Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Lattin, James M.
Análise de dados multivariados / James M. Lattin, J. Douglas Carroll, Paul E. Green ; [tradução Harue Avritscher]. -- São Paulo : Cengage Learning, 2011.

Título original: Analyzing multivariate data.
ISBN 978-85-221-0901-2

1. Análise multivariada I. Carroll, J. Douglas. II. Green, Paul E. III. Título.

10-12603 CDD-519.535

Índice para catálogo sistemático:

1. Análise de dados multivariados : Matemática 519.535

Análise de Dados Multivariados

James M. Lattin
Pós-Graduação na Business School da Stanford University

J. Douglas Carroll
Pós-Graduação na Escola de Administração da Rutgers University

Paul E. Green
Wharton School, Pennsylvania State University

Revisão Técnica
Flavio Soares Corrêa da Silva
PhD em Inteligência Artificial pela Edinburgh University, livre-docente e professor associado do Departamento de Ciência da Computação no Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP)

CENGAGE Learning™

Austrália • Brasil • Japão • Coreia • México • Cingapura • Espanha • Reino Unido • Estados Unidos

5

Análise fatorial exploratória

5.1 INTRODUÇÃO

No capítulo anterior, discutimos a análise dos componentes principais, um método para se reduzir a dimensionalidade dos dados multivariados e que facilita o entendimento dos padrões de associação entre as variáveis. Neste capítulo, discutimos a *análise fatorial exploratória*, um método similar com base em um modelo subjacente diferente, conhecido como *modelo de fator comum*. Como os métodos são similares, a análise fatorial exploratória é frequentemente usada para atingir os mesmos objetivos da análise dos componentes principais. Nossa meta aqui é enfatizar as *diferenças* conceituais entre eles. O modelo de fator comum assume pressupostos explícitos sobre como é medida cada variável no conjunto de dados. O modelo sustenta que a variância observada em cada medida pode ser atribuída a um número relativamente pequeno de *fatores comuns* (isto é, características comuns não observáveis em duas ou mais variáveis) e a um único *fator específico* (não relacionado com qualquer outro fator subjacente do modelo). Embora possa haver, de fato, mais de um fator específico para uma variável, é impossível distinguir estatisticamente entre um e mais de um fator específico. O objetivo da análise fatorial exploratória é identificar os fatores comuns (separados dos fatores específicos) e explicar sua relação com os dados observados. Portanto, mesmo que o procedimento da solução que usamos para realizar a análise fatorial exploratória seja similar ao dos componentes principais, os modelos subjacentes são diferentes.

Com a análise fatorial exploratória, permitimos que os padrões observados de associação nos dados determinem a solução de fator. No Capítulo 6, retornaremos à *análise fatorial confirmatória*, na qual começamos com uma noção anterior sobre a estrutura da solução do fator e então testamos para determinar se ela é consistente com os dados. Na análise confirmatória, o modelo subjacente é o mesmo, mas o procedimento de solução é bem diferente do da análise fatorial exploratória.

O modelo de fator comum fornece um quadro de referência explícito que podemos usar para avaliar as propriedades da medida de nossos dados. Nos exemplos fornecidos neste capítulo, vamos pressupor que a variação em um fator específico seja interpretável puramente como erro de medida. Em geral, para distinguir entre variância específica (isto é, a variância do fator específico) e a variância do erro (isto é, a variância do erro de medida associada com cada variável), é necessário que haja algumas avaliações independentes da confiabilidade da medida (para distinguir-se entre dois fatores específicos separados).

Na ausência dessa avaliação independente, assumimos que a variância no fator específico reflete a quantidade de variância na medida atribuível ao erro. Isso nos dá algum sentido de confiabilidade da medida: quanto menor o erro da variância, mais confiável é a medida. Formalizaremos a noção de confiabilidade no Capítulo 6.

Em nosso desenvolvimento deste capítulo, prestamos particular atenção ao uso da rotação para facilitar a interpretação de uma solução analítica de fator. Como a orientação da solução de fator exploratório é arbitrária, às vezes faz sentido escolher a solução que exiba uma estrutura simples (isto é, aquela que é a mais facilmente interpretável, em um sentido que foi definido inicialmente por Thurstone, 1947). De fato, o uso da rotação é igualmente aplicável quando a análise dos componentes principais (apresentada no Capítulo 4) é utilizada para analisar fatorialmente os dados. Todos os diferentes métodos de rotação desenvolvidos aqui podem ser aplicados à solução dos componentes principais também.

5.1.1 Aplicações potenciais

A análise fatorial exploratória pode ser aplicada na maior parte das mesmas aplicações em que se usa a análise de componentes principais; o modelo de fator comum pode ser preferível quando os pressupostos explícitos do modelo de medida são adequados. Relacionadas, a seguir, estão duas aplicações ilustrativas da análise fatorial exploratória: identificação de "traços latentes" ou "características não observáveis" e uso de escores de fatores na análise de dependência.

Identificando "traços latentes" ou "características não observáveis"

Às vezes, é importante distinguir entre uma variável de dado e o conceito ou ideia que ela se propõe a medir. Quando se está lidando com características físicas, como comprimento ou peso (e o instrumento de medida é altamente preciso), tal distinção pode ser desnecessária porque a propriedade do objeto é quase que perfeitamente observável. No entanto, quando se está lidando com atitudes, crenças, percepções e outras noções psicológicas, nossos instrumentos de medida são, no melhor dos casos, imperfeitos.

Na área de marketing, os pesquisadores podem interessar-se por obter um conceito particular (por exemplo, "satisfação do cliente") para entendê-lo melhor e definir como ele é influenciado pelas ações da empresa. Pode ser difícil, se não impossível, projetar uma única pergunta na pesquisa que capte exatamente um constructo como a satisfação do cliente. Em lugar disso, o pesquisador pode elaborar todo um questionário com várias questões, cada uma insuficiente, mas projetada para captar alguma faceta da satisfação do consumidor. Com o uso da análise fatorial, é possível identificar a fonte da variância comum subjacente a essas questões (as quais presumivelmente refletem a satisfação subjacente do consumidor) e separar os erros não sistemáticos na medida. Os escores dos fatores do modelo de fator comum podem, então, servir como um índice (ou talvez mais do que um índice, dependendo do número dos fatores subjacentes) da satisfação do consumidor em análise subsequente e desenvolvimento de modelo.

EXEMPLO *Identificando as diferentes dimensões da personalidade de marca (Aaker, 1997).*

Aaker (1997) usou a análise fatorial exploratória para identificar as diferentes dimensões da personalidade de marca. Trabalhando com um conjunto de 114 traços de personalidade (triados de um conjunto de 309 traços possíveis gerados por estudos de personalidade em psicologia e pesquisa de mercado), Aaker pediu aos respondentes que classificassem cada marca de um grupo de 10 marcas em relação a cada um dos 114 traços de personalidade, usando uma escala de cinco pontos (de 1 = absolutamente não descritivo até 5 = extremamente descritivo). Ela utilizou quatro conjuntos diferentes de marcas, cada um contendo uma marca em foco (a marca de jeanswear Levi's) e outras nove (todas de marcas nacionais destacadas e bem conhecidas, representando uma ampla variedade de categorias comerciais diferentes), perfazendo um total de 37 marcas diferentes. Aaker calculou a média das classificações dos indivíduos (cada marca foi classificada em relação a cada traço por aproximadamente 150 ou 160 respondentes, com a exceção da Levi's, que foi classificada por todos) e depois realizou a análise fatorial da matriz de correlação 114 × 114 dos traços.

Aaker escolheu uma solução de cinco fatores, que eram responsáveis por mais de 90% da variação dos traços. Após rodar a solução, ela rotulou os fatores da seguinte maneira: *sinceridade* (responsável por 26,5% da variância), *emoção* (responsável por 25,1%), *competência* (17,5%), *sofisticação* (11,9%) e *robustez* (8,8%). A Tabela 5.1 mostra exemplos dos traços que foram mais altamente correlacionados com cada um dos cinco fatores. Com base nos resultados desse estudo, Aaker continuou a construir e validar um instrumento de 42 itens para medir esses cinco componentes diferentes da personalidade de marca.

Tabela 5.1 Traços associados às diferentes dimensões da personalidade de marca

Sinceridade	Emoção	Competência	Sofisticação	Robustez
Honesta	Ousada	Confiável	Glamourosa	Dura
Genuína	Entusiasmada	Responsável	Pretensiosa	Forte
Animadora	Imaginativa	Segura	Encantadora	Ao ar livre
Pés no chão	Moderna	Eficiente	Romântica	Masculina
Amistosa	Legal	Inteligente	De alta classe	
		Bem-sucedida	Suave	

Usando escores fatoriais na análise de dependência

Como nos componentes principais, reduzir o número de dimensões é frequentemente uma importante preocupação na análise fatorial, pois isso facilita a visualização dos dados e ajuda a aumentar a parcimônia do modelo. Por exemplo, como parte de um extenso estudo de mercado envolvendo o conceito de um novo modelo de carro de luxo, Roberts (1984) pesquisou 162 consumidores alvo para determinar suas percepções a respeito de automóveis. Mediu suas percepções sobre modelos de carro familiar em vista de nove dimensões diferentes: luxo, estilo, confiabilidade, economia de combustível, segurança, custos de manutenção, qualidade, durabilidade e desempenho na estrada. Seu principal objetivo era construir um modelo que relacionasse as percepções às preferências; no entanto, o número disponível de graus de liberdade para calibrar tal modelo era limitado. Portanto, ele usou a análise fatorial na tentativa de identificar um número menor de fatores comuns subjacentes que pudessem ser usados em seu lugar.

EXEMPLO *Mapeamento da posição perceptiva do conceito de um novo carro de luxo (Roberts, 1984; Roberts e Urban, 1988) LUXURY_CAR*

Roberts descobriu que uma solução de dois fatores explicava mais de 60% da variância dos nove atributos. Ele, então, rotacionou a solução (usando um procedimento varimax) para facilitar a interpretação dos fatores. As correlações entre os atributos e os fatores, chamadas de *cargas fatoriais*, são apresentadas na Tabela 5.2; as correlações mais altas estão sublinhadas. O padrão sugere que o primeiro fator (que é mais altamente correlacionado com os atributos luxo, estilo, segurança e desempenho na estrada) reflete o apelo emocional do carro, enquanto o segundo fator (correlacionado com confiabilidade, economia de combustível, manutenção, qualidade e durabilidade) reflete o bom-senso econômico. Roberts rotulou os dois fatores como *Atraente* e *Sensato*.

Usando esse modelo, Roberts calculou o valor médio de cada fator para cada modelo de carro classificado e depois fez a regressão dos escores desse fator contra as preferências declaradas dos respondentes. Graças ao uso dos dois fatores, ele foi capaz de explicar 30% da variância em preferência (em termos de R^2 ajustado); ambos os fatores foram altamente significativos. Em contraste, quando Roberts fez a regressão das avaliações dos nove atributos em relação à preferência, sua capacidade de ajustar os dados cresceu apenas ligeiramente (um R^2 ajustado de 33%), enquanto os erros padrão de suas estimativas de parâmetros aumentaram substancialmente devido à multicolinearidade dos atributos (de fato, somente dois dos nove coeficientes eram significativos em nível 0,05). Usando esse

Tabela 5.2 Solução de dois fatores para os dados de Roberts: matriz de cargas de fator

	Atraente	**Sensato**
Luxo	0,884	−0,051
Estilo	0,748	0,153
Confiabilidade	0,396	0,691
Economia de combustível	−0,202	0,786
Segurança	0,720	0,172
Manutenção	0,149	0,756
Qualidade	0,501	0,650
Durabilidade	0,386	0,677
Desempenho	0,686	0,391

modelo de fator parcimonioso, Roberts foi capaz de avaliar a posição relativa do conceito do novo carro (apresentado na Figura 5.1) e prever com exatidão sua participação no mercado.

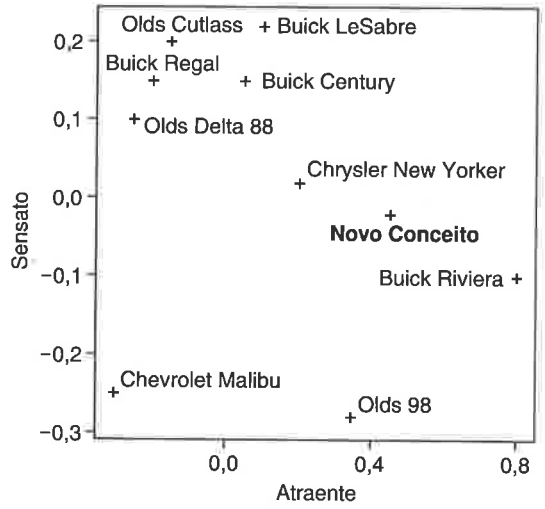

Figura 5.1 Mapa dos escores fatoriais de automóveis existentes e do conceito do novo carro (Fonte: Roberts, 1984).

5.2 ANÁLISE FATORIAL EXPLORATÓRIA: SEU FUNCIONAMENTO

5.2.1 Intuição

Assim como nos componentes principais, a intuição subjacente à análise fatorial é mais bem explicada com um exemplo simples. Considere o teste psicológico com crianças como o conduzido por Holzinger e Swineford (1939). Foram administrados vários testes diferentes em crianças de 7ª e 8ª séries ($n = 145$) e, para simplificar, foquemos os seguintes cinco testes: compreensão de parágrafo (*PARA*), complementação de sentenças (*SENT*), significado das palavras (*WORD*), adição (*ADD*) e contagem de pontos (*DOTS*). Usamos as variáveis X_1 a X_5 para representar esses cinco testes diferentes. As correlações entre os testes (com base em uma amostra de 145 crianças) são apresentadas na Tabela 5.3.

Com a análise fatorial, nosso pressuposto é que os escores observados no teste podem ser descritos como uma função de um pequeno número de fatores comuns subjacentes e um conjunto de fatores específicos (um fator específico para cada teste). Digamos, por exemplo, que acreditamos haver um único fator comum que subjaz os escores de testes dos estudantes, que representamos por ξ (a letra minúscula grega "xi"). Esse fator pode refletir a inteligência ou a capacidade inata de cada estudante de se fazer o teste. Nosso modelo de um fator diz que os escores observados X_i para o teste i são uma

Tabela 5.3 Matriz de correlação para os dados do teste psicológico

	PARA	*SENT*	*WORD*	*ADD*	*DOTS*
PARA	1,000				
SENT	0,722	1,000			
WORD	0,714	0,685	1,000		
ADD	0,203	0,246	0,170	1,000	
DOTS	0,095	0,181	0,113	0,585	1,000

função do fator ξ (que é comum a todos os cinco testes) e um fator específico – vamos chamá-lo δ_i (a letra minúscula grega "delta") – que é específico do teste i. Em várias abordagens para estimar os parâmetros do modelo de fator comum, tudo que se pode estimar realmente é a variância única (ou a "singularidade"), que é a soma da variância desse fator específico e o erro de medida. Assumiremos que essa variância específica se deve inteiramente ao erro de medida, refletindo a incapacidade do teste de captar perfeitamente o fator comum subjacente. Como os fatores específicos afetam somente suas medidas específicas (isto é, o fator δ_i afeta somente o teste i), procederemos com a suposição de que esses fatores específicos δ são mutuamente não correlacionados (isto é, a correlação entre δ_i e δ_j é igual a zero para todos os i e j) e não correlacionados com o fator comum subjacente ξ. Ambas são suposições padrão do modelo de fator comum.

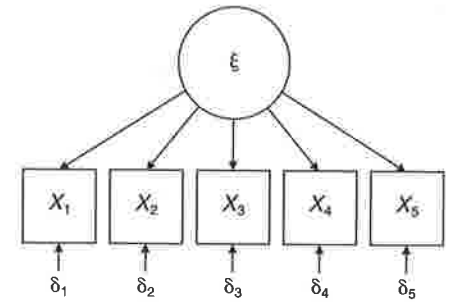

Figura 5.2 Diagrama de caminhos de um modelo de fator único com cinco variáveis.

O modelo de um fator pode ser representado pelo diagrama da Figura 5.2. As flechas no diagrama que apontam para cada escore do teste (essas são as nossas medidas observadas, representadas pelas caixas) indicam uma fonte de variação que contribui para a medida. Nesse caso, cada medida X_i possui duas fontes que contribuem para a variação (ambas não observáveis): o fator comum ξ e o fator específico δ_i. Em forma de equação, podemos escrever o seguinte:

$$\begin{aligned} X_1 &= \lambda_1\xi + \delta_1 \\ X_2 &= \lambda_2\xi + \delta_2 \\ X_3 &= \lambda_3\xi + \delta_3 \\ X_4 &= \lambda_4\xi + \delta_4 \\ X_5 &= \lambda_5\xi + \delta_5 \end{aligned} \tag{5.1}$$

Nessas equações, os coeficientes λ refletem o grau com que cada medida X reflete o fator comum subjacente ξ. Pressupondo que tanto X quanto ξ sejam variáveis padronizadas (isto é, com média zero e variância 1), a variância de X_i pode ser decomposta como segue:

$$\text{var}(X_i) = \text{var}(\lambda_i\xi + \delta_i) = \lambda_i^2 + \text{var}(\delta_i) = 1$$

Como as variáveis são padronizadas, λ_i é interpretável como um coeficiente de correlação. O termo λ_i^2 é interpretável como a proporção da variação em X_i explicada pelo fator comum ξ, e é chamado de *comunalidade* de X_i. A variância remanescente em X_i é explicada pelo fator específico δ_i.

Se usarmos $\theta_{ii}^2 = \text{var}(\delta_i)$ para representar essa variância do fator específico (que pressupomos refletir o erro de medida em X_i), a comunalidade de X_i é igual a

$$1 - \theta_{ii}^2$$

À medida que a comunalidade de X_i se aproxima de 1 (isto é, à medida que a variância do erro vai a zero), sugere-se que X_i é quase uma medida perfeita do fator comum subjacente ξ. Se, por outro lado, ξ não for de maneira alguma captado pela medida de X_i (isso poderia acontecer se X_i fosse um teste mal projetado da inteligência do estudante), o coeficiente λ poderia ser próximo de zero e quase todas as variações em X_i poderiam ser explicadas pelo fator específico δ_i.

Nesta ilustração particular, iniciamos com a suposição de que um único fator comum (como a inteligência geral) é subjacente aos escores do teste no estudo de Holzinger e Swineford. Entretanto, talvez o desempenho individual no teste seja uma função de mais de uma capacidade subjacente. Por exemplo, poderíamos acreditar que o desempenho individual de um estudante em testes é determinado por duas fontes diferentes de capacidade inata – a *aptidão verbal* do estudante (que representamos como ξ_1) e a *aptidão quantitativa* (representada por ξ_2) – e que esses dois fatores diferentes entram em jogo em diferentes níveis, dependendo do tipo de teste. Esse modelo de dois fatores poderia ser representado como apresentado no diagrama da Figura 5.3.

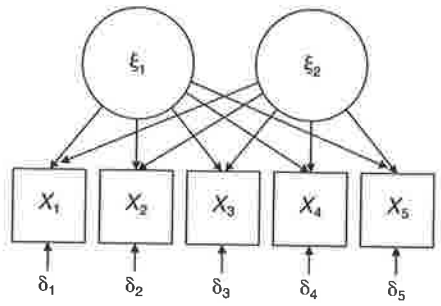

Figura 5.3 Diagrama dos caminhos de um modelo de dois fatores com cinco variáveis.

No modelo de dois fatores, três fontes agora contribuem para a variação observada em cada escore de medida do teste X_i: os dois fatores comuns ξ_1 e ξ_2 e um fator específico δ_i. Em forma de equação, o modelo de dois fatores poderia ser escrito como

$$\begin{aligned} X_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \delta_1 \\ X_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \delta_2 \\ X_3 &= \lambda_{31}\xi_1 + \lambda_{32}\xi_2 + \delta_3 \\ X_4 &= \lambda_{41}\xi_1 + \lambda_{42}\xi_2 + \delta_4 \\ X_5 &= \lambda_{51}\xi_1 + \lambda_{52}\xi_2 + \delta_5 \end{aligned} \qquad \textbf{(5.2)}$$

Como anteriormente, o coeficiente λ reflete a extensão com que cada fator comum contribui com a variância de cada escore de variável no teste. Se X e ξ forem padronizados, os parâmetros λ são interpretáveis como coeficiente de correlação. Se, além disso, assumirmos que os fatores comuns subjacentes não são correlacionados (outra suposição padrão do modelo de fator comum), a comunalidade de cada medida do teste é dada pela soma das cargas de fator ao quadrado para essa variável; portanto, a comunalidade para X_1 é dada por $\lambda_{11}^2 + \lambda_{12}^2$.

Considere o que ocorre, sob o modelo de dois fatores, quando temos um estudante com alta aptidão verbal (alto ξ_1) e baixa aptidão quantitativa (baixo ξ_2). Podemos esperar que tal estudante se saia bem em testes que exigem mais capacidade verbal do que capacidade quantitativa. Se o desempenho do estudante na tarefa de completar uma sentença (medida por X_1) depender somente de sua aptidão verbal, devemos esperar um valor de λ_{11} próximo de 1 e um valor de λ_{12} próximo de zero.

Em um cenário exploratório – aquele em que não há previamente informações fortes a respeito da estrutura do fator subjacente de desempenho do estudante no teste –, desejamos inferir, a partir dos dados, o número apropriado de fatores subjacentes e os valores dos coeficientes nas equações do modelo de fator comum. A abordagem, que detalhamos mais formalmente na próxima seção, é semelhante à dos componentes principais, já que também aqui tentamos extrair um número menor de fatores que irá representar adequadamente a matriz de correlação observada. A diferença em relação ao modelo de fator comum é que devemos agora explicar os fatores específicos δ que não estavam presentes nos componentes principais.

Procedimento para solução

Da mesma forma que nos componentes principais, o procedimento para solução da análise fatorial exploratória foca a decomposição da covariância ou matriz de correlação de **X**. A diferença entre as duas abordagens são os fatores específicos que formam o modelo fatorial comum. Como assume-se que esses fatores específicos são mutuamente não correlacionados e independentes dos fatores comuns subjacentes, eles somente contribuem para a diagonal da matriz de covariância. Isso é facilmente observado quando examinamos o elemento diagonal da matriz de covariância de qualquer medida de X_i.

$$\text{var}(X_i) = \text{var}(\lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \delta_i) \qquad \textbf{(5.3)}$$

Há nove termos de produtos cruzados na expressão para a variância de X_i. Seguindo a partir dos pressupostos de independência impostos pelo modelo de fator comum (e assumindo, além disso, que os fatores comuns foram padronizados com variância igual a 1), todos os termos da covariância abandonam o modelo, resultando em

$$\text{var}(X_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \theta_{ii}^2 = 1 \qquad \textbf{(5.4)}$$

onde $\theta_{ii}^2 = \text{var}(\delta_i)$. Portanto, o único lugar em que os fatores específicos aparecem na matriz de covariância de **X** é na diagonal, onde contribuem com uma parcela da variância de cada medida.

O que aconteceria se conhecêssemos a variância de cada fator específico antecipadamente? Poderíamos, então, subtrair esses valores da diagonal da matriz de covariância. Restaria, nesse caso, uma matriz na qual a única fonte de variação e covariação seria atribuível aos fatores comuns subjacentes. Poderíamos, assim, utilizar uma abordagem de componentes principais para decompor essa matriz e encontrar os fatores comuns. Esse é essencialmente o procedimento de solução que usamos na análise fatorial exploratória. Na análise de componentes principais, decompomos a matriz de correlação **R** (que tem 1 em sua diagonal), mas, para o modelo de fator comum, decompomos uma matriz de correlação com elementos diagonais $1 - \theta_{ii}^2$. Com efeito, começamos subtraindo a variação atribuível aos fatores específicos no modelo (lembre-se de que tal variação pode ser interpretável como erro de medida ou qualquer outra fonte de variação específica para os testes específicos e não correlacionada a outros fatores no modelo), deixando apenas as comunalidades na diagonal. Então, tentamos explicar a variação restante com os fatores comuns. Se soubermos os valores das comunalidades, poderemos então solucionar o problema, lançando mão dos mesmos procedimentos que usamos com os componentes principais. Esse processo é chamado, às vezes, de *fatoração de eixo principal* (ou simplesmente *abordagem de componentes principais* ao modelo de fator comum). A estimativa das comunalidades é um aspecto não trivial desse procedimento de solução, e há muitas abordagens diferentes.

Vamos supor, por enquanto, que sabemos alguma coisa sobre o montante da variação atribuível aos fatores específicos em nosso exemplo ilustrativo (mais tarde, nesta seção, discutiremos o que fazer quando esses valores não são conhecidos). Para esse conjunto específico de testes, digamos que aproximadamente metade da variação em cada fator é atribuível ao fator específico – isto é, começamos por estabelecer $\theta_{ii}^2 = 0{,}50$ para todos os i. Naturalmente, não é necessário que as variâncias do erro sejam iguais em todas as medidas (e, em geral, elas não são); esse é simplesmente um ponto de partida que usamos para ilustrar a aplicação do método. Substituímos os elementos diagonais de **R** pelos valores da comunalidade $1 - \theta_{ii}^2 = 0{,}50$ para todos os i e efetuamos uma decomposição matricial da matriz modificada. Os autovalores são apresentados a seguir:

$$\lambda_1 = 2{,}187 \qquad \lambda_2 = 1{,}022 \qquad \lambda_3 = -0{,}135 \qquad \lambda_4 = -0{,}089 \qquad \lambda_5 = 0{,}015$$

A primeira coisa que observamos é que esses autovalores parecem diferentes daqueles da análise de componentes principais. Nem todos os autovalores são positivos e sua soma não é mais p (o número de variáveis na análise). A razão para isso é que, subtraindo a variação atribuível aos fatores específicos, reduzimos as informações restantes de que necessitamos para determinar os fatores comuns. A variação já explicada pelos fatores específicos no modelo é 0,50 + 0,50 + 0,50 + 0,50 + 0,50 = 2,5, que é 2,5/5,0 = 50% da variação total nos escores dos cinco testes originais. A soma dos elementos diagonais é também 2,5, que é o valor remanescente da variação a ser explicada pelos fatores comuns.

Agora, a questão permanece: de quantos fatores comuns necessitamos? Observe que o critério para seleção do número de fatores a serem extraídos modificou-se com relação à análise de componentes principais. Tendo já determinado alguma variação com os fatores específicos, nossa meta agora é inferir a maior variação restante possível com os fatores comuns. Estamos, portanto, buscando todos os autovalores que sejam significativamente diferentes de zero. Nesse caso, há dois, sugerindo que extraiamos $c = 2$ fatores comuns.

A matriz de *cargas fatoriais* (isto é, a matriz cujos valores são as correlações entre as variáveis originais e os fatores comuns) para a solução de dois fatores é apresentada na Tabela 5.4. O padrão da estrutura do fator mostra que os primeiros três testes (compreensão de parágrafo, complementação da sentença e significado da palavra) possuem cargas mais altas no primeiro fator comum (com todas as correlações ao redor de 0,8), enquanto os dois últimos testes (adição e contagem de pontos) têm cargas mais altas no segundo fator comum (com correlações ao redor de 0,6). Isso é consistente com a interpretação de que o primeiro fator reflete a aptidão verbal do estudante enquanto o segundo fator indica aptidão quantitativa.

Tabela 5.4 Matriz das cargas fatoriais para um modelo de dois fatores (com comunalidades aproximadas)

	Fator 1	Fator 2
PARA	0,7722	−0,2351
SENT	0,7838	−0,1576
WORD	0,7562	−0,2372
ADD	0,4293	0,6017
DOTS	0,3476	0,6506

Utilizando as cargas fatoriais, podemos também calcular a proporção da variação em cada teste explicada pelos dois fatores comuns. Essa proporção é chamada de *comunalidade*. Por exemplo, para o teste de compreensão de parágrafo (X_1), os fatores comuns determinam $(0,77)^2 + (-0,24)^2 = 0,65$ ou 65% da variação no teste. O valor dessa comunalidade é mais alto que o valor de $1 - \theta_{11}^2 = 0,50$ com o qual iniciamos nossa análise, devido ao fato de nosso ponto de partida ter sido baseado em alguma noção anterior que era somente aproximada. Para uma amostra particular de estudantes, é provável que o valor exato seja diferente.

É possível refinar nossas estimativas iniciais através de um processo iterativo e, assim, finalmente chegar a um resultado consistente. Substituímos nossa estimativa inicial de comunalidade pelos resultados da análise fatorial da primeira rodada e efetuamos novamente a análise. Assim, poderíamos substituir o valor de 0,50 na diagonal para X_1 (compreensão de parágrafo) por um novo valor 0,65, e assim por diante para as variáveis restantes. Podemos continuar com esse processo por mais uma iteração, duas iterações ou até que haja convergência – isto é, até que a mudança nas comunalidades entre as rodadas seja suficientemente pequena. Esse processo iterativo é frequentemente usado para a estimativa de comunalidade. A abordagem resultante é, às vezes, chamada de *método de fator principal*.

E se não tivéssemos fortes informações preliminares sobre o erro de medida em nossos dados (isto é, a variação não sistemática em nossas medidas não relacionada aos fatores comuns subjacentes)? O que usaríamos, então, para uma estimativa inicial de comunalidade? Uma medida amplamente utilizada é a *correlação múltipla quadrática* (SMC – squared multiple correlation), que é o montante

de variação em uma variável explicada por todas as outras variáveis no conjunto de dados. Se, por exemplo, quisermos usar a SMC como nossa estimativa inicial de comunalidade de X_1 (compreensão de parágrafo), podemos fazer a regressão de X_1 sobre as variáveis restantes X_2, X_3, X_4 e X_5 e usar o valor de R^2.

O que torna a SMC uma boa estimativa da comunalidade? Lembre-se de que a comunalidade é a proporção da variância em X explicada pelos fatores comuns ξ. Embora preferíssemos, por causa disso, fazer a regressão de X sobre os fatores comuns ξ, temos o problema de os fatores comuns não serem observáveis. O que temos, no entanto, são as variáveis restantes X, e cada uma das quais reflete os fatores ξ subjacentes (embora imperfeitamente). Como são medidos com erro, sua capacidade de explicar a variância em X é atenuada. Assim, a SMC serve como um limite inferior à verdadeira comunalidade. Em geral, quanto mais confiáveis as medidas (isto é, quanto mais baixa a variância do fator específico), mais precisa é a SMC como uma estimativa da comunalidade.

A Tabela 5.5 mostra os resultados da análise de fator comum dos dados de escore do teste com o uso da SMC como estimativa inicial de comunalidade. Utilizamos um processo iterativo como aquele descrito acima: substitua a estimativa de comunalidade refinada para o valor inicial e continue até que haja uma mudança insignificante nos sucessivos valores de comunalidade.

Tabela 5.5 Análise fatorial usando SMCs como estimativas iniciais

Estimativa Prévia de Comunalidade: SMC				
PARA	*SENT*	*WORD*	*ADD*	*DOTS*
0,6158	0,5914	0,5701	0,3672	0,3493

	Autovalores Finais				
	1	2	3	4	5
Autovalor	2,2826	1,0273	0,0252	−0,0010	−0,0247

	Padrão de Fator	
	Fator 1	Fator 2
PARA	0,8349	−0,2418
SENT	0,8253	−0,1398
WORD	0,7898	−0,2274
ADD	0,4146	0,6503
DOTS	0,3297	0,6890
	Variância Explicada por Cada Fator	
	2,2826	1,0273

As diferenças entre as cargas de fator iniciais na Tabela 5.4 e os resultados finais na Tabela 5.5 não são substanciais. O padrão de fator leva à mesma interpretação substantiva dos fatores. A divisão da variação é também quase a mesma: o montante da variação explicado pelos fatores comuns é ainda aproximadamente de 66%. Comparadas com as nossas estimativas iniciais, as comunalidades associadas com os dois últimos testes (adição e contagem de pontos) são mais baixas do que com os três primeiros testes: para X_4 e X_5, as comunalidades são aproximadamente 0,60; para X_1, X_2 e X_3, as comunalidades são aproximadamente 0,70. Se considerarmos essa fonte de variação um erro de medida, podemos concluir que os dois últimos testes são consideravelmente menos confiáveis do que os três primeiros.

Uma palavra de cautela é necessária em se tratando da abordagem iterativa para estimar comunalidades. E se o teste de contagem de pontos for deixado de lado, sobrando a adição como única medida de aptidão quantitativa do estudante? Nessas circunstâncias, nossa estimativa inicial de comunalidade

para a adição, baseada em SMC, será bastante baixa (porque nenhum dos três primeiros testes capta o segundo fator comum). Assim, é mais provável concluirmos que a adição é um teste altamente não confiável quando, de fato, é a única medida de um constructo subjacente importante. É esse problema de estimativa de comunalidade que leva alguns pesquisadores a preferir a análise de componentes principais em lugar da análise fatorial.

Rodando a solução fatorial

Como apresentamos na Seção 5.2.2 a seguir, o modelo de fator comum efetivamente possui um número infinito de soluções, cada uma equivalente no que diz respeito a sua capacidade de reproduzir a matriz de covariância observada. A razão para isso é que a orientação da solução fatorial (isto é, a escolha dos vetores de base que descrevem o sistema de coordenadas) é, no final das contas, arbitrária. Isso é chamado de *indeterminação rotacional* do modelo de fator comum. Com a análise de componentes principais, definimos o problema de modo a resolver a indeterminação: o primeiro componente principal é a combinação linear dos dados originais (apropriadamente escalonados) com a maior variância, o segundo componente principal é a combinação linear com a maior variância a seguir (sujeita a ser não correlacionada com a primeira), e assim por diante. Isso assegura uma única (embora arbitrária) solução.

Se a orientação da solução fatorial for definitivamente arbitrária, por que não escolher uma que nos ajude a entendê-la e interpretá-la melhor? Tentar propor uma interpretação clara dos fatores comuns subjacentes usando a matriz de cargas fatoriais pode ser uma tarefa atemorizadora. Seria vantajoso poder escolher uma orientação diferente, rodando-se a solução fatorial, de tal modo que a matriz de cargas fatoriais fosse simplificada. A questão permanece: como escolher tal rotação? Como operacionalizar a meta de simplificar a matriz de cargas fatoriais e depois encontrar uma rotação da solução que nos deixará mais próximos dessa meta?

A abordagem mais popular para se encontrar a matriz **T** de rotação (por enquanto, nos limitaremos à discussão das *rotações ortogonais*, que preservam a independência dos fatores comuns subjacentes) baseia-se nos princípios da *estrutura simples* desenvolvida por Thurstone (1947). Thurstone acreditava que a maioria dos domínios de conteúdo provavelmente envolveria diversos fatores latentes (isto é, subjacentes ou não observados). Ele também supôs que qualquer variável observada isoladamente poderia estar correlacionada a apenas uns poucos fatores subjacentes, e que qualquer fator individualmente poderia estar associado com apenas algumas variáveis. A ideia geral, então, era encontrar agrupamentos de variáveis na medida do possível, com cada um definindo somente um fator. De modo mais generalizado, gostaríamos de encontrar uma orientação dos eixos do fator de tal modo que cada teste (ou outra variável) tenha cargas relativamente altas (positivas ou negativas) em apenas alguns fatores, sendo a maioria das outras cargas próximas de zero.

Se o raciocínio subjacente de estrutura simples for válido, a matriz de cargas fatoriais deve exibir um tipo particular de padrão (Comrey, 1973):

1. A maioria das cargas de qualquer fator específico (coluna) deve ser pequena (tão próxima de zero quanto possível), e somente algumas cargas devem possuir valor absoluto elevado.
2. Uma linha específica da matriz de cargas, contendo as cargas de uma dada variável com cada fator, deve exibir cargas diferentes de zero em apenas um ou não mais que em alguns poucos fatores.
3. Qualquer par de fatores (colunas) deve exibir diferentes padrões de cargas. De outra forma, não se poderia distinguir os dois fatores representados por essas colunas.

Um exemplo hipotético que demonstra a noção de estrutura simples é apresentado na Tabela 5.6 e na Figura 5.4. Imagine um estudo das percepções do consumidor de analgésicos, no qual cada sujeito é solicitado a classificar sua marca preferida de acordo com seis atributos:

1. Não indispõe o estômago.
2. Não possui efeitos colaterais indesejáveis.
3. Acaba com a dor.
4. Funciona rapidamente.

5. Mantém-me acordado.
6. Oferece alívio limitado.

Tabela 5.6 Cargas fatoriais para analgésicos

	Solução não rodada		Solução rodada	
Atributo	Fator 1	Fator 2	Fator 1	Fator 2
Não indispõe o estômago	0,579	−0,452	0,139	0,721
Sem efeitos colaterais indesejáveis	0,522	−0,572	0,017	0,774
Acaba com a dor	0,645	0,436	0,772	0,097
Funciona rapidamente	0,542	0,542	0,764	−0,051
Mantém-me acordado	−0,476	0,596	−0,034	−0,762
Alívio limitado	−0,613	−0,439	−0,750	−0,074
	Variância explicada por		**Variância explicada por**	
	1,921	1,562	1,765	1,718

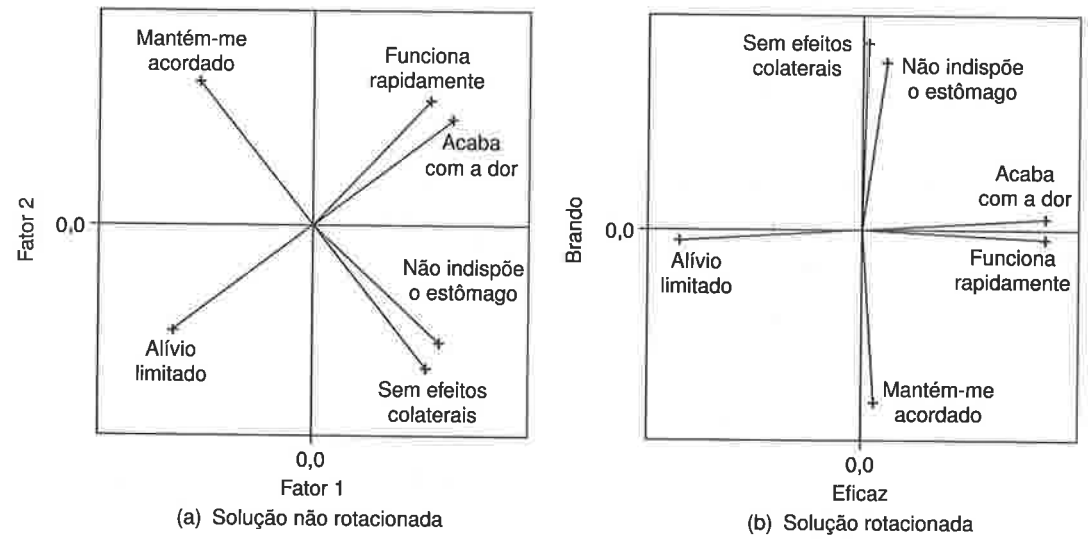

Figura 5.4 Gráfico das cargas fatoriais para atributos dos analgésicos: rodados e não rodados.

A matriz de cargas fatoriais na primeira parte da Tabela 5.6 mostra a solução não rodada de uma análise fatorial com dois fatores comuns. (Lembre-se de que esses fatores não rodados são orientados de tal forma que o primeiro fator comum explique o valor máximo de variação; o segundo fator comum explica tanta variação residual quanto possível, e pode não estar correlacionado com o primeiro fator comum.) Observe que todos os 12 elementos da matriz de cargas fatoriais são relativamente altos (todos maiores que 0,4 em valor absoluto). Com tantas cargas cruzadas substanciais, é difícil obter uma interpretação clara dos fatores.

A Figura 5.4 apresenta as cargas fatoriais e mostra como é possível rodar o espaço fatorial de modo que os atributos se alinhem mais proximamente aos fatores comuns. A ideia é escolher o ângulo de rotação para que as projeções de cada atributo sobre os fatores comuns sejam ou elevados (isto é, próximos a 1 em valor absoluto) ou reduzidos (próximos a zero). Isso reduz essencialmente as cargas cruzadas e desloca a matriz de cargas na direção de uma estrutura simples.

A matriz de cargas fatoriais rodada na Tabela 5.6 agora exibe algo próximo a uma estrutura simples. Os atributos 3 e 4 ("Acaba com a dor" e "Funciona rapidamente") carregam-se positivamente no primeiro fator comum, e o atributo 6 ("Fornece alívio limitado") carrega-se negativamente. Os atributos

1 e 2 ("Não indispõe o estômago" e "Sem efeitos colaterais indesejáveis") carregam-se positivamente sobre o segundo fator comum e o atributo 5 ("Mantém-me acordado") carrega-se negativamente. Todas as cargas restantes são baixas em valor absoluto. Dados os conjuntos de atributos que definem cada fator, podemos nomear o primeiro fator como "Eficácia" e o segundo fator como "Brandura". (As cargas negativas refletem correlações negativas. Por exemplo, quanto maior a classificação em "Mantém-me acordado", menor a "Brandura" do analgésico.)

Observe que, após a rotação, as características advindas da maximização da variância da orientação do fator inicial são perdidas; isto é, embora os fatores retidos no total expliquem exatamente tanta variância no conjunto original de dados quanto antes, essa variância é agora dividida diferentemente através das novas dimensões da configuração rodada. Portanto, não é mais o caso de o primeiro fator (rodado) explicar a variância mais alta e assim por diante. Como o valor da variação explicada por cada fator normalmente não é uma fonte importante de preocupação, a troca é considerada válida se a solução rodada for mais fácil de interpretar.

5.2.2 Mecânica

Nesse tratamento mais formal do modelo de análise fatorial, é interessante retornarmos ao nosso desenvolvimento do modelo de componentes principais do Capítulo 4. Mostramos que resolver o problema de componentes principais era equivalente a fazer uma decomposição do valor único da matriz $\mathbf{X}$ de dados padronizados, como apresentado a seguir:

$$\mathbf{X} = \mathbf{Z}_s\mathbf{D}^{1/2}\mathbf{U}' \tag{5.5}$$

onde $\mathbf{Z}_s$ representa os componentes principais padronizados (todos mutuamente não correlacionados), $\mathbf{D}^{1/2}$ é uma matriz diagonal com o desvio-padrão dos componentes principais definindo a diagonal e $\mathbf{U}$ é a matriz de autovalores (todos mutuamente ortogonais). Utilizando essa descoberta, podemos reescrever a matriz de correlação da amostra $\mathbf{R}$ como uma função dos autovalores e autovetores da decomposição do valor único. A matriz de correlação é dada por

$$\mathbf{R} = 1/(n-1)\mathbf{X}'\mathbf{X} \tag{5.6}$$

Substituindo-se a decomposição de valor único da Equação (5.5) na Equação (5.6) e simplificando, temos

$$\begin{aligned}\mathbf{R} &= 1/(n-1)(\mathbf{Z}_s\mathbf{D}^{1/2}\mathbf{U}')'(\mathbf{Z}_s\mathbf{D}^{1/2}\mathbf{U}') \\ &= 1/(n-1)\mathbf{U}\mathbf{D}^{1/2}(\mathbf{Z}_s'\mathbf{Z}_s)\mathbf{D}^{1/2}\mathbf{U}' \\ &= (\mathbf{U}\mathbf{D}^{1/2})(\mathbf{U}\mathbf{D}^{1/2})'\end{aligned} \tag{5.7}$$

porque $1/(n-1)\mathbf{Z}_s'\mathbf{Z}_s$ é apenas uma matriz identidade. Se, além disso, lembrarmos que o produto da matriz $\mathbf{U}\mathbf{D}^{1/2}$ é apenas a matriz $\mathbf{F}$ de cargas fatoriais (isto é, a matriz de correlação entre a matriz original de dados $\mathbf{X}$ e a matriz de componentes principais $\mathbf{Z}$), podemos simplificar a reexpressão de $\mathbf{R}$ como segue:

$$\mathbf{R} = \mathbf{F}\mathbf{F}' \tag{5.8}$$

Uma vez que nosso objetivo com a análise de componentes principais é, com frequência, obter alguma forma de redução de dimensão, tentamos extrair um subconjunto de componentes c (onde $c < p$, o número de variáveis de $\mathbf{X}$) que se aproxime muito de $\mathbf{R}$. Assim, nos componentes principais, podemos determinar

$$\mathbf{R} \approx \mathbf{F}_c\mathbf{F}_c' \tag{5.9}$$

onde $\mathbf{F}_c$ representa somente as c primeiras colunas da matriz $\mathbf{F}$ de cargas fatoriais.

Em análise fatorial exploratória, também tentamos aproximar a matriz de correlação $\mathbf{R}$, mas usando um modelo diferente. Em lugar de substituir a decomposição de valor único para $\mathbf{X}$ na Equação (5.6),

usamos o modelo de fator comum esboçado anteriormente. O modelo geral com c fatores comuns pode ser escrito como segue:

$$\begin{aligned}X_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \cdots + \lambda_{1c}\xi_c + \delta_1 \\ X_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \cdots + \lambda_{2c}\xi_c + \delta_2 \\ X_3 &= \lambda_{31}\xi_1 + \lambda_{32}\xi_2 + \cdots + \lambda_{3c}\xi_c + \delta_3 \\ &\vdots \\ X_p &= \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \cdots + \lambda_{pc}\xi_c + \delta_p\end{aligned} \tag{5.10}$$

Na notação matricial, o modelo de fator comum torna-se

$$\mathbf{X} = \boldsymbol{\Xi}\boldsymbol{\Lambda}_c' + \boldsymbol{\Delta} \tag{5.11}$$

onde $\boldsymbol{\Xi} = [\xi_1, \xi_2, \ldots \xi_c]$ e $\boldsymbol{\Delta} = [\delta_1, \delta_2, \ldots \delta_p]$ é uma matriz de coeficientes $p \times c$. Além disso, fazemos os seguintes três pressupostos sobre os componentes do modelo de fator comum.

PRESSUPOSTOS PARA O MODELO DE FATOR COMUM

1. Os fatores comuns ξ são mutuamente não correlacionados com variância unitária:
$$\frac{1}{(n-1)}\boldsymbol{\Xi}'\boldsymbol{\Xi} = \mathbf{I}$$
2. Os fatores específicos δ são mutuamente não correlacionados com a matriz de covariância diagonal:
$$\boldsymbol{\Theta} = \frac{1}{(n-1)}\boldsymbol{\Delta}'\boldsymbol{\Delta} = \text{diag}(\theta_{11}^2, \theta_{22}^2, \ldots, \theta_{pp}^2)$$
3. Os fatores comuns ξ e os fatores específicos δ são não correlacionados:
$$\boldsymbol{\Xi}'\boldsymbol{\Delta} = \mathbf{0}$$

Substituimos agora o modelo de fator comum da Equação (5.11) na Equação (5.6) para obter nossa aproximação à matriz de correlação $\mathbf{R}$:

$$\begin{aligned}\mathbf{R} &= 1/(n-1)(\boldsymbol{\Xi}\boldsymbol{\Lambda}_c' + \boldsymbol{\Delta})'(\boldsymbol{\Xi}\boldsymbol{\Lambda}_c' + \boldsymbol{\Delta}) \\ &= 1/(n-1)(\boldsymbol{\Lambda}_c\boldsymbol{\Xi}'\boldsymbol{\Xi}\boldsymbol{\Lambda}_c' + \boldsymbol{\Delta}'\boldsymbol{\Xi}\boldsymbol{\Lambda}_c' + \boldsymbol{\Delta}_c\boldsymbol{\Xi}'\boldsymbol{\Delta} + \boldsymbol{\Delta}'\boldsymbol{\Delta})\end{aligned} \tag{5.12}$$

Pelo pressuposto 3 do modelo de fator comum, o segundo e o terceiro termos na expressão entre parênteses acima vão para zero. Pelo pressuposto 1, podemos substituir a expressão $1/(n-1)\boldsymbol{\Xi}\boldsymbol{\Xi}'$ do primeiro termo pela matriz identidade $\mathbf{I}$. E, pelo pressuposto 2, o último termo na expressão se torna $\boldsymbol{\Theta}$. Fazendo essas simplificações, chegamos a

$$\mathbf{R} = \boldsymbol{\Lambda}_c\boldsymbol{\Lambda}_c' + \boldsymbol{\Theta} \tag{5.13}$$

ou

$$\mathbf{R} - \boldsymbol{\Theta} = \boldsymbol{\Lambda}_c\boldsymbol{\Lambda}_c' \tag{5.14}$$

Uma comparação da Equação (5.9) do modelo de componentes principais com a Equação (5.14) do modelo de fator comum revela a semelhança entre as duas abordagens – e sua diferença essencial. Em cada caso, fazemos uma decomposição de matriz de uma matriz simetricamente quadrada. A matriz $\boldsymbol{\Lambda}_c$ é exatamente análoga à matriz $\mathbf{F}_c$: é uma matriz de cargas fatoriais, cujos elementos são interpretáveis como as correlações entre as variáveis originais $\mathbf{X}$ e o os fatores comuns extraídos c.

Indeterminação rotacional

Nos componentes principais, escolhemos cada componente de modo sequencial para explicar o valor máximo possível de variação em nossos dados originais, com limitação de não haver correlação com todos os componentes previamente selecionados. Isso garantia uma solução única (embora algo arbitrária em termos de orientação escolhida). Com o modelo de fator comum, no entanto, não impusemos tal limitação. Como consequência, há efetivamente um número infinito de soluções que são idênticas no seu grau de aproximação da matriz $\mathbf{R} - \boldsymbol{\Theta}$. Referimo-nos a essa propriedade como *indeterminação rotacional* do modelo de fator comum.

Demonstraremos inicialmente essa propriedade com um exemplo. Considere a solução de dois fatores para os dados de escore de teste apresentados na Tabela 5.5. Um gráfico das cargas fatoriais da tabela é mostrado na Figura 5.5. Mudamos agora a orientação dos fatores rodando-os 30 graus em sentido horário. A rotação (que é executada pela multiplicação da matriz) preserva a ortogonalidade dos dois fatores. Do Capítulo 2, sabemos que uma matriz, para realizar uma rotação ortogonal (em duas dimensões), toma a seguinte forma:

$$\mathbf{T} = \begin{bmatrix} \cos\alpha & -\operatorname{sen}\alpha \\ \operatorname{sen}\alpha & \cos\alpha \end{bmatrix} \tag{5.15}$$

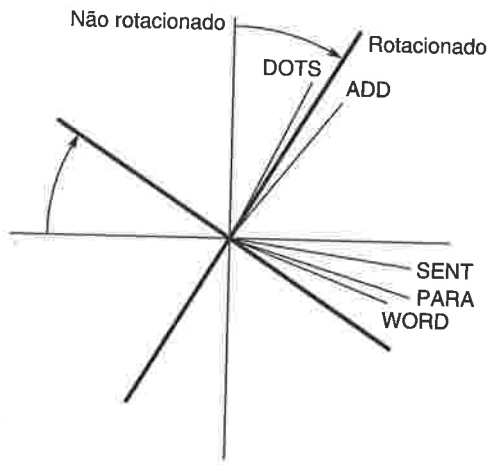

Figura 5.5 Diagrama mostrando uma rotação no sentido horário de 30 graus.

Quando o ângulo de rotação $\alpha = -30$ graus, obtemos a seguinte matriz $\mathbf{T}$ de rotação ortogonal:

$$\mathbf{T} = \begin{bmatrix} 0{,}866 & 0{,}500 \\ -0{,}500 & 0{,}866 \end{bmatrix} \tag{5.16}$$

Mudando a orientação dos eixos que representam os fatores comuns, também mudamos os valores das cargas fatoriais. Podemos calcular as novas cargas (representadas por $\boldsymbol{\Lambda}_c^*$), que são apenas as correlações entre os fatores rodados ($\boldsymbol{\Xi}\mathbf{T}$) e as variáveis originais $\mathbf{X}$ (dadas por $\boldsymbol{\Xi}\boldsymbol{\Lambda}_c' + \boldsymbol{\Delta}$), simplesmente como

$$\boldsymbol{\Lambda}_c^* = 1/(n-1)(\boldsymbol{\Xi}\boldsymbol{\Lambda}_c' + \boldsymbol{\Delta})'\boldsymbol{\Xi}\mathbf{T}$$

ou

$$\boldsymbol{\Lambda}_c^* = \boldsymbol{\Lambda}_c\mathbf{T}$$

porque $1/(n-1)\boldsymbol{\Xi}'\boldsymbol{\Xi} = \mathbf{I}$ e $\boldsymbol{\Lambda}'\boldsymbol{\Xi} = \mathbf{0}$. As cargas fatoriais rodadas $\boldsymbol{\Lambda}_c^*$ são apresentadas na Tabela 5.7 ao lado de cargas originais não rodadas. Como antes, as principais cargas na matriz (isto é, aquelas com os valores absolutos mais elevados) não se modificaram dramaticamente: os três primeiros testes carregam sobre o primeiro fator e os dois últimos carregam no segundo. Mas observe que as cargas cruzadas na solução rodada mudaram. Os três primeiros testes agora carregam positivamente sobre o

segundo fator (ao invés de negativamente), e os dois últimos testes carregam quase exclusivamente sobre o segundo fator (em vez de positivamente sobre os dois fatores).

Tabela 5.7 Cargas fatoriais iniciais e cargas fatoriais rodadas –30 graus

	Cargas iniciais		Cargas rodadas –30 graus	
	Fator 1	Fator 2	Fator 1	Fator 2
PARA	0,811	–0,200	0,801	0,232
SENT	0,811	–0,106	0,754	0,313
WORD	0,779	–0,195	0,773	0,221
ADD	0,375	0,586	0,032	0,695
DOTS	0,295	0,614	–0,052	0,679

Há duas propriedades importantes da solução de fator rodado que devem ser observadas. Primeira, embora a variância explicada por cada um dos fatores se modifique, a variância total explicada por ambos os fatores permanece exatamente a mesma. Como a solução não rodada é obtida via decomposição da matriz $\mathbf{R} - \boldsymbol{\Theta}$, tal solução é orientada para que o primeiro fator comum explique o valor máximo de variação e o segundo fator comum explique a variação restante. A rotação muda a orientação, o que garante que a variação explicada pelo primeiro fator vai diminuir. No entanto, a rotação somente muda a orientação dos eixos atravessando o espaço dos fatores comuns; assim, ela não influencia o valor total da variação explicada.

A segunda propriedade da solução rodada é que as comunalidades não são alteradas pela rotação. Essa propriedade é observada mais facilmente reconstruindo-se a matriz $\mathbf{R} - \boldsymbol{\Theta}$ a partir das cargas fatoriais rodadas. Temos

$$\begin{aligned} \mathbf{R} - \boldsymbol{\Theta} &= \boldsymbol{\Lambda}_c^*\boldsymbol{\Lambda}_c^{*\prime} \\ &= \boldsymbol{\Lambda}_c\mathbf{T}(\boldsymbol{\Lambda}_c\mathbf{T})' \\ &= \boldsymbol{\Lambda}_c\mathbf{T}\mathbf{T}'\boldsymbol{\Lambda}_c' \end{aligned} \tag{5.17}$$

Multiplicando as matrizes, fica fácil ver que $\mathbf{T}\mathbf{T}' = \mathbf{I}$:

$$\begin{bmatrix} \cos^2\alpha + \operatorname{sen}^2\alpha = 1 & -\cos\alpha\operatorname{sen}\alpha + \operatorname{sen}\alpha\cos\alpha = 0 \\ \operatorname{sen}\alpha\cos\alpha - \cos\alpha\operatorname{sen}\alpha = 0 & \cos^2\alpha + \operatorname{sen}^2\alpha = 1 \end{bmatrix}$$

Mais genericamente, $\mathbf{T}\mathbf{T}' = \mathbf{I}$ é sempre verdadeiro para uma rotação ortogonal, em qualquer número de dimensões, pois a matriz transposta age no sentido de rodar os eixos de volta na direção oposta, deixando-nos com a orientação inicial.

O resultado importante – que as comunalidades, a variância explicada pelos fatores comuns e a adequação do modelo de fator comum não são afetados pela rotação ortogonal – será útil mais tarde, quando considerarmos os modos de melhorar a capacidade de interpretação da solução fatorial.

Rotação fatorial

Para encontrar a rotação apropriada, devemos determinar um meio de quantificar o que queremos dizer com estrutura simples na forma de uma função objetiva. Então, procuramos através de todos os ângulos possíveis de rotação e selecionamos a matriz de rotação $\mathbf{T}$, de tal forma que a matriz de cargas fatoriais rodada $\mathbf{A} = \boldsymbol{\Lambda}_c\mathbf{T}$ exiba o maior valor da função objetiva para uma estrutura simples. Há muitas maneiras diferentes de se quantificar a estrutura simples, e cada uma envolve uma forma ligeiramente diferente de rotação. No restante desta seção, discutiremos apenas duas: a rotação varimax de Kaiser e a rotação quartimax.

Rotação varimax de Kaiser. Lembre-se de que cada elemento da matriz de cargas rodada a_{ik} pode ser entendido como a correlação entre a variável i e o fator comum k. A carga ao quadrado a_{ik}^2 é a proporção da variação na variável i atribuível ao fator comum k. Já que escolhemos os fatores comuns como não correlacionados, o valor total da variação explicada por todos os fatores comuns – que denominamos de

comunalidade – é dado pela soma das cargas ao quadrado, ou $h_i^2 = \Sigma_k a_{ik}^2$. Para obter uma estrutura simples, gostaríamos de encontrar uma rotação que determine que as cargas ao quadrado a_{ik}^2 sejam próximas a 1 ou zero.

O procedimento varimax tenta fazer isso focando-se nas colunas de **A**: ele escolhe a matriz de rotação **T** para maximizar a variância total da coluna a_{ik}^2. A k-ésima variância de coluna é dada pela expressão

$$V_k = \frac{1}{p}\sum_{i=1}^{p}(a_{ik}^2)^2 - \frac{1}{p^2}\left(\sum_{i=1}^{p} a_{ik}^2\right)^2 \tag{5.18}$$

Maximizar a soma dessas variâncias de coluna V_k para todos os fatores k equivale a maximizar a seguinte expressão:

$$V = \sum_{k=1}^{c}\sum_{i=1}^{p} a_{ik}^4 - \frac{1}{p}\sum_{k=1}^{c}\left(\sum_{i=1}^{p} a_{ik}^2\right)^2 \tag{5.19}$$

Observe que a variância máxima é alcançada quando os valores de a_{ik}^2 são levados a zero ou 1; por definição, quando o valor de a_{ik}^2 se aproxima de 1 para algum fator k, todos os outros elementos naquela linha da matriz são levados a tender a zero.

É também possível construir a função objetiva para uma rotação varimax utilizando-se as cargas ao quadrado normalizadas, a_{ik}^2/h_i^2. Quando as cargas são normalizadas dessa maneira, o elemento a_{ik}^2/h_i^2 é interpretável como a proporção da variância comum na variável i atribuível ao fator comum k. Usar essa normalização garante que todas as variáveis recebam peso igual na escolha da rotação (que pode não ser o caso para cargas ao quadrado não normalizadas, quando algumas variáveis possuem baixas comunalidades).

Rotação quartimax. Em contraste à varimax, que foca as colunas da matriz de cargas fatoriais rodadas **A**, a rotação quartimax foca as linhas. A função objetiva para a quartimax depende de as comunalidades de uma variável não serem alteradas pela rotação ortogonal; assim, a expressão $\Sigma_k a_{ik}^2$ é constante, independente da matriz de rotação **T**. É também verdade que a soma das comunalidades ao quadrado por todas as variáveis $\Sigma_i\ (\Sigma_k a_{ik}^2)^2$ é igualmente uma constante. Expandindo essa expressão, o resultado é

$$\sum_{i=1}^{p}\sum_{k=1}^{c} a_{ik}^4 + \sum_{i=1}^{p}\left(\sum_{k=1}^{c}\sum_{j\neq k} a_{ik}^2 a_{ij}^2\right) \tag{5.20}$$

O segundo termo na expressão acima é simplesmente a soma dos produtos cruzados das cargas ao quadrado. Quando uma matriz exibe uma estrutura simples, esse produto deve ser tão pequeno quanto possível (isto é, quando uma variável carrega-se altamente sobre algum fator k, a carga deve ser próxima de zero para todos os outros fatores j). Uma vez que a expressão anterior é uma constante para todas as matrizes de rotação **T**, maximizar o seu primeiro termo é um modo de se garantir que os termos do produto cruzado sejam pequenos. Portanto, a quartimax escolhe uma matriz de rotação ortogonal **T** para maximizar

$$Q = \sum_{i=1}^{p}\sum_{k=1}^{c} a_{ik}^4 \tag{5.21}$$

Como na varimax, é possível normalizar as cargas ao quadrado antes de realizar a rotação, dividindo-se pela comunalidade de cada variável.

Levantamos a questão da rotação oblíqua (em que a ortogonalidade mútua dos fatores não é preservada pela rotação) na Seção 5.4.1 a seguir.

Escores fatoriais

Com frequência, a análise fatorial não é um fim em si mesma, mas um passo intermediário no caminho de uma análise posterior dos dados. Para realizar uma análise subsequente, precisamos da localização de cada uma das observações originais no reduzido espaço fatorial. Esses valores são chamados de *escores fatoriais*.

Obter escores fatoriais de um modelo de fator comum não é tão fácil quanto obter escores da análise de componentes principais. Lembre-se de que os escores dos componentes principais são combinações lineares de variáveis originais que podem ser calculadas diretamente, utilizando-se os coeficientes dos autovalores da matriz de correlação. Na análise de fator comum, os escores não podem ser calculados exatamente devido à indeterminação introduzida pelos fatores específicos. Em outras palavras, com o modelo de fator comum, não podemos determinar ξ como uma função de X sem conhecer δ.

Portanto, é necessário calcular os coeficientes dos escores fatoriais que usaremos para calcular os escores fatoriais. Usamos a seguinte expressão linear para aproximar $\boldsymbol{\Xi}$:

$$\boldsymbol{\Xi} = \mathbf{X}_s\mathbf{B} \tag{5.22}$$

onde **B** é a matriz de coeficientes de escore de fator. Como os valores de $\boldsymbol{\Xi}$ não são diretamente observados, não podemos escolher **B** usando a regressão dos mínimos quadrados. No entanto, se pré-multiplicarmos cada lado da equação acima por $1/(n-1)\mathbf{X}_s'$, teremos

$$\frac{1}{(n-1)}\mathbf{X}_s'\boldsymbol{\Xi} = \frac{1}{(n-1)}\mathbf{X}_s'\mathbf{X}_s\mathbf{B} \tag{5.23}$$

ou

$$\boldsymbol{\Lambda}_c = \mathbf{R}\mathbf{B}$$

Podemos garantir que essa igualdade se mantém caso pré-multipliquemos ambos os lados dessa equação por $\mathbf{R}^{-1}$, levando à expressão

$$\mathbf{B} = \mathbf{R}^{-1}\boldsymbol{\Lambda}_c \tag{5.24}$$

para os coeficientes do escore fatorial. Substituindo essa expressão de **B** na Equação (5.22), teremos a seguinte expressão para os escores fatoriais estimados:

$$\boldsymbol{\Xi} = \mathbf{X}_s\mathbf{R}^{-1}\boldsymbol{\Lambda}_c \tag{5.25}$$

Observe que, como a matriz de cargas $\boldsymbol{\Lambda}_c$ está sujeita à indeterminação rotacional, os escores de fator identificados na Equação (5.25) não são únicos e são sujeitos à mesma indeterminação rotacional. O que não varia é o produto $\hat{\mathbf{X}}_c = \boldsymbol{\Xi}\boldsymbol{\Lambda}_c'$, em que $\hat{\mathbf{X}}_c$ é a parte apropriada de **X** explicada pelos fatores comuns c.

5.3 EXEMPLO DE PROBLEMA: PERCEPÇÕES DE CEREAIS PRONTOS PARA COMER

5.3.1 Dados

Como parte de um estudo das impressões do consumidor sobre cereais prontos para comer, patrocinado pela Kellogg da Austrália, Roberts e Lattin (1991) pesquisaram a percepção desses consumidores em relação às suas marcas favoritas de cereais. Solicitou-se a cada respondente que avaliasse suas três marcas preferidas tendo em mente cada um dos 25 diferentes atributos. Uma escala de cinco pontos foi utilizada para indicar a extensão de cada atributo em relação a cada marca.

> **EXEMPLO** *Criação de um mapa de percepção das marcas de cereais prontos para comer no mercado australiano (Roberts e Lattin, 1991) RTE_CEREAL [RTE = Pronto para comer]*

Para os propósitos dessa ilustração, selecionamos um subconjunto de dados coletados por Roberts e Lattin, refletindo as avaliações das 12 marcas de cereais mais frequentemente citadas na amostra (no estudo original, um total de 40 marcas diferentes foram avaliadas por 121 respondentes, mas apenas um pequeno número de consumidores classificou a maior parte das marcas). Os 25 atributos e as 12 marcas estão relacionados na Tabela 5.8. No total, 116 respondentes forneceram 235 observações sobre as 12 marcas selecionadas.

Tabela 5.8 Lista de 25 atributos de cereais RTE e 12 marcas de cereais RTE

Cereais	Atributos	
1. All Bran	Satisfaz	Família
2. Cerola Muesli	Natural	Calorias
3. Just Right	Fibra	Simples
4. Kellogg's Corn Flakes	Doce	Crocante
5. Komplete	Fácil	Regular
6. NutriGrain	Sal	Açúcar
7. Purina Muesli	Gratificante	Fruta
8. Rice Bubbles	Energia	Processo
9. Special K	Divertido	Qualidade
10. Sustain	Crianças	Prazer
11. Vitabrit	Encharcado	Chato
12. Weetbix	Econômico	Nutritivo
	Saúde	

Um objetivo do estudo era caracterizar o comportamento do consumidor de cereal (isto é, explicar quais marcas o consumidor poderia pensar em adquirir e quais delas sequer consideraria) como uma função das características subjacentes das marcas disponíveis. Para alcançar tal objetivo em nível individual, é necessário um modelo relativamente parcimonioso. Roberts e Lattin, portanto, propuseram usar a análise fatorial para reduzir a dimensionalidade dos dados de atributo e revelar um número menor de fatores subjacentes que explicassem um valor importante da variância nas medidas originais. Vamos agora analisar nosso subconjunto escolhido dos dados de Roberts e Lattin, usando análise fatorial exploratória, e discutir os resultados.

5.3.2 Resultados

Começamos examinando um gráfico scree dos autovalores da análise de componentes principais para determinar o número apropriado de fatores para o modelo de fator comum. O gráfico scree, apresentado na Figura 5.6, sugere que extração de quatro fatores é justificável; observe que três deles parecem formar um "ângulo" no valor da variância explicada após o quarto autovalor. Usando um critério de proporcionalidade (isto é, cada fator comum deve explicar pelo menos tanta variação quanto uma das variáveis originais da análise, o que é diretamente análogo à lógica subjacente da regra de Kaiser para reter fatores na análise de componentes principais), a inclusão do quinto fator é, no melhor dos casos, marginal.

Em seguida, executamos o modelo de fator comum, usando a SMC como nossas estimativas iniciais das comunalidades. As SMC, apresentadas na Tabela 5.9, variam em valor, de um baixo 0,23 (para o

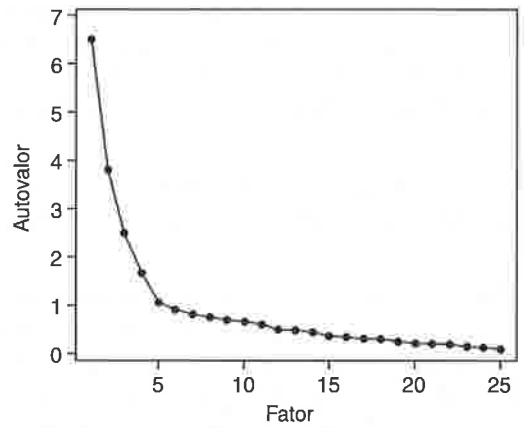

Figura 5.6 Gráfico scree para os dados de cereal RTE.

atributo "Fácil") a um elevado 0,75 (para o atributo "Saúde"). A soma das SMC de todos os 25 atributos é 13,0. Portanto, começamos com uma estimativa inicial que sugere que os fatores comuns explicam mais da metade (13/25 = 52%) da variação nos dados de atributo.

Tabela 5.9 SMC para cada um dos atributos de cereal RTE

Satisfaz	Natural	Fibra	Doce	Fácil	Sal	Gratificante
0,6333	0,6359	0,7035	0,6116	0,2304	0,4481	0,6150
Energia	Divertido	Crianças	Encharcado	Econômico	Saúde	Família
0,5711	0,4833	0,6360	0,3143	0,3722	0,7500	0,5989
Calorias	Simples	Crocante	Regular	Açúcar	Fruta	Processo
0,4168	0,4083	0,4105	0,5211	0,6642	0,4877	0,2975
Qualidade	Prazer	Chato	Nutritivo			
0,6144	0,5607	0,3340	0,7107			

A solução de quatro fatores, incluindo cargas fatoriais não rodadas, é apresentada na Tabela 5.10 a seguir. Observe que alguns dos atributos têm valores relativamente baixos de comunalidade. Por exemplo, o atributo "Fácil" possui um valor final de comunalidade de 0,17 na solução de quatro fatores. Deste modo, temos que 83% da variação nessa medida particular são únicos para o atributo. Isso é potencialmente uma preocupação: por essa variável particular, não estamos captando quase nenhuma de suas informações na solução de fator comum.

Há várias razões potenciais para que isso ocorra. Uma pode ser o fato de termos optado por extrair muito poucos fatores e, como consequência, há uma quantidade limitada de informações que podemos explicar nessa medida particular. Em geral, não é má ideia tentar extrair um fator adicional para determinar se ele altera os resultados (isto é, a comunalidade do atributo aumenta?). Em caso positivo, a capacidade de explicar a variação adicional em alguma medida importante pode justificar a complexidade de uma solução com número mais elevado de fatores.

Uma segunda razão potencial tem a ver com o problema do cálculo da comunalidade mencionada anteriormente. Se "Fácil" é o único atributo que capta a facilidade de uso subjacente do produto, nossa estimativa inicial da comunalidade desse atributo (baseada na SMC) será então baixa. Sem ao menos outra variável projetada para medir o mesmo fator comum (ou alguma avaliação independente da confiabilidade do atributo), não podemos averiguar a verdadeira importância do atributo na análise.

Por fim, uma comunalidade baixa pode também ser tomada como indicação de uma medida ruim (isto é, uma medida na qual há ruído considerável associado a erro de medida). Nesse caso particular, parece que "Fácil" é um atributo que se aplica igualmente bem a todos os itens na categoria do produto (afinal de contas, qualquer cereal pronto para consumo deve ser fácil de preparar). Dentro da categoria, pode ser muito difícil para os consumidores distinguir entre as marcas de cereais com base na facilidade de uso. Se esse for o caso, é de se esperar que a maior parte da variância nesse atributo seja não sistemática; como consequência, a solução fatorial teria um componente de variância específica relativamente alta. Em geral, alguma forma de pré-testagem para se estabelecer a propriedade da medida em um dado contexto de estudo ajudará a evitar o problema.

Agora, usamos a rotação como tentativa de facilitar a interpretação da solução de quatro fatores para cereais RTE. A Tabela 5.11 mostra a solução de fator rodado varimax (a solução quartimax seria muito similar). Para tornar mais fácil a visualização da estrutura nas cargas fatoriais, sublinhamos todos os elementos maiores que 0,40 em cada matriz (se nenhum elemento na fila exceder 0,40, marcamos o elemento de valor mais alto com um sublinhado com pontos).

Voltamos agora à interpretação. Nosso objetivo ao nominar cada fator é sugerir um termo que descreva o domínio do conteúdo das variáveis que carregam altamente naquele fator. Dar nome a um fator é quase sempre uma forma de arte em si. Em geral, é muito pobre utilizar-se apenas dos nomes das variáveis para denominar os fatores, embora, algumas vezes, isso seja inevitável.

Tabela 5.10 Solução fatorial não rodada para dados de cereal RTE

	Fator 1	Fator 2	Fator 3	Fator 4
Satisfaz	0,7276	0,1030	−0,0674	0,1960
Natural	0,7323	−0,2422	−0,1077	0,1052
Fibra	0,7226	−0,2385	−0,3098	0,1607
Doce	0,0850	0,7407	−0,2053	0,1510
Fácil	0,3181	0,1454	0,2064	0,1074
Sal	−0,2062	0,4995	−0,1385	0,3971
Gratificante	0,7207	0,1809	0,1627	0,1704
Energia	0,7022	0,1328	−0,0657	0,1264
Divertido	0,3886	0,4959	0,2101	−0,1579
Crianças	0,2133	0,2759	0,7335	0,1118
Encharcado	−0,0984	−0,2319	0,1515	0,4257
Econômico	0,1508	−0,2361	0,4890	0,1002
Saúde	0,8081	−0,3075	−0,1068	0,0713
Família	0,3074	0,2179	0,6731	0,0271
Calorias	−0,1603	0,5663	−0,1721	0,2151
Simples	−0,3055	−0,3560	0,2178	0,4086
Crocante	0,2888	0,4544	0,2143	−0,1927
Regular	0,5943	−0,1354	−0,1927	0,0682
Açúcar	−0,2476	0,7213	−0,2467	0,2416
Fruta	0,3733	0,2534	−0,4802	−0,1569
Processo	−0,3033	0,2721	0,0030	0,2470
Qualidade	0,7324	−0,1394	0,0515	−0,0328
Prazer	0,4648	0,5632	0,0758	−0,2018
Chato	−0,3789	−0,2619	−0,0988	0,3392
Nutritivo	0,7978	−0,2176	−0,1422	0,1252
Variância explicada por cada fator				
	6,1085	3,3528	2,0456	1,1184

Estimativas finais de comunalidade

Satisfaz	Natural	Fibra	Doce	Fácil	Sal	Gratificante
0,5829	0,6176	0,7008	0,6208	0,1765	0,4689	0,6077
Energia	Divertido	Crianças	Encharcado	Econômico	Saúde	Família
0,5311	0,4660	0,6722	0,2676	0,3277	0,7641	0,5958
Calorias	Simples	Crocante	Regular	Açúcar	Fruta	Processo
0,4222	0,4344	0,3730	0,4133	0,7008	0,4588	0,2270
Qualidade	Prazer	Chato	Nutritivo			
0,5596	0,5798	0,3370	0,7197			

Os atributos que carregam altamente sobre o primeiro fator comum são Satisfaz, Natural, Fibra, Gratificante, Energia, Saudável, Regular, Qualidade e Nutritivo. Roberts e Lattin denominaram esse fator de "Saudável". Embora o termo possa superenfatizar a contribuição de um entre vários atributos, ele capta a natureza dos vários benefícios oferecidos. Outros rótulos possíveis seriam "Benéfico" (embora, em marketing, a maioria dos atributos de produtos é pensada de forma a transmitir ideia de benefícios ao consumidor, portanto, "Benéfico" pode não ser específico o suficiente) ou "Saudável" (embora "Saudável" possa dar ideia de um produto totalmente natural e, assim, terminaria em conflito com a definição do segundo fator). A frase que vem à mente quando se lê a lista de atributos associados com o primeiro fator é: "é bom para você"!

Tabela 5.11 Rotação varimax de solução de fator para dados de cereal RTE

	Fator 1	Fator 2	Fator 3	Fator 4
Satisfaz	0,7057	0,0885	0,1998	0,1512
Natural	0,7529	−0,2084	0,0555	0,0366
Fibra	0,8214	−0,1156	−0,1200	0,0211
Doce	0,0684	0,7020	0,0744	0,3469
Fácil	0,2382	0,0634	0,3252	0,0643
Sal	−0,0924	0,6855	0,0161	−0,0836
Gratificante	0,6254	0,0770	0,4240	0,1703
Energia	0,6594	0,0786	0,1930	0,2098
Divertido	0,1629	0,1765	0,4175	0,4781
Crianças	−0,0252	0,0340	0,8502	0,0102
Encharcado	0,0330	0,0143	0,0942	−0,4806
Econômico	0,0686	−0,2809	0,4149	−0,2289
Saúde	0,8287	−0,2877	0,0519	0,0460
Família	0,0616	−0,0554	0,7612	0,0899
Calorias	−0,1141	0,6267	−0,0072	0,1203
Simples	−0,1466	−0,0622	0,0680	−0,6566
Crocante	0,0734	0,1458	0,3732	0,4362
Regular	0,6132	−0,0996	−0,0270	0,0889
Açúcar	−0,1844	0,8166	−0,0522	0,1651
Fruta	0,3764	0,1874	−0,2669	0,4429
Processo	−0,2363	0,3737	0,0262	−0,1259
Qualidade	0,6466	−0,2444	0,2048	0,1704
Prazer	0,2444	0,2335	0,3368	0,6019
Chato	−0,1646	0,0668	−0,2255	−0,5048
Nutritivo	0,8315	−0,1764	0,0517	0,0558
Variância explicada por cada fator				
	5,2026	2,6598	2,4747	2,2837

Os atributos que carregam altamente sobre o segundo fator comum são: Doce, Sal, Calorias e Açúcar. Processado também sobrecarrega mais intensamente sobre o fator 2, com uma correlação um pouco abaixo de 0,40. Roberts e Lattin usaram o rótulo "Artificial" para descrever esse fator. Esse rótulo particular é consistente com o fato de os atributos Natural, Saudável e Nutritivo também carregarem negativamente sobre esse fator. Como esse fator não está correlacionado com o primeiro fator, seria inapropriado rotulá-lo como "Não saudável" (que sugere uma correlação forte, se não perfeita, com o fator 1).

Os atributos que carregam mais altamente sobre o terceiro fator comum são Crianças, Econômico e Família. Os atributos Gratificante e Divertido têm também cargas próximas de 0,40 para esse fator (mas possuem cargas mais elevadas sobre outros fatores). Esse padrão geral é consistente com um fator que reflete a "Popularidade" ou "Amplo Apelo" do cereal. Roberts e Lattin usaram o termo "Não Adulto" para refletir o grau em que o cereal não era percebido como estando posicionado exclusivamente no nicho de cereal destinado a adultos.

Por fim, os atributos que carregam mais altamente sobre o quarto fator comum são Divertido, (Encharcado), (Simples), Crocante, Fruta, Prazer e (Chato). Os atributos entre parênteses apresentam cargas negativas; por exemplo, quanto mais encharcado o cereal, menor o valor sobre o fator 4. Roberts e Lattin usaram o termo "Interessante" (talvez tenham sentido que o termo "Empolgante" era exaltado demais para cereais matinais). De qualquer modo, isso deve fornecer algumas impressões para o processo de interpretar a solução fatorial e nomear os fatores.

O resultado é que reduzimos um problema envolvendo 25 medidas de atributo a um que envolve somente quatro fatores comuns interpretáveis. Fazendo isso, podemos explicar mais da metade da variação nos dados originais.

5.4 QUESTÕES RELATIVAS À APLICAÇÃO DA ANÁLISE FATORIAL

5.4.1 Posso obter uma solução com fatores correlacionados?

Quando extraímos fatores ou componentes (na análise de componentes principais ou no modelo de fator comum), somos forçados a assumir que haja ortogonalidade (isto é, pressupor por imposição que os fatores comuns não são mutuamente correlacionados) para identificar as soluções. No entanto, uma vez que tenhamos extraído os fatores, não há razão, em teoria (embora haja algumas considerações práticas discutidas a seguir), por que não possamos escolher uma rotação subsequente que permita que os fatores sejam correlacionados. Não limitar que a rotação seja ortogonal pode garantir melhor aproximação à estrutura simples na matriz de cargas transformada e, desta maneira, incrementa-se a capacidade de interpretação da solução. Referimo-nos a essas rotações não ortogonais como *rotações oblíquas*. A intenção por trás de uma rotação oblíqua é alinhar os eixos do fator tão proximamente quanto possível a grupos de variáveis do conjunto original de dados – ou, de outra maneira, facilitar a capacidade de interpretação da configuração resultante –, sejam os fatores resultantes correlacionados ou não. Desse modo, a função objetiva para a rotação oblíqua é similar àquela para a rotação ortogonal (por exemplo, minimizar a soma dos produtos cruzados das cargas fatoriais ao quadrado), mas sem a limitação da ortogonalidade.

Em geral, uma transformação oblíqua conduz a várias questões relativas à interpretação da solução fatorial que não são motivo de preocupação quando se usa uma rotação ortogonal. A questão de princípio é que, com uma rotação oblíqua, faz-se necessária uma distinção entre as chamadas *cargas de estrutura* e *cargas de padrão*. A Figura 5.7 ilustra a distinção. Suponha que tenhamos algum ponto **p** no sistema oblíquo que é obtido pela rotação de $\mathbf{\Lambda}_c$, a solução fatorial original (ortogonal) da nova matriz $\mathbf{A}^* = \mathbf{\Lambda}_c\mathbf{T}^*$, em que $\mathbf{T}^*$ representa a matriz de transformação oblíqua. O ângulo ψ reflete a correlação entre os fatores $\mathbf{a}_1^*$ e $\mathbf{a}_2^*$ no sistema oblíquo.

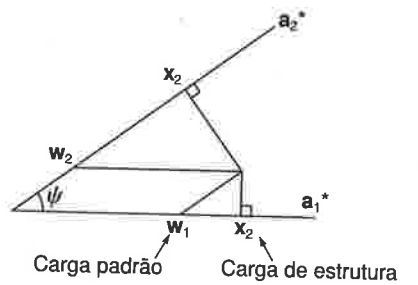

Figura 5.7 Diagrama representando a diferença entre cargas de estrutura e cargas de padrão.

Observe que há dois meios pelos quais podemos registrar a posição do ponto **p** no espaço fatorial oblíquo. Por um lado, é possível registrar a projeção (perpendicular) do ponto **p** nos eixos $\mathbf{a}_1^*$ e $\mathbf{a}_2^*$, dada por x_1 e x_2 na figura. No jargão da análise fatorial, há as *cargas de estrutura*. Por outro lado, podemos registrar as coordenadas do ponto **p** no sistema de referência oblíquo, dadas por w_1 e w_2 na figura. Essas são as *cargas de padrão*. Quando os fatores são ortogonais, a distinção entre cargas de estrutura e cargas de padrão desaparece.

Cargas de estrutura são correlações entre variáveis e fatores (portanto, são limitadas a uma variação entre −1 e +1). Esses são os valores dados na matriz de cargas transformada $\mathbf{A}^* = \mathbf{\Lambda}_c\mathbf{T}^*$. Em geral, as cargas de estrutura não são úteis na interpretação da solução fatorial porque a estrutura simples é disfarçada pelas correlações do fator subjacente. Observe na figura que tanto x_1 quanto x_2 são relativamente amplos devido à correlação entre os dois fatores. Em contraste, as cargas de padrão

são como coeficientes de regressão parcial, no sentido de que levam em consideração a variação explicada por outros fatores. Em outras palavras, elas são coeficientes em uma combinação linear dos fatores oblíquos que nos permitem reconstruir a configuração original. Como resultado, as cargas de padrão não são limitadas a variar entre −1 e +1, mas são geralmente mais úteis na interpretação da solução fatorial. Como apresentado na figura, w_1 é maior que w_2, o que sugere que o ponto **p** está mais proximamente alinhado com o fator um. As cargas de padrão são obtidas das cargas de estrutura através da seguinte expressão:

$$\mathbf{A}^*\mathbf{P} \tag{5.26}$$

onde **P** é uma matriz padrão de direção dos cossenos (apropriadamente normalizados), que capta os ângulos entre os fatores correlacionados no sistema de referência oblíquo. No exemplo bidimensional dado na Figura 5.7, a relação entre cargas de padrão e de estrutura é dada como segue:

$$\begin{bmatrix} w_1 & w_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & \cos\psi \\ \cos\psi & 1 \end{bmatrix} \tag{5.27}$$

Outra advertência a se ter em mente quando se trabalha com as soluções oblíquas: como os fatores podem estar correlacionados, a comunalidade de uma variável não pode ser computada como a soma das cargas ao quadrado na linha i da matriz de cargas fatoriais. Além disso, a variância explicada por um fator não pode ser computada como a soma de cargas ao quadrado na coluna k da matriz de cargas fatoriais.

Agora, ilustramos a rotação oblíqua voltando ao exemplo ilustrativo no teste psicológico. Examinando a matriz de cargas fatoriais não rodada (apresentada na Tabela 5.5), a interpretação da solução de dois fatores é ainda bastante clara: os três primeiros testes carregam principalmente sobre o primeiro fator (que podemos interpretar como aptidão verbal) e os dois últimos carregam principalmente sobre o segundo fator (que podemos interpretar como aptidão quantitativa). A questão é a seguinte: pode-se obter um retrato mais completo da estrutura simples com a aplicação de alguma rotação oblíqua à solução?

Para responder a essa questão, usaremos uma abordagem à transformação oblíqua conhecida como ajuste de matriz alvo. A matriz alvo consiste em uma representação específica da estrutura simples subjacente à solução do problema. Ela pode ser construída sobre base teórica ou pode representar a estrutura de uma matriz de cargas fatoriais interpretada. Nesse caso particular, nossa matriz alvo é uma matriz simples, dada por

$$\mathbf{G} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

Nossa meta é encontrar uma matriz de transformação que rodará a matriz de cargas fatoriais $\mathbf{\Lambda}_c$ com a congruência mais próxima possível da matriz alvo **G**. Se usarmos $\mathbf{T}^*$ para representar a matriz de transformação, nossa meta é escolher $\mathbf{T}^*$ para que a transformação resultante $\mathbf{\Lambda}_c\mathbf{T}^*$ seja tão próxima quanto possível da matriz alvo **G**. A transformação que representa o melhor ajuste de mínimos quadrados a **G** é dada por

$$\mathbf{T}^* = (\mathbf{\Lambda}_c'\mathbf{\Lambda}_c)^{-1}\mathbf{\Lambda}_c'\mathbf{G} \tag{5.28}$$

O procedimento não é mais elaborado do que regredir cada coluna de **G** sobre $\mathbf{\Lambda}_c$ e depois normalizar as colunas resultantes de $\mathbf{T}^*$. Observe que as colunas de $\mathbf{T}^*$ não serão, em geral, ortogonais. Esse tipo de transformação é, às vezes, conhecido como *rotação procrusteana* (veja Hurley e Cattell, 1962).

Os resultados da rotação oblíqua da solução de dois fatores para os dados do teste psicológico são apresentados na Tabela 5.12. As cargas de padrão mostram uma aproximação muito grande à estrutura simples de alvo: as cargas cruzadas no padrão fatorial estão todas próximas de 0,02 a 0,06 em valor absoluto. Portanto, temos uma solução clara na qual o primeiro fator é medido pelos três primeiros testes e o segundo fator é medido pelos dois últimos.

Tabela 5.12 Resultados da rotação oblíqua dos dados do teste psicológico

	Cargas de padrão		Cargas de estrutura	
	Fator 1	Fator 2	Fator 1	Fator 2
PARA	0,8480	−0,0300	0,8687	0,1906
SENT	0,7905	0,0666	0,8344	0,2753
WORD	0,8016	−0,0271	0,8215	0,1815
ADD	0,0499	0,7312	0,2428	0,7696
DOTS	−0,0432	0,7488	0,1510	0,7626

Correlação interfatorial	
	Fator 2
Fator 1	0,2528

A transformação oblíqua resulta em uma correlação entre os dois fatores de 0,25. Essa modesta correlação positiva sugere que os estudantes com alta aptidão verbal (da forma como foi mensurada pelos testes de compreensão de parágrafo, complementação de sentenças e significado das palavras) terão também uma aptidão maior do que a aptidão quantitativa média (e vice-versa). Note que essa correlação entre fatores é refletida nas correlações de estrutura: todas as cargas cruzadas na matriz estão na proximidade de 0,25. Dessa maneira, trocamos um pouco da simplicidade de uma rotação ortogonal por uma solução de fator comprovadamente "mais completa". Em geral, não é má ideia conduzir tanto a rotação ortogonal quanto a oblíqua para determinar qual delas serve melhor aos objetivos da análise.

Uma rotação oblíqua dos dados do cereal RTE não oferece nem uma interpretação diferente nem um quadro substancialmente mais completo da estrutura simples. No espaço oblíquo, todos os quatro fatores são positivamente intercorrelacionados (possivelmente, reflexo de algum efeito halo subjacente), com uma exceção: os fatores 1 e 2 exibem uma pequena correlação negativa. Isso é consistente com a noção de que um cereal percebido como "Saudável" (isto é, alto em fator 1) é mais provavelmente percebido como não "Artificial" (isto é, abaixo da média no fator 2).

5.4.2 Como posso usar os resultados em análises subsequentes?

Com frequência, a análise fatorial não é um fim em si, mas um passo intermediário no caminho para a análise posterior de dados. Por exemplo, Roberts e Lattin usaram a análise fatorial para identificar um número menor de dimensões subjacentes que poderiam usar na construção de um modelo de consideração para os cereais RTE. Para esse tipo de análise subsequente, é necessário que se conheça a localização de cada observação original no reduzido espaço fatorial. Esses valores são chamados de *escores fatoriais*.

Os coeficientes de escore fatorial da solução rodada varimax de quatro fatores para os dados dos cereais RTE são dados na Tabela 5.13. Usamos esses coeficientes para calcular os escores fatoriais de cada uma das 235 observações no conjunto de dados e depois calculamos os escores médios fatoriais de cada uma das 12 marcas classificadas. Esses escores médios (apresentados na Tabela 5.14) são esquematizados na Figura 5.8; para economizar espaço, mostramos apenas dois gráficos de pares (Fator 1 *versus* Fator 2 e Fator 3 *versus* Fator 4).

Os escores fatoriais médios são consistentes com nossas noções anteriores acerca das posições relativas desses cereais. Por exemplo, Weetbix e Vitabrits (que têm o formato de um biscoito de farelo de trigo) são percebidos de maneira semelhante: saudáveis, naturais e desinteressantes. As três diferentes marcas de musli (Cerola, Komplete e Purina) também são agrupadas no espaço do fator.

5.4.3 Como avaliar a validade da estrutura de fator?

Em todos os procedimentos exploratórios, é preciso avaliar o grau com que nossa análise descritiva capta características da população como um todo e não apenas as características de nossa amostra

Tabela 5.13 Coeficientes de escores fatoriais para os dados dos cereais RTE

	Coeficientes de escore fatorial			
	Fator 1	Fator 2	Fator 3	Fator 4
Satisfaz	0,1456	0,0915	0,0228	−0,0420
Natural	0,1370	−0,0142	−0,0112	−0,0417
Fibra	0,2160	0,0542	−0,1280	−0,0746
Doce	0,0425	0,2261	−0,0030	0,0745
Fácil	0,0088	0,0176	0,0663	−0,0115
Sal	0,0301	0,2010	0,0158	−0,1196
Gratificante	0,1101	0,0731	0,1540	−0,0361
Energia	0,0918	0,0752	0,0266	0,0126
Divertido	−0,0024	0,0017	0,1017	0,1316
Crianças	−0,0433	0,0151	0,3735	−0,0789
Encharcado	0,0390	0,0646	0,0402	−0,1995
Econômico	−0,0043	−0,0453	0,1190	−0,0855
Saúde	0,2155	−0,0792	−0,0091	−0,0590
Família	−0,0294	−0,0453	0,2700	−0,0043
Calorias	0,0259	0,1567	−0,0080	−0,0253
Simples	0,0319	0,0699	0,0740	−0,2534
Crocante	−0,0372	−0,0136	0,0864	0,1507
Regular	0,0672	−0,0267	−0,0316	0,0129
Açúcar	0,0444	0,3824	−0,0310	−0,0384
Fruta	0,0376	0,0321	−0,1393	0,1689
Processo	0,0091	0,1020	0,0196	−0,0824
Qualidade	0,0576	−0,0674	0,0531	0,0527
Prazer	−0,0279	0,0329	0,0654	0,2471
Chato	0,0327	0,0680	−0,0178	−0,1752
Nutritivo	0,2004	0,0045	−0,0229	−0,0711

Tabela 5.14 Escores fatoriais médios para os 12 cereais RTE

Marca	Número de observações	Fator 1	Fator 2	Fator 3	Fator 4
All Bran	15	0,3490	−0,3185	−0,8861	−0,3754
Cerola	13	0,5174	0,4994	−0,2332	0,6582
Corn Flakes	27	−0,5541	0,1165	0,5738	0,0241
Just Right	16	−0,0133	0,2278	−0,4196	0,4753
Komplete	14	0,5464	0,2338	−1,0084	0,6024
NutriGrain	24	−0,4255	0,8086	0,5482	0,2852
Purina	18	0,6226	0,6838	−0,4095	0,5754
Rice Bubbles	21	−1,1650	−0,4295	0,6041	0,0627
Special K	23	−0,3058	−0,2142	0,1541	−0,0549
Sustain	12	0,6772	−0,3021	−0,3234	0,8566
Vitabrits	25	0,3649	−0,5921	0,2180	−0,9316
Weetbix	27	0,3266	−0,4120	−0,0713	−0,8837

selecionada. Portanto, necessitamos fazer a pergunta: os resultados da análise são generalizáveis? Como a interpretação da estrutura fatorial subjacente é um objetivo importante da análise fatorial, devemos também perguntar: se eu tivesse escolhido uma amostra diferente da mesma população, teria alcançado uma interpretação substancialmente igual?

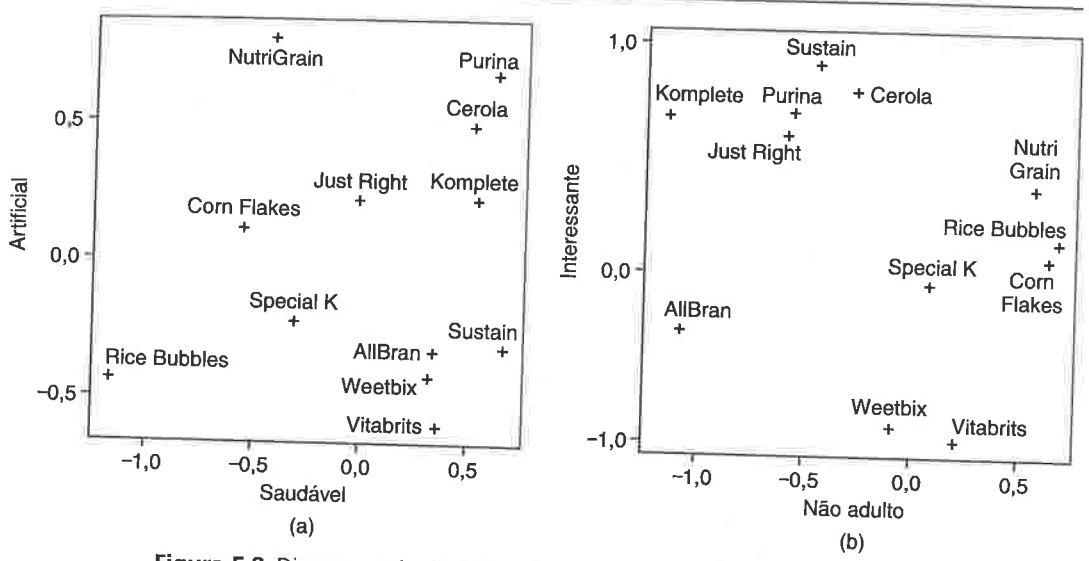

Figura 5.8 Diagramas de dispersão dos escores fatoriais para cereais RTE.

Pode-se responder a essa questão através do uso de uma amostra de teste. Se o tamanho de nossa amostra de corte transversal for suficientemente grande, podemos simplesmente dividir os dados em dois grupos separados de tamanho aproximadamente igual (utilizando atribuição aleatória para assegurar a comparabilidade). Poderíamos, então, efetuar uma análise fatorial em cada um e comparar as matrizes de cargas fatoriais por similaridade de interpretação. Uma comparação baseada em inspeção visual casual, no entanto, é raramente satisfatória. Pode haver centenas de números envolvidos, o que torna difícil tal comparação. Ademais, a tarefa é subjetiva e sujeita ao pensamento desejoso (isto é, o pesquisador pode apresentar a tendência de se apegar às semelhanças entre as matrizes, minimizando os contrastes). Uma abordagem mais rigorosa envolve o uso da análise fatorial confirmatória para testar a estrutura fatorial específica identificada na primeira amostra, utilizando-se os dados da amostra de teste para tanto. Se a estrutura fatorial específica não for rejeitada (isto é, o modelo fornece um ajuste razoável aos dados do teste), podemos então concluir que a interpretação se estende para além da primeira amostra de dados. A análise fatorial confirmatória é apresentada e discutida no Capítulo 6.

Outra abordagem envolve avaliar o impacto da variação da amostragem sobre a estrutura fatorial (via uso de uma amostra de teste ou amostragem bootstrap). Vamos supor que usamos o seguinte procedimento para identificar a estrutura subjacente de uma solução com c fatores comuns: depois de algum tipo de rotação (vamos dizer varimax), identificamos a carga mais alta (em valor absoluto) em cada linha da matriz de cargas fatoriais e associamos a variável àquele fator. Então, interpretamos cada fator comum com base nas variáveis associadas a ele (isto é, ou uma elevada carga positiva ou uma elevada carga negativa). Efetivamente, isso leva-nos a substituir a matriz de cargas fatoriais por uma matriz de zeros e 1 indicativa da estrutura simples subjacente à solução fatorial. O resultado é uma divisão na qual cada variável é associada a um e somente um fator comum. Essa "estruturação" da matriz de cargas fatoriais forma a base da nossa interpretação. Usamos $\mathbf{S}_0$ para representar essa transformação de estrutura simples da matriz de cargas fatoriais para os dados originais. A matriz $\mathbf{S}_0$ para os dados do cereal RTE é apresentada na Tabela 5.15. Note que os atributos têm sido ressequenciados para que os atributos que carregam sobre o mesmo fator sejam listados conjuntamente.

Agora, fazemos outra pergunta: com uma amostra diferente da mesma população, teríamos uma interpretação semelhante? Para responder, pode-se fazer mais uma pergunta: com uma amostra diferente, encontraríamos o mesmo conjunto de variáveis associadas aos fatores comuns, se seguíssemos o procedimento descrito no parágrafo anterior? Com uma amostra de teste, podemos fazer isso, conduzindo uma análise fatorial dos dados, extraindo c fatores (o mesmo número que a solução fatorial nos dados originais) e criando uma matriz de cargas fatoriais de estrutura simples, representada por $\mathbf{S}_1$, seguindo o procedimento descrito no parágrafo anterior.

Tabela 5.15 Matriz zero/um de cargas fatoriais máximas

	Fator 1	Fator 2	Fator 3	Fator 4
Satisfaz	1	0	0	0
Natural	1	0	0	0
Fibra	1	0	0	0
Gratificante	1	0	0	0
Energia	1	0	0	0
Saúde	1	0	0	0
Regular	1	0	0	0
Qualidade	1	0	0	0
Nutritivo	1	0	0	0
Doce	0	1	0	0
Sal	0	1	0	0
Calorias	0	1	0	0
Açúcar	0	1	0	0
Processo	0	1	0	0
Fácil	0	0	1	0
Crianças	0	0	1	0
Econômico	0	0	1	0
Família	0	0	1	0
Divertido	0	0	0	1
Encharcado	0	0	0	1
Simples	0	0	0	1
Crocante	0	0	0	1
Fruta	0	0	0	1
Prazer	0	0	0	1
Chato	0	0	0	1

Temos agora duas divisões, $\mathbf{S}_0$ e $\mathbf{S}_1$, cada uma representando uma atribuição de variáveis aos fatores comuns subjacentes. Um modo de compará-los é criar uma tabulação cruzada $c \times c$, em que a célula (i, j) da tabela representa o número de variáveis associadas ao fator comum i na análise dos dados originais e o fator j na análise dos dados do teste. Usando notação matricial, essa tabulação cruzada é dada pelo produto das matrizes $\mathbf{S}_0'\mathbf{S}_1$. Caso a interpretação da análise fatorial original seja válida, ela deve ser estendida para a amostra de teste. Portanto, é de se esperar uma forte correspondência entre as partições $\mathbf{S}_0$ e $\mathbf{S}_1$, com registros elevados na diagonal da tabulação cruzada e registros discretos ou de valor zero fora da diagonal. Há medidas, como a estatística Rand modificada de Hubert e Arabie (1985), para comparar partições, que podem ser usadas para quantificar a força da correspondência entre as duas estruturas fatoriais.

É possível também avaliar a estabilidade de uma interpretação particular de estrutura fatorial através do uso de amostragem bootstrap. Uma amostra bootstrap é simplesmente uma amostra de tamanho n efetuada com a recolocação das n observações originais. Note que algumas observações não serão incluídas na amostra bootstrap, enquanto outras podem ser incluídas duas ou três vezes. A amostra bootstrap nos fornece uma ideia dos efeitos de variação de amostragem que seriam de se esperar ao assumirmos que o conjunto original de dados é representativo da população subjacente. Para cada amostra bootstrap k, efetuamos uma análise fatorial e formamos uma matriz $\mathbf{S}_k$ de estrutura simples. Depois, avaliamos a estabilidade de nossa interpretação tendo em vista a variação na correspondência entre as matrizes de estrutura $\mathbf{S}_0$ e $\mathbf{S}_k$ por todas as amostras bootstrap k.

Para avaliar a variabilidade através das múltiplas divisões, é mais fácil formar uma tabulação cruzada $p \times p$ de todas as variáveis no modelo, em que a célula (i, j) indica se as variáveis i e j

estão associadas com o mesmo fator. Na notação matricial, essa tabulação cruzada é dada pelo produto de matriz $\mathbf{S}_0\mathbf{S}'_k$. Podemos efetuar um grande número de amostras bootstrap (digamos, K) e então tomar a soma $\Sigma_{k=1}^{K}\ \mathbf{S}_0\mathbf{S}'_k$. A tabulação cruzada resultante contém o número de vezes que cada par de variáveis é identificado com o mesmo fator comum subjacente. Se nossa interpretação for estável (isto é, relativamente não afetada pela variação da amostragem em nossos dados), devemos esperar que as variáveis que carregaram juntas em nossos dados originais também sejam associadas ao mesmo fator comum na maioria ou em todas as nossas amostras bootstrap.

Como ilustração, usamos esse procedimento para avaliar a estabilidade da estrutura simples subjacente à nossa interpretação da solução fatorial para os dados do cereal RTE (apresentados na Tabela 5.15). Efetuamos um total de $K = 25$ amostras bootstrap. A tabulação cruzada dada pela soma dos produtos cruzados da matriz $\Sigma_k\ \mathbf{S}_0\mathbf{S}'_k$ é mostrada na Tabela 5.16. Os padrões na tabela revelam muitos aspectos da solução fatorial, que não podem ser avaliados olhando-se apenas para a matriz de cargas fatoriais dos dados originais. Em primeiro lugar, observamos que a estrutura subjacente aos fatores 1 e 2 ("Saudável" e "Artificial") é extremamente estável, pelo menos mais do que a estrutura subjacente aos fatores 3 e 4 ("Não Adulto" e "Interessante"). Dos 14 atributos associados aos primeiros dois fatores na solução original, todos, menos dois, exibiram exatamente o mesmo padrão de carga da amostra original em todas as 25 amostras bootstrap. Os dois atributos restantes (Gratificante e Processado) são consistentes com a estrutura original em mais de 80% das amostras bootstrap (22 e 21 vezes de 25, respectivamente).

Parece haver mais variabilidade na estrutura subjacente ao terceiro e quarto fatores comuns. Por exemplo, os atributos Crocante e Divertido, que foram associados ao quarto fator ("Interessante") na amostra original, são associados aos atributos Crianças e Família (do terceiro fator comum "Não Adulto") em mais da metade das amostras bootstrap. Além disso, o atributo "Fruta" carrega com as variáveis do quarto fator somente em aproximadamente metade do tempo; no restante do tempo, ele carrega principalmente sobre o primeiro fator (e ocasionalmente sobre o segundo ou o terceiro fator).

Não temos estatística para aplicar aos números da Tabela 5.16 e testar a confiabilidade da estrutura fatorial escolhida. Como muitos dos outros passos da análise fatorial exploratória (por exemplo, escolher o número de fatores subjacentes e interpretar a solução), a avaliação da validade da interpretação fatorial é um julgamento subjetivo (na medida em que ele sequer é praticado!). Mesmo que nossa interpretação substantiva não se modifique, a análise é útil no sentido de revelar a extensão com que a estrutura pode mudar com as variações na amostra.

Pode ser tentador, depois de um exercício de validação como o anterior, "ajustar-se" a solução fatorial. Isso normalmente se dá pelo descarte das variáveis que pareçam menos estáveis, em termos de sua posição na solução fatorial, para que assim se chegue a uma solução fatorial mais completamente interpretável. Portanto, no contexto dos dados do cereal RTE, é uma tentação abandonarmos as variáveis Divertido, Crocante e talvez Fruta para refazer a análise fatorial, esperando uma solução final que evite algumas das potenciais ambiguidades reveladas na solução anterior.

Essa é uma tentação à qual é melhor resistir, em virtude de pelo menos duas razões. Primeira, a retenção seletiva de variáveis é ainda baseada em padrões de variância na amostra, que consiste em uma combinação de variação atribuível à população subjacente mais a variância associada com essa amostra particular. O ajuste pode melhorar o ajuste da amostra de dentro da solução fatorial, mas põe em risco sua capacidade de generalização para além da amostra. No mínimo, se alguma análise fatorial exploratória for utilizada seletivamente para se reduzir o número de variáveis na análise, a solução fatorial subsequente deve então ser reavaliada com o uso de uma amostra de teste.

A segunda razão é que a estrutura simples nem sempre é uma exigência para se chegar a uma boa solução analítica fatorial. Lembre-se de que o objetivo da análise fatorial exploratória é deixar que os dados sugiram a estrutura subjacente. Uma solução que exiba uma estrutura simples é mais fácil de interpretar, mas nem sempre reflete o que é o mundo.

Tabela 5.16 Tabulação cruzada de 25 amostras bootstrap

Satisfaz	25	25	25	22	25	25	25	25	25	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0
Natural	25	25	25	22	25	25	25	25	25	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0
Fibra	25	25	25	22	25	25	25	25	25	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0
Gratificante	22	22	22	25	22	22	22	22	22	0	0	0	0	0	2	0	7	0	0	0	0	0	0	0	0
Energia	25	25	25	22	25	25	25	25	25	0	0	0	0	0	0	1	7	0	0	0	0	0	0	0	0
Saúde	25	25	25	22	25	25	25	25	25	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0
Regular	25	25	25	22	25	25	25	25	25	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0
Qualidade	25	25	25	22	25	25	25	25	25	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0
Nutritivo	25	25	25	22	25	25	25	25	25	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0
Doce	0	0	0	0	0	0	0	0	0	25	25	25	25	25	25	25	14	1	1	13	2	4	4	14	14
Sal	0	0	0	0	0	0	0	0	0	25	25	25	25	25	25	25	14	0	0	12	3	3	3	21	22
Calorias	0	0	0	0	0	0	0	0	0	25	25	25	25	25	25	25	8	2	2	9	6	2	1	22	22
Açúcar	0	0	0	0	0	0	0	0	0	25	25	25	25	25	25	25	14	0	0	12	3	3	3	24	25
Processado	2	2	2	2	2	2	2	2	2	21	21	21	21	21	21	25	14	1	1	13	2	4	4	14	14
Fácil	3	3	3	6	3	3	3	3	3	0	0	0	0	0	0	0	14	1	1	13	2	4	4	14	14
Crianças	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	12	3	3	3	21	22
Econômico	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	8	2	2	9	6	2	1	22	22
Família	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	12	3	3	3	24	25
Divertido	0	0	0	2	0	0	0	0	0	25	25	25	25	25	25	25	14	1	1	13	2	4	4	14	14
Encharcado	0	0	0	0	0	0	0	0	0	25	25	25	25	25	25	25	14	0	0	12	3	3	3	21	22
Simples	0	0	0	0	0	0	0	0	0	25	25	25	25	25	25	25	14	0	0	12	3	3	3	24	25
Crocante	0	0	0	1	0	0	0	0	0	25	25	25	25	25	25	25	14	1	1	13	2	4	4	14	14
Fruta	7	7	7	5	7	7	7	7	7	1	1	1	1	1	1	1	14	1	1	13	2	4	4	14	14
Prazer	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	14	0	0	12	3	3	3	21	22
Chato	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	14	0	0	12	3	3	3	24	25

5.5 RESUMO DA APRENDIZAGEM

- A análise fatorial exploratória, como a análise de componentes principais, é um método que pode ser usado para reduzir-se a dimensionalidade dos dados multivariados e para melhor entendimento dos padrões de associação entre as variáveis. Também é útil para a compreensão dos padrões de correlação nos dados e para identificar os traços ou características subjacentes.
- Na análise fatorial, lançamos mão de um conjunto de pressupostos sobre as fontes de variância em nossos dados; isso é conhecido como modelo de fator comum. O modelo considera que a variação em X_i é atribuível a um número relativamente pequeno de fatores comuns (representados por $\xi_1, \xi_2, \ldots, \xi_c$) e um único fator específico (representado por δ_i). Assim,

$$X_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \ldots + \lambda_{ic}\xi_c + \delta_i$$

  onde c é o número de fatores comuns no modelo (em geral, muito menor que o número de variáveis em **X**).
  - Os *fatores comuns* são assim chamados porque também contribuem para a variação em todas as variáveis em **X**.
  - O fator específico δ pode ser pensado como um termo de erro (é, na realidade, uma combinação de erro de variância e variância específica ao fator).
- Para determinar uma solução para o modelo de fator comum, consideramos os seguintes pressupostos:
  - Os fatores comuns não são correlacionados entre si – isto é, $\text{corr}(\xi_i, \xi_j) = 0$.
  - Os fatores específicos não são correlacionados entre si – isto é, $\text{corr}(\delta_i, \delta_j) = 0$.
  - Os fatores comuns e os fatores específicos não são correlacionados entre si – isto é, $\text{corr}(\xi_i, \delta_j) = 0$.
- Supondo que as variáveis **X** sejam padronizadas [isto é, que $\text{var}(X_i) = 1$ para todos os i], a proporção de variância em X_i atribuível aos fatores comuns, chamada de *comunalidade*, é dada por

$$h_i^2 = \Sigma_i \lambda_{ik}^2 = 1 - \theta_{ii}^2$$

  onde $\theta_{ii}^2 = \text{var}(\delta_i)$ é a variância do fator específico δ_i.
- Se as comunalidades h_i^2 são conhecidas, é possível determinar uma solução para o modelo de fator comum, realizando-se uma análise de componentes principais da matriz de correlação **R** com comunalidades h_i^2 na diagonal. Efetivamente, isso é o mesmo que realizar uma decomposição de valor único na matriz de covariância, após remoção da variância atribuível aos fatores específicos.
  - Se as comunalidades não são conhecidas (que é normalmente o caso), é possível aproximá-las usando-se a correlação múltipla quadrática (SMC).
- A solução para o modelo de fator comum é determinada pela orientação do primeiro fator, de modo que ele explique a maior variância possível; do segundo fator, de modo que ele explique a maior variância possível que está sujeita a não ser correlacionada com o primeiro fator comum, e assim por diante.
  - As correlações entre as variáveis **X** e os fatores comuns Ξ são chamadas *cargas fatoriais* e são representadas por Λ.
- Às vezes, é desejável reorientar-se a solução fatorial para que a matriz de cargas fatoriais exiba uma estrutura mais simples, fazendo com que os fatores sejam mais fáceis de interpretar.
  - Uma solução fatorial exibe estrutura simples quando qualquer variável única é altamente correlacionada com um ou alguns fatores comuns e qualquer fator comum único é altamente correlacionado com somente algumas variáveis.
- A solução fatorial é reorientada através de um processo chamado *rotação*. Dois tipos de rotação são discutidos neste capítulo:
  - A *rotação ortogonal* preserva a ortogonalidade dos eixos (isto é, os fatores rodados permanecem não correlacionados). Dois métodos amplamente usados para efetuar-se a rotação ortogonal são a *rotação varimax* (que tenta construir a estrutura simples focando-se nas colunas da matriz de cargas fatoriais) e a *rotação quartimax* (que enfoca as linhas).

  - A *rotação oblíqua* permite uma correlação entre os fatores rodados. Uma abordagem envolve rotar a solução de modo que ela corresponda à matriz alvo que exibe estrutura simples. Alguns cuidados sobre as propriedades das soluções do fator oblíquo:
  a. Como os fatores comuns são agora correlacionados, as comunalidades não são mais dadas pela soma das cargas fatoriais ao quadrado.
  b. Em uma solução oblíqua, há uma distinção entre *carga de estrutura* (isto é, a correlação entre uma variável e um fator) e uma *carga de padrão* (isto é, a correlação parcial entre uma variável e um fator comum controlando outros fatores comuns).
- Os escores fatoriais representam as posições das observações no espaço do fator comum. Em razão da indeterminação introduzida pelos fatores específicos, esses escores devem ser estimados. Os coeficientes do escore fatorial são dados por

$$\mathbf{B} = \mathbf{R}^{-1}\mathbf{\Lambda}_c$$

  em que $\mathbf{R}^{-1}$ é o inverso da matriz de correlação observada e $\mathbf{\Lambda}_c$ é a matriz das cargas fatoriais para a solução fatorial com c fatores comuns. Os escores fatoriais são então calculados como

$$\Xi = \mathbf{X}_s\mathbf{B}$$

LEITURAS SELECIONADAS

Geral

HARMAN, H. H. *Modern factor analysis.* 3. ed. Chicago: University of Chicago Press, 1976.

MULAIK, S. A. *The foundation of factor analysis.* Nova York: McGraw-Hill, 1972.

RUMMEL, R. J. *Applied factor analysis*. Evanston: Northwestern University Press, 1970.

SPEARMAN, C. "General intelligence objectively determined and measured", *American Journal of Psychology*, v. 15, p. 201-293, 1904.

THURSTONE, L. L. *Multiple factor analysis*. Chicago: University of Chicago Press, 1947.

Rotação

CURETON, E. E.; MULAIK, S.A. "The weighted varimax rotation and the promax rotation", *Psychometrika*, v. 40, p. 183-195, 1975.

HENDRICKSON, A. E.; WHITE, P. O."Promax: a quick method for rotation to oblique simple structure", *British Journal of Mathematical and Statistical Psychology*, v. 17, p. 65-70, 1964.

HURLEY, J. L.; CATTELL, R. B. "The procrustes program: producing direct rotation to test a hypothesized factor structure", *Behavorial Science*, v. 7, p. 258-262, 1962.

KAISER, H. F. "The varimax criterion for analytic rotation in factor analysis", *Psychometrika*, v. 23, p. 187-200, 1958.

NEUHAUS, J. O.; Wrigley, C. "The quartimax method: an analytical approach to orthogonal simple structure", *British Journal of Mathematical and Statistical Psychology*, v. 7, p. 81-91, 1954.

EXERCÍCIOS

5.1 Para fins de ilustração no capítulo, usamos um subconjunto de dados de teste psicológico coletados por Holzinger e Swineford (1939). De fato, Holzinger e Swineford coletaram dados de 145 crianças da 7º e 8º séries em um total de 26 diferentes testes psicológicos. Um conjunto maior de nove testes (incluindo os cinco utilizados no capítulo) é apresentado a seguir:

X_1 Percepção visual
X_2 Cubos
X_3 Losangos
X_4 Compreensão de parágrafos
X_5 Complementação de sentenças
X_6 Significado de palavras
X_7 Adição
X_8 Contagem de pontos
X_9 Letras maiúsculas retas e curvas

A matriz de correlação (mostrada na Tabela 5.17) está disponível no arquivo *PSYCH_TEST*. Analise esses dados utilizando a análise fatorial. Quantos fatores há? Como você os interpretaria? Como esses resultados diferem daqueles baseados nos cinco testes apresentados no capítulo?

5.2 Os dados brutos sobre os cereais prontos para comer coletados por Roberts e Lattin (usados como exemplo de problema no capítulo) estão disponíveis no arquivo *RTE_CEREAL*. O arquivo contém 27 variáveis definidas como segue: a primeira coluna contém a identidade do objeto, a segunda contém o número de identidade do cereal avaliado e as 25 colunas restantes contêm as avaliações de cada um dos 25 atributos. O número de identidade do cereal e a lista ordenada de atributos são dados na Tabela 5.8.

a. Conduza sua própria análise fatorial dos dados do *RTE_CEREAL*. Tente extrair e rotar cinco fatores, determinando se isso faz alguma diferença em sua interpretação dos dados.

Tabela 5.17

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1	1								
X_2	0,318	1							
X_3	0,436	0,419	1						
X_4	0,335	0,234	0,323	1					
X_5	0,304	0,157	0,283	0,722	1				
X_6	0,326	0,195	0,350	0,714	0,685	1			
X_7	0,116	0,057	0,056	0,203	0,246	0,17	1		
X_8	0,314	0,145	0,229	0,095	0,181	0,113	0,585	1	
X_9	0,489	0,239	0,361	0,309	0,345	0,28	0,408	0,512	1

5.3 Sessenta estudantes classificaram 10 marcas de refrigerantes (Coca-Cola, Diet Pepsi, Dr. Pepper, Mt. Dew, Pepsi, Royal Crown, 7Up, Sprite, Diet 7Up, Tab) em relação a quatro atributos (calorias, doçura, saciar a sede e popularidade com os outros) em dois momentos diferentes durante o semestre (setembro e novembro). As variáveis no conjunto de dados são definidas como segue:

X_1 Calorias (setembro)
X_2 Calorias (novembro)
X_3 Doçura (setembro)
X_4 Doçura (novembro)
X_5 Sacia a sede (setembro)
X_6 Sacia a sede (novembro)
X_7 Popularidade (setembro)
X_8 Popularidade (novembro)

A matriz de correlação é dada na Tabela 5.18 (e no arquivo *SOFT_DRINKS*).
Analise os dados usando a análise fatorial. Quantos fatores há? Como você os interpretaria?

5.4 Imagine que você foi contratado como assistente de pesquisa por um professor universitário que está fazendo uma análise empírica. Esse professor possui um conjunto de dados com seis variáveis correlacionadas ($n = 100$) e ele solicitou a você que conduzisse uma análise fatorial desses dados. Você recebeu as seguintes informações:

- As variáveis são simplesmente denominadas X_1 a X_6. O professor quer que você seja informado somente pelos padrões de associação que observa nos dados (e não pelos nomes das variáveis).

Tabela 5.18 Matriz de correlação para os dados de *SOFT_DRINKS*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1							
X_2	0,886	1						
X_3	0,649	0,597	1					
X_4	0,588	0,621	0,649	1				
X_5	0,067	0,034	−0,08	−0,136	1			
X_6	0,054	0,076	−0,075	−0,092	0,542	1		
X_7	0,037	0,029	−0,018	−0,054	0,446	0,274	1	
X_8	0,075	0,102	0,089	0,069	0,225	0,267	0,730	1

- De acordo com o professor, a maioria das pesquisas empíricas publicadas em sua área é baseada no modelo de fator comum.
- O professor pretende usar os dados em uma análise subsequente. Portanto, é importante que ele possa interpretar facilmente os resultados que você apresentar a ele.

A matriz de correlação é apresentada na Tabela 5.19. Os dados brutos estão no arquivo *SIX_VARIABLES*. O professor solicitou-lhe que forneça a ele as respostas às seguintes questões:

a. Quantos fatores você extrairia dessas seis variáveis? Explique claramente as razões por trás de sua decisão.
b. Quanta informação do conjunto original das seis variáveis é explicada por esses fatores?
c. Explique claramente (mas de modo sucinto) a relação entre os fatores escolhidos e as variáveis originais.
d. Usando sua solução fatorial proposta, como você descreveria as diferenças entre as duas primeiras observações na amostra?

5.5 A matriz de correlação para os dados coletados por Roberts (1984) descrita no início do capítulo é apresentada na Tabela 5.20 (e está disponível no arquivo *LUXURY_CAR*):

X_1 Luxo e conforto
X_2 Estilo e modelo
X_3 Confiabilidade
X_4 Economia de combustível
X_5 Segurança
X_6 Custos de manutenção
X_7 Qualidade
X_8 Durabilidade e valor de revenda
X_9 Desempenho na estrada

Analise os dados de Roberts usando o modelo de fator comum. Extraia e faça a rotação de três fatores comuns e descreva suas descobertas (isto é, forneça uma interpretação concisa e perceptiva dos resultados de sua análise). Como os seus resultados diferem dos resultados de dois fatores apresentados no começo do capítulo?

5.6 Uma estudante de doutorado em Pesquisa de Alimento possui um conjunto de dados com 10 variáveis. Ela está preocupada com o fato de 10 variáveis ser um número elevado demais para uma análise subsequente que necessita realizar. Ela pretende reduzir de alguma forma seu conjunto de dados – talvez omitindo algumas variáveis – para não mais que cinco variáveis, mas sem prejuízo extremo das informações originais contidas nos dados.

a. Como você faria para ajudar essa estudante a resolver seu problema? Que abordagem você recomendaria? Qual é a sua solução? Lembre-se de que o resultado final deve ser um conjunto de dados que a estudante possa usar em análise subsequente. Os dados brutos estão disponíveis no arquivo *FOOD_RESEARCH_A*.
b. A estudante tem a sorte de ter separado um conjunto de dados para servir como amostra de teste com o mesmo número de observações. Como você usaria esses dados para aumentar a confiança da estudante na abordagem proposta no item "a"? Os dados do teste estão disponíveis no arquivo *FOOD_RESEARCH_B*.

Tabela 5.19 Matriz de correlação para os dados *SIX_VARIABLES*

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1,000					
X_2	0,849	1,000				
X_3	0,462	0,442	1,000			
X_4	0,416	0,439	0,909	1,000		
X_5	0,409	0,360	0,499	0,501	1,000	
X_6	0,455	0,334	0,478	0,459	0,862	1,000

Tabela 5.20 Matriz de correlação para os dados *LUXURY_CAR*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1	1,000								
X_2	0,591	1,000							
X_3	0,356	0,350	1,000						
X_4	−0,098	0,072	0,38	1,000					
X_5	0,573	0,408	0,382	0,062	1,000				
X_6	0,156	0,232	0,517	0,424	0,303	1,000			
X_7	0,400	0,414	0,611	0,320	0,401	0,479	1,000		
X_8	0,282	0,375	0,512	0,346	0,308	0,463	0,605	1,000	
X_9	0,519	0,484	0,467	0,167	0,455	0,311	0,574	0,557	1,000

5.7 Em janeiro de 1998, 303 estudantes de MBA foram entrevistados a respeito de suas avaliações e preferências sobre 10 diferentes automóveis. Os automóveis, listados em ordem de apresentação na pesquisa, foram BMW 328i, Ford Explorer, Infiniti J30, Jeep Grand Cherokee, Lexus ES300, Chrysler Town & Country, Mercedes C280, Saab 9000, Porsche Boxster e Volvo V90. Cada estudante classificou todos os 10 carros. Para os fins deste exercício, um carro foi selecionado aleatoriamente por cada um dos estudantes, resultando em um tamanho de amostra de 303 avaliações.

Os estudantes classificaram cada carro em relação a 16 atributos. As primeiras oito questões pediam que os estudantes avaliassem o grau em que cada uma das seguintes palavras descrevia um determinado carro (em que 5 = "Extremamente descritiva" e 1 = "Não descreve de maneira alguma"): excitante, confiável, luxuoso, próprio para o ar livre, poderoso, estiloso, confortável e vigoroso. As oito questões seguintes solicitavam que os estudantes classificassem seu nível de concordância com cada uma das seguintes afirmações sobre um determinado carro (em que 5 = "Concordo totalmente" e 1 = "Discordo totalmente"):

"Este carro é gostoso de dirigir".
"Este carro é seguro".
"Este carro tem um grande desempenho".
"Este carro é para a família".
"Este carro é versátil".
"Este carro é esportivo".
"Este carro é um carro de alto padrão".
"Este carro é prático".

Os dados brutos estão disponíveis no arquivo *MBA_CAR*. Há 18 variáveis no arquivo, definidos como segue:

Identidade do estudante (disfarçado)
Identidade do carro:

1 = BMW 328i
2 = Ford Explorer
3 = Infiniti J30
4 = Jeep Grand Cherokee
5 = Lexus ES300
6 = Chrysler Town & Country
7 = Mercedes C280
8 = Saab 9000
9 = Porsche Boxster
10 = Volvo V90

X_1 Excitante
X_2 Confiável

X_3 Luxuoso
X_4 Próprio para o ar livre
X_5 Poderoso
X_6 Estiloso
X_7 Confortável
X_8 Vigoroso
X_9 Gostoso
X_{10} Seguro
X_{11} Desempenho
X_{12} Família
X_{13} Versátil
X_{14} Esportivo
X_{15} *Status*
X_{16} Prático

a. Realize uma análise de fator comum do conjunto de dados. Quantos fatores você reteria? Como você os interpretaria?
b. Salve os escores fatoriais e esquematize os escores fatoriais médios para cada um dos 10 carros avaliados pelos estudantes. O que o gráfico diz sobre as semelhanças e as diferenças dos 10 modelos de carro?

5.8 Westbrook e Oliver (1991) investigaram os tipos de respostas emocionais para a experiência de consumo. Eles realizaram um estudo de uma amostra de proprietários de carros recém-adquiridos. Um total de 125 questionários foi preenchido com respeito às reações emocionais dos respondentes em relação aos seus automóveis. Os autores usaram a medida DES-II de Izard (1997), que contém 10 subescalas que representam a frequência com a qual os sujeitos experimentam cada uma das 10 emoções relacionadas a seguir:

X_1 Interesse
X_2 Alegria
X_3 Surpresa
X_4 Tristeza
X_5 Raiva
X_6 Desgosto
X_7 Desprezo
X_8 Medo
X_9 Vergonha
X_{10} Culpa

A matriz de correlação, reproduzida na Tabela 5.21, está disponível no arquivo *EMOTIONS*. Em sua opinião, qual é a dimensionalidade do espaço psicológico que contém os vários padrões de resposta emocional? Como você interpretaria a(s) dimensão(ões) que encontrar?

Tabela 5.21 Matriz de correlação para os dados *EMOTIONS*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1									
X_2	0,20	1								
X_3	0,08	0,30	1							
X_4	0,13	−0,30	0,37	1						
X_5	0,27	−0,22	0,28	0,80	1					
X_6	0,22	−0,23	0,39	0,84	0,85	1				
X_7	0,22	−0,20	0,45	0,76	0,82	0,92	1			
X_8	0,33	−0,07	0,46	0,67	0,72	0,80	0,83	1		
X_9	0,31	−0,08	0,40	0,55	0,60	0,67	0,77	0,76	1	
X_{10}	0,25	−0,21	0,48	0,74	0,68	0,78	0,78	0,7	0,70	1