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

4

Análise de componentes principais

4.1 INTRODUÇÃO

Atualmente, os pesquisadores defrontam-se com dezenas ou mesmo centenas de variáveis diferentes em suas análises. Com tantas dimensões diferentes, é difícil abarcar ou sequer visualizar os padrões de associação entre elas. O processo é, além do mais, complicado pelo fato de frequentemente haver redundância substancial entre dimensões, o que leva a altos níveis de correlação e multicolinearidade.

A *análise de componentes principais* é um método para reexpressarem-se dados multivariados. Ela permite que o pesquisador reoriente os dados de modo que as primeiras poucas dimensões expliquem o maior número possível de informações disponíveis. Se houver a presença de redundância substancial no conjunto de dados, pode ser possível explicar a maior parte das informações no conjunto original de dados com um número relativamente pequeno de dimensões. Essa redução de dimensões torna a visualização dos dados mais direta e a sua análise subsequente mais administrável. O pesquisador deve decidir quantos componentes principais reter para a análise subsequente, fazendo uma difícil escolha entre simplicidade (isto é, um pequeno número de dimensões é mais fácil de administrar) e completude (isto é, um grande número de dimensões capta uma quantidade maior de informações disponíveis). A solução de componentes principais possui a propriedade de cada componente não estar correlacionado com todos os outros, o que tem a vantagem de eliminar a multicolinearidade quando se usa o resultado em uma análise de dependência (por exemplo, análise de regressão).

Neste livro, fazemos uma distinção entre análise de componentes principais e *análise fatorial exploratória* (que é coberta no Capítulo 5). Embora ambos os tipos de métodos possam ser usados para alcançar os mesmos fins (por exemplo, redução da dimensão dos dados multivariados), os modelos subjacentes são diferentes. A análise fatorial, como veremos, preocupa-se em identificar as fontes subjacentes de variação comum a duas ou mais variáveis (chamadas *fatores comuns*). Um pressuposto explícito nesse modelo de fator comum é que a variação em cada variável observada é atribuível aos fatores comuns subjacentes e a um fator específico (frequentemente interpretável como erro de medida). Em contraste, com componentes principais, estamos preocupados principalmente em reexpressar os dados. Não há modelo de medida subjacente. Cada componente principal é uma combinação linear exata (isto é, soma ponderada) das variáveis originais.

4.1.1 Aplicações potenciais

Redução de dimensão

EXEMPLO *Mapeamento da distribuição de 95 genes na Europa e Oriente Médio (Cavalli-Sforza, Menozzi e Piazzi, 1994).*

Sempre que o tamanho do conjunto de dados torna-se difícil de manejar (em termos do número de variáveis), os componentes principais podem ser úteis na redução dessa dimensionalidade. Trabalhar com menos dimensões torna mais fácil visualizar os dados e identificar padrões interessantes. Um exemplo é a aplicação de componentes principais para o mapeamento genético relatada por Cavalli-Sforza (2000). Antes do desenvolvimento da agricultura, o tamanho das populações era pequeno e podia-se prever uma tendência genética que levava à produção de diferentes frequências de genes de uma região para outra. Uma vez que a migração afeta todos os genes igualmente, deveria ser possível reconstruir as migrações por meio de mapas informativos sobre a frequência de genes; quanto maior o número de genes estudados, mais confiáveis os resultados. Cavalli-Sforza e colegas reuniram informações sobre a frequência de 95 diferentes genes em toda a Europa e o Oriente Médio. Realizaram, então, uma análise de componentes principais para simplificar os dados, tomando combinações ponderadas de informações sobre a frequência de todos os genes para captar o maior número de informações possível. Em um estudo anterior, usando as frequências de 39 genes, eles descobriram que os três primeiros componentes principais explicavam aproximadamente a metade das informações disponíveis nos dados. Desistindo de cerca da metade das informações nos dados, esses pesquisadores foram capazes de reduzir a dimensionalidade do problema em até 92%. Ao invés de tentar interpretar 39 mapas (cada um representando a frequência de um gene diferente), eles precisaram examinar somente três mapas (um para cada componente principal).

Os mapas dos dois primeiros componentes principais do estudo de Cavalli-Sforza são apresentados nas Figuras 4.1 e 4.2. Observe que as áreas mais escuras (ou mais densamente sombreadas) do mapa representam os escores mais altos do componente principal. De acordo com Cavalli-Sforza, o mapa do primeiro componente principal na Figura 4.1 corresponde quase que exatamente aos mapas que

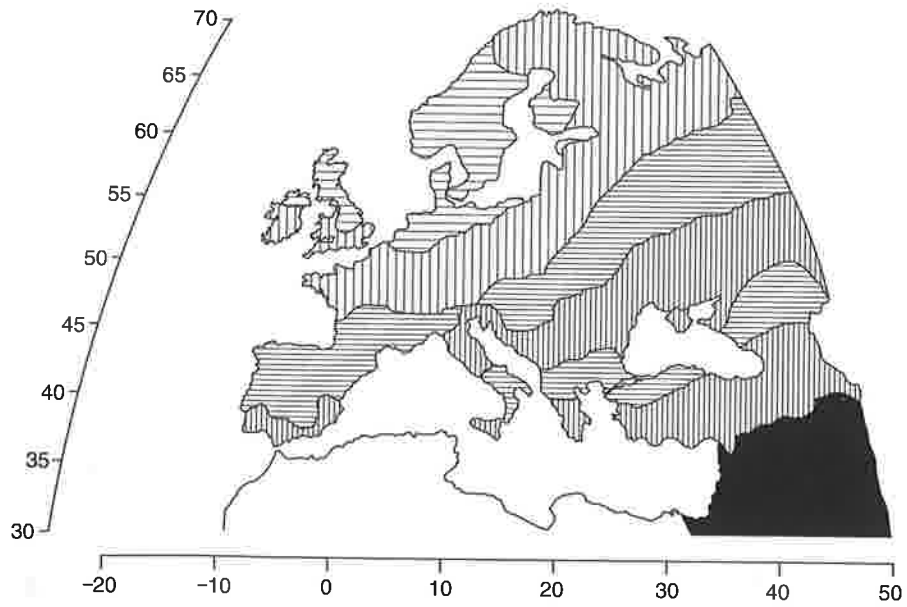

Figura 4.1 Mapa do primeiro componente principal da análise dos 95 genes na Europa. (Fonte: Cavalli-Sforza, L. Luca; Menozzi, Paolo e Alberto Piazzi. *The History and Geography of Human Genes*. Copyright © 1994 da Princeton University Press. Reimpresso com permissão da Princeton University Press.)

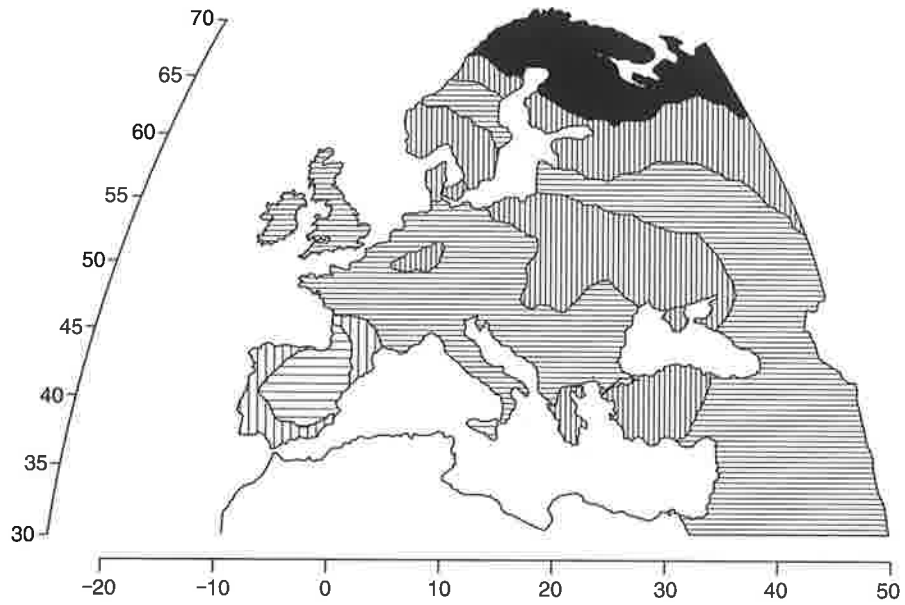

Figura 4.2 Mapa do segundo componente principal da análise dos 95 genes na Europa. (Fonte: Cavalli-Sforza; L. Luca, Menozzi, Paolo e Alberto Piazzi. *The History and Geography of Human Genes*. Copyright © 1994 da Princeton University Press. Reimpresso com permissão da Princeton University Press.)

acompanham a disseminação da agricultura por toda a Europa (obtidos graças a diagramas registrando as datas de chegada de cereais, baseados em estimativas via radiocarbono). A correlação entre o mapa arqueológico e o mapa genético fundamenta a interpretação de uma expansão de fazendeiros do Oriente Médio para a Europa, que se misturaram com a população local de caçadores-coletores.

O mapa do segundo componente principal na Figura 4.2 parece ser mais estritamente correlacionado com a distribuição das línguas por toda a Europa. Cavalli-Sforza formulou a hipótese de que o segundo componente principal pode corresponder a mudanças genéticas ocorridas em virtude de adaptação às latitudes mais frias do norte depois da Idade do Gelo ou, talvez, à chegada de populações que falavam a língua ural do oeste da Sibéria.

EXEMPLO *Medindo a necessidade de cognição do indivíduo (Ofir e Simonson, 2001) COGNIÇÃO*

Outro exemplo de redução da dimensão é a criação de índices de dados de levantamentos ou de dados experimentais. Pesquisadores frequentemente usam muitas questões diferentes para obter uma propriedade ou característica particular dos respondentes de pesquisas. Por exemplo, Ofir e Simonson (2001) usaram uma bateria de 18 questões desenvolvida por Cacioppo, Petty e Kao (1984) para obter a "necessidade de cognição" de cada sujeito (isto é, a extensão com que o sujeito gosta e se envolve no raciocínio e na resolução de problemas). A lista de questões é apresentada na Tabela 4.1 (observe que algumas são "codificadas de modo invertido", de modo que um escore mais alto para a questão reflete uma necessidade mais baixa de cognição); a matriz de correlação é apresentada na Tabela 4.2.

Criar um único índice de "necessidade de cognição" é não somente mais parcimonioso do que usar várias medidas separadas, mas também mais confiável do que usar uma única medida (porque somos menos suscetíveis ao erro de medida associado a uma questão particular do estudo). Uma abordagem possível para se criar tal índice é tomar a soma dos escores (subtrair ao invés de somar as medidas codificadas de modo invertido) de todas as 18 variáveis. Uma abordagem alternativa é a análise de componentes principais: o primeiro componente principal será a combinação linear das 18 questões que captar a quantidade máxima de informações nos dados. Os resultados de uma análise de componentes principais de dados da Tabela 4.2 – apenas para o primeiro componente – são apresentados na Tabela 4.3.

Tabela 4.1 Dezoito itens utilizados na mensuração de um levantamento da "necessidade de cognição" do respondente

Item	Resposta
C_1	Eu prefiro problemas complexos a problemas simples.
C_2	Gosto de ser responsável ao lidar com uma situação que requeira muito raciocínio.
C_3	Raciocinar não é minha ideia de diversão. (R)
C_4	Eu prefiro fazer algo que exija pouco raciocínio do que algo que com certeza desafiaria minhas capacidades de raciocínio. (R)
C_5	Tento prever e evito situações em que haja uma provável chance de eu ter que pensar profundamente sobre alguma coisa. (R)
C_6	Encontro satisfação em debater profundamente por longas horas.
C_7	Penso apenas o necessário. (R)
C_8	Prefiro pensar em pequenos projetos diários do que em projetos em longo prazo. (R)
C_9	Gosto de tarefas que exijam pouca reflexão após eu tê-las aprendido. (R)
C_{10}	A ideia de confiar em meu raciocínio para chegar ao topo me atrai.
C_{11}	Eu realmente gosto de uma tarefa que envolva fornecer novas soluções para os problemas.
C_{12}	Aprender novas maneiras de raciocinar não me anima muito. (R)
C_{13}	Prefiro que minha vida esteja repleta de problemas que devo resolver.
C_{14}	A noção de pensar abstratamente é atraente para mim.
C_{15}	Prefiro tarefas que envolvam raciocínio, sejam difíceis e importantes a aquelas que não exigem muito raciocínio.
C_{16}	Sinto alívio ao invés de satisfação quando termino uma tarefa que exigiu muito esforço mental. (R)
C_{17}	É suficiente para mim que algo termine o trabalho: não me importo como ou por que funciona. (R)
C_{18}	Normalmente, acabo por debater as questões, mesmo que não me afetem pessoalmente.

(R) indica um item de código reverso
Fonte: Cacioppo, Petty e Kao (1984).

Observe que o primeiro componente principal explica quase um terço da variação nas 18 diferentes questões do estudo (32%). Observe também que os pesos são consistentes com a interpretação do índice de "necessidade de cognição" (isto é, todas as medidas codificadas ao inverso possuem pesos negativos). Parece haver alguns desvios na ponderação das medidas: os pesos usados para criar o primeiro componente principal varia de um baixo 0,087 (para a questão C_{18}) a um elevado 0,309 (para a questão C_2).

Componentes principais também podem ser usados em pesquisas de mercado para simplificar a dimensionalidade dos perfis de consumidores com base em dados demográficos e psicográficos. As companhias de pesquisa de mercado, como a A. C. Nielsen e a Information Resources, Inc., mantêm grandes painéis de consumidores (dezenas de milhares de famílias) e os utilizam para acompanhar as atividades do mercado. Cada família fornece uma enorme bateria de informações: padrões demográficos como a idade e a escolaridade de cada membro da família, posse de utensílios, leitura de revistas (e, em alguns casos, até mesmo sobre o hábito de assistir à TV), comportamento de compra em muitas categorias diferentes de produtos e até mesmo medidas de atitudes. Utilizando-se a análise de componente principal, pode-se criar um perfil do consumidor que consiste em um número muito menor de dimensões que, apesar disso, capta a maior parte das informações contidas no conjunto original de dados. O comércio eletrônico (através do qual as empresas agora tornaram-se capazes de acompanhar e obter um retrato altamente detalhado de suas interações com os consumidores, o que leva a uma tecnologia de relação custo-benefício cada vez melhor) é outra área que está "madura" para a simplificação possibilitada pelos componentes principais.

Tabela 4.2 Matriz de correlação de 18 itens que mede a necessidade de cognição (n = 201)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}
C_2	0,445																
C_3	−0,239	−0,454															
C_4	−0,270	−0,375	0,365														
C_5	−0,326	−0,487	0,421	0,558													
C_6	0,248	0,274	−0,187	−0,187	−0,250												
C_7	−0,268	−0,338	0,319	0,345	0,343	−0,235											
C_8	−0,270	−0,355	0,228	0,306	0,340	−0,165	0,314										
C_9	−0,320	−0,328	0,389	0,415	0,310	−0,174	0,221	0,312									
C_{10}	0,289	0,375	−0,411	−0,338	−0,336	0,279	−0,268	−0,299	−0,380								
C_{11}	0,364	0,516	−0,325	−0,405	−0,384	0,297	−0,262	−0,148	−0,338	0,520							
C_{12}	−0,366	−0,415	0,258	0,373	0,555	−0,275	0,310	0,132	0,245	−0,261	−0,392						
C_{13}	0,393	0,382	−0,245	−0,288	−0,322	0,220	−0,172	−0,245	−0,250	0,321	0,418	−0,350					
C_{14}	0,341	0,376	−0,257	−0,250	−0,295	0,291	−0,231	−0,331	−0,145	0,338	0,310	−0,397	0,454				
C_{15}	0,268	0,354	−0,228	−0,185	−0,165	0,186	−0,066	−0,181	−0,177	0,223	0,236	−0,199	0,274	0,230			
C_{16}	−0,280	−0,192	0,166	0,320	0,242	−0,155	0,212	0,150	0,214	−0,147	−0,242	0,192	−0,058	−0,132	−0,071		
C_{17}	−0,273	−0,425	0,282	0,412	0,332	−0,232	0,251	0,364	0,373	−0,226	−0,361	0,284	−0,114	−0,090	−0,174	0,331	
C_{18}	0,126	0,166	−0,140	−0,069	−0,105	0,162	−0,131	−0,042	−0,007	−0,021	0,100	−0,119	0,029	0,209	0,084	−0,194	−0,087

Tabela 4.3 Resultados da análise de componentes principais dos dados da necessidade de cognição na Tabela 4.2

	$\mathbf{u_1}$		$\mathbf{u_1}$
C_1	0,251	C_{10}	0,259
C_2	0,309	C_{11}	0,282
C_3	−0,253	C_{12}	−0,259
C_4	−0,275	C_{13}	0,232
C_5	−0,289	C_{14}	0,230
C_6	0,183	C_{15}	0,169
C_7	−0,227	C_{16}	−0,164
C_8	−0,206	C_{17}	−0,229
C_9	−0,234	C_{18}	0,087

Autovalor $\lambda_1 = 5{,}7794$
Proporção da variância responsável por 32,1%.

EXEMPLO *Criação de uma taxonomia sobre leitura de revista com o propósito de planejar a mídia (Cannon e Williams, 1988).*

Identificando padrões de associação entre as variáveis. Além de reduzir a dimensionalidade, a análise de componentes principais também é útil para a obtenção de um *insight* sobre os padrões de associação. Com frequência, é difícil identificar, apenas pela inspeção de uma grande matriz de correlação, as variáveis que "vão juntas" em virtude de um alto grau de covariação mútua. Essa tarefa é enormemente simplificada examinando-se o relacionamento entre a variável original e os componentes principais. Canton e Williams (1988) usaram os componentes principais para criar a taxonomia da leitura de revistas. Utilizando dados do estudo Target Group Index (TGI), da Axiom Market Research Bureau, que consistiu em 26.818 entrevistas com adultos de 18 anos ou mais, Cannon e Williams usaram os componentes principais para analisar a leitura de revistas por meio de uma lista com 21 diferentes publicações. Seguindo o conjunto precedente de um estudo anterior (Rentz e Reynolds, 1979), eles extraíram sete componentes principais e depois examinaram suas correlações com as variáveis originais (isto é, a leitura de cada revista) para definir os componentes. Essa matriz de correlações, conhecida como *matriz de cargas*, é apresentada na Tabela 4.4. Para facilitar a identificação de cargas relativamente grandes, as correlações acima de 0,40 estão sublinhadas.

Analisando-se cada coluna da tabela, é possível definir cada componente, de acordo com as variáveis às quais está mais fortemente associado. O primeiro componente, por exemplo, exibe as mais fortes correlações (acima de 0,50) com as seguintes revistas: *Better Homes & Garden*, *Family Circle*, *Good Housekeeping*, *Ladies Home Journal*, *McCall's* e *Woman's Day*. As características desse conjunto de revistas (por exemplo, artigos para o lar sobre "receitas" e "como fazer", uma "orientação sobre tarefas domésticas incluindo ideias para manter eficientemente uma casa") levaram Cannon e Williams a nomear o primeiro componente principal como "Operações do Lar". O segundo componente principal – correlacionado acima de 0,60 com a *Cosmopolitan*, *Glamour* e *Vogue* – foi chamado por eles de "Glamour". Cannon e Williams argumentam que uma compreensão dos padrões de leitura dessas revistas levará a uma tomada de decisão mais eficaz por parte dos planejadores de mídia.

Pelo exemplo anterior, o leitor astuto reconhecerá que a interpretação de cada componente principal é mais fácil quando as cargas tomam valores que são elevados (em valor absoluto) ou próximos de zero. Para facilitar a interpretação, é possível rodar os componentes retidos (isto é, mudar sua orientação) para que as cargas da matriz adquiram uma estrutura mais simples. Discutimos os métodos de rotação no Capítulo 5, no contexto da análise fatorial exploratória, mas deve-se ter em mente que os métodos lá discutidos são igualmente aplicáveis à análise de componentes principais.

Tabela 4.4 Matriz de cargas de uma análise dos componentes principais da leitura de 21 revistas diferentes

	Fatores*						
Revistas	F_1	F_2	F_3	F_4	F_5	F_6	F_7
American Home	0,46	0,07	−0,04	−0,06	0,02	0,06	0,26
Better Homes & Garden	0,59	0,14	0,14	0,03	−0,06	−0,05	0,22
Cosmopolitan	0,19	0,63	0,16	0,04	0,04	0,07	−0,13
Family Circle	0,71	0,06	0,06	0,07	−0,06	−0,09	−0,12
Forbes	−0,04	0,00	0,09	0,08	0,02	0,03	0,84
Glamour	0,20	0,69	−0,05	0,04	0,02	0,10	0,03
Good Housekeeping	0,67	0,12	0,06	−0,01	0,08	0,11	0,06
House & Garden	0,45	0,27	0,02	−0,06	0,00	0,06	0,28
Ladies Home Journal	0,66	0,10	0,04	0,00	0,12	0,09	0,00
McCall's	0,65	0,09	0,03	0,01	0,12	0,13	−0,06
National Geographic	0,19	0,00	0,61	0,25	−0,06	−0,22	0,15
Parents	0,17	−0,01	0,14	0,07	0,02	0,84	0,06
Photoplay	0,09	0,10	0,13	−0,05	0,73	−0,10	−0,04
Playboy	−0,18	0,20	0,69	−0,06	−0,01	0,25	−0,05
Psychology Today	0,02	0,22	0,10	0,65	−0,06	0,19	−0,06
Reader's Digest	0,35	−0,11	0,58	0,01	0,08	−0,01	0,05
Redbook	0,48	0,17	0,09	0,09	0,15	0,24	−0,15
Scientific American	0,00	−0,06	0,01	0,81	0,05	−0,08	0,10
True Story	0,05	−0,01	−0,11	0,03	0,71	0,12	0,05
Vogue	0,13	0,67	−0,02	0,06	0,04	−0,15	0,13
Woman's Day	0,70	0,08	0,03	0,07	−0,03	−0,10	−0,12

*Os nomes dos fatores foram atribuídos como segue:
F_1: Operações do Lar
F_2: Glamour
F_3: Cultural/Interesse Literário
F_4: Científico
F_5: Sensacionalismo
F_6: Operações do Lar/Mães Jovens
F_7: Negócios
As cargas sublinhadas representam as revistas consideradas para inclusão em cada fator, quando se desenvolvem os nomes do fator.

4.2 FUNCIONAMENTO DOS COMPONENTES PRINCIPAIS

4.2.1 Intuição

A intuição subjacente à análise de componentes principais é mais bem ilustrada com a ajuda de um exemplo com base em dados artificiais. A Tabela 4.5 mostra a matriz de correlação para três variáveis: X_1, X_2 e X_3. Observe que esses dados foram padronizados (isto é, todos têm uma média igual a zero e um desvio-padrão igual a 1). As implicações de lidar-se com dados não padronizados serão discutidas na Seção 4.4.2.

Tabela 4.5 Matriz de correlação para X_1, X_2 e X_3

	X_1	X_2	X_3
X_1	1,000	0,562	0,704
X_2	0,562	1,000	0,304
X_3	0,704	0,304	1,000
	$\text{var}(X_1) = 1{,}00$	$\text{var}(X_2) = 1{,}00$	$\text{var}(X_3) = 1{,}00$

É difícil representar esses dados tridimensionais em uma página impressa bidimensional de tal forma que o formato da configuração possa ser claramente discernido. A Figura 4.3 mostra uma visão estilizada de um diagrama de dispersão de três direções das variáveis X_1, X_2 e X_3. Em três dimensões, o contorno da configuração tem um formato elíptico (isto é, parecido com uma bola de futebol americano amassada, com um corte transversal oval ao invés de circular). A projeção dessa forma em cada plano das coordenadas resulta em um diagrama de dispersão bidimensional (apresentados como sombras na figura para X_1 *versus* X_2, X_1 *versus* X_3 e X_2 *versus* X_3).

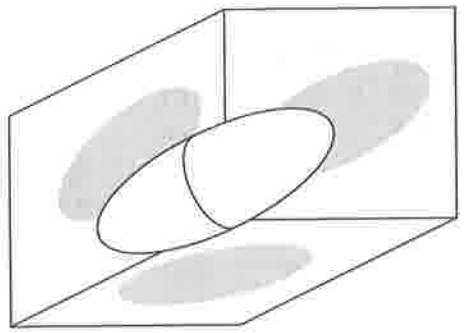

Figura 4.3 Visão estilizada tridimensional do formato de distribuição de X_1, X_2 e X_3. Sombras representam os formatos em duas dimensões.

Um diagrama de dispersão tridimensional dos dados reais é apresentado na Figura 4.4. Os diagramas de dispersão em pares (que poderiam ser projetados nas paredes do gráfico na Figura 4.4) são apresentados na Figura 4.5. Esses gráficos mostram uma correlação positiva entre todos os três pares de variáveis: a mais fraca é entre X_2 e X_3 ($r = 0{,}30$) e a mais forte, entre X_1 e X_3 ($r = 0{,}70$).

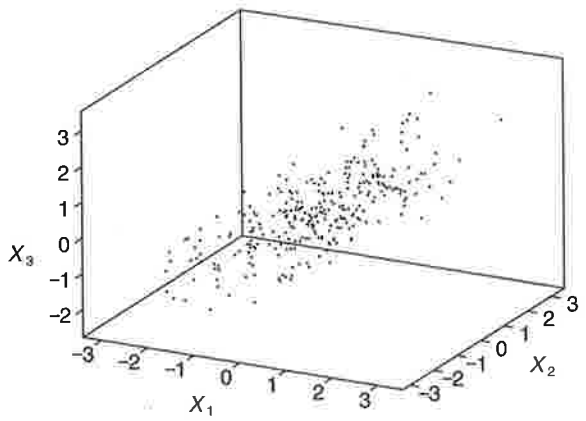

Figura 4.4 Diagrama de dispersão tridimensional dos valores reais de X_1, X_2 e X_3

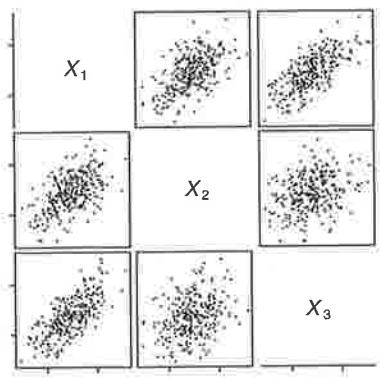

Figura 4.5 Diagrama de dispersão em pares de X_1 *versus* X_2, X_1 *versus* X_3 e X_2 *versus* X_3.

Observando esse padrão de correlação positiva, somos tentados a fazer a seguinte pergunta: seria possível usar uma única dimensão para captar e transmitir a maior parte (se não todas) das informações contidas nas variáveis X_1, X_2 e X_3? Essencialmente, é disso que a análise de componentes principais trata: encontrar um número menor de dimensões que explique uma quantidade suficiente de informações nas variáveis originais.

Dito de maneira mais formal, o primeiro componente principal é a combinação linear de X_1, X_2 e X_3 que exibe variância máxima (lembre-se de que a variância é apenas uma medida de dispersão dos dados). Recorde-se de que uma combinação linear é simplesmente a projeção de todos os pontos no espaço tridimensional em um único eixo. Escolhendo a combinação linear com a variância máxima, estamos explicando o máximo possível de informações contidas em X_1, X_2 e X_3 via uma combinação linear.

A Figura 4.6 mostra a orientação do primeiro componente principal. É o eixo mais longo da configuração de formato em elipse (isto é, a linha que corre ao longo do comprimento da bola de futebol americano, conectando uma extremidade à outra). Se projetarmos cada ponto do dado sobre este eixo, a variável resultante – vamos chamá-la de Z_1 – exibirá a maior variância possível. Em outras palavras, essa é a combinação linear de X_1, X_2 e X_3 que explica a quantidade máxima de variação ou de informação no conjunto original de variáveis. Por enquanto, não vamos entrar em detalhes sobre como conseguimos a combinação linear exata de X_1, X_2 e X_3 que produz Z_1. Suponhamos que $\mathbf{u}_1 = (u_{11}, u_{12}, u_{13})$ represente um vetor de unidade de comprimento orientado ao longo do eixo mais longo do elipsoide, de modo que $\mathbf{z}_1 = \mathbf{X}\,\mathbf{u}_1$, em que $\mathbf{z}_1$ é um vetor n-dimensional consistindo em valores de Z_1 e $\mathbf{X}$ é a matriz $n \times 3$, cujos elementos são os valores de X_1, X_2 e X_3.

A variância da nova variável Z_1 é 2,05, mais que o dobro da variância de qualquer uma das variáveis originais. Podemos pensar na variância de Z_1 como *variância explicada pelo* primeiro componente principal; quanto maior esse valor, tanto mais "informações" dos dados originais estarão contidas nesse único componente. Em razão do padrão geral da correlação positiva nas três variáveis (pense nele como uma redundância – todas as três variáveis originais refletem algo da mesma variação subjacente), uma única combinação linear de X_1, X_2 e X_3 pode explicar toda essa variação compartilhada.

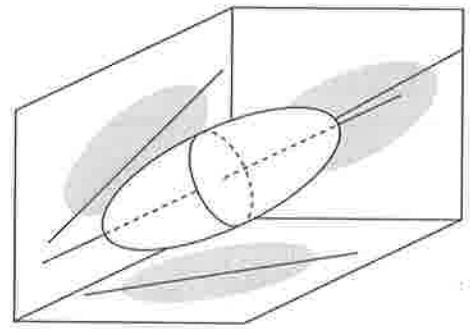

Figura 4.6 Visão tridimensional estilizada identificando o primeiro componente principal.

Embora Z_1 explique uma grande parte da variação em X_1, X_2 e X_3, não explica toda a variação. Muito embora Z_1 nos diga onde, ao longo do comprimento da bola de futebol americano, podemos encontrar cada observação, ela não diz nada sobre se está no meio da bola ou perto de sua margem externa. Pela aplicação sequencial da lógica acima, podemos agora identificar uma segunda combinação linear que explica a maior parte possível da variação residual (isto é, aquela que não é explicada por Z_1).

Como se parecem os dados depois de gerarem as informações em Z_1? Em razão de Z_1 fornecer-nos informações sobre a localização de cada observação ao longo do eixo mais longo do elipsoide, estamos agora preocupados apenas com as dimensões restantes que não são explicadas por Z_1. Essas duas dimensões restantes são dadas pelo plano perpendicular ao eixo central (isto é, a linha orientada na direção do vetor $\mathbf{u}_1$). Essa representação é essencialmente o "corte transversal" da bola de futebol, com todos os pontos projetados sobre o mesmo corte. Uma visão estilizada desse plano (e o formato da configuração dos pontos nesse plano) é apresentada na Figura 4.7; um diagrama de dispersão tridimensional dos dados reais (com Z_1 removido) é apresentado na Figura 4.8.

Podemos obter uma ideia melhor do formato da configuração no corte transversal, projetando-o de volta no espaço da coordenada original e olhando para o diagrama de dispersão em pares de X_1 *versus* X_2, X_1 *versus* X_3 e X_2 *versus* X_3. Esses diagramas de dispersão, apresentados na Figura 4.9, são diretamente análogos aos diagramas de dispersão apresentados na Figura 4.5, com uma importante exceção: a variação nos dados explicada por Z_1 foi removida.

Os diagramas de dispersão da Figura 4.9 revelam algo que não era óbvio no início. A matriz de correlação na Tabela 4.5 mostra um padrão positivo de associação entre todas as três variáveis, então, não é surpreendente que o componente que explicaria a maior parte das informações nos dados *consistisse* em uma combinação linear positiva de X_1, X_2 e X_3. Depois de explicar Z_1, os diagramas de dispersão sugerem que os padrões remanescentes de associação refletem uma forte associação negativa

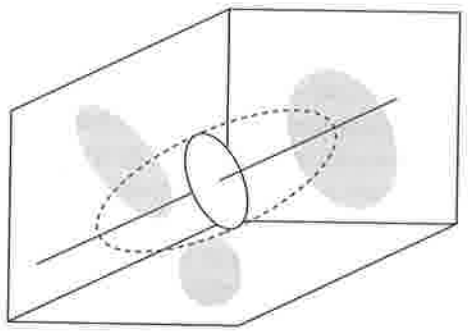

Figura 4.7 Visão tridimensional estilizada após a remoção das informações explicadas pelo primeiro componente principal.

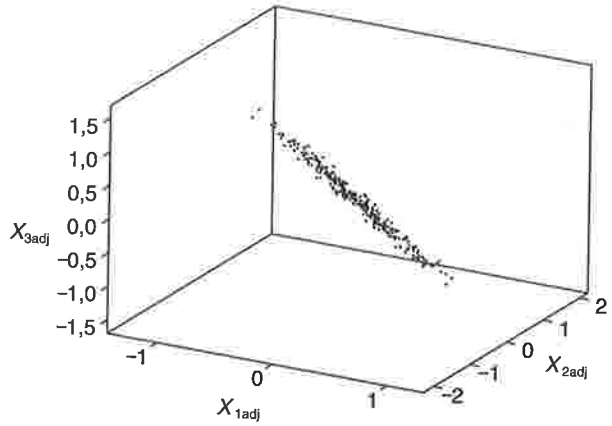

Figura 4.8 Diagrama de dispersão tridimensional de X_1, X_2 e X_3, após a remoção das informações em Z_1.

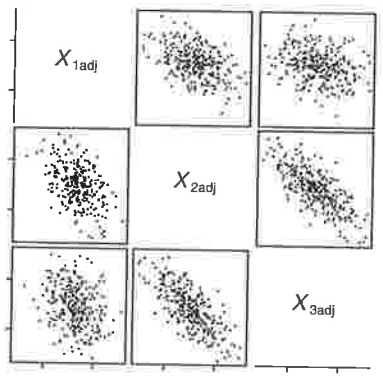

Figura 4.9 Diagramas de dispersão de pares X_1 *versus* X_2, X_1 *versus* X_3 e X_2 *versus* X_3, após remoção das informações em Z_1.

entre X_2 e X_3. Isso seria imediatamente óbvio no exame da matriz de correlação, e é este exatamente o motivo pelo qual os componentes principais provam-se úteis na redução da dimensionalidade e oferecem *insights* aos padrões de associação em grandes conjuntos de dados.

Podemos agora escolher uma segunda combinação linear das três variáveis que explique a quantidade máxima possível da variância restante e não explicada por Z_1. De fato, estamos encontrando o eixo mais longo da configuração em forma de elipse dos pontos sobre o corte transversal bidimensional dos dados originais. A variável resultante – vamos chamá-la de Z_2 – agora explica o valor máximo da variação ainda não explicada pela primeira variável Z_1. Como antes, consideremos $\mathbf{u}_2$ um vetor de unidade de comprimento orientado na direção do eixo longo do corte transversal, de modo que $\mathbf{z}_2 = \mathbf{X}\,\mathbf{u}_2$, em que $\mathbf{z}_2$ é um vetor n-dimensional com os valores Z_2. Observe que $\mathbf{u}_2$ é ortogonal a $\mathbf{u}_1$ (isto é, os vetores são perpendiculares um ao outro) e que a nova variável Z_2 não é correlacionada com Z_1. Isso se deve ao fato de termos restringido nossa escolha de $\mathbf{z}_2$ ao plano perpendicular a $\mathbf{z}_1$. Portanto, não há como as informações que temos da posição dos pontos no corte transversal do elipsoide nos dizerem qualquer coisa sobre onde eles caem no seu eixo longo.

A variância de Z_2 (igual a 0,72) é consideravelmente menor que a de Z_1. Isso ocorre propositalmente, pois deliberadamente escolhemos Z_1 para explicar o máximo possível de variação em X_1, X_2 e X_3. No entanto, também observamos que a variância explicada por Z_1 é menor que a de qualquer uma das variáveis originais (todas elas foram padronizadas com variância igual a 1).

A variação residual restante – a qual não é explicada nem por Z_1 nem por Z_2 – é inteiramente explicada por uma terceira variável Z_3. Isso é devido ao fato de termos começado com um problema tridimensional, e após a escolha de $\mathbf{u}_1$ e $\mathbf{u}_2$ para serem mutuamente ortogonais, há somente um meio

restante que se pode escolher para que $\mathbf{u}_3$ seja ortogonal a $\mathbf{u}_1$ e $\mathbf{u}_2$. O vetor final, dado por $\mathbf{z}_3 = \mathbf{X}\mathbf{u}_3$, define o terceiro componente principal; ele explica a variação restante e é ortogonal tanto a $\mathbf{z}_2$ quanto a $\mathbf{z}_1$. Observe que a variância de Z_3 é somente 0,23, o que sugere que a maior parte das informações de X_1, X_2 e X_3 pode ser explicada pelos dois primeiros componentes apenas, Z_1 e Z_2.

As três novas variáveis estão resumidas na Tabela 4.6. Uma representação estilizada do formato da distribuição de Z_1, Z_2 e Z_3 é apresentada na Figura 4.10. O diagrama de dispersão dos valores reais na Figura 4.11 mostra que a configuração original dos pontos não mudou, apenas a sua orientação. A transformação $\mathbf{Z} = \mathbf{XU}$ serviu simplesmente para rodar os eixos do diagrama de dispersão original, enquanto preserva sua ortogonalidade (devido à ortogonalidade mútua de $\mathbf{u}_1$, $\mathbf{u}_2$ e $\mathbf{u}_3$). Recorde-se, do Capítulo 2, que essa forma de multiplicação de matriz serve simplesmente para reorientar os vetores de base.

Tabela 4.6 Matriz de correlação para Z_1, Z_2 e Z_3

	Z_1	Z_2	Z_3
Z_1	1,000	0,000	0,000
Z_2	0,000	1,000	0,000
Z_3	0,000	0,000	1,000
	$\text{var}(Z_1) = 2{,}05$	$\text{var}(Z_2) = 0{,}72$	$\text{var}(Z_3) = 0{,}23$

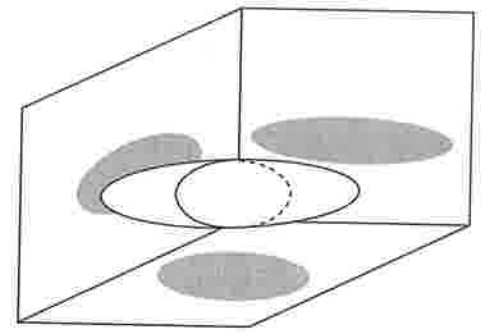

Figura 4.10 Visão tridimensional estilizada da forma de distribuição de Z_1, Z_2 e Z_3.

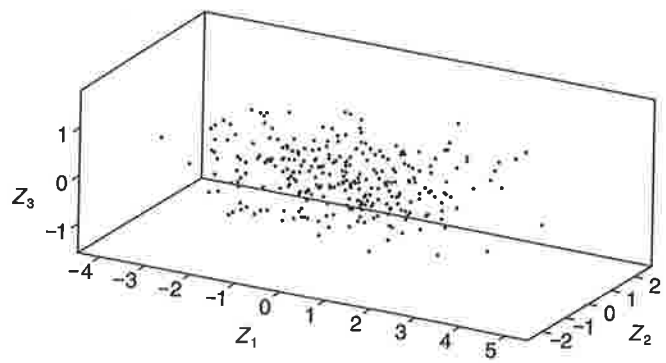

Figura 4.11 Diagrama de dispersão tridimensional dos valores reais de Z_1, Z_2 e Z_3.

A solução possui as seguintes propriedades: as novas variáveis Z_1, Z_2 e Z_3 são mutuamente não correlacionadas, e cada nova variável é escolhida para explicar a máxima quantidade possível de variação ainda não explicada pelas variáveis anteriormente escolhidas. Embora leve as três variáveis novas em conta para explicar todas as informações contidas em X_1, X_2 e X_3, observe que apenas Z_1 e

Z_2 juntas explicam uma variância de 2,77 (comparada a uma variância de 0,23 fornecida por Z_3). O verdadeiro valor da redução de dimensões é difícil de ilustrar em um exemplo de problema com três variáveis, mas é fácil imaginar os benefícios proporcionados pela redução de 20 variáveis para quatro ou cinco dimensões e ainda explicar uma grande porcentagem de variação nas variáveis originais.

Normalmente, é instrutivo examinar a relação entre os componentes principais Z e as variáveis originais X. Uma maneira de se fazer isso é olhar para a matriz de correlação, apresentada na Tabela 4.7. Note que, embora Z_1, Z_2 e Z_3 não sejam mutuamente correlacionadas, elas são relacionadas a X_1, X_2 e X_3 em padrões consistentes com a análise precedente.

Tabela 4.7 Cargas do componente principal. Correlações entre os componentes principais e os dados originais

	Cargas do Componente Principal		
	Z_1	Z_2	Z_3
X_1	0,9279	−0,0798	−0,3641
X_2	0,7255	0,6696	0,1590
X_3	0,8222	−0,5008	0,2706

Essas correlações, às vezes conhecidas como *cargas dos componentes principais*, podem ajudar na interpretação dos componentes principais. Por exemplo, vemos que Z_1 é correlacionada positivamente a X_1, X_2 e X_3; isso é consistente com a ideia expressa acima de que Z_1 reflete um componente de variância compartilhada subjacente a todas as variáveis originais. Se houvesse atributos de produto, Z_1 poderia estar captando um *efeito halo* – isto é, uma fonte de viés positivo atribuível a alguma qualidade de produto intangível, como a imagem da marca. Z_1 poderia também estar captando algum tipo de *efeito de método* (por exemplo, se X_1, X_2 e X_3 fossem todas medidas utilizando-se o mesmo instrumento). O segundo componente principal é mais interessante: Z_2 é positivamente relacionado a X_3 e negativamente relacionado a X_2 (a correlação com X_1 é próxima demais de zero para influenciar significativamente a interpretação). Assim, podemos concluir que Z_2 capta uma dimensão que diferencia X_2 de X_3.

As cargas do componente principal são também úteis para nos dizer quanto da variância em cada uma das variáveis originais X é explicado pelos componentes principais. Lembre-se, do Capítulo 3, de que o quadrado do coeficiente de correlação, que representamos por R^2, conta-nos que proporção da variância em uma variável é responsável ou explicada por outra variável. Desta maneira, olhando para a correlação entre X_1 e Z_1 na matriz de cargas de fator, podemos dizer que a proporção da variância em X_1 que é explicada pelo primeiro componente principal é $0{,}93^2 = 0{,}86$, ou 86%. Da mesma maneira, a variância em X_1 explicada pelo segundo componente principal é menor que 1% $(-0{,}08)^2$. Uma vez que Z_1 e Z_2 não são correlacionados, podemos somar os valores de R^2 para determinar o valor da variância em X_1 explicada pelos primeiros dois componentes principais. Nesse caso, a proporção é $0{,}93^2 + (-0{,}08)^2 = 0{,}87$. Nessa ilustração particular, notamos que os dois primeiros componentes principais explicam quase 90% (ou mais) da variância em cada variável original X_1, X_2 e X_3.

4.2.2 Mecânica

Para mostrar como são determinados os vetores do componente principal, apresentamos agora um tratamento mais matemático do problema. Em geral, em problemas de componentes principais, nosso objetivo é encontrar uma combinação linear das variáveis originais $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ com variância máxima. Para os fins dessa derivação, digamos que $\mathbf{X}$ seja padronizado (isto é, cada variável é normalizada para média zero e variância unitária). Se representarmos a combinação linear pelo vetor $\mathbf{u} = (u_1, u_2, \ldots, u_p)'$, nossa meta é escolher $\mathbf{u}$ para maximizar a variância dos elementos de $\mathbf{z} = \mathbf{Xu}$, que pode ser escrita como segue:

$$\text{var}(\mathbf{z}) = \frac{1}{(n-1)}\mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u} \tag{4.1}$$

Como $\mathbf{X}$ é padronizado, o termo $1/(n-1)\mathbf{X}'\mathbf{X}$ é apenas a matriz de correlação da amostra $\mathbf{R}$. Substituindo, temos

$$\text{var}(\mathbf{z}) = \mathbf{u}'\mathbf{R}\mathbf{u} \tag{4.2}$$

Observe que a instrução de maximização de problema ainda não é suficiente, porque podemos escolher os componentes de $\mathbf{u}$ (isto é, o comprimento do vetor) como sendo arbitrariamente elevados e, desse modo, levar var($\mathbf{z}$) ao infinito. Por isso, impomos uma limitação da unidade de comprimento do vetor, que é estabelecido como $\mathbf{u}'\mathbf{u} = 1$. O problema assim modificado pode agora ser apresentado da seguinte maneira:

$$\begin{gathered}\text{Escolha } \mathbf{u} \text{ para maximizar } \mathbf{u}'\mathbf{R}\mathbf{u} \\ \text{sujeito à limitação } \mathbf{u}'\mathbf{u} = 1\end{gathered} \tag{4.3}$$

Podemos resolver esse problema de otimização limitada formando o Lagrangiano, tomando a primeira derivada, atribuindo-lhe valor zero e resolvendo. O Lagrangiano é dado por

$$\text{L} = \mathbf{u}'\mathbf{R}\mathbf{u} - \lambda(\mathbf{u}'\mathbf{u} - 1) \tag{4.4}$$

onde λ é chamado de *multiplicador Lagrangiano*. Observe que λ pode ser escolhido para penalizar a função objetiva se a igualdade limitada ($\mathbf{u}'\mathbf{u} = 1$) não for satisfeita. Tomando a derivada de L em relação aos elementos de $\mathbf{u}$, tem-se

$$\frac{\partial \text{L}}{\partial \mathbf{u}} = 2\mathbf{R}\mathbf{u} - 2\lambda\mathbf{u} \tag{4.5}$$

Estabelecer a Equação (4.5) igual a zero e resolvê-la requer que as seguintes condições sejam satisfeitas:

$$\mathbf{R}\mathbf{u} = \lambda\mathbf{u} \quad \text{ou} \quad (\mathbf{R} - \lambda\mathbf{I})\mathbf{u} = 0 \tag{4.6}$$

A equação acima possui uma estrutura especial: ela é conhecida como *problema autovalor – autovetor*. O vetor $\mathbf{u}$ é chamado de *autovetor* (às vezes, de *vetor característico*) e o escalar λ é chamado de *autovalor*. Dado que a matriz $\mathbf{R}$ seja de posições completas (isto é, estabelecendo-se que não há multicolinearidade perfeita entre as variáveis observadas $\mathbf{X}$), a solução consistirá em p autovalores positivos e autovetores associados.

Resolver um problema autovalor–autovetor analiticamente é uma empreitada complicada porque envolve encontrar as raízes de uma equação polinomial de ordem p-ésima. Felizmente, várias rotinas de computador que fornecem soluções numéricas estão disponíveis. Usamos programas de computador (tal como o PROC IML em SAS) para resolver a autoestrutura dos dados na Tabela 4.1. Os resultados são apresentados na Tabela 4.8. É fácil verificar que, após multiplicar os dados originais de $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]$ pelos autovetores $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]$, forma-se a matriz dos escores dos componentes principais dimensionados $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3]$ apresentada na Tabela 4.8.

Tabela 4.8 Autovalores (λ) e autovetores ($\mathbf{u}$) de X_1, X_2 e X_3

$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$
0,65	0,09	−0,76
0,51	0,80	0,33
0,57	−0,59	0,56
$\lambda_1 = 2{,}05$	$\lambda_2 = 0{,}72$	$\lambda_3 = 0{,}23$

Variância explicada pelos componentes principais

Em primeiro lugar, é interessante observar que os autovalores λ_1, λ_2 e λ_3 são exatamente iguais às variâncias para os três componentes principais relatados na Tabela 4.6. Poderia ser coincidência? A resposta é não. Para saber por quê, voltamos à função objetiva do problema dos componentes principais.

Vale lembrar que nosso objetivo é maximizar a variância de **z**, dada por **u′Ru**. Observe que, para que as condições de primeira ordem se mantenham, devemos ter $\mathbf{Ru} = \lambda\mathbf{u}$. Substituindo $\lambda\mathbf{u}$ por **Ru**, temos

$$\text{var}(\mathbf{z}) = \mathbf{u}'\mathbf{Ru} = \mathbf{u}'\lambda\mathbf{u} = \lambda\mathbf{u}'\mathbf{u} = \lambda \tag{4.7}$$

porque $\mathbf{u}'\mathbf{u} = 1$. Portanto, o autovalor λ é exatamente a variância do componente principal associado. Usaremos **D** para representar a matriz de covariância diagonal dos componentes principais; nesse exemplo, $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2 \text{ e } \lambda_3)$.

Como os componentes principais são mutuamente não correlacionados, a variância da soma é simplesmente a soma das variâncias dos componentes individuais. Neste exemplo, $\text{var}(\mathbf{z}_1 + \mathbf{z}_2 + \mathbf{z}_3) = \lambda_1 + \lambda_2 + \lambda_3 = 3{,}0$. É uma coincidência que a soma das variâncias dos componentes principais seja a mesma do número de variáveis no problema? Novamente, a resposta é não. A variância total por todos os componentes principais é dada por traço(**D**), em que o traço é uma função da matriz que simplesmente dá a soma dos elementos da diagonal. Já que **D** = **U′RU**, como derivado anteriormente, podemos também representar a variância total como traço(**U′RU**), em que **U′**, **R** e **U** são todos matrizes quadradas.

Uma propriedade da função traço é que o valor é o mesmo para todas as permutações cíclicas do produto da matriz. Em outras palavras, para o produto de matriz **ABC**, temos traço(**ABC**) = traço(**CAB**) = traço (**BCA**). Portanto, traço(**U′RU**) = traço(**RUU′**). Como os autovetores que compõem a matriz **U** são mutuamente ortogonais (e de posições completas), sabemos que **UU′** = **I**, o que significa que a soma das variâncias de todos os componentes principais se reduz a traço(**RUU′**) = traço(**RI**) = traço(**R**). Esta é simplesmente a soma daqueles que se situam ao longo da diagonal da matriz de correlação.

Essa propriedade da solução dos componentes principais é particularmente útil quando se trata de expressar a quantidade de variação explicada por algum subconjunto dos componentes principais. Por esse simples exemplo ilustrativo, podemos dizer que os dois primeiros componentes principais, Z_1 e Z_2, explicam $(2{,}05 + 0{,}72)/3 = 92\%$ da variação no conjunto inteiro de dados. Essa forma de expressão se tornará especialmente útil quando se tratar de decidir quantos componentes principais será necessário reter para a análise subsequente.

Cargas do componente principal

Outro subproduto útil da solução de componentes principais é a matriz de correlação do escore do componente principal (**Z**) com os dados originais (**X**). Ela ajuda a interpretar **Z**, se soubermos o padrão de relacionamento com os dados **X** existentes. A matriz de correlação é dada pela seguinte expressão:

$$\text{corr}(\mathbf{X}, \mathbf{Z}) = \frac{1}{(n-1)}\mathbf{X}'\mathbf{Z}_s \tag{4.8}$$

onde **X** é a matriz dos dados originais (por convenção, padronizados) e $\mathbf{Z}_s = \mathbf{ZD}^{-1/2}$ é a matriz dos componentes principais padronizados. Substituindo **XU** por **Z** na expressão acima, temos

$$\text{corr}(\mathbf{X}, \mathbf{Z}) = \frac{1}{(n-1)}\mathbf{X}'\mathbf{XUD}^{-1/2} \tag{4.9}$$

Uma vez que $1/(n-1)\mathbf{X}'\mathbf{X}$ é simplesmente a matriz **R** de correlação da amostra e sabemos (da solução dos componentes principais) que **R** pode ser reexpresso como **UDU′**, podemos substituir na expressão acima para obter:

$$\text{corr}(\mathbf{X}, \mathbf{Z}) = (\mathbf{UDU}')\mathbf{UD}^{-1/2} = \mathbf{UD}^{1/2} \tag{4.10}$$

porque **U′U** = **I**. Assim, temos como resultado que as correlações entre os componentes principais Z e as variáveis originais X, às quais iremos nos referir como as *cargas do componente principal* e representaremos como

$$\mathbf{F} = \mathbf{UD}^{1/2} \tag{4.11}$$

são determinadas por um escalonamento simples dos autovetores subjacentes.

Das cargas do componente principal, podemos determinar a quantidade de variância em cada variável original explicada por qualquer número dos componentes principais. A expressão geral para a variância explicada na variável X_i pelo primeiro componente principal c é

$$\sum_{j=1}^{c} f_{ij}^2 \tag{4.12}$$

onde c é o número dos componentes principais retidos e f_{ij} é a correlação entre X_i e Z_j da matriz **F**. Observe que, quando $c = p$ (isto é, todos os componentes principais estão retidos), $\sum_j f_{ij}^2 = 1{,}0$.

Chegando à solução dos componentes principais

Para obter a matriz padronizada dos componentes principais (representada por $\mathbf{Z}_s$), simplesmente multiplicamos depois $\mathbf{Z} = \mathbf{XU}$ por $\mathbf{D}^{-1/2}$, em que **D** é a matriz diagonal da variância – covariância de **Z**, que resulta em

$$\mathbf{Z}_s = \mathbf{XUD}^{-1/2} \tag{4.13}$$

Com um pouco de álgebra, podemos reescrever a equação anterior para expressar **X** como uma função de $\mathbf{Z}_s$. Multiplicando depois por $\mathbf{D}^{1/2}$ e então por **U′** (observe que $\mathbf{UU}' = \mathbf{U}'\mathbf{U} = \mathbf{I}$ por causa da ortogonalidade mútua de $\mathbf{u}_1$, $\mathbf{u}_2$ e $\mathbf{u}_3$), temos

$$\mathbf{X} = \mathbf{Z}_s\mathbf{D}^{1/2}\mathbf{U}' \tag{4.14}$$

O que isso revela é que qualquer matriz de dados **X** pode ser expressa como produto da matriz de três matrizes mais simples. $\mathbf{Z}_s$ é uma matriz de variáveis não correlacionadas (cada uma com variância unitária). $\mathbf{D}^{1/2}$ é uma matriz diagonal que executa uma transformação extensora (essencialmente "despadronizando" $\mathbf{Z}_s$, multiplicando-a pelos desvios-padrão de **Z**) e **U′** é uma matriz de transformação que realiza uma rotação ortogonal. Esse modo de expressar **X** é conhecido como uma *decomposição em valores singulares* (SVD – singular value decomposition). Assim, pode-se realizar uma análise de componentes principais através de uma análise do autovalor (algumas vezes, chamada de *decomposição espectral*) da matriz **R** ou uma decomposição de valor único da matriz **X**.

4.3 EXEMPLO DE PROBLEMA: PRODUTO ESTADUAL BRUTO

Após apresentarmos a intuição subjacente e alguns fundamentos analíticos da análise de componentes principais usando uma situação empírica artificial, oferecemos agora um exemplo ilustrativo que utiliza um conjunto de dados reais.

4.3.1 Dados

Os dados (expressos em milhões de dólares) são valores do produto estadual bruto (GSP – Gross State Product) de cada uma das 13 diferentes áreas de atividade econômica em 1996:

EXEMPLO *Dados sobre o produto estadual bruto (GSP) para 13 áreas de atividade econômica, GSP_RAW e GSP_SHARE*

1. agricultura, silvicultura e pesca;
2. mineração;
3. construção civil;
4. indústria (bens duráveis);
5. indústria (bens não duráveis ou bens de consumo);
6. transporte;
7. comunicações;
8. eletricidade, gás e saneamento;
9. comércio atacadista;

Tabela 4.9 Correlação entre 13 medidas diferentes de atividades econômicas (dados brutos em milhões de dólares)

	AGRI-CULTURA	*MINERAÇÃO*	*CONS-TRUÇÃO CIVIL*	*IND. BENS DURÁVEIS*	*IND. BENS NÃO DURÁVEIS*	*TRANS-PORTE*	*COMUNI-CAÇÃO*	*SERVIÇOS URBANOS*	*COMÉRCIO ATACADISTA*	*COMÉRCIO VAREJISTA*	*FIDUCIÁRIO*	*SERVIÇOS*
MINERAÇÃO	0,248											
CONSTRUÇÃO CIVIL	0,804	0,415										
IND. BENS DURÁVEIS	0,749	0,262	0,873									
IND. BENS NÃO DURÁVEIS	0,662	0,391	0,879	0,841								
TRANSPORTE	0,813	0,458	0,976	0,862	0,896							
COMUNICAÇÃO	0,716	0,356	0,930	0,742	0,835	0,932						
SERVIÇOS URBANOS	0,674	0,525	0,951	0,837	0,907	0,942	0,903					
COMÉRCIO ATACADISTA	0,815	0,346	0,977	0,877	0,883	0,973	0,953	0,930				
COMÉRCIO VAREJISTA	0,848	0,343	0,984	0,886	0,862	0,965	0,929	0,920	0,984			
FIDUCIÁRIO	0,740	0,190	0,902	0,790	0,793	0,876	0,946	0,851	0,944	0,933		
SERVIÇOS	0,804	0,269	0,955	0,846	0,829	0,932	0,950	0,894	0,979	0,978	0,983	
GOVERNO	0,814	0,344	0,972	0,843	0,868	0,949	0,957	0,915	0,974	0,982	0,949	0,978

10. comércio varejista;
11. fiduciário, seguros e setor imobiliário;
12. serviços; e
13. governo.

Há 50 observações no conjunto de dados, uma para cada um dos 50 estados dos Estados Unidos. A matriz de correlação para os dados brutos é apresentada na Tabela 4.9.

Os registros na matriz de correlação tornam claro que há uma quantidade considerável de correlação nesses dados. Uma boa quantidade de correlação positiva não é surpreendente, porque são de se esperar altos níveis de GSP em todas as indústrias dos estados maiores e mais prósperos (e níveis abaixo da média de GSP nos estados menores e menos populosos). Para esse exemplo particular, pode ser verdade que uma boa quantidade da covariação subjacente possa ser explicada por um único componente principal que capta o tamanho de cada estado (e com ele a magnitude da atividade econômica).

Podemos também considerar que vale a pena examinar a *participação* do GSP atribuível a cada área de atividade econômica. Por exemplo, será que em alguns estados advém da mineração e da agricultura uma maior proporção de seu GSP? Certas áreas da atividade econômica (por exemplo, comércio varejista e atacadista) são positivamente correlacionadas? Podemos olhar para esses números tomando os dados originais e dividindo cada variável pelo GSP total de cada estado. O resultado é que agora cada variável é escalonada como uma proporção entre zero e um (efetivamente, removemos as informações sobre o tamanho do estado). A matriz de correlação para os dados do GSP (expressos em participações) é apresentada na Tabela 4.10.

Fica claro a partir da matriz de correlação da Tabela 4.10 que os registros não são tão amplamente correlacionados como na Tabela 4.9. Embora alguns sejam tão elevados como 0,5 ou 0,6, a maioria varia entre 0,0 e 0,3 e muitos são negativos. Isso sugere que um único componente principal pode não ser suficiente para explicar todas as variações subjacentes. Caso se conclua que diferentes estados se "especializam" principalmente em diferentes combinações de atividades econômicas, podemos confiar no componente principal para nos ajudar a identificar esses padrões subjacentes.

É também importante reconhecer que a matriz de correlação na Tabela 4.10 não é de posições completas. Uma vez que as variáveis agora são expressas como proporções, elas somam-se em um 100% constante. Portanto, mesmo havendo 13 variáveis, a classificação da matriz é somente 12 (com efeito, removemos uma dimensão com a divisão pelo total do GSP). Isso significa que os autovalores dos 12 primeiros componentes principais somarão 13; o último componente principal para esses dados terá um autovalor de exatamente zero.

4.3.2 Resultados

Dados brutos

Os resultados básicos da análise de componentes principais dos dados do *GPS_RAW* (Produto Estadual Bruto) são apresentados na Tabela 4.11. (Para economizar espaço, apresentamos apenas as cargas dos três primeiros componentes.) Uma olhada nos autovalores revela que o primeiro componente principal explica uma quantidade substancial de variação nos dados: $\lambda_1 = 10{,}9$. Como o total de variância do conjunto original de variáveis é 13, isso significa que somente o primeiro autovalor explica 10,9/13 = 84% da variação. O próximo maior autovalor é somente $\lambda_2 = 0{,}98$, que explica aproximadamente 7% da variância dos dados originais.

Uma olhada nas cargas confirma que o primeiro componente principal é uma combinação positiva de todas as 13 medidas da atividade econômica (a única medida da atividade econômica que não tem peso elevado é a mineração). O primeiro componente é essencialmente interpretável como um índice da atividade econômica total, que varia de um elevado 14,7 para a Califórnia ao modesto −2,73 de Vermont. É interessante observar que a correlação entre o primeiro componente principal e a população estimada em cada estado para 1995 é de 0,995.

Dados compartilhados

A análise de componentes principais dos dados do *GSP_SHARE* revela alguns padrões mais interessantes de associação. Dividir pelo GSP total efetivamente remove da análise a informação sobre

Tabela 4.10 Correlação entre 13 diferentes medidas da atividade econômica (dados expressos como participação no GSP total)

	AGRI-CULTURA	*MINERAÇÃO*	*CONS-TRUÇÃO CIVIL*	*IND. BENS DURÁVEIS*	*IND. BENS NÃO DURÁVEIS*
MINERAÇÃO	−0,064				
CONSTRUÇÃO CIVIL	0,085	−0,021			
IND. BENS DURÁVEIS	0,032	−0,424	−0,130		
IND. BENS NÃO DURÁVEIS	−0,145	−0,138	−0,318	0,204	
TRANSPORTE	0,279	0,612	0,075	−0,357	−0,176
COMUNICAÇÃO	−0,184	−0,193	−0,023	−0,317	−0,100
SERVIÇOS URBANOS	0,043	0,390	0,013	−0,051	0,071
COMÉRCIO ATACADISTA	0,245	−0,553	−0,087	0,271	0,039
COMÉRCIO VAREJISTA	0,095	−0,396	0,401	0,195	−0,121
FIDUCIÁRIO	−0,301	−0,406	−0,253	−0,182	−0,133
SERVIÇOS	−0,322	−0,460	0,324	−0,159	−0,458
GOVERNO	0,110	0,231	0,181	−0,411	−0,237

	TRANS-PORTE	*COMUNI-CAÇÃO*	*SERVIÇOS URBANOS*	*COMÉRCIO ATACADISTA*	*COMÉRCIO VAREJISTA*	*FIDUCIÁRIO*	*SERVIÇOS*
MINERAÇÃO							
CONSTRUÇÃO CIVIL							
IND. BENS DURÁVEIS							
IND. BENS NÃO DURÁVEIS							
TRANSPORTE	−0,049						
COMUNICAÇÃO	−0,056	−0,169					
SERVIÇOS URBANOS	−0,214	0,330	−0,267				
COMÉRCIO ATACADISTA	−0,148	0,125	0,030	0,166			
COMÉRCIO VAREJISTA	−0,503	0,120	−0,379	0,040	−0,309		
FIDUCIÁRIO	−0,422	0,309	−0,314	0,239	0,202	0,519	
SERVIÇOS	0,428	0,193	0,045	−0,343	0,287	−0,351	−0,180
GOVERNO							

o tamanho do estado. Os resultados são apresentados na Tabela 4.12 (novamente, para economizar espaço, apresentamos apenas as cargas para os três primeiros componentes). Nesse caso, nenhuma dimensão dominante única capta a variabilidade mais relevante nas medidas originais. Os três primeiros autovalores explicam (3,24+2,23+1,96)/13 = 57% do total da variância; os cinco primeiros autovalores (todos maiores que 1) explicam 77%.

Sem que tomemos uma decisão sobre quantos componentes reter (guardemos isso para depois), examinamos os primeiros dois componentes principais na tentativa de obter alguma ideia de sua interpretação. Analisar as cargas pode ajudar. Vemos que o primeiro componente principal está positivamente correlacionado com mineração e transporte e negativamente com o comércio atacadista e o fiduciário, seguros e atividade imobiliária. O segundo componente principal é positivamente correlacionado com a indústria de bens duráveis e de consumo (ou não duráveis) e negativamente com construção e comunicação. Observe que tanto Z_1 quanto Z_2 são negativamente correlacionados com serviços, no entanto, Z_1 exibe uma correlação positiva com o governo, enquanto Z_2 tem uma forte correlação negativa.

Às vezes, é bastante útil desenhar as cargas do componente principal. A Figura 4.12 mostra um gráfico das cargas dos dois primeiros componentes principais. Cada ponto na Figura 4.12 representa um par de correlações: a correlação da variável original com cada um dos dois componentes principais. O gráfico revela quais das medidas originais estão mais proximamente associadas a cada um dos componentes, bem como quais medidas são mais semelhantes (refletidas por pontos que são próximos no mapa). Por exemplo, ambas as atividades da indústria (bens duráveis e de consumo) ocupam um perfil relativamente similar nas primeiras duas dimensões do espaço do componente principal.

Um diagrama dos escores do componente principal pode nos dar uma ideia da localização dos vários estados no espaço do componente principal. Um diagrama de dispersão baseado nos dois primeiros componentes principais é apresentado na Figura 4.13. A primeira coisa que se destaca no gráfico é a existência de duas discrepâncias muito substanciais. Tanto Alasca quanto Wyoming têm valores de 6,6 no primeiro componente principal; o estado seguinte mais próximo possui um valor de menos de 3,0. Esses valores elevados podem ser atribuídos aos valores extremos de duas dimensões da atividade econômica local: mineração e transporte. Em Wyoming, 31,6% do Produto Estadual Bruto vêm da mineração. No Alasca, a porcentagem é de 22,4%. Isso é comparado ao nível médio de todos os 50 estados com menos de 1%. De maneira semelhante, para o transporte, as porcentagens do Alasca e de Wyoming são 12,1% e 6,4%, respectivamente (comparadas a uma média de 3%). Essas são algumas das discrepâncias mais substanciais em todo o conjunto de dados. Já que o objetivo dos componentes principais é explicar tantas variâncias quanto possível, não é surpreendente

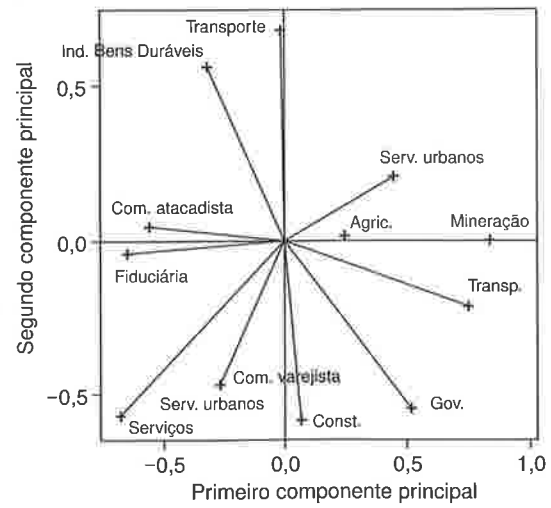

Figura 4.12 Gráfico das cargas de fator para os dois primeiros componentes principais dos dados de *GSP_SHARE*.

Tabela 4.11 Resultado da análise dos componentes principais dos dados de *GPS_RAW*: autovalores e cargas

	1	2	3	4	5	6	7	8	9	10	11	12	13
Autovalor	10,9443	0,9794	0,4001	0,3427	0,1392	0,0694	0,0401	0,0333	0,0251	0,0105	0,0075	0,0061	0,0022
Cumulativo	0,8419	0,9172	0,9480	0,9743	0,9851	0.9904	0,9935	0,9960	0,9980	0,9988	0,9994	0,9998	1,0000

	Cargas		
	Z_1	Z_2	Z_3
AGRICULTURA	0,82452	−0,14508	0,51730
MINERAÇÃO	0,39706	0,90347	0,08516
CONSTRUÇÃO CIVIL	0,98718	−0,03238	0,00568
IND. BENS DURÁVEIS	0,88799	−0,08722	0,13649
IND. DE BENS DE CONSUMO	0,90347	0,09041	−0,12308
TRANSPORTE	0,98010	0,08473	0,03860
COMUNICAÇÃO	0,94977	−0,02280	−0,19607
SERVIÇOS URBANOS	0,95031	0,19483	−0,14271
ATACADO	0,99204	−0,05292	−0,01980
VAREJO	0,98999	−0,06360	0,05374
FIDUCIÁRIO	0,93650	−0,21819	−0,15567
SERVIÇOS	0,97547	−0,14711	−0,05229
GOVERNO	0,98436	−0,05950	−0,02916

Variância explicada por		
Z_1	Z_2	Z_3
10,9443	0,9794	0,4001

Tabela 4.12 Resultados da análise dos componentes principais dos dados do *GSP_SHARE*: autovalores e cargas

	1	2	3	4	5	6	7	8	9	10	11	12	13
Autovalor	3,2355	2,2365	1,9598	1,3603	1,1574	0,8683	0,7245	0,6158	0,3182	0,2354	0,1517	0,1365	0,0000
Cumulativo	0,2489	0,4209	0,5717	0,6763	0,7654	0,8321	0,8879	0,9352	0,9597	0,9778	0,9895	1,0000	1,0000

	Cargas		
	Z_1	Z_2	Z_3
AGRICULTURA	0,24251	−0,01116	0,53899
MINERAÇÃO	0,84487	−0,00222	−0,36357
CONSTRUÇÃO CIVIL	0,06347	0,58840	0,36005
IND. BENS DURÁVEIS	−0,32981	−0,56192	0,52553
IND. DE BENS DE CONSUMO	−0,01746	−0,68617	0,04997
TRANSPORTE	0,75273	0,21978	0,00890
COMUNICAÇÃO	−0,27324	0,47225	−0,11488
SERVIÇOS URBANOS	0,44418	−0,20634	0,09641
ATACADO	−0,56709	−0,04233	0,40618
VAREJO	−0,16213	0,39039	0,71010
FIDUCIÁRIO	−0,65308	0,04526	−0,62523
SERVIÇOS	−0,68331	0,57414	−0,17846
GOVERNO	0,51955	0,55114	0,11951

Variância explicada por		
Z_1	Z_2	Z_3
3,2355	2,2365	1,9598

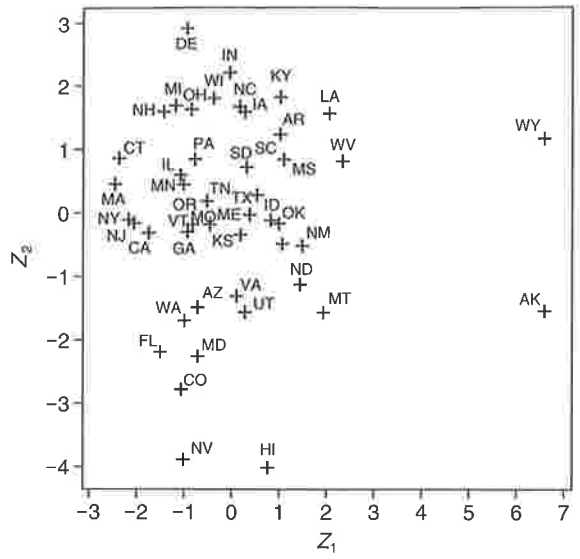

Figura 4.13 Gráfico dos escores dos dois primeiros componentes principais dos dados de *GSP_SHARE*.

que o primeiro componente principal reflita as dimensões subjacentes em que essas observações discrepantes ocorrem.

4.4 QUESTÕES RELATIVAS À APLICAÇÃO DOS COMPONENTES PRINCIPAIS

4.4.1 Quando é apropriado usar os componentes principais?

Até agora, assumimos que os dados sob análise são altamente correlacionados para justificar o uso dos componentes principais na redução das dimensões. Entretanto, se as variáveis forem altamente independentes umas das outras, os componentes principais podem não ser apropriados.

O teste de esfericidade, desenvolvido por Bartlett (1950), coloca diretamente a questão: a matriz de correlação deve ser decomposta em fatores, em primeiro lugar? O teste é um teste de qui-quadrado aproximado com um teste estatístico que é função do determinante da matriz de correlação **R**, como apresentado a seguir:

TESTE DE ESFERICIDADE DE BARTLETT

$$\chi^2\left[\frac{(p^2-p)}{2}\right] = -\left[(n-1)-\frac{(2p+5)}{6}\right]\ln|\mathbf{R}|$$

onde: $\ln|\mathbf{R}|$ = log natural do determinante da matriz de correlação,
$(p^2-p)/2$ = número de graus de liberdade associado ao teste estatístico de qui-quadrado,
p = número de variáveis,
n = número de observações.

A lógica do teste é bastante simples de descrever. O determinante da matriz de correlação é uma medida generalizada da variância. Ele pode ser calculado tomando-se o produto dos autovalores da matriz, isto é,

$$|\mathbf{R}| = \prod_{j=1}^{p} \lambda_j \tag{4.15}$$

Quando as variáveis na análise são de fato mutuamente independentes, esperamos que **R** se aproxime da matriz identidade **I** (isto é, elementos diagonais igual a 1 e elementos fora da diagonal perto de zero em valor absoluto). Quando esse é o caso, um diagrama de dispersão dos dados apresenta

formato esférico em vez de oval, como uma bola de futebol americano (daí o nome do teste). Todos os autovalores da matriz são próximos de 1,0 (indicando que o eixo mais longo do elipsoide não é significativamente mais longo do que qualquer outro eixo); portanto, $|\mathbf{R}|$ é próximo de 1,0 e $\ln|\mathbf{R}|$ é próximo de 0. À medida que o montante de correlação entre as variáveis aumenta, o diagrama de dispersão dos dados começa a parecer mais elipsoidal. Nesse caso, alguns dos autovalores de **R** são maiores que 1 e alguns são mais próximos de zero. O resultado é que, à medida que o nível de correlação aumenta, o produto dos autovalores torna-se mais próximo de zero, o que significa que $\ln|\mathbf{R}|$ torna-se um número negativo maior.

Agora fica fácil perceber como funciona o teste de Bartlett. O qui-quadrado é um teste de qualidade de ajuste, com a hipótese nula de que a verdadeira matriz de correlação da população subjacente (da qual a amostra é retirada) é a *matriz identidade* (isto é, um conjunto perfeitamente esférico de dados multivariados). Quando o determinante da matriz de correlação da amostra é próximo de 1, o teste de qui-quadrado de Bartlett é próximo de zero, indicando um bom ajuste e não nos dando nenhuma razão para rejeitar a hipótese nula de esfericidade. Se não pudermos rejeitar a hipótese nula, podemos concluir que não é apropriado reduzir a dimensionalidade dos dados. No entanto, quanto mais correlação está presente nos dados, $|\mathbf{R}|$ torna-se mais próximo de 0,0 (o que significa que $\ln|\mathbf{R}|$ torna-se um número negativo grande). Nesse caso, o teste de qui-quadrado de Bartlett adquire um grande valor positivo, indicando um mau ajuste dos dados à hipótese nula e levando-nos a rejeitar a hipótese nula de esfericidade.

EXEMPLO *Determinantes da orientação estratégica de uma unidade de negócio (Burke, 1984)*
STRATEGIC_THRUST

O exemplo de problema baseado no produto estadual bruto foi escolhido por sua fatorabilidade óbvia. Não deveria surpreender-nos descobrir que o teste de Bartlett para esfericidade é cabalmente rejeitado (de fato, o teste estatístico é $\chi^2 = 190{,}4$ com 66 graus de liberdade, o que é significativo em um valor de p de menos de 0,0001). Mas o que acontece com um problema em que o padrão de correlação nos dados é muito menos pronunciado? Burke (1984) estudou os determinantes da orientação estratégica de uma unidade comercial. Entre as variáveis de seu estudo, há seis variáveis ambientais potencialmente importantes; uma lista de definições de variáveis e a matriz de correlação são dadas na Tabela 4.13 (os resultados são baseados em uma amostra de $n = 86$ unidades comerciais distintas). A questão que precisamos responder é essa: as seis variáveis ambientais representam seis dimensões independentes que podem ser usadas para caracterizar a orientação estratégica de uma unidade comercial?

Tabela 4.13 Matriz de correlação para os dados de Burke

	MA	*EX*	*EN*	*OS*	*EU*	*SI*
Atratividade do mercado (MA)	1,00					
Barreiras para a saída (EX)	−0,09	1,00				
Barreiras para entrada (EN)	0,02	0,31	1,00			
Simetria da organização (OS)	0,07	0,01	0,10	1,00		
Incerteza ambiental (EU)	0,04	0,09	−0,24	−0,06	1,00	
Importância de curto prazo (SI)	0,21	0,13	0,12	0,23	0,08	1,00

Uma olhada casual nas correlações da matriz sugere que a resposta pode ser sim. A correlação mais alta na tabela é de 0,31 e mais da metade das correlações são menores que 0,10 em valor absoluto.

Uma análise de componente principal da matriz de correlação apresenta os seguintes autovalores: $\lambda_1 = 1{,}491$, $\lambda_2 = 1{,}266$, $\lambda_3 = 1{,}122$, $\lambda_4 = 0{,}905$, $\lambda_5 = 0{,}671$ e $\lambda_6 = 0{,}545$. Calculando-se o determinante da matriz de correlação, temos $|\mathbf{R}| = 0{,}701$ e $\ln|\mathbf{R}| = -0{,}355$. Substituindo no teste estatístico de Bartlett (com $p = 6$ e $n = 86$), temos

$$\chi^2\left[\frac{(36-6)}{2}\right] = -\left[(86-1) - \frac{(12+5)}{6}\right](-0{,}355)$$

ou

$$\chi^2(15) = 29{,}2$$

Consultando uma tabela de qui-quadrado, observa-se que o valor crítico para um χ^2 estatístico com 15 graus de liberdade em um nível de confiança de 95% (isto é, $\alpha = 0{,}05$) é de 25,0. Uma vez que o teste estatístico é maior que o valor crítico, rejeitamos a hipótese nula de esfericidade, o que significa ser justificável perseguir alguma forma de redução de dimensão.

No entanto, é importante lembrar que, embora o teste estatístico de Bartlett seja um teste aproximado, sua força parece ser bastante elevada. O teste de qui-quadrado é notoriamente sensível ao tamanho da amostra; como resultado, é incomum encontrar um exemplo prático no qual a esfericidade não seja rejeitada. Isso tende a ser desvantajoso na prática, porque dados perfeitamente esféricos são raramente encontrados. Isso significa que, para qualquer conjunto de dados de tamanho razoável, é raro encontrar um caso em que a esfericidade não seja rejeitada. Esse exemplo nos ajuda a apontar uma das limitações do teste de Bartlett. Definitivamente, a decisão de se reduzir a dimensionalidade de um conjunto de dados é baseada mais na necessidade prática de haver uma representação parcimoniosa do que em um teste estatístico de independência.

4.4.2 Como os dados deveriam ser escalonados?

Até agora, todas as nossas demonstrações envolveram o uso de dados padronizados; isto é, nosso foco tem sido a matriz de correlação em vez da matriz de covariância. Há uma boa razão para isso, embora ela certamente não seja necessária. Como a análise de componentes principais busca maximizar a variância, ela pode ser sensível às diferenças de escala entre as variáveis. Com a inclusão de uma variável como "Vendas Corporativas", expressa em dólares (em que a variação entre as observações mais e menos elevadas pode ser da ordem de milhões), teremos um impacto muito maior sobre a variância da combinação linear do que uma variável como "Número de Empregados" (em que a variação provavelmente será da ordem de milhares), independente do padrão de covariação entre as variáveis. A padronização garante que os dados sejam expressos em unidades comparáveis.

Isso é especialmente pertinente no exemplo que envolve medidas do produto estadual bruto. Entre nossas 13 medidas de atividade econômica (mesmo quando expressas em proporções do GSP total), as medidas que exibem a quantidade média maior de atividade econômica também exibem a maior variância. Por exemplo, a proporção média do GSP atribuível ao fiduciário, seguros e atividade imobiliária é de 17,1% para todos os 50 estados; o desvio-padrão é de 5,2 pontos percentuais. Para a construção civil, o nível médio é de 4,3% e o desvio-padrão é de somente 0,9. Nesse contexto, conduzir os componentes principais sobre os dados de GSP não padronizados tenderia a agrupar as grandes áreas da atividade econômica, não necessariamente as variáveis com o padrão mais forte de intercorrelação.

Em outras circunstâncias, há razões para evitar-se a padronização. Considere uma pesquisa de mercado, por exemplo, na qual todas as perguntas foram medidas em uma escala Likert de cinco pontos (a escala vai de "Concordo Totalmente" a "Discordo Totalmente"). É possível que algumas questões sejam mais ou menos informativas do que outras, talvez porque estejam mal redigidas ou não claramente entendidas pelos respondentes, ou porque há homogeneidade substancial com respeito à resposta. Imagine uma questão que tenha como quase todas as respostas justamente o meio da escala ("Não concordo nem discordo"). Como essa questão nos fornece dados que não são informativos, é preferível dar a essa variável menos peso na análise subsequente usando os dados não padronizados. Em casos em que houver razão para se acreditar que a variância de uma variável é um indicador de sua importância em geral (ou alto valor de informação), podem ser apropriadas a utilização de dados não padronizados e a análise de componentes principais sobre a matriz de covariância. Embora isso mude de alguma forma a natureza de seus resultados – por exemplo, as cargas dos componentes principais serão graduadas como covariâncias em vez de correlações –, a mesma intuição geral se aplica.

4.4.3 Quantos componentes devem ser retidos?

Caso seja apropriado reduzir-se a dimensionalidade dos dados originais, a próxima pergunta que naturalmente surge é: quantos componentes devem ser retidos? É possível responder essa questão com o emprego de uma aplicação sequencial de uma versão modificada do teste de Bartlett. Se a esfericidade for rejeitada, extraímos o maior componente principal e depois testamos a matriz de correlação residual (isto é, $\mathbf{R} - \lambda_1 \mathbf{u}_1 \mathbf{u}'_1$) para examinar se seu determinante é diferente de zero. Continuamos extraindo os componentes até que a matriz residual não seja estatisticamente significativa. Infelizmente, devido ao poder do teste de qui-quadrado, essa abordagem tende a reter um grande número de componentes. Uma vez que, na prática, o pesquisador está interessado na possibilidade de interpretação e na significância operacional das soluções dos componentes principais, outros critérios têm sido adotados para auxiliar no processo de seleção do número de componentes a ser retidos. Discutimos, a seguir, três dessas regras de ouro. Em cada caso, a aplicação da regra deve ser acompanhada por uma porção substancial de discernimento.

Gráfico Scree

Essa abordagem gráfica foi proposta por Cattell (1966). Ela envolve desenhar a variância explicada por cada componente principal, na ordem do maior para o menor. Então, buscamos um "ângulo" na curva – isto é, um ponto após o qual os autovalores remanescentes declinam de modo aproximadamente linear – e retemos somente aqueles componentes que estão acima do ângulo. Portanto, o *teste scree* requer um julgamento relativo do valor da variância explicada pelos componentes retidos.

O gráfico scree para os dados do GSP é apresentado na Figura 4.14. Uma inclinação digna de nota no gráfico ocorre após o terceiro componente principal (com um autovalor de 1,96); do quarto componente em diante, o declínio nos autovalores é aproximadamente linear e mais plano do que na primeira parte do gráfico. Com base nesse gráfico scree, podemos decidir-nos por reter somente os primeiros três componentes principais. Infelizmente, na prática, a identificação do ângulo raramente é tão clara e pouco ambígua. Com muita frequência, o declínio na variância se parece mais com uma curva suave do que com um ângulo agudo. O gráfico scree para os dados do estudo de Burke, apresentado na Figura 4.15, é um exemplo. Você diria que o ângulo está em qual lugar desta curva?

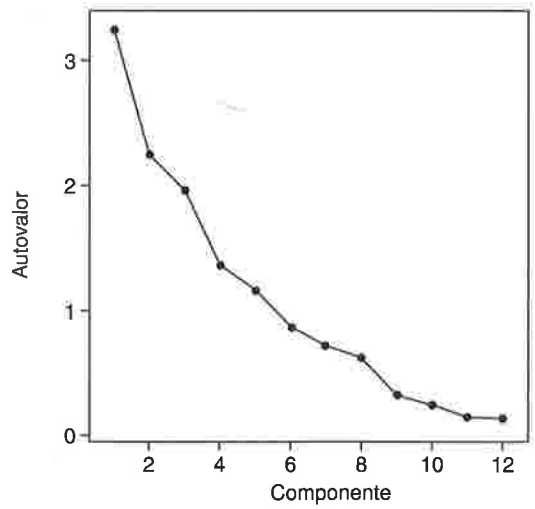

Figura 4.14 Gráfico scree para os dados *GSP_SHARE*.

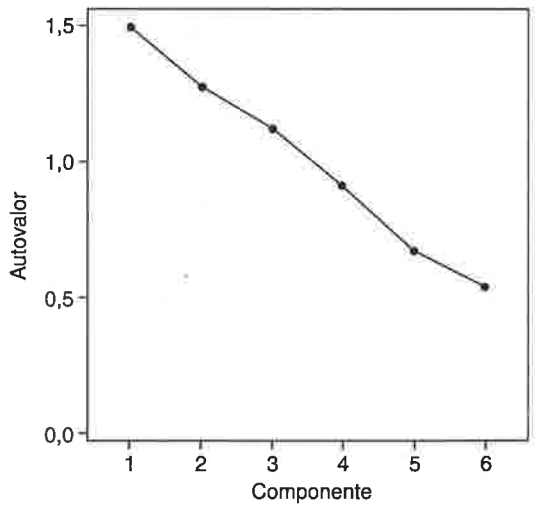

Figura 4.15 Gráfico scree para os dados de Burke.

Regra de Kaiser

Quando o gráfico scree não oferece diagnóstico, a *regra de Kaiser* pode ser conveniente. Kaiser (1959) recomendou somente a retenção dos componentes principais com autovalores que excedam a unidade (isto é, supondo que estamos tratando com variáveis padronizadas na análise). A regra reflete uma noção de bom-senso: qualquer componente principal, por se tratar de uma medida da variância comum,

deveria explicar pelo menos tantas variações quanto qualquer uma das variáveis originais X. A regra de Kaiser, portanto, busca um julgamento absoluto em relação ao valor de variância explicada por cada componente principal.

No caso dos dados do GSP, a regra de Kaiser busca reter os primeiros cinco componentes principais: $\lambda_1 = 3{,}24$, $\lambda_2 = 2{,}23$, $\lambda_3 = 1{,}96$, $\lambda_4 = 1{,}37$, $\lambda_5 = 1{,}15$. Entretanto, dois componentes têm autovalores próximos a 1,0: $\lambda_5 = 1{,}15$ e $\lambda_6 = 0{,}86$. Pode-se justificar (em razão da possibilidade de interpretação) o estabelecimento de um limite pouco superior a 1,0 e a retenção somente de uma solução quadridimensional. Ou pode-se argumentar que, para explicar os 90% de variação nos dados originais, necessitamos estabelecer um limite um pouco abaixo de 1,0 para que retenhamos os primeiros seis componentes principais. Como qualquer uma dessas abordagens, a regra de Kaiser deve ser encarada como uma orientação, não como uma lei inviolável.

Procedimento de Horn

Horn (1965) sugeriu um tipo diferente de regra para o estabelecimento de limite. Em lugar de usar um valor fixo de 1,0, Horn propôs o uso de um autovalor como limite, a partir da análise de componentes principais de dados randômicos com o mesmo número de variáveis e o mesmo número de observações. O procedimento é baseado na percepção de que a análise de componentes principais, em sua tentativa de explicar o máximo possível da variância, irá capitalizar sobre a variação da amostra randômica dentro dos dados. Mesmo quando os dados são retirados de uma população subjacente com matriz de covariância $\mathbf{I}$, o erro de amostragem levará a desvios menores de $\mathbf{I}$ que são apreendidos pelos componentes principais. Como consequência, o autovalor do primeiro componente principal será sempre maior que 1. O *procedimento de Horn* essencialmente ajusta o limite para explicar essa tendência e capitalizar sobre o erro da amostra.

Para calcular o limite de Horn, geramos uma matriz de dados da amostra de ordem $n \times p$ usando um gerador de números randômicos. Depois, aplicamos os componentes principais aos dados randômicos e desenhamos os autovalores em um gráfico scree junto com os autovalores dos dados reais. Somente aqueles componentes principais com autovalores que excederem o limite de Horn são retidos. Adicionamos uma modificação de atualização ao procedimento de Horn: em lugar de usar um gerador de números randômicos (com alguma distribuição paramétrica subjacente), podemos melhorar a situação a partir dos dados reais. Em outras palavras, fazemos testes com a recolocação de cada variável no conjunto de dados. Observe que devemos avançar e melhorar a situação a partir de cada variável separadamente (isto é, da distribuição marginal de cada variável e não da distribuição conjunta) para gerar uma amostra na qual a distribuição conjunta subjacente seja completamente independente.

Ilustramos a aplicação do procedimento de Horn usando os dados relativos ao produto estadual bruto. Geramos uma matriz de amostra aleatória de ordem 50×13 avançando a partir dos dados reais. Aplicamos os componentes principais para dados randômicos e reproduzimos os autovalores obtidos em um gráfico scree ao longo dos autovalores da análise dos componentes principais dos dados reais. O gráfico scree é apresentado na Figura 4.16.

Observe que o autovalor do primeiro componente principal dos dados incrementados é 2,17. Assim, mesmo quando os dados são inteiramente gerados a partir de uma população subjacente, com completa independência de todas as variáveis, o erro de amostragem conduz a um componente principal com mais de duas vezes a variância do que qualquer variável do conjunto original de dados. Isso é um reflexo da capitalização sobre a variação ao acaso que se pode esperar com uma técnica de maximização de variância tais como os componentes principais (como é o caso com a regressão e outras técnicas que estudaremos em capítulos posteriores).

O ponto de interseção no gráfico da Figura 4.16 sugere que retenhamos somente os três primeiros componentes (em lugar dos cinco componentes sugeridos pela aplicação da regra de Kaiser). De fato, o procedimento de Horn irá sempre fornecer um autovalor de limite maior que 1. Dessa maneira, a aplicação do procedimento de Horn sempre resultará na retenção do mesmo ou de menor número de componentes principais do que reteríamos aplicando a regra de Kaiser.

Um dos principais benefícios do procedimento de Horn está em tornar as pessoas conscientes do potencial de capitalização pelos componentes principais sobre a variação ao acaso na amostra.

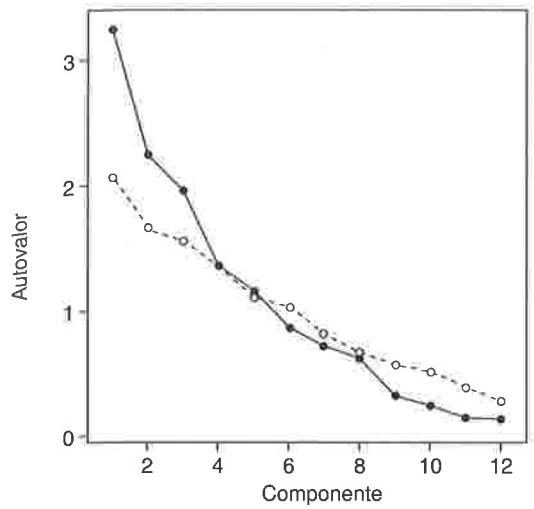

Figura 4.16 Gráfico scree dos dados de *GSP_SHARE* (linha contínua) *versus* benchmark de Horn (linha pontilhada).

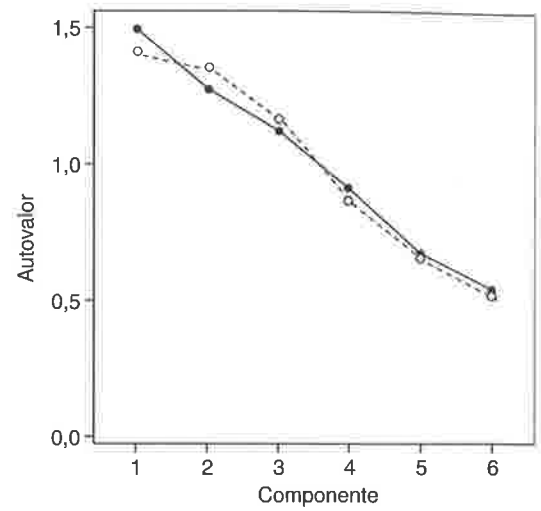

Figura 4.17 Gráfico scree para os dados de Burke (linha contínua) *versus* benchmark de Horn (linha pontilhada).

Considere, por exemplo, o resultado da aplicação do procedimento nos dados de Burke (apresentado na Figura 4.17): o benchmark de Horn é quase indistinguível do gráfico scree dos autovalores dos dados reais. Isso sugere que as variáveis no estudo de Burke podem, de fato, ser independentes e que os componentes principais com autovalores maiores que 1 são resultado da capitalização sobre a correlação ao acaso que existe na amostra. O grau dessa capitalização torna-se ainda mais extremo com amostras de tamanhos menores e maior número de variáveis; o procedimento de Horn nos ajuda a conhecer a extensão dessa "inflação" quando escolhemos o número de fatores para reter na análise.

Variância explicada

Às vezes, será importante reter um número suficiente de componentes para reconstruir-se adequadamente o conjunto de dados originais (isto é, para que sejamos capazes de explicar determinado valor da variância em cada uma das variáveis originais). Lembre-se de que a relação entre os dados originais $\mathbf{X}$ e os componentes principais $\mathbf{Z}$ pode ser expressa como $\mathbf{X} = \mathbf{Z}_s\mathbf{D}^{1/2}\mathbf{U}'$. Se retivermos somente c componentes padronizados, a relação é somente aproximada; isto é,

$$\mathbf{X} \approx [\mathbf{z}_{1s}, \mathbf{z}_{2s}, \ldots, \mathbf{z}_{cs}]\mathbf{D}^{1/2}\mathbf{U}'$$

Podemos impor o seguinte tipo de limitação à nossa análise: devemos reter um número suficiente de componentes para que possamos explicar pelo menos 50% (ou alguma outra porcentagem especificada) da variância em cada variável original. O exame dos autovalores diz se podemos explicar 50% da variância em todas as variáveis, mas não se captamos pelo menos 50% da variância de cada variável. Quando tal critério de porcentagem é usado, ele deve ser baseado em uma avaliação da confiabilidade (e, portanto, a porcentagem da variação de erro) das variáveis que estão sendo analisadas.

Para cada área da atividade econômica nos dados *GSP_SHARE*, a Tabela 4.14 mostra a variância explicada pelos três primeiros componentes principais. Em 10 das 13 variáveis, o valor é de quase 50% ou mais (por exemplo, mais de 80% da variância nos setores fiduciário, de seguros e atividade imobiliária é explicado pelos três primeiros componentes). A tabela também mostra que menos de 40% da variância original é captada em três das 13 variáveis: agricultura, silvicultura e pesca; comunicação; e gás, eletricidade e saneamento (para o último dos três, a proporção da variância explicada é inferior a 25%). Se essas variáveis forem consideradas confiáveis e importantes para as análises subsequentes, faz sentido reter um quarto componente principal e determinar se a variância explicada nessas variáveis sofre algum aumento.

Tabela 4.14 Variância explicada pelos três primeiros componentes principais em cada medida da atividade econômica dos dados do GSP

AGRICULTURA	0,349
MINERAÇÃO	0,846
CONSTRUÇÃO CIVIL	0,480
IND. BENS DURÁVEIS	0,701
IND. BENS DE CONSUMO	0,474
TRANSPORTE	0,615
COMUNICAÇÃO	0,311
SERVIÇOS URBANOS	0,249
ATACADO	0,488
VAREJO	0,683
FIDUCIÁRIO	0,819
SERVIÇOS	0,828
GOVERNO	0,588

4.4.4 Como avaliar a validade da solução?

Em geral, quando abordamos a questão da validade, estamos questionando a possibilidade de generalização dos resultados de nossa análise além dessa amostra particular. Quando se lida com um conjunto de dados de corte transversal de amostra de tamanho n, pode-se perguntar até que ponto os achados de nossa análise (com base em uma amostra dada) se estende a uma nova amostra de tamanho n retirada da mesma população subjacente. Se não tentarmos responder a essa questão, não poderemos ter certeza da extensão com que nossas descobertas se relacionam à população subjacente ou somente a essa amostra particular.

Uma forma de resolver essa questão é usar uma *amostra de teste*. Antes de realizar a análise de componentes principais, colocamos de lado uma parte dos dados escolhida aleatoriamente. Uma regra de ouro: use dois terços dos dados para a análise e um terço como amostra de teste. A amostra de teste nos dá um conjunto de observações completamente diferente daquele usado na análise (e, no entanto, retirado da mesma população subjacente) com o qual podemos verificar os resultados do modelo.

Um aspecto da solução de componentes principais que produz validação é a variância explicada por cada componente principal. Se nossa solução de componentes principais se mostrar válida, é de se esperar que ela se generalize para uma amostra diferente, retirada da mesma população subjacente. Desse modo, poderemos tomar o autovetor $\mathbf{u}_1$ derivado da análise, formar a combinação linear $\mathbf{X}_s\mathbf{u}_1$ usando dados padronizados da segunda amostra (teste) e ter a variância do componente assim formada aproximadamente igual à variância de $\mathbf{Z}_1$ da análise dos componentes principais da primeira amostra.

Vimos do procedimento de Horn que mesmo que não haja associação na população subjacente (isto é, a verdadeira matriz de correlação da população é $\mathbf{I}$), a análise dos componentes principais resultará em uma solução em que os primeiros poucos componentes explicam mais do que uma quantidade proporcional de variação na amostra. Quando os dados são verdadeiramente independentes, no entanto, pode-se esperar que a combinação linear $\mathbf{X}_s\mathbf{u}_1$ (formada pelo uso do primeiro autovetor da análise de componentes principais e dos dados da amostra de teste) apresente uma variância de aproximadamente 1, porque nenhum dos padrões de associação na amostra original é comum à segunda amostra (de teste) graças à verdadeira independência na população subjacente.

A questão ainda permanece: como validar os resultados da nossa análise dos dados reais de GSP? Quanto de variação no primeiro componente principal é atribuível à variação da capitalização sobre o acaso na amostra? Essa é uma questão difícil, porque já analisamos os dados da população inteira (todos os 50 estados). Criar uma amostra de teste é possível (por exemplo, desconsiderando um terço dos dados ou 16 estados escolhidos aleatoriamente), mas isso reduz o tamanho do conjunto de dados usados na análise dos componentes principais e, na realidade, aumenta a probabilidade de haver

capitalização sobre o acaso (o que reduz o número de observações relativas ao número de variáveis). Podemos usar os dados de um ano diferente, mas quaisquer diferenças entre as duas análises poderiam ser atribuídas a mudanças na atividade econômica no decorrer do tempo (em vez de puramente um erro de amostragem). No restante desta seção, discutimos duas abordagens possíveis para validação: a *jackknife* e a *bootstrap*.

Validação jackknife

No caso de amostras de tamanho pequeno como essa, a validação *jackknife* (também conhecida como *abordagem por método-U por amostra de teste*; veja, por exemplo, Crask e Perreault, 1977) é especialmente útil. A ideia é: retemos uma observação, conduzimos a análise de componentes principais sobre as 49 observações restantes e, depois, utilizamos o vetor $\mathbf{u}_1$ para calcular o valor de $\mathbf{Z}_1$ na única observação retida. Depois, repetimos esse processo 50 vezes, conduzindo uma análise diferente de componentes principais para cada possível subconjunto de 49 estados. Podemos, então, comparar a variância dos valores "jackknificados" à variância do primeiro componente principal (determinado pelo uso de todas as 50 observações). A vantagem dessa abordagem é que não se usam quaisquer das informações da i-ésima observação para calcular-se o escore do seu componente principal, e cada análise é baseada em uma amostra quase do mesmo tamanho do conjunto de dados originais. A desvantagem, naturalmente, é que é mais complicado realizar a análise. Os cálculos não demandam muito tempo – hoje em dia, é possível realizar dezenas de análises de componentes desse tamanho em segundos –, mas muitos pacotes de programas estatísticos não fazem provisões explícitas para uma abordagem baseada em reamostragem.

Os resultados desse exercício de validação mostram que a variância do componente "jackknificado" diminui apenas ligeiramente para 3,08 (de um autovalor de $\lambda_1 = 3{,}24$ da análise dos componentes principais originais). Portanto, o primeiro componente principal parece explicar a variância que é sistemática e generalizável para além da amostra; se captasse puramente o erro de amostragem, poderíamos esperar que a variância do componente "jackknificado" declinasse para um valor próximo a 1,0. Lembre-se de que a variância captada por esse primeiro componente principal é amplamente determinada por dois estados (Alasca e Wyoming) que apresentam participações desproporcionalmente grandes em duas áreas da atividade econômica: mineração e transporte. Parte da razão pela qual obtemos uma indicação de validade com a abordagem jackknife se dá porque uma dessas duas discrepâncias está sempre presente na amostra. Quando o Alasca é retido, Wyoming determina a direção do primeiro componente principal, e vice-versa. Em geral, seria de se esperar uma redução mais elevada na variância do componente principal "jackknificado" na presença de uma discrepância. Embora os resultados desse exemplo pareçam estáveis, é importante entender que a variância explicada por esse primeiro componente principal é atribuível a um subconjunto minúsculo de observações.

Validação bootstrap

Podemos avaliar a validade de nossos resultados caso seja possível aplicá-los a outra amostra de dados. Na ausência de uma amostra de teste, podemos obter um novo conjunto de observações reamostrando os dados, um processo conhecido como *bootstrapping*. Se supusermos que os dados em nossa amostra são representativos de alguma população subjacente, retirar várias amostras de tamanho n (com recolocação, o que significa que algumas observações serão tiradas mais de uma vez e algumas, nenhuma vez) poderia imitar a variabilidade introduzida pela amostragem da população como um todo. O bootstrapping permite-nos avaliar a distribuição de nossos resultados de análise dos componentes principais mesmo quando suas propriedades estatísticas não sejam bem conhecidas.

Nesse caso, estamos interessados em avaliar quanto da variância do primeiro componente principal é atribuível à variação ao acaso na amostra. Para isso, criamos nova amostra de dados (isto é, tiramos uma amostra de 50 observações com recolocação) e depois formamos uma combinação linear dos dados "bootstrappados", usando o vetor $\mathbf{u}_1$ do conjunto inicial de dados. Podemos, então, comparar a variância dessa combinação linear com a variância do primeiro componente principal da amostra bootstrapped (que sabemos ser a combinação linear com a variância máxima). Se a comparação for próxima – por exemplo, se o quociente entre a variância anterior e a última for próximo de 1 –, podemos então concluir que a variação que estamos captando é sistemática em ambas as amostras (em outras

palavras, comum à população subjacente). Se o quociente for pequeno, podemos então concluir que nosso achado não pode ser generalizado para fora da amostra.

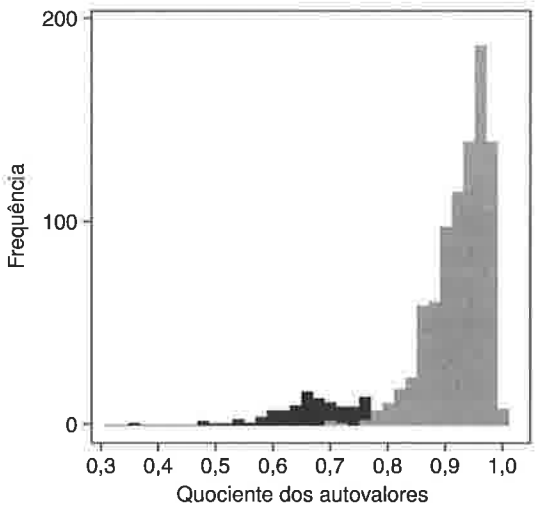

Figura 4.18 Histograma de 1.000 amostras bootstrapped dos dados do GSP. As barras negras denotam valores em que nem as observações de AK nem de WY aparecem na amostra bootstrapped.

A Figura 4.18 mostra a distribuição do quociente descrito acima das 1.000 amostras bootstrapped. Na média, a combinação linear formada usando-se os pesos da análise de componentes principais dos dados originais capta aproximadamente 90% da variância, tanto quanto o primeiro componente principal da amostra bootstrapped. Isso parece sugerir que os resultados são razoavelmente bem generalizáveis além da amostra. Entretanto, há também uma pequena cauda na distribuição (representada pela região escurecida do histograma na figura). Para aproximadamente 12% das amostras bootstrapped, obtemos um conjunto de observações sem a representação de Alasca e de Wyoming. Para essas amostras, o valor médio do quociente é de somente 67%. Isso sugere que as observações de Alasca e Wyoming não são apenas valores extremos, mas também atípicos em relação aos restantes 48 estados.

Neste capítulo, discutimos a validação no contexto da variância explicada pela solução dos componentes principais. No Capítulo 5, discutiremos mais amplamente a avaliação da validade da matriz de cargas, que é instrumental na formatação de nossas conclusões sobre os padrões de associação nos dados.

4.5 RESUMO DA APRENDIZAGEM

- *A análise dos componentes principais* é um método de se reexpressarem os dados multivariados. Ela permite que o pesquisador reoriente os dados para que algumas poucas primeiras dimensões expliquem tantas informações quanto possível. A análise dos componentes principais é também útil na identificação e compreensão dos padrões de associação entre as variáveis.
- O primeiro componente principal, representado por Z_1, é dado pela combinação linear das variáveis originais $\mathbf{X} = [X_1, X_2, \ldots, X_p]$ com a maior variância possível (em que a variância é interpretável como as informações contidas nos dados).
  - O segundo componente principal, representado por Z_2, é dado pela combinação linear de $\mathbf{X}$ que explica a maioria das informações (variância mais alta) não captada ainda por Z_1; isto é, Z_2 é escolhido para não ser correlacionado a Z_1.
  - Todos os componentes principais subsequentes são escolhidos para que não sejam correlacionados a todos os componentes principais anteriores.

- Em razão de a análise de componentes principais buscar maximizar a variância, ela pode ser altamente sensível às diferenças de escala entre variáveis. Portanto, normalmente (mas não sempre) é uma boa ideia padronizar os dados (a abordagem tratada neste capítulo) e representá-los por $\mathbf{X}_s$.
- A solução para o problema dos componentes principais enunciado anteriormente é obtida realizando-se uma decomposição de autovalor da matriz de correlação (isto é, a matriz de covariância dos dados padronizados). Intuitivamente, isso corresponde a encontrar os eixos principais do formato (com frequência, aproximadamente elipsoidal) adquirido pelo diagrama de dispersão dos dados.
  - Cada autovetor, indicado por $\mathbf{u}_i$, representa a direção de um desses eixos principais. O vetor $\mathbf{u}$ controla os pesos usados para formar a combinação linear de $\mathbf{X}_s$, que resulta nos pontos do componente principal; isto é, $\mathbf{z}_i = \mathbf{X}_s\mathbf{u}_i$.
  - Cada autovalor, representado por λ_i, é igual à variância do componente principal Z_i. Propositalmente, a solução é escolhida para que $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$.
  - A matriz de covariância para os componentes principais, representada por $\mathbf{D}$, é uma matriz diagonal com $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ na diagonal.
- A soma das variâncias de todos os componentes principais é igual a p, o número das variáveis na matriz $\mathbf{X}$. Portanto, a proporção da variação explicada pelos primeiros componentes principais c é dada por

$$\sum_{i-1}^{c} \frac{\lambda_i}{p}$$

- Quando se tenta interpretar os componentes principais, é útil olhar para as *cargas dos componentes principais*, que são as correlações entre as variáveis originais $\mathbf{X}$ e os componentes principais $\mathbf{Z}$. A matriz das cargas, representada por $\mathbf{F}$, pode ser calculada como

$$\mathbf{F} = \mathbf{U}\mathbf{D}^{1/2}$$

onde $\mathbf{U}$ é a matriz dos autovetores e $\mathbf{D}$ é a matriz de covariância dos componentes principais.
  - Cada componente principal é definido pelas variáveis com as quais é mais altamente correlacionado.
  - É possível também rotar a solução do componente principal para simplificar a estrutura da matriz de cargas e facilitar a interpretação. Trata-se da rotação no Capítulo 5.
- Em geral, só faz sentido usar a análise dos componentes principais quando os dados não são independentes (isto é, quando a matriz de correlação $\mathbf{R}$ é diferente da matriz identidade $\mathbf{I}$). Bartlett fornece um teste de qui-quadrado para determinar a esfericidade dos dados, representado por

$$\chi^2 = -\left[(n-1) - \frac{(2p+5)}{6}\right] \ln |\mathbf{R}|$$

com $(p^2 - p)/2$ graus de liberdade, onde p é o número de variáveis e n é o tamanho da amostra.
- Há várias regras de ouro a serem usadas na determinação do número de componentes principais retidos para análise posterior:
  - O *gráfico scree*. Esquematiza a variância explicada por cada componente principal (do maior para o menor) e retém todos os componentes principais acima do "ângulo" no gráfico.
  - *Regra de Kaiser*. Retém todos os componentes principais que explicam a quantidade mais do que proporcional da variância; isto é, todos os componentes em que $\lambda_i > 1$.
  - *Procedimento de Horn*. Retém todos os componentes principais em que a variância for maior que um benchmark baseado em dados aleatórios. Essa abordagem reterá menos componentes do que a regra de Kaiser porque leva em consideração a tendência dos componentes principais de capitalizar sobre a variação da amostragem idiossincrática.

- A proporção da variância em cada variável original X_i explicada pelos primeiros componentes principais c é dada pela soma das cargas do componente ao quadrado; isto é,

$$\sum_{k=1}^{c} f_{ik}^2$$

Às vezes, faz sentido reter um número suficiente de componentes principais para explicar adequadamente a variância em cada variável original.

- Como a análise dos componentes principais é baseada na explicação da variação nos dados, é altamente suscetível à presença de discrepâncias e observações influentes. É uma boa ideia usar a validação de dados (ou outros métodos, como *jackknifing* ou *bootstrapping*) para validar a solução.

LEITURAS SELECIONADAS

Geral

CARROLL, J. Douglas; GREEN, Paul E. *Mathematical tools for applied multivariate analysis.* Nova York: Academic Press, 1997. Com contribuições de Anil Chaturvedi.

DUNTEMAN, George H. "Principal components analysis." In: *Quantitative applications in the social science.* Newbury Park: Sage Publications, 1989.

HOTELLING, H. "Analysis of a complex of statistical variables into principal components", *Journal of Educational Psychology*, v. 24, p. 417-441, 498-520, 1933.

Quantos componentes?

BARTLETT, M. S. "Tests of significance of factor analysis", *British Journal of Psychology,* Statistical Section, v. 3, p. 77-85, 1950.

CATTELL, R.B. "The scree test for the number of factors", *Multivariate Behavioral Research*, v. 1, p. 245-276, 1966.

HORN, J. L. "A rationale and test for the number of factors in factor analysis", *Psychometrika*, v. 30, p. 179-186, 1965.

KAISER, H. F. "The application of electronic computers to factor analysis". In *Symposium on the Application of Computers to Psychological Problems*, American Psychological Association, 1959.

Jackknife e Bootstrapping

EFRON, Bradley; GONG, Gail. "A leisurely look at the bootstrap, the jackknife, and cross-validation", *American Statistician*, v. 37, p. 36-48, fev. 1983.

EXERCÍCIOS

4.1 Em análise de regressão, os pesquisadores às vezes usam o *índice de condição* (dado pela raiz quadrada da razão do maior e do menor autovalor da matriz de covariância das variáveis independentes no modelo de regressão) como um indicador de multicolinearidade problemática. Explique a lógica por trás de tal abordagem, usando como referência os componentes principais.

4.2 Os dados da ilustração do começo do capítulo (com variáveis X_1, X_2 e X_3) estão disponíveis no arquivo *PCA_EXAMPLE*. Faça sua própria análise de componentes principais usando esses dados e replique os resultados apresentados no capítulo.

4.3 Os dados sobre o produto estadual bruto (usado como exemplo de problema no capítulo) estão disponíveis em dois arquivos: *GSP_RAW* (representado em milhões de dólares) e *GSP_SHARE* (expresso como uma proporção do total do GSP para cada estado). Cada arquivo consiste em 50 linhas (uma para cada estado) e 14 colunas (a primeira contendo a identidade do estado e os restantes 13 correspondendo às diferentes áreas de atividade econômica listadas na Seção 4.3.1 do capítulo).

a. Usando os dados do *GSP_SHARE*, replique a análise dos componentes principais realizada no capítulo.

b. O que acontecerá se você remover as observações correspondentes ao Alasca e ao Wyoming? Repita a análise e esquematize os escores para os primeiros dois componentes principais. Como você interpretaria os resultados?

4.4 Um pesquisador coletou informações sobre 100 respondentes acerca das seis variáveis de opinião política apresentadas a seguir:

X_1 O governo deveria investir mais dinheiro em escolas
X_2 O governo deveria investir mais dinheiro para reduzir o desemprego
X_3 O governo deveria controlar os grandes negócios
X_4 O governo deveria acelerar o fim da discriminação racial através de transporte escolar
X_5 O governo deveria zelar para que as minorias obtenham suas respectivas quotas de emprego
X_6 O governo deveria expandir o programa Head Start*

A matriz de correlação para esses dados é apresentada na Tabela 4.15 e está disponível no arquivo *GOVERNMENT_1*.

Tabela 4.15 Matriz de correlação para os dados *GOVERNO_1*

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1,0000					
X_2	0,6008	1,0000				
X_3	0,4984	0,4749	1,0000			
X_4	0,1920	0,2196	0,2079	1,0000		
X_5	0,1959	0,1912	0,2010	0,4334	1,0000	
X_6	0,3466	0,2979	0,2445	0,3197	0,4207	1,0000

a. Analise esses dados usando a análise de componentes principais. Quanta variação é explicada pelos dois primeiros componentes principais? Como você interpretaria os dois componentes?

4.5 Um dos mais famosos conjuntos de dados em estatística são os dados da íris de Fisher. O conjunto (disponível no arquivo *IRIS*) contém medidas de 50 espécimes de cada uma das três espécies diferentes de íris – *Iris setosa*, *Iris versicolor* e *Iris virginica* – com as seguintes dimensões (as medidas estão em milímetros):

X_1 Espécies (1 = *Iris setosa*, 2 = *Iris versicolor*, 3 = *Iris virginica)*
X_2 Comprimento da sépala
X_3 Largura da sépala
X_4 Comprimento da pétala
X_5 Largura da pétala

a. Analise os dados em *IRIS* (variáveis $X_2 - X_5$) utilizando a análise de componentes principais. De quantos componentes você precisa para descrever adequadamente os dados? Como você os interpretaria?
b. Desenhe os escores médios do componente principal de cada um dos três diferentes tipos de íris para os primeiros dois componentes principais. Descreva as suas descobertas.

4.6 Os dados brutos colhidos por Ofir e Simonson (2001) para medir "necessidade de cognição" estão disponíveis no arquivo *COGNITION*. O arquivo de dados contém 19 variáveis: o número de identidade de um respondente e os escores para os 18 itens listados na Tabela 4.1.

a. Analise os dados para a necessidade de cognição usando os componentes principais. Quanta variância é explicada pelo primeiro componente? Sua análise sugere que um único componente realiza um trabalho adequado de captação da necessidade de cognição? Por que sim ou por que não?
b. Os pesquisadores, em geral, formam a escala da necessidade de cognição, adicionando os escores dos itens (itens de códigos reversos são subtraídos). Como essa medida se compara com o primeiro componente principal da análise no item "a"? Qual você usaria e por quê?

4.7 Golding e Seidman (1974) estudaram os interesses vocacionais de 231 estudantes universitários do sexo masculino. Cada respondente classificou a força de seus interesses em 22 áreas vocacionais, listadas a seguir:

* Programa para melhorar a vida de crianças de baixa renda, fornecendo serviços de qualidade abrangentes para o desenvolvimento infantil focados na família, incluindo educação, saúde, nutrição e saúde mental. http://www.headstart.net.index2.htm, acessado em 11.05.2010. (NT)

X_1	falar em público	X_9	matemática	X_{17}	serviço social
X_2	direito e política	X_{10}	ciências	X_{18}	atividades religiosas
X_3	gestão de negócio	X_{11}	mecânica	X_{19}	ensino
X_4	vendas	X_{12}	natureza	X_{20}	música
X_5	merchandising	X_{13}	agricultura	X_{21}	arte
X_6	trabalho em escritório	X_{14}	aventura	X_{22}	escrita
X_7	atividades militares	X_{15}	liderança recreacional		
X_8	supervisão técnica	X_{16}	serviço médico		

A matriz de correlação está disponível no arquivo *VOCATIONS*.

a. Analise os dados usando os componentes principais. Aparentemente, há mais de uma dimensão descrevendo os interesses vocacionais entre os estudantes masculinos? Como você descreveria a(s) dimensão(ões) subjacente(s)? Quais interesses vocacionais parecem assemelhar-se? Quais parecem mais diferentes?

4.8 O arquivo *RECORDS* contém dados sobre os recordes atléticos de cada um dos 55 países dos seguintes eventos:

X_1	100 m (em segundos)	X_5	1.500 m (min)
X_2	200 m (s)	X_6	5.000 m (min)
X_3	400 m (s)	X_7	10.000 m (min)
X_4	800 m (min)	X_8	maratona (min)

Analise esses dados usando componentes principais e responda às seguintes questões:

a. Quanta variação é explicada pelos dois primeiros componentes principais?
b. Quantos componentes principais são necessários para explicar 80% da variação nos dados originais?
c. Quantos componentes você extrairia? Como interpretaria os componentes?

4.9 Huba et al. (1981) coletaram dados sobre a taxa de uso de drogas junto a 1.634 estudantes entre a 7ª e 9ª séries em 11 escolas da área metropolitana da grande Los Angeles. Cada participante respondeu um questionário sobre o número de vezes em que uma substância particular havia sido usada. As respostas foram registradas em uma escala de cinco pontos: 1 = nunca experimentei, 2 = somente uma vez, 3 = algumas vezes, 4 = muitas vezes e 5 = regularmente. O estudo mediu o uso das seguintes substâncias:

X_1	cigarros	X_8	heroína ou outros opiáceos
X_2	cerveja	X_9	maconha
X_3	vinho	X_{10}	haxixe
X_4	bebida alcoólica	X_{11}	inalantes (cola, gasolina etc.)
X_5	cocaína	X_{12}	alucinógenos (LSD, mescalina etc.)
X_6	tranquilizantes	X_{13}	estimulantes anfetamínicos
X_7	medicamentos que causam euforia		

A matriz de correlação está disponível no arquivo *DRUG_USE* e é apresentada na Tabela 4.16.

O que a análise dos componentes principais diz sobre o uso dos diferentes tipos de drogas nessa população específica de estudantes?

4.10 Utilizando o jornal diário *The New York Times*, Rummel (1966) coletou dados sobre a existência de conflitos entre 341 pares de nações. Para cada par de países, Rummel registrou a incidência dos seguintes itens:

X_1	violência	X_9	comunicações escritas
X_2	violência planejada	X_{10}	comunicações orais
X_3	incidência de violência	X_{11}	comunicações escritas e orais
X_4	atos militares	X_{12}	comunicações não classificadas
X_5	dias de violência	X_{13}	acusações
X_6	atos negativos	X_{14}	protestos
X_7	rejeição diplomática	X_{15}	avisos
X_8	comunicações negativas	X_{16}	atos xenofóbicos

As entradas da matriz de correlação (disponível no arquivo *FOREIGN_CONFLICT*) refletem a extensão com que os pares tendem a aparecer juntos na descrição do estado de conflito entre dois países quaisquer.

a. Baseando-se nos dados de Rummel, há mais de uma dimensão distinta de conflitos estrangeiros? Como você descreveria e interpretaria os resultados de sua análise desses dados?

Tabela 4.16 Matriz de correlação para os dados de *DRUG_USE*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
X_1	1,000												
X_2	0,447	1,000											
X_3	0,422	0,619	1,000										
X_4	0,435	0,604	0,583	1,000									
X_5	0,114	0,068	0,053	0,115	1,000								
X_6	0,203	0,146	0,139	0,258	0,349	1,000							
X_7	0,091	0,103	0,110	0,122	0,209	0,221	1,000	1,000					
X_8	0,082	0,063	0,066	0,097	0,321	0,355	0,201	0,154	1,000				
X_9	0,513	0,445	0,365	0,482	0,186	0,315	0,150	0,219	0,534	1,000			
X_{10}	0,304	0,318	0,240	0,368	0,303	0,377	0,163	0,288	0,301	0,302	1,000		
X_{11}	0,245	0,203	0,183	0,255	0,272	0,323	0,310	0,320	0,204	0,368	0,340	1,000	
X_{12}	0,101	0,088	0,074	0,139	0,279	0,367	0,232	0,314	0,394	0,467	0,392	0,511	1,000
X_{13}	0,245	0,199	0,184	0,293	0,278	0,545	0,232	0,314	0,394	0,467	0,392	0,511	1,000