

# Análise Fatorial por Componentes Principais

*O amor e a verdade estão tão unidos entre si que é praticamente impossível separá-los.  
São como duas faces da mesma moeda.*

**Mahatma Gandhi**

Ao final deste capítulo, você terá condições de:

- Estabelecer as circunstâncias a partir das quais a técnica de análise fatorial por componentes principais pode ser utilizada.
- Entender o conceito de fator.
- Saber avaliar a adequação global da análise fatorial por meio da estatística KMO e do teste de esfericidade de Bartlett.
- Compreender os conceitos de autovalores e autovetores em matrizes de correlação de Pearson.
- Saber calcular e interpretar os *scores* fatoriais e, a partir dos mesmos, definir fatores.
- Determinar e interpretar cargas fatoriais e comunalidades.
- Construir *loading plots*.
- Entender os conceitos referentes à rotação de fatores e elaborar a rotação ortogonal Varimax.
- Construir *rankings* de desempenho a partir do comportamento conjunto de variáveis.
- Elaborar a técnica de análise fatorial por componentes principais de maneira algébrica e por meio do IBM SPSS Statistics Software® e do Stata Statistical Software® e interpretar seus resultados.

## 10.1. INTRODUÇÃO

As técnicas exploratórias de **análise fatorial** são muito úteis quando há a intenção de se trabalhar com variáveis que apresentem, entre si, **coeficientes de correlação** relativamente elevados e se deseja estabelecer novas variáveis que captem o comportamento conjunto das variáveis originais. Cada uma dessas novas variáveis é chamada de **fator**, que pode ser entendido como o **agrupamento de variáveis** a partir de critérios estabelecidos. Nesse sentido, a análise fatorial é uma técnica multivariada que procura identificar uma quantidade relativamente pequena de fatores que representam o comportamento conjunto de variáveis originais interdependentes. Assim, enquanto a análise de agrupamentos estudada no capítulo anterior faz uso de medidas de distância ou de semelhança para agrupar observações e formar *clusters*, a análise fatorial utiliza coeficientes de correlação para agrupar variáveis e gerar fatores.

Dentre os métodos para determinação de fatores, o conhecido como **componentes principais** é, sem dúvida, o mais utilizado em análise fatorial, já que se baseia no pressuposto de que podem ser extraídos **fatores não correlacionados** a partir de **combinações lineares das variáveis originais**. A análise fatorial por componentes principais permite, portanto, que, a partir de um conjunto de variáveis originais correlacionadas entre si, seja determinado outro conjunto de variáveis (fatores) resultantes da combinação linear do primeiro conjunto.

Embora na literatura, como sabemos, apareça com certa frequência o termo **análise fatorial confirmatória**, a análise fatorial é, em essência, uma **técnica multivariada exploratória**, ou de **interdependência**, visto que não possui caráter preditivo para outras observações não presentes inicialmente na amostra, e a inclusão de novas observações no banco de dados torna necessária a reaplicação da técnica, para que sejam gerados novos

fatores mais precisos e atualizados. Conforme discute Reis (2001), a análise fatorial pode ser utilizada tanto com o objetivo exploratório de redução da dimensão dos dados, com foco na criação de fatores a partir de variáveis originais, quanto com o objetivo de se confirmar uma hipótese inicial de que os dados poderão ser reduzidos a determinado fator, ou determinada dimensão, previamente estabelecido. Independentemente da natureza do objetivo, a análise fatorial continuará exploratória. Caso um pesquisador tenha a intenção de utilizar uma técnica para, de fato, confirmar as relações encontradas na análise fatorial, poderá fazer uso, por exemplo, de **modelos de equações estruturais**.

A análise fatorial por componentes principais apresenta quatro objetivos principais: (1) identificação de correlações entre variáveis originais para a criação de fatores que representam a combinação linear daquelas variáveis (**redução estrutural**); (2) verificação da **validade de constructos** previamente estabelecidos, tendo em vista a alocação das variáveis originais em cada fator; (3) **elaboração de rankings** por meio da criação de indicadores de desempenho a partir dos fatores; e (4) extração de fatores ortogonais para posterior uso em técnicas multivariadas confirmatórias que necessitam de **ausência de multicolinearidade**.

Imagine que um pesquisador tenha interesse em estudar a relação de interdependência entre diversas variáveis quantitativas que traduzem o comportamento socioeconômico dos municípios de uma nação. Nessa situação, podem ser determinados fatores que eventualmente consigam explicar o comportamento das variáveis originais, e, nesse sentido, a análise fatorial é utilizada para a redução estrutural dos dados e para posterior elaboração de um indicador socioeconômico que capte o comportamento conjunto dessas variáveis. A partir desse indicador, pode inclusive ser criado um *ranking* de desempenho dos municípios, e os próprios fatores podem ser utilizados em uma eventual análise de agrupamentos.

Em outra situação, fatores extraídos a partir de variáveis originais podem ser utilizados como variáveis explicativas de outra variável (dependente), inicialmente não considerada na análise. Por exemplo, fatores obtidos a partir do comportamento conjunto das notas escolares em determinadas disciplinas do último ano do ensino médio podem ser utilizados como variáveis explicativas da classificação geral dos estudantes no vestibular ou do fato de o estudante ter ou não sido aprovado. Note, nessas situações, que os fatores (ortogonais entre si) são utilizados, em vez das próprias variáveis originais, como variáveis explicativas de determinado fenômeno em modelos multivariados confirmatórios, como regressão múltipla ou regressão logística, a fim de que sejam eliminados eventuais problemas de multicolinearidade. É importante ressaltar, entretanto, que esse procedimento somente faz sentido quando há o intuito de elaborar um **diagnóstico** acerca do comportamento da variável dependente, sem a intenção de previsões para outras observações não presentes inicialmente na amostra. Como novas observações não apresentam os correspondentes valores dos fatores gerados, a obtenção desses valores somente é possível ao se incluírem tais observações em nova análise fatorial.

Em uma terceira situação, imagine que uma empresa varejista esteja interessada em avaliar o nível de satisfação dos clientes por meio da aplicação de um questionário em que as perguntas tenham sido previamente classificadas em determinados grupos. Por exemplo, as perguntas A, B e C foram classificadas no grupo *qualidade do atendimento*, as perguntas D e E, no grupo *percepção positiva de preços*, e as perguntas F, G, H e I, no grupo *variedade do sortimento de produtos*. Após a aplicação do questionário em uma amostra significativa de consumidores, em que essas nove variáveis são levantadas por meio da atribuição de notas que variam de 0 a 10, a empresa varejista decide elaborar uma análise fatorial por componentes principais para verificar se, de fato, a combinação das variáveis reflete o constructo previamente estabelecido. Se isso ocorrer, a análise fatorial terá sido utilizada para validar o constructo, apresentando objetivo de natureza confirmatória.

Podemos perceber, em todas essas situações, que as variáveis originais a partir das quais serão extraídos fatores são quantitativas, visto que a análise fatorial parte do estudo do comportamento dos coeficientes de correlação de Pearson entre as variáveis. É comum, entretanto, que pesquisadores façam uso do **incorreto procedimento de ponderação arbitrária** em variáveis qualitativas, como variáveis em **escala Likert**, para, a partir de então, ser aplicada uma análise fatorial. **Trata-se de um erro grave!** Existem técnicas exploratórias destinadas exclusivamente ao estudo do comportamento de variáveis qualitativas como, por exemplo, a análise de correspondência a ser estudada no próximo capítulo, e a análise fatorial definitivamente não se apresenta para tal finalidade!

Em um contexto histórico, o desenvolvimento da análise fatorial é devido, em parte, aos trabalhos pioneiros de Pearson (1896) e Spearman (1904). Enquanto Karl Pearson desenvolveu um tratamento matemático rigoroso acerca do que se convencionou chamar de correlação, Charles Edward Spearman publicou, no início do século

XX, um seminal trabalho em que eram avaliadas as inter-relações entre os desempenhos de estudantes em diversas disciplinas, como Francês, Inglês, Matemática e Música. Como as notas dessas disciplinas apresentavam forte correlação, Spearman propôs que *scores* oriundos de testes aparentemente incompatíveis compartilhavam um fator geral único, e estudantes que apresentavam boas notas possuíam algum componente psicológico ou de inteligência mais desenvolvido. De modo geral, Spearman destacou-se profundamente pela aplicação de métodos matemáticos e estudos de correlação para a análise da mente humana.

Décadas mais tarde, o estatístico matemático e influente teórico econômico Harold Hotelling convencionou chamar, em 1933, de *Principal Component Analysis* a análise que determina componentes a partir da maximização da variância de dados originais. Ainda na primeira metade do século XX, o psicólogo Louis Leon Thurstone, a partir da investigação sobre as ideias de Spearman e com base na aplicação de determinados testes psicológicos cujos resultados foram submetidos à análise fatorial, identificou sete aptidões primárias das pessoas: aptidões espaciais e visuais, compreensão verbal, fluidez verbal, rapidez perceptual, aptidão numérica, raciocínio e memória. Na psicologia, o termo *fatores mentais* é inclusive destinado a variáveis que apresentam maior influência sobre determinado comportamento.

Atualmente, a análise fatorial é utilizada em diversos campos do conhecimento, como marketing, economia, estratégia, finanças, contabilidade, atuária, engenharia, logística, psicologia, medicina, ecologia e bioestatística, entre outros.

A análise fatorial por componentes principais deve ser definida com base na teoria subjacente e na experiência do pesquisador, de modo que seja possível aplicar a técnica de forma correta e analisar os resultados obtidos.

Neste capítulo, trataremos da técnica de análise fatorial por componentes principais, com os seguintes objetivos: (1) introduzir os conceitos; (2) apresentar, de maneira algébrica e prática, o passo a passo da modelagem; (3) interpretar os resultados obtidos; e (4) propiciar a aplicação da técnica em SPSS e Stata. Seguindo a lógica proposta no livro, será inicialmente elaborada a solução algébrica de um exemplo vinculada à apresentação dos conceitos. Somente após a introdução dos conceitos, serão apresentados os procedimentos para a elaboração da técnica em SPSS e Stata.

## 10.2. ANÁLISE FATORIAL POR COMPONENTES PRINCIPAIS

Muitos são os procedimentos inerentes à análise fatorial, com diferentes métodos para a determinação (**extração**) de fatores a partir da matriz de correlações de Pearson. O método mais utilizado, adotado para a extração dos fatores neste capítulo, é conhecido por componentes principais, em que a consequente redução estrutural é também chamada de **transformação de Karhunen-Loève**.

Nas seções seguintes, apresentaremos o desenvolvimento teórico da técnica, bem como a elaboração de um exemplo prático. Enquanto nas seções 10.2.1 a 10.2.5 serão apresentados os principais conceitos, a seção 10.2.6 é destinada à resolução de um exemplo prático por meio de solução algébrica, a partir de um banco de dados.

### 10.2.1. Correlação linear de Pearson e conceito de fator

Imaginemos um banco de dados que apresente  $n$  observações e, para cada observação  $i$  ( $i = 1, \dots, n$ ), valores correspondentes a cada uma das  $k$  variáveis métricas  $X$ , conforme mostra a Tabela 10.1.

**Tabela 10.1** Modelo geral de um banco de dados para elaboração de análise fatorial.

Observação $i$	$X_{1i}$	$X_{2i}$	...	$X_{ki}$
1	$X_{11}$	$X_{21}$	...	$X_{k1}$
2	$X_{12}$	$X_{22}$		$X_{k2}$
3	$X_{13}$	$X_{23}$		$X_{k3}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$n$	$X_{1n}$	$X_{2n}$		$X_{kn}$

A partir do banco de dados, e dada a intenção de que sejam extraídos fatores a partir das  $k$  variáveis  $X$ , devemos definir a **matriz de correlações**  $\rho$  que apresenta os valores da **correlação linear de Pearson** entre cada par de variáveis, conforme mostra a expressão (10.1).

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} \quad (10.1)$$

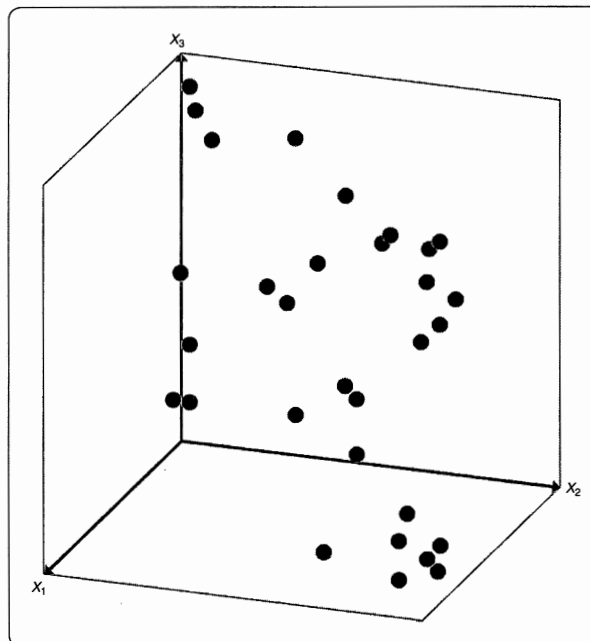
A matriz de correlações  $\rho$  é simétrica em relação à diagonal principal que, obviamente, apresenta valores iguais a 1. Para, por exemplo, as variáveis  $X_1$  e  $X_2$ , a correlação de Pearson  $\rho_{12}$  pode ser calculada com base na expressão (10.2).

$$\rho_{12} = \frac{\sum_{i=1}^n (X_{1i} - \bar{X}_1) \cdot (X_{2i} - \bar{X}_2)}{\sqrt{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2} \cdot \sqrt{\sum_{i=1}^n (X_{2i} - \bar{X}_2)^2}} \quad (10.2)$$

em que  $\bar{X}_1$  e  $\bar{X}_2$  representam, respectivamente, as médias das variáveis  $X_1$  e  $X_2$ .

Logo, como a correlação de Pearson é uma medida do grau da relação linear entre duas variáveis métricas, podendo variar entre  $-1$  e  $1$ , um valor mais próximo de um desses extremos indica a existência de relação linear entre as duas variáveis em análise, que, dessa forma, podem contribuir significativamente para a extração de um único fator. Por outro lado, um valor da correlação de Pearson muito próximo de  $0$  indica que a relação linear entre as duas variáveis é praticamente inexistente; portanto, diferentes fatores podem ser extraídos.

Imaginemos uma situação hipotética em que determinado banco de dados apresente apenas três variáveis ( $k = 3$ ). Um gráfico de dispersão tridimensional pode ser elaborado a partir dos valores de cada variável para cada observação. O gráfico encontra-se, de maneira exemplificada, na Figura 10.1.



**Figura 10.1** Gráfico de dispersão tridimensional para situação hipotética com três variáveis.

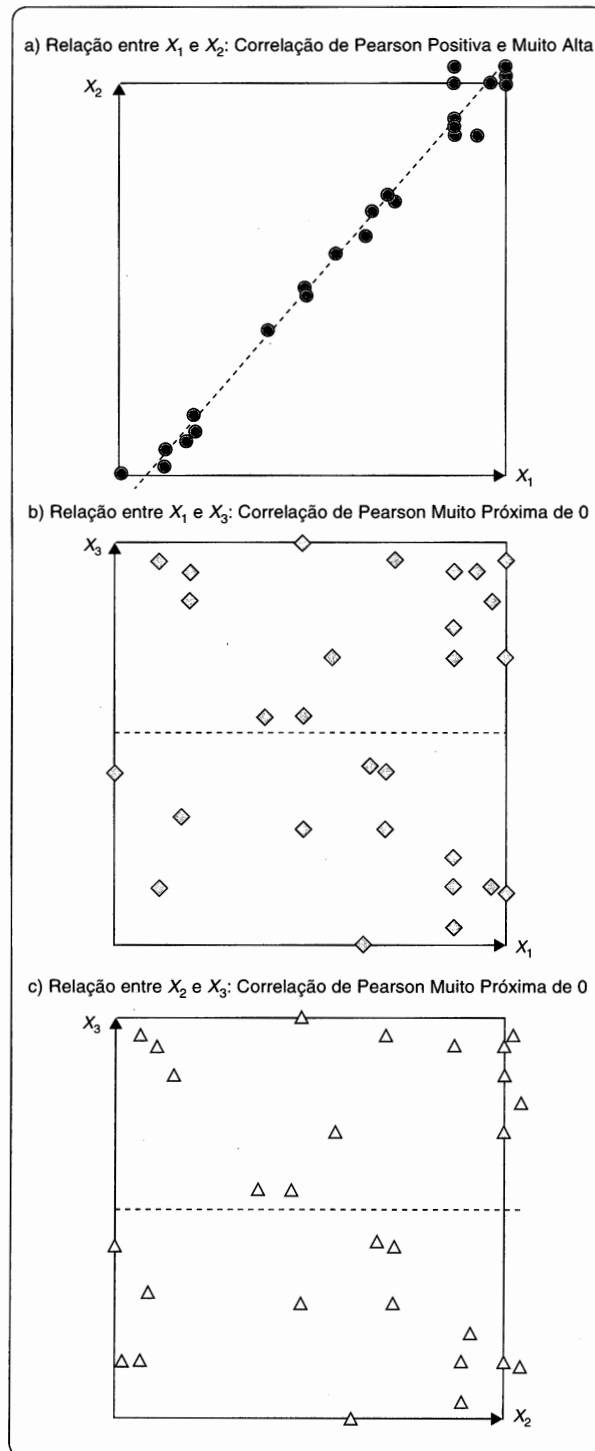
Com base apenas na análise visual do gráfico da Figura 10.1, é difícil avaliar o comportamento das relações lineares entre cada par de variáveis. Nesse sentido, a Figura 10.2 apresenta a projeção dos pontos correspondentes a cada observação em cada um dos planos formados pelos pares de variáveis, com destaque, em tracejado, para o ajuste que representa a relação linear entre as respectivas variáveis.

Enquanto a Figura 10.2a mostra que existe considerável relação linear entre as variáveis  $X_1$  e  $X_2$  (correlação de Pearson muito alta), as Figuras 10.2b e 10.2c explicitam que não existe relação linear entre  $X_3$  e essas variáveis.

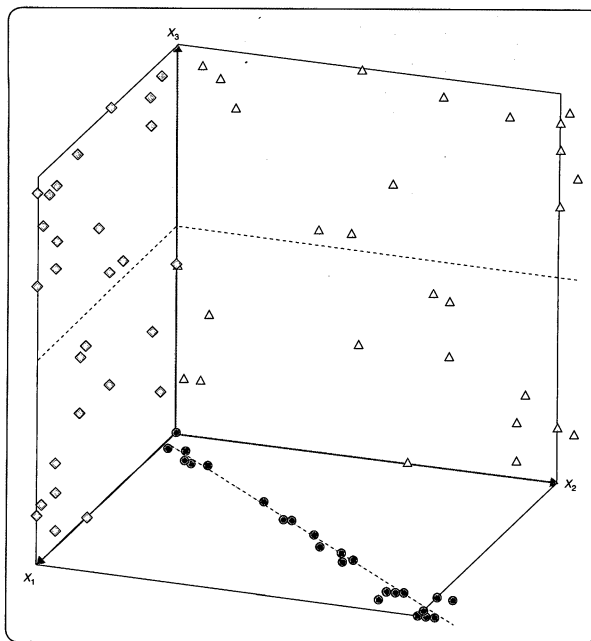
A Figura 10.3 mostra essas projeções no gráfico tridimensional, com os respectivos ajustes lineares em cada plano (retas tracejadas).

Dessa forma, nesse exemplo hipotético, enquanto as variáveis  $X_1$  e  $X_2$  poderão ser representadas de maneira bastante significativa por um único fator, que chamaremos de  $F_1$ , a variável  $X_3$  poderá ser representada por outro fator,  $F_2$ , ortogonal a  $F_1$ . A Figura 10.4 apresenta, de maneira tridimensional, a extração desses novos fatores.

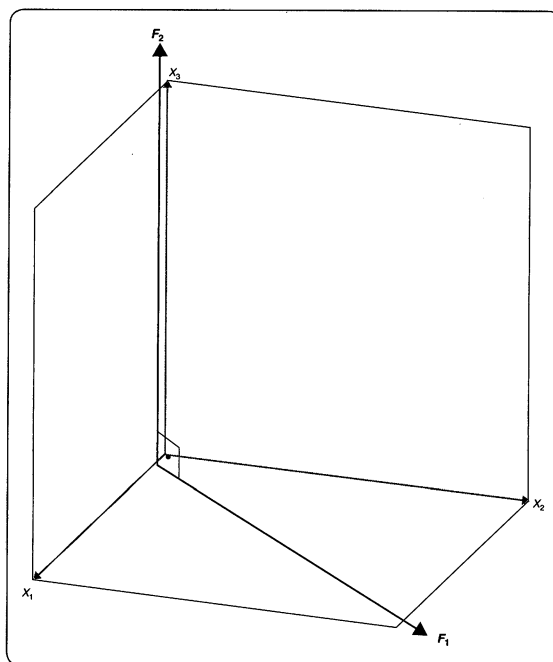
Logo, os fatores podem ser entendidos como **representações de dimensões latentes** que explicam o comportamento de variáveis originais.



**Figura 10.2** Projeção dos pontos em cada plano formado por determinado par de variáveis.



**Figura 10.3** Projeção dos pontos em gráfico tridimensional com ajustes lineares por plano.



**Figura 10.4** Extração de fatores.

Apresentados esses conceitos iniciais, é importante salientar que, em muitos casos, o pesquisador pode optar por não extrair um fator representado de maneira considerável por apenas uma variável (neste caso, o fator  $F_2$ ), e o que vai definir a extração de cada um dos fatores é o cálculo dos autovalores da matriz de correlações  $\rho$ , conforme será estudado na seção 10.2.3. Antes disso, entretanto, será necessário que se verifique a **adequação global da análise fatorial**, a ser discutida na próxima seção.

### 10.2.2. Adequação global da análise fatorial: estatística Kaiser-Meyer-Olkin (KMO) e teste de esfericidade de Bartlett

Uma adequada extração de fatores a partir de variáveis originais requer que a matriz de correlações  $\rho$  apresente valores relativamente elevados e estatisticamente significantes. Conforme discutem Hair *et al.* (2009), embora a inspeção visual da matriz de correlações  $\rho$  não revele se a extração de fatores será, de fato, adequada, uma

quantidade substancial de valores inferiores a 0,30 representa um preliminar indício de que a análise fatorial poderá ser inapropriada.

Para que seja verificada a adequação global propriamente dita da extração dos fatores, devemos recorrer à **estatística Kaiser-Meyer-Olkin (KMO)** e ao **teste de esfericidade de Bartlett**.

A estatística KMO fornece a proporção de variância considerada comum a todas as variáveis na amostra em análise, ou seja, que pode ser atribuída à existência de um fator comum. Essa estatística varia de 0 a 1, e, enquanto valores mais próximos de 1 indicam que as variáveis compartilham um percentual de variância bastante elevado (correlações de Pearson altas), valores mais próximos de 0 são decorrentes de correlações de Pearson baixas entre as variáveis, o que pode indicar que a análise fatorial será inadequada. A estatística KMO, apresentada inicialmente por Kaiser (1970), pode ser calculada por meio da expressão (10.3).

$$KMO = \frac{\sum_{l=1}^k \sum_{c=1}^k \rho_{lc}^2}{\sum_{l=1}^k \sum_{c=1}^k \rho_{lc}^2 + \sum_{l=1}^k \sum_{c=1}^k \varphi_{lc}^2}, l \neq c \quad (10.3)$$

em que  $l$  e  $c$  representam, respectivamente, as linhas e colunas da matriz de correlações  $\rho$ , e os termos  $\varphi$  representam os **coeficientes de correlação parcial** entre duas variáveis. Enquanto os coeficientes de correlação de Pearson  $\rho$  são também chamados de **coeficientes de correlação de ordem zero**, os coeficientes de correlação parcial  $\varphi$  são também conhecidos por **coeficientes de correlação de ordem superior**. Para três variáveis, são também chamados de **coeficientes de correlação de primeira ordem**, para quatro variáveis, de **coeficientes de correlação de segunda ordem** e assim sucessivamente.

Imaginemos outra situação hipotética em que determinado banco de dados apresenta novamente três variáveis ( $k = 3$ ). **É possível que  $\rho_{12}$  reflita, de fato, o grau de relação linear entre  $X_1$  e  $X_2$ , estando a variável  $X_3$  relacionada com as outras duas?** Nessa situação,  $\rho_{12}$  pode não representar o verdadeiro grau de relação linear entre  $X_1$  e  $X_2$  na presença de  $X_3$ , o que pode fornecer uma falsa impressão sobre a natureza da relação entre as duas primeiras. É nesse sentido que os coeficientes de correlação parcial podem contribuir com a análise, visto que, segundo Gujarati e Porter (2008), são utilizados quando se deseja conhecer a correlação entre duas variáveis, controlando-se ou desconsiderando-se os efeitos de outras variáveis presentes na base de dados. Para nossa situação hipotética, é o coeficiente de correlação independente da influência, se é que ela existe, de  $X_3$  sobre  $X_1$  e  $X_2$ .

Dessa maneira, para três variáveis  $X_1$ ,  $X_2$  e  $X_3$ , podemos definir da seguinte forma os coeficientes de correlação de primeira ordem:

$$\varphi_{12,3} = \frac{\rho_{12} - \rho_{13} \cdot \rho_{23}}{\sqrt{(1 - \rho_{13}^2) \cdot (1 - \rho_{23}^2)}} \quad (10.4)$$

em que  $\varphi_{12,3}$  representa a correlação entre  $X_1$  e  $X_2$ , mantendo-se  $X_3$  constante,

$$\varphi_{13,2} = \frac{\rho_{13} - \rho_{12} \cdot \rho_{23}}{\sqrt{(1 - \rho_{12}^2) \cdot (1 - \rho_{23}^2)}} \quad (10.5)$$

em que  $\varphi_{13,2}$  representa a correlação entre  $X_1$  e  $X_3$ , mantendo-se  $X_2$  constante, e

$$\varphi_{23,1} = \frac{\rho_{23} - \rho_{12} \cdot \rho_{13}}{\sqrt{(1 - \rho_{12}^2) \cdot (1 - \rho_{13}^2)}} \quad (10.6)$$

em que  $\varphi_{23,1}$  representa a correlação entre  $X_2$  e  $X_3$ , mantendo-se  $X_1$  constante.

De maneira geral, um coeficiente de correlação de primeira ordem pode ser obtido por meio da seguinte expressão:

$$\varphi_{ab,c} = \frac{\rho_{ab} - \rho_{ac} \cdot \rho_{bc}}{\sqrt{(1 - \rho_{ac}^2) \cdot (1 - \rho_{bc}^2)}} \quad (10.7)$$

em que  $a$ ,  $b$  e  $c$  podem assumir valores 1, 2 ou 3, correspondentes às três variáveis em análise.

Já, para uma situação em que estejam presentes na análise quatro variáveis, a expressão geral de determinado coeficiente de correlação parcial (coeficiente de correlação de segunda ordem) é dada por:

$$\varphi_{ab,cd} = \frac{\varphi_{ab,c} - \varphi_{ad,c} \cdot \varphi_{bd,c}}{\sqrt{(1 - \varphi_{ad,c}^2) \cdot (1 - \varphi_{bd,c}^2)}} \quad (10.8)$$

em que  $\varphi_{ab,cd}$  representa a correlação entre  $X_a$  e  $X_b$ , mantendo-se  $X_c$  e  $X_d$  constantes, sabendo-se que  $a, b, c$  e  $d$  podem assumir valores 1, 2, 3 ou 4, correspondentes às quatro variáveis em análise.

A obtenção de coeficientes de correlação de ordens superiores, em que são consideradas na análise cinco ou mais variáveis, deverá ser feita sempre com base na determinação dos coeficientes de correlação parcial de ordens mais baixas. Na seção 10.2.6, elaboraremos um exemplo prático com a utilização de quatro variáveis, em que a solução algébrica da estatística KMO será obtida por meio da expressão (10.8).

É importante ressaltar que, mesmo que o coeficiente de correlação de Pearson entre duas variáveis seja 0, o coeficiente de correlação parcial entre elas pode não ser igual a 0, dependendo dos valores dos coeficientes de correlação de Pearson entre cada uma dessas variáveis e as demais presentes na base de dados.

Para que uma análise fatorial seja considerada adequada, os coeficientes de correlação parcial entre as variáveis devem ser baixos. Esse fato denota que as variáveis compartilham um percentual de variância elevado, e a desconsideração de uma ou mais delas na análise pode prejudicar a qualidade da extração dos fatores. Neste sentido, o Quadro 10.1 apresenta, segundo critério já bastante aceito na literatura, um indicativo sobre a relação entre a estatística KMO e a adequação global da análise fatorial.

**Quadro 10.1** Relação entre a estatística KMO e a adequação global da análise fatorial.

Estatística KMO	Adequação Global da Análise Fatorial
Entre 1,00 e 0,90	Muito boa
Entre 0,90 e 0,80	Boa
Entre 0,80 e 0,70	Média
Entre 0,70 e 0,60	Razoável
Entre 0,60 e 0,50	Má
Menor do que 0,50	Inaceitável

Já o teste de esfericidade de Bartlett (Bartlett, 1954) consiste em comparar a matriz de correlações  $\rho$  com uma matriz identidade  $\mathbf{I}$  de mesma dimensão. Se as diferenças entre os valores correspondentes fora da diagonal principal de cada matriz não forem estatisticamente diferentes de 0, a determinado nível de significância, poderemos considerar que a extração dos fatores não será adequada. Nesse caso, em outras palavras, as correlações de Pearson entre cada par de variáveis são estatisticamente iguais a 0, o que inviabiliza qualquer tentativa de extração de fatores a partir de variáveis originais. Logo, podemos definir as hipóteses nula e alternativa do teste de esfericidade de Bartlett da seguinte maneira:

$$H_0: \rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} = \mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$H_1: \rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} \neq \mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

A estatística correspondente ao teste de esfericidade de Bartlett é uma estatística  $\chi^2$ , que apresenta a seguinte expressão:



$$\chi^2_{\text{Bartlett}} = - \left[ (n-1) - \left( \frac{2 \cdot k + 5}{6} \right) \right] \cdot \ln |D| \quad (10.9)$$

com  $\frac{k \cdot (k-1)}{2}$  graus de liberdade. Sabemos que  $n$  é o tamanho da amostra, e  $k$ , o número de variáveis. Além disso,  $D$  representa o determinante da matriz de correlações  $\rho$ .

O teste de esfericidade de Bartlett permite, portanto, que verifiquemos, para determinado número de graus de liberdade e determinado nível de significância, se o valor total da estatística  $\chi^2_{\text{Bartlett}}$  é maior que o valor crítico da estatística. Se for o caso, poderemos afirmar que as correlações de Pearson entre os pares de variáveis são estatisticamente diferentes de 0 e que, portanto, podem ser extraídos fatores a partir das variáveis originais, sendo a análise fatorial apropriada. Quando da elaboração de um exemplo prático, na seção 10.2.6, também apresentaremos os cálculos da estatística  $\chi^2_{\text{Bartlett}}$  e o resultado do teste de esfericidade de Bartlett.

Ressalta-se que **deve ser sempre preferido o teste de esfericidade de Bartlett à estatística KMO para efeitos de decisão sobre a adequação global da análise fatorial**, visto que, enquanto o primeiro é um teste com determinado nível de significância, o segundo é apenas um coeficiente (estatística) calculado sem distribuição de probabilidades determinada e hipóteses que permitam avaliar o nível correspondente de significância para efeitos de decisão.

Além disso, é importante mencionarmos que, para apenas duas variáveis originais, a estatística KMO será sempre igual a 0,50, ao passo que a estatística  $\chi^2_{\text{Bartlett}}$  poderá indicar a rejeição ou não da hipótese nula do teste de esfericidade, dependendo da magnitude da correlação de Pearson entre as duas variáveis. Logo, enquanto a estatística KMO será 0,50 nessas situações, será o teste de esfericidade de Bartlett que permitirá que o pesquisador decida sobre a extração ou não de um fator a partir das duas variáveis originais. Já, para três variáveis originais, é muito comum que o pesquisador extraia dois fatores com significância estatística do teste de esfericidade de Bartlett, porém com estatística KMO menor que 0,50. Essas duas situações enfatizam ainda mais a maior relevância do teste de esfericidade de Bartlett em relação à estatística KMO para efeitos de tomada de decisão.

Por fim, vale mencionar que comumente encontramos na literatura a recomendação de que seja estudada a magnitude da medida conhecida por **alpha de Cronbach**, de forma anterior ao estudo da adequação global da análise fatorial, a fim de que seja avaliada a fidedignidade com que um fator pode ser extraído a partir de variáveis originais. Ressaltamos que o alpha de Cronbach oferece ao pesquisador indícios apenas sobre a consistência interna das variáveis do banco de dados para que seja extraído um único fator. Assim, sua determinação não representa um requisito obrigatório para a elaboração da análise fatorial, visto que essa técnica permite a extração de mais fatores. Entretanto, para efeitos didáticos, discutiremos os principais conceitos sobre o alpha de Cronbach no apêndice deste capítulo, com determinação algébrica e correspondentes aplicações nos softwares SPSS e Stata.

Discutidos esses conceitos e verificada a adequação global da análise fatorial, podemos partir para a definição dos fatores.

### 10.2.3. Definição dos fatores por componentes principais: determinação dos autovalores e autovetores da matriz de correlações $\rho$ e cálculo dos scores fatoriais

Como um fator representa a combinação linear de variáveis originais, podemos definir, para  $k$  variáveis, um número máximo de  $k$  fatores ( $F_1, F_2, \dots, F_k$ ), de maneira análoga à quantidade máxima de agrupamentos que podem ser definidos a partir de uma amostra com  $n$  observações, conforme estudamos no capítulo anterior, visto que um fator também pode ser entendido com o resultado do **agrupamento de variáveis**. Dessa forma, para  $k$  variáveis, temos:

$$\begin{aligned} F_{1i} &= s_{11} \cdot X_{1i} + s_{21} \cdot X_{2i} + \dots + s_{k1} \cdot X_{ki} \\ F_{2i} &= s_{12} \cdot X_{1i} + s_{22} \cdot X_{2i} + \dots + s_{k2} \cdot X_{ki} \\ &\vdots \\ F_{ki} &= s_{1k} \cdot X_{1i} + s_{2k} \cdot X_{2i} + \dots + s_{kk} \cdot X_{ki} \end{aligned} \quad (10.10)$$

em que os termos  $s$  são conhecidos por **scores fatoriais**, que representam os parâmetros de um modelo linear que relaciona determinado fator com as variáveis originais. O cálculo dos *scores* fatoriais é de fundamental

importância dentro do contexto da técnica de análise fatorial e é elaborado a partir da determinação dos autovalores e autovetores da matriz de correlações  $\rho$ . Na expressão (10.11), reproduzimos a matriz de correlações  $\rho$  já apresentada na expressão (10.1).

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} \quad (10.11)$$

Essa matriz de correlações, com dimensões  $k \times k$ , apresenta  $k$  autovalores  $\lambda^2$  ( $\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_k^2$ ), que podem ser obtidos a partir da solução da seguinte equação:

$$\det(\lambda^2 \cdot \mathbf{I} - \rho) = 0 \quad (10.12)$$

em que  $\mathbf{I}$  é a matriz identidade, também com dimensões  $k \times k$ .

Como determinado fator representa o resultado do agrupamento de variáveis, é importante ressaltar que:

$$\lambda_1^2 + \lambda_2^2 + \dots + \lambda_k^2 = k \quad (10.13)$$

A expressão (10.12) pode ser reescrita da seguinte maneira:

$$\begin{vmatrix} \lambda^2 - 1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda^2 - 1 & \cdots & -\rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda^2 - 1 \end{vmatrix} = 0 \quad (10.14)$$

de onde podemos definir a matriz de autovalores  $\Lambda^2$  da seguinte forma:

$$\Lambda^2 = \begin{pmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k^2 \end{pmatrix} \quad (10.15)$$

Para que sejam definidos os autovetores da matriz  $\rho$  com base nos autovalores, devemos resolver os seguintes sistemas de equações para cada autovalor  $\lambda^2$  ( $\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_k^2$ ):

- Determinação de Autovetores  $v_{11}, v_{21}, \dots, v_{k1}$  a partir do Primeiro Autovalor ( $\lambda_1^2$ ):

$$\begin{pmatrix} \lambda_1^2 - 1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda_1^2 - 1 & \cdots & -\rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda_1^2 - 1 \end{pmatrix} \cdot \begin{pmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{k1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (10.16)$$

de onde vem que:

$$\begin{cases} (\lambda_1^2 - 1) \cdot v_{11} - \rho_{12} \cdot v_{21} \cdots - \rho_{1k} \cdot v_{k1} = 0 \\ -\rho_{21} \cdot v_{11} + (\lambda_1^2 - 1) \cdot v_{21} \cdots - \rho_{2k} \cdot v_{k1} = 0 \\ \vdots \\ -\rho_{k1} \cdot v_{11} - \rho_{k2} \cdot v_{21} \cdots + (\lambda_1^2 - 1) \cdot v_{k1} = 0 \end{cases} \quad (10.17)$$

- Determinação de Autovetores  $v_{12}, v_{22}, \dots, v_{k2}$  a partir do Segundo Autovalor ( $\lambda_2^2$ ):

$$\begin{pmatrix} \lambda_2^2 - 1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda_2^2 - 1 & \cdots & -\rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda_2^2 - 1 \end{pmatrix} \cdot \begin{pmatrix} v_{12} \\ v_{22} \\ \vdots \\ v_{k2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (10.18)$$

de onde vem que:

$$\begin{cases} (\lambda_2^2 - 1) \cdot v_{12} - \rho_{12} \cdot v_{22} \cdots - \rho_{1k} \cdot v_{k2} = 0 \\ -\rho_{21} \cdot v_{12} + (\lambda_2^2 - 1) \cdot v_{22} \cdots - \rho_{2k} \cdot v_{k2} = 0 \\ \vdots \\ -\rho_{k1} \cdot v_{12} - \rho_{k2} \cdot v_{22} \cdots + (\lambda_2^2 - 1) \cdot v_{k2} = 0 \end{cases} \quad (10.19)$$

- Determinação de Autovetores  $v_{1k}, v_{2k}, \dots, v_{kk}$  a partir do  $k$ -ésimo Autovalor ( $\lambda_k^2$ ):

$$\begin{pmatrix} \lambda_k^2 - 1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda_k^2 - 1 & \cdots & -\rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda_k^2 - 1 \end{pmatrix} \cdot \begin{pmatrix} v_{1k} \\ v_{2k} \\ \vdots \\ v_{kk} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (10.20)$$

de onde vem que:

$$\begin{cases} (\lambda_k^2 - 1) \cdot v_{1k} - \rho_{12} \cdot v_{2k} \cdots - \rho_{1k} \cdot v_{kk} = 0 \\ -\rho_{21} \cdot v_{1k} + (\lambda_k^2 - 1) \cdot v_{2k} \cdots - \rho_{2k} \cdot v_{kk} = 0 \\ \vdots \\ -\rho_{k1} \cdot v_{1k} - \rho_{k2} \cdot v_{2k} \cdots + (\lambda_k^2 - 1) \cdot v_{kk} = 0 \end{cases} \quad (10.21)$$

Dessa forma, podemos calcular os *scores* fatoriais de cada fator com base na determinação dos autovalores e autovetores da matriz de correlações  $\rho$ . Os vetores dos *scores* fatoriais podem ser definidos da seguinte forma:

- *Scores* Fatoriais do Primeiro Fator:

$$\mathbf{S}_1 = \begin{pmatrix} s_{11} \\ s_{21} \\ \vdots \\ s_{k1} \end{pmatrix} = \begin{pmatrix} \frac{v_{11}}{\sqrt{\lambda_1^2}} \\ \frac{v_{21}}{\sqrt{\lambda_1^2}} \\ \vdots \\ \frac{v_{k1}}{\sqrt{\lambda_1^2}} \end{pmatrix} \quad (10.22)$$

- *Scores* Fatoriais do Segundo Fator:

$$\mathbf{S}_2 = \begin{pmatrix} s_{12} \\ s_{22} \\ \vdots \\ s_{k2} \end{pmatrix} = \begin{pmatrix} \frac{v_{12}}{\sqrt{\lambda_2^2}} \\ \frac{v_{22}}{\sqrt{\lambda_2^2}} \\ \vdots \\ \frac{v_{k2}}{\sqrt{\lambda_2^2}} \end{pmatrix} \quad (10.23)$$

- Scores Fatoriais do  $k$ -ésimo Fator:

$$\mathbf{S}_k = \begin{pmatrix} s_{1k} \\ s_{2k} \\ \vdots \\ s_{kk} \end{pmatrix} = \begin{pmatrix} \frac{v_{1k}}{\sqrt{\lambda_k^2}} \\ \frac{v_{2k}}{\sqrt{\lambda_k^2}} \\ \vdots \\ \frac{v_{kk}}{\sqrt{\lambda_k^2}} \end{pmatrix} \quad (10.24)$$

Como os *scores* fatoriais de cada fator são padronizados pelos respectivos autovalores, os fatores do conjunto de equações apresentado na expressão (10.10) devem ser obtidos pela multiplicação de cada *score* fatorial pela correspondente variável original, padronizada por meio do procedimento *Zscores*. Dessa forma, podemos obter cada um dos fatores com base nas seguintes equações:

$$\begin{aligned} F_{1i} &= \frac{v_{11}}{\sqrt{\lambda_1^2}} \cdot ZX_{1i} + \frac{v_{21}}{\sqrt{\lambda_1^2}} \cdot ZX_{2i} + \dots + \frac{v_{k1}}{\sqrt{\lambda_1^2}} \cdot ZX_{ki} \\ F_{2i} &= \frac{v_{12}}{\sqrt{\lambda_2^2}} \cdot ZX_{1i} + \frac{v_{22}}{\sqrt{\lambda_2^2}} \cdot ZX_{2i} + \dots + \frac{v_{k2}}{\sqrt{\lambda_2^2}} \cdot ZX_{ki} \\ F_{ki} &= \frac{v_{1k}}{\sqrt{\lambda_k^2}} \cdot ZX_{1i} + \frac{v_{2k}}{\sqrt{\lambda_k^2}} \cdot ZX_{2i} + \dots + \frac{v_{kk}}{\sqrt{\lambda_k^2}} \cdot ZX_{ki} \end{aligned} \quad (10.25)$$

em que  $ZX_i$  representa o valor padronizado de cada variável  $X$  para determinada observação  $i$ . Ressalta-se que todos os fatores extraídos apresentam, entre si, correlações de Pearson iguais a 0, ou seja, **são ortogonais entre si**.

Um pesquisador mais atento notará que os *scores* fatoriais de cada fator correspondem exatamente aos parâmetros estimados de um **modelo de regressão linear múltipla** que apresenta, como variável dependente, o próprio fator e, como variáveis explicativas, as variáveis padronizadas.

Matematicamente, é possível ainda verificar a relação existente entre os autovetores, a matriz de correlações  $\rho$  e a matriz de autovalores  $\Lambda^2$ . Logo, definindo-se a matriz de autovetores  $\mathbf{V}$  da seguinte forma:

$$\mathbf{V} = \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1k} \\ v_{21} & v_{22} & \dots & v_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & \dots & v_{kk} \end{pmatrix} \quad (10.26)$$

podemos comprovar que:

$$\mathbf{V}' \cdot \rho \cdot \mathbf{V} = \Lambda^2 \quad (10.27)$$

ou:

$$\begin{pmatrix} v_{11} & v_{21} & \dots & v_{k1} \\ v_{12} & v_{22} & \dots & v_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1k} & v_{2k} & \dots & v_{kk} \end{pmatrix} \cdot \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \dots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \dots & 1 \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1k} \\ v_{21} & v_{22} & \dots & v_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & \dots & v_{kk} \end{pmatrix} = \begin{pmatrix} \lambda_1^2 & 0 & \dots & 0 \\ 0 & \lambda_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_k^2 \end{pmatrix} \quad (10.28)$$

Na seção 10.2.6 apresentaremos um exemplo prático a partir do qual essa relação poderá ser verificada.

Enquanto na seção 10.2.2, discutimos a adequação global da análise fatorial, nesta seção apresentamos os procedimentos para a extração dos fatores, no caso de a técnica se mostrar apropriada. Mesmo sabendo, para  $k$  variáveis, que o número máximo de fatores é também igual a  $k$ , é de fundamental importância que o pesquisador defina, com base em determinado critério, a quantidade adequada de fatores que, de fato, representam as variáveis originais. Em nosso exemplo hipotético da seção 10.2.1, vimos que apenas dois fatores ( $F_1$  e  $F_2$ ) seriam suficientes para representar as três variáveis originais ( $X_1$ ,  $X_2$  e  $X_3$ ).

Embora o pesquisador tenha liberdade para definir, de forma preliminar, a quantidade de fatores a serem extraídos na análise, visto que pode ter a intenção de verificar, por exemplo, a validade de um constructo previamente estabelecido (procedimento conhecido por **critério a priori**), é de fundamental importância que seja feita uma análise com base na magnitude dos autovalores calculados a partir da matriz de correlações  $\rho$ .

Como os autovalores correspondem ao percentual de variância compartilhada pelas variáveis originais para a formação de cada fator, conforme discutiremos na seção 10.2.4, como  $\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_k^2$  e sabendo-se que os fatores  $F_1, F_2, \dots, F_k$  são obtidos a partir dos respectivos autovalores, fatores extraídos a partir de autovalores menores são formados a partir de menores percentuais de variância compartilhada pelas variáveis originais. Visto que um fator representa determinado agrupamento de variáveis, fatores extraídos a partir de autovalores menores que 1 possivelmente não conseguem representar o comportamento de sequer uma variável original (claro que para a regra existem exceções, que ocorrem para os casos em que determinado autovalor é menor mas muito próximo a 1). O critério de escolha da quantidade de fatores, em que são levados em consideração apenas os fatores correspondentes a autovalores maiores que 1, é comumente utilizado e conhecido por **critério da raiz latente** ou **critério de Kaiser**.

O método para a extração de fatores apresentado neste capítulo é conhecido como componentes principais, e o primeiro fator  $F_1$ , formado pelo maior percentual de variância compartilhada pelas variáveis originais, é também chamado de **fator principal**. Esse método é profundamente referenciado na literatura e utilizado na prática quando o pesquisador deseja elaborar uma redução estrutural dos dados para a criação de fatores ortogonais, definir *rankings* de observações por meio dos fatores gerados e até mesmo verificar a validade de constructos previamente estabelecidos. Outros métodos para extração dos fatores, como aqueles conhecidos por **mínimos quadrados generalizados**, **mínimos quadrados ponderados**, **máxima verossimilhança**, **alpha factoring** e **image factoring**, apresentam diferentes critérios e determinadas particularidades e, embora também possam ser encontrados na literatura, não serão abordados neste livro.

Além disso, é comum que se discuta sobre a necessidade de que a análise fatorial seja aplicada a variáveis que apresentem **normalidade multivariada** dos dados, para que haja consistência quando da determinação dos *scores* fatoriais. Entretanto, é importante ressaltar que a normalidade multivariada é uma suposição bastante rígida, sendo necessária somente para alguns métodos de extração dos fatores, como o método de máxima verossimilhança. A maioria dos métodos de extração de fatores não requer a suposição de normalidade multivariada dos dados e, conforme discute Gorsuch (1983), a análise fatorial por componentes principais parece ser, na prática, bastante robusta contra violações de normalidade.

#### 10.2.4. Cargas fatoriais e comunalidades

Estabelecidos os fatores, podemos definir as **cargas fatoriais**, que nada mais são que **correlações de Pearson entre as variáveis originais e cada um dos fatores**. A Tabela 10.2 apresenta as cargas fatoriais para cada par variável-fator.

**Tabela 10.2** Cargas fatoriais entre variáveis originais e fatores.

Variável \ Fator	$F_1$	$F_2$	...	$F_k$
$X_1$	$c_{11}$	$c_{12}$	...	$c_{1k}$
$X_2$	$c_{21}$	$c_{22}$		$c_{2k}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$X_k$	$c_{k1}$	$c_{k2}$		$c_{kk}$

Com base no critério da raiz latente (em que são considerados apenas fatores oriundos de autovalores maiores que 1), é de se supor que as cargas fatoriais entre os fatores correspondentes a autovalores menores que 1 e todas as variáveis originais sejam baixas, visto que já terão apresentado correlações de Pearson (cargas) mais elevadas com fatores extraídos anteriormente a partir de autovalores maiores. Do mesmo modo, variáveis originais que compartilhem apenas uma pequena parcela de variância com as demais variáveis apresentarão cargas fatoriais elevadas apenas em um único fator. Caso isso ocorra para todas as variáveis originais, não existirão diferenças significativas entre a matriz de correlações  $\rho$  e a matriz identidade  $\mathbf{I}$ , tornando a estatística  $\chi^2_{\text{Bartlett}}$  muito baixa. Esse fato permite afirmar que a análise fatorial será inapropriada, e, nessa situação, o pesquisador poderá optar por não extrair fatores a partir das variáveis originais.

Como as cargas fatoriais são as correlações de Pearson entre cada variável e cada fator, a somatória dos quadrados dessas cargas em cada linha da Tabela 10.2 será sempre igual a 1, visto que cada variável compartilha parte do seu percentual de variância com todos os  $k$  fatores, e a somatória dos percentuais de variância (cargas fatoriais ou correlações de Pearson ao quadrado) será 100%.

Por outro lado, caso seja extraída uma quantidade de fatores menor que  $k$ , em função do critério da raiz latente, a somatória dos quadrados das cargas fatoriais em cada linha não chegará a ser igual a 1. A essa somatória, dá-se o nome de **comunalidade**, que representa a **variância total compartilhada de cada variável em todos os fatores extraídos a partir de autovalores maiores que 1**. Logo, podemos escrever que:

$$\begin{aligned} c_{11}^2 + c_{12}^2 + \dots &= \text{comunalidade } X_1 \\ c_{21}^2 + c_{22}^2 + \dots &= \text{comunalidade } X_2 \\ &\vdots \\ c_{k1}^2 + c_{k2}^2 + \dots &= \text{comunalidade } X_k \end{aligned} \quad (10.29)$$

O objetivo principal da análise das comunalidades é verificar se alguma variável acaba por não compartilhar um significativo percentual de variância com os fatores extraídos. Embora não haja um ponto de corte a partir do qual determinada comunalidade possa ser considerada alta ou baixa, visto que o tamanho da amostra pode interferir nesse julgamento, a existência de comunalidades consideravelmente baixas em relação às demais pode sugerir que o pesquisador reconsidere a inclusão da respectiva variável na análise fatorial.

Logo, definidos os fatores com base nos *scores* fatoriais, podemos afirmar que as cargas fatoriais serão exatamente iguais aos parâmetros estimados de um modelo de regressão linear múltipla que apresenta, como variável dependente, determinada variável padronizada  $ZX$  e, como variáveis explicativas, os próprios fatores, sendo o **coeficiente de ajuste  $R^2$**  de cada modelo igual à própria comunalidade da respectiva variável original.

A somatória dos quadrados das cargas fatoriais em cada coluna da Tabela 10.2, por outro lado, será igual ao respectivo autovalor, visto que a razão entre cada autovalor e a quantidade total de variáveis pode ser entendida como o percentual de variância compartilhada por todas as  $k$  variáveis originais para a formação de cada fator. Logo, podemos escrever que:

$$\begin{aligned} c_{11}^2 + c_{21}^2 + \dots + c_{k1}^2 &= \lambda_1^2 \\ c_{12}^2 + c_{22}^2 + \dots + c_{k2}^2 &= \lambda_2^2 \\ &\vdots \\ c_{1k}^2 + c_{2k}^2 + \dots + c_{kk}^2 &= \lambda_k^2 \end{aligned} \quad (10.30)$$

Após a determinação dos fatores e do cálculo das cargas fatoriais, é possível ainda que algumas variáveis apresentem correlações de Pearson (cargas fatoriais) intermediárias (nem tão altas, nem tão baixas) com todos os fatores extraídos, embora sua comunalidade não seja relativamente tão baixa. Nesse caso, embora a solução da análise fatorial já tenha sido obtida de forma adequada e considerada finalizada, o pesquisador pode, para os casos em que a tabela de cargas fatoriais apresentar valores intermediários para uma ou mais variáveis em todos os fatores, elaborar uma rotação desses fatores, a fim de que sejam aumentadas as correlações de Pearson entre as variáveis originais e novos fatores gerados. Na próxima seção, trataremos especificamente da rotação de fatores.

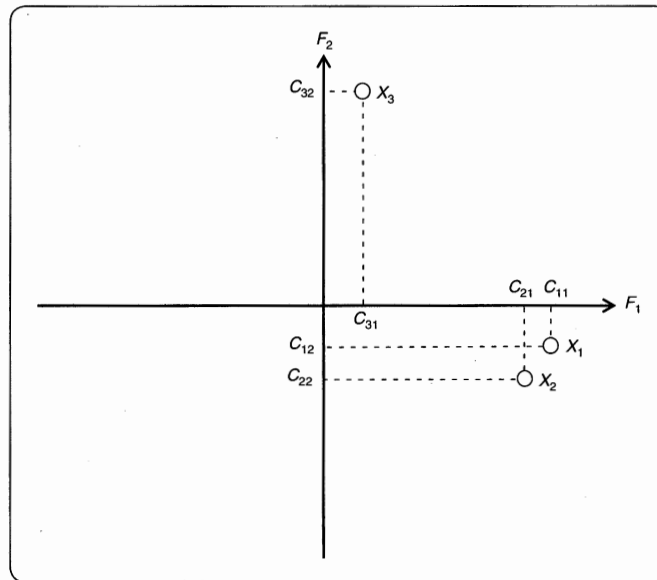
### 10.2.5. Rotação de fatores

Imaginemos novamente uma situação hipotética em que determinado banco de dados apresenta apenas três variáveis ( $k = 3$ ). Após a elaboração da análise fatorial por componentes principais, são extraídos dois fatores, ortogonais entre si, com cargas fatoriais (correlações de Pearson) com cada uma das três variáveis originais, de acordo com a Tabela 10.3.

**Tabela 10.3** Cargas fatoriais entre três variáveis e dois fatores.

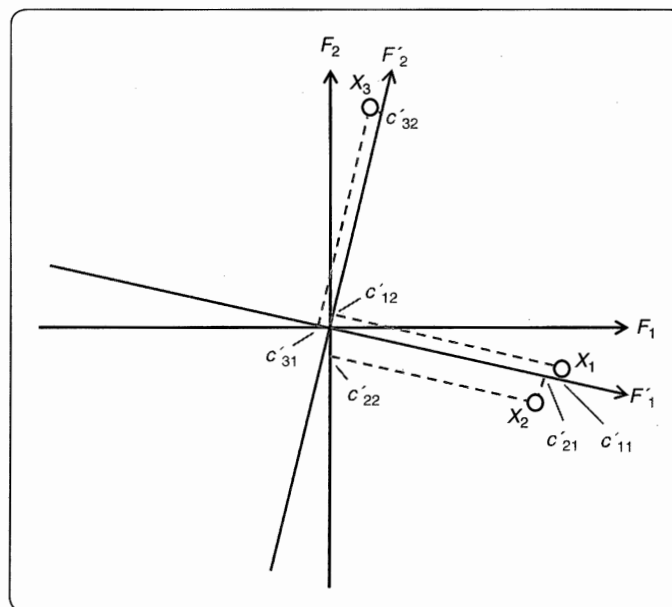
<b>Variável \ Fator</b>	<b><math>F_1</math></b>	<b><math>F_2</math></b>
$X_1$	$c_{11}$	$c_{12}$
$X_2$	$c_{21}$	$c_{22}$
$X_3$	$c_{31}$	$c_{32}$

A fim de que possa ser elaborado um gráfico com as posições relativas de cada variável em cada fator (gráfico conhecido como **loading plot**), podemos considerar as cargas fatoriais coordenadas (abscissas e ordenadas) das variáveis em um plano cartesiano formado pelos dois fatores ortogonais. Esse gráfico encontra-se, de maneira exemplificada, na Figura 10.5.



**Figura 10.5** Loading plot para situação hipotética com três variáveis e dois fatores.

Para que tenhamos melhor visualização das variáveis mais representadas por determinado fator, podemos pensar em uma rotação, em torno da origem, dos fatores originalmente extraídos  $F_1$  e  $F_2$ , de modo a aproximar os pontos correspondentes às variáveis  $X_1$ ,  $X_2$  e  $X_3$  de um dos novos fatores, chamados de **fatores rotacionados**  $F'_1$  e  $F'_2$ . A Figura 10.6 apresenta essa situação de forma exemplificada.



**Figura 10.6** Definição dos fatores rotacionados a partir dos fatores originais.

Com base na Figura 10.6, podemos verificar, para cada variável em análise, que, enquanto a carga para um fator é aumentada, para o outro, é diminuída. A Tabela 10.4 mostra a redistribuição de cargas para nossa situação hipotética.

**Tabela 10.4** Cargas fatoriais originais e rotacionadas para a nossa situação hipotética.

Variável \ Fator	Cargas Fatoriais Originais		Cargas Fatoriais Rotacionadas	
	$F_1$	$F_2$	$F'_1$	$F'_2$
$X_1$	$c_{11}$	$c_{12}$	$ c'_{11}  >  c_{11} $	$ c'_{12}  <  c_{12} $
$X_2$	$c_{21}$	$c_{22}$	$ c'_{21}  >  c_{21} $	$ c'_{22}  <  c_{22} $
$X_3$	$c_{31}$	$c_{32}$	$ c'_{31}  <  c_{31} $	$ c'_{32}  >  c_{32} $

Logo, para uma situação genérica, podemos afirmar que a rotação é um procedimento que maximiza as cargas de cada variável em determinado fator, em detrimento dos demais. Nesse sentido, o efeito final da rotação é a redistribuição das cargas fatoriais para fatores que inicialmente apresentavam menores percentuais de variância compartilhada por todas as variáveis originais. O objetivo principal é minimizar a quantidade de variáveis com altas cargas em determinado fator, já que cada um dos fatores passará a ter cargas mais expressivas somente com algumas das variáveis originais. Consequentemente, a rotação pode simplificar a interpretação dos fatores.

Embora as communalidades e o percentual total de variância compartilhada por todas as variáveis em todos os fatores não sejam alterados com a rotação (tampouco as estatísticas KMO e  $\chi^2_{\text{Bartlett}}$ ), o percentual de variância compartilhada pelas variáveis originais em cada fator é redistribuído e, portanto, alterado. Em outras palavras, são determinados novos autovalores  $\lambda'$  ( $\lambda'_1, \lambda'_2, \dots, \lambda'_k$ ) a partir das **cargas fatoriais rotacionadas**. Assim, podemos escrever que:

$$\begin{aligned}
 c'^2_{11} + c'^2_{12} + \dots &= \text{comunalidade } X_1 \\
 c'^2_{21} + c'^2_{22} + \dots &= \text{comunalidade } X_2 \\
 &\vdots \\
 c'^2_{k1} + c'^2_{k2} + \dots &= \text{comunalidade } X_k
 \end{aligned} \tag{10.31}$$

e que:

$$\begin{aligned}
 c'^2_{11} + c'^2_{21} + \dots + c'^2_{k1} &= \lambda'^2_1 \neq \lambda^2_1 \\
 c'^2_{12} + c'^2_{22} + \dots + c'^2_{k2} &= \lambda'^2_2 \neq \lambda^2_2 \\
 &\vdots \\
 c'^2_{1k} + c'^2_{2k} + \dots + c'^2_{kk} &= \lambda'^2_k \neq \lambda^2_k
 \end{aligned} \tag{10.32}$$

mesmo sendo respeitada a expressão (10.13), ou seja:

$$\lambda^2_1 + \lambda^2_2 + \dots + \lambda^2_k = \lambda'^2_1 + \lambda'^2_2 + \dots + \lambda'^2_k = k \tag{10.33}$$

Além disso, a partir da rotação dos fatores, são obtidos novos **scores fatoriais rotacionados**,  $s'$ , de modo que as expressões finais dos fatores rotacionados serão:

$$\begin{aligned}
 F'_{1i} &= s'_{11} \cdot ZX_{1i} + s'_{21} \cdot ZX_{2i} + \dots + s'_{k1} \cdot ZX_{ki} \\
 F'_{2i} &= s'_{12} \cdot ZX_{1i} + s'_{22} \cdot ZX_{2i} + \dots + s'_{k2} \cdot ZX_{ki} \\
 &\vdots \\
 F'_{ki} &= s'_{1k} \cdot ZX_{1i} + s'_{2k} \cdot ZX_{2i} + \dots + s'_{kk} \cdot ZX_{ki}
 \end{aligned} \tag{10.34}$$

É importante ressaltar que a adequação global da análise fatorial (estatística KMO e teste de esfericidade de Bartlett) não é alterada com a rotação, já que a matriz de correlações  $\rho$  continua a mesma.

Embora existam diversos métodos de rotação fatorial, o mais utilizado e que será adotado quando da elaboração prática de um exemplo neste capítulo refere-se ao **método de rotação ortogonal** conhecido por **Varimax**, cuja principal finalidade é minimizar a quantidade de variáveis que apresentam elevadas cargas em determinado fator por meio da redistribuição das cargas fatoriais e maximização da variância compartilhada em fatores correspondentes a autovalores mais baixos. Daí decorre a nomenclatura Varimax, proposta por Kaiser (1958).



O algoritmo por trás do método de rotação Varimax consiste em determinar um ângulo de rotação  $\theta$  em que pares de fatores são rotacionados igualmente. Logo, conforme discute Harman (1968), para determinado par de fatores  $F_1$  e  $F_2$ , por exemplo, as cargas fatoriais rotacionadas  $c'$  entre os dois fatores e as  $k$  variáveis originais são obtidas a partir das cargas fatoriais originais  $c$ , por meio da seguinte multiplicação matricial:

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ \vdots & \vdots \\ c_{k1} & c_{k2} \end{pmatrix} \cdot \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} = \begin{pmatrix} c'_{11} & c'_{12} \\ c'_{21} & c'_{22} \\ \vdots & \vdots \\ c'_{k1} & c'_{k2} \end{pmatrix} \quad (10.35)$$

em que  $\theta$ , ângulo de rotação no sentido anti-horário, é obtido pela seguinte expressão:

$$\theta = 0,25 \cdot \arctan \left[ \frac{2 \cdot (D \cdot k - A \cdot B)}{C \cdot k - (A^2 - B^2)} \right] \quad (10.36)$$

sendo:

$$A = \sum_{l=1}^k \left( \frac{c_{1l}^2}{\text{comunalidade}_l} - \frac{c_{2l}^2}{\text{comunalidade}_l} \right) \quad (10.37)$$

$$B = \sum_{l=1}^k \left( 2 \cdot \frac{c_{1l} \cdot c_{2l}}{\text{comunalidade}_l} \right) \quad (10.38)$$

$$C = \sum_{l=1}^k \left[ \left( \frac{c_{1l}^2}{\text{comunalidade}_l} - \frac{c_{2l}^2}{\text{comunalidade}_l} \right)^2 - \left( 2 \cdot \frac{c_{1l} \cdot c_{2l}}{\text{comunalidade}_l} \right)^2 \right] \quad (10.39)$$

$$D = \sum_{l=1}^k \left[ \left( \frac{c_{1l}^2}{\text{comunalidade}_l} - \frac{c_{2l}^2}{\text{comunalidade}_l} \right)^2 \cdot \left( 2 \cdot \frac{c_{1l} \cdot c_{2l}}{\text{comunalidade}_l} \right) \right] \quad (10.40)$$

Na seção 10.2.6, faremos uso dessas expressões do método de rotação Varimax para determinar as cargas fatoriais rotacionadas a partir das cargas originais.

Além da rotação Varimax, outros métodos de rotação ortogonal também podem ser mencionados, como o **Quartimax** e o **Equimax**, embora sejam menos referenciados na literatura e utilizados com menor intensidade na prática. Além deles, o pesquisador ainda pode fazer uso de **métodos de rotação oblíqua**, em que são gerados fatores não ortogonais. Embora não sejam abordados neste capítulo, merecem menção nesta categoria os chamados **Direct Oblimin** e **Promax**.

Como os métodos de rotação oblíqua podem, por vezes, ser utilizados quando se deseja validar determinado constructo, cujos fatores iniciais sejam não correlacionados, recomenda-se que um método de rotação ortogonal seja utilizado para uso subsequente dos fatores extraídos em outras técnicas multivariadas, como determinados modelos confirmatórios em que é exigida a premissa de ausência de multicolinearidade de variáveis explicativas.

### 10.2.6. Exemplo prático de análise fatorial por componentes principais

Imagine que nosso mesmo professor, bastante engajado com atividades acadêmicas e didáticas, tenha agora o interesse em estudar como se comportam as notas de seus alunos para, em sequência, propor um *ranking* de desempenho escolar.

Para tanto, ele fez um levantamento sobre as notas finais, que variam de 0 a 10, de cada um de seus 100 alunos nas disciplinas de Finanças, Custos, Marketing e Atuária. Parte do banco de dados elaborado encontra-se na Tabela 10.5.

**Tabela 10.5** Exemplo: Notas finais de Finanças, Custos, Marketing e Atuária.

Estudante	Nota final de Finanças ( $X_{1i}$ )	Nota final de Custos ( $X_{2i}$ )	Nota final de Marketing ( $X_{3i}$ )	Nota final de Atuária ( $X_{4i}$ )
Gabriela	5,8	4,0	1,0	6,0
Luiz Felipe	3,1	3,0	10,0	2,0
Patrícia	3,1	4,0	4,0	4,0
Gustavo	10,0	8,0	8,0	8,0
Letícia	3,4	2,0	3,2	3,2
Ovídio	10,0	10,0	1,0	10,0
Leonor	5,0	5,0	8,0	5,0
Dalila	5,4	6,0	6,0	6,0
Antônio	5,9	4,0	4,0	4,0
...				
Estela	8,9	5,0	2,0	8,0

O banco de dados completo pode ser acessado por meio do arquivo **NotasFatorial.xls**.

Por meio desse banco de dados, é possível que seja elaborada a Tabela 10.6, que apresenta os coeficientes de correlação de Pearson entre cada par de variáveis, calculados por meio da lógica apresentada na expressão (10.2).

**Tabela 10.6** Coeficientes de correlação de Pearson para cada par de variáveis.

	<i>finanças</i>	<i>custos</i>	<i>marketing</i>	<i>atuária</i>
<i>finanças</i>	1,000	0,756	-0,030	0,711
<i>custos</i>	0,756	1,000	0,003	0,809
<i>marketing</i>	-0,030	0,003	1,000	-0,044
<i>atuária</i>	0,711	0,809	-0,044	1,000

Dessa forma, podemos escrever a expressão matriz de correlações  $\rho$  conforme segue:

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{pmatrix} = \begin{pmatrix} 1,000 & 0,756 & -0,030 & 0,711 \\ 0,756 & 1,000 & 0,003 & 0,809 \\ -0,030 & 0,003 & 1,000 & -0,044 \\ 0,711 & 0,809 & -0,044 & 1,000 \end{pmatrix}$$

que apresenta determinante  $D = 0,137$ .

Com base na análise da matriz de correlações  $\rho$ , é possível verificar que apenas as notas correspondentes à variável *marketing* não apresentam correlações com as notas das demais disciplinas, representadas pelas outras variáveis. Por outro lado, estas apresentam correlações relativamente elevadas entre si (0,756 entre *finanças* e *custos*, 0,711 entre *finanças* e *atuária* e 0,809 entre *custos* e *atuária*), o que indica que poderão compartilhar significativa variância para a formação de um fator. Embora essa análise preliminar seja importante, não pode representar mais que um simples diagnóstico, visto que a adequação global da análise fatorial precisa ser elaborada com base na estatística KMO e, principalmente, por meio do resultado do teste de esfericidade de Bartlett.

Conforme discutimos na seção 10.2.2, a estatística KMO fornece a proporção de variância considerada comum a todas as variáveis presentes na análise, e, para que seja estabelecido seu cálculo, precisamos determinar os coeficientes de correlação parcial  $\phi$  entre cada par de variáveis que, neste caso, serão coeficientes de correlação de segunda ordem, visto que estamos trabalhando com quatro variáveis simultaneamente.

Logo, com base na expressão (10.7), precisamos determinar, inicialmente, os coeficientes de correlação de primeira ordem utilizados para o cálculo dos coeficientes de correlação de segunda ordem. A Tabela 10.7 apresenta esses coeficientes.

**Tabela 10.7** Coeficientes de correlação de primeira ordem.

$\varphi_{12,3} = \frac{\rho_{12} - \rho_{13} \cdot \rho_{23}}{\sqrt{(1-\rho_{13}^2) \cdot (1-\rho_{23}^2)}} = 0,756$	$\varphi_{13,2} = \frac{\rho_{13} - \rho_{12} \cdot \rho_{23}}{\sqrt{(1-\rho_{12}^2) \cdot (1-\rho_{23}^2)}} = -0,049$	$\varphi_{14,2} = \frac{\rho_{14} - \rho_{12} \cdot \rho_{24}}{\sqrt{(1-\rho_{12}^2) \cdot (1-\rho_{24}^2)}} = 0,258$
$\varphi_{14,3} = \frac{\rho_{14} - \rho_{13} \cdot \rho_{34}}{\sqrt{(1-\rho_{13}^2) \cdot (1-\rho_{34}^2)}} = 0,711$	$\varphi_{23,1} = \frac{\rho_{23} - \rho_{12} \cdot \rho_{13}}{\sqrt{(1-\rho_{12}^2) \cdot (1-\rho_{13}^2)}} = 0,039$	$\varphi_{24,1} = \frac{\rho_{24} - \rho_{12} \cdot \rho_{14}}{\sqrt{(1-\rho_{12}^2) \cdot (1-\rho_{14}^2)}} = 0,590$
$\varphi_{24,3} = \frac{\rho_{24} - \rho_{23} \cdot \rho_{34}}{\sqrt{(1-\rho_{23}^2) \cdot (1-\rho_{34}^2)}} = 0,810$	$\varphi_{34,1} = \frac{\rho_{34} - \rho_{13} \cdot \rho_{14}}{\sqrt{(1-\rho_{13}^2) \cdot (1-\rho_{14}^2)}} = -0,033$	$\varphi_{34,2} = \frac{\rho_{34} - \rho_{23} \cdot \rho_{24}}{\sqrt{(1-\rho_{23}^2) \cdot (1-\rho_{24}^2)}} = -0,080$

Dessa maneira, a partir desses coeficientes e fazendo uso da expressão (10.8), podemos calcular os coeficientes de correlação de segunda ordem considerados na expressão da estatística KMO. A Tabela 10.8 apresenta esses coeficientes.

**Tabela 10.8** Coeficientes de correlação de segunda ordem.

$\varphi_{12,34} = \frac{\varphi_{12,3} - \varphi_{14,3} \cdot \varphi_{24,3}}{\sqrt{(1-\varphi_{14,3}^2) \cdot (1-\varphi_{24,3}^2)}} = 0,438$	
$\varphi_{13,24} = \frac{\varphi_{13,2} - \varphi_{14,2} \cdot \varphi_{34,2}}{\sqrt{(1-\varphi_{14,2}^2) \cdot (1-\varphi_{34,2}^2)}} = -0,029$	$\varphi_{23,14} = \frac{\varphi_{23,1} - \varphi_{24,1} \cdot \varphi_{34,1}}{\sqrt{(1-\varphi_{24,1}^2) \cdot (1-\varphi_{34,1}^2)}} = 0,072$
$\varphi_{14,23} = \frac{\varphi_{14,2} - \varphi_{13,2} \cdot \varphi_{34,2}}{\sqrt{(1-\varphi_{13,2}^2) \cdot (1-\varphi_{34,2}^2)}} = 0,255$	$\varphi_{24,13} = \frac{\varphi_{24,1} - \varphi_{23,1} \cdot \varphi_{34,1}}{\sqrt{(1-\varphi_{23,1}^2) \cdot (1-\varphi_{34,1}^2)}} = 0,592$
	$\varphi_{34,12} = \frac{\varphi_{34,1} - \varphi_{23,1} \cdot \varphi_{24,1}}{\sqrt{(1-\varphi_{23,1}^2) \cdot (1-\varphi_{24,1}^2)}} = -0,069$

Portanto, com base na expressão (10.3), podemos calcular a estatística KMO. Os termos da expressão são dados por:

$$\sum_{l=1}^k \sum_{c=1}^k \rho_{lc}^2 = (0,756)^2 + (-0,030)^2 + (0,711)^2 + (0,003)^2 + (0,809)^2 + (-0,044)^2 = 1,734$$

$$\sum_{l=1}^k \sum_{c=1}^k \varphi_{lc}^2 = (0,438)^2 + (-0,029)^2 + (0,255)^2 + (0,072)^2 + (0,592)^2 + (-0,069)^2 = 0,619$$

de onde vem que:

$$KMO = \frac{1,734}{1,734 + 0,619} = 0,737$$

O valor da estatística KMO indica, com base no critério apresentado no Quadro 10.1, que a adequação global da análise fatorial é **média**. Para testarmos se, de fato, a matriz de correlações  $\rho$  é estatisticamente diferente da matriz identidade  $\mathbf{I}$  de mesma dimensão, devemos recorrer ao teste de esfericidade de Bartlett, cuja estatística  $\chi_{\text{Bartlett}}^2$  é dada pela expressão (10.9). Temos, para  $n = 100$  observações,  $k = 4$  variáveis e determinante da matriz de correlações  $\rho$   $D = 0,137$ , que:

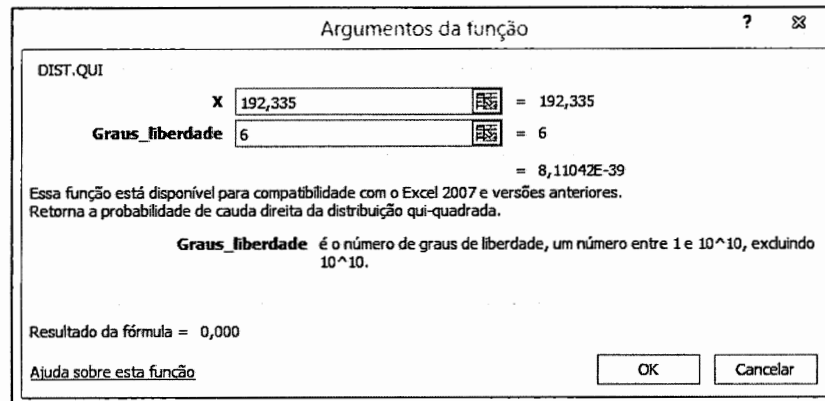
$$\chi_{\text{Bartlett}}^2 = - \left[ (100-1) - \left( \frac{2 \cdot 4 + 5}{6} \right) \right] \cdot \ln(0,137) = 192,335$$

com  $\frac{4 \cdot (4-1)}{2} = 6$  graus de liberdade. Logo, por meio da Tabela D do apêndice do livro, temos que  $\chi_c^2 = 12,592$  ( $\chi^2$  crítico para 6 graus de liberdade e para o nível de significância de 5%). Dessa forma, como  $\chi_{\text{Bartlett}}^2 = 192,335 > \chi_c^2 = 12,592$ , podemos rejeitar a hipótese nula de que a matriz de correlações  $\rho$  seja estatisticamente igual à matriz identidade  $\mathbf{I}$ , ao nível de significância de 5%.

Softwares como o SPSS e o Stata não oferecem o  $\chi^2_c$  para os graus de liberdade definidos e determinado nível de significância. Todavia, oferecem o nível de significância do  $\chi^2_{\text{Bartlett}}$  para esses graus de liberdade. Dessa forma, em vez de analisarmos se  $\chi^2_{\text{Bartlett}} > \chi^2_c$ , devemos verificar se o nível de significância do  $\chi^2_{\text{Bartlett}}$  é menor que 0,05 (5%) a fim de darmos continuidade à análise fatorial. Assim:

Se *valor-P* (ou *P-value* ou *Sig.*  $\chi^2_{\text{Bartlett}}$  ou *Prob.*  $\chi^2_{\text{Bartlett}}$ ) < 0,05, a matriz de correlações  $\rho$  não é estatisticamente igual à matriz identidade **I** de mesma dimensão.

O nível de significância do  $\chi^2_{\text{Bartlett}}$  pode ser obtido no Excel por meio do comando **Fórmulas → Inserir Função → DIST.QUI**, que abrirá uma caixa de diálogo, conforme mostra a Figura 10.7.



**Figura 10.7** Obtenção do nível de significância de  $\chi^2$  (comando **Inserir Função**).

Conforme podemos observar por meio da Figura 10.7, o *valor-P* da estatística  $\chi^2_{\text{Bartlett}}$  é consideravelmente menor que 0,05 (*valor-P*  $\chi^2_{\text{Bartlett}} = 8,11 \times 10^{-39}$ ), ou seja, as correlações de Pearson entre os pares de variáveis são estatisticamente diferentes de 0 e, portanto, podem ser extraídos fatores a partir das variáveis originais, sendo a análise fatorial bastante apropriada. Para um pesquisador interessado, todos esses cálculos estão apresentados diretamente no arquivo **NotasFatorialCálculosKMOBartlett.xls**.

Verificada a adequação global da análise fatorial, podemos partir para a definição propriamente dita dos fatores. Para tanto, devemos inicialmente determinar os quatro autovalores  $\lambda^2$  ( $\lambda_1^2 \geq \lambda_2^2 \geq \lambda_3^2 \geq \lambda_4^2$ ) da matriz de correlações  $\rho$ , que podem ser obtidos a partir da solução da expressão (10.12). Sendo assim, temos que:

$$\begin{vmatrix} \lambda^2 - 1 & -0,756 & 0,030 & -0,711 \\ -0,756 & \lambda^2 - 1 & -0,003 & -0,809 \\ 0,030 & -0,003 & \lambda^2 - 1 & 0,044 \\ -0,711 & -0,809 & 0,044 & \lambda^2 - 1 \end{vmatrix} = 0$$

de onde vem que:

$$\begin{cases} \lambda_1^2 = 2,519 \\ \lambda_2^2 = 1,000 \\ \lambda_3^2 = 0,298 \\ \lambda_4^2 = 0,183 \end{cases}$$

Logo, com base na expressão (10.15), a matriz de autovalores  $\Lambda^2$  pode ser escrita da seguinte forma:

$$\Lambda^2 = \begin{pmatrix} 2,519 & 0 & 0 & 0 \\ 0 & 1,000 & 0 & 0 \\ 0 & 0 & 0,298 & 0 \\ 0 & 0 & 0 & 0,183 \end{pmatrix}$$

Note que a expressão (10.13) é satisfeita, ou seja:

$$\lambda_1^2 + \lambda_2^2 + \dots + \lambda_k^2 = 2,519 + 1,000 + 0,298 + 0,183 = 4$$

Como os autovalores correspondem ao percentual de variância compartilhada pelas variáveis originais para a formação de cada fator, podemos elaborar uma tabela de variância compartilhada (Tabela 10.9).

**Tabela 10.9** Variância compartilhada pelas variáveis originais para a formação de cada fator.

Fator	Autovalor $\lambda^2$	Variância Compartilhada (%)	Variância Compartilhada Acumulada (%)
1	2,519	$\left(\frac{2,519}{4}\right) \cdot 100 = 62,975$	62,975
2	1,000	$\left(\frac{1,000}{4}\right) \cdot 100 = 25,010$	87,985
3	0,298	$\left(\frac{0,298}{4}\right) \cdot 100 = 7,444$	95,428
4	0,183	$\left(\frac{0,183}{4}\right) \cdot 100 = 4,572$	100,000

Por meio da análise da Tabela 10.9, podemos afirmar que, enquanto 62,975% da variância total são compartilhados para a formação do primeiro fator, 25,010% são compartilhados para a formação do segundo. O terceiro e o quarto fatores, cujos autovalores são menores que 1, são formados por meio de menores percentuais de variância compartilhada. Como o critério mais adotado para a escolha da quantidade de fatores é o critério da raiz latente (critério de Kaiser), em que são levados em consideração apenas os fatores correspondentes a autovalores maiores que 1, o pesquisador pode optar por elaborar toda a análise subsequente apenas com os dois primeiros fatores, formados pelo compartilhamento de 87,985% da variância total das variáveis originais, ou seja, com perda total de variância de 12,015%. Para efeitos didáticos, entretanto, vamos apresentar os cálculos dos *scores* fatoriais por meio da determinação dos autovetores correspondentes aos quatro autovalores.

Logo, para que sejam definidos os autovetores da matriz  $\rho$  com base nos quatro autovalores calculados, devemos resolver os seguintes sistemas de equações para cada autovalor, com base nas expressões (10.16) a (10.21):

- Determinação de Autovetores  $v_{11}, v_{21}, v_{31}, v_{41}$  a partir do Primeiro Autovalor ( $\lambda_1^2 = 2,519$ ):

$$\begin{cases} (2,519 - 1,000) \cdot v_{11} - 0,756 \cdot v_{21} + 0,030 \cdot v_{31} - 0,711 \cdot v_{41} = 0 \\ -0,756 \cdot v_{11} + (2,519 - 1,000) \cdot v_{21} - 0,003 \cdot v_{31} - 0,809 \cdot v_{41} = 0 \\ 0,030 \cdot v_{11} - 0,003 \cdot v_{21} + (2,519 - 1,000) \cdot v_{31} + 0,044 \cdot v_{41} = 0 \\ -0,711 \cdot v_{11} - 0,809 \cdot v_{21} + 0,044 \cdot v_{31} + (2,519 - 1,000) \cdot v_{41} = 0 \end{cases}$$

de onde vem que:

$$\begin{pmatrix} v_{11} \\ v_{21} \\ v_{31} \\ v_{41} \end{pmatrix} = \begin{pmatrix} 0,5641 \\ 0,5887 \\ -0,0267 \\ 0,5783 \end{pmatrix}$$

- Determinação de Autovetores  $v_{12}, v_{22}, v_{32}, v_{42}$  a partir do Segundo Autovalor ( $\lambda_2^2 = 1,000$ ):

$$\begin{cases} (1,000 - 1,000) \cdot v_{12} - 0,756 \cdot v_{22} + 0,030 \cdot v_{32} - 0,711 \cdot v_{42} = 0 \\ -0,756 \cdot v_{12} + (1,000 - 1,000) \cdot v_{22} - 0,003 \cdot v_{32} - 0,809 \cdot v_{42} = 0 \\ 0,030 \cdot v_{12} - 0,003 \cdot v_{22} + (1,000 - 1,000) \cdot v_{32} + 0,044 \cdot v_{42} = 0 \\ -0,711 \cdot v_{12} - 0,809 \cdot v_{22} + 0,044 \cdot v_{32} + (1,000 - 1,000) \cdot v_{42} = 0 \end{cases}$$

de onde vem que:

$$\begin{pmatrix} v_{12} \\ v_{22} \\ v_{32} \\ v_{42} \end{pmatrix} = \begin{pmatrix} 0,0068 \\ 0,0487 \\ 0,9987 \\ -0,0101 \end{pmatrix}$$

- Determinação de Autovetores  $v_{13}, v_{23}, v_{33}, v_{43}$  a partir do Terceiro Autovalor ( $\lambda_3^2 = 0,298$ ):

$$\begin{cases} (0,298-1,000) \cdot v_{13} - 0,756 \cdot v_{23} + 0,030 \cdot v_{33} - 0,711 \cdot v_{43} = 0 \\ -0,756 \cdot v_{13} + (0,298-1,000) \cdot v_{23} - 0,003 \cdot v_{33} - 0,809 \cdot v_{43} = 0 \\ 0,030 \cdot v_{13} - 0,003 \cdot v_{23} + (0,298-1,000) \cdot v_{33} + 0,044 \cdot v_{43} = 0 \\ -0,711 \cdot v_{13} - 0,809 \cdot v_{23} + 0,044 \cdot v_{33} + (0,298-1,000) \cdot v_{43} = 0 \end{cases}$$

de onde vem que:

$$\begin{pmatrix} v_{13} \\ v_{23} \\ v_{33} \\ v_{43} \end{pmatrix} = \begin{pmatrix} 0,8008 \\ -0,2201 \\ -0,0003 \\ -0,5571 \end{pmatrix}$$

- Determinação de Autovetores  $v_{14}, v_{24}, v_{34}, v_{44}$  a partir do Quarto Autovalor ( $\lambda_4^2 = 0,183$ ):

$$\begin{cases} (0,183-1,000) \cdot v_{14} - 0,756 \cdot v_{24} + 0,030 \cdot v_{34} - 0,711 \cdot v_{44} = 0 \\ -0,756 \cdot v_{14} + (0,183-1,000) \cdot v_{24} - 0,003 \cdot v_{34} - 0,809 \cdot v_{44} = 0 \\ 0,030 \cdot v_{14} - 0,003 \cdot v_{24} + (0,183-1,000) \cdot v_{34} + 0,044 \cdot v_{44} = 0 \\ -0,711 \cdot v_{14} - 0,809 \cdot v_{24} + 0,044 \cdot v_{34} + (0,183-1,000) \cdot v_{44} = 0 \end{cases}$$

de onde vem que:

$$\begin{pmatrix} v_{14} \\ v_{24} \\ v_{34} \\ v_{44} \end{pmatrix} = \begin{pmatrix} 0,2012 \\ -0,7763 \\ 0,0425 \\ 0,5959 \end{pmatrix}$$

Determinados os autovetores, um pesquisador mais curioso poderá comprovar a relação apresentada na expressão (10.27), ou seja:

$$\mathbf{V}' \cdot \boldsymbol{\rho} \cdot \mathbf{V} = \boldsymbol{\Lambda}^2$$

$$\begin{pmatrix} 0,5641 & 0,5887 & -0,0267 & 0,5783 \\ 0,0068 & 0,0487 & 0,9987 & -0,0101 \\ 0,8008 & -0,2201 & -0,0003 & -0,5571 \\ 0,2012 & -0,7763 & 0,0425 & 0,5959 \end{pmatrix} \cdot \begin{pmatrix} 1,000 & 0,756 & -0,030 & 0,711 \\ 0,756 & 1,000 & 0,003 & 0,809 \\ -0,030 & 0,003 & 1,000 & -0,044 \\ 0,711 & 0,809 & -0,044 & 1,000 \end{pmatrix} = \begin{pmatrix} 2,519 & 0 & 0 & 0 \\ 0 & 1,000 & 0 & 0 \\ 0 & 0 & 0,298 & 0 \\ 0 & 0 & 0 & 0,183 \end{pmatrix}$$

Com base nas expressões (10.22) a (10.24), podemos calcular os *scores* fatoriais correspondentes a cada uma das variáveis padronizadas para cada um dos fatores. Dessa forma, temos condições de escrever, a partir da expressão (10.25), as expressões dos fatores  $F_1, F_2, F_3$  e  $F_4$ , conforme segue:

$$\begin{aligned}
F_{1i} &= \frac{0,5641}{\sqrt{2,519}} \cdot Z_{\text{finanças}_i} + \frac{0,5887}{\sqrt{2,519}} \cdot Z_{\text{custos}_i} - \frac{0,0267}{\sqrt{2,519}} \cdot Z_{\text{marketing}_i} + \frac{0,5783}{\sqrt{2,519}} \cdot Z_{\text{atuária}_i} \\
F_{2i} &= \frac{0,0068}{\sqrt{1,000}} \cdot Z_{\text{finanças}_i} + \frac{0,0487}{\sqrt{1,000}} \cdot Z_{\text{custos}_i} + \frac{0,9987}{\sqrt{1,000}} \cdot Z_{\text{marketing}_i} - \frac{0,0101}{\sqrt{1,000}} \cdot Z_{\text{atuária}_i} \\
F_{3i} &= \frac{0,8008}{\sqrt{0,298}} \cdot Z_{\text{finanças}_i} - \frac{0,2201}{\sqrt{0,298}} \cdot Z_{\text{custos}_i} - \frac{0,0003}{\sqrt{0,298}} \cdot Z_{\text{marketing}_i} - \frac{0,5571}{\sqrt{0,298}} \cdot Z_{\text{atuária}_i} \\
F_{4i} &= \frac{0,2012}{\sqrt{0,183}} \cdot Z_{\text{finanças}_i} - \frac{0,7763}{\sqrt{0,183}} \cdot Z_{\text{custos}_i} + \frac{0,0425}{\sqrt{0,183}} \cdot Z_{\text{marketing}_i} + \frac{0,5959}{\sqrt{0,183}} \cdot Z_{\text{atuária}_i}
\end{aligned}$$

de onde vem que:

$$\begin{aligned}
F_{1i} &= 0,355 \cdot Z_{\text{finanças}_i} + 0,371 \cdot Z_{\text{custos}_i} - 0,017 \cdot Z_{\text{marketing}_i} + 0,364 \cdot Z_{\text{atuária}_i} \\
F_{2i} &= 0,007 \cdot Z_{\text{finanças}_i} + 0,049 \cdot Z_{\text{custos}_i} + 0,999 \cdot Z_{\text{marketing}_i} - 0,010 \cdot Z_{\text{atuária}_i} \\
F_{3i} &= 1,468 \cdot Z_{\text{finanças}_i} - 0,403 \cdot Z_{\text{custos}_i} - 0,001 \cdot Z_{\text{marketing}_i} - 1,021 \cdot Z_{\text{atuária}_i} \\
F_{4i} &= 0,470 \cdot Z_{\text{finanças}_i} - 1,815 \cdot Z_{\text{custos}_i} + 0,099 \cdot Z_{\text{marketing}_i} + 1,394 \cdot Z_{\text{atuária}_i}
\end{aligned}$$

Com base nas expressões dos fatores e nas variáveis padronizadas, podemos calcular os valores correspondentes a cada fator para cada observação. A Tabela 10.10 mostra esses resultados para parte do banco de dados.

**Tabela 10.10** Cálculo dos fatores para cada observação.

Estudante	$Z_{\text{finanças}_i}$	$Z_{\text{custos}_i}$	$Z_{\text{marketing}_i}$	$Z_{\text{atuária}_i}$	$F_{1i}$	$F_{2i}$	$F_{3i}$	$F_{4i}$
Gabriela	-0,011	-0,290	-1,650	0,273	0,016	-1,665	-0,176	0,739
Luiz Felipe	-0,876	-0,697	1,532	-1,319	-1,076	1,503	0,342	-0,831
Patrícia	-0,876	-0,290	-0,590	-0,523	-0,600	-0,603	-0,634	-0,672
Gustavo	1,334	1,337	0,825	1,069	1,346	0,887	0,327	-0,228
Letícia	-0,779	-1,104	-0,872	-0,841	-0,978	-0,922	0,161	0,379
Ovídio	1,334	2,150	-1,650	1,865	1,979	-1,553	-0,812	-0,841
Leonor	-0,267	0,116	0,825	-0,125	-0,111	0,829	-0,312	-0,429
Dalila	-0,139	0,523	0,118	0,273	0,242	0,139	-0,694	-0,623
Antônio	0,021	-0,290	-0,590	-0,523	-0,281	-0,597	0,682	-0,250
...								
Estela	0,982	0,113	-1,297	1,069	0,802	-1,293	0,305	1,616
<b>Média</b>	<b>0,000</b>	<b>0,000</b>	<b>0,000</b>	<b>0,000</b>	<b>0,000</b>	<b>0,000</b>	<b>0,000</b>	<b>0,000</b>
<b>Desvio-Padrão</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>

Para, por exemplo, a primeira observação da amostra (**Gabriela**), podemos verificar que:

$$\begin{aligned}
F_{1\text{Gabriela}} &= 0,355 \cdot (-0,011) + 0,371 \cdot (-0,290) - 0,017 \cdot (-1,650) + 0,364 \cdot (0,273) = 0,016 \\
F_{2\text{Gabriela}} &= 0,007 \cdot (-0,011) + 0,049 \cdot (-0,290) + 0,999 \cdot (-1,650) - 0,010 \cdot (0,273) = -1,665 \\
F_{3\text{Gabriela}} &= 1,468 \cdot (-0,011) - 0,403 \cdot (-0,290) - 0,001 \cdot (-1,650) - 1,021 \cdot (0,273) = -0,176 \\
F_{4\text{Gabriela}} &= 0,470 \cdot (-0,011) - 1,815 \cdot (-0,290) + 0,099 \cdot (-1,650) + 1,394 \cdot (0,273) = 0,739
\end{aligned}$$

Ressalta-se que todos os fatores extraídos apresentam, entre si, correlações de Pearson iguais a 0, ou seja, **são ortogonais entre si**.

Um pesquisador mais curioso poderá ainda verificar que os *scores* fatoriais correspondentes a cada fator são exatamente os parâmetros estimados de um modelo de regressão linear múltipla que apresenta, como variável dependente, o próprio fator, e como variáveis explicativas, as variáveis padronizadas.

Estabelecidos os fatores, podemos definir as cargas fatoriais, que correspondem aos coeficientes de correlação de Pearson entre as variáveis originais e cada um dos fatores. A Tabela 10.11 apresenta as cargas fatoriais para os dados do nosso exemplo.

**Tabela 10.11** Cargas fatoriais (coeficientes de correlação de Pearson) entre variáveis e fatores.

<b>Variável \ Fator</b>	<b>F<sub>1</sub></b>	<b>F<sub>2</sub></b>	<b>F<sub>3</sub></b>	<b>F<sub>4</sub></b>
<i>finanças</i>	0,895	0,007	0,437	0,086
<i>custos</i>	0,934	0,049	-0,120	-0,332
<i>marketing</i>	-0,042	0,999	0,000	0,018
<i>atuária</i>	0,918	-0,010	-0,304	0,255

Para cada variável original, foi destacado na Tabela 10.11 o maior valor da carga fatorial. Logo, podemos verificar que, enquanto as variáveis *finanças*, *custos* e *atuária* apresentam maiores correlações com o primeiro fator, apenas a variável *marketing* apresenta maior correlação com o segundo fator. Isso comprova a necessidade de um segundo fator para que todas as variáveis compartilhem percentuais significativos de variância. Entretanto, o terceiro e quarto fatores apresentam correlações relativamente baixas com as variáveis originais, o que explica que os respectivos autovalores sejam menores que 1. Caso a variável *marketing* não tivesse sido inserida na análise, apenas o primeiro fator seria necessário para explicar o comportamento conjunto das demais variáveis, e os demais fatores também apresentariam respectivos autovalores menores que 1.

Logo, conforme discutimos na seção 10.2.4, podemos verificar que cargas fatoriais entre fatores correspondentes a autovalores menores que 1 são relativamente baixas, visto que já apresentaram correlações de Pearson mais elevadas com fatores extraídos anteriormente a partir de autovalores maiores.

Com base na expressão (10.30), podemos verificar que a somatória dos quadrados das cargas fatoriais em cada coluna da Tabela 10.11 será o respectivo autovalor que, conforme discutimos, pode ser entendido como o percentual de variância compartilhada pelas quatro variáveis originais para a formação de cada fator. Logo, temos que:

$$(0,895)^2 + (0,934)^2 + (-0,042)^2 + (0,918)^2 = 2,519$$

$$(0,007)^2 + (0,049)^2 + (0,999)^2 + (-0,010)^2 = 1,000$$

$$(0,437)^2 + (-0,120)^2 + (0,000)^2 + (-0,304)^2 = 0,298$$

$$(0,086)^2 + (-0,332)^2 + (0,018)^2 + (0,255)^2 = 0,183$$

de onde podemos comprovar que o segundo autovalor somente atingiu o valor 1 por conta da alta carga fatorial existente para a variável *marketing*.

Além disso, a partir das cargas fatoriais apresentadas na Tabela 10.11, podemos também calcular as comunicações, que representam a variância total compartilhada de cada variável em todos os fatores extraídos a partir de autovalores maiores que 1. Logo, podemos escrever, com base na expressão (10.29), que:

$$\text{comunidade}_{\text{finanças}} = (0,895)^2 + (0,007)^2 = 0,802$$

$$\text{comunidade}_{\text{custos}} = (0,934)^2 + (0,049)^2 = 0,875$$

$$\text{comunidade}_{\text{marketing}} = (-0,042)^2 + (0,999)^2 = 1,000$$

$$\text{comunidade}_{\text{atuária}} = (0,918)^2 + (-0,010)^2 = 0,843$$

Logo, embora a variável *marketing* seja a única que apresenta carga fatorial elevada com o segundo fator, é a variável em que menor percentual de variância é perdido para a formação dos dois fatores. Por outro lado, a variável *finanças* é a que apresenta maior perda de variância para a formação desses dois fatores (cerca de 19,8%). Se tivéssemos considerado as cargas fatoriais dos quatro fatores, obviamente todas as comunicações seriam iguais a 1.



Conforme discutimos na seção 10.2.4, pode-se verificar que as cargas fatoriais são exatamente os parâmetros estimados de um modelo de regressão linear múltipla, que apresenta, como variável dependente, determinada variável padronizada e, como variáveis explicativas, os próprios fatores, sendo o coeficiente de ajuste  $R^2$  de cada modelo igual à comunalidade da respectiva variável original.

Para os dois primeiros fatores, portanto, podemos elaborar um gráfico em que são plotadas as cargas fatoriais de cada variável em cada um dos eixos ortogonais que representam, respectivamente, os fatores  $F_1$  e  $F_2$ . Esse gráfico, conhecido por *loading plot*, encontra-se na Figura 10.8.

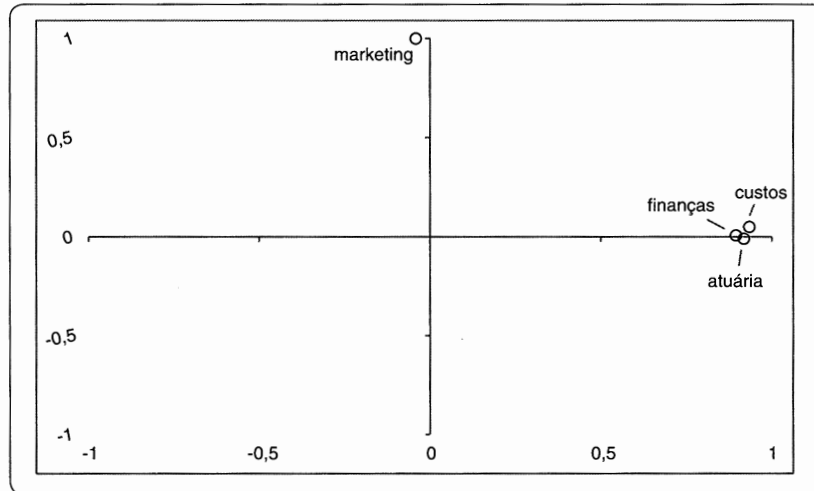


Figura 10.8 Loading plot.

Por meio da análise do *loading plot*, fica claro o comportamento das correlações. Enquanto as variáveis *finanças*, *custos* e *atuária* apresentam elevada correlação com o primeiro fator (eixo das abscissas), a variável *marketing* apresenta forte correlação com o segundo fator (eixo das ordenadas). Um pesquisador mais curioso poderá investigar as razões por que ocorre esse fenômeno, visto que, por vezes, enquanto as disciplinas Finanças, Custos e Atuária são ministradas de forma mais quantitativa, a disciplina Marketing pode ser ministrada com apelo mais qualitativo e comportamental. É importante mencionar, contudo, que a definição de fatores não obriga o pesquisador a nomeá-los, já que frequentemente não é tarefa simples. **A análise fatorial não tem, como um de seus objetivos, a nomeação de fatores**, e, caso haja a intenção de fazê-lo, é necessário que o pesquisador tenha profundo conhecimento sobre o fenômeno em estudo, e **técnicas confirmatórias** podem auxiliá-lo nessa empreitada.

Podemos considerar, neste momento, encerrada a elaboração da análise fatorial por componentes principais. Entretanto, conforme discutimos na seção 10.2.5, caso o pesquisador deseje obter melhor visualização das variáveis mais representadas por determinado fator, pode elaborar uma rotação por meio do método ortogonal Varimax, que maximiza as cargas de cada variável em determinado fator. Como, em nosso exemplo, já temos uma excelente ideia das variáveis com altas cargas em cada fator, sendo o *loading plot* (Figura 10.8) já bastante claro, a rotação pode ser considerada desnecessária. Será elaborada, portanto, apenas para efeitos didáticos, visto que, por vezes, o pesquisador pode se deparar com situações em que tal fenômeno não se apresente de forma tão clara.

Logo, com base nas cargas fatoriais para os dois primeiros fatores (duas primeiras colunas da Tabela 10.11), obtaremos as cargas fatoriais rotacionadas  $c'$  após a rotação dos dois fatores por um ângulo  $\theta$ . Sendo assim, com base na expressão (10.35), podemos escrever que:

$$\begin{pmatrix} 0,895 & 0,007 \\ 0,934 & 0,049 \\ -0,042 & 0,999 \\ 0,918 & -0,010 \end{pmatrix} \cdot \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} c'_{11} & c'_{12} \\ c'_{21} & c'_{22} \\ \vdots & \vdots \\ c'_{k1} & c'_{k2} \end{pmatrix}$$

em que o ângulo de rotação no sentido anti-horário  $\theta$  é obtido a partir da expressão (10.36). Antes, entretanto, devemos determinar os valores dos termos  $A$ ,  $B$ ,  $C$  e  $D$  presentes nas expressões (10.37) a (10.40). A construção das Tabelas 10.12 a 10.15 nos auxilia para essa finalidade.

**Tabela 10.12** Obtenção do termo  $A$  para cálculo do ângulo de rotação  $\theta$ .

Variável	$c_1$	$c_2$	comunalidade	$\left( \frac{c_{1i}^2}{\text{comunalidade}_i} - \frac{c_{2i}^2}{\text{comunalidade}_i} \right)$
<i>finanças</i>	0,895	0,007	0,802	1,000
<i>custos</i>	0,934	0,049	0,875	0,995
<i>marketing</i>	-0,042	0,999	1,000	-0,996
<i>atuária</i>	0,918	-0,010	0,843	1,000
			$A$ (soma)	1,998

**Tabela 10.13** Obtenção do termo  $B$  para cálculo do ângulo de rotação  $\theta$ .

Variável	$c_1$	$c_2$	comunalidade	$\left( 2 \cdot \frac{c_{1i} \cdot c_{2i}}{\text{comunalidade}_i} \right)$
<i>finanças</i>	0,895	0,007	0,802	0,015
<i>custos</i>	0,934	0,049	0,875	0,104
<i>marketing</i>	-0,042	0,999	1,000	-0,085
<i>atuária</i>	0,918	-0,010	0,843	-0,022
			$B$ (soma)	0,012

**Tabela 10.14** Obtenção do termo  $C$  para cálculo do ângulo de rotação  $\theta$ .

Variável	$c_1$	$c_2$	comunalidade	$\left( \frac{c_{1i}^2}{\text{comunalidade}_i} - \frac{c_{2i}^2}{\text{comunalidade}_i} \right)^2 - \left( 2 \cdot \frac{c_{1i} \cdot c_{2i}}{\text{comunalidade}_i} \right)^2$
<i>finanças</i>	0,895	0,007	0,802	1,000
<i>custos</i>	0,934	0,049	0,875	0,978
<i>marketing</i>	-0,042	0,999	1,000	0,986
<i>atuária</i>	0,918	-0,010	0,843	0,999
			$C$ (soma)	3,963

**Tabela 10.15** Obtenção do termo  $D$  para cálculo do ângulo de rotação  $\theta$ .

Variável	$c_1$	$c_2$	comunalidade	$\left( \frac{c_{1i}^2}{\text{comunalidade}_i} - \frac{c_{2i}^2}{\text{comunalidade}_i} \right) \cdot \left( 2 \cdot \frac{c_{1i} \cdot c_{2i}}{\text{comunalidade}_i} \right)$
<i>finanças</i>	0,895	0,007	0,802	0,015
<i>custos</i>	0,934	0,049	0,875	0,103
<i>marketing</i>	-0,042	0,999	1,000	0,084
<i>atuária</i>	0,918	-0,010	0,843	-0,022
			$D$ (soma)	0,181

Logo, levando em consideração as  $k = 4$  variáveis, e com base na expressão (10.36), podemos calcular o ângulo de rotação no sentido anti-horário  $\theta$  da seguinte forma:

$$\theta = 0,25 \cdot \arctan \left[ \frac{2 \cdot (D \cdot k - A \cdot B)}{C \cdot k - (A^2 - B^2)} \right] = 0,25 \cdot \arctan \left\{ \frac{2 \cdot [(0,181) \cdot 4 - (1,998) \cdot (0,012)]}{(3,9636) \cdot 4 - [(1,998)^2 - (0,012)^2]} \right\} = 0,029 \text{ rad}$$

E, por fim, podemos calcular as cargas fatoriais rotacionadas:

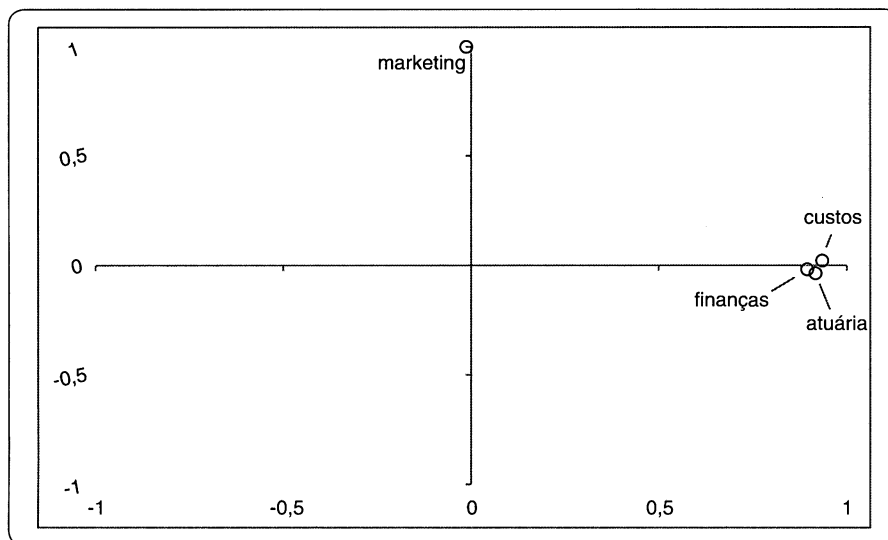
$$\begin{pmatrix} 0,895 & 0,007 \\ 0,934 & 0,049 \\ -0,042 & 0,999 \\ 0,918 & -0,010 \end{pmatrix} \cdot \begin{pmatrix} \cos 0,029 & -\sin 0,029 \\ \sin 0,029 & \cos 0,029 \end{pmatrix} = \begin{pmatrix} c'_{11} & c'_{12} \\ c'_{21} & c'_{22} \\ c'_{31} & c'_{32} \\ c'_{41} & c'_{42} \end{pmatrix} = \begin{pmatrix} 0,895 & -0,019 \\ 0,935 & 0,021 \\ -0,013 & 1,000 \\ 0,917 & -0,037 \end{pmatrix}$$

A Tabela 10.16 apresenta, de forma consolidada, as cargas fatoriais rotacionadas pelo método Varimax para os dados de nosso exemplo.

**Tabela 10.16** Cargas fatoriais rotacionadas pelo método Varimax.

<b>Variável \ Fator</b>	<b><math>F'_1</math></b>	<b><math>F'_2</math></b>
<i>finanças</i>	0,895	-0,019
<i>custos</i>	0,935	0,021
<i>marketing</i>	-0,013	1,000
<i>atuária</i>	0,917	-0,037

Conforme já mencionamos, embora os resultados sem a rotação já demonstrassem quais variáveis apresentavam elevadas cargas em cada fator, a rotação acabou por distribuir, ainda que levemente para os dados do nosso exemplo, as cargas das variáveis em cada um dos fatores rotacionados. Um novo *loading plot* (agora com cargas rotacionadas) também pode demonstrar essa situação (Figura 10.9).



**Figura 10.9** Loading plot com cargas rotacionadas.

Embora os gráficos das Figuras 10.8 e 10.9 sejam muito parecidos, visto que o ângulo de rotação  $\theta$  é bastante pequeno neste exemplo, é comum que o pesquisador encontre situações em que a rotação irá contribuir consideravelmente para a elaboração de uma leitura mais fácil das cargas, o que pode, consequentemente, simplificar a interpretação dos fatores.

É importante frisarmos que a rotação não altera as communalidades, ou seja, a expressão (10.31) pode ser verificada:

$$\text{comunalidade}_{\text{finanças}} = (0,895)^2 + (-0,019)^2 = 0,802$$

$$\text{comunalidade}_{\text{custos}} = (0,935)^2 + (0,021)^2 = 0,875$$

$$\text{comunalidade}_{\text{marketing}} = (-0,013)^2 + (1,000)^2 = 1,000$$

$$\text{comunalidade}_{\text{atuária}} = (0,917)^2 + (-0,037)^2 = 0,843$$

Entretanto, a rotação altera os autovalores correspondentes a cada fator. Sendo assim, temos, para os dois fatores rotacionados, que:

$$(0,895)^2 + (0,935)^2 + (-0,013)^2 + (0,917)^2 = \lambda_1'^2 = 2,518$$

$$(-0,019)^2 + (0,021)^2 + (1,000)^2 + (-0,037)^2 = \lambda_2'^2 = 1,002$$

A Tabela 10.17 apresenta, com base nos novos autovalores  $\lambda_1^2$  e  $\lambda_2^2$ , os percentuais de variância compartilhada pelas variáveis originais para a formação dos dois fatores rotacionados.

**Tabela 10.17** Variância compartilhada pelas variáveis originais para a formação dos dois fatores rotacionados.

Fator	Autovalor $\lambda^2$	Variância Compartilhada (%)	Variância Compartilhada Acumulada (%)
1	2,518	$\left(\frac{2,518}{4}\right) \cdot 100 = 62,942$	62,942
2	1,002	$\left(\frac{1,002}{4}\right) \cdot 100 = 25,043$	87,985

Em comparação à Tabela 10.9, podemos perceber que, embora não haja alteração do compartilhamento de 87,985% da variância total das variáveis originais para a formação dos fatores rotacionados, a rotação redistribui a variância compartilhada pelas variáveis em cada fator.

Conforme discutimos, as cargas fatoriais correspondem aos parâmetros estimados de um modelo de regressão linear múltipla que apresenta, como variável dependente, determinada variável padronizada e, como variáveis explicativas, os próprios fatores. Dessa forma, podemos, por meio de operações algébricas, chegar às expressões dos *scores* fatoriais a partir das cargas, visto que eles representam parâmetros estimados dos respectivos modelos de regressão que têm, como variável dependente, os fatores e, como variáveis explicativas, as variáveis padronizadas. Logo, chegamos, a partir das cargas fatoriais rotacionadas (Tabela 10.16), às seguintes expressões dos fatores rotacionados  $F'_1$  e  $F'_2$ .

$$F'_{1i} = 0,355 \cdot Z_{finanças_i} + 0,372 \cdot Z_{custos_i} + 0,012 \cdot Z_{marketing_i} + 0,364 \cdot Z_{atuária_i}$$

$$F'_{2i} = -0,004 \cdot Z_{finanças_i} + 0,038 \cdot Z_{custos_i} + 0,999 \cdot Z_{marketing_i} - 0,021 \cdot Z_{atuária_i}$$

Por fim, o professor deseja criar um *ranking* de desempenho escolar de seus alunos. Como os dois fatores rotacionados,  $F'_1$  e  $F'_2$ , são formados pelos maiores percentuais de variância compartilhada pelas variáveis originais (no caso, 62,942% e 25,043% da variância total, respectivamente, conforme mostra a Tabela 10.17) e correspondem a autovalores maiores que 1, serão utilizados para que seja elaborado o desejado *ranking* de desempenho escolar.

Um critério bastante aceito e utilizado para a formação de *rankings* a partir de fatores é conhecido como **critério da soma ponderada e ordenamento**, em que são somados, para cada observação, os valores obtidos de todos os fatores (que possuem autovalores maiores que 1) ponderados pelos respectivos percentuais de variância compartilhada, com o subsequente ordenamento das observações com base nos resultados obtidos. Esse critério é bastante aceito por considerar o desempenho em todas as variáveis originais, visto que a consideração apenas do primeiro fator (**critério do fator principal**) pode não levar em conta, por exemplo, o desempenho positivo obtido em determinada variável que eventualmente compartilhe um considerável percentual de variância com o segundo fator. A Tabela 10.18 mostra, para 10 alunos escolhidos na amostra, o resultado do *ranking* de desempenho escolar resultante do ordenamento elaborado após a soma dos valores obtidos dos fatores ponderados pelos respectivos percentuais de variância compartilhada.

O *ranking* completo pode ser acessado no arquivo **NotasFatorialRanking.xls**.

É de fundamental importância ressaltar que a criação de *rankings* de desempenho a partir de variáveis originais é um procedimento considerado **estático**, visto que a inclusão de novas observações ou variáveis pode alterar os *scores* fatoriais, o que torna obrigatória a elaboração de uma nova análise fatorial. A própria evolução temporal dos fenômenos representados pelas variáveis pode alterar a matriz de correlações, o que torna necessária a reaplicação da técnica para que sejam gerados novos fatores obtidos a partir de *scores* mais precisos e atualizados. Aqui cabe, portanto, uma crítica a indicadores socioeconômicos que utilizam *scores* estáticos previamente estabelecidos para cada variável no cálculo do fator a ser utilizado para a definição do *ranking* em situações em que novas observações sejam constantemente incluídas; mais que isso, em situações em que haja a evolução temporal, que altera a matriz de correlações das variáveis originais em cada período.

**Tabela 10.18** *Ranking* de desempenho escolar pelo critério da soma ponderada e ordenamento.

Estudante	$Z_{finanças_i}$	$Z_{custos_i}$	$Z_{marketing_i}$	$Z_{atuária_i}$	$F_{1i}$	$F_{2i}$	$(F_{1i} \cdot 0,62942) + (F_{2i} \cdot 0,25043)$	ranking
Adelino	1,30	2,15	1,53	1,86	1,959	1,568	1,626	1
Renata	0,60	2,15	1,53	1,86	1,709	1,570	1,469	2
Ovídio	1,33	2,15	-1,65	1,86	1,932	-1,611	0,813	13
Kamal	1,33	2,07	-1,65	1,86	1,902	-1,614	0,793	14
Itamar	-1,29	-0,55	1,53	-1,04	-1,022	1,536	-0,259	57
Luiz Felipe	-0,88	-0,70	1,53	-1,32	-1,032	1,535	-0,265	58
Gabriela	-0,01	-0,29	-1,65	0,27	-0,032	-1,665	-0,437	73
Marina	0,50	-0,50	-0,94	-1,16	-0,443	-0,939	-0,514	74
Viviane	-1,64	-1,16	-1,01	-1,00	-1,390	-1,029	-1,133	99
Gilmar	-1,52	-1,16	-1,40	-1,44	-1,512	-1,409	-1,304	100

Vale comentar que os fatores extraídos são variáveis quantitativas e, portanto, a partir deles, podem ser elaboradas outras técnicas multivariadas exploratórias, como análise de agrupamentos, dependendo dos objetivos do pesquisador. Além disso, cada fator também pode ser transformado em uma variável qualitativa, por meio, por exemplo, de sua categorização em faixas estabelecidas com base em determinado critério e, a partir de então, ser elaborada uma análise de correspondência, a fim de avaliar uma eventual associação entre as categorias criadas e as categorias de outras variáveis qualitativas, conforme estudaremos no próximo capítulo.

Os fatores podem também ser utilizados como variáveis explicativas de determinado fenômeno em modelos multivariados confirmatórios como, por exemplo, modelos de regressão múltipla, visto que a ortogonalidade elimina problemas de multicolinearidade. Por outro lado, tal procedimento somente faz sentido quando há o intuito de um diagnóstico acerca do comportamento da variável dependente, sem a intenção de previsões. Como novas observações não apresentam os correspondentes valores dos fatores gerados, sua obtenção somente é possível ao se incluírem tais observações em nova análise fatorial, a fim de se obterem novos *scores* fatoriais, já que se trata de uma técnica exploratória.

Além disso, uma variável qualitativa obtida por meio da categorização em faixas de determinado fator também pode ser inserida como variável dependente de um modelo de regressão logística multinomial, permitindo que o pesquisador avalie as probabilidades que cada observação tem de pertencer a cada faixa, em função do comportamento de outras variáveis explicativas não inicialmente consideradas na análise fatorial. Ressaltamos, da mesma forma, que esse procedimento apresenta caráter de diagnóstico do comportamento das variáveis na amostra para as observações existentes, sem finalidade preditiva.

Na sequência, esse mesmo exemplo será elaborado nos softwares SPSS e Stata. Enquanto na seção 10.3 serão apresentados os procedimentos para elaboração da análise fatorial por componentes principais no SPSS, assim como seus resultados, na seção 10.4 serão apresentados os comandos para a elaboração da técnica no Stata, com respectivos *outputs*.

### 10.3. ANÁLISE FATORIAL POR COMPONENTES PRINCIPAIS NO SOFTWARE SPSS

Nesta seção, apresentaremos o passo a passo para a elaboração do nosso exemplo no IBM SPSS Statistics Software®. Seguindo a lógica proposta no livro, o principal objetivo é propiciar ao pesquisador uma oportunidade de elaborar a análise fatorial por componentes principais neste software, dada sua facilidade de manuseio e a didática das operações. A cada apresentação de um *output*, faremos menção ao respectivo resultado obtido quando da solução algébrica da técnica na seção anterior, a fim de que o pesquisador possa compará-los e formar seu conhecimento e erudição sobre o tema. A reprodução das imagens nesta seção tem autorização da International Business Machines Corporation®.

Voltando ao exemplo apresentado na seção 10.2.6, lembremos que o professor tem interesse em elaborar um *ranking* de desempenho escolar de seus alunos com base no comportamento conjunto das notas finais de quatro disciplinas. Os dados encontram-se no arquivo **NotasFatorial.sav** e são exatamente iguais aos apresentados parcialmente na Tabela 10.5 da seção 10.2.6.

Para que seja elaborada, portanto, a análise fatorial, vamos clicar em **Analyze → Dimension Reduction → Factor....** Uma caixa de diálogo como a apresentada na Figura 10.10 será aberta.

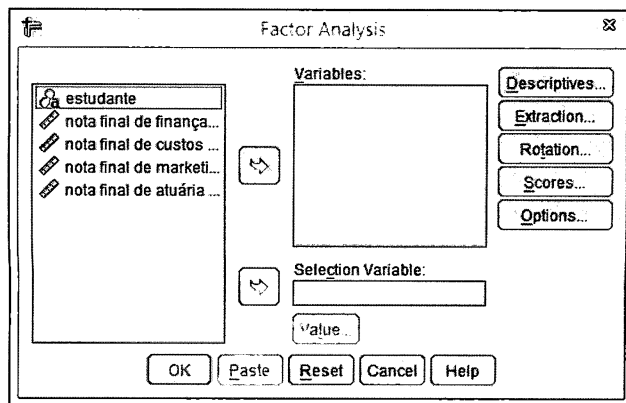


Figura 10.10 Caixa de diálogo para elaboração da análise fatorial no SPSS.

Na sequência, devemos inserir as variáveis originais *finanças*, *custos*, *marketing* e *atuária* em **Variables**, conforme mostra a Figura 10.11.

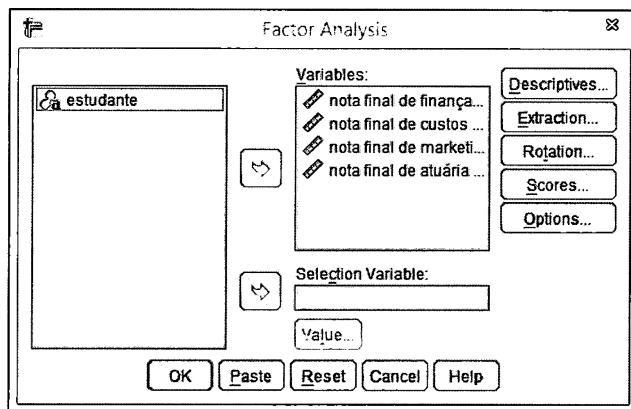


Figura 10.11 Seleção das variáveis originais.

Ao contrário do discutido no capítulo anterior, quando da elaboração da análise de agrupamentos, é importante mencionar que o pesquisador não precisa se preocupar com a padronização *Zscores* das variáveis originais para a elaboração da análise fatorial, visto que as correlações entre variáveis originais ou entre suas correspondentes variáveis padronizadas são exatamente as mesmas. Mesmo assim, **caso o pesquisador opte por padronizar cada uma das variáveis, irá perceber que os outputs serão exatamente os mesmos.**

No botão **Descriptives...**, marcaremos primeiramente a opção **Initial solution** em **Statistics**, que faz com que sejam apresentados nos *outputs* todos os autovalores da matriz de correlações, mesmo os menores que 1. Além disso, vamos também selecionar as opções **Coefficients**, **Determinant** e **KMO and Bartlett's test of sphericity** em **Correlation Matrix**, conforme mostra a Figura 10.12.

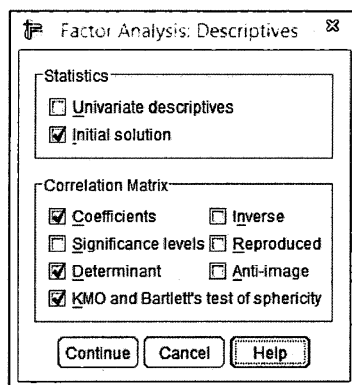
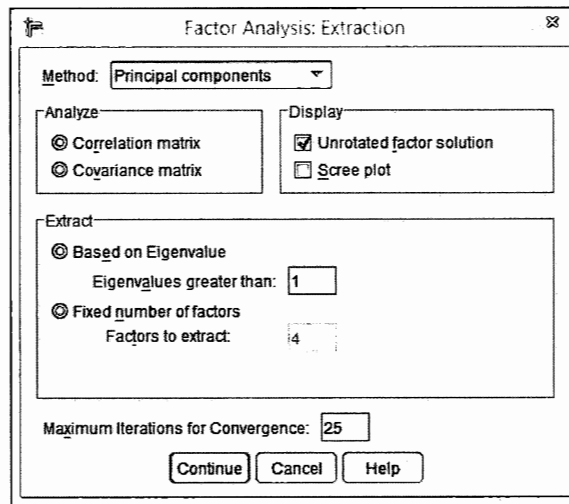


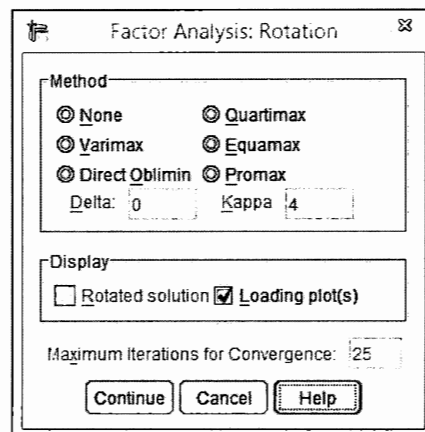
Figura 10.12 Seleção das opções iniciais para elaboração da análise fatorial.

Ao clicarmos em **Continue**, voltaremos para a caixa de diálogo principal da análise fatorial. Na sequência, devemos clicar no botão **Extraction...** Conforme mostra a Figura 10.13, iremos manter selecionadas as opções referentes ao método de extração dos fatores (**Method: Principal components**) e ao critério de escolha da quantidade de fatores. Nesse caso, conforme discutimos na seção 10.2.3, serão levados em consideração apenas os fatores correspondentes a autovalores maiores que 1 (critério da raiz latente ou critério de Kaiser), e, portanto, devemos manter selecionada a opção **Based on Eigenvalue** → **Eigenvalues greater than: 1** em **Extract**. Além disso, vamos também manter selecionadas as opções **Unrotated factor solution**, em **Display**, e **Correlation matrix**, em **Analyze**.



**Figura 10.13** Escolha do método de extração dos fatores e do critério para determinação da quantidade de fatores.

Da mesma forma, vamos clicar em **Continue** para que retornemos à caixa de diálogo principal da análise fatorial. Em **Rotation...**, vamos, por enquanto, selecionar a opção **Loading plot(s)** em **Display**, mantendo ainda selecionada a opção **None** em **Method**, conforme mostra a Figura 10.14.



**Figura 10.14** Caixa de diálogo para seleção do método de rotação e do *loading plot*.

A opção pela extração de fatores ainda não rotacionados neste momento é didática, visto que os *outputs* gerados poderão ser comparados com os obtidos algebricamente na seção 10.2.6. O pesquisador já pode, entretanto, optar por extrair fatores rotacionados já nesta oportunidade.

Após clicarmos em **Continue**, podemos selecionar o botão **Scores...** na caixa de diálogo principal da técnica. Neste momento, selecionaremos apenas a opção **Display factor score coefficient matrix**, conforme mostra a Figura 10.15, que faz com que sejam apresentados, nos *outputs*, os *scores* fatoriais correspondentes a cada fator extraído.

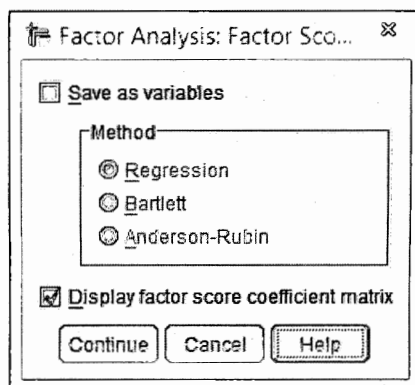


Figura 10.15 Seleção da opção para apresentação dos scores fatoriais.

Na sequência, podemos clicar em **Continue** e em **OK**.

O primeiro *output* (Figura 10.16) apresenta a matriz de correlações  $\rho$ , igual à da Tabela 10.6 da seção 10.2.6, por meio da qual podemos verificar que a variável *marketing* é a única que apresenta baixos coeficientes de correlação de Pearson com todas as demais variáveis. Conforme discutimos, é um primeiro indício de que as variáveis *finanças*, *custos* e *atuária* podem ser correlacionadas com determinado fator, enquanto a variável *marketing* pode se correlacionar fortemente com outro.

Correlation Matrix <sup>a</sup>					
		nota final de finanças (0 a 10)	nota final de custos (0 a 10)	nota final de marketing (0 a 10)	nota final de atuária (0 a 10)
Correlation	nota final de finanças (0 a 10)	1,000	,756	-,030	,711
	nota final de custos (0 a 10)	,756	1,000	,003	,809
	nota final de marketing (0 a 10)	-,030	,003	1,000	-,044
	nota final de atuária (0 a 10)	,711	,809	-,044	1,000

a. Determinant = ,137

Figura 10.16 Coeficientes de correlação de Pearson.

Podemos também verificar que o *output* da Figura 10.16 ainda traz o valor do determinante da matriz de correlações  $\rho$ , utilizado para o cálculo da estatística  $\chi^2_{\text{Bartlett}}$ , conforme discutimos quando da apresentação da expressão (10.9).

A fim de estudarmos a adequação global da análise fatorial, vamos analisar os *outputs* da Figura 10.17, que apresenta os resultados dos cálculos correspondentes à estatística KMO e  $\chi^2_{\text{Bartlett}}$ . Enquanto a primeira indica, com base no critério apresentado no Quadro 10.1, que a adequação global da análise fatorial é considerada **mé-dia** (KMO = 0,737), a estatística  $\chi^2_{\text{Bartlett}} = 192,335$  (Sig.  $\chi^2_{\text{Bartlett}} < 0,05$  para 6 graus de liberdade) permite-nos rejeitar, ao nível de significância de 5% e com base nas hipóteses do teste de esfericidade de Bartlett, que a matriz de correlações  $\rho$  seja estatisticamente igual à matriz identidade **I** de mesma dimensão. Logo, podemos concluir que a análise fatorial é apropriada.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,737
Bartlett's Test of Sphericity	Approx. Chi-Square	192,335
	df	6
	Sig.	,000

Figura 10.17 Resultados da estatística KMO e do teste de esfericidade de Bartlett.



Os valores das estatísticas KMO e  $\chi^2_{\text{Bartlett}}$  são calculados, respectivamente, por meio das expressões (10.3) e (10.9) apresentadas na seção 10.2.2, e são exatamente iguais aos obtidos algebricamente na seção 10.2.6.

Na sequência, a Figura 10.18 apresenta os quatro autovalores da matriz de correlações  $\rho$  correspondentes a cada um dos fatores extraídos inicialmente, com os respectivos percentuais de variância compartilhada pelas variáveis originais.

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,519	62,975	62,975	2,519	62,975	62,975
2	1,000	25,010	87,985	1,000	25,010	87,985
3	,298	7,444	95,428			
4	,183	4,572	100,000			

Extraction Method: Principal Component Analysis.

**Figura 10.18** Autovalores e variância compartilhada pelas variáveis originais para a formação de cada fator.

Note que os autovalores são exatamente iguais aos obtidos algebricamente na seção 10.2.6, de modo que:

$$\lambda_1^2 + \lambda_2^2 + \dots + \lambda_k^2 = 2,519 + 1,000 + 0,298 + 0,183 = 4$$

Como consideraremos na análise apenas os fatores cujos autovalores sejam maiores que 1, a parte direita da Figura 10.18 mostra o percentual de variância compartilhada pelas variáveis originais para a formação apenas desses fatores. Logo, de forma análoga ao apresentado na Tabela 10.9, podemos afirmar que, enquanto 62,975% da variância total são compartilhados para a formação do primeiro fator, 25,010% são compartilhados para a formação do segundo. Portanto, para a formação desses dois fatores, a perda total de variância das variáveis originais é igual a 12,015%.

Extraídos dois fatores, a Figura 10.19 apresenta os *scores* fatoriais correspondentes a cada uma das variáveis padronizadas para cada um desses fatores.

Component Score Coefficient Matrix		
	Component	
	1	2
nota final de finanças (0 a 10)	,355	,007
nota final de custos (0 a 10)	,371	,049
nota final de marketing (0 a 10)	-,017	,999
nota final de atuária (0 a 10)	,364	-,010

Extraction Method: Principal Component Analysis.

**Figura 10.19** Scores fatoriais.

Dessa forma, temos condições de escrever as expressões dos fatores  $F_1$  e  $F_2$  conforme segue:

$$F_{1i} = 0,355 \cdot Z_{\text{finanças}_i} + 0,371 \cdot Z_{\text{custos}_i} - 0,017 \cdot Z_{\text{marketing}_i} + 0,364 \cdot Z_{\text{atuária}_i}$$

$$F_{2i} = 0,007 \cdot Z_{\text{finanças}_i} + 0,049 \cdot Z_{\text{custos}_i} + 0,999 \cdot Z_{\text{marketing}_i} - 0,010 \cdot Z_{\text{atuária}_i}$$

Note que as expressões são idênticas às obtidas na seção 10.2.6 a partir da definição algébrica dos *scores* fatoriais não rotacionados.

A Figura 10.20 apresenta as cargas fatoriais, que correspondem aos coeficientes de correlação de Pearson entre as variáveis originais e cada um dos fatores. Os valores presentes na Figura 10.20 são iguais aos apresentados nas duas primeiras colunas da Tabela 10.11.

**Component Matrix<sup>a</sup>**

	Component	
	1	2
nota final de finanças (0 a 10)	,895	,007
nota final de custos (0 a 10)	,934	,049
nota final de marketing (0 a 10)	-,042	,999
nota final de atuária (0 a 10)	,918	-,010

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

**Figura 10.20** Cargas fatoriais.

Em destaque para cada variável está a maior carga fatorial, e, portanto, podemos verificar que, enquanto as variáveis *finanças*, *custos* e *atuária* apresentam maiores correlações com o primeiro fator, apenas a variável *marketing* apresenta maior correlação com o segundo fator.

Conforme também discutimos na seção 10.2.6, a somatória dos quadrados das cargas fatoriais em coluna resulta no autovalor do correspondente fator, ou seja, representa o percentual de variância compartilhada pelas quatro variáveis originais para a formação de cada fator. Sendo assim, podemos verificar que:

$$(0,895)^2 + (0,934)^2 + (-0,042)^2 + (0,918)^2 = 2,519$$

$$(0,007)^2 + (0,049)^2 + (0,999)^2 + (-0,010)^2 = 1,000$$

Por outro lado, a somatória dos quadrados das cargas fatoriais em linha resulta na comunalidade da respectiva variável, ou seja, representa o percentual de variância compartilhada de cada variável original nos dois fatores extraídos. Nesse sentido, podemos também verificar que:

$$\text{comunalidade}_{\text{finanças}} = (0,895)^2 + (0,007)^2 = 0,802$$

$$\text{comunalidade}_{\text{custos}} = (0,934)^2 + (0,049)^2 = 0,875$$

$$\text{comunalidade}_{\text{marketing}} = (-0,042)^2 + (0,999)^2 = 1,000$$

$$\text{comunalidade}_{\text{atuária}} = (0,918)^2 + (-0,010)^2 = 0,843$$

Nos *outputs* do SPSS também é apresentada a tabela de comunalidades, conforme mostra a Figura 10.21.

**Communalities**

	Initial	Extraction
nota final de finanças (0 a 10)	1,000	,802
nota final de custos (0 a 10)	1,000	,875
nota final de marketing (0 a 10)	1,000	1,000
nota final de atuária (0 a 10)	1,000	,843

Extraction Method: Principal Component Analysis.

**Figura 10.21** Comunalidades.

O *loading plot*, que apresenta a posição relativa de cada variável em cada fator, com base nas respectivas cargas fatoriais, também é apresentado nos *outputs*, conforme mostra a Figura 10.22 (equivalente à Figura 10.8 da seção 10.2.6), em que o eixo das abscissas representa o fator  $F_1$ , e o das ordenadas, o fator  $F_2$ .

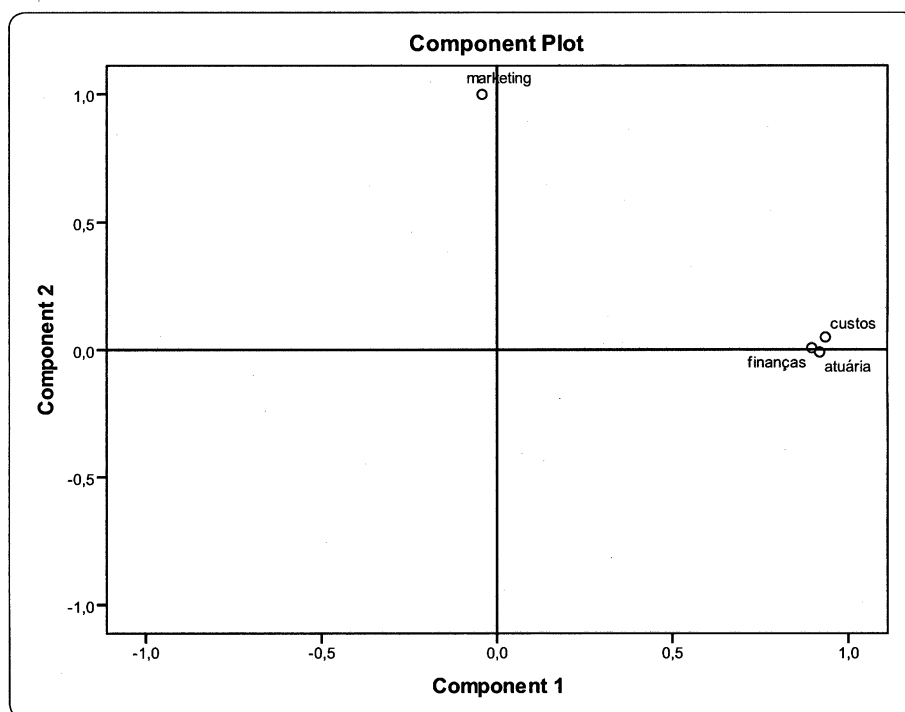


Figura 10.22 Loading plot.

Embora seja bastante clara a posição relativa das variáveis em cada eixo, ou seja, as magnitudes das correlações entre cada uma delas e cada fator, para efeitos didáticos optamos por elaborar a rotação dos eixos, que, por vezes, pode facilitar a interpretação dos fatores por propiciar melhor distribuição das cargas fatoriais das variáveis em cada fator.

Assim, vamos novamente clicar em **Analyze** → **Dimension Reduction** → **Factor...** e, no botão **Rotation...**, selecionar a opção **Varimax**, conforme mostra a Figura 10.23.

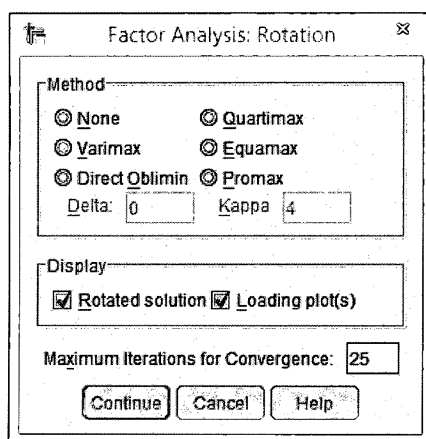
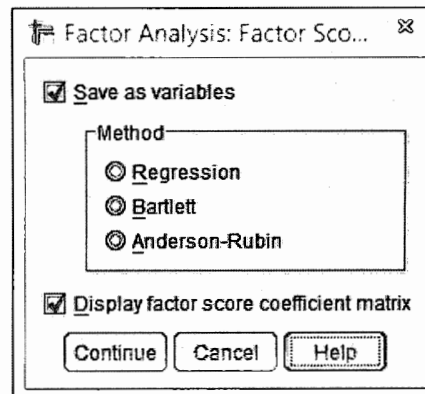


Figura 10.23 Seleção do método de rotação ortogonal Varimax.

Ao clicarmos em **Continue**, retornaremos à caixa de diálogo principal da análise fatorial. No botão **Scores...**, vamos agora selecionar a opção **Save as variables**, conforme mostra a Figura 10.24, a fim de que os fatores gerados, agora rotacionados, sejam disponibilizados no banco de dados como novas variáveis. A partir desses fatores, será elaborado o *ranking* de desempenho escolar dos alunos.



**Figura 10.24** Seleção da opção para salvar os fatores como novas variáveis no banco de dados.

Na sequência, podemos clicar em **Continue** e em **OK**.

As Figuras 10.25 a 10.29 mostram os *outputs* que apresentam diferenças, em relação aos anteriores, decorrentes da rotação. Nesse sentido, não são novamente apresentados os resultados da matriz de correlações, da estatística KMO, do teste de esfericidade de Bartlett e da tabela de communalidades que, embora calculadas a partir das cargas rotacionadas, não apresentam alterações em seus valores.

A Figura 10.25 apresenta estas cargas fatoriais rotacionadas e, por meio delas, é possível verificar, ainda que de forma tênue, certa redistribuição das cargas das variáveis em cada fator.

Rotated Component Matrix <sup>a</sup>		
	Component	
	1	2
nota final de finanças (0 a 10)	,895	-,019
nota final de custos (0 a 10)	,935	,021
nota final de marketing (0 a 10)	-,013	1,000
nota final de atuária (0 a 10)	,917	-,037

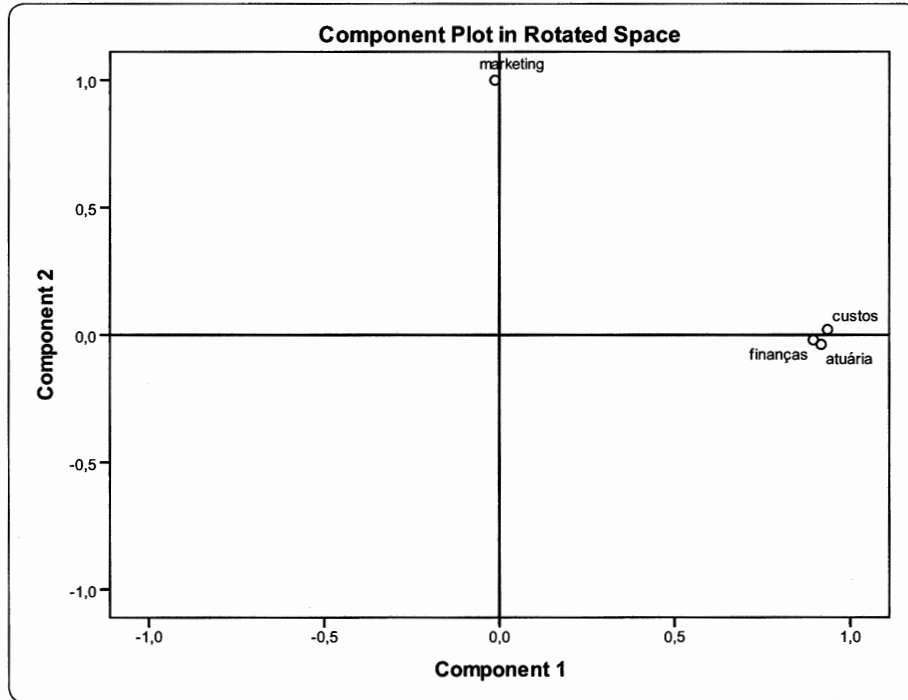
Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

**Figura 10.25** Cargas fatoriais rotacionadas pelo método Varimax.

Note que as cargas fatoriais rotacionadas da Figura 10.25 são exatamente iguais às obtidas algebricamente na seção 10.2.6, a partir das expressões (10.35) a (10.40), e apresentadas na Tabela 10.16.

O novo *loading plot*, construído a partir das cargas fatoriais rotacionadas e equivalente à Figura 10.9, encontra-se na Figura 10.26.



**Figura 10.26** Loading plot com cargas rotacionadas.

O ângulo de rotação calculado algebricamente na seção 10.2.6 também faz parte dos *outputs* do SPSS e pode ser encontrado na Figura 10.27.

Component Transformation Matrix		
Component	1	2
1	1,000	-,029
2	,029	1,000

Extraction Method: Principal  
Component Analysis.  
Rotation Method: Varimax with  
Kaiser Normalization.

**Figura 10.27** Ângulo de rotação (em radianos).

Conforme discutimos, a partir das cargas fatoriais rotacionadas, podemos verificar que não existem alterações nos valores das comunicações das variáveis consideradas na análise, ou seja:

$$\text{comunidade}_{\text{finanças}} = (0,895)^2 + (-0,019)^2 = 0,802$$

$$\text{comunidade}_{\text{custos}} = (0,935)^2 + (0,021)^2 = 0,875$$

$$\text{comunidade}_{\text{marketing}} = (-0,013)^2 + (1,000)^2 = 1,000$$

$$\text{comunidade}_{\text{atuária}} = (0,917)^2 + (-0,037)^2 = 0,843$$

Por outro lado, os novos autovalores podem ser obtidos da seguinte forma:

$$(0,895)^2 + (0,935)^2 + (-0,013)^2 + (0,917)^2 = \lambda_1'^2 = 2,518$$

$$(-0,019)^2 + (0,021)^2 + (1,000)^2 + (-0,037)^2 = \lambda_2'^2 = 1,002$$

A Figura 10.28 apresenta, em **Rotation Sums of Squared Loadings**, os resultados dos autovalores para os dois primeiros fatores rotacionados, com os respectivos percentuais de variância compartilhada pelas quatro variáveis originais. Os resultados estão de acordo com os apresentados na Tabela 10.17.

Total Variance Explained									
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,519	62,975	62,975	2,519	62,975	62,975	2,518	62,942	62,942
2	1,000	25,010	87,985	1,000	25,010	87,985	1,002	25,043	87,985
3	,298	7,444	95,428						
4	,183	4,572	100,000						

Extraction Method: Principal Component Analysis.

**Figura 10.28** Autovalores e variância compartilhada pelas variáveis originais para a formação dos dois fatores rotacionados.

Em comparação com os resultados obtidos antes da rotação, podemos perceber que, embora não haja alteração do compartilhamento de 87,985% da variância total das variáveis originais para a formação dos dois fatores rotacionados, a rotação redistribuiu a variância compartilhada pelas variáveis em cada fator.

A Figura 10.29 apresenta os *scores* fatoriais rotacionados, a partir dos quais podem ser obtidas as expressões dos novos fatores.

Component Score Coefficient Matrix		
	Component	
	1	2
nota final de finanças (0 a 10)	,355	-,004
nota final de custos (0 a 10)	,372	,038
nota final de marketing (0 a 10)	,012	,999
nota final de atuária (0 a 10)	,364	-,021

Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.  
Component Scores.

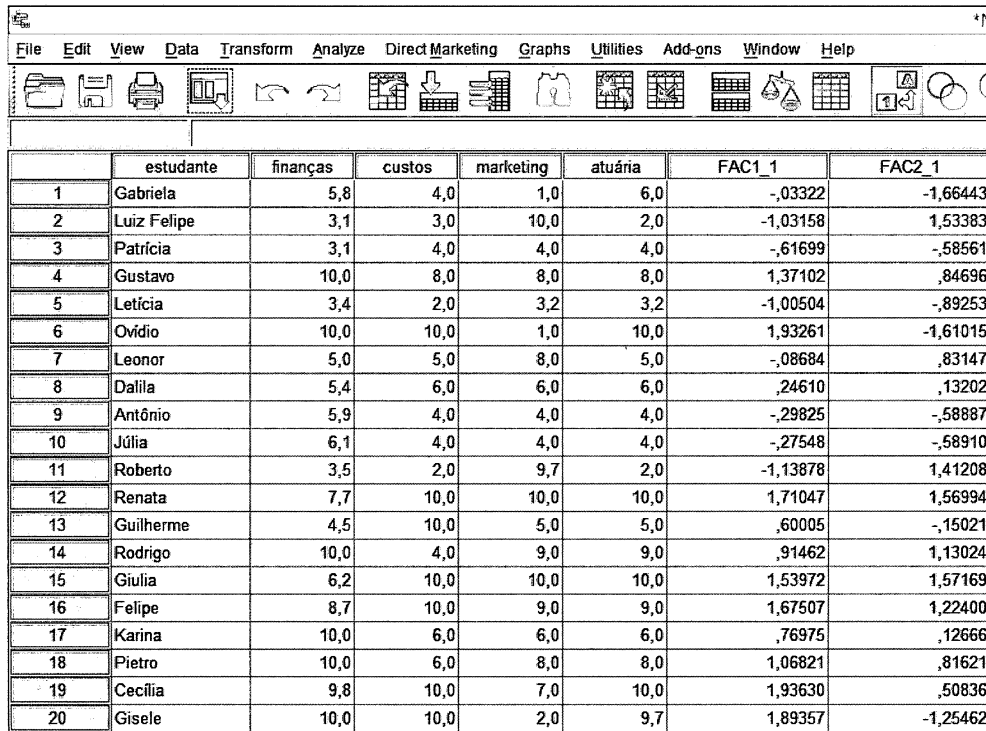
**Figura 10.29** Scores fatoriais rotacionados.

Portanto, podemos escrever as seguintes expressões dos fatores rotacionados:

$$F'_{1i} = 0,355 \cdot Z_{\text{finanças}_i} + 0,372 \cdot Z_{\text{custos}_i} + 0,012 \cdot Z_{\text{marketing}_i} + 0,364 \cdot Z_{\text{atuária}_i}$$

$$F'_{2i} = -0,004 \cdot Z_{\text{finanças}_i} + 0,038 \cdot Z_{\text{custos}_i} + 0,999 \cdot Z_{\text{marketing}_i} - 0,021 \cdot Z_{\text{atuária}_i}$$

Ao elaborarmos o procedimento descrito, podemos verificar que são geradas duas novas variáveis no banco de dados, chamadas pelo SPSS de *FAC1\_1* e *FAC2\_1*, conforme mostra a Figura 10.30 para as 20 primeiras observações.



	estudante	finanças	custos	marketing	atuária	FAC1_1	FAC2_1
1	Gabriela	5,8	4,0	1,0	6,0	-,03322	-,166443
2	Luiz Felipe	3,1	3,0	10,0	2,0	-,103158	,153383
3	Patrícia	3,1	4,0	4,0	4,0	-,61699	-,58561
4	Gustavo	10,0	8,0	8,0	8,0	,137102	,84696
5	Letícia	3,4	2,0	3,2	3,2	-,100504	-,89253
6	Ovídio	10,0	10,0	1,0	10,0	,193261	-,161015
7	Leonor	5,0	5,0	8,0	5,0	-,08684	,83147
8	Dalila	5,4	6,0	6,0	6,0	,24610	,13202
9	Antônio	5,9	4,0	4,0	4,0	-,29825	-,58887
10	Júlia	6,1	4,0	4,0	4,0	-,27548	-,58910
11	Roberto	3,5	2,0	9,7	2,0	-,113878	,141208
12	Renata	7,7	10,0	10,0	10,0	,171047	,156994
13	Guilherme	4,5	10,0	5,0	5,0	,60005	-,15021
14	Rodrigo	10,0	4,0	9,0	9,0	,91462	,113024
15	Giulia	6,2	10,0	10,0	10,0	,153972	,157169
16	Felipe	8,7	10,0	9,0	9,0	,167507	,122400
17	Karina	10,0	6,0	6,0	6,0	,76975	,12666
18	Pietro	10,0	6,0	8,0	8,0	,106821	,81621
19	Cecília	9,8	10,0	7,0	10,0	,193630	,50836
20	Gisele	10,0	10,0	2,0	9,7	,189357	-,125462

Figura 10.30 Banco de dados com os valores de  $F_1$  (FAC1\_1) e  $F_2$  (FAC2\_1) por observação.

Essas novas variáveis, que apresentam os valores dos dois fatores rotacionados para cada uma das observações do banco de dados, são ortogonais entre si, ou seja, apresentam coeficiente de correlação de Pearson igual a 0. Isso pode ser verificado ao clicarmos em **Analyze** → **Correlate** → **Bivariate...** Na caixa de diálogo que será aberta, devemos inserir as quatro variáveis originais em **Variables** e selecionar as opções **Pearson** (em **Correlation Coefficients**) e **Two-tailed** (em **Test of Significance**), conforme mostra a Figura 10.31.

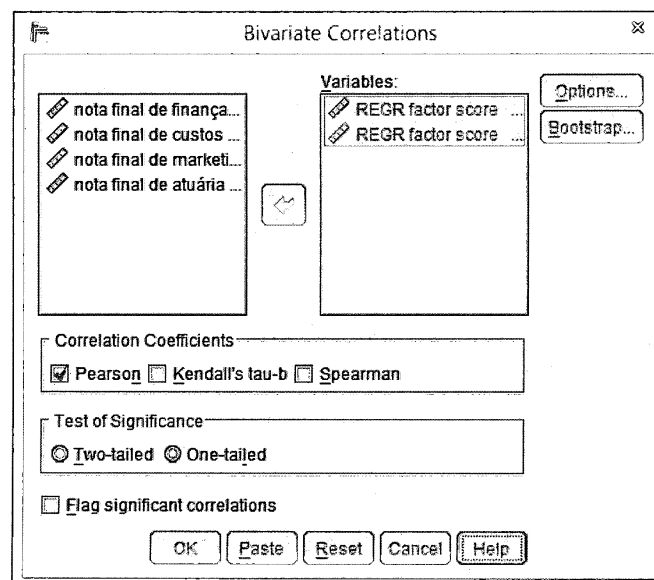


Figura 10.31 Caixa de diálogo para determinação do coeficiente de correlação de Pearson entre os dois fatores rotacionados.

Ao clicarmos em **OK**, será apresentado o *output* da Figura 10.32, em que é possível verificar que o coeficiente de correlação de Pearson entre os dois fatores rotacionados é igual a 0.

Correlations		REGR factor score 1 for analysis 1	REGR factor score 2 for analysis 1
REGR factor score 1 for analysis 1	Pearson Correlation	1	,000
	Sig. (2-tailed)		1,000
	N	100	100
REGR factor score 2 for analysis 1	Pearson Correlation	,000	1
	Sig. (2-tailed)	1,000	
	N	100	100

**Figura 10.32** Coeficiente de correlação de Pearson entre os dois fatores rotacionados.

De acordo com o estudado nas seções 10.2.4 e 10.2.6, um pesquisador mais curioso poderá ainda verificar que os *scores* fatoriais rotacionados podem ser obtidos por meio da estimação de dois modelos de regressão linear múltipla, em que é considerado, como variável dependente em cada um deles, determinado fator, e como variáveis explicativas, as variáveis padronizadas. Os *scores* fatoriais serão os parâmetros estimados em cada modelo.

Do mesmo modo, também é possível verificar que as cargas fatoriais rotacionadas também podem ser obtidas por meio da estimação de quatro modelos de regressão linear múltipla, em que é considerada, em cada um deles, determinada variável padronizada como variável dependente, e os fatores, como variáveis explicativas. Enquanto as cargas fatoriais serão os parâmetros estimados em cada modelo, as comunalidades serão os respectivos coeficientes de ajuste  $R^2$ . Portanto, podem ser obtidas as seguintes expressões:

$$Z_{finanças_i} = 0,895 \cdot F'_{1i} - 0,019 \cdot F'_{2i} + u_i, R^2 = 0,802$$

$$Z_{custos_i} = 0,935 \cdot F'_{1i} + 0,021 \cdot F'_{2i} + u_i, R^2 = 0,875$$

$$Z_{marketing_i} = -0,013 \cdot F'_{1i} + 1,000 \cdot F'_{2i} + u_i, R^2 = 1,000$$

$$Z_{atuária_i} = 0,917 \cdot F'_{1i} - 0,037 \cdot F'_{2i} + u_i, R^2 = 0,843$$

em que os termos  $u_i$  representam **fontes adicionais de variação**, além dos fatores  $F'_1$  e  $F'_2$ , para explicar o comportamento de cada variável, sendo também chamados de **termos de erro** ou **resíduos**.

Caso surja o interesse em verificar esses fatos, devemos obter as variáveis padronizadas, clicando em **Analyze** → **Descriptive Statistics** → **Descriptives....** Ao selecionarmos todas as variáveis originais, devemos clicar em **Save standardized values as variables**. Embora esse procedimento específico não seja mostrado aqui, após clicarmos em **OK**, as variáveis padronizadas serão geradas no próprio banco de dados.

Com base nos fatores gerados, temos condições, portanto, de elaborar o desejado *ranking* de desempenho escolar. Para tanto, faremos uso do critério descrito na seção 10.2.6, conhecido por critério da soma ponderada e ordenamento, em que uma nova variável é gerada a partir da multiplicação dos valores de cada fator pelos respectivos percentuais de variância compartilhada pelas variáveis originais. Neste sentido, esta nova variável, que chamaremos de *ranking*, apresenta a seguinte expressão:

$$ranking_i = 0,62942 \cdot F'_{1i} + 0,25043 \cdot F'_{2i}$$

em que os parâmetros 0,62942 e 0,25043 correspondem, respectivamente, aos percentuais de variância compartilhada pelos dois primeiros fatores, conforme mostra a Figura 10.28.



Para que a variável seja gerada no banco de dados, devemos clicar em **Transform → Compute Variable....** Em **Target Variable**, devemos digitar o nome da nova variável (*ranking*) e, em **Numeric Expression**, devemos digitar a expressão de soma ponderada  $(FAC1\_1 \cdot 0.62942) + (FAC2\_1 \cdot 0.25043)$ , conforme mostra a Figura 10.33. Ao clicarmos em **OK**, a variável *ranking* aparecerá no banco de dados.

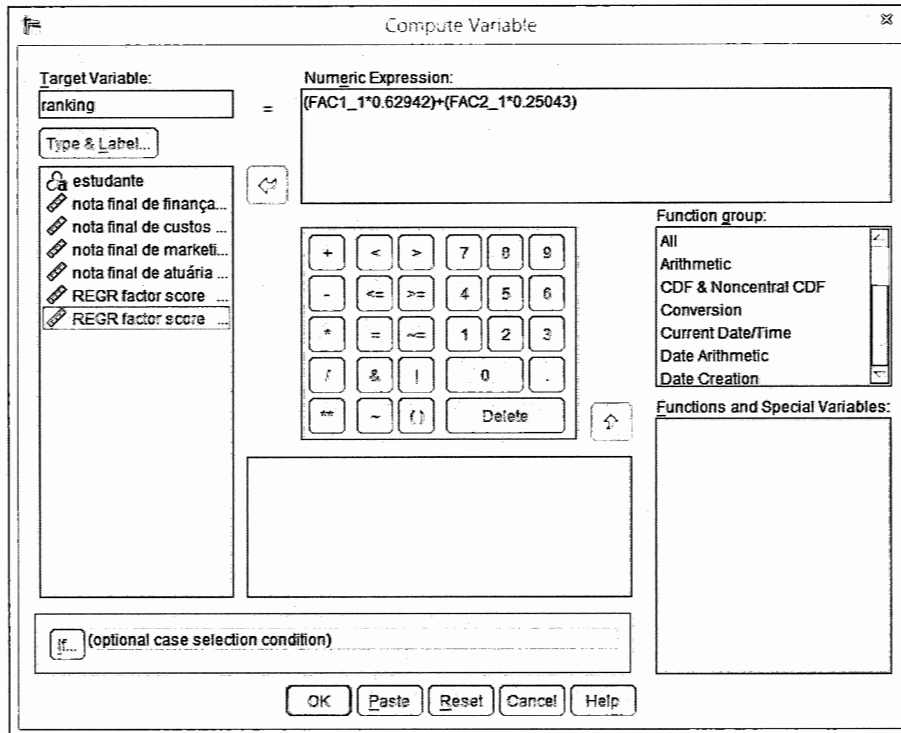


Figura 10.33 Criação de nova variável (*ranking*).

Por fim, para elaborarmos o ordenamento da variável *ranking*, devemos clicar em **Data → Sort Cases....** Além de selecionarmos a opção **Descending**, devemos inserir a variável *ranking* em **Sort by**, conforme mostra a Figura 10.34. Ao clicarmos em **OK**, as observações aparecerão ordenadas no banco de dados, do maior para o menor valor da variável *ranking*, conforme mostra a Figura 10.35 para as 20 observações com melhor desempenho escolar.

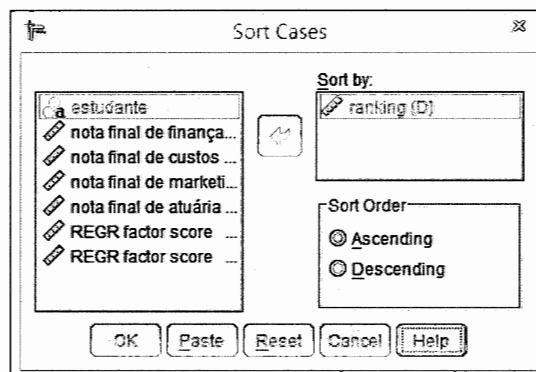
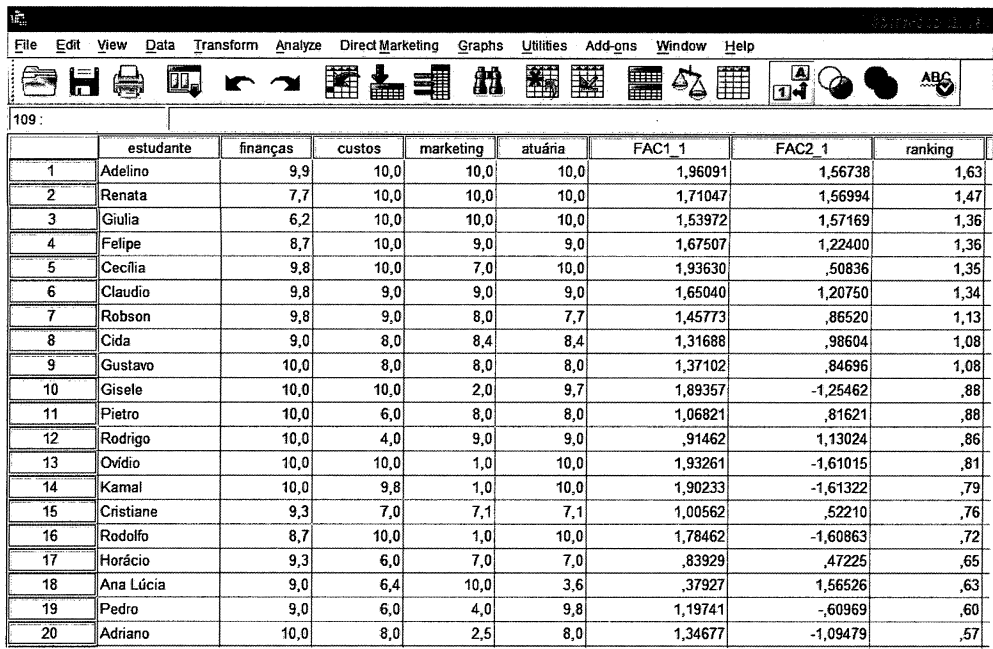


Figura 10.34 Caixa de diálogo para ordenamento das observações pela variável *ranking*.



	estudante	finanças	custos	marketing	atuária	FAC1_1	FAC2_1	ranking
1	Adelino	9,9	10,0	10,0	10,0	1,96091	1,56738	1,63
2	Renata	7,7	10,0	10,0	10,0	1,71047	1,56994	1,47
3	Giulia	6,2	10,0	10,0	10,0	1,53972	1,57169	1,36
4	Felipe	8,7	10,0	9,0	9,0	1,67507	1,22400	1,36
5	Cecília	9,8	10,0	7,0	10,0	1,93630	,50836	1,35
6	Claudio	9,8	9,0	9,0	9,0	1,65040	1,20750	1,34
7	Robson	9,8	9,0	8,0	7,7	1,45773	,86520	1,13
8	Cida	9,0	8,0	8,4	8,4	1,31688	,98604	1,08
9	Gustavo	10,0	8,0	8,0	8,0	1,37102	,84696	1,08
10	Gisele	10,0	10,0	2,0	9,7	1,89357	-1,25462	,88
11	Pietro	10,0	6,0	8,0	8,0	1,06821	,81621	,88
12	Rodrigo	10,0	4,0	9,0	9,0	,91462	1,13024	,86
13	Ovídio	10,0	10,0	1,0	10,0	1,93261	-1,61015	,81
14	Kamal	10,0	9,8	1,0	10,0	1,90233	-1,61322	,79
15	Cristiane	9,3	7,0	7,1	7,1	1,00562	,52210	,76
16	Rodolfo	8,7	10,0	1,0	10,0	1,78462	-1,60863	,72
17	Horácio	9,3	6,0	7,0	7,0	,83929	,47225	,65
18	Ana Lúcia	9,0	6,4	10,0	3,6	,37927	1,56526	,63
19	Pedro	9,0	6,0	4,0	9,8	1,19741	-,60969	,60
20	Adriano	10,0	8,0	2,5	8,0	1,34677	-1,09479	,57

Figura 10.35 Banco de dados com o ranking de desempenho escolar.

Podemos verificar que o *ranking* construído pelo critério da soma ponderada e ordenamento aponta para **Adelino** como o estudante com melhor desempenho escolar no conjunto de disciplinas, seguido por **Renata**, **Giulia**, **Felipe** e **Cecília**.

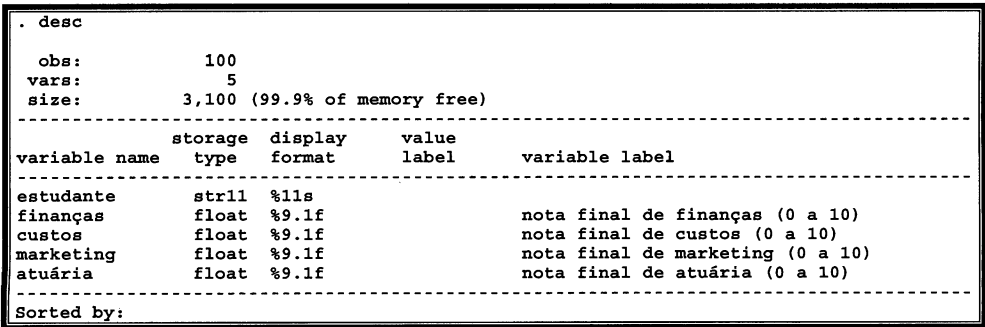
Apresentados os procedimentos para aplicação da análise fatorial por componentes principais no SPSS, partiremos para a elaboração da técnica no Stata, seguindo o padrão adotado no livro.

10.4. ANÁLISE FATORIAL POR COMPONENTES PRINCIPAIS NO SOFTWARE STATA

Apresentaremos agora o passo a passo para a elaboração de nosso exemplo no Stata Statistical Software®. Nosso objetivo, nesta seção, não é discutir novamente os conceitos pertinentes à análise fatorial por componentes principais, porém propiciar ao pesquisador uma oportunidade de elaborar a técnica por meio dos comandos desse software. A cada apresentação de um *output*, faremos menção ao respectivo resultado obtido quando da aplicação da técnica de forma algébrica e também por meio do SPSS. A reprodução das imagens apresentadas nesta seção tem autorização da StataCorp LP®.

Já partiremos, portanto, para o banco de dados construído pelo professor a partir dos questionamentos feitos a cada um dos 100 estudantes. Este banco de dados encontra-se no arquivo **NotasFatorial.dta** e é exatamente igual ao apresentado parcialmente na Tabela 10.5 da seção 10.2.6.

Inicialmente, podemos digitar o comando **desc**, que possibilita a análise das características do banco de dados, como a quantidade de observações, a quantidade de variáveis e a descrição de cada uma delas. A Figura 10.36 apresenta esse primeiro *output* do Stata.



. desc				
obs:	100			
vars:	5			
size:	3,100	(99.9% of memory free)		
-----				
variable name	storage type	display format	value label	variable label
-----				
estudante	str11	%11s		
finanças	float	%9.1f		nota final de finanças (0 a 10)
custos	float	%9.1f		nota final de custos (0 a 10)
marketing	float	%9.1f		nota final de marketing (0 a 10)
atuária	float	%9.1f		nota final de atuária (0 a 10)
-----				
Sorted by:				

Figura 10.36 Descrição do banco de dados **NotasFatorial.dta**.

O comando **pwcorr ... , sig** gera os coeficientes de correlação de Pearson entre cada par de variáveis, com os respectivos níveis de significância. Vamos, portanto, digitar o seguinte comando:

**pwcorr finanças custos marketing atuária, sig**

A Figura 10.37 apresenta o *output* gerado.

pwcorr finanças custos marketing atuária, sig				
	finanças	custos	marketing	atuária
finanças	1.0000			
custos	0.7558 0.0000	1.0000		
marketing	-0.0297 0.7695	0.0031 0.9759	1.0000	
atuária	0.7109 0.0000	0.8091 0.0000	-0.0443 0.6617	1.0000

**Figura 10.37** Coeficientes de correlação de Pearson e respectivos níveis de significância.

Os *outputs* da Figura 10.37 mostram que as correlações entre a variável *marketing* e cada uma das demais variáveis são relativamente baixas e não estatisticamente significantes, ao nível de significância de 5%. Por outro lado, as demais variáveis apresentam, entre si, correlações elevadas e estatisticamente significantes a esse nível de significância, o que representa um primeiro indício de que a análise fatorial poderá agrupá-las em determinado fator, sem que haja perda substancial de suas variâncias, enquanto a variável *marketing* poderá apresentar alta correlação com outro fator. Essa figura está em conformidade com o apresentado na Tabela 10.6 da seção 10.2.6 e também na Figura 10.16, quando da elaboração da técnica no SPSS (seção 10.3).

A adequação global da análise fatorial pode ser avaliada pelos resultados da estatística KMO e do teste de esfericidade de Bartlett, que podem ser obtidos por meio do comando **factortest**. Logo, vamos digitar:

**factortest finanças custos marketing atuária**

Os *outputs* gerados encontram-se na Figura 10.38.

. factortest finanças custos marketing atuária	
Determinant of the correlation matrix	
Det	= 0.137
Bartlett test of sphericity	
Chi-square	= 192.335
Degrees of freedom	= 6
p-value	= 0.000
H0: variables are not intercorrelated	
Kaiser-Meyer-Olkin Measure of Sampling Adequacy	
KMO	= 0.737

**Figura 10.38** Resultados da estatística KMO e do teste de esfericidade de Bartlett.

Com base no resultado da estatística KMO, a adequação global da análise fatorial pode ser considerada **mé-dia**. Porém, mais importante que essa informação é o resultado do teste de esfericidade de Bartlett. A partir do resultado da estatística  $\chi^2_{\text{Bartlett}}$ , podemos afirmar, para o nível de significância de 5% e 6 graus de liberdade, que a matriz de correlações de Pearson é estatisticamente diferente da matriz identidade de mesma dimensão, visto que  $\chi^2_{\text{Bartlett}} = 192,335$  ( $\chi^2$  calculado para 6 graus de liberdade) e  $\text{Prob. } \chi^2_{\text{Bartlett}} (p\text{-value}) < 0,05$ . Note que os resultados dessas estatísticas são condizentes com os calculados algebricamente na seção 10.2.6 e também apresentados na Figura 10.17 da seção 10.3. A Figura 10.38 ainda apresenta o valor do determinante da matriz de correlações, utilizado para o cálculo da estatística  $\chi^2_{\text{Bartlett}}$ .

O Stata ainda permite que sejam obtidos os autovalores e autovetores da matriz de correlações. Para tanto, devemos digitar o seguinte comando:

**pca finanças custos marketing atuária**

A Figura 10.39 apresenta esses autovalores e autovetores, exatamente iguais aos calculados algebricamente na seção 10.2.6. Como ainda não elaboramos o procedimento de rotação dos fatores gerados, podemos verificar que os percentuais de variância compartilhada pelas variáveis originais para a formação de cada fator correspondem aos apresentados na Tabela 10.9.

```
. pca finanças custos marketing atuária
```

Principal components/correlation					Number of obs	=	100
					Number of comp.	=	4
					Trace	=	4
Rotation: (unrotated = principal)					Rho	=	1.0000
Component	Eigenvalue	Difference	Proportion	Cumulative			
Comp1	2.51899	1.51859	0.6297	0.6297			
Comp2	1.0004	.702642	0.2501	0.8798			
Comp3	.297753	.114889	0.0744	0.9543			
Comp4	.182864	.	0.0457	1.0000			

Principal components (eigenvectors)					
Variable	Comp1	Comp2	Comp3	Comp4	Unexplained
finanças	0.5641	0.0068	0.8008	0.2012	0
custos	0.5887	0.0487	-0.2201	-0.7763	0
marketing	-0.0267	0.9987	-0.0003	0.0425	0
atuária	0.5783	-0.0101	-0.5571	0.5959	0

Figura 10.39 Autovalores e autovetores da matriz de correlações.

Apresentados estes primeiros *outputs*, podemos elaborar a análise fatorial por componentes principais propriamente dita, digitando o seguinte comando, cujos resultados são apresentados na Figura 10.40.

**factor finanças custos marketing atuária, pcf**

em que o termo **pcf** se refere ao método de componentes principais (em inglês, *principal-components factor method*).

Enquanto a parte superior da Figura 10.40 apresenta novamente os autovalores da matriz de correlações com os respectivos percentuais de variância compartilhada das variáveis originais, já que o pesquisador pode optar por não fazer uso do comando **pca**, a parte inferior da figura mostra as cargas fatoriais, que representam as correlações entre cada variável e os fatores que apresentam apenas autovalores maiores que 1. Portanto, podemos perceber que o Stata considera, automaticamente, o critério da raiz latente (critério de Kaiser) para a escolha da quantidade de fatores. Se, por alguma razão, o pesquisador optar por extrair uma quantidade de fatores levando em conta um autovalor menor, a fim de que sejam extraídos mais fatores, deverá digitar o termo **mineigen(#)** ao final do comando **factor**, em que # será um número correspondente ao autovalor a partir do qual fatores serão extraídos.

```
. factor finanças custos marketing atuária, pcf
(obs=100)
```

Factor analysis/correlation					Number of obs	=	100
Method: principal-component factors					Retained factors	=	2
Rotation: (unrotated)					Number of params	=	6
Factor	Eigenvalue	Difference	Proportion	Cumulative			
Factor1	2.51899	1.51859	0.6297	0.6297			
Factor2	1.00040	0.70264	0.2501	0.8798			
Factor3	0.29775	0.11489	0.0744	0.9543			
Factor4	0.18286	.	0.0457	1.0000			

LR test: independent vs. saturated: chi2(6) = 194.32 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances			
Variable	Factor1	Factor2	Uniqueness
finanças	0.8953	0.0068	0.1983
custos	0.9343	0.0487	0.1246
marketing	-0.0424	0.9989	0.0003
atuária	0.9179	-0.0101	0.1573

Figura 10.40 Outputs da análise fatorial por componentes principais no Stata.

As cargas fatoriais apresentadas na Figura 10.40 são iguais às das duas primeiras colunas da Tabela 10.11 da seção 10.2.6, e da Figura 10.20 da seção 10.3. Por meio delas, podemos verificar que, enquanto as variáveis *finanças*, *custos* e *atuária* apresentam elevadas correlações com o primeiro fator, a variável *marketing* apresenta forte correlação com o segundo. Além disso, na matriz de cargas fatoriais ainda é apresentada uma coluna chamada **Uniqueness**, ou **exclusividade**, cujos valores representam, para cada variável, o percentual de variância perdida para compor os fatores extraídos, ou seja, corresponde a  $(1 - \text{comunalidade})$  de cada variável. Sendo assim, temos que:

$$\text{uniqueness}_{\text{finanças}} = 1 - [(0,8953)^2 + (0,0068)^2] = 0,1983$$

$$\text{uniqueness}_{\text{custos}} = 1 - [(0,9343)^2 + (0,0487)^2] = 0,1246$$

$$\text{uniqueness}_{\text{marketing}} = 1 - [(-0,0424)^2 + (0,9989)^2] = 0,0003$$

$$\text{uniqueness}_{\text{atuária}} = 1 - [(0,9179)^2 + (-0,0101)^2] = 0,1573$$

Logo, pelo fato de a variável *marketing* apresentar baixas correlações com cada um das demais variáveis originais, acaba por possuir elevada correlação de Pearson com o segundo fator. Isso faz seu valor de *uniqueness* ser muito baixo, visto que seu percentual de variância compartilhada com o segundo fator é quase igual a 100%.

Sabendo que são extraídos dois fatores, vamos, neste momento, partir para a rotação por meio do método Varimax. Para tanto, devemos digitar o seguinte comando:

**rotate, varimax horst**

em que o termo **horst** define o ângulo de rotação a partir das cargas fatoriais padronizadas. Esse procedimento está de acordo com o elaborado algebricamente na seção 10.2.6. Os *outputs* gerados encontram-se na Figura 10.41.

```
. rotate, varimax horst
```

Factor analysis/correlation		Number of obs =	100
Method: principal-component factors		Retained factors =	2
Rotation: orthogonal varimax (Kaiser on)		Number of params =	6

Factor	Variance	Difference	Proportion	Cumulative
Factor1	2.51768	1.51598	0.6294	0.6294
Factor2	1.00170	.	0.2504	0.8798

LR test: independent vs. saturated: chi2(6) = 194.32 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
finanças	0.8951	-0.0195	0.1983
custos	0.9354	0.0213	0.1246
marketing	-0.0131	0.9997	0.0003
atuária	0.9172	-0.0370	0.1573

Factor rotation matrix

	Factor1	Factor2
Factor1	0.9996	-0.0293
Factor2	0.0293	0.9996

**Figura 10.41** Rotação dos fatores pelo método Varimax.

A partir da Figura 10.41, podemos verificar, conforme já discutimos, que o percentual de variância compartilhada por todas as variáveis para a formação dos dois fatores é igual a 87,98%, embora o autovalor de cada fator rotacionado seja diferente do obtido anteriormente. O mesmo pode ser dito em relação aos valores de *uniqueness* de cada variável, mesmo sendo diferentes as cargas fatoriais rotacionadas em relação às correspondentes não rotacionadas, visto que o método Varimax maximiza as cargas de cada variável em determinado fator. A Figura 10.41 ainda mostra, ao final, o ângulo de rotação. Todos esses *outputs* são idênticos aos calculados na seção 10.2.6 e também apresentados quando da elaboração da técnica no SPSS, nas Figuras 10.25, 10.27 e 10.28.

Dessa forma, podemos escrever que:

$$uniqueness_{finanças} = 1 - [(0,8951)^2 + (-0,0195)^2] = 0,1983$$

$$uniqueness_{custos} = 1 - [(0,9354)^2 + (0,0213)^2] = 0,1246$$

$$uniqueness_{marketing} = 1 - [(-0,0131)^2 + (0,9997)^2] = 0,0003$$

$$uniqueness_{atuária} = 1 - [(0,9172)^2 + (-0,0370)^2] = 0,1573$$

e que:

$$(0,8951)^2 + (0,9354)^2 + (-0,0131)^2 + (0,9172)^2 = \lambda_1'^2 = 2,51768$$

$$(-0,0195)^2 + (0,0213)^2 + (0,9997)^2 + (-0,0370)^2 = \lambda_2'^2 = 1,00170$$

Caso o pesquisador deseje, o Stata ainda permite que sejam comparadas, em uma mesma tabela, as cargas fatoriais rotacionadas com aquelas obtidas antes da rotação. Para tanto, é necessário digitar o seguinte comando, após a elaboração da rotação:

**estat rotatecompare**

Os *outputs* gerados encontram-se na Figura 10.42.

```
. estat rotatecompare
```

Rotation matrix -- orthogonal varimax (Kaiser on)

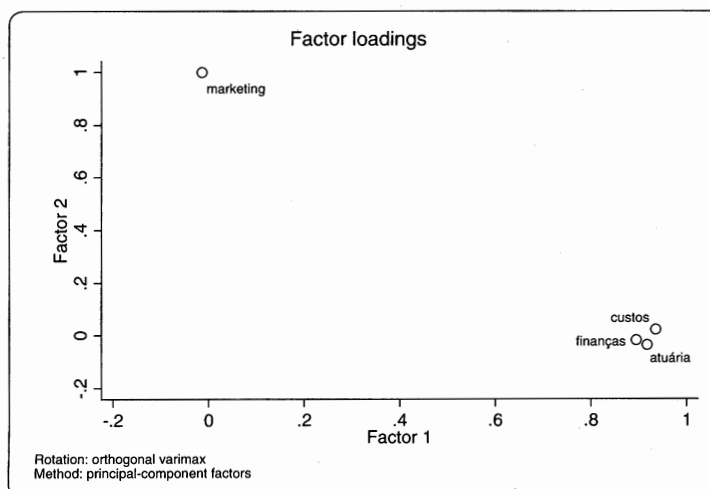
Variable	Factor1	Factor2
Factor1	0.9996	-0.0293
Factor2	0.0293	0.9996

Factor loadings

Variable	Rotated		Unrotated	
	Factor1	Factor2	Factor1	Factor2
finanças	0.8951	-0.0195	0.8953	0.0068
custos	0.9354	0.0213	0.9343	0.0487
marketing	-0.0131	0.9997	-0.0424	0.9989
atuária	0.9172	-0.0370	0.9179	-0.0101

**Figura 10.42** Comparação das cargas fatoriais rotacionadas e não rotacionadas.

O *loading plot* das cargas fatoriais rotacionadas pode ser obtido, neste momento, por meio da digitação do comando **loadingplot**. Esse gráfico, que corresponde aos apresentados nas Figuras 10.9 e 10.26, encontra-se na Figura 10.43.



**Figura 10.43** Loading plot com cargas rotacionadas.

Elaborados esses procedimentos, o pesquisador pode desejar criar duas novas variáveis no banco de dados, correspondentes aos fatores rotacionados obtidos pela análise fatorial. Nesse sentido, é preciso digitar o seguinte comando:

**predict f1 f2**

em que **f1** e **f2** são os nomes das variáveis correspondentes, respectivamente, ao primeiro e ao segundo fatores. Ao digitarmos o comando, além de serem criadas as duas novas variáveis no banco de dados, será também gerado um *output* como o da Figura 10.44, em que são apresentados os *scores* fatoriais rotacionados.

```
. predict f1 f2
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2
finanças	0.35548	-0.00364
custos	0.37219	0.03780
marketing	0.01247	0.99861
atuária	0.36395	-0.02078

**Figura 10.44** Geração dos fatores no banco de dados e *scores* fatoriais rotacionados.

Os resultados apresentados na Figura 10.44 são equivalentes aos do SPSS (Figura 10.29). Além disso, é possível também verificar que os dois fatores gerados são ortogonais, ou seja, apresentam coeficiente de correlação de Pearson igual a 0. Para tanto, vamos digitar:

**estat common**

que fornece o *output* da Figura 10.45.

```
. estat common

Correlation matrix of the varimax rotated common factors
```

Factors	Factor1	Factor2
Factor1	1	
Factor2	0	1

**Figura 10.45** Coeficiente de correlação de Pearson entre os dois fatores rotacionados.

Apenas para fins didáticos, iremos agora obter os *scores* e as cargas fatoriais rotacionados a partir de modelos de regressão linear múltipla. Para tanto, vamos inicialmente gerar, no banco de dados, as variáveis padronizadas por meio do procedimento *Zscores*, a partir de cada uma das variáveis originais, digitando a seguinte sequência de comandos:

```
egen zfinanças = std(finanças)
egen zcustos = std(custos)
egen zmarketing = std(marketing)
egen zatuária = std(atuária)
```

Feito isso, podemos digitar os dois seguintes comandos, que representam dois modelos de regressão linear múltipla, em que cada um deles apresenta determinado fator como variável dependente e as variáveis padronizadas como variáveis explicativas.

```
reg f1 zfinanças zcustos zmarketing zatuária
reg f2 zfinanças zcustos zmarketing zatuária
```

Os resultados desses modelos encontram-se na Figura 10.46.

```
. reg f1 zfinanças zcustos zmarketing zatuária
```

Source	SS	df	MS	Number of obs = 100		
Model	98.9999996	4	24.7499999	F( 4, 95) = .		
Residual	0	95	0	Prob > F = .		
Total	98.9999996	99	.999999996	R-squared = 1.0000		
				Adj R-squared = 1.0000		
				Root MSE = 0		

f1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
zfinanças	.3554795	.	.	.	.
zcustos	.3721907	.	.	.	.
zmarketing	.0124719	.	.	.	.
zatuária	.3639452	.	.	.	.
_cons	1.96e-09	.	.	.	.

```
. reg f2 zfinanças zcustos zmarketing zatuária
```

Source	SS	df	MS	Number of obs = 100		
Model	99.0000001	4	24.75	F( 4, 95) = .		
Residual	0	95	0	Prob > F = .		
Total	99.0000001	99	1	R-squared = 1.0000		
				Adj R-squared = 1.0000		
				Root MSE = 0		

f2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
zfinanças	-.0036389	.	.	.	.
zcustos	.0377955	.	.	.	.
zmarketing	.9986053	.	.	.	.
zatuária	-.020781	.	.	.	.
_cons	9.08e-11	.	.	.	.

Figura 10.46 Outputs dos modelos de regressão linear múltipla com fatores como variáveis dependentes.

Note, a partir da análise da Figura 10.46, que os parâmetros estimados em cada modelo correspondem aos *scores* fatoriais rotacionados para cada variável, de acordo com o já apresentado na Figura 10.44. Assim, como todos os parâmetros do intercepto são praticamente iguais a 0, podemos escrever que:

$$F'_{1i} = 0,3554795 \cdot Z_{finanças_i} + 0,3721907 \cdot Z_{custos_i} + 0,0124719 \cdot Z_{marketing_i} + 0,3639452 \cdot Z_{atuária_i}$$

$$F'_{2i} = -0,0036389 \cdot Z_{finanças_i} + 0,0377955 \cdot Z_{custos_i} + 0,9986053 \cdot Z_{marketing_i} - 0,020781 \cdot Z_{atuária_i}$$

Obviamente, como as quatro variáveis compartilham variâncias para a formação de cada fator, os coeficientes de ajuste  $R^2$  de cada modelo são iguais a 1.

Já para a obtenção das cargas fatoriais rotacionadas, devemos digitar os quatro seguintes comandos, que representam quatro modelos de regressão linear múltipla, em que cada um deles apresenta determinada variável padronizada como variável dependente, e os fatores rotacionados, como variáveis explicativas.

**reg zfinanças f1 f2**

**reg zcustos f1 f2**

**reg zmarketing f1 f2**

**reg zatuária f1 f2**

Os resultados desses modelos encontram-se na Figura 10.47.

Note agora, a partir da análise dessa figura, que os parâmetros estimados em cada modelo correspondem às cargas fatoriais rotacionadas para cada fator, de acordo com o já apresentado na Figura 10.41. Nesse sentido, como todos os parâmetros do intercepto são praticamente iguais a 0, podemos escrever que:

$$Z_{finanças_i} = 0,895146 \cdot F'_{1i} - 0,0194694 \cdot F'_{2i} + u_i, R^2 = 1 - uniqueness = 0,8017$$

$$Z_{custos_i} = 0,935375 \cdot F'_{1i} + 0,0212916 \cdot F'_{2i} + u_i, R^2 = 1 - uniqueness = 0,8754$$

$$Z_{marketing_i} = -0,013053 \cdot F'_{1i} + 0,9997495 \cdot F'_{2i} + u_i, R^2 = 1 - uniqueness = 0,9997$$

$$Z_{atuária_i} = 0,917223 \cdot F'_{1i} - 0,0370175 \cdot F'_{2i} + u_i, R^2 = 1 - uniqueness = 0,8427$$



em que os termos  $u_i$  representam fontes adicionais de variação, além dos fatores  $F'_1$  e  $F'_2$ , para explicar o comportamento de cada variável, visto que outros dois fatores com autovalores menores que 1 também poderiam ter sido extraídos. Os coeficientes de ajuste  $R^2$  de cada modelo diferentes de 1 correspondem aos valores das communalidades de cada variável, ou seja, a  $(1 - \text{uniqueness})$ .

<b>. reg zfinanças f1 f2</b>						
Source	SS	df	MS	Number of obs = 100		
Model	79.3648681	2	39.682434	F( 2, 97) = 196.04		
Residual	19.6351317	97	.202424038	Prob > F = 0.0000		
Total	98.9999997	99	.999999997	R-squared = 0.8017		
				Adj R-squared = 0.7976		
				Root MSE = .44992		
zfinanças	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
f1	.895146	.0452182	19.80	0.000	.8054003	.9848916
f2	-.0194694	.0452182	-0.43	0.668	-.109215	.0702763
_cons	-4.42e-09	.0449916	-0.00	1.000	-.0892958	.0892958
<b>. reg zcustos f1 f2</b>						
Source	SS	df	MS	Number of obs = 100		
Model	86.662589	2	43.3312945	F( 2, 97) = 340.68		
Residual	12.3374069	97	.127189762	Prob > F = 0.0000		
Total	98.9999959	99	.999999958	R-squared = 0.8754		
				Adj R-squared = 0.8728		
				Root MSE = .35664		
zcustos	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
f1	.935375	.0358433	26.10	0.000	.8642359	1.006514
f2	.0212916	.0358433	0.59	0.554	-.0498475	.0924307
_cons	-3.38e-09	.0356637	-0.00	1.000	-.0707825	.0707825
<b>. reg zmarketing f1 f2</b>						
Source	SS	df	MS	Number of obs = 100		
Model	98.9672733	2	49.4836367	F( 2, 97) = .		
Residual	.032725878	97	.00033738	Prob > F = 0.0000		
Total	98.9999992	99	.999999992	R-squared = 0.9997		
				Adj R-squared = 0.9997		
				Root MSE = .01837		
zmarketing	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
f1	-.013053	.001846	-7.07	0.000	-.0167169	-.0093892
f2	.9997495	.001846	541.56	0.000	.9960856	1.003413
_cons	7.10e-11	.0018368	0.00	1.000	-.0036455	.0036455
<b>. reg zatuária f1 f2</b>						
Source	SS	df	MS	Number of obs = 100		
Model	83.4241641	2	41.7120821	F( 2, 97) = 259.77		
Residual	15.5758359	97	.160575627	Prob > F = 0.0000		
Total	99	99	1	R-squared = 0.8427		
				Adj R-squared = 0.8394		
				Root MSE = .40072		
zatuária	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
f1	.917223	.0402738	22.77	0.000	.8372907	.9971553
f2	-.0370175	.0402738	-0.92	0.360	-.1169498	.0429147
_cons	2.40e-09	.0400719	0.00	1.000	-.0795316	.0795316

Figura 10.47 Outputs dos modelos de regressão linear múltipla com variáveis padronizadas como variáveis dependentes.

Embora o pesquisador possa optar por não elaborar os modelos de regressão linear múltipla quando da aplicação da análise fatorial, visto que se trata apenas de procedimento de verificação, acreditamos que seu caráter didático tem fundamental importância para o completo entendimento da técnica.

A partir dos fatores rotacionados extraídos (variáveis  $f1$  e  $f2$ ), podemos definir o desejado *ranking* de desempenho escolar. Assim como elaborado quando da aplicação da técnica no SPSS, faremos uso do critério descrito

na seção 10.2.6, conhecido por critério da soma ponderada e ordenamento, em que uma nova variável é gerada a partir da multiplicação dos valores de cada fator pelos respectivos percentuais de variância compartilhada pelas variáveis originais. Vamos digitar o seguinte comando:

```
gen ranking = f1*0.6294+f2*0.2504
```

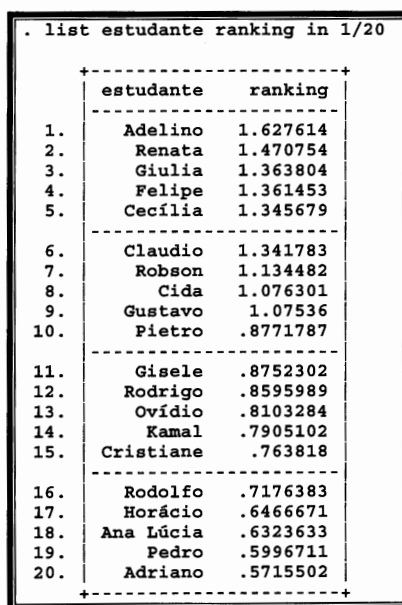
em que os termos 0.6294 e 0.2504 correspondem, respectivamente, aos percentuais de variância compartilhada pelos dois primeiros fatores, conforme mostra a Figura 10.41. A nova variável gerada no banco de dados chama-se *ranking*. Na sequência, podemos ordenar as observações, do maior para o menor valor da variável *ranking*, digitando o seguinte comando:

```
gsort -ranking
```

Na sequência, podemos listar, a título de exemplo, o *ranking* de desempenho escolar dos 20 melhores alunos, com base no comportamento conjunto das notas finais das quatro disciplinas. Para tanto, podemos digitar o seguinte comando:

```
list estudante ranking in 1/20
```

A Figura 10.48 mostra o *ranking* dos 20 estudantes mais bem posicionados.



	estudante	ranking
1.	Adelino	1.627614
2.	Renata	1.470754
3.	Giulia	1.363804
4.	Felipe	1.361453
5.	Cecília	1.345679
6.	Claudio	1.341783
7.	Robson	1.134482
8.	Cida	1.076301
9.	Gustavo	1.07536
10.	Pietro	.8771787
11.	Gisele	.8752302
12.	Rodrigo	.8595989
13.	Ovídio	.8103284
14.	Kamal	.7905102
15.	Cristiane	.763818
16.	Rodolfo	.7176383
17.	Horácio	.6466671
18.	Ana Lúcia	.6323633
19.	Pedro	.5996711
20.	Adriano	.5715502

**Figura 10.48** *Ranking* de desempenho escolar dos 20 melhores estudantes.

## 10.5. CONSIDERAÇÕES FINAIS

Muitas são as situações em que o pesquisador deseja agrupar variáveis em um ou mais fatores, verificar a validade de constructos previamente estabelecidos, criar fatores ortogonais para posterior uso em técnicas multivariadas confirmatórias que necessitam de ausência de multicolinearidade ou elaborar *rankings* por meio da criação de indicadores de desempenho. Nessas situações, os procedimentos relacionados à análise fatorial são bastante indicados, sendo o mais utilizado o conhecido como componentes principais.

A análise fatorial permite, portanto, que sejam aprimorados os processos decisórios com base no comportamento e na relação de interdependência entre variáveis quantitativas que apresentam relativa intensidade de correlação. Como os fatores gerados a partir das variáveis originais também são variáveis quantitativas, os *outputs* da análise fatorial podem servir de *inputs* em outras técnicas multivariadas, como análise de agrupamentos. A própria estratificação de cada fator em faixas pode permitir que seja avaliada a associação entre essas faixas e as categorias de outras variáveis qualitativas, por meio da análise de correspondência.

O uso dos fatores em técnicas multivariadas confirmatórias também pode fazer sentido quando o pesquisador tem a intenção de elaborar diagnósticos sobre o comportamento de determinada variável dependente e utiliza os fatores extraídos como variáveis explicativas, fato que elimina eventuais problemas de multicolinearidade por

serem os fatores ortogonais. A própria consideração de determinada variável qualitativa obtida com base na estratificação em faixas de determinado fator pode ser utilizada, por exemplo, em um modelo de regressão logística multinomial, o que permite a elaboração de um diagnóstico sobre as probabilidades que cada observação tem de pertencer a cada faixa, em função do comportamento de outras variáveis explicativas não inicialmente consideradas na análise fatorial.

Seja qual for o objetivo principal da aplicação da técnica, a análise fatorial pode propiciar a colheita de bons e interessantes frutos de pesquisa úteis à tomada de decisão. Sua elaboração deve ser sempre feita por meio do correto e consciente uso do software escolhido para a modelagem, com base na teoria subjacente e na experiência e intuição do pesquisador.

## 10.6. EXERCÍCIOS

1. A partir de uma base de dados que contém determinadas variáveis dos clientes (pessoas físicas), os analistas do departamento de CRM (*Customer Relationship Management*) de um banco elaboraram uma análise fatorial por componentes principais, com o intuito de estudar o comportamento conjunto dessas variáveis para, na sequência, propor a criação de um indicador de perfil de investimento. As variáveis utilizadas para a elaboração da modelagem foram:

Variável	Descrição
<i>idade</i>	Idade do cliente <i>i</i> (anos).
<i>rfixa</i>	Percentual de recursos aplicado em fundos de renda fixa (%).
<i>rvariável</i>	Percentual de recursos aplicado em fundos de renda variável (%).
<i>pessoas</i>	Quantidade de pessoas que mora na residência.

Em determinado relatório gerencial, os analistas apresentaram as cargas fatoriais (coeficientes de correlação de Pearson) entre cada variável original e os dois fatores extraídos por meio do critério da raiz latente ou critério de Kaiser. Essas cargas fatoriais encontram-se na tabela a seguir:

Variável	Fator 1	Fator 2
<i>idade</i>	0,917	0,047
<i>rfixa</i>	0,874	0,077
<i>rvariável</i>	-0,844	0,197
<i>pessoas</i>	0,031	0,979

Pede-se:

- Quais os autovalores correspondentes aos dois fatores extraídos?
- Quais os percentuais de variância compartilhada por todas as variáveis para a composição de cada fator? Qual o percentual total de variância perdida das quatro variáveis para a extração desses dois fatores?
- Para cada variável, qual o percentual de variância compartilhada para a formação dos dois fatores (comunalidade)?
- Qual a expressão de cada variável padronizada em função dos dois fatores extraídos?
- Elabore o *loading plot* a partir das cargas fatoriais.
- Interprete os dois fatores com base na distribuição das cargas de cada variável.

2. Um estudioso do comportamento de indicadores sociais e econômicos de nações deseja investigar a relação eventualmente existente entre variáveis relacionadas com corrupção, violência, renda e educação, e, para tanto, levantou dados de 50 países, considerados desenvolvidos ou emergentes, em dois anos consecutivos. Os dados encontram-se nos arquivos **IndicadorPaíses.sav** e **IndicadorPaíses.dta**, que apresentam as seguintes variáveis:

Variável	Período	Descrição
<i>país</i>		Variável <i>string</i> que identifica o país <i>i</i> .
<i>cpi1</i>	ano 1	<i>Corruption Perception Index</i> , que corresponde à percepção dos cidadãos em relação ao abuso do setor público sobre os benefícios privados de uma nação, cobrindo aspectos administrativos e políticos. Quanto menor o índice, maior a percepção de corrupção no país (Fonte: Transparência Internacional).
<i>cpi2</i>	ano 2	
<i>violência1</i>	ano 1	Quantidade de assassinatos a cada 100.000 habitantes (Fontes: Organização Mundial da Saúde, Escritório das Nações Unidas para Drogas e Crime e <i>GIMD Global Burden of Injuries</i> ).
<i>violência2</i>	ano 2	
<i>pib_capita1</i>	ano 1	PIB <i>per capita</i> em US\$ ajustado pela inflação, com ano base 2000 (Fonte: Banco Mundial).
<i>pib_capita2</i>	ano 2	
<i>escol1</i>	ano 1	Quantidade média de anos de escolaridade por pessoas com mais de 25 anos, incluindo ensinos primário, secundário e superior (Fonte: <i>Institute for Health Metrics and Evaluation</i> ).
<i>escol2</i>	ano 2	

A fim de que seja criado, para cada ano, um indicador socioeconômico que dê origem a um *ranking* de países, o estudioso decide elaborar uma análise fatorial por componentes principais a partir das variáveis de cada período. Com base nos resultados obtidos, pede-se:

- Por meio da estatística KMO e do teste de esfericidade de Bartlett, é possível afirmar que a análise fatorial por componentes principais é apropriada para cada um dos anos de estudo? No caso do teste de esfericidade de Bartlett, utilize o nível de significância de 5%.
- Quantos fatores são extraídos na análise em cada um dos anos, levando-se em consideração o critério da raiz latente? Qual(is) o(s) autovalor(es) correspondente(s) ao(s) fator(es) extraído(s) em cada ano, bem como o(s) percentual(is) de variância compartilhada por todas as variáveis para a composição desse(s) fator(es)?
- Para cada variável, qual a carga fatorial e o percentual de variância compartilhada para a formação do(s) fator(es) em cada ano? Ocorreram alterações nas comunalidades de cada variável de um ano para o outro?
- Qual(is) a(s) expressão(ões) do(s) fator(es) extraído(s) em cada ano, em função das variáveis padronizadas? De um ano para o outro, ocorreram alterações nos *scores* fatoriais das variáveis em cada fator? Discuta a importância de se elaborar uma análise fatorial específica em cada ano para a criação de indicadores.
- Considerando o fator principal extraído como indicador socioeconômico, elabore o *ranking* dos países a partir desse indicador em cada um dos anos. Houve alterações de um ano para o outro nas posições relativas dos países no *ranking*?

3. O gerente-geral de uma loja pertencente a uma rede de drogarias deseja conhecer a percepção dos consumidores em relação a oito atributos, descritos a seguir:

Atributo (Variável)	Descrição
<i>sortimento</i>	Percepção sobre o sortimento de produtos.
<i>reposição</i>	Percepção sobre a qualidade e rapidez na reposição dos produtos.
<i>layout</i>	Percepção sobre o <i>layout</i> da loja.
<i>conforto</i>	Percepção sobre conforto térmico, acústico e visual na loja.
<i>limpeza</i>	Percepção sobre a limpeza geral da loja.
<i>atendimento</i>	Percepção sobre a qualidade do atendimento prestado.
<i>preço</i>	Percepção sobre o nível de preços praticados em relação à concorrência.
<i>desconto</i>	Percepção sobre política de descontos.

Para tanto, realizou, durante determinado período, uma pesquisa com 1.700 clientes no ponto de venda, cujo questionário foi estruturado por grupo de atributos, e a pergunta correspondente a cada atributo solicitava que o consumidor atribuisse uma nota de 0 a 10 para sua percepção em relação àquele atributo, em que 0 correspondia a uma percepção totalmente negativa, e 10, à melhor percepção possível. Por ter certa experiência, o gerente-geral da loja decidiu de antemão juntar as questões em três grupos, de modo que o questionário completo ficasse de seguinte forma:

<i>Com base em sua percepção, preencha o questionário a seguir com notas de 0 a 10, em que a nota 0 significa que sua percepção é totalmente negativa em relação a determinado atributo, e a nota 10, que sua percepção é a melhor possível.</i>	<i>Nota</i>
<b>Produtos e Ambiente de Loja</b>	
Dê uma nota de 0 a 10 para o sortimento de produtos.	
Dê uma nota de 0 a 10 para a qualidade e rapidez na reposição dos produtos.	
Dê uma nota de 0 a 10 para o layout da loja.	
Dê uma nota de 0 a 10 para o conforto térmico, acústico e visual na loja.	
Dê uma nota de 0 a 10 para a limpeza geral da loja.	
<b>Atendimento</b>	
Dê uma nota de 0 a 10 para a qualidade do atendimento prestado.	
<b>Preços e Política de Descontos</b>	
Dê uma nota de 0 a 10 para o nível de preços praticados em relação à concorrência.	
Dê uma nota de 0 a 10 para a política de descontos.	

O banco de dados completo elaborado pelo gerente-geral da loja encontra-se nos arquivos **Percepção-Drogaria.sav** e **PercepçãoDrogaria.dta**. Pede-se:

- Apresente a matriz de correlações entre cada par de variáveis. Com base na magnitude dos valores dos coeficientes de correlação de Pearson, é possível identificar um primeiro indício de que a análise fatorial poderá agrupar as variáveis em fatores?
- Por meio do resultado do teste de esfericidade de Bartlett, é possível afirmar, ao nível de significância de 5%, que a análise fatorial por componentes principais é apropriada?
- Quantos fatores são extraídos na análise, levando-se em consideração o critério da raiz latente? Qual(is) o(s) autovalor(es) correspondente(s) ao(s) fator(es) extraído(s), bem como o(s) percentual(is) de variância compartilhada por todas as variáveis para a composição desse(s) fator(es)?
- Qual o percentual total de perda de variância das variáveis originais resultante da extração do(s) fator(es) com base no critério da raiz latente?
- Para cada variável, qual a carga e o percentual de variância compartilhada para a formação do(s) fator(es)?
- Com a imposição da extração de três fatores, em detrimento do critério da raiz latente, e com base nas novas cargas fatoriais, é possível confirmar o constructo do questionário proposto pelo gerente-geral da loja? Em outras palavras, as variáveis de cada grupo do questionário acabam, de fato, por apresentar maior compartilhamento de variância com um fator comum?
- Discuta o impacto da decisão de extração de três fatores sobre os valores das communalidades?
- Elabore uma rotação Varimax e discuta novamente, com base na redistribuição das cargas fatoriais, o constructo inicialmente proposto no questionário pelo gerente-geral da loja.
- Apresente o *loading plot* 3D com as cargas fatoriais rotacionadas.

## APÊNDICE

## Alpha de Cronbach

## A. Breve Apresentação

A estatística **alpha**, proposta por Cronbach (1951), é uma medida utilizada para se avaliar a **consistência interna** das variáveis de um banco de dados, ou seja, é uma medida do **grau de confiabilidade** (*reliability*) com a qual determinada escala, adotada para a definição das variáveis originais, produz resultados consistentes sobre a relação dessas variáveis. Segundo Nunnally e Bernstein (1994), o grau de confiabilidade é definido a partir do comportamento das correlações entre as variáveis originais (ou padronizadas), e, portanto, o alpha de Cronbach pode ser utilizado para se avaliar a fidedignidade com a qual um fator pode ser extraído a partir dessas variáveis, sendo, assim, relacionado com a análise fatorial.

Segundo Rogers, Schmitt e Mullins (2002), embora o alpha de Cronbach não seja a única medida de confiabilidade existente, visto que apresenta restrições relacionadas com a multidimensionalidade, ou seja, com a identificação de múltiplos fatores, pode ser definido como a medida que possibilita avaliar a intensidade com a qual determinado constructo ou fator está presente nas variáveis originais. Dessa forma, um banco de dados com variáveis que compartilhem um único fator tende a apresentar elevado alpha de Cronbach.

Nesse sentido, o alpha de Cronbach não pode ser utilizado para a avaliação da adequação global da análise fatorial, ao contrário da estatística KMO e do teste de esfericidade de Bartlett, visto que sua magnitude oferece ao pesquisador indícios apenas sobre a consistência interna da escala utilizada para a extração de um único fator. Caso seu valor seja baixo, sequer o primeiro fator poderá ser adequadamente extraído, principal razão por que alguns pesquisadores optam por estudar a magnitude do alpha de Cronbach antes da elaboração da análise fatorial, embora essa decisão não represente um requisito obrigatório para a elaboração da técnica.

O alpha de Cronbach pode ser definido por meio da seguinte expressão:

$$\alpha = \frac{k}{k-1} \cdot \left[ 1 - \frac{\sum \text{var}_k}{\text{var}_{\text{soma}}} \right] \quad (10.41)$$

em que:

$\text{var}_k$  é a variância da  $k$ -ésima variável, e

$$\text{var}_{\text{soma}} = \frac{\sum_{i=1}^n \left( \sum_k X_{ki} \right)^2 - \frac{\left( \sum_{i=1}^n \sum_k X_{ki} \right)^2}{n}}{n-1} \quad (10.42)$$

que representa a variância da soma de cada linha do banco de dados, ou seja, a variância da soma dos valores correspondentes a cada observação. Além disso, sabemos que  $n$  é o tamanho da amostra, e  $k$ , o número de variáveis  $X$ .

Logo, podemos afirmar que, se ocorrerem consistências nos valores das variáveis, o termo  $\text{var}_{\text{soma}}$  será grande o suficiente para que alpha ( $\alpha$ ) tenda a 1. Por outro lado, variáveis que apresentam correlações baixas, possivelmente decorrentes da presença de valores aleatórios nas observações, farão o termo  $\text{var}_{\text{soma}}$  regredir à soma das variâncias de cada variável ( $\text{var}_k$ ), o que fará alpha ( $\alpha$ ) tender a 0.

Embora não haja um consenso na literatura sobre o valor de  $\alpha$  a partir do qual exista consistência interna das variáveis do banco de dados, é interessante que o resultado obtido seja maior que 0,6 quando da aplicação de técnicas exploratórias.

Na sequência, apresentaremos o cálculo do  $\alpha$  de Cronbach para os dados do exemplo utilizado ao longo do capítulo.

## B. Determinação Algébrica do Alpha de Cronbach

A partir das variáveis padronizadas do exemplo estudado ao longo do capítulo, podemos elaborar a Tabela 10.19, que nos ajuda para o cálculo do  $\alpha$  de Cronbach.

**Tabela 10.19** Procedimento para cálculo do  $\alpha$  de Cronbach.

Estudante	$Z_{finanças_i}$	$Z_{custos_i}$	$Z_{marketing_i}$	$Z_{atuária_i}$	$\sum_{k=4} X_{ki}$	$\left(\sum_{k=4} X_{ki}\right)^2$
Gabriela	-0,011	-0,290	-1,650	0,273	-1,679	2,817
Luiz Felipe	-0,876	-0,697	1,532	-1,319	-1,360	1,849
Patrícia	-0,876	-0,290	-0,590	-0,523	-2,278	5,191
Gustavo	1,334	1,337	0,825	1,069	4,564	20,832
Letícia	-0,779	-1,104	-0,872	-0,841	-3,597	12,939
Ovídio	1,334	2,150	-1,650	1,865	3,699	13,682
Leonor	-0,267	0,116	0,825	-0,125	0,549	0,301
Dalila	-0,139	0,523	0,118	0,273	0,775	0,600
Antônio	0,021	-0,290	-0,590	-0,523	-1,382	1,909
...						
Estela	0,982	0,113	-1,297	1,069	0,868	0,753
<b>Variância</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>	$\left(\sum_{i=1}^{100} \sum_{k=4} X_{ki}\right)^2 = 0$	$\sum_{i=1}^{100} \left(\sum_{k=4} X_{ki}\right)^2 = 832,570$

Logo, com base na expressão (10.42), temos que:

$$\text{var}_{\text{soma}} = \frac{832,570}{99} = 8,410$$

e, fazendo uso da expressão (10.41), podemos calcular o  $\alpha$  de Cronbach:

$$\alpha = \frac{4}{3} \cdot \left[ 1 - \frac{4}{8,410} \right] = 0,699$$

Podemos considerar esse valor aceitável para a consistência interna das variáveis de nosso banco de dados. Entretanto, conforme veremos quando da determinação do  $\alpha$  de Cronbach no SPSS e no Stata, existe perda considerável de confiabilidade pelo fato de as variáveis originais não estarem medindo o mesmo fator, ou seja, a mesma dimensão, visto que esta estatística apresenta restrições relacionadas com a multidimensionalidade. Ou seja, caso não incluíssemos a variável *marketing* no cálculo do  $\alpha$  de Cronbach, seu valor seria consideravelmente maior, o que indica que essa variável não contribui para o constructo, ou para o primeiro fator, formado pelas demais variáveis (*finanças*, *custos* e *atuária*).

A planilha completa com o cálculo do  $\alpha$  de Cronbach pode ser acessada por meio do arquivo **AlphaCronbach.xls**.

De maneira análoga ao realizado ao longo do capítulo, apresentaremos, na sequência, os procedimentos para obtenção do  $\alpha$  de Cronbach no SPSS e no Stata.

## C. Determinação do Alpha de Cronbach no SPSS

Vamos novamente fazer uso do arquivo **NotasFatorial.sav**.

Para que possamos determinar o  $\alpha$  de Cronbach com base nas variáveis padronizadas, devemos inicialmente padronizá-las pelo procedimento *Zscores*. Para tanto, vamos clicar em **Analyze** → **Descriptive Statistics** → **Descriptives....** Ao selecionarmos todas as variáveis originais, devemos clicar em **Save standardized values**

as **variables**. Embora esse procedimento específico não seja mostrado aqui, após clicarmos em **OK**, as variáveis padronizadas serão geradas no próprio banco de dados.

Na sequência, vamos clicar em **Analyze → Scale → Reliability Analysis...** Uma caixa de diálogo será aberta. Devemos inserir as variáveis padronizadas em **Items**, conforme mostra a Figura 10.49.

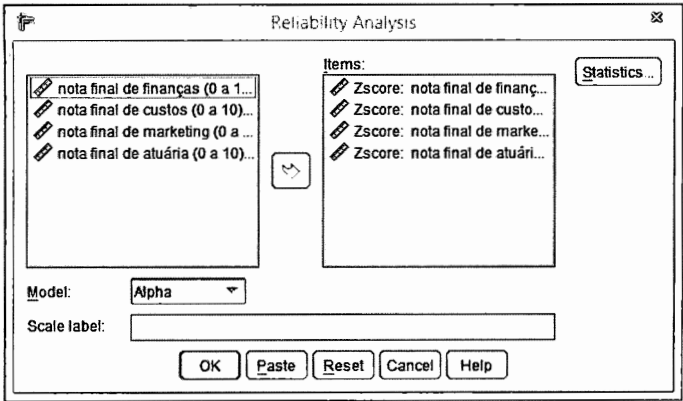


Figura 10.49 Caixa de diálogo para determinação do alpha de Cronbach no SPSS.

Na sequência, em **Statistics...**, devemos marcar a opção **Scale if item deleted**, conforme mostra a Figura 10.50. Essa opção faz com que sejam calculados os diferentes valores de alpha de Cronbach quando se elimina cada variável da análise. O termo **item** é bastante referenciado no trabalho de Cronbach (1951) e utilizado como sinônimo de **variável**.

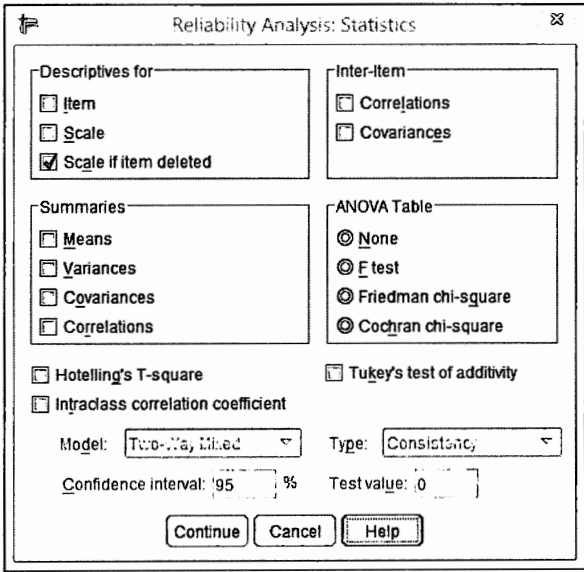


Figura 10.50 Seleção da opção para cálculo do alpha ao se excluir determinada variável.

Em seguida, podemos clicar em **Continue** e em **OK**. A Figura 10.51 apresenta o resultado do alpha de Cronbach, cujo valor é exatamente igual ao calculado por meio das expressões (10.41) e (10.42) e mostrado na seção anterior.

Reliability Statistics	
Cronbach's Alpha	N of Items
,699	4

Figura 10.51 Resultado do alpha de Cronbach no SPSS.

Além disso, a Figura 10.52 ainda apresenta na última coluna os valores que seriam obtidos do alpha de Cronbach, caso determinada variável fosse excluída da análise. Assim, podemos verificar que a presença da variável



*marketing* contribui negativamente para a identificação de apenas um fator, pois, conforme sabemos, essa variável apresenta forte correlação com o segundo fator extraído pela análise de componentes principais elaborada ao longo do capítulo. Como o alpha de Cronbach é uma medida de confiabilidade unidimensional, a exclusão da variável *marketing* faria seu valor chegar a 0,904.

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Zscore: nota final de finanças (0 a 10)	,0000000	4,536	,675	,508
Zscore: nota final de custos (0 a 10)	,0000000	4,274	,758	,447
Zscore: nota final de marketing (0 a 10)	,0000000	7,552	-,026	,904
Zscore: nota final de atuária (0 a 10)	,0000000	4,458	,699	,491

**Figura 10.52** Alpha de Cronbach quando da exclusão de cada variável.

Na sequência, obteremos os mesmos *outputs* por meio da aplicação de comandos específicos no Stata.

#### D. Determinação do Alpha de Cronbach no Stata

Vamos agora abrir o arquivo **NotasFatorial.dta**.

A fim de que seja calculado o alpha de Cronbach, devemos digitar o seguinte comando:

**alpha finanças custos marketing atuária, asis std**

em que o termo **std** faz com que seja calculado o alpha de Cronbach a partir das variáveis padronizadas, mesmo que tenham sido consideradas as variáveis originais no comando **alpha**.

O *output* gerado encontra-se na Figura 10.53.

```
. alpha finanças custos marketing atuária, asis std
Test scale = mean(standardized items)
Average interitem correlation:      0.3675
Number of items in the scale:      4
Scale reliability coefficient:      0.6992
```

**Figura 10.53** Resultado do alpha de Cronbach no Stata.

Caso o pesquisador opte por obter os valores do alpha de Cronbach quando da exclusão de cada uma das variáveis, assim como realizado no SPSS, poderá digitar o seguinte comando:

**alpha finanças custos marketing atuária, asis std item**

Os novos *outputs* são apresentados na Figura 10.54, em que os valores da última coluna são exatamente iguais aos apresentados na Figura 10.52, o que corrobora o fato de que as variáveis *finanças*, *custos* e *atuária* apresentam elevada consistência interna para a determinação de um único fator.

. alpha finanças custos marketing atuária, asis std item						
Test scale = mean(standardized items)						
Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem correlation	alpha
finanças	100	+	0.8404	0.6748	0.2559	0.5079
custos	100	+	0.8855	0.7585	0.2123	0.4471
marketing	100	+	0.3204	-0.0258	0.7586	0.9041
atuária	100	+	0.8537	0.6989	0.2431	0.4907
Test scale					0.3675	0.6992

**Figura 10.54** Consistência interna ao se excluir cada variável – Última coluna.