<i>Passo 9 – Interpretação dos fatores</i>	<i>63</i>
<i>Exemplo 2 – Uso de escores fatoriais em um modelo de regressão</i>	<i>67</i>
<i>Utilização dos escores fatoriais</i>	<i>69</i>
Referências bibliográficas	72

INTRODUÇÃO: PONTOS PRINCIPAIS DO LIVRO

Este livro é uma introdução ao método estatístico conhecido como análise fatorial exploratória (AFE).

Com relação ao conteúdo a ser abordado, os principais tópicos são:

- ✓ Noções básicas da análise fatorial.
- ✓ Análise de componentes principais e análise fatorial.
- ✓ Diferença entre análise fatorial exploratória e confirmatória.
- ✓ Tipos de correlação e estimação de parâmetros.
- ✓ Rotação de fatores: rotação ortogonal e oblíqua.
- ✓ Autovalor, cargas fatoriais e escores fatoriais.
- ✓ Análise multivariada de dados: técnicas de dependência e de interdependência.

✓ Principais aplicações da análise fatorial e combinação com outras técnicas multivariadas.

- ✓ Principais limitações da análise fatorial.
- ✓ Uso do pacote estatístico R.

Nesse sentido, os principais objetivos deste livro são:

✓ Possibilitar uma maior compreensão sobre o uso da análise fatorial, tendo como foco sua aplicação na área de Ciências Humanas e Sociais.

✓ Orientar quanto aos conceitos e aos mecanismos de aplicação da análise fatorial.

✓ Permitir uma leitura crítica e aplicada dos principais aspectos relacionados à análise fatorial, estabelecendo relações com outras técnicas multivariadas.

São ainda os resultados esperados após a leitura deste livro:

- ✓ Diferenciar a análise fatorial de outras técnicas multivariadas.
- ✓ Planejar e implementar a análise fatorial.
- ✓ Distinguir entre o uso exploratório e confirmatório da análise fatorial.

- ✓ Identificar como determinar o número de fatores a extrair.
- ✓ Entender o conceito de rotação de fatores.
- ✓ Explicar a função dos escores fatoriais e como usá-los.
- ✓ Identificar as principais limitações da análise fatorial.
- ✓ Realizar análise fatorial exploratória utilizando o *software* R.

ANÁLISE FATORIAL

No campo das Ciências Humanas e Sociais, frequentemente tentamos mensurar fenômenos que não são diretamente observáveis, que chamamos de variáveis latentes ou construtos. Inteligência, personalidade, motivação, nível socioeconômico, democracia e vulnerabilidade social são alguns exemplos de variáveis latentes. Essas variáveis são inferidas, por meio de um modelo matemático, de outras variáveis que são observáveis (medidas diretamente). É muito comum serem necessárias muitas variáveis para medir um único construto latente. Tomemos o nível socioeconômico (NSE) como exemplo. Para mensurá-lo, fazemos perguntas sobre questões como ocupação e escolaridade dos pais, bens domésticos, bens culturais e recursos educacionais da casa. A partir dessas variáveis diretamente observáveis estimamos o construto nível socioeconômico. Ou seja: o NSE é uma síntese feita a partir da combinação de vários elementos e por isso são necessários instrumentos como testes ou questionários que se associem a essa variável latente (SOARES, 2005). Dessa forma, o “NSE visa mensurar empiricamente um construto teórico que situa os indivíduos em classes ou estratos sociais, nos quais eles compartilham algumas características semelhantes tais como ocupação, renda ou educação” (ALVES *et al.*, 2013, p. 16).

Entretanto, operacionalizar conceitos em indicadores empíricos é algo bastante complexo. Continuando no nosso exemplo sobre o NSE, existem métodos diferentes de mensurar o nível socioeconômico e discordâncias na literatura sobre os elementos que precisam ser considerados para estimar esse construto. Um exemplo é o Programa Internacional de Avaliação de estudantes (PISA), que estima o NSE por meio de um índice chamado *International Socio-Economic Index of Occupational Status* (ISEI) (GANZEBOOM *et al.*, 1992). Nesse sentido, um mesmo conceito pode ser operacionalizado empiricamente de muitas maneiras, sendo a análise fatorial uma delas.

A análise fatorial (AF) é utilizada para investigar os padrões ou relações latentes para um número grande de variáveis e determinar se a informação pode ser resumida a um conjunto menor de fatores. Através da AF é possível “reduzir o número de dimensões necessárias para se descrever dados derivados de um grande número de medidas” (URBINA, 2007, p. 176). O fator pode ser definido como uma combinação linear das variáveis originais. Os fatores representam as dimensões latentes (construtos) que resumem o conjunto original de variáveis, mantendo a representatividade das características das variáveis originais (essa é uma grande vantagem dessa técnica, que será demonstrada no decorrer deste livro). Assim, a análise fatorial é usada para investigar as relações entre um grande número de variáveis e organizá-las em um conjunto menor de fatores (HAIR *et al.*, 2005). Portanto, os dois principais usos da análise fatorial são resumo e redução dos dados, que podem ser muito úteis à medida que o número de variáveis utilizadas em técnicas multivariadas aumenta.

Além disso, a análise fatorial (AF) é diferente de métodos de dependência, como a regressão múltipla, nos quais uma variável é considerada como dependente (resposta) e as outras independentes (explicativas). A AF é um método de interdependência, no qual todas as variáveis são consideradas simultaneamente. Cada variável é prevista por todas as outras. Assim, técnicas de dependência visam à previsão e à explicação, e as de interdependência visam à identificação de estrutura (HAIR *et al.*, 2005).

As técnicas fatoriais podem atingir seus objetivos por uma perspectiva exploratória (análise fatorial exploratória – AFE) ou por uma perspectiva confirmatória (análise fatorial confirmatória – AFC). Na AFE, deixamos os dados observados determinarem o modelo fatorial subjacente *a posteriori* (raciocínio indutivo para inferir um modelo a partir dos dados observados). Já na AFC, derivamos um modelo fatorial *a priori* (raciocínio dedutivo para fazer hipóteses de uma estrutura antecipadamente) (BRYANT; YARNOLD, 2000). Nesse sentido, uma técnica exploratória “deixa os dados falarem por eles mesmos”, não existe uma

intervenção do pesquisador predeterminando uma estrutura. Confiamos puramente na empiria dos dados e não estabelecemos restrições sobre a estimação ou número de componentes. Entretanto, em algumas situações, temos ideias sobre a estrutura dos dados, a partir de algum referencial teórico e/ou pesquisas anteriores. Nesses casos, temos a expectativa de que a AF desempenhe uma função confirmatória. A análise fatorial confirmatória determina como um modelo testado consegue se ajustar aos dados (KLEM, 2000). Portanto, na AFC precisamos especificar antecipadamente toda a estrutura a ser testada. Nesse caso, é essencial que a pesquisa seja dirigida pela teoria e/ou pesquisas anteriores.

Para Urbina (2007), a AFC é uma perspectiva mais nova e “busca hipóteses ou confirmar teorias a respeito de fatores presumidamente existentes” (URBINA, 2007, p. 176). Já a AFE tem por finalidade “descobrir quais fatores (isto é, variáveis latentes ou constructos) subjazem às variáveis em análise” (Urbina, 2007, p. 176). A AFE geralmente é usada nas fases mais embrionárias da pesquisa, para literalmente explorar os dados. Nessa etapa, o pesquisador explora a relação entre um conjunto de variáveis, identificando padrões de correlação. A AFC é usada para testar hipóteses, onde o pesquisador é guiado por teoria e testa em que medida as variáveis são representativas de um conceito/dimensão (FIGUEIREDO; SILVA, 2010).

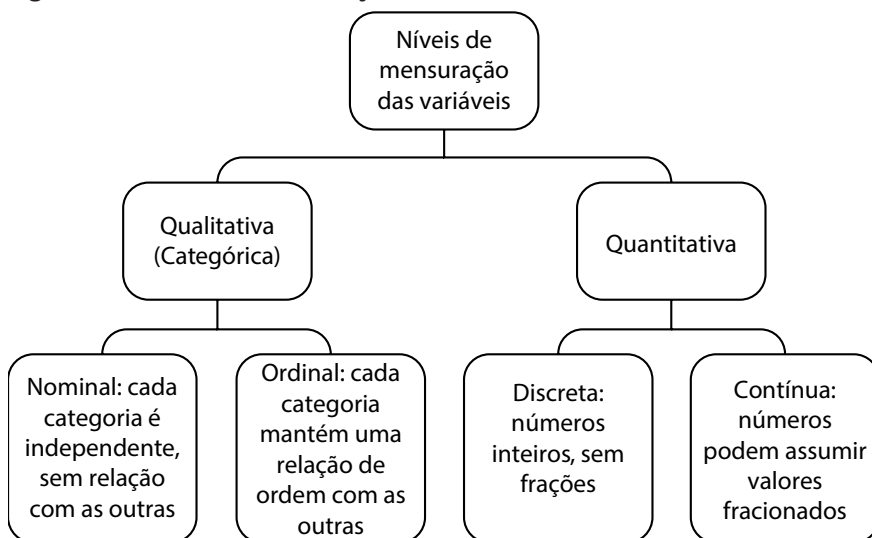
Neste livro, abordaremos apenas a análise fatorial exploratória. Isso porque o nível de complexidade da análise fatorial confirmatória foge ao escopo desta obra, pois a AFC pode ser considerada um tipo especial de modelagem de equações estruturais.

Níveis de mensuração das variáveis e diferentes tipos de correlação

Antes de explicar a análise fatorial propriamente dita, discutiremos alguns aspectos importantes sobre os tipos de variáveis e de correlação. Existem basicamente dois tipos de dados: não métricos (qualitativos) e métricos (quantitativos).

As medidas não métricas podem ser efetuadas em uma escala nominal ou ordinal. A escala nominal designa números utilizados simplesmente para rotular indivíduos ou objetos. Assim, os números ou símbolos designados aos objetos não possuem significado quantitativo além de indicar a presença ou ausência de um determinado atributo. São exemplos a religião ou partido político de um grupo de indivíduos. A escala ordinal reflete uma ordenação entre as categorias que constituem uma variável. Embora essas medidas sejam frequentemente representadas por números em uma escala, esses números possuem o significado de simplesmente indicar uma ordem. Assim, as variáveis são ordenadas em relação à quantia do atributo possuída. Podemos citar como exemplo uma escala Likert do tipo: quase nunca, raramente, às vezes, frequentemente e quase sempre (HAIR *et al.*, 2005).

As escalas intervalares e escalas de razão (ambas métricas) fornecem um nível mais alto de precisão de medida, permitindo que quase todas as operações matemáticas sejam realizadas. As duas escalas têm unidades constantes de medida. Assim, diferenças entre quaisquer dois pontos em qualquer parte da escala são iguais. A única diferença entre as duas é que as escalas intervalares têm um ponto zero arbitrário, enquanto as escalas de razão têm um ponto zero absoluto. Um exemplo de escala intervalar são as escalas de temperatura. Um exemplo de escalas de razão são as balanças para medir pesos. As escalas de razão representam a maior precisão possível de medida, pois todas as operações matemáticas são possíveis com elas (HAIR *et al.*, 2005). Essa mesma nomenclatura é utilizada por Babbie (1999) para os níveis de mensuração das variáveis. No entanto, dependendo da fonte consultada, podem ocorrer pequenas variações quanto a esses nomes. A Figura 1, além de sintetizar as informações sobre os níveis de mensuração, apresenta pequenas diferenças nos nomes:

Figura 1 – Níveis de mensuração das variáveis

Fonte: Matos (2010).

Como indicado na Figura 1, as variáveis qualitativas são frequentemente denominadas como categóricas. As variáveis quantitativas também podem ser classificadas em discretas e contínuas. Além disso, outro termo é usado para a classificação de um tipo especial de variável categórica: variável dicotômica ou binária (*dummy variable*). Como o próprio nome indica, só existem duas categorias (exemplo: escola privada e pública). Vale ainda destacar que uma variável categórica que apresenta mais de duas categorias pode ser representada por um conjunto de variáveis dicotômicas.

O pesquisador precisa entender os diversos níveis de medida por duas razões. Primeiro, ao identificar a escala de medida de cada variável que está sendo utilizada, devemos estar atentos para que dados qualitativos não sejam erroneamente usados como quantitativos e vice-versa. Segundo, o nível de medida é crucial para determinar que técnicas multivariadas são as mais adequadas aos dados, considerando tanto variáveis independentes como dependentes (HAIR *et al.*, 2005). Como exemplo, podemos citar a regressão. Nessa técnica, o que

determina o tipo de regressão a ser executada é a variável dependente: se a variável dependente for contínua, devemos conduzir uma regressão linear; se for dicotômica, uma regressão logística, e assim por diante. No caso da análise fatorial, quanto ao nível de mensuração, a literatura mais conservadora recomenda apenas o uso de variáveis contínuas ou discretas (FIGUEIREDO; SILVA, 2010). No entanto, devido à sofisticação atual dos *softwares* estatísticos, veremos que é possível realizar análise fatorial com variáveis categóricas. Isso é importante, pois nas Ciências Humanas e Sociais trabalhamos muito com variáveis categóricas. Porém, nem todos os *softwares* estatísticos possuem estimadores robustos para a análise de indicadores categóricos. Portanto, o pesquisador precisa ter clareza da análise que fez e de como ela foi calculada, inclusive para relatar eventuais limitações dos resultados. Todas essas questões serão evidenciadas e demonstradas no decorrer do livro.

Como existem tipos diferentes de variáveis, também existem tipos diferentes de correlação. Muitos métodos estatísticos se baseiam na utilização da correlação entre os dados para estimar seus resultados. A correlação mais conhecida é a de Pearson. Esse tipo de correlação requer que as variáveis tenham um nível de medida quantitativo. No entanto, existem outros tipos de correlação: bisserial, policórica, polisserial e tetracórica. Fundamentalmente, essas correlações são consequência de combinações de tipos de variáveis diferentes (binárias, ordinais, quantitativas) (HAIR *et al.*, 2005):

- ✓ **Correlação bisserial:** uma variável métrica é associada com uma medida binária.

- ✓ **Correlação policórica:** ambas as variáveis são medidas ordinais com três ou mais categorias.

- ✓ **Correlação polisserial:** uma variável métrica é associada com uma medida ordinal.

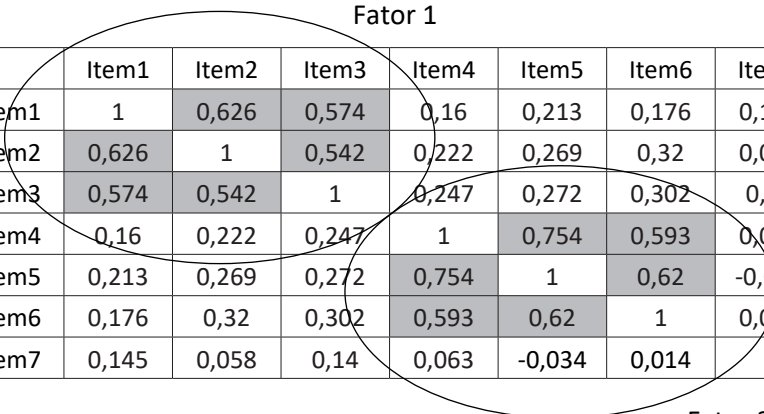
- ✓ **Correlação tetracórica:** ambas as variáveis são medidas binárias.

Como veremos na próxima seção, a correlação é um conceito estatístico muito importante para entender a análise fatorial.

Fatores, suas diferentes formas de representação e cargas fatoriais

Quando medimos um conjunto de variáveis, as correlações entre os pares de variáveis podem ser organizadas em uma matriz de correlações (matriz-R), que é uma tabela de coeficientes de correlação. O fato de existirem alguns coeficientes de correlação altos entre subconjuntos de variáveis sugere que elas podem estar medindo uma mesma dimensão subjacente, que é denominada como fator ou variável latente. Na análise fatorial, tentamos reduzir a matriz-R à sua dimensão subjacente investigando quais variáveis parecem se agrupar de maneira significativa (FIELD, 2009). Apresentamos a seguir um exemplo de análise de uma matriz de correlação na Tabela 1.

Tabela 1 – Matriz de correlações



	Item1	Item2	Item3	Item4	Item5	Item6	Item7
Item1	1	0,626	0,574	0,16	0,213	0,176	0,145
Item2	0,626	1	0,542	0,222	0,269	0,32	0,058
Item3	0,574	0,542	1	0,247	0,272	0,302	0,14
Item4	0,16	0,222	0,247	1	0,754	0,593	0,063
Item5	0,213	0,269	0,272	0,754	1	0,62	-0,034
Item6	0,176	0,32	0,302	0,593	0,62	1	0,014
Item7	0,145	0,058	0,14	0,063	-0,034	0,014	1

Fonte: elaboração própria.

Como apontamos anteriormente, a análise fatorial visa à redução de um conjunto de variáveis em um conjunto menor. Isso pode ser obtido identificando as variáveis que apresentam correlações altas com um grupo de variáveis específicas, mas que não se correlacionam (ou possuem correlações baixas) com as variáveis fora daquele grupo. Na

Tabela 1, parecem existir dois conjuntos dentro dessa lógica: variáveis inter-relacionadas, que devem estar medindo uma dimensão subjacente comum (fator). Esses grupos são: itens 1, 2 e 3 (fator 1), itens 4, 5 e 6 (fator 2). Ainda merecem destaque aqui dois aspectos: a) existe uma expectativa, na identificação de fatores, de que uma variável tenha correlações altas com algumas variáveis e baixas com outras. Isso não significa que as correlações com variáveis fora do seu fator sejam zero. Isso é algo difícil de acontecer, principalmente na área de Ciências Humanas e Sociais; b) o item 7 no nosso exemplo apresenta correlações baixas com todas as outras variáveis, indicando que ele não pertence a nenhum fator. Isso significa que, provavelmente, ao ajustar uma análise fatorial, esse item precisaria ser excluído. Em outras palavras, já na exploração inicial dos dados, podemos ter dicas sobre variáveis que provavelmente não irão se comportar bem na análise fatorial.

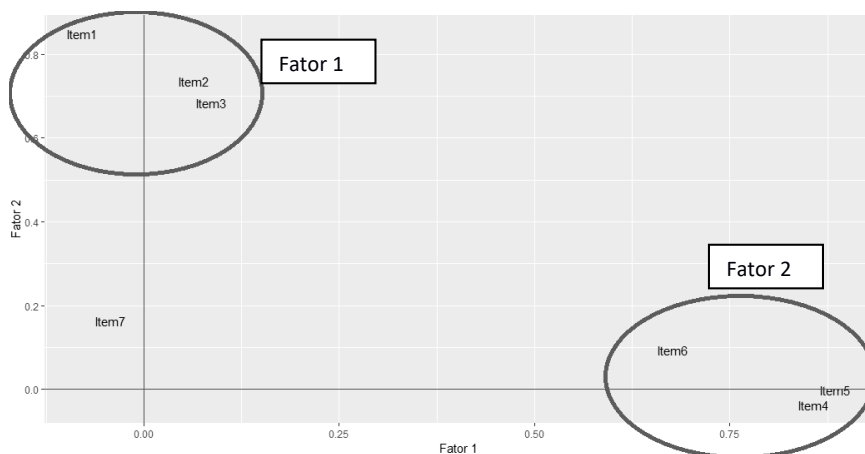
Aqui voltamos à nossa ideia do uso da análise fatorial para reduzir os dados. Dessa forma, temos vários itens para medir cada construto latente (fator), que não é observável diretamente. Portanto, existe um pressuposto de que os itens que medem o mesmo fator são altamente correlacionados. Sabemos que o coeficiente de correlação é a covariância corrigida pelas diferenças em desvio padrão. Dessa forma, a AF “se baseia no pressuposto fundamental de que alguns fatores subjacentes, que são em menor número que as variáveis observadas, são responsáveis pela covariação entre as variáveis” (KIM; MUELLER, 1978, p. 12). Ela identifica variáveis que “caminham juntas” (covariância), ou seja, variáveis que apresentam a mesma estrutura subjacente (TABACHINICK; FIDELL, 2007).

Os fatores podem ser representados de formas diferentes. Isso é importante para que possamos entendê-los melhor. Nesse sentido, apresentamos a seguir mais duas formas de representação dos fatores: gráfica e matemática.

Fatores são “abstrações empíricas” (entidades estatísticas) que podemos visualizar como eixos de um sistema de coordenadas onde representamos as variáveis. Se imaginarmos fatores como sendo os eixos de um gráfico, então podemos traçar variáveis ao longo desses eixos

(FIELD, 2009). A Figura 2 mostra os mesmos dados da Tabela 1, só que agora representados graficamente.

Figura 2 – Representação gráfica dos fatores



Fonte: elaboração própria.

A Figura 2 mostra um gráfico no qual existem apenas dois fatores. Vale destacar que, para os dois fatores, a linha do eixo varia entre -1 e 1 (limites do coeficiente de correlação). Os eixos do gráfico representam o peso ou a contribuição de cada um dos itens em cada um dos fatores. A proximidade entre os pontos indica variáveis que estão altamente correlacionadas: fator 1 – eixo vertical (itens 1, 2 e 3) e fator 2 – eixo horizontal (itens 4, 5 e 6). Essa representação gráfica, portanto, apenas mostra de outra maneira a mesma estrutura subjacente da matriz-R (Tabela 1), inclusive para o item 7. No caso desse item, na matriz-R, ele havia apresentado correlações baixas com todas as outras variáveis. No gráfico, isso é sinalizado da seguinte forma: ele se encontra espacialmente distante dos dois grupos de itens. Vejam que espacialmente os itens do fator 1 e do fator 2 estão muito próximos. Vale ainda destacar que, se tivéssemos um terceiro fator, ele seria representado por um terceiro eixo, o que resultaria em um gráfico 3-D. No entanto, a visualização se torna mais complexa, sendo que, para fins didáticos, o gráfico com duas dimensões é mais facilmente inteligível.

Se cada eixo do gráfico representa um fator, as coordenadas das variáveis ao longo de cada eixo representam a força da relação entre a variável e cada fator. Em uma situação ideal, uma variável deveria ter uma coordenada alta para um dos eixos e coordenadas baixas para todos os outros eixos (fatores). Isso pode ser considerado como um indicativo de que a variável se relaciona somente com um fator (ou muito fortemente com esse fator específico). A lógica aqui é: variáveis que têm coordenadas altas no mesmo eixo devem estar medindo aspectos de uma mesma dimensão comum subjacente (fator) (FIELD, 2009).

Aqui já podemos introduzir um conceito importante na análise fatorial: a carga fatorial. A coordenada de uma variável ao longo do eixo é conhecida como carga fatorial (*factor loading*). A carga fatorial pode ser definida como a correlação da variável com o fator¹. Se essa carga assume um valor positivo, significa que a variável está positivamente correlacionada com o fator, e, se assume valor negativo, essa correlação é negativa. Nesse segundo caso, a variável apresenta um sentido de variação oposto ao do construto. Para sabermos a proporção de variância explicada, devemos elevar a correlação (carga fatorial) ao quadrado (quando elevamos ao quadrado, o sinal não faz diferença). Isso nos dá um indicativo da importância de cada uma das variáveis para um fator específico. Como colocado anteriormente, os fatores resumem os dados, mas mantêm a representatividade das variáveis originais. Ou seja: os itens contribuem de maneira desigual para o fator: quanto maior a carga fatorial, maior a contribuição do item para o fator. Esse é um aspecto muito rico e interessante da análise fatorial: manter a representatividade das variáveis originais. Isso não acontece em outras técnicas mais simples de elaboração de índices que representam um conjunto de variáveis, que têm como pressuposto o fato de que todas as variáveis contribuem igualmente.

Quanto à representação matemática dos fatores, os eixos desenhados no gráfico anterior (Figura 2) são linhas retas e, portanto, podem ser descritos matematicamente pela equação de uma linha

¹ Essa é uma definição simplificada de carga fatorial. Por ora, ela é suficiente, mas será mais detalhada no decorrer do livro.

reta. Assim, fatores também podem ser representados em termos de uma equação, mais especificamente uma equação de um modelo linear (Equação 1). Aqui, dois pontos merecem ser destacados: não existe intercepto na equação, isso porque as linhas se interceptam em zero (dessa forma, o intercepto é também zero); os b s na equação representam as cargas fatoriais (FIELD, 2009).

Equação 1

$$Fator_i = b_1Variável_1 + b_2Variável_2 + \dots + b_nVariável_n + \varepsilon_i$$

Passamos a descrever a equação que representa os dados do nosso exemplo anterior² (Equação 2), onde temos dois fatores cujas equações são:

Equação 2

$$Fator_1 = b_1Item1_i + b_2Item2_i + b_3Item3_i + b_4Item4_i + b_5Item5_i + b_6Item6_i + \varepsilon_i$$

$$Fator_2 = b_1Item1_i + b_2Item2_i + b_3Item3_i + b_4Item4_i + b_5Item5_i + b_6Item6_i + \varepsilon_i$$

Fica evidente que as duas equações são iguais, pois ambas incluem todas as variáveis medidas (itens 1 ao 6). Porém, os valores de b nas equações serão diferentes, pois a importância de cada variável para os fatores varia. Nesse sentido, podemos substituir os valores de b nas equações pelas cargas fatoriais (as coordenadas das variáveis na Figura 2 (Equação 3)).

Equação 3

$$Fator_1 = 0,84Item1_i + 0,73Item2_i + 0,68Item3_i + 0,03Item4_i + 0,00Item5_i + 0,09Item6_i + \varepsilon_i$$

$$Fator_2 = -0,07Item1_i + 0,06Item2_i + 0,08Item3_i + 0,85Item4_i + 0,88Item5_i + 0,67Item6_i + \varepsilon_i$$

² No gráfico anterior, usamos o banco de dados com sete variáveis. A intenção era didática: mostrar como uma variável (item 7) não parecia estar associada a nenhum fator. No entanto, para a representação matemática do fator, optamos por não incluir o item 7.

A partir das duas equações, fica claro que, para o fator 1, os valores de b são altos para item 1, item 2 e item 3. Já nas outras variáveis (item 4, item 5 e item 6), os valores de b são baixos. Isso já foi mostrado anteriormente na representação gráfica. Nesse sentido, o gráfico e as equações estão representando a mesma coisa. Ou seja, as cargas fatoriais no gráfico são os valores de b nas equações. Para o fator 2, temos a situação oposta: os valores de b são altos para item 4, item 5 e item 6 e baixos para item 1, item 2 e item 3. Portanto, isso significa que, em cada um dos fatores, três variáveis são muito importantes (aquelas com valores altos de b) e outras três são pouco importantes (aquelas com valores baixos de b). Em uma situação considerada ideal, as variáveis precisam ter valores de b altos para um fator e valores de b baixos para os outros fatores.

As cargas fatoriais podem ser inseridas em uma matriz onde as colunas representam cada fator e as linhas representam as cargas fatoriais de cada uma das variáveis nos fatores. Essa matriz é denominada matriz fatorial (*factor matrix*) ou matriz de componentes (*component matrix*), no caso da análise de componentes principais³ (FIELD, 2009). Nos nossos dados, essa matriz tem duas colunas (representando dois fatores) e seis linhas (representando as seis variáveis).

$$A = \begin{pmatrix} -0,07 & 0,84 \\ 0,06 & 0,73 \\ 0,08 & 0,68 \\ 0,85 & -0,03 \\ 0,88 & 0,00 \\ 0,67 & 0,09 \end{pmatrix}$$

Para interpretar a matriz, devemos relacioná-la com as cargas descritas na equação 3. Por exemplo: a primeira linha representa o item 1 (a primeira variável), que tem uma carga fatorial de -0,07 para o fator 2 e uma carga de 0,84 para o fator 1. Aqui é importante destacar que, mesmo que uma variável apresente carga baixa em um fator (no nosso exemplo,

³ A diferença entre análise fatorial e análise de componentes principais será abordada no decorrer deste livro.

em alguns casos ela é próxima de zero), é esperado que todas as variáveis carreguem em todos fatores. Entretanto, lembramos mais uma vez: em uma boa solução na análise fatorial, esperamos sempre um padrão: uma variável deve ter carga fatorial alta no fator ao qual pertence e carga baixa nos demais fatores.

Por fim, precisamos esclarecer alguns pontos sobre as cargas fatoriais: ao tratarmos das representações gráfica e matemática, em alguns momentos descrevemos as cargas como a correlação entre a variável e um fator específico e em outros falamos das cargas como coeficientes da regressão (*b*). Primeiramente, de uma maneira geral, tanto o coeficiente de correlação quanto o de regressão representam o relacionamento entre a variável e o modelo linear. No entanto, a carga de um fator, dependendo da análise, pode ser tanto um coeficiente de correlação quanto um de regressão. Veremos, mais adiante no livro, que podemos assumir duas posições distintas: os fatores são ou não são correlacionados entre si⁴. Quando os fatores não são correlacionados, eles são assumidos como independentes e os valores dos coeficientes de correlação e de regressão são idênticos. Nessa situação, a interpretação da carga fatorial é mais simples. Adicionalmente, como em qualquer correlação, basta elevar a carga fatorial ao quadrado para saber a proporção de variância explicada. No entanto, existem casos nos quais os fatores são correlacionados entre si, onde as correlações entre as variáveis e os fatores são diferentes dos coeficientes de regressão. Nesse caso, existem dois conjuntos distintos (matrizes) de cargas fatoriais: os coeficientes de correlação entre as variáveis e os fatores (matriz de estrutura fatorial – *factor structure matrix*) e os coeficientes de regressão das variáveis em cada fator (matriz padrão fatorial – *factor pattern matrix*). Esses coeficientes podem ter interpretações diferentes, mas a maior parte dos pesquisadores interpreta a matriz padrão porque ela é normalmente mais simples (FIELD, 2009; GRAHAM *et. al.*, 2003).

⁴ Ver seção Rotação de fatores. Veremos dois tipos de rotação de fatores: ortogonal (os fatores são independentes, não correlacionados) e oblíqua (os fatores são correlacionados entre si).

Em síntese, se os fatores são correlacionados na análise fatorial exploratória, a carga fatorial é um coeficiente de regressão da variável no fator, não uma correlação. Portanto, por se tratar de um coeficiente de regressão, a carga fatorial pode inclusive apresentar um valor maior do que 1 (JORESOG, 1999). Caso isso aconteça em uma análise, apenas verifique a variância residual (*residual variance*) da variável cuja carga fatorial deu maior do que 1. Se a variância residual for negativa, a solução é inadmissível e a variável não deve ser usada. Variância residual negativa pode sugerir que fatores demais estão sendo extraídos. Caso a variância residual seja positiva, a variável com carga fatorial maior do que 1 pode ser utilizada (MUTHEN, 2005).

Outra forma de pensar é que a carga fatorial será sempre um coeficiente de regressão, porque para cada variável temos uma regressão com o fator latente como preditor. O modelo é $Y = \tau + \lambda Fator + \varepsilon$ onde Y é a variável, τ é o intercepto e λ é a carga fatorial. Como em qualquer regressão com apenas um preditor, o coeficiente padronizado é a correlação entre a variável dependente e a variável independente. Se tivermos mais de um fator, o coeficiente padronizado é uma correlação parcial controlando pelo efeito do outro fator.

Entretanto, a ideia mais importante que você deve ter em mente é: as cargas fatoriais nos dizem o quanto uma variável contribui para o fator, indicando que algumas contribuem mais e outras, menos. Essa é a questão-chave aqui.

Escores fatoriais

Estimadas as equações que representam cada um dos fatores, é possível calcular o valor daquele fator para cada um dos indivíduos de uma base de dados. Esses valores são conhecidos como escores fatoriais (*factor scores*). São uma espécie de média ponderada das variáveis observadas em cada uma das unidades amostrais, onde os pesos são dados justamente pelas cargas fatoriais. Para o exemplo anterior (Equação 3), elas seriam definidas por:

$$\begin{aligned}
 Fator_1 &= 0,84Item1_i + 0,73Item2_i + 0,68Item3_i + \\
 &\quad -0,03Item4_i + 0,00Item5_i + 0,09Item6_i + \varepsilon_i \\
 Fator_2 &= -0,07Item1_i + 0,06Item2_i + 0,08Item3_i + \\
 &\quad 0,85Item4_i + 0,88Item5_i + 0,67Item6_i + \varepsilon_i
 \end{aligned}$$

Essa é uma maneira bem simples de se calcular os escores fatoriais, mas pode apresentar problemas caso as variáveis estejam em escalas distintas. Nesse caso, os valores obtidos dos escores para cada um dos fatores irão depender da escala das variáveis originais e não poderão ser comparados entre si (Field, 2009). Além disso, as cargas fatoriais dependem da rotação que é feita nos fatores. Dessa maneira, utilizá-las como pesos diretamente no cálculo dos escores fatoriais pode levar a resultados ambíguos de acordo com a rotação feita por cada pesquisador. Distefano *et al.* (2009) chamam esse tipo de método de não refinado. Incluem ainda, nessa classe, a soma não ponderada das variáveis levando em conta apenas o sinal que carregam em cada fator, ou ainda a soma não ponderada das variáveis considerando apenas aquelas que possuem uma carga fatorial mínima. Os autores explicam que esse tipo de metodologia, apesar de simples de calcular e fácil de interpretar, é muito instável por depender fortemente da amostra em particular que está sendo analisada. Além disso, esse tipo de método produz escores fatoriais que estão correlacionados, mesmo quando os fatores são ortogonais no modelo ajustado. A segunda classe de métodos é denominada pelos autores como métodos refinados. Ao contrário dos anteriores, eles levam em conta tanto a variância compartilhada entre as variáveis e os fatores (comunalidade), bem como aquela que não é compartilhada. A maioria deles produz escores fatoriais que possuem uma distribuição normal padronizada e, portanto, variam no intervalo de -3 a 3. Esses métodos garantem resultados mais estáveis e mais confiáveis, que mantêm ainda o tipo de correlação existente entre os fatores, sejam eles ortogonais ou não. Distefano *et al.* (2009) classificam dentro desse conjunto as seguintes técnicas: método da regressão, escores de Bartlett, escores de

Anderson-Rubin. Iremos apresentar no livro o mais usado deles: método da regressão.

Destacamos ainda que, após a extração dos fatores e o cálculo dos escores fatoriais, estes podem ser utilizados como substitutos das variáveis originais nas análises. Por exemplo: se os fatores representam indicadores de sustentabilidade financeira dos municípios, podemos utilizá-los para fazer testes que comparem grupos de municípios distintos. Outro exemplo de situação em que podemos usar os escores fatoriais é quando ajustamos um modelo de regressão no qual existe multicolinearidade entre as variáveis. Esse tipo de comportamento pode levar a estimativas com uma variabilidade muito grande e muito sensíveis a mudanças sutis nos dados. Dessa maneira, a técnica de componentes principais pode ser usada para reduzir as variáveis originais a um número menor de componentes não correlacionados entre si. O modelo reajustado considerando os escores fatoriais como variáveis explicativas não irá mais apresentar o problema de multicolinearidade, e todos os resultados e interpretações serão agora baseados nos fatores (Farrar; Glauber, 1967). Em outras palavras, existem possibilidades ricas de combinar a análise fatorial com outras técnicas estatísticas, inclusive análises multivariadas.

Método da regressão

Nesse método, as equações que definem os fatores permanecem as mesmas: o que muda são as cargas fatoriais, ou seja, os valores de b . O método da regressão recalcula esses valores levando em conta as correlações entre as variáveis originais. Dessa maneira, os resultados dos escores se tornam invariantes à escala dessas variáveis (Field, 2009).

A ideia por trás desse método está no fato de considerar um modelo de regressão cujas variáveis explicativas são as variáveis observadas padronizadas e a variável resposta são os escores fatoriais. O peso de cada variável no cálculo dos escores fatoriais será dado pelos coeficientes desse modelo. Esses pesos são calculados multiplicando-se a matriz de cargas fatoriais pelo inverso da matriz de correlações R quando os fatores

são ortogonais entre si, e pelo inverso da matriz de correlação entre os fatores quando eles são oblíquos. Isso é uma maneira de se dividir os pesos que as variáveis têm nos fatores pelas suas correlações. Segundo Field (2009), a matriz de cargas resultante representa uma medida mais pura da relação única entre o fator e a variável.

No exemplo anterior, a matriz de cargas fatoriais original é dada por:

$$A = \begin{pmatrix} -0,07 & 0,84 \\ 0,06 & 0,73 \\ 0,08 & 0,68 \\ 0,85 & -0,03 \\ 0,88 & 0,00 \\ 0,67 & 0,09 \end{pmatrix}$$

O inverso da matriz de correlações é dado por:

$$R^{-1} = \begin{pmatrix} 1,92 & -0,88 & -0,66 & 0,04 & -0,16 & 0,21 \\ -0,88 & 1,88 & -0,42 & 0,05 & -0,04 & -0,32 \\ -0,66 & -0,42 & 1,69 & -0,09 & -0,02 & -0,19 \\ 0,04 & 0,05 & -0,09 & 2,47 & -1,55 & -0,50 \\ -0,16 & -0,04 & -0,02 & -1,55 & 2,63 & -0,67 \\ 0,21 & -0,32 & -0,19 & -0,50 & -0,67 & 1,83 \end{pmatrix}$$

Portanto, a matriz de coeficientes dos fatores fica igual a:

$$\begin{aligned} B &= R^{-1} A = \\ &\begin{pmatrix} 1,92 & -0,88 & -0,66 & 0,04 & -0,16 & 0,21 \\ -0,88 & 1,88 & -0,42 & 0,05 & -0,04 & -0,32 \\ -0,66 & -0,42 & 1,69 & -0,09 & -0,02 & -0,19 \\ 0,04 & 0,05 & -0,09 & 2,47 & -1,55 & -0,50 \\ -0,16 & -0,04 & -0,02 & -1,55 & 2,63 & -0,67 \\ 0,21 & -0,32 & -0,19 & -0,50 & -0,67 & 1,83 \end{pmatrix} \begin{pmatrix} -0,07 & 0,84 \\ 0,06 & 0,73 \\ 0,08 & 0,68 \\ 0,85 & -0,03 \\ 0,88 & 0,00 \\ 0,67 & 0,09 \end{pmatrix} \\ &= \begin{pmatrix} -0,22 & 0,55 \\ -0,06 & 0,32 \\ -0,05 & 0,28 \\ 0,40 & -0,12 \\ 0,56 & -0,19 \\ 0,17 & 0,00 \end{pmatrix} \end{aligned}$$

Comunalidade, análise fatorial e análise de componentes principais

Para os fins da análise fatorial, existem três tipos de variância total: 1) variância comum: variância compartilhada com outras variáveis na AF; 2) variância específica: variância de cada variável, única e que não é explicada ou associada com outras variáveis na AF; e 3) variância do erro: variância de uma variável devido a erros na coleta de dados ou na medida (HAIR *et al.*, 2005). Como mencionamos anteriormente, existe um pressuposto de que as variáveis que medem o mesmo fator são altamente correlacionadas, sendo que a correlação é a covariância corrigida pelas diferenças em desvio padrão. A AF identifica variáveis que “caminham juntas” (covariância). Nesse sentido, o tipo de variância mais importante para a AF é a variância comum.

Nesse ponto, podemos introduzir um conceito importante para a AF: a comunalidade, que pode ser definida como a “quantia total de variância que uma variável original compartilha com todas as outras variáveis incluídas na análise” (HAIR *et al.*, 2005, p. 90). Em outras palavras, a comunalidade é a proporção de variância comum presente numa determinada variável. Dessa forma, uma variável que não apresente variância específica ou de erro, teria uma comunalidade de 1, enquanto uma variável que não compartilhe variância com nenhuma outra variável teria uma comunalidade de valor 0. A literatura geralmente indica um valor mínimo de 0,5 para a comunalidade ser considerada satisfatória. Portanto, para uma variável funcionar bem em uma AF, ela precisa ter uma grande proporção de variância comum.

Existem muitas formas para estimar a comunalidade. A mais usada é a correlação múltipla ao quadrado (CMQ) de cada variável com todas as outras variáveis. Assim, se hipoteticamente temos um banco de dados com seis variáveis, suponha que você rodou uma regressão múltipla usando uma variável (variável 1) como dependente e as outras cinco variáveis como preditores: o R^2 múltiplo resultante é usado como estimativa da comunalidade para a variável 1. Esse é o procedimento da análise fatorial

e são essas estimativas que permitem que a AF seja realizada. Quando os fatores são extraídos, novas comunalidades (correlação múltipla entre cada variável e os fatores) podem ser calculadas. Logo, a comunalidade é uma medida da proporção de variância explicada pelos fatores (FIELD, 2009).

Podemos considerar duas abordagens para encontrar dimensões subjacentes nos dados: análise fatorial (AF) e análise de componentes principais (ACP). As duas técnicas tentam gerar combinações lineares das variáveis que capturem ao máximo possível a variância dessas variáveis observadas. A diferença é que na ACP toda a variância é utilizada, enquanto na AF apenas a variância compartilhada é usada (DANCEY; REIDY, 2006).

Dessa forma, somente a AF pode estimar os fatores subjacentes, pois a ACP se preocupa apenas em determinar que componentes lineares existem nos dados e como uma variável específica contribui com o componente (tecnicamente, temos produtos distintos nas duas técnicas: fatores na AF e componentes principais na ACP). Porém, para quem não é especialista em estatística, a diferença entre um componente principal e um fator pode ser complicada de explicar, pois ambos são modelos lineares e as diferenças estão principalmente nos cálculos. Adicionalmente, a ACP é conceitualmente menos complicada do que a AF e possui diversas similaridades com a análise discriminante (FIELD, 2009). Em síntese, podemos considerar que “se você estiver interessado numa solução teórica não contaminada por variabilidade de erro, a análise fatorial deve ser sua escolha. Se você quiser simplesmente um resumo empírico do conjunto de dados, a análise de componentes principais é uma escolha melhor” (TABACHINICK; FIDELL, 2007, p. 608).

Além disso, na maior parte dos casos, tanto a ACP quanto a AF chegam aos mesmos resultados se o número de variáveis superar 30 ou se as comunalidades excederem 0,60 para a maioria das variáveis (HAIR *et al.*, 2005). Já Stevens (1992) indica que, com 30 ou mais variáveis e comunalidades maiores do que 0,7 para todas as variáveis, as soluções provavelmente serão muito próximas. No entanto, para o autor, com um número menor do que 20 variáveis e com comunalidades baixas ($< 0,4$), podem acontecer resultados diferentes.

Por fim, é importante destacar que essa comparação entre AF e ACP é uma questão polêmica na literatura. Foge ao escopo deste livro aprofundar o debate. Apenas informamos o leitor sobre essas questões, com o intuito de esclarecer que tecnicamente as duas abordagens são diferentes.

Autovalores, diagrama de inclinação e porcentagem de variância acumulada

Para o ajuste do modelo de análise fatorial, o primeiro passo é definir o número de fatores que serão extraídos. Essa tarefa é complexa e consiste em encontrar a quantidade de fatores que representa melhor o padrão de correlação entre as variáveis. Nesse sentido, o pesquisador enfrenta um dilema entre parcimônia e explicação. Quanto mais fatores extrairmos, menor é o grau de parcimônia, entretanto, maior é a quantidade total de variância explicada pelos fatores. Inversamente, quanto menos fatores extrairmos, maior é o grau de parcimônia, no entanto, menor será a quantidade total de variância explicada. Portanto, a solução ótima seria encontrar o número mínimo de fatores que maximiza a quantidade de variância total explicada (FIGUEIREDO; SILVA, 2010).

Não existe um único critério consensual para determinar quantos fatores devemos extrair. Neste livro, abordamos os principais métodos indicados na literatura:

- ✓ critério do autovalor (*eigenvalue*);
- ✓ critério do diagrama de inclinação (Scree test);
- ✓ critério da porcentagem de variância acumulada.

A seguir explicamos detalhadamente cada um dos critérios, já adiantando que eles devem ser empregados conjuntamente.

O autovalor (*eigenvalue*) pode ser definido da seguinte forma:

o *eigenvalue* de um dado fator mede a variância em todas as variáveis que é devida ao fator. A razão de *eigenvalues* é a razão da importância explicativa dos fatores em relação às variáveis. Se um fator tem um *eigenvalue* baixo, ele contribui pouco para a explicação das variâncias

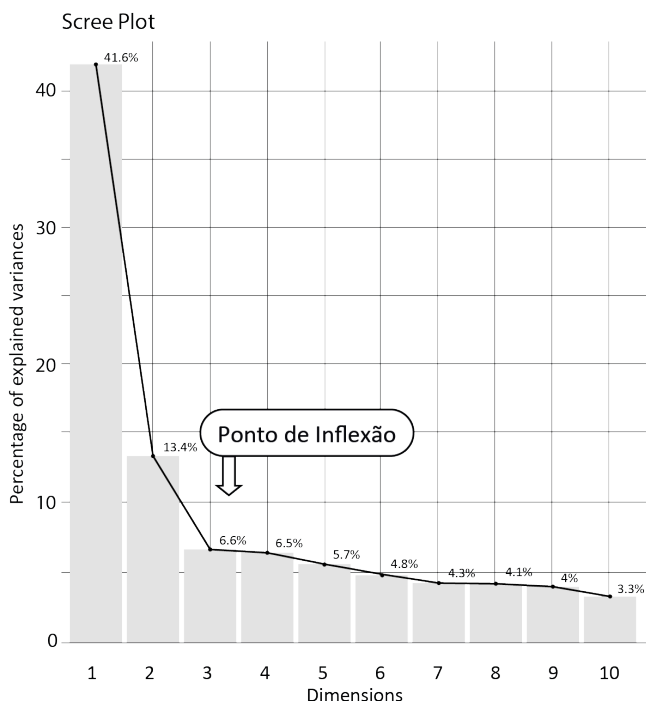
nas variáveis e pode ser ignorado como redundante em relação a fatores mais importantes (GARSON, 2009).

Como princípio básico da análise fatorial, devemos reter apenas fatores com autovalores grandes. Nesse sentido, existe uma regra bastante utilizada (critério de Kaiser) que sugere que devemos extrair somente os fatores com autovalor maior do que 1. Esse critério tem como base o raciocínio de que autovalores representam a quantidade de variação explicada por um fator e que um autovalor de 1 representa uma quantidade substancial de variação (FIELD, 2009). Esse critério de Kaiser costuma funcionar melhor quando o investigador analisa entre 20 e 50 variáveis (TABACHINICK; FIDELL, 2007). Já Stevens (1992) indica que o critério de Kaiser é mais preciso quando a quantidade de variáveis é menor do que 30 e as comunalidades após a extração são todas maiores do que 0,7. Ou quando o tamanho da amostra é maior do que 250 e a média da comunalidade é maior ou igual a 0,6. Entretanto, também existem alguns autores que consideram o critério de Kaiser muito rígido e conservador, sugerindo usar autovalores menores. Em síntese, quanto a esse primeiro método: o critério de Kaiser (autovalor maior do que 1) é o mais utilizado.

Outro método é o diagrama de declividade (*Scree test*) (Catell, 1966). Ele é um gráfico dos autovalores (eixo-y) e seus fatores associados (eixo-x). Podemos obter um fator para cada uma das variáveis existentes, e cada uma tem um autovalor associado. Se representarmos em um gráfico os autovalores, a importância de cada um deles fica bastante evidente. Nesse sentido, sempre temos alguns fatores com autovalores bastante altos e muitos fatores com autovalores considerados baixos. Portanto, esse gráfico tem um formato bem característico: uma inclinação bastante acentuada na curva seguida de uma cauda praticamente horizontal (FIELD, 2009). Cattell (1966) argumentou que o ponto de corte para decidir sobre o número de fatores deve ser no ponto de inflexão dessa curva. Ou seja: a curva da variância individual de cada fator se torna horizontal ou então sofre uma queda abrupta (Catell, 1966). Isso é um sinal de que muita variância foi perdida e devemos parar a extração

de fatores. A Figura 3 mostra um exemplo desse gráfico em que a altura das barras representa a proporção da variância explicada por cada um dos fatores e com a seta é destacado o ponto de inflexão. Segundo Stevens (1992), com amostras de mais de 200 sujeitos, o diagrama de declividade é um critério confiável.

Figura 3- Diagrama de declividade (Scree test)



Fonte: elaboração própria.

O terceiro critério é a porcentagem de variância acumulada para definir a quantidade de fatores que devemos extrair. Nesse método, a extração dos fatores continua até que um patamar específico seja obtido. Hair *et al.* (2005) sugerem o patamar de 60% como aceitável nas Ciências Humanas e Sociais. A Tabela 2 ilustra esse critério. Assim, pela porcentagem de variância acumulada, deveríamos reter 3 fatores, pois nesse ponto atingimos o patamar de 60%.

Tabela 2 – Autovalor e porcentagem de variância acumulada

Autovalor	Variância (%)	Variância acumulada (%)
4,99	41,61	41,61
1,61	13,45	55,06
0,80	6,63	61,69
0,78	6,52	68,20
0,68	5,66	73,86
0,58	4,82	78,68
0,51	4,26	82,94
0,50	4,13	87,08
0,48	3,97	91,05
0,39	3,27	94,32
0,35	2,88	97,19
0,34	2,81	100,00

Fonte: elaboração própria.

Ainda sobre a Tabela 2, verificamos uma questão interessante: segundo o critério de Kaiser (autovalor maior do que 1), deveríamos reter 2 fatores. Mas, como afirmamos anteriormente, pelo critério da variância acumulada seriam 3. Ainda, pelo diagrama de declividade (*Scree test*), esse número também seria de 3 para esse mesmo banco de dados (Figura 3). É interessante mencionar que, como regra geral, o teste *scree* resulta em pelo menos 1, e às vezes 2 ou 3 fatores a mais, em relação ao critério de Kaiser (autovalor >1) (HAIR *et al.*, 2005).

O que isso nos diz? Primeiro, que na prática não costumamos usar um único critério. Eles são usados conjuntamente. Segundo, nem sempre todos os critérios são concordantes. No exemplo anterior, isso ficou evidente. Assim, para decidir sobre o número de fatores a extrair, voltamos a um aspecto já colocado: a parcimônia é importante (a solução ótima é encontrar o número mínimo de fatores que maximiza a quantidade de variância total explicada).

Aqui também vale sempre lembrar aspectos básicos de pesquisa: volte nos seus objetivos e na teoria. Pergunte-se: teoricamente faz mais sentido essas variáveis estarem agrupadas, por exemplo, em 2 ou 3 fatores? Isso é muito importante, pois é comum encontrarmos mais de uma solução empírica aceitável, do ponto de vista estatístico, quanto ao número de fatores. Assim, a decisão final pode acabar sendo teórica. Lembre sempre que as variáveis de um fator devem medir um mesmo construto latente e precisam estar associadas entre si.

Por fim, uma síntese sobre os critérios para definir o número de fatores que serão extraídos:

- ✓ critério do autovalor (eigenvalue) (critério de Kaiser): apenas os fatores que tem autovalor >1 ;

- ✓ critério do diagrama de inclinação (Scree test): a curva da variância individual de cada fator se torna horizontal ou sofre uma queda abrupta (ponto de inflexão da curva);

- ✓ critério da porcentagem de variância: patamar de 60% da variância acumulada.

(Os critérios devem ser usados em conjunto.)

Extração de fatores

Após a definição do número de fatores do modelo, o passo seguinte é decidir qual técnica será utilizada para o cálculo das cargas fatoriais. Essa é a chamada extração dos fatores. Existem várias técnicas para extrair os fatores. A escolha dentre as várias possibilidades depende do tipo de dado que está sendo analisado e do objetivo da análise. Entretanto, segundo Tabachnick e Fidell (2007), para uma base de dados com um número razoável de observações e variáveis, os resultados não podem ser muito discrepantes entre si. Essa inclusive é uma das formas de validar as estimativas encontradas, de mostrar que os resultados são estáveis. Quanto às opções disponíveis para os métodos de extração dos fatores, apresentamos a seguir alguns dos principais métodos que estão implementados nos *softwares* estatísticos: componentes principais (*principal components*),

fatores principais (*principal factors*), máxima verossimilhança (*maximum likelihood*), mínimos quadrados ordinários (*ordinary least squares*), mínimos quadrados generalizados (*generalized least squares*).

O método de componentes principais é um dos mais comuns e produz combinações lineares das variáveis originais que sejam independentes entre si e expliquem o máximo da variabilidade dos dados. A primeira componente explica a maior parte dessa variância, a segunda é a que possui o segundo maior poder de explicação, e assim por diante. Juntas, todas as componentes explicam toda a variabilidade dos dados. Para Tabachnick e Fidell (2007), essa técnica deve ser escolhida se o interesse do pesquisador é resumir um grande número de variáveis em um conjunto menor. Ela ainda pode servir como um passo inicial na análise fatorial e ajudar na determinação do número de fatores⁵.

A técnica de fatores principais é semelhante à anterior. O objetivo é maximizar a variabilidade explicada pelos fatores. A diferença é que a comunalidade das variáveis é estimada de maneira iterativa. Como o objetivo é apenas o de maximizar a variância, pode ser que a matriz de correlações estimada não fique muito próxima da matriz real (TABACHNICK; FIDELL, 2007).

A máxima verossimilhança encontra as cargas fatoriais que maximizem a probabilidade da amostra gerar a matriz de correlações observada. Esse método é um dos mais utilizados e está implementado em praticamente todos os *softwares* estatísticos (COSTELLO; OSBORNE, 2005).

A técnica dos mínimos quadrados ordinários encontra os fatores de forma que a soma do quadrado da diferença entre a matriz observada e a estimada seja mínima. Nessa soma, são considerados apenas os termos fora da diagonal principal, visto que essa parte, que contém as comunalidades das variáveis, será estimada posteriormente (TABACHNICK; FIDELL, 2007).

O procedimento dos mínimos quadrados generalizados é bem semelhante ao anterior. A diferença é que nesse método as variáveis

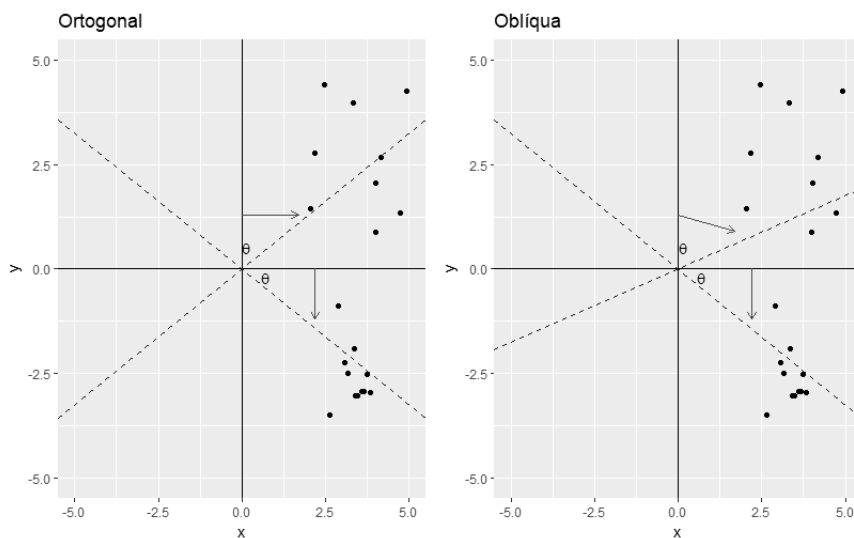
⁵ Recomendamos a leitura da seção Comunalidade, análise fatorial e análise de componentes principais para relembrar as diferenças entre as duas técnicas.

são **ponderadas de acordo com sua** comunalidade. Variáveis que apresentarem um valor alto para essa medida, ou seja, apresentam grande parte da sua variância compartilhada com as demais, receberão um peso maior. Por outro lado, aquelas que não podem ser explicadas pelas demais de maneira razoável receberão um peso menor (TABACHNICK; FIDELL, 2007).

Rotação de fatores

Após a extração dos fatores, podemos calcular o grau de adaptação das variáveis aos fatores por meio das cargas fatoriais. Normalmente, o que acontece é que a maior parte das variáveis tem cargas altas no fator mais importante e cargas baixas nos outros fatores. Isso torna a interpretação mais difícil. Nesse sentido, a técnica de rotação de fatores é utilizada para atingir uma melhor distinção entre os fatores (Field, 2009). Para entender melhor essa técnica, precisamos retomar a representação gráfica (geométrica) dos fatores (Figura 4).

Figura 4 – Ilustração da rotação de fatores



Fonte: elaboração própria.

Se tomarmos um fator como um eixo ao longo do qual podemos traçar as variáveis, o método de rotação dos fatores gira esses eixos de forma que as variáveis estejam carregadas ao máximo somente em um fator (FIELD, 2009). Assim, podemos dizer que “o método de rotação se refere ao método matemático que rotaciona os eixos no espaço geométrico. Isso torna mais fácil determinar quais variáveis são carregadas em quais componentes” (SCHAWB, 2007).

A Figura 4 ilustra essa técnica por meio de um exemplo em que temos apenas dois fatores. Nessa representação gráfica, as linhas contínuas representam os fatores, e as linhas pontilhadas, a rotação dos fatores (os eixos são girados). Portanto, depois da rotação, as cargas fatoriais das variáveis são maximizadas sobre um fator (aquele que intercepta o aglomerado) e minimizadas nos outros fatores (FIELD, 2009). Em outras palavras, podemos dizer que, quanto mais próximas as variáveis estiverem do fator (eixo), menor será o resíduo do modelo.

Existem dois tipos de rotação de fatores: ortogonal e oblíqua. Na rotação fatorial ortogonal, cada fator é independente (ortogonal) em relação a todos os outros (a correlação entre eles é assumida como sendo zero). Já a rotação fatorial oblíqua é calculada de maneira que os fatores extraídos são correlacionados (HAIR *et al.*, 2005). A rotação ortogonal está ilustrada no lado esquerdo da Figura 4. A palavra ortogonal significa não relacionado (não existe correlação entre os fatores). Assim, os fatores são mantidos independentes durante a rotação (os eixos são mantidos perpendiculares, com um ângulo reto). Por outro lado, na rotação oblíqua (lado direito da Figura 4), os fatores estão correlacionados (os eixos não são mantidos perpendiculares).

Em síntese, o principal objetivo da rotação dos fatores é tornar o resultado empírico encontrado mais facilmente interpretável, conservando as suas propriedades estatísticas (FIGUEIREDO; SILVA, 2010). A rotação de fatores é um método usado após a extração de fatores para maximizar cargas altas entre os fatores e as variáveis e minimizar as cargas baixas (TABACHNICK; FIDELL, 2007). Essa otimização gerada pela rotação de fatores também pode ser ilustrada de outras formas (Tabela 3).

Tabela 3 – Efeito da rotação de fatores nos autovalores e porcentagem de variância

Fator	Sem rotação			Com rotação		
	Autovalor	Variância explicada (%)	Variância acumulada	Autovalor	Variância explicada (%)	Variância acumulada
1	4,74	0,4	0,4	3,38	0,28	0,28
2	1,3	0,11	0,51	2,66	0,22	0,50

Fonte: elaboração própria.

A Tabela 3 apresenta o resultado do ajuste do modelo para os dados do exemplo gráfico anterior e mostra como a rotação de fatores “equilibra” o modelo final. Quando comparamos a solução rotacionada com a solução inicial, percebemos uma mudança principal: o valor dos autovalores e a porcentagem de variância explicada por eles são distribuídos de uma maneira mais equilibrada na solução rotacionada. A modificação é percebida em ambos os fatores. Antes, o fator 1 tinha uma importância bem maior do que o fator 2. Depois, a sua importância explicativa diminui, assim como aumenta a importância do fator 2.

Já a Tabela 4 mostra o efeito da rotação de fatores nas cargas fatoriais.

Tabela 4 – Exemplo matriz de cargas rotacionadas e não rotacionadas

Variável	Sem rotação		Com rotação	
	Fator 1	Fator 2	Fator 1	Fator 2
Item 1	0,4941	-0,1382	0,4598	0,0936
Item 2	0,6994	-0,1758	0,6321	0,1533
Item 3	0,6099	-0,2233	0,6173	0,0603
Item 4	0,6173	-0,2144	0,6138	0,0731
Item 5	0,6555	-0,3963	0,8109	-0,0994
Item 6	0,7020	-0,3215	0,7712	0,0017
Item 7	0,6269	-0,3272	0,7265	-0,0406
Item 8	0,5406	0,0538	0,3097	0,3175

	Sem rotação		Com rotação	
Item 9	0,6022	0,3754	0,0475	0,6846
Item 10	0,7024	0,3971	0,0937	0,7559
Item 11	0,5816	0,4938	-0,0779	0,7989
Item 12	0,6731	0,4888	-0,0122	0,8380

Fonte: elaboração própria.

Nas duas primeiras colunas da Tabela 4, percebemos que algumas variáveis (por exemplo, os itens 9, 10, 11 e 12) possuem cargas fatoriais altas em ambos os fatores. Isso é chamado de indeterminação fatorial, pelo fato de não conseguirmos identificar a qual fator a variável pertence. Já nas últimas colunas, que apresentam o resultado com a rotação, essas variáveis passaram a apresentar carga alta no fator 2 e baixa no fator 1, resolvendo o problema da indeterminação fatorial. Ou seja, a solução do modelo foi otimizada. Além disso, geralmente, nas análises, mesmo em variáveis com uma diferenciação mais clara entre os fatores, as cargas também tendem a ficar melhores (aumentar em um fator e diminuir em outro) após a rotação.

Vale ainda destacar que, mesmo após a rotação dos fatores, variáveis com indeterminação fatorial podem persistir. No nosso exemplo, isso é ilustrado pelo item 8. Nas partes posteriores de análises do livro, daremos exemplos de como lidar com esse tipo de situação. Por enquanto, estamos apenas ilustrando mais uma das propriedades da rotação de fatores.

A escolha do método de rotação dos fatores

Quanto à escolha do tipo de rotação, essa é uma questão bastante polêmica na literatura. Existem autores que argumentam que, geralmente, as técnicas de rotação ortogonal e oblíqua levam a resultados semelhantes, principalmente quando o padrão de correlação entre as variáveis é claro (PALLANT, 2007; HAIR *et al.*, 2005). Entretanto, Field (2009) nos alerta para uma série de questões: a) a escolha da rotação vai depender se temos uma boa razão teórica para supor que os fatores sejam relacionados

ou independentes; b) usar rotação ortogonal com dados de Ciências Humanas e Sociais não parece ter nenhum sentido. Nessas áreas, as variáveis quase sempre são correlacionadas. Assim, existem autores que defendem que esse tipo de rotação fatorial nunca deveria ser utilizado em Ciências Humanas e Sociais. Portanto, para usar rotação ortogonal, o pesquisador precisaria ter evidências teóricas ou empíricas muito fortes de que os fatores não são correlacionados.

As rotações ortogonais são mais fáceis de relatar e interpretar. Entretanto, devemos assumir que os construtos são independentes. Na prática, esse pressuposto é difícil de ser atendido. Já as rotações oblíquas permitem que os fatores sejam correlacionados, mas são consideradas mais difíceis de descrever e interpretar (TABACHINICK; FIDELL, 2007).

Quanto às opções disponíveis para cada um desses tipos de rotação, o *software* R possui quatro métodos de rotação ortogonal (*varimax*, *quartimax*, *BentlerT* e *geominT*) e cinco métodos de rotação oblíqua (*oblimin*, *promax*, *simplimax*, *BentlerQ* e *geominQ*). Os métodos diferem na forma de rotacionar os fatores e produzem resultados distintos.

Os dois métodos mais importantes de rotação ortogonal são *quartimax* e *varimax*. A rotação *quartimax* visa maximizar a dispersão da carga dos fatores de uma variável por todos os fatores. No entanto, isso geralmente ocasiona muitas variáveis com cargas altas em um único fator. Já a rotação *varimax* é o procedimento oposto, ou seja, ela tenta maximizar a dispersão das cargas dentro dos fatores. Assim, ela tenta carregar um número menor de variáveis altamente em cada fator, o que resulta em grupos de fatores mais interpretáveis. *Varimax* tenta evitar que muitas variáveis tenham cargas altas em um único fator. O método *varimax* é o mais utilizado (FIELD *et. al.*, 2012).

Os dois métodos mais importantes de rotação oblíqua são *promax* e *oblimin*. *Promax* é um método mais rápido desenvolvido para bancos de dados muito grandes. A rotação *oblimin* é geralmente a mais indicada (FIELD *et al.*, 2012).

Em síntese, as recomendações gerais são: se você espera que os fatores sejam independentes, use *varimax*. Se você espera que os

fatores sejam correlacionados, use *oblimin*. Isso diminui as chances de você errar. No entanto, os métodos podem variar dependendo dos objetivos e da análise.

Planejamento e etapas da análise fatorial exploratória

Após toda essa primeira parte teórica do livro, podemos fazer uma síntese, um planejamento sobre a análise fatorial exploratória, que possui vários passos. É muito importante cumprir cada um deles adequadamente para que o resultado da análise seja satisfatório. A seguir, dividimos o planejamento em 2 etapas, que são discutidas detalhadamente.

Etapla 1 – Verificação da adequação da base de dados

Antes de se iniciar a análise fatorial, é necessário verificar se os dados são adequados para esse tipo de modelo. Nessa etapa, alguns pontos são importantes: amostra, nível de mensuração das variáveis, padrão de correlações, Teste de Bartlett e Teste de Kaiser-Meyer-Olkin.

Tamanho da amostra

O tamanho mínimo da amostra para se ajustar uma AF depende do número de variáveis que estão sendo analisadas. Quanto maior esse número, mais dados devem ser coletados, pois mais parâmetros precisam ser estimados. Segundo Costello e Osborne (2005), grande parte dos trabalhos recomendam um mínimo de 10 observações para cada variável coletada. Segundo esses autores, o tamanho também vai depender da natureza dos dados observados. Se as variáveis se separam muito bem nos fatores (não apresentam cargas fatoriais cruzadas) e apresentam comunalidade alta, não são necessários muitos dados.

Hair *et al.* (2005) indicam que dificilmente conseguimos realizar uma AF com uma amostra menor do que 50 observações, sugerindo que, de preferência, o tamanho deve ser maior ou igual a 100. Os autores ainda fazem as seguintes recomendações: como regra geral, ter

pelo menos cinco vezes mais observações do que o número de variáveis analisadas, mas sendo ideal uma proporção de dez para um; sempre tentar ter a maior razão possível de casos por variável, o que pode ser obtido sendo parcimonioso com a escolha das variáveis (seleção teórica e prática das variáveis).

Segundo Pasquali (1999), técnicas estatísticas como a análise fatorial fazem exigências grandes dos dados. Assim, eles precisam produzir variância suficiente para que a análise seja consistente. O autor faz a seguinte recomendação sobre a AF: qualquer análise com menos de 200 observações dificilmente pode ser levada em consideração.

Já Field *et al.* (2012) argumentam que a amostra pode variar em função de vários pontos, recomendando que o número de observações coletadas siga os critérios apresentados na Tabela 5:

Tabela 5 – Relação entre o tamanho da amostra e outros dados da AF

Cargas fatoriais	Tamanho mínimo de amostra
4 ou mais cargas maiores que 0,6 no fator	Não existe tamanho mínimo
10 ou mais cargas maiores que 0,4 nos fatores	150
Fatores com algumas cargas baixas	300
Comunalidade	Tamanho de amostra
Todas maiores que 0,6	Mesmo amostras pequenas (menos de 100) podem ser adequadas
Em torno de 0,5	Entre 100 e 200
Muito abaixo de 0,5	Acima de 500

Fonte: Guadagnoli e Velicer (1988); MacCallum, Widaman, Zhang e Hong (1999).

Em síntese, quanto ao tamanho da amostra, não existe uma regra única na literatura. Entretanto, se pudermos definir uma regra de ouro, ela seria: precisamos trabalhar com amostras grandes e tentar obter sempre a maior razão possível de casos por variável (por isso não devemos incluir variáveis aleatoriamente no modelo). Nem sempre é fácil delimitar o que seja uma “amostra grande”, mas as recomendações anteriores nos ajudam a entender essa questão. Para Field *et al.* (2012), uma amostra

de 300 ou mais provavelmente resultará numa solução estável da análise fatorial, mas é importante o pesquisador medir variáveis suficientes para todos os fatores esperados teoricamente.

Nível de mensuração das variáveis

Esse tópico foi tratado em detalhes na seção *Níveis de mensuração das variáveis e diferentes tipos de correlação*. Aqui apenas vamos lembrar alguns pontos importantes. Existem basicamente dois tipos de dados: não métricos (qualitativos) e métricos (quantitativos). As medidas não métricas (categóricas) podem ser efetuadas em uma escala nominal, ordinal e dicotômica. As variáveis quantitativas podem ser classificadas em discretas e contínuas.

No caso da AF, a literatura mais conservadora recomenda apenas o uso de variáveis contínuas ou discretas (FIGUEIREDO; SILVA, 2010). Entretanto, com a sofisticação atual dos *softwares* estatísticos, é possível realizar análise fatorial com variáveis categóricas, algo essencial nas Ciências Humanas e Sociais. No R, as análises podem ser feitas, por exemplo, usando a função *polychor()*, do pacote *psych*. Mas nem todos os *softwares* estatísticos possuem estimadores robustos para a análise de indicadores categóricos. Portanto, o pesquisador deve ter clareza da análise realizada e de suas eventuais limitações.

Tudo isso nos importa para pensar sobre os tipos diferentes de correlação. Além da correlação de Pearson, na qual as variáveis precisam ter um nível de medida quantitativo, existem outros tipos de correlação (HAIR *et al.*, 2005):

- ✓ **Correlação bisserial:** uma variável métrica é associada com uma medida binária.
- ✓ **Correlação policórica:** ambas as variáveis são medidas ordinais com três ou mais categorias.
- ✓ **Correlação polisserial:** uma variável métrica é associada com uma medida ordinal.
- ✓ **Correlação tetracórica:** ambas as variáveis são medidas binárias.

Correlação

A AF só faz sentido se as variáveis analisadas forem altamente correlacionadas entre si. Uma vez definido o nível de mensuração das variáveis e o tipo de correlação adequada, antes de se iniciarem as análises, precisamos verificar a matriz de correlações. Como apresentado anteriormente, variáveis que medem o mesmo construto devem apresentar uma correlação alta. Field *et al.* (2012) sugerem que a maioria das entradas da matriz devem estar acima de 0,3. Se algumas variáveis tiverem muitas correlações abaixo desse valor, elas são candidatas a serem excluídas das análises. É importante lembrar que essa é apenas uma inspeção inicial dos dados e que sempre devemos utilizar uma combinação de critérios para tomar decisões sobre a AF. Entretanto, correlações baixas são um primeiro indício de alerta sobre possíveis variáveis problemáticas. Além da análise visual da matriz de correlação, o caso mais extremo em que todas as variáveis são independentes entre si pode ser verificado por meio do Teste de Bartlett, que será apresentado a seguir.

Segundo Field *et al.* (2012) a situação no outro extremo, em que as variáveis são perfeitamente correlacionadas, também causa problemas na estimação do modelo. Esse problema é conhecido como multicolinearidade e também ocorre quando ajustamos modelos de regressão. Acontece que, quando duas variáveis têm uma correlação quase perfeita, fica inviável separar o peso delas em cada um dos fatores. Os autores recomendam verificar se existem muitos casos em que a correlação é superior a 0,8. Yong e Pearce (2013) recomendam calcular, para cada variável, a SMC (*Squared Multiple Correlation*), que é uma espécie de medida inicial da comunalidade, que mede o quanto da variabilidade de cada variável pode ser explicada pelas demais. Segundo os autores, variáveis com SMC muito próxima de zero (completamente independente das demais) ou um (tem toda sua variabilidade explicada pelas outras) são possíveis candidatas a serem descartadas da base de dados.

Teste de Bartlett (Bartlett's test of sphericity – BTS)

Na situação extrema de independência perfeita entre todas as variáveis, a matriz de correlação se reduz à matriz identidade, pois todos os elementos fora da diagonal principal são iguais a zero. Isso significa que as variáveis não se agrupam para formar nenhum construto e, portanto, a construção dos fatores perde todo seu sentido. O Teste de Bartlett tem essa situação como sua hipótese nula e, caso ela seja rejeitada, pode-se concluir que existe algum tipo de associação entre as variáveis e que elas podem, de fato, representar conjuntamente um ou mais traços latentes. Portanto, o Teste de Bartlett deve ser estatisticamente significativo ($p < 0,05$).

Segundo Field *et al.* (2012), o Teste de Bartlett, como todo teste de hipótese, depende muito do tamanho amostral e tende a rejeitar a hipótese nula para amostras grandes. Como no caso da AF não podemos trabalhar com amostras pequenas, a significância desse teste não é uma garantia de que todas as variáveis vão se agrupar em fatores. Os autores recomendam, portanto, excluir aquelas que apresentam uma correlação muito baixa com todas as demais.

Teste de Kaiser-Meyer-Olkin (KMO)

Além do Teste de Bartlett, outro teste deve ser realizado para verificar a adequabilidade da amostra: o KMO (Kaiser-Meyer-Olkin). Essa medida, que varia entre 0 e 1, representa a proporção da variância das variáveis que pode ser explicada pelos fatores ou traços latentes. Quanto mais próximo esse valor estiver de 1, mais adequados os dados estão para se ajustar uma AF. Field *et al.* (2012) recomendam utilizar os critérios apresentados na Tabela 6.

Tabela 6 – Critério de corte dos valores do KMO

KMO	Adequabilidade da amostra
< 0,5	Inaceitável
[0,5 - 0,7]	Medíocre
[0,7 - 0,8]	Bom
[0,8 - 0,9]	Ótimo

KMO	Adequabilidade da amostra
>0,9	Excelente

Fonte: Hutcheson e Sofroniou (1999).

Hair *et al.* (2005) e Kaiser (1974) indicam 0,5 como valor mínimo aceitável (valores abaixo disso sugerem a necessidade de coletar mais dados ou repensar quais variáveis devem ser incluídas).

A Tabela 7 sintetiza a primeira etapa de planejamento da análise fatorial exploratória.

Tabela 7 – Síntese da etapa 1: adequação da base de dados

Tamanho da amostra	Amostras grandes (maior do que 100) e pelo menos 5 vezes mais observações do que o número de variáveis (o ideal seria pelo menos 10 vezes mais observações).
Nível de mensuração das variáveis	Variáveis categóricas (ordinal e dicotômica) e quantitativas (discretas e contínuas). Definir o tipo de correlação: Pearson, bisserial, policórica, polisserial e tetracórica.
Matriz de correlação	A maioria dos coeficientes de correlação deve ter valores maiores do que 0,3.
Teste de Bartlett (BTS)	Deve ser estatisticamente significativo: $p < 0,05$.
Teste de Kaiser-Meyer-Olkin (KMO)	Quanto mais próximo de 1 melhor. 0,5 como valor mínimo aceitável, mas o ideal seria um valor a partir de 0,7.

Fonte: elaboração própria.

Etapa 2 – Número de fatores, extração, rotação e interpretação

Após a exploração inicial dos dados e a verificação de que a base é adequada, devemos realizar a análise fatorial propriamente dita. Aqui estão envolvidos os seguintes passos: determinação do número de fatores, extração das cargas fatoriais, rotação e interpretação dos fatores. Todos esses passos já foram abordados em seções anteriores. A única exceção foi o passo da interpretação dos fatores, que passamos a explicar em seguida.

A análise fatorial é uma técnica puramente empírica. Nesse sentido, um fator (construto latente) pode ser considerado como meramente uma “abstração empírica”. Entretanto, esperamos que o fator encontrado faça sentido do ponto de vista substantivo, teórico. Assim, a fase final da análise fatorial consiste em examinar como as variáveis se agrupam e nomear os fatores, justificando teoricamente como as variáveis se relacionam com os fatores. Devemos examinar todas as variáveis pertencentes a um determinado fator, em especial aquelas com cargas mais altas, e dar um nome para o fator que reflita da maneira mais adequada possível o conjunto de variáveis pertencentes a ele. Caso você não consiga nomear algum fator, isso pode ser um sinal para considerar modelos alternativos, por exemplo, com números de fatores diferentes. Como vimos, existem vários critérios para determinar o número de fatores a extrair, nem sempre concordantes.

Aqui um ponto importante merece ser destacado: mesmo que você esteja fazendo uma análise fatorial exploratória, não espere que ela faça milagres. Você deve sempre se valer de teoria e/ou de pesquisas anteriores. Portanto, mesmo sendo uma técnica exploratória, algum tipo de hipótese sobre o agrupamento das variáveis sempre deve existir. Adicionalmente, assim como em outras técnicas estatísticas, toda variável que você insere vai interferir no resultado do modelo como um todo e também nas outras variáveis. Só devemos inserir no modelo variáveis com algum fundamento teórico ou empírico.

A Tabela 8 sintetiza a segunda etapa de planejamento da análise fatorial exploratória.

Tabela 8 – Síntese da etapa 2: número de fatores, extração, rotação e interpretação

Número de fatores	<p>Critério do autovalor (critério de Kaiser): apenas os fatores que tem autovalor >1.</p> <p>Critério do diagrama de inclinação: a curva da variância individual de cada fator se torna horizontal ou sofre uma queda abrupta (ponto de inflexão da curva).</p> <p>Critério de porcentagem de variância: patamar de 60% da variância acumulada.</p> <p>Os critérios devem ser usados em conjunto.</p>
Método de extração	Componentes principais, fatores principais, máxima verossimilhança, mínimos quadrados ordinários, mínimos quadrados generalizados.
Rotação dos fatores	<p>Rotação ortogonal: <i>varimax</i>, <i>quartimax</i>, <i>BentlerT</i> e <i>geominT</i> (mais usado: <i>varimax</i>).</p> <p>Rotação oblíqua: <i>oblimin</i>, <i>promax</i>, <i>simplimax</i>, <i>BentlerQ</i> e <i>geominQ</i> (mais usado <i>oblimin</i>).</p>
Interpretação dos fatores	Examinar como as variáveis se agrupam e nomear os fatores, justificando teoricamente como as variáveis se relacionam com os fatores.

Fonte: elaboração própria.

Exemplos de pesquisa

Exemplo 1 – Concepções de avaliação

Como primeiro exemplo do livro, selecionamos um questionário educacional desenvolvido na Nova Zelândia (BROWN, 2003, 2006) e adaptado para a realidade brasileira por Matos (2010), intitulado *Students' Conceptions of Assessment (SCoA) version VI* (Concepções de Avaliação de Estudantes). Ele avalia as concepções de alunos sobre a avaliação. O formato de resposta do questionário é uma escala Likert de seis pontos: discordo fortemente, discordo na maior parte, concordo ligeiramente, concordo moderadamente, concordo na maior parte,

concordo fortemente. O instrumento original em inglês possui 33 itens distribuídos em oito fatores⁶.

Nesse sentido, selecionamos alguns itens do questionário que serão nossas variáveis analisadas por meio da análise fatorial:

Item 1- A avaliação é sem valor.

Item 2- Eu desconsidero os meus resultados de avaliação.

Item 3- A avaliação tem um impacto pequeno no meu aprendizado.

Item 4- A avaliação encoraja a minha turma a trabalhar junta e a ajudar uns aos outros.

Item 5- A avaliação me motiva e aos meus colegas a ajudarem uns aos outros.

Item 6- A avaliação faz a nossa turma cooperar mais uns com os outros.

Item 7- Eu presto atenção nos meus resultados de avaliação para me concentrar no que eu posso melhorar da próxima vez.

Item 8- Eu faço uso do *feedback* que recebo para melhorar meu aprendizado.

Item 9- Eu observo o que eu fiz de errado ou de maneira insuficiente para guiar o que eu deveria aprender em seguida.

Item 10- Eu uso as avaliações para assumir responsabilidade para as minhas próximas etapas de aprendizagem.

Item 11- Eu uso as avaliações para identificar o que eu preciso estudar em seguida.

Item 12- A avaliação interfere no meu aprendizado.

A base de dados usada é composta por estudantes de 18 cursos de graduação ($N= 756$; 216 homens e 540 mulheres) de duas IES de Minas Gerais: uma Universidade Federal ($N= 297$) e um Centro Universitário privado ($N= 459$).

⁶ Para maiores detalhes sobre esse instrumento, ver Matos (2010), Matos *et. al.* (2012) e Matos *et. al.* (2013).

Passo 1 – Tamanho da amostra

A amostra é composta por 756 indivíduos. Como estamos analisando 12 variáveis (itens), o mínimo recomendado seriam 60 sujeitos (pelo menos 5 vezes mais observações do que o número de variáveis). Portanto, esse requisito foi satisfeito para o conjunto de dados analisado. O código em R a seguir faz a leitura da base de dados e verifica se o critério do tamanho amostral é satisfeito.

```
# Carrega o pacote para ler o banco de dados
library(foreign)
respostas_instrumento= read.spss("Exemplo 1.sav", to.data.frame=TRUE)
dim(respostas_instrumento)
dim(respostas_instrumento)[1]>5*dim(respostas_instrumento)[2]
```

Apesar disso, ainda quanto ao tamanho da amostra, lembramos que o recomendado para a análise fatorial é usar amostras grandes (pelo menos maior do que 100). Nesse caso, nosso banco de dados também atende ao critério.

Passo 2 – Nível de mensuração

Todas as variáveis analisadas nesse exemplo são categóricas (ordinais). Portanto, para o cálculo das correlações devemos utilizar a correlação policórica. O cálculo das correlações será realizado por meio da função *polychoric()* do pacote *psych*. Para esse cálculo, o R exige que as variáveis estejam em escala numérica. Nesse sentido, primeiramente é necessário converter cada uma das variáveis para um vetor numérico ordenado de acordo com as categorias e em seguida calcular as correlações. A seguir, encontra-se o código que realiza essas duas tarefas. Caso as variáveis categóricas já estejam em formato numérico, não é necessário fazer essa transformação.

```
# Converte cada uma das variaveis para escala numerica e salva em uma nova base
# de dados chama 'respostas_instrumento_numerica'
respostas_instrumento_numerica<-respostas_instrumento
for(i in 1:ncol(respostas_instrumento)){
  levels(respostas_instrumento_numerica[,i])<-1:6
  respostas_instrumento_numerica[,i]<-as.numeric(respostas_instrumento_numerica[,i])
}
```

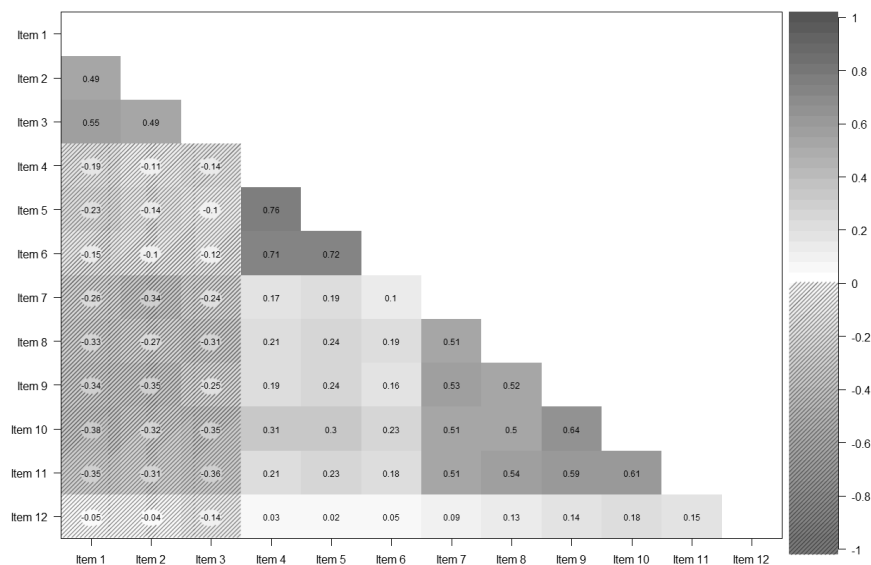
Passo 3 – Matriz de correlações

Após calcularmos as correlações entre as variáveis, vamos verificar se elas satisfazem o seguinte critério: a maioria dos coeficientes de correlação apresenta valores maiores do que 0,3. Podemos verificar a matriz numérica por meio das seguintes linhas de código:

```
# Carrega o pacote
require(psych)
# Calcula a correlacao
correlacao<-polychoric(respostas_instrumento_numerica)
# Imprime as correlacoes com 2 casas decimais
round(correlacao$rho,2)
```

Para facilitar a visualização dessas correlações, o R possui ferramentas gráficas bastante úteis para representá-las. O código a seguir gera o gráfico com essas correlações (Figura 5).

```
corPlot(correlacao$rho,numbers=TRUE,upper=FALSE,diag=FALSE)
```

Figura 5 – Matriz de correlações Exemplo 1⁷

Fonte: elaboração própria.

A Figura 5 mostra que grande parte das correlações está acima de 0,3. Quanto mais forte o cinza, tanto as áreas plana quanto as áreas hachuradas, mais forte é a correlação. A área plana representa correlação positiva, enquanto a área hachurada sinaliza correlação negativa. Apenas o item 12 apresenta uma correlação baixa com todos os demais. Esse é um primeiro sinal de alerta: é um indicativo de que ele pode apresentar problemas no ajuste do modelo, uma vez que já sabemos que para a análise fatorial é muito importante que as variáveis estejam correlacionadas.

Passo 4 – Teste de Bartlett (BTS)

Vamos agora testar se a matriz de correlações é significativamente diferente da matriz identidade. As linhas de código a seguir realizam o teste e mostram seu resultado.

⁷ O gráfico apresentado no livro está cinza, mas o gráfico gerado pelo R apresenta originalmente 2 cores: azul e vermelho. No gráfico colorido, quanto mais forte a cor, tanto azul quanto vermelho, mais forte é a correlação. O azul representa correlação positiva, enquanto o vermelho sinaliza correlação negativa.

```
# Teste bartlett
cortest.bartlett(correlacao$rho,n=nrow(respostas_instrumento))
$chisq
[1] 3834.734
$p.value
[1] 0
$df
[1] 66
```

O resultado desse teste deve ser estatisticamente significativo: $p < 0,05$. No nosso exemplo, o p-valor do teste ($\$p.value$) foi aproximadamente zero. Isso significa que a hipótese nula deve ser rejeitada com uma significância de 5% e que a matriz de correlações é diferente da identidade. Assim, mais um critério de adequação da base de dados foi atendido.

Passo 5 – Teste de Kaiser-Meyer-Olkin (KMO)

Passamos agora para o último passo da verificação da adequação da base de dados: o cálculo do teste KMO. Ele apresenta um valor global, para todas as variáveis (no nosso exemplo, os itens do questionário), e um valor individual para cada variável. As linhas de código a seguir realizam o teste e mostram seus resultados.

```
# KMO
KMO(correlacao$rho)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = correlacao$rho)
Overall MSA = 0.85
MSA for each item =
```

Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
0.84	0.83	0.78	0.78	0.78	0.80	0.89	0.92	0.87
Item 10	Item 11	Item 12						
0.89	0.90	0.81						

Os resultados mostram que tanto o KMO global (Overall MSA = 0.85) quanto o KMO de cada um dos itens (MSA for each item) foram bem altos, superiores a 0,8 em sua maioria. Como apontado anteriormente, para o KMO, quanto mais próximo de 1 melhor: 0,5 é o valor mínimo

aceitável, mas o ideal seria um valor a partir de 0,7. Como nosso KMO global foi de 0,85, esse critério também foi atendido.

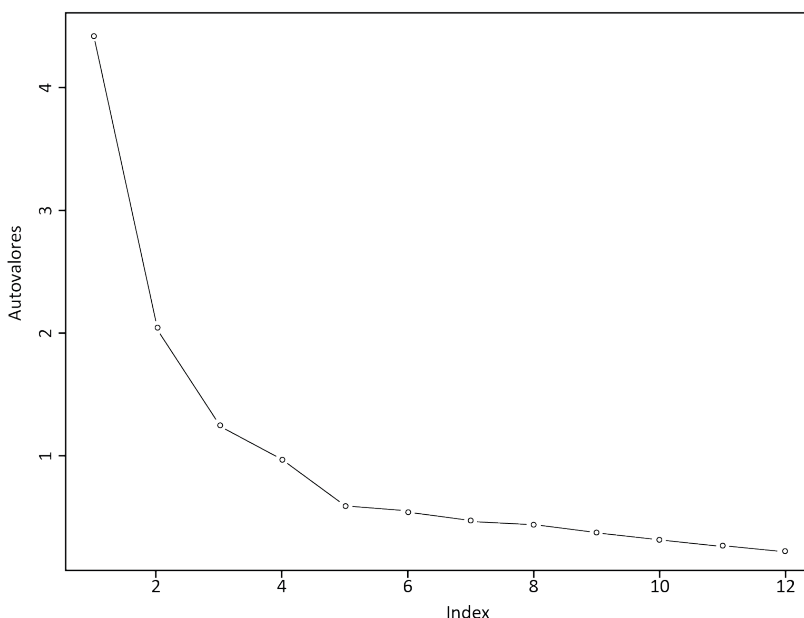
Portanto, todos os critérios de adequação da base de dados foram atendidos e a Etapa 1 da análise fatorial (verificação da adequação da base de dados) está concluída.

Passo 6 – Determinação do número de fatores

Aqui iniciamos a etapa 2 da análise fatorial, que envolve mais uma série de passos: número de fatores, extração, rotação e interpretação.

Como explicado anteriormente, a definição do número de fatores que serão extraídos é uma tarefa complexa. Não existe um único critério consensual para determinar o número de fatores. O código a seguir calcula os principais métodos indicados na literatura: critério do autovalor (*eigenvalue*), critério do diagrama de inclinação (*Scree test*), critério da porcentagem de variância acumulada.

```
# Calculo dos autovalores
round(eigen(correlacao$rho)$values,2)
[1] 4.44 2.05 1.25 0.98 0.60 0.55 0.48 0.44 0.38 0.32 0.27 0.23
# Numero de autovalores maiores que 1
sum(eigen(correlacao$rho)$values>1)
[1] 3
# Scree-plot
plot(eigen(correlacao$rho)$values, type = "b",ylab='Autovalores')
# Calculo da proporcao explicada por cada fator
proporcao_explicacao<-eigen(correlacao$rho)$values/
sum(eigen(correlacao$rho)$values)
# Calculo da proporcao de explicacao acumulada
proporcao_acumulada <- cumsum(proporcao_explicacao)
# Imprime a proporcao acumulada com 2 casas decimais
round(proporcao_acumulada,2)
[1] 0.37 0.54 0.65 0.73 0.78 0.82 0.86 0.90 0.93 0.96 0.98 1.00
```

Figura 6 – Scree-plot para o Exemplo 1

Fonte: elaboração própria.

Segundo o critério de Kaiser (autovalor maior do que 1), devemos reter 3 fatores. De acordo com o critério da variância acumulada (patamar de 60% da variância acumulada), também retemos 3 fatores, pois eles são capazes de explicar 65% da variabilidade dos dados. Portanto, o poder de explicação do modelo está adequado. Finalmente, pelo diagrama de declividade (*Scree test*), esse número poderia ser de até 5 fatores (ponto de inflexão: a curva da variância individual de cada fator se torna horizontal ou então sofre uma queda abrupta) (Figura 6). Isso acontece porque, como regra geral, o teste *scree* resulta em pelo menos 1, e às vezes 2 ou 3 fatores a mais, em relação ao critério de Kaiser (autovalor >1) (HAIR *et al.*, 2005). No nosso exemplo, pelos resultados encontrados, decidimos reter 3 fatores. Adicionalmente, sobre a variância explicada, é importante destacar que, nesse passo 6, ela é calculada como a razão dos autovalores sobre a soma de todos os autovalores. Nesse sentido, após

o ajuste do modelo (ver passos seguintes), a porcentagem de variância explicada pode variar um pouco.

Como já apontado anteriormente, na prática não costumamos usar um único critério. Eles são usados conjuntamente e nem sempre todos eles são concordantes. Portanto, para decidir sobre o número de fatores a extrair, a parcimônia é importante: a solução ótima é encontrar o número mínimo de fatores que maximiza a quantidade de variância total explicada. Nesse sentido, volte no seu referencial teórico tendo sempre em mente: teoricamente faz mais sentido essas variáveis estarem agrupadas em quantos fatores? Isso porque é um fenômeno comum encontrarmos mais de uma solução empírica aceitável, do ponto de vista estatístico, quanto ao número de fatores. Assim, a decisão final pode acabar sendo teórica. Lembre sempre que as variáveis de um fator devem medir um mesmo construto latente e precisam estar associadas entre si, inclusive teoricamente.

Passo 7 – Extração das cargas fatoriais

Após a definição do número de fatores, o passo seguinte é decidir qual técnica será utilizada para o cálculo das cargas fatoriais (extração dos fatores). Como apontado previamente, existem várias técnicas para a extração dos fatores, dependendo do tipo de dado que está sendo analisado e do objetivo da análise. Para o nosso exemplo, vamos usar a função *fa()* do pacote *psych* e o método da máxima verossimilhança, que na função é definido como 'ml'.

```
## Extracao dos 3 fatores sem rotacao e usando o metodo ml
fa_sem_rotacao <- fa(correlacao$rho,3,rotate="none", fm="ml")
# Mostra as cargas fatoriais
loadings(fa_sem_rotacao)
```

Loadings:

	ML1	ML2	ML3
Item 1	-0.480	0.333	0.432
Item 2	-0.402	0.368	0.357
Item 3	-0.413	0.386	0.526
Item 4	0.720	0.474	
Item 5	0.748	0.459	

Item 6	0.666	0.490	
Item 7	0.501	-0.411	0.204
Item 8	0.548	-0.372	0.140
Item 9	0.593	-0.466	0.272
Item 10	0.660	-0.396	0.170
Item 11	0.592	-0.450	0.166
Item 12	0.133	-0.129	
	ML1	ML2	ML3
SS loadings	3.793	1.972	0.786
Proportion Var	0.316	0.164	0.066
Cumulative Var	0.316	0.480	0.546

Na saída do R, primeiramente temos três colunas representando os 3 fatores (ML1, ML2 e ML3). Em cada coluna, temos as cargas fatoriais (*loadings*) de todos os itens nos 3 fatores. Para facilitar a visualização dos resultados, o R omite todas as cargas menores que 0,1. Abaixo, a linha “*SS loadings*” apresenta a soma das cargas fatoriais ao quadrado. Já as linhas “*Proportion Var*” e “*Cumulative Var*” representam, respectivamente, a proporção de variância explicada por cada fator e a variância acumulada.

Adicionalmente, é importante mencionar que as cargas fatoriais não rotacionadas não são apresentadas nos resultados de pesquisa. Os fatores serão rotacionados no passo seguinte, e só então essas cargas e seus respectivos fatores serão interpretados.

Passo 8 – Rotação dos fatores

Após a extração dos fatores, a técnica de rotação é utilizada para atingir uma melhor distinção entre eles. A rotação maximiza cargas altas entre os fatores e as variáveis e minimiza as cargas baixas. A solução rotacionada é usada para relatar os resultados finais da pesquisa.

Como indicado anteriormente, existem dois tipos de rotação: ortogonal (os fatores não são correlacionados) e oblíqua (os fatores são correlacionados). Primeiramente, ilustramos uma rotação ortogonal com o método *varimax* (um dos mais usados nesse tipo de rotação).

Rotação ortogonal

```
fa_com_rotacao_varimax <- fa(correlacao$rho,3,rotate="varimax", fm="ml")
loadings(fa_com_rotacao_varimax)
```

Loadings:

	ML2	ML1	ML3
Item 1	-0.263	-0.128	0.666
Item 2	-0.277		0.588
Item 3	-0.206		0.743
Item 4	0.129	0.849	
Item 5	0.170	0.858	
Item 6		0.821	
Item 7	0.656		-0.161
Item 8	0.626	0.139	-0.217
Item 9	0.780	0.103	-0.157
Item 10	0.724	0.202	-0.239
Item 11	0.713	0.116	-0.240
Item 12	0.177		
SS loadings	ML2	ML1	ML3
Proportion Var	2.735	2.243	1.574
Cumulative Var	0.228	0.187	0.131
	0.228	0.415	0.546

Para usar rotação ortogonal, o pesquisador precisa ter evidências teóricas ou empíricas muito fortes de que os fatores não são correlacionados. Esse não é o caso do nosso exemplo, onde sabemos que as concepções de avaliação de estudantes estão correlacionadas. Em Ciências Humanas e Sociais, as variáveis quase sempre estão associadas. Portanto, incluímos esse exemplo de rotação ortogonal apenas por se tratar de um livro didático. Em seguida, analisamos mais detalhadamente um exemplo de rotação oblíqua.

Rotação oblíqua

Agora, ilustramos uma rotação oblíqua com o método *oblimin* (um dos mais usados nesse tipo de rotação).

```
## Extracao dos 3 fatores com rotacao e usando o metodo ml
fa_com_rotacao <- fa(correlacao$rho,3,rotate="oblimin", fm="ml")
# Mostra as cargas fatoriais
loadings(fa_com_rotacao)
```

Loadings:			
	ML2	ML1	ML3
Item 1			0.690
Item 2			0.599
Item 3			0.803
Item 4		0.861	
Item 5		0.865	
Item 6		0.841	
Item 7	0.699		
Item 8	0.635		
Item 9	0.844		
Item 10	0.732		
Item 11	0.731		
Item 12	0.182		
	ML2	ML1	ML3
SS loadings	2.725	2.211	1.493
Proportion Var	0.227	0.184	0.124
Cumulative Var	0.227	0.411	0.536

Na saída do R, temos as cargas fatoriais (*loadings*) dos itens nos 3 fatores. Para facilitar a visualização dos resultados, o R omite todas as cargas menores que 0,1. Assim, conseguimos identificar claramente quais itens pertencem a cada um dos fatores por meio das cargas fatoriais mais altas: ML3 (itens 1, 2 e 3), ML1 (itens 4, 5 e 6) e ML2 (itens 7, 8, 9, 10, 11). Observamos que, após a rotação, as cargas se distribuem bem melhor entre os fatores (compare com os resultados do passo 7 – extração das cargas fatoriais).

Apenas o item 12 apresenta uma carga fatorial baixa em todos os fatores (indeterminação fatorial: não conseguimos identificar a qual fator a variável pertence⁸). Por isso, vamos verificar também as comunalidades de cada um dos itens.

```
# Mostra as comunalidades
round(fa_com_rotacao$communalities,3)
```

Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
0.529	0.425	0.597	0.744	0.771	0.686	0.462	0.458	0.643
Item 10	Item 11	Item 12						
0.622	0.580	0.036						

⁸ A indeterminação fatorial também pode acontecer com variáveis que apresentam cargas altas em mais de um fator.

Notamos que o item 12, além de apresentar cargas fatoriais baixas em todos os fatores, ainda apresenta um valor de comunalidade muito reduzido. Portanto, baseado nesses dois critérios (indeterminação fatorial e comunalidade baixa), o item 12 será excluído e o modelo reajustado. Vale ainda destacar que, desde o passo 3 da etapa 1 (matriz de correlações), já tínhamos sinais de que esse item seria problemático, pois ele apresentou uma correlação baixa com todos os outros itens.

As cargas foram extraídas novamente e os resultados após a rotação oblíqua são apresentados a seguir. Percebemos que agora todos os itens se distribuem bem entre os fatores, mantendo o padrão anteriormente identificado: ML3 (itens 1, 2 e 3), ML1 (itens 4, 5 e 6) e ML2 (itens 7, 8, 9, 10, 11). O poder de explicação do modelo também aumentou para 58,1% (isso evidencia que o item 12 estava gerando ruído no modelo).

Além disso, após a exclusão de variáveis problemáticas, é muito comum que os valores das cargas fatoriais melhorem (aumentem). No nosso exemplo, isso foi pouco visível, provavelmente por termos excluído apenas uma variável. Outra questão relevante é que o critério da comunalidade maior do que 0,5 não deve ser utilizado isoladamente e de maneira muito rígida. Em outras palavras, temos variáveis no modelo com um valor um pouco abaixo desse critério, mas que apresentam cargas fatoriais altas (exemplo: item 8). Assim, o pesquisador deve sempre usar mais de um critério.

Loadings:			
	ML2	ML1	ML3
Item 1			0.691
Item 2			0.599
Item 3			0.802
Item 4		0.861	
Item 5		0.864	
Item 6		0.842	
Item 7	0.702		
Item 8	0.635		
Item 9	0.844		
Item 10	0.729		
Item 11	0.731		
	ML2	ML1	ML3
SS loadings	2.691	2.210	1.492
Proportion Var	0.245	0.201	0.136
Cumulative Var	0.245	0.446	0.581

Ainda sobre a exclusão de variáveis do modelo: dependendo do seu banco de dados, pode ser necessária a exclusão de muitas variáveis. Uma recomendação importante é ir excluindo as variáveis aos poucos, nas seguintes etapas: a) use principalmente os dois critérios de indeterminação fatorial e comunalidade baixa para identificar itens problemáticos; b) selecione as variáveis que você considera as mais problemáticas de acordo com os critérios anteriores; c) exclua essas variáveis e ajuste o modelo novamente; d) caso persistam variáveis problemáticas, repetir os passos anteriores.

Por que isso é importante? Como outras análises estatísticas, a análise fatorial é um procedimento de modelagem dos dados. Assim, qualquer variável incluída ou não no modelo interfere no resultado de todas as outras. E, como acabamos de dizer, após a exclusão de variáveis problemáticas, é bastante comum que os valores das cargas fatoriais das variáveis que permaneceram no modelo aumentem.

Síntese dos resultados

O R oferece a possibilidade de apresentar todos os resultados do ajuste do modelo em uma espécie de resumo das análises. Isso pode ser feito usando a função *print()*, como mostrado a seguir:

```
print(fa_com_rotacao)
> print(fa_com_rotacao)
Factor Analysis using method = ml
Call: fa(r = correlacao$rho, nfactors = 3, rotate = "oblimin", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
```

	ML2	ML1	ML3	h2	u2	com
Item 1	-0.04	-0.06	0.69	0.53	0.47	1.0
Item 2	-0.10	0.03	0.60	0.43	0.57	1.1
Item 3	0.05	0.02	0.80	0.59	0.41	1.0
Item 4	0.00	0.86	-0.01	0.74	0.26	1.0
Item 5	0.05	0.86	0.02	0.77	0.23	1.0
Item 6	-0.05	0.84	-0.01	0.69	0.31	1.0
Item 7	0.70	-0.04	0.02	0.46	0.54	1.0
Item 8	0.63	0.03	-0.05	0.46	0.54	1.0
Item 9	0.84	-0.03	0.06	0.64	0.36	1.0
Item 10	0.73	0.08	-0.05	0.62	0.38	1.0
Item 11	0.73	-0.01	-0.06	0.58	0.42	1.0

Aqui, já sabemos que as colunas ML2, ML1 e ML3 representam as cargas fatoriais padronizadas de todas as variáveis em todos os fatores. Já as colunas h2, u2 e com representam, respectivamente, a variância compartilhada, variância específica e a soma das duas.

	ML2	ML1	ML3
SS loadings	2.75	2.22	1.5
Proportion Var	0.25	0.20	0.14
Cumulative Var	0.25	0.45	0.59
Proportion Explained	0.42	0.34	0.24
Cumulative Proportion	0.42	0.76	1.00

With factor correlations of			
	ML2	ML1	ML3
ML2	1.00	0.32	-0.56
ML1	0.32	1.00	-0.22
ML3	-0.56	-0.22	1.00

Como usamos uma rotação oblíqua, o R gera uma matriz de correlação entre os fatores, que deve ser interpretada como qualquer matriz de correlação. Vamos realizar a interpretação dessa matriz no passo 9 – interpretação dos fatores.

Mean item complexity = 1			
Test of the hypothesis that 3 factors are sufficient.			
The degrees of freedom for the null model are 55 and the objective function was 5.06			
The degrees of freedom for the model are 25 and the objective function was 0.13			
The root mean square of the residuals (RMSR) is 0.02			
The df corrected root mean square of the residuals is 0.03			
Fit based upon off diagonal values = 1			
Measures of factor score adequacy			
	ML2	ML1	ML3
Correlation of (regression) scores with factors	0.93	0.95	0.89
Multiple R square of scores with factors	0.87	0.90	0.79
Minimum correlation of possible factor scores	0.74	0.79	0.57

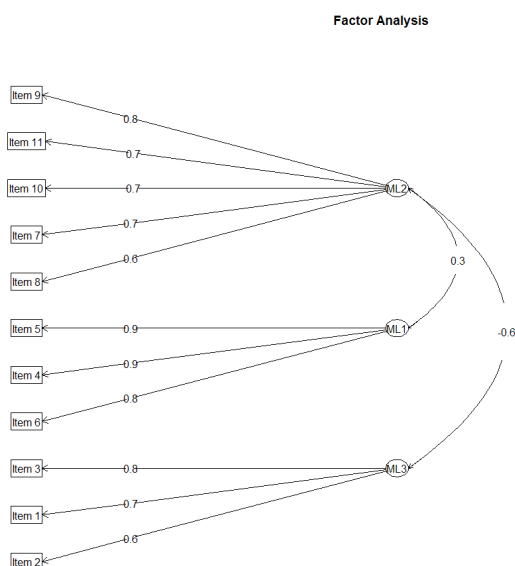
Outras ferramentas do R para a análise fatorial

O R apresenta ainda algumas ferramentas adicionais de visualização e apresentação das cargas fatoriais. É possível representar os fatores em um diagrama usando o seguinte comando:

```
# Diagrama dos fatores
fa.diagram(fa_com_rotacao)
```

O resultado é apresentado na Figura 7. As setas duplas entre os fatores representam as correlações entre eles. Por padrão, o R omite as correlações abaixo de 0,3. Isso aconteceu com a correlação entre os fatores 1 e 3, que foi 0,22. Já as setas entre o fator e as variáveis indicam as cargas fatoriais.

Figura 7 – Representação dos fatores do Exemplo 1

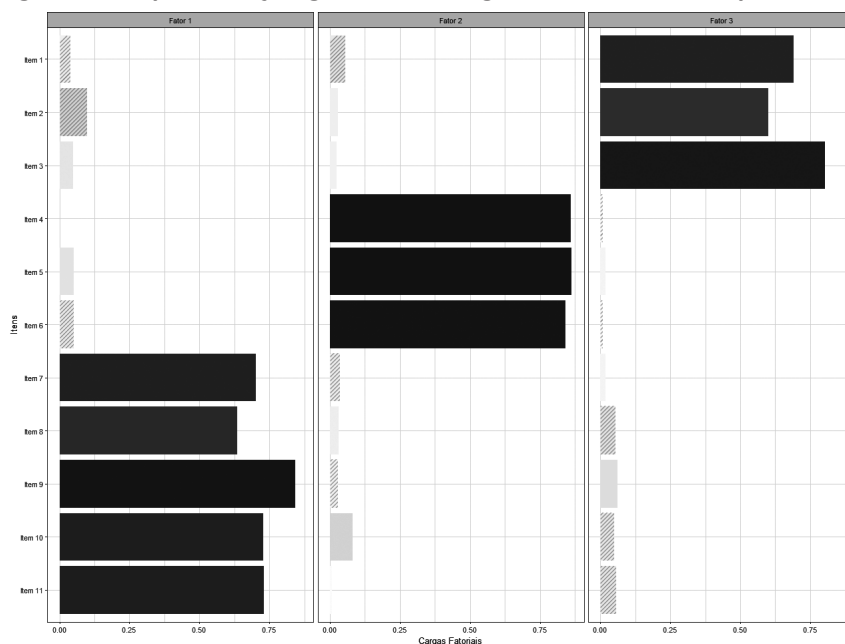


Fonte: elaboração própria.

Um outro tipo de gráfico, criado pelo usuário Dan Mirman (https://rpubs.com/danmirman/plotting_factor_analysis), apresenta uma

representação dos fatores a partir de gráficos de barras. O código para criar esse gráfico, apresentado na Figura 8, se encontra a seguir.

```
## Gráfico representacao cargas fatoriais
# Salva as cargas em uma tabela
cargas<-data.frame(loadings(fa_oblimin)[,c(1,2,3)])
# Cria uma coluna com o nomes dos itens
cargas$Itens<-row.names(cargas)
# Reordena os itens
Ord <- 11:1
cargas$Itens <- reorder(cargas$Itens, Ord)
# Muda o nomes das colunas
colnames(cargas)[1:3]<-c("Fator 1","Fator 2","Fator 3")
# Muda o formato dos dados
# Carrega o pacote
require(reshape2)
loadings.m <- melt(cargas, id="Itens",
  measure=c("Fator 1","Fator 2","Fator 3"),
  variable.name="Fator", value.name="Cargas")
# Carrega o pacotes
require(ggplot2)
# Plota o gráfico
ggplot(loadings.m, aes(Itens, abs(Cargas), fill=Cargas)) +
facet_wrap(~ Fator, nrow=1) + #coloca os fatores em caixas distintas
geom_bar(stat="identity") + #faz as barras
coord_flip() + #inverte os eixos
#define as cores
scale_fill_gradient2(name = "Cargas",
  high = "blue", mid = "white", low = "red",
  midpoint=0, guide=F) +
ylab("Cargas Fatoriais") + #muda o o nome do eixo
theme_bw(base_size=10) #muda o tamanho da fonte
```

Figura 8 – Representação gráfica das cargas fatoriais do Exemplo 1⁹

Fonte: elaboração própria.

Passo 9 – Interpretação dos fatores

Como abordado anteriormente, a análise fatorial é uma técnica puramente empírica. No entanto, os fatores precisam fazer sentido do ponto de vista teórico. Assim, essa é a fase final da análise fatorial: examinar como as variáveis se agrupam e nomear os fatores, justificando teoricamente. Devemos dar um nome para o fator que represente da melhor maneira possível o conjunto de variáveis pertencentes a ele.

No nosso exemplo 1, obtivemos os seguintes resultados: fator 1 (itens 4, 5 e 6), fator 2 (itens 7, 8, 9, 10, 11) e fator 3 (itens 1, 2 e 3). A seguir, nomeamos e explicamos teoricamente cada um dos fatores. Lembrando

⁹ O gráfico apresentado no livro está preto e branco, mas o gráfico gerado pelo R originalmente é colorido. Inclusive o código apresentado anteriormente para criar esse gráfico faz menção a essas cores: `scale_fill_gradient2(name = "Cargas", high = "blue", mid = "white", low = "red", midpoint=0, guide=F) +`

que, no caso do exemplo 1, estamos analisando as diferentes concepções de alunos sobre a avaliação. Nossos fatores irão abordar algumas dessas concepções descritas na literatura educacional.

Fator 1 – Impacto positivo na sala de aula

Item 4- A avaliação encoraja a minha turma a trabalhar junta e a ajudar uns aos outros

Item 5- A avaliação me motiva e aos meus colegas a ajudarem uns aos outros

Item 6- A avaliação faz a nossa turma cooperar mais uns com os outros

O fator 1 representa uma concepção de que a avaliação tem um impacto emocionalmente positivo nos alunos. Ele agrega itens que estão voltados para um impacto positivo na sala de aula como um todo. A concepção de que a avaliação possui um impacto emocionalmente positivo nos estudantes enfatiza as respostas emocionais deles com relação a formatos diferentes de avaliação (BROWN *et al.*, 2009). Respostas emocionais positivas, como a motivação dos colegas de sala de aula, estão incluídas nessa concepção. Pesquisas indicam que alguns estudantes relatam respostas emocionais positivas com relação à avaliação, embora essa concepção seja menos comum do que as outras (MATOS, 2010).

Fator 2 – Melhora da aprendizagem do aluno

Item 7- Eu presto atenção nos meus resultados de avaliação para me concentrar no que eu posso melhorar da próxima vez

Item 8- Eu faço uso do *feedback* que recebo para melhorar meu aprendizado

Item 9- Eu observo o que eu fiz de errado ou de maneira insuficiente para guiar o que eu deveria aprender em seguida

Item 10- Eu uso as avaliações para assumir responsabilidade para as minhas próximas etapas de aprendizagem

Item 11- Eu uso as avaliações para identificar o que eu preciso estudar em seguida

O fator 2 representa uma concepção de que a avaliação melhora o ensino e a aprendizagem. Ele agrega itens que estão voltados para o uso da avaliação por parte do aluno para a melhora do processo ensino e aprendizagem. A concepção de melhora enfatiza a expectativa dos estudantes de que a avaliação leva a uma aprendizagem melhor, mais eficaz (HARRIS; BROWN; HARNETT, 2009). Essa é uma concepção associada com uma função pedagógica e formativa da avaliação (MATOS, 2010).

Fator 3 - Irrelevância da avaliação

Item 1- A avaliação é sem valor

Item 2- Eu desconsidero os meus resultados de avaliação

Item 3- A avaliação tem um impacto pequeno no meu aprendizado

O fator 3 representa uma concepção de que a avaliação é irrelevante e os alunos respondem negativamente a ela. Ele agrega itens que estão voltados para considerar a avaliação como algo a ser ignorado ou como algo ruim. A concepção de irrelevância enfatiza a percepção de que a avaliação é negativa, ruim ou injusta. Os estudantes diferem com relação ao grau de percepção negativa da avaliação. Assim, os alunos se mostram sensíveis quanto à avaliação que eles percebem como injusta, ruim ou irrelevante para eles (MATOS, 2010).

Nesse último passo, vale lembrar mais uma vez: mesmo em uma análise fatorial exploratória, devemos sempre nos valer de teoria e/ou de pesquisas anteriores. Assim, sempre temos algum tipo de hipótese sobre o agrupamento das variáveis. Adicionalmente, só devemos inserir no modelo variáveis com algum fundamento teórico ou empírico, pois toda variável interfere no resultado do modelo como um todo e nas outras variáveis. Portanto, se você tiver dificuldade em nomear os fatores, isso pode ser um sinal para considerar modelos alternativos, por exemplo, com números de fatores diferentes. Como vimos, existem vários critérios para determinar o número de fatores a extrair, nem sempre concordantes.

Em outras palavras, se você encontrar fatores que não façam sentido teórico algum, os resultados não serão uteis. Nesse caso, pode ser necessário até mesmo repensar e refazer a pesquisa. Apenas a título de ilustração, no exemplo 1, seria muito estranho encontrar um fator com as seguintes variáveis:

Item 2 - Eu desconsidero os meus resultados de avaliação

Item 3 - A avaliação tem um impacto pequeno no meu aprendizado

Item 8 - Eu faço uso do *feedback* que recebo para melhorar meu aprendizado

Item 11 - Eu uso as avaliações para identificar o que eu preciso estudar em seguida

Isso porque as variáveis estão associadas a concepções diferentes de avaliação presentes na literatura. Dessa forma, seria impossível dar um nome para esse fator que representasse o conjunto de variáveis pertencentes a ele.

Finalmente, podemos analisar também as correlações entre os fatores, que já foram calculadas em passos anteriores.

Matriz de correlação entre os fatores

	ML2	ML1	ML3
ML2	1.00		
ML1	0.32	1.00	
ML3	-0.56	-0.22	1.00

O maior valor foi a correlação negativa entre o fator 2 (Melhora da aprendizagem do aluno) e o fator 3 (Irrelevância da avaliação) ($r = -0.56$). O fato de o sinal ser negativo faz sentido teoricamente, pois o fator 2 representa uma concepção de que a avaliação melhora o ensino e a aprendizagem e o fator 3 uma concepção de que a avaliação é irrelevante e os alunos respondem negativamente a ela. Ou seja, são concepções opostas. A correlação entre o fator 1 (Impacto positivo na sala de aula) e o fator 3 (Irrelevância da avaliação) também é negativa ($r = -0.22$). A interpretação é similar a anterior: são concepções opostas, sendo que o fator 1 representa uma concepção de que a avaliação tem um impacto emocionalmente positivo nos alunos. Já a correlação entre o fator 1

(Impacto positivo na sala de aula) e o fator 2 (Melhora da aprendizagem do aluno) é positiva ($r = 0.32$), o que mais uma vez faz todo sentido do ponto de vista teórico: esperamos que um impacto emocionalmente positivo nos alunos esteja associado positivamente com a melhora do ensino e aprendizagem. Tomados em conjunto, esses resultados reforçam a validade dos nossos dados. Vale ainda destacar que as correlações encontradas entre os fatores evidenciam que a escolha por uma rotação oblíqua foi acertada.

Exemplo 2 – Uso de escores fatoriais em um modelo de regressão

No exemplo 1, explicamos passo a passo todas as etapas da análise fatorial. Nesse sentido, o exemplo 2 cumpre apenas um objetivo: demonstrar como a análise fatorial pode ser usada em combinação com outras técnicas. Aqui, faremos a combinação com uma regressão multinível.

Como explicado ao longo do livro, os escores fatoriais podem ser utilizados de diversas maneiras. Uma delas seria em modelos de regressão, atuando como variáveis explicativas ou como variável resposta (no nosso exemplo, usamos como variável explicativa). O objetivo é ajustar um modelo multinível para explicar a proficiência em Língua Portuguesa dos alunos do quinto ano que fizeram a Prova Brasil em 2015. Quando fazem essa avaliação, os alunos respondem ainda a um questionário contextual que contém várias perguntas sobre o seu perfil. Dentre essas questões, algumas dizem respeito ao seu nível socioeconômico. Alves, Soares e Xavier (2013) utilizaram um modelo de Teoria de Resposta ao Item (TRI) para criar um indicador de nível socioeconômico a partir desses itens. Nesse exemplo, utilizamos um modelo de análise fatorial para criar esse indicador, visto que ele é um traço latente que está sendo representado por uma série de perguntas respondidas pelos alunos¹⁰. A Tabela 9 apresenta os itens utilizados nesta análise.

¹⁰ Tanto a TRI quanto a análise fatorial são técnicas que lidam com traços latentes.

Tabela 9 – Itens referentes ao nível socioeconômico dos alunos

Item	Pergunta
1	Na sua casa tem televisão em cores?
2	Na sua casa tem aparelho de rádio?
3	Na sua casa tem videocassete e/ou DVD?
4	Na sua casa tem geladeira?
5	Na sua casa tem freezer (parte da geladeira duplex)?
6	Na sua casa tem freezer separado da geladeira?
7	Na sua casa tem máquina de lavar roupa (o tanquinho NÃO deve ser considerado)?
8	Na sua casa tem carro?
9	Na sua casa tem computador?
10	Na sua casa tem banheiro?
11	Em sua casa trabalha empregado(a) doméstico(a) pelo menos cinco dias por semana?
12	Até que série sua mãe, ou a mulher responsável por você, estudou?
13	Até que série seu pai, ou o homem responsável por você, estudou?

Nesse exemplo 2, os passos são os mesmos do exemplo 1. A amostra é composta por 2.497.431 alunos e temos 13 variáveis (itens). Todas as variáveis analisadas nesse caso são categóricas. Portanto, para o cálculo das correlações devemos usar correlação policórica. Na etapa da determinação do número de fatores, os resultados mostram que 1 fator parece ser adequado para esse caso (apenas 1 autovalor maior que 1 e explicação de 66% da variabilidade).

As cargas fatoriais encontradas foram:

Loadings:

	ML1
Item1	0.791
Item2	0.749
Item3	0.778
Item4	0.846
Item5	0.814
Item6	0.790
Item7	0.829
Item8	0.814
Item9	0.820
Item10	0.821
Item11	0.839
Item12	0.723
Item13	0.697

Nesse caso, é interessante destacar que apenas um fator já é suficiente para explicar a variabilidade dos dados. Nesse sentido, a rotação não precisa ser executada quando encontramos apenas 1 fator. O fator encontrado foi denominado como nível socioeconômico.

Utilização dos escores fatoriais

Como discutido previamente, é possível calcular o valor do fator para cada um dos sujeitos da base de dados. Esses valores são os escores fatoriais. Eles são uma espécie de média ponderada das variáveis observadas em cada uma das unidades amostrais, onde os pesos são dados pelas cargas fatoriais. Após a extração dos fatores e o cálculo dos escores fatoriais, estes podem ser utilizados como substitutos das variáveis originais nas análises. Assim, existem possibilidades ricas de combinar a análise fatorial com outras técnicas estatísticas, inclusive análises multivariadas.

Nesse exemplo 2, feito o ajuste do modelo e verificada sua adequabilidade, vamos agora utilizar os resultados obtidos em uma regressão multinível, que é a mais adequada para dados educacionais por causa da sua estrutura hierárquica (alunos agrupados em escolas).

Os comandos a seguir calculam os escores fatoriais e os guardam em um vetor denominado NSE. O método utilizado aqui para o cálculo dos escores foi o método da regressão.

```
# Calculo dos escores fatoriais
my.scores <- factor.scores(respostas_numericas,fa_sem_rotacao,
method="regression")
dados_2015$NSE<-my.scores$scores[,1]
```

Agora que os escores fatoriais já estão salvos na base de dados, em uma variável denominada NSE, podemos realizar o ajuste do modelo multinível. Além do NSE, será utilizada aqui também a variável sexo do aluno (TX_RESP_Q001). As linhas de código a seguir recodificam essa variável para que seja utilizada no modelo. Isso foi necessário porque, na base disponibilizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), as variáveis não apresentam os rótulos, que aparecem em um arquivo separado, intitulado dicionário de variáveis.

```
# Recodifica a variavel sexo
dados_2015$TX_RESP_Q001<-factor(dados_2015$TX_RESP_Q001)
levels(dados_2015$TX_RESP_Q001)<-c(NA,"Masculino","Feminino")
```

O pacote utilizado para o ajuste do modelo multinível será o *lme4* e a função utilizada é a *lmer()*. Calculamos um modelo de 2 níveis (alunos agrupados em escolas). Não entraremos em detalhes sobre esse modelo, visto que esse não é o foco deste livro. Nossa intenção aqui é apenas ilustrar a combinação da análise fatorial com outras técnicas.

```
# Carrega o pacote
require(lme4)
modelo_hierarquico <- lmer(PROFICIENCIA_LP ~ TX_RESP_Q001+NSE+
                           (1 | ID_ESCOLA), data=dados_2015)
summary(modelo_hierarquico)
```

A seguir, apresentamos os resultados dos efeitos fixos do modelo ajustado.

Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	-0.949538	0.002296	-413.5
TX_RESP_Q001Feminino	0.188618	0.001083	174.2
NSE	0.005467	0.001023	5.3

O NSE apresenta um efeito positivo sobre a proficiência dos alunos. Além disso, as meninas apresentam, em média, resultados melhores do que os meninos.

Por fim, algumas vantagens e características da análise fatorial merecem ser novamente destacadas:

- Mesmo que o fator seja criado a partir de variáveis categóricas, ele sempre será analisado como uma variável contínua. O motivo acaba de ser ilustrado, pois utilizamos os escores fatoriais em vez das variáveis originais. Nesse sentido, o fator deve sempre ser interpretado como uma variável contínua. Por exemplo: em uma análise hipotética, um fator é colocado como variável resposta em uma regressão. Nesse caso, devemos fazer uma regressão linear.

- O objetivo da redução dos dados foi atingido no exemplo 2 por meio da análise fatorial. Em vez de incluir 13 variáveis no modelo de regressão, incluímos apenas uma: o fator nível socioeconômico. E o mais importante: os fatores mantêm a representatividade das variáveis originais. Dessa forma, os itens contribuem de maneira desigual para o fator: quanto maior a carga fatorial, maior a contribuição do item para o fator. No exemplo 2, o Item4 (carga fatorial 0.846) apresenta a maior contribuição para o fator, enquanto o Item13 (carga fatorial 0.697) apresenta a menor contribuição. Esse é um aspecto muito rico da análise fatorial: manter a representatividade das variáveis originais. Isso não acontece em outras técnicas mais simples de elaboração de índices.

REFERÊNCIAS BIBLIOGRÁFICAS

- ALVES, Maria Teresa Gonzaga; SOARES, José Francisco; XAVIER, Flavia Pereira. O nível socioeconômico das escolas de educação básica brasileiras. IN: REUNIÃO DA ASSOCIAÇÃO BRASILEIRA DE AVALIAÇÃO EDUCACIONAL - ABAVE, 2013, Brasília. p. 15-32.
- BABBIE, E. R. *Métodos de pesquisas de survey*. Belo Horizonte: Ed. UFMG, 1999. 519 p.
- BROWN, G. T. L. *Secondary school students' conceptions of assessment: a survey of four schools*. Conceptions of Assessment and Feedback Project Report #5. Auckland, NZ: University of Auckland, 2006.
- Brown, G. T. L. *Students' conceptions of assessment (SCoA) inventory* (Versions 1-6). Unpublished test. Auckland, NZ: University of Auckland, 2003.
- BROWN, G. T. L. *et al.* Use of interactive-informal assessment practices: New Zealand secondary students' conceptions of assessment. *Learning & Instruction*, v. 19, n. 2, p. 97-111, 2009.
- BRYANT, F. b.; YARNOLD, P. R. Principal-components analysis and exploratory and confirmatory factor analysis. IN: GRIMM, L. G.; YARNOLD, P. R. (Eds.). *Reading and understanding multivariate statistics*. Washington, DC: APA, 2000. cap. 4, p. 99-136.
- CATTELL, R. B. The scree test for the number of factors. *Multivariate Behavioral Research*. v. 1, p. 245-76, 1966.
- COSTELLO, Anna B.; OSBORNE, Jason W. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, v. 10, n. 7, p. 1-9, 2005.
- DANCEY, C; REIDY, J. *Estatística sem Matemática para Psicologia: usando SPSS para Windows*. Porto Alegre: Artmed, 2006.
- DISTEFANO, Christine; ZHU, Min; MINDRILA, Diana. Understanding and using factor scores: considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, v. 14, n. 20, p. 1-11, 2009.
- FARRAR, Donald E.; GLAUBER, Robert R. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, p. 92-107, 1967.
- FIELD, A. *Descobrimos a Estatística usando o SPSS*. 2. ed. Porto Alegre: Bookman, 2009.
- FIELD, A.; MILES, J.; FIELD, Z. *Discovering statistics using R*. Sage Publications, 2012.
- FIGUEIREDO, D. B.; SILVA, J. A. Visão além do alcance: uma introdução à análise fatorial. *Opinião Pública*, Campinas, v. 16, n. 1, p. 160-185, jun. 2010.
- GANZEBOOM, Harry B. G.; DE GRAAF, Paul M.; TREIMAN, Donald J. A standard international socio-economic index of occupational status. *Social Science Research*, Amsterdam, v. 21, n. 1, p. 1-56, 1992.
- GARSON, G. D. *Statnotes: topics in multivariate analysis*. 2009. [Online] Disponível em: <<http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm>> Acesso em: [22 jan. 2018].
- GRAHAM, J. M.; GUTHRIE, A. C.; THOMPSON, B. Consequences of not interpreting structure coefficients in published CFA research: a reminder. *Structural Equation Modeling*, v. 10, n. 1, p. 142-153, 2003.
- GUADAGNOLI, E.; VELICER, W. Relation of sample size to the stability of component patterns.

Psychological Bulletin. v. 103, p. 265-275, 1988.

JÖRESKOG, K. G. How large can a standardized coefficient be? 1999. Disponível em: <http://www.ssicentral.com/lisrel/techdocs/HowLargeCanaStandardizedCoefficientbe.pdf>. Acesso em: [22 jul. 2018].

HAIR, J. F. *et al.* *Análise multivariada de dados*. 5. ed. Porto Alegre: Bookman, 2005. 593 p.

HUTCHESON, G.; SOFRONIOU, N. *The multivariate social scientist*. London: Sage, 1999.

KAISER, H. F. An index of factorial simplicity. *Psychometrika*, v. 39, p. 31-36, 1974.

HARRIS, L. R.; BROWN, G. T. L.; HARNETT, J. A. Assessment from students' perspectives: using pupil drawings to examine their conceptions of assessment. In: MCINERNEY, D. M.; BROWN, G. T. L.; Liem, G. A. D. (Eds.). *Student perspectives on assessment: what students can tell us about improving school outcomes*. Greenwich, CT: Information Age Press, 2009.

KIM, J; MUELLER, C. W. *Factor analysis: statistical methods and practical issues*. Beverly Hills, CA: Sage, 1978.

KLEM, L. Structural equation modeling. In: GRIMM, L. G.; YARNOLD, P. R. (Eds.). *Reading and understanding more multivariate statistics*. Washington, DC: APA, 2000. cap. 7, p. 227-260.

MACCALLUM, R. C. *et al.* Sample size in factor analysis. *Psychological Methods*. v. 4, n. 1, p. 84-99, 1999.

MATOS, D. A. S. *A avaliação no ensino superior: concepções múltiplas de estudantes brasileiros*, 2010. Tese (Doutorado em Educação) – Faculdade de Educação, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.

MATOS, D. A. S.; BROWN, G. T. L.; CIRINO, S. D. Concepções de avaliação de alunos universitários: uma revisão da literatura. *Estudos em Avaliação Educacional*, v. 23, p. 204-231, 2012.

MATOS, D. A. S. *et al.* Avaliação no ensino superior: concepções múltiplas de estudantes brasileiros. *Estudos em Avaliação Educacional*, v. 24, p. 172-193, 2013.

MUTHÉN, L. K.; MUTHÉN, B. O. *Mplus* (Version 4.2). Los Angeles, CA: Muthén & Muthén, 2005.

PALLANT, J. *SPSS Survival Manual*. Open University Press, 2007.

PASQUALI, L. (Org.) *Instrumentos Psicológicos: manual prático de elaboração*. Brasília: LabPAM, 1999.

SOARES, Tufi Machado. Utilização da teoria da resposta ao item na produção de indicadores sócio-econômicos. *Pesquisa Operacional*, Rio de Janeiro, v. 25, n. 1, p. 83-112, 2005.

STEVENS, J. P. *Applied multivariate statistics for the Social Sciences*. 2. ed. Hillsdale (NJ): Erlbaum, 1992.

SCHAWB, A.J. *Eletronic classroom*. [Online] Disponível em: <<http://www.utexas.edu/ssw/eclassroom/schwab.html>> Acesso em: [22 jan. 2018].

TABACHNICK, B.; FIDELL, L. *Using multivariate statistics*. Needham Heights: Allyn & Bacon, 2007.

URBINA, S. *Fundamentos da testagem Psicológica*. Porto Alegre, RS: Artmed, 2007.

YONG, An Gie; PEARCE, Sean. A beginner's guide to factor analysis: focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, v. 9, n. 2, p. 79-94, 2013.

