Luiz Paulo Fávero
Patrícia Belfiore
Fabiana Lopes da Silva
Betty Lilian Chan

# análise de dados

## MODELAGEM MULTIVARIADA PARA TOMADA DE DECISÕES



CAMPUS

CAPÍTULO 7

# Análise Fatorial

A classificação é uma atividade conceitual básica dos seres humanos.

ELISABETH REIS (2001)

***AO FINAL DESTE CAPÍTULO, VOCÊ SERÁ CAPAZ DE:***

- Estabelecer as circunstâncias a partir das quais uma análise fatorial pode ser aplicada.
- Entender as premissas subjacentes à utilização da análise fatorial.
- Diferenciar a análise fatorial das outras técnicas multivariadas.
- Entender os principais métodos rotacionais.
- Determinar o número de fatores a serem extraídos, bem como nomeá-los.
- Entender os conceitos de cargas fatoriais e escores fatoriais.
- Entender os conceitos de *eigenvalues* (autovalores) e comunalidades.
- Saber estratificar observações a partir de escores fatoriais.

## 7.1. APRESENTAÇÃO DO CAPÍTULO

A análise fatorial, ou análise do fator comum, é uma técnica multivariada de interdependência que busca sintetizar as relações observadas entre um conjunto de variáveis inter-relacionadas, buscando identificar fatores comuns. A idéia básica reside na premissa de que é possível representar um conjunto de variáveis originais observadas por meio de um número menor de fatores intrínsecos.

Neste sentido, a maior vantagem da análise fatorial é permitir a simplificação ou a redução de um grande número de dados, por intermédio da determinação das dimensões latentes, também conhecidas por fatores. E, como consequência, possibilitar ao pesquisador a criação de indicadores inicialmente não observáveis compostos do agrupamento de variáveis.

As possibilidades de utilização da análise fatorial (AF) são inúmeras. Um pesquisador pode estar interessado em como se comportam e se relacionam variáveis econômico-financeiras de uma amostra de empresas do setor petroquímico. Um formulador de políticas públicas pode desejar criar um indicador sociodemográfico de distritos municipais a partir de variáveis relacionadas a renda, taxa de desemprego, taxa de mortalidade infantil, índice de criminalidade e nível de escolaridade. Um estudante de psicologia pode investigar a relação entre diversas variáveis que medem alguns aspectos do comportamento de adolescentes e criar fatores psicológicos. Por fim, um executivo de recursos humanos pode estar interessado em criar um indicador de desempenho dos funcionários de sua empresa em função de características de produtividade, pontualidade, assiduidade e relacionamento interpessoal. Todos esses problemas podem ser resolvidos por meio da aplicação da análise fatorial.

Os três maiores objetivos deste capítulo são: (1) introduzir a natureza, a filosofia e as condições da análise fatorial; (2) apresentar a aplicação da técnica; e (3) discutir os resultados obtidos.

## 7.2. UMA INTRODUÇÃO À ANÁLISE FATORIAL

A análise fatorial (AF) é uma técnica multivariada que busca identificar um número relativamente pequeno de fatores comuns que podem ser utilizados para representar relações entre um grande número de variáveis inter-relacionadas.

De acordo com Maroco (2007), a análise fatorial (AF) é uma técnica de análise exploratória de dados que tem por objetivo descobrir e analisar a estrutura de um conjunto de variáveis inter-relacionadas, de modo a construir uma escala de medida para fatores (intrínsecos) que, de alguma forma (mais ou menos explícita), controla as variáveis originais.

A partir das correlações observadas entre as variáveis originais, a AF estima os fatores comuns que são subjacentes às variáveis e não diretamente observáveis.

O maior objetivo da análise fatorial é permitir a simplificação ou redução de um grande número de variáveis por meio da determinação das dimensões latentes comuns (fatores). Neste sentido, a técnica transforma um conjunto de variáveis correlacionadas em outro grupo que pode ser não correlacionado, de maneira a reduzir a complexidade e facilitar a interpretação dos dados. Esta técnica busca verificar quantos fatores há no modelo e o que eles representam, embora nomeá-los não seja uma tarefa objetiva.

Segundo Johnson e Wichern (2007), na análise do fator comum, as variáveis são agrupadas em função de suas correlações. Isso significa que as variáveis que compõem determinado fator devem ser altamente correlacionadas entre si e fracamente correlacionadas com as variáveis que entram na composição de outro fator qualquer. Portanto, para facilitar a interpretação, a análise fatorial pode ser entendida como uma técnica de agrupamento de variáveis ou colunas de um banco de dados.

Maroco (2007) destaca que o objetivo primordial da AF é atribuir um escore (quantificação) a *constructos*, ou fatores, que não são diretamente observáveis. Este novo escore é uma representação parcimoniosa da informação presente nas diferentes variáveis e é capaz de resumir a informação presente em muitas variáveis em um número reduzido de fatores não diretamente observáveis. Esses fatores permitem identificar as relações estruturais entre as variáveis que, de outra forma, passariam despercebidas no conjunto vasto de variáveis originais.

De acordo com Hair, Anderson, Tatham e Black (2005), um fator "representa a combinação linear (variável estatística) das variáveis originais. Os fatores também representam as dimensões latentes (*constructos*) que resumem ou explicam o conjunto original de variáveis observadas".

As suposições em análise fatorial, de acordo com Pestana e Gageiro (2005), Hair, Anderson, Tatham e Black (2005) e Ho (2006), são:

- **Normalidade e linearidade**: desvios na normalidade e na linearidade podem reduzir as correlações observadas entre as variáveis e, portanto, prejudicar a solução.
- **Matriz de correlações com valores significativos**: o pesquisador deve garantir que a matriz de correlações apresente valores altos o suficiente para justificar a aplicação da análise fatorial. Se a inspeção visual da matriz de correlações não indicar um número substancial de valores superiores a 0,30, sua utilização provavelmente não será apropriada.



Quando da aplicação da AF, assim como em outras técnicas que serão abordadas ao longo deste livro, o pesquisador deverá ter certa preocupação com a premissa da distribuição normal multivariada das variáveis. A normalidade multivariada pode ser testada por meio de uma rotina do *software* SAS, que está disponível na página virtual do livro na Web (www.campus.com.br).

Além disso, antes da utilização da análise fatorial, deve-se identificar a existência de *outliers* e se a distribuição dos dados é viesada. Esses dois fenômenos podem distorcer os resultados, uma vez que alteram as estimativas das médias e dos desvios padrão e, assim, afetam as estimativas das covariâncias e das correlações.

Sobre o tamanho da amostra, Hair, Anderson, Tatham e Black (2005) afirmam que dificilmente se aplica uma análise fatorial com uma amostra de 50 observações, e que, de preferência, o tamanho da amostra deve ser igual ou superior a 100 observações. Como regra geral, utiliza-se um mínimo de 5 vezes mais observações do que o número de variáveis que compõem o banco de dados, sendo recomendável, todavia, que este coeficiente seja de 10 observações para cada variável. Porém, ressalta-se que o dimensionamento amostral, conforme discutido no Capítulo 5 (Inferência Estatística), resguarda o pesquisador quanto ao número de observações para a aplicação das técnicas multivariadas.

A análise fatorial pode ser exploratória ou confirmatória. A análise fatorial confirmatória é frequentemente estudada como um caso particular da técnica de equações estruturais, já que o pesquisador possui algum conhecimento prévio de como as variáveis se comportam e se relacionam e, desta forma, assume que a estrutura dos fatores é conhecida.

Na análise fatorial exploratória, por outro lado, o pesquisador tem pouco ou nenhum conhecimento prévio acerca da estrutura dos fatores. Neste capítulo, abordaremos apenas a análise fatorial exploratória.

Basicamente, a AF pode ser dividida nas seguintes etapas:

- Análise da matriz de correlações e adequação da utilização da AF.
- Extração dos fatores iniciais e determinação do número de fatores.
- Rotação dos fatores.
- Interpretação dos fatores.

Para fins didáticos, os tópicos a serem abordados no presente capítulo seguirão as etapas descritas anteriormente, a fim de facilitar a visualização e o acompanhamento lógico da técnica por parte do pesquisador. Entretanto, faz-se necessário, primeiramente, o desenvolvimento do modelo matemático e lógico da AF, que será apresentado na seção seguinte.

## 7.3. MODELAGEM DA ANÁLISE FATORIAL

O desenvolvimento da AF resulta do trabalho de Charles Spearman, que, em 1904, publicou o texto intitulado "*General intelligence, objectively determined and measured*". Segundo Manly (2004), enquanto estudava a correlação entre a pontuação obtida pelos alunos em vários testes, Spearman notou que muitas correlações observadas poderiam ser quantificadas por meio de um modelo mais simples. Assim, criou uma hipótese de que os desempenhos dos alunos em várias disciplinas são inter-relacionados, e essas inter-relações podem ser explicadas pelo nível de inteligência geral dos estudantes.

A tabela a seguir mostra uma matriz de correlações dos resultados de testes de inteligência, em um dos casos analisados no trabalho de Spearman (1904).

De acordo com Manly (2004), Spearman notou que a matriz possuía uma propriedade interessante: a razão entre duas linhas quaisquer é aproximadamente proporcional, se desconsiderada a diagonal principal da matriz.

**Tabela 7.1:** Matriz de correlações entre os resultados dos testes de inteligência

| | Clássicos | Francês | Inglês | Matemática | Discriminação de tom | Música |
|---|---|---|---|---|---|---|
| Clássicos | 1,00 | 0,83 | 0,78 | 0,70 | 0,66 | 0,63 |
| Francês | 0,83 | 1,00 | 0,67 | 0,67 | 0,65 | 0,57 |
| Inglês | 0,78 | 0,67 | 1,00 | 0,64 | 0,54 | 0,51 |
| Matemática | 0,70 | 0,67 | 0,64 | 1,00 | 0,45 | 0,51 |
| Discriminação de tom | 0,66 | 0,65 | 0,54 | 0,45 | 1,00 | 0,40 |
| Música | 0,63 | 0,57 | 0,51 | 0,51 | 0,40 | 1,00 |

*Fonte:* Manly (2004).

Por exemplo, para as linhas *Clássicos* e *Inglês*, tem-se:

$$\frac{0{,}83}{0{,}67} \approx \frac{0{,}70}{0{,}64} \approx \frac{0{,}66}{0{,}54} \approx \frac{0{,}63}{0{,}51} \approx 1{,}2$$

Assim, Spearman sugeriu que cada um dos seis testes de inteligência (variáveis) pudesse ser descrito pela seguinte expressão:

$$X_i = a_i F + \varepsilon_i \tag{7.1}$$

Sendo que:

- $X_i$ é o *i*-ésimo escore da variável analisada (no caso, os escores dos testes *Clássicos, Francês, Inglês, Matemática, Discriminação de tom* e *Música*) depois de efetuada a padronização para a obtenção de médias iguais a zero e desvios padrão iguais a 1 (Z scores);
- $F$ é o fator aleatório comum (inteligência geral) para todas as variáveis medidas;
- $\varepsilon_i$ é um componente aleatório específico para cada teste de inteligência;
- $a_i$ é a constante chamada de **carga fatorial (*loading*)**, que mede a importância dos fatores na composição de cada variável (correlação).

De acordo com Manly (2004), a variância de $X_i$ é dada por:

$$\begin{aligned} Var(X_i) &= Var(a_i F + \varepsilon_i) \\ &= Var(a_i F) + Var(\varepsilon_i) \\ &= a_i^2\, Var(F) + Var(\varepsilon_i) \\ &= a_i^2\, Var(\varepsilon_i) \end{aligned} \tag{7.2}$$

Em que $a_i$ é uma constante, $F$ e $\varepsilon_i$ são independentes e a variância de $F$ é igual a 1. Como $Var(X_i) = 1$, tem-se:

$$1 = a_i^2 + Var(\varepsilon_i) \tag{7.3}$$

O quadrado da $a_i$ (carga fatorial) representa a proporção da variância de $X_i$, que é explicada pelo fator comum e também é denominada de **comunalidade.**

Segundo Maroco (2007), "Spearman defendia que a performance de uma criança em um teste qualquer podia ser obtida pela soma de um fator geral $F$ com uma habilidade específica $\varepsilon_i$".

Generalizando a proposta de Spearman (1904), tem-se que o modelo de análise fatorial considera que as $p$ variáveis observáveis ($X_1, X_2, X_3, \ldots, X_p$), extraídas de uma população com vetor de média $\mu$ e matriz de covariância $\Sigma$, são linearmente dependentes de algumas variáveis não observáveis $F_1, F_2, F_3, \ldots, F_m$, denominadas de fatores comuns, e de $p$ fontes adicionais de variação $\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots, \varepsilon_p$, denominadas de erros ou fatores específicos (JOHNSON e WICHERN, 2007).



O modelo de análise fatorial é, portanto, apresentado da seguinte forma:

$$\begin{aligned} X_1 &= \mu_1 + a_{11}F_1 + a_{12}F_2 + \ldots + a_{1m}F_m + \varepsilon_1 \\ X_2 &= \mu_2 + a_{21}F_1 + a_{22}F_2 + \ldots + a_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p &= \mu_p + a_{p1}F_1 + a_{p2}F_2 + \ldots + a_{pm}F_m + \varepsilon_p \end{aligned} \tag{7.4}$$

Conforme já apontado anteriormente, o coeficiente $a_{ij}$ é denominado de *loading* ou carga fatorial, e representa o peso da variável $i$ no fator $j$, ou *seja*, o grau de correlação entre as variáveis originais e os fatores.

Efetuando a padronização de $X$ (média 0 e desvio padrão 1), o modelo fatorial passa a ser escrito, genericamente, da seguinte forma:

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \ldots a_{im}F_m + \varepsilon_i \quad (i = 1, \ldots, p) \tag{7.5}$$

Neste caso, $X_i$ representa as variáveis padronizadas, $a_i$ as cargas fatoriais, $F_m$ os fatores comuns e $\varepsilon_i$ os fatores específicos. Entretanto, de acordo com Maroco (2007), o modelo anterior assume as seguintes premissas:

1. Os fatores comuns ($F_k$) são independentes (ortogonais) e igualmente distribuídos, com média 0 e variância 1 ($k = 1, \ldots, m$).
2. Os fatores específicos ($\varepsilon_i$) são independentes e igualmente distribuídos, com média 0 e variância $\psi_i$ ($i = 1, \ldots, p$).
3. $F_k$ e $\varepsilon_i$ são independentes.

O termo $\psi_i$ representa a variância de $\varepsilon_i$, ou seja, $Var(\varepsilon_i) = \psi_i$.

Se as três premissas anteriores forem verificadas, estaremos diante de um modelo fatorial ortogonal. Caso contrário, se $F_k$ e $\varepsilon_i$ estiverem correlacionados, o modelo fatorial será oblíquo.

Já os fatores podem ser estimados por combinação linear das variáveis, da seguinte forma:

$$\begin{aligned} F_1 &= d_{11}X_1 + d_{12}X_2 + \ldots + d_{1mi}X_i \\ F_2 &= d_{21}X_1 + d_{22}X_2 + \ldots + d_{2mi}X_i \\ &\vdots \\ F_m &= d_{m1}X_1 + d_{m2}X_2 + \ldots + d_{mi}X_i \end{aligned} \tag{7.6}$$

Sendo $F_m$ os fatores comuns, $d_{mi}$ os coeficientes dos escores fatoriais e $X_i$ as variáveis originais.

O **escore fatorial** resulta da multiplicação dos coeficientes $d_{mi}$ pelo valor das variáveis originais. Na existência de mais de um fator, o escore fatorial corresponderá às coordenadas da variável em relação aos eixos (fatores).

Retornando à Expressão (7.5), sua variância será dada por:

$$\begin{aligned} Var(X_i) &= Var(a_{i1}F_1 + a_{i2}F_2 + \ldots + a_{im}F_m + \varepsilon_i) = 1 \\ &= 1 = a_{i1}^2 Var(F_1) + a_{i2}^2 Var(F_2) + \ldots + a_{im}^2 Var(F_m) + \psi_i \\ &= a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2 + \psi_i \end{aligned} \tag{7.7}$$

Portanto, observa-se que a variância de $X_i$ pode ser decomposta em duas partes:

$$Var(X_i) = \underbrace{a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2}_{comunalidade} + \underbrace{\psi_i}_{variância\ específica} \quad (7.8)$$

Sendo:

$$h_i^2 = a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2 \quad (7.9)$$

A **comunalidade** ($h_i^2 = a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2$) representa uma estimativa da variância de $X_i$ que é explicada pelos fatores comuns; e $\psi_i$ é chamada de especificidade de $X_i$, pois não está ligada ao fator comum. Portanto, a comunalidade é um índice da variabilidade total explicada por todos os fatores para cada variável. Assim:

$$Var(X_i) = h_i^2 + \psi_i \qquad (i = 1, 2, \ldots, p) \quad (7.10)$$

Em notação matricial, o modelo fatorial apresentado por meio da Expressão (7.4) pode ser expresso por:

$$X - \mu = \Lambda F + \varepsilon \quad (7.11)$$

Em que:

$(X-\mu)_{px1} = \begin{bmatrix} X_1 & - & \mu_1 \\ X_2 & - & \mu_2 \\ & \vdots & \\ X_p & - & \mu_p \end{bmatrix}$ representa o vetor das $p$ variáveis padronizadas;

$F_{px1} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_p \end{bmatrix}$ representa o vetor de fatores comuns;

$\varepsilon_{px1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$ representa o vetor dos fatores específicos;

e

$\Lambda_{pxm} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}$ representa a matriz dos pesos fatoriais.

## 7.4. ADEQUAÇÃO DA UTILIZAÇÃO DA ANÁLISE FATORIAL

Para que a utilização da análise fatorial seja adequada, o pesquisador deve efetuar os seguintes passos: analisar a matriz de correlações, verificar a estatística KMO e o teste de esfericidade de Bartlett e analisar a matriz anti-imagem.



### 7.4.1. Análise da Matriz de Correlações

Como a AF é baseada nas correlações entre as variáveis, o primeiro passo é examinar a matriz de correlações e verificar se existem valores significativos para justificar a utilização da técnica. Caso as correlações entre todas as variáveis sejam baixas, talvez a análise fatorial não seja apropriada. Além disso, é de se esperar que as variáveis que apresentam alta correlação tendem a compartilhar o mesmo fator.

A matriz de correlações mede a associação linear entre as variáveis, por meio do coeficiente de correlação de Pearson.

De acordo com Hair, Anderson, Tatham e Black (2005), se a inspeção visual da matriz de correlações não revelar um número substancial de valores superiores a 0,30, há fortes indícios de que a utilização da técnica não é apropriada.

### 7.4.2. KMO e Teste de Esfericidade de Bartlett

Uma forma de examinar a matriz de correlações e verificar a adequação da AF consiste na aplicação do teste de esfericidade de Bartlett. Utiliza-se tal teste com o intuito de avaliar a hipótese de que a matriz das correlações pode ser a matriz identidade com determinante igual a 1, conforme apresentado a seguir:

$$\begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \quad (7.12)$$

Se a matriz de correlações for igual à matriz identidade, isso significa que as inter-relações entre as variáveis são iguais a 0 e, neste caso, deve-se reconsiderar a utilização de análise fatorial.

Se a hipótese nula ($H_0$: a matriz de correlações é uma matriz identidade) não for rejeitada, isso significa que as variáveis não estão correlacionadas e, nesta situação, não é adequada a utilização da AF. Por outro lado, se a hipótese nula for rejeitada, haverá indícios de que existem correlações significativas entre as variáveis originais. Vale destacar que este teste requer que as variáveis apresentem normalidade multivariada (PESTANA e GAGEIRO, 2005).

Uma estatística usual é a de Kaiser-Meyer-Olkin (KMO), que compara as correlações simples com as correlações parciais. Esta medida é dada pela seguinte expressão:

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum_{i \neq j} a_{ij}^2} \quad (7.13)$$

Em que:

$r_{ij}$ = coeficiente de correlação entre variáveis;

$a_{ij}$ = coeficiente de correlação parcial.

A estatística KMO, cujos valores variam entre 0 e 1, avalia a adequação da amostra quanto ao grau de correlação parcial entre as variáveis, que deve ser pequeno. O valor de KMO próximo de 0 indica que a análise fatorial pode não ser adequada, pois existe uma correlação fraca entre as variáveis. Por outro lado, quanto mais próximo de 1 o seu valor, mais adequada é a utilização da técnica. Os intervalos de análise dos valores de KMO podem ser observados no quadro a seguir.

**Quadro 7.1:** Estatística KMO (Kaiser-Meyer-Olkin)

| KMO | Análise Fatorial |
|---|---|
| 1 – 0,9 | Muito boa |
| 0,8 – 0,9 | Boa |
| 0,7 – 0,8 | Média |
| 0,6 – 0,7 | Razoável |
| 0,5 – 0,6 | Má |
| < 0,5 | Inaceitável |

Assim, valores para a estatística KMO iguais ou inferiores a 0,60 indicam que a análise fatorial pode ser inadequada.

### 7.4.3. Matriz Anti-imagem

A matriz de correlações anti-imagem contém os valores negativos das correlações parciais e é uma forma de obter indícios acerca da necessidade de eliminação de determinada variável do modelo.

Pode-se calcular uma Medida de Adequação da Amostra, ou *Measure of Sampling Adequacy* (MSA), para cada variável, de forma similar à estatística KMO. Esta medida inclui apenas os coeficientes que se deseja analisar. Sua expressão é apresentada a seguir:

$$MSA = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2} \qquad (7.14)$$

O pesquisador deve analisar primeiramente os valores de MSA para cada variável individualmente e excluir as que se encontram no domínio inaceitável, segundo Hair, Anderson, Tatham e Black (2005).

A diagonal principal da matriz anti-imagem gerada como *output* pelo *software* SPSS fornece os valores de MSA. É de se esperar que, quanto maiores forem tais valores, melhor será a utilização da AF. Entretanto, se alguma variável apresentar baixo valor na diagonal principal e alto valor fora dela, talvez haja necessidade de excluí-la do modelo.

Cabe observar que, por vezes, a baixa correlação de determinada variável com as demais não necessariamente implica sua eliminação, uma vez que esta variável pode representar um fator isoladamente.

## 7.5. EXTRAÇÃO DOS FATORES INICIAIS

Conforme já comentado, o objetivo da AF consiste em determinar um número reduzido de fatores que podem representar a estrutura das variáveis originais.

Nesta etapa, é determinado o número de fatores comuns necessários para descrever adequadamente os dados. Assim, o pesquisador deve decidir: (1) o método de extração dos fatores; e (2) o número de fatores selecionados para representar a estrutura latente dos dados.

### 7.5.1. Método de Extração

Há, basicamente, dois métodos principais que podem ser utilizados para a obtenção de fatores: Análise dos Componentes Principais (ACP) e Análise dos Fatores Comuns (AFC).

A ACP considera a variância total dos dados, ao contrário da AFC, cujos fatores são estimados com base na variância comum. Conforme abordado anteriormente, a variância pode ser decomposta nos seguintes termos: comum, específica e erro. A variância comum, também chamada de comunalidade, é aquela compartilhada entre as variáveis, a variância específica é aquela ligada à variável individual e o termo do erro representa a variância ligada aos fatores aleatórios.



A ACP procura uma combinação linear das variáveis observadas, de maneira a maximizar a variância total explicada. Se determinadas variáveis $X_1$, $X_2$ , $X_3$ e $X_4$ forem altamente correlacionadas, elas serão combinadas de modo a formar um fator que explicará a maior quantidade de variância na amostra. O segundo componente terá a segunda maior quantidade de variância e não será correlacionado com o primeiro e, assim, sucessivamente.

Neste sentido, um conjunto de variáveis correlacionadas é transformado em um conjunto de variáveis não correlacionadas. A AF permite que seus resultados possam ser utilizados como *inputs* de outras técnicas multivariadas, como, por exemplo, em técnicas de regressão múltipla, análise discriminante e regressão logística, a fim de possibilitar ao pesquisador a eliminação de eventuais problemas de multicolinearidade das variáveis explicativas, conforme será estudado, respectivamente, nos Capítulos 10, 11 e 12.

A escolha do método de extração depende do objetivo do pesquisador. Se a escolha for pela redução dos dados para obtenção do mínimo número de fatores necessários para explicar o máximo de variância representada pelas variáveis originais, a análise de componentes principais é apropriada. Quando o objetivo primário é identificar fatores ou dimensões latentes que refletem o que as variáveis têm em comum, a análise de fatores comuns é o método mais apropriado, de acordo com Hair, Anderson, Tatham e Black (2005) e Ho (2006).

Além da ACP e AFC, Reis (2001) ainda destaca os seguintes métodos:

- Máxima verossimilhança: indicado quando se trata de uma amostra de indivíduos retirados de uma população normal e se pretende explicar a estrutura latente da matriz de correlações.
- Mínimos quadrados ordinários e generalizados (OLS e GLS): objetivos semelhantes aos do método anterior.
- Alpha: parte do pressuposto de que as variáveis em estudo constituem uma amostra do universo de variáveis existentes e de que os indivíduos compõem toda a população.

### 7.5.2. Escolha do Número de Fatores

A AF busca transformar um conjunto de variáveis em fatores e, para isso, o método extrai, primeiramente, a combinação linear que explica a maior parte da variância dos dados. Na sequência, o modelo busca uma nova combinação entre as variáveis que explique um montante de variância cada vez menor e, assim, sucessivamente.

Neste ponto, o pesquisador deve decidir quantos fatores reter e, para facilitar essa decisão, explicitamos os seguintes critérios que podem auxiliá-lo:

- Critério da raiz latente (critério de Kaiser).
- Critério *a priori*.
- Critério de percentagem de variância.
- Critério do gráfico Scree;

Pelo **critério da raiz latente (critério de Kaiser)**, escolhe-se o número de fatores a reter, em função do número de valores próprios acima de 1. Os valores próprios, também chamados de autovalores ou *eigenvalues*, são ordenados por dimensão. Os *eigenvalues* mostram a variância explicada por cada fator, ou seja, quanto cada fator consegue explicar da variância total. No método de extração de componentes principais, a soma dos valores próprios iguala o número de variáveis.

A escolha dos componentes que apresentam *eigenvalues* maior que 1 decorre do fato de que, no mínimo, o componente deve explicar a variância de uma variável utilizada no modelo, uma vez que estamos trabalhando com variáveis padronizadas, com média 0 e variância igual a 1. Logo, somente os fatores que possuírem valores próprios (*eigenvalues*) superiores a 1 serão significativos. Ressalta-se que o SPSS apresenta ao pesquisador a opção de alteração do valor 1 como critério de "corte" para o número de fatores.

O **critério *a priori*** é o método mais simples, pois, neste caso, o pesquisador já sabe quantos fatores extrair da AF.

O **critério da porcentagem de variância** consiste em escolher, como número de fatores, um número mínimo necessário para que o percentual de variância explicada alcance o nível satisfatório desejado. Neste caso, o pesquisador escolherá qual o nível que acredita ser satisfatório para sua pesquisa.

O **critério do gráfico Scree** é utilizado para identificar o número ótimo de fatores que podem ser extraídos antes que a quantia de variância única comece a dominar a estrutura de variância comum (HAIR, ANDERSON, TATHAM e BLACK, 2005).

O gráfico Scree é realizado por meio da plotagem dos valores da raiz latente, ou *eigenvalues*, no eixo Y e o número de fatores no eixo X, de acordo com a ordem de extração. Assim, o ponto a partir do qual o gráfico passa a se tornar "mais horizontal" reflete um indicativo do número máximo de fatores a serem extraídos.

O gráfico a seguir mostra, apenas como exemplo, o comparativo entre o critério do gráfico Scree e o critério de Kaiser.

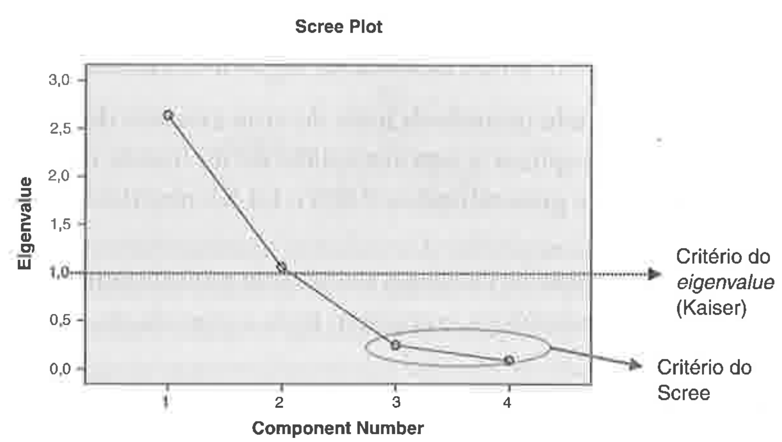


**Figura 7.1:** Gráfico Scree.

### 7.6. ROTAÇÃO DOS FATORES

Normalmente, os fatores produzidos na fase de extração nem sempre são facilmente interpretados. A aplicação de um método de rotação tem como objetivo principal a transformação dos coeficientes dos componentes principais retidos em uma estrutura simplificada (THURSTONE *apud* REIS, 2001).

Segundo Corrar, Paulo e Dias Filho (2007), "a rotação dos fatores é possível, pois as cargas fatoriais podem ser representadas como pontos entre eixos (que, neste caso, são os próprios fatores). Esses eixos podem ser girados sem alterar a distância entre os pontos. Todavia, as coordenadas do ponto em relação aos eixos são alteradas, ou seja, as cargas fatoriais (relação entre fator e variável) são alteradas na rotação".

Os métodos de rotação podem ser ortogonais ou oblíquos. Os métodos ortogonais produzem fatores que não estão correlacionados entre si, chamados de fatores ortogonais, sendo interpretados a partir de suas cargas (*loadings*). Na rotação oblíqua, por outro lado, os fatores estão correlacionados e, para a interpretação da solução, torna-se necessária a consideração simultânea das correlações e dos *loadings*.

Para os métodos rotacionais ortogonais, merecem destaque o Varimax, o Quartimax e o Equamax. O mais utilizado é o Varimax, que busca minimizar o número de variáveis que têm altas cargas em um fator, simplificando a interpretação dos fatores. Sobre o método Varimax, Reis (2001) afirma que "é um método ortogonal e pretende que, para cada componente principal, existam apenas alguns pesos significativos e todos os outros sejam próximos de zero, isto é, o objetivo é maximizar a variação entre os pesos de cada componente principal, daí o nome Varimax".



O método Quartimax busca simplificar as linhas de uma matriz fatorial (número de fatores), ou seja, seu objetivo é tornar os pesos de cada variável elevados para um pequeno número de componentes, e próximos de zero para todos os demais componentes (REIS, 2001). Ou seja, este método busca minimizar o número de fatores necessários para explicar uma variável.

Já o método Equamax congrega características dos métodos anteriores, ou seja, seu objetivo é simplificar as linhas e colunas simultaneamente (simplificação dos fatores e das variáveis).

Os métodos de rotação oblíqua gerados pelo SPSS são o Direct Oblimin e o Promax. Nos métodos de rotação oblíqua, as comunalidades são preservadas, porém os fatores gerados apresentam-se de forma mais fortemente correlacionada.

Do ponto de vista teórico, sobre os métodos rotacionais ortogonais ou oblíquos, não há razão analítica para favorecer um método em detrimento de outro. Entretanto, se o objetivo da pesquisa for reduzir o número de variáveis originais, independentemente de quão significativos os fatores resultantes possam ser, a solução apropriada provavelmente será gerada por meio de um método ortogonal.

A matriz de componentes, após a rotação ortogonal, tem como objetivo extremar os valores das cargas fatoriais (*loadings*), de modo que cada variável se associe apenas a um fator. A rotação transforma a matriz inicial das cargas fatoriais em outra mais fácil de ser interpretada. Na AF, as variáveis com baixas cargas fatoriais (*loadings*) devem ser eliminadas, de forma que sejam utilizadas apenas as variáveis com elevados *loadings*.

Vale destacar que a rotação não afeta a qualidade de ajuste do modelo fatorial, as comunalidades e o total da variância explicada pelos fatores. Entretanto, o percentual da variância explicada em cada fator muda após a rotação.

As Figuras 7.2 e 7.3 exemplificam os métodos de rotação ortogonal e oblíquo.

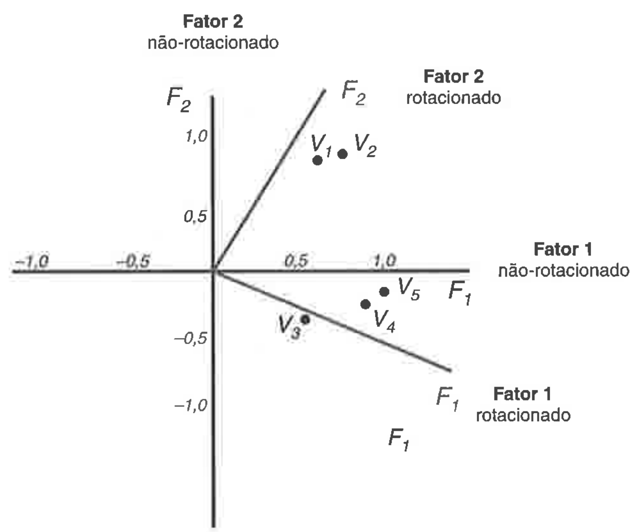


**Figura 7.2:** Método de rotação ortogonal.

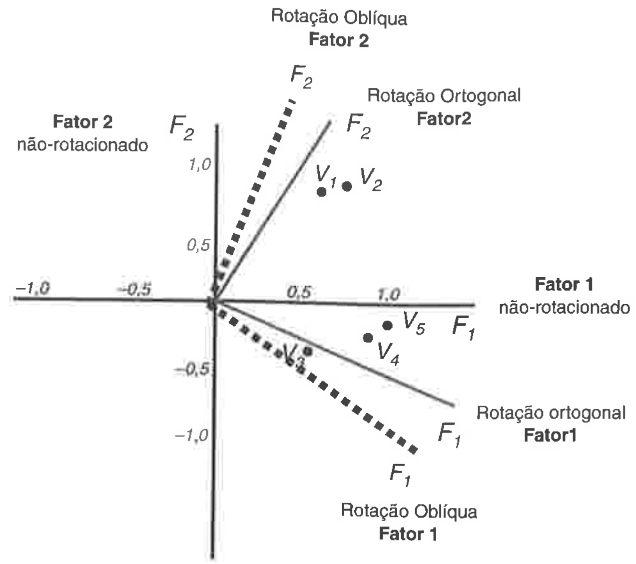


**Figura 7.3:** Método de rotação oblíqua.

## 7.7. INTERPRETAÇÃO DOS FATORES

A última etapa da técnica refere-se, justamente, à interpretação e nomeação dos fatores por meio das cargas fatoriais. Entretanto, o pesquisador deve decidir quais cargas fatoriais valem a pena ser consideradas. Como regra geral, considera-se que as cargas fatoriais maiores que 0,30 atingem o nível mínimo; cargas de 0,40 são consideradas mais importantes; se forem maiores que 0,50, serão consideradas estatisticamente significativas (HAIR, ANDERSON, TATHAM e BLACK, 2005).

A Tabela 7.2 a seguir apresenta orientações para a identificação de cargas fatoriais significantes, ao nível de 5% de significância, com base no tamanho da amostra.

Vale destacar que o valor da carga fatorial ao quadrado representa a quantia de variância total da variável explicada pelo fator. Além disso, a comunalidade é a proporção da variância de cada variável explicada pelo conjunto dos fatores selecionados, sendo calculada pela somatória dos quadrados das cargas de cada fator na variável. Tais conceitos serão mais bem trabalhados e aplicados na seção a seguir, que apresenta um exemplo prático de elaboração da técnica.

**Tabela 7.2:** Relação entre Cargas Fatoriais e Tamanho da Amostra

| Carga fatorial | Tamanho da amostra |
|---|---|
| 0,30 | 350 |
| 0,35 | 250 |
| 0,40 | 200 |
| 0,45 | 150 |
| 0,50 | 120 |
| 0,55 | 100 |
| 0,60 | 85 |
| 0,65 | 70 |
| 0,70 | 60 |
| 0,75 | 50 |

*Fonte:* Hair, Anderson, Tatham e Black (2005).



## 7.8. ANÁLISE FATORIAL: UM EXEMPLO PRÁTICO

Um analista de mercado quer estudar as relações estruturais entre quatro indicadores financeiros provenientes de 45 empresas. Os indicadores selecionados foram:

- Código da empresa (Cód_Emp).
- Prazo médio de recebimento das vendas (PMRV, em dias).
- Endividamento (em %).
- Vendas (em R$ × mil).
- Margem líquida das vendas (em %).

Para desenvolver o exemplo, abra o arquivo **Fatorial.sav**.

Para a utilização da AF, o pesquisador deve verificar a existência de *outliers* e se os dados apresentam distribuição viesada. O pacote estatístico SPSS não dispõe de um teste estatístico para avaliar a normalidade multivariada, mas apenas univariada. Entretanto, para Hair, Anderson, Tatham e Black (2005), "apesar de a normalidade univariada não garantir a normalidade multivariada, se todas as variáveis atendem a essa condição, então quaisquer desvios da normalidade multivariada geralmente são inócuos".

O *software* SPSS não apresenta um teste robusto para validar essa premissa, mas dispõe de ferramentas alternativas que sugerem a existência ou não de distribuição normal multivariada dos dados, como, por exemplo, os testes de Kolmogorov-Smirnov e de Shapiro-Wilk, bem como o ferramental para a elaboração de Box-Plots.

Para os pesquisadores mais interessados e com acesso ao *software* SAS, a página virtual do livro no site www.campus.com.br traz a rotina para elaboração do teste de normalidade multivariada nesse pacote computacional.

Para testar a normalidade de cada variável individualmente, selecione **Analyze → Descriptive Statistics → Explore**, de acordo com a Figura 7.4 e conforme já realizado no Capítulo 5.

Na sequência, insira as variáveis *PMRV*, *Endividamento*, *Vendas* e *Margem_líquida* em **Dependent List**. Em **Plots**, marque a opção **Normality plots with tests** e, por fim, clique em **Continue** e em **OK**, conforme a Figura 7.5.

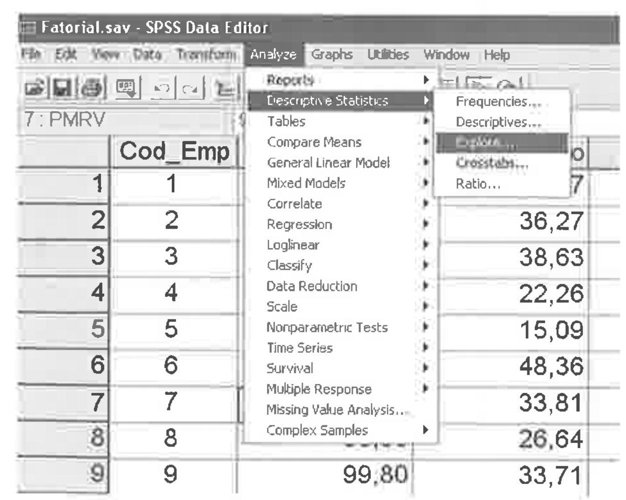

**Figura 7.4:** Testes de normalidade univariada – Kolmogorov-Smirnov e Shapiro-Wilk.

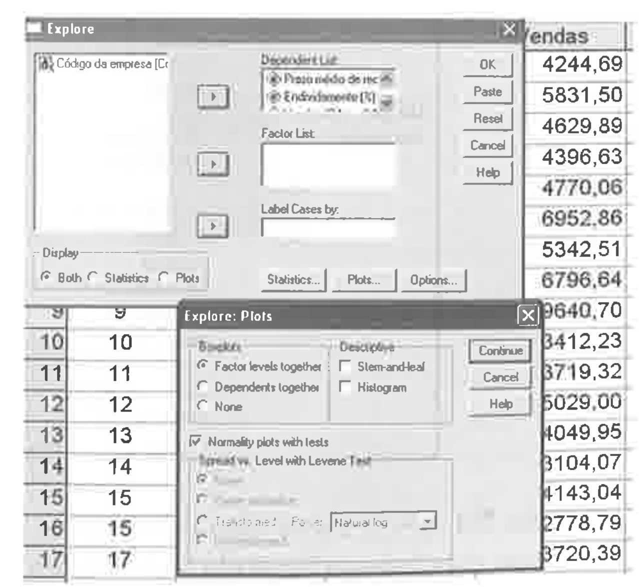

**Figura 7.5:** Caixa de diálogo – testes de normalidade univariada.

Na Tabela 7.3, são apresentadas algumas medidas de posição e de dispersão para cada uma das variáveis.

**Tabela 7.3:** Algumas Medidas de Posição e de Dispersão

**Descriptives**

| | | | Statistic | Std. Error |
|---|---|---|---|---|
| Prazo médio de recebimento das vendas (dias) | Mean | | 53,1260 | 4,62726 |
| | 95% Confidence Interval for Mean | Lower Bound | 43,8004 | |
| | | Upper Bound | 62,4516 | |
| | 5% Trimmed Mean | | 53,0905 | |
| | Median | | 49,2200 | |
| | Variance | | 963,519 | |
| | Std. Deviation | | 31,04060 | |
| | Minimum | | 6,42 | |
| | Maximum | | 99,80 | |
| | Range | | 93,38 | |
| | Interquartile Range | | 60,99 | |
| | Skewness | | ,145 | ,354 |
| | Kurtosis | | -1,427 | ,695 |
| Endividamento (%) | Mean | | 31,7116 | 1,87093 |
| | 95% Confidence Interval for Mean | Lower Bound | 27,9409 | |
| | | Upper Bound | 35,4822 | |
| | 5% Trimmed Mean | | 31,0350 | |
| | Median | | 29,7460 | |
| | Variance | | 157,518 | |
| | Std. Deviation | | 12,55060 | |
| | Minimum | | 14,77 | |
| | Maximum | | 69,44 | |
| | Range | | 54,68 | |
| | Interquartile Range | | 19,31 | |
| | Skewness | | ,783 | ,354 |
| | Kurtosis | | ,313 | ,695 |
| Vendas (R$ x mil) | Mean | | 3989,2929 | 238,92561 |
| | 95% Confidence Interval for Mean | Lower Bound | 3507,7700 | |
| | | Upper Bound | 4470,8158 | |
| | 5% Trimmed Mean | | 3869,7144 | |
| | Median | | 3719,3200 | |
| | Variance | | 2568845 | |
| | Std. Deviation | | 1602,762 | |
| | Minimum | | 1980,57 | |
| | Maximum | | 9640,70 | |
| | Range | | 7660,13 | |
| | Interquartile Range | | 2044,77 | |
| | Skewness | | 1,274 | ,354 |
| | Kurtosis | | 2,206 | ,695 |
| Margem líquida das vendas (%) | Mean | | 13,2216 | ,49542 |
| | 95% Confidence Interval for Mean | Lower Bound | 12,2231 | |
| | | Upper Bound | 14,2200 | |
| | 5% Trimmed Mean | | 13,2069 | |
| | Median | | 13,0540 | |
| | Variance | | 11,045 | |
| | Std. Deviation | | 3,32340 | |
| | Minimum | | 8,45 | |
| | Maximum | | 18,19 | |
| | Range | | 9,74 | |
| | Interquartile Range | | 6,85 | |
| | Skewness | | ,152 | ,354 |
| | Kurtosis | | -1,464 | ,695 |

A Tabela 7.4 apresenta os resultados dos testes de normalidade. Com base no teste de Kolmogorov-Smirnov, estudado no Capítulo 5 (Inferência Estatística), observa-se que todas as variáveis possuem distribuição normal (*p-value* > 0,01), para um nível de significância de 1%. Neste caso, preferiu-se o teste de Kolmogorov-Smirnov ao de Shapiro-Wilk, já que o tamanho da amostra é maior do que 30.



**Tabela 7.4:** Resultados dos Testes de Normalidade Univariada

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| Prazo médio de recebimento das vendas (dias) | ,107 | 45 | ,200* | ,914 | 45 | ,003 |
| Endividamento (%) | ,120 | 45 | ,100 | ,940 | 45 | ,021 |
| Vendas (R$ x mil) | ,126 | 45 | ,073 | ,907 | 45 | ,002 |
| Margem líquida das vendas (%) | ,148 | 45 | ,015 | ,906 | 45 | ,001 |

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

A próxima etapa consiste na geração do Box-Plot, que pode possibilitar a obtenção de indícios de normalidade multivariada. Entretanto, antes de elaborarmos o Box-Plot, é importante padronizarmos (Z scores) as variáveis.

Após a obtenção das variáveis padronizadas, selecione **Graphs → Legacy Dialogs → Boxplot,** conforme mostrado na Figura 7.6.

Na sequência, selecione **Simple, Summaries of separate variables**, de acordo com a Figura 7.7.

Conforme apresentado na Figura 7.8, selecione as variáveis padronizadas *Zscore(PMRV)*, *Zscore(Endividamento)*, *Zscore(Vendas)* e *Zscore(Margem_líquida)* na caixa **Boxes Represent** e clique em **OK.**

O Box-Plot é apresentado na Figura 7.9 que, em consonância com os resultados da Tabela 7.4, mostram o comportamento das distribuições das variáveis que, aparentemente, seguem uma distribuição normal multivariada.

Além disso, o Box-Plot permite a identificação de possíveis *outliers*. Caso haja elementos atípicos, talvez seja necessária sua exclusão.

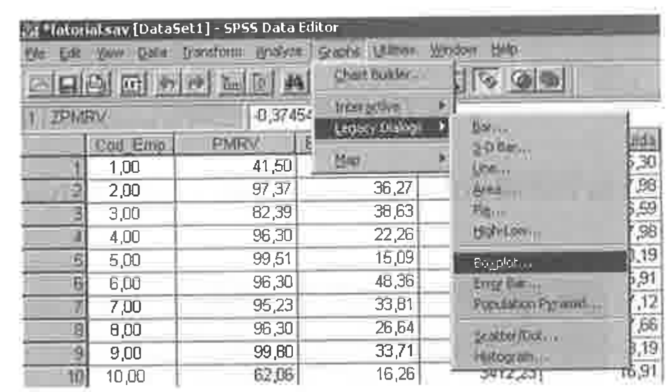

**Figura 7.6:** Box-Plot.


**Figura 7.7:** Caixa de diálogo do Box-Plot.

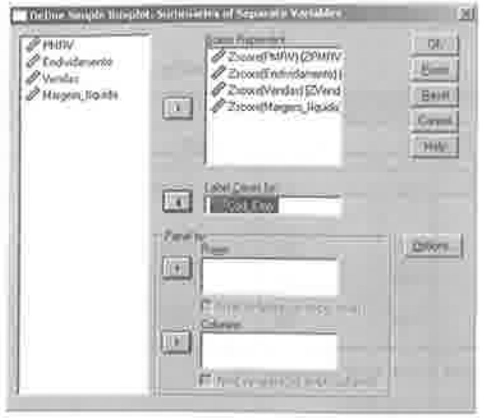

**Figura 7.8:** Seleção de variáveis.

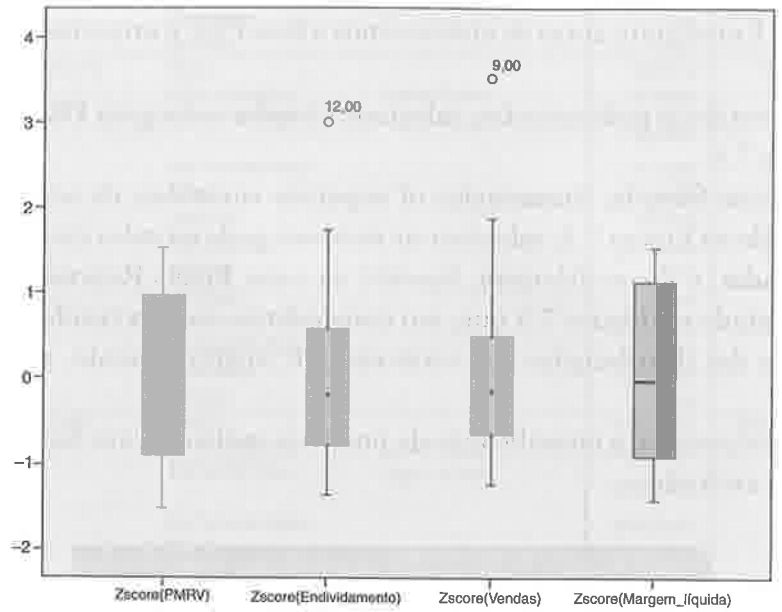


**Figura 7.9:** Gráfico Box-Plot.

Agora, vamos gerar a matriz de correlações que constam do menu **Analyze → Correlate → Bivariate** de acordo com a Figura 7.10.

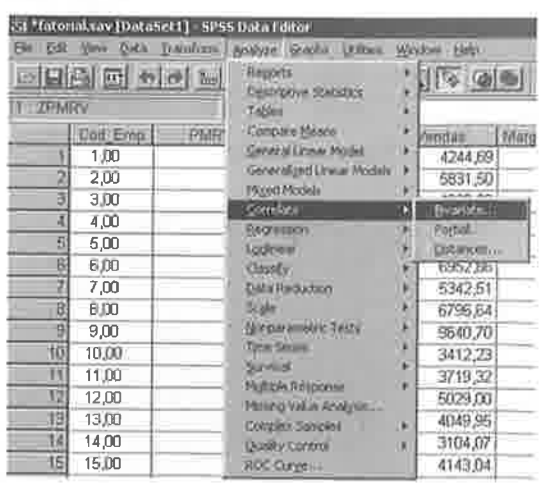

**Figura 7.10:** Análise de correlação bivariada.



Conforme a Figura 7.11, em **Variables**, selecione as quatro variáveis originais e clique no coeficiente de correlação de Pearson.

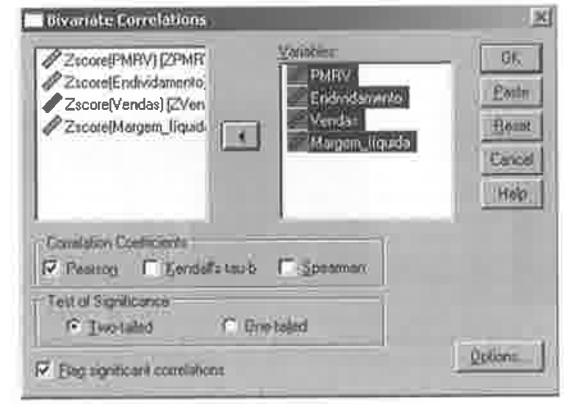


**Figura 7.11:** Correlação: seleção de variáveis.

**Correlations**

| | | PMRV | Endividamento | Vendas | Margem_líquida |
|---|---|---|---|---|---|
| PMRV | Pearson Correlation | 1 | ,235 | ,625** | ,598** |
| | Sig. (2-tailed) | | ,121 | ,000 | ,000 |
| | N | 45 | 45 | 45 | 45 |
| Endividamento | Pearson Correlation | ,235 | 1 | ,238 | -,098 |
| | Sig. (2-tailed) | ,121 | | ,115 | ,523 |
| | N | 45 | 45 | 45 | 45 |
| Vendas | Pearson Correlation | ,625** | ,238 | 1 | ,580** |
| | Sig. (2-tailed) | ,000 | ,115 | | ,000 |
| | N | 45 | 45 | 45 | 45 |
| Margem_líquida | Pearson Correlation | ,598** | -,098 | ,580** | 1 |
| | Sig. (2-tailed) | ,000 | ,523 | ,000 | |
| | N | 45 | 45 | 45 | 45 |

** Correlation is significant at the 0.01 level (2-tailed).

**Figura 7.12: Matriz de correlações de Pearson.**

Por meio da análise da matriz da Figura 7.12, pode-se observar que há altas correlações entre as variáveis *Vendas*, *PMRV* e *Margem_líquida*, que são significantes ao nível de confiança de 5%.

Assim sendo, os requisitos iniciais de existência de considerável número de correlações com valores superiores a 0,30 e de normalidade multivariada (apenas indícios, já que o teste de normalidade multivariada não foi efetivamente realizado) foram atendidos, o que nos permite dar continuidade à aplicação da técnica de AF.

Para realizarmos a AF no SPSS, selecione **Analyze → Data Reduction → Factor...**, conforme apresentado na Figura 7.13.

Na sequência, em **Variables**, selecione as variáveis *PMRV*, *Endividamento*, *Vendas* e *Margem_líquida*, conforme Figura 7.14. Ressalta-se que não há diferenças nos resultados, caso seja feita a seleção das variáveis padronizadas.

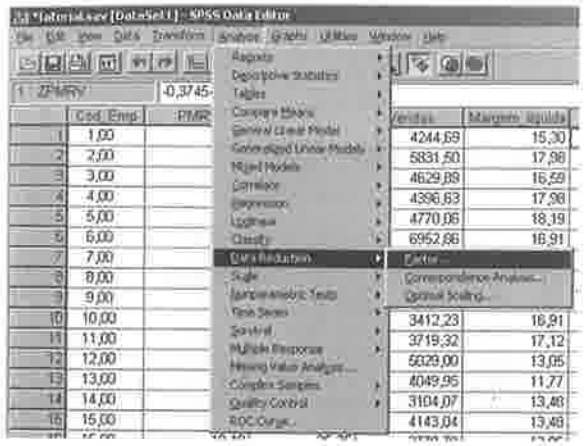

**Figura 7.13:** Procedimento para elaboração da análise fatorial.

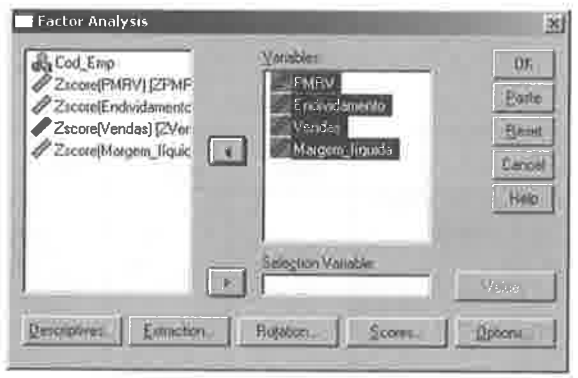


**Figura 7.14:** Análise fatorial – seleção de variáveis.

Em seguida, clique em **Descriptives...**, cujo procedimento abrirá a caixa de diálogo apresentada por meio da Figura 7.15, e selecione, em **Statistics**, as opções **Univariate descriptives** e **Initial solution**, para obter as análises descritivas das variáveis analisadas e a solução inicial, respectivamente. Além disso, selecione, em **Correlation Matrix**, as opções **Coefficients, Significance levels, Determinant, KMO and Bartlett´s test of sphericity** e **Anti-image**. Por fim, clique em **Continue**.

Na sequência, clique em **Extraction...** para selecionar o método de extração dos fatores, conforme a Figura 7.16.

O SPSS fornece diversos métodos de extração, conforme já discutido anteriormente. No nosso exemplo, em **Method**, selecione **Principal Components** e, em **Display**, selecione **Unrotated factor solution**

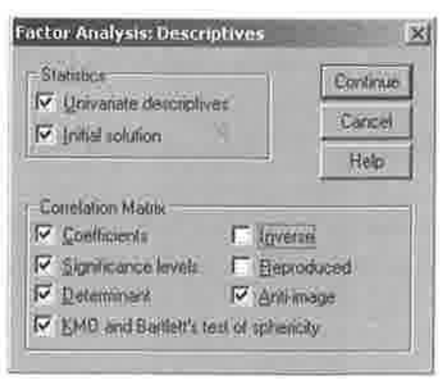


**Figura 7.15:** Análise fatorial – menu **Descriptives**.



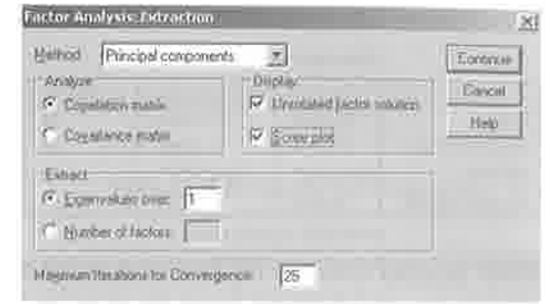


**Figura 7.16:** Análise fatorial – definição do método de extração dos fatores.

(para ter acesso à solução inicial da análise fatorial) e **Scree Plot** (que auxilia na seleção do número de fatores a serem extraídos). Em **Extract**, o pesquisador pode decidir quantos fatores serão extraídos pela AF, em função dos valores de "corte" do *eigenvalue*. O SPSS traz, como *default*, que a extração dos fatores seja efetuada para valores de *eigenvalues* superiores a 1, entretanto, o pesquisador pode utilizar outros valores de "corte". Clique em **Continue**.

Em seguida, clique em **Rotation...**, para abrir a caixa de diálogo, conforme mostrado na Figura 7.17, e selecione o método Varimax. Em **Display**, selecione **Rotated solution** e **Loading plot(s)**. Por fim, clique em **Continue**.

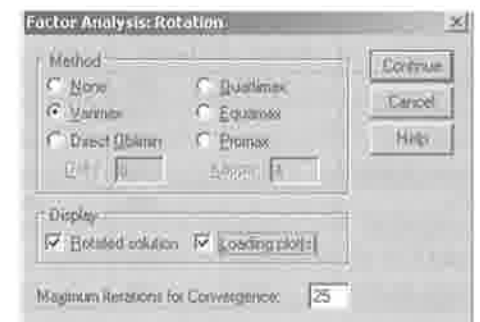


**Figura 7.17:** Análise fatorial – definição do método de rotação.

Na sequência, clique em **Scores...** para selecionar o método de cálculo dos escores (**Regression**), salvar os fatores na base de dados (**Save as variables**) e solicitar a matriz dos coeficientes dos escores fatoriais (**Display factor score coefficient matrix**) , conforme a Figura 7.18.

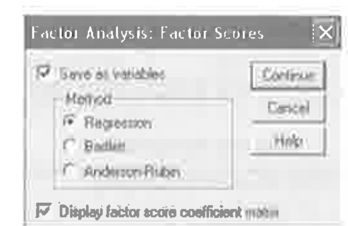


**Figura 7.18:** Análise fatorial – definição do método de cálculo dos escores fatoriais.

Finalmente, clique no botão **OK** para obter os *outputs*.

A tabela a seguir mostra a média e o desvio padrão para cada variável analisada.

**Tabela 7.5:** Estatísticas Descritivas

**Descriptive Statistics**

| | Mean | Std. Deviation | Analysis N |
|---|---|---|---|
| PMRV | 53,1260 | 31,04060 | 45 |
| Endividamento | 31,7116 | 12,55060 | 45 |
| Vendas | 3989,2929 | 1602,76171 | 45 |
| Margem_líquida | 13,2216 | 3,32340 | 45 |

A matriz de correlações apresenta os coeficientes de Pearson entre as variáveis e os *p-values* para as hipóteses $H_0$: $p = 0$ *versus* $h_1$: $p > 0$. O desejável é que ocorram altos coeficientes de correlação entre as variáveis.

Com base na Tabela 7.6, podemos afirmar que existem fortes correlações entre PMRV e Vendas (0,625), entre Vendas e Margem Líquida (0,580) e entre PMRV e Margem Líquida (0,598), todas significativas ao nível de confiança de 1%. Entretanto, não verificamos a existência de correlação entre Endividamento e as demais variáveis. Portanto, podemos supor que existe algum compartilhamento de fatores em comum entre as três primeiras variáveis, o que justifica a utilização da análise fatorial, e que a quarta variável (Endividamento) poderá ser excluída ou compor isoladamente outro fator.

**Tabela 7.6:** Matriz de Correlações

**Correlation Matrix**[a]

| | | PMRV | Endividamento | Vendas | Margem_líquida |
|---|---|---|---|---|---|
| Correlation | PMRV | 1,000 | ,235 | ,625 | ,598 |
| | Endividamento | ,235 | 1,000 | ,238 | -,098 |
| | Vendas | ,625 | ,238 | 1,000 | ,580 |
| | Margem_líquida | ,598 | -,098 | ,580 | 1,000 |
| Sig. (1-tailed) | PMRV | | ,060 | ,000 | ,000 |
| | Endividamento | ,060 | | ,057 | ,261 |
| | Vendas | ,000 | ,057 | | ,000 |
| | Margem_líquida | ,000 | ,261 | ,000 | |

a. Determinant = ,281

Na Tabela 7.7, são apresentados os resultados da estatística KMO e do teste de esfericidade de Bartlett. Este último é utilizado para verificar a hipótese de a matriz das correlações ser a matriz identidade (determinante igual a 1). Já a estatística KMO (Kaiser-Meyer-Olkin), que varia entre 0 e 1, compara as correlações simples com as parciais observadas entre as variáveis. Conforme já discutido anteriormente, valores de KMO próximos de zero indicam que a análise fatorial pode não ser adequada, já que explicitam a existência de uma fraca correlação entre as variáveis. Ao contrário, quanto mais próximos de 1, maior será a qualidade da AF.

**Tabela 7.7:** Resultados da Estatística KMO e do Teste de Esfericidade de Bartlett

**KMO and Bartlett's Test**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | ,631 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 53,165 |
| | df | 6 |
| | Sig. | ,000 |

No nosso exemplo, o KMO de 0,631 torna razoável a aplicação da análise fatorial, de acordo com os critérios apresentados no Quadro 7.1, o que permite sua utilização. O nível de significância do teste de esfericidade de Bartlett (*p-value* = 0,000) conduz à rejeição da hipótese de a matriz de correlações ser a identidade, evidenciando, portanto, que há correlações entre as variáveis. Mais uma vez, corrobora a utilização da AF.



Os resultados da Tabela 7.8 também podem ser utilizados para explicitar a adequação da utilização da AF no nosso exemplo.

**Tabela 7.8:** Matriz Anti-imagem

**Anti-image Matrices**

| | | PMRV | Endividamento | Vendas | Margem_líquida |
|---|---|---|---|---|---|
| Anti-image Covariance | PMRV | ,492 | -,159 | -,169 | -,210 |
| | Endividamento | -,159 | ,805 | -,158 | ,233 |
| | Vendas | -,169 | -,158 | ,510 | -,194 |
| | Margem_líquida | -,210 | ,233 | -,194 | ,495 |
| Anti-image Correlation | PMRV | ,691[a] | -,252 | -,338 | -,427 |
| | Endividamento | -,252 | ,318[a] | -,246 | ,369 |
| | Vendas | -,338 | -,246 | ,707[a] | -,387 |
| | Margem_líquida | -,427 | ,369 | -,387 | ,601[a] |

a. Measures of Sampling Adequacy(MSA)

As matrizes anti-imagem de variâncias-covariâncias e de correlações apresentam os valores negativos das covariâncias e das correlações parciais entre as variáveis. Tais valores estimam as correlações entre as variáveis que não decorrem dos fatores comuns. Na matriz anti-imagem de correlações, os valores da diagonal principal também representam uma medida de adequação dos dados à análise fatorial, conhecida por MSA (medida de adequação da amostra), para cada uma das variáveis em análise. Caso algum valor esteja abaixo de 0,5, tal fato indica que esta variável específica pode não se ajustar à estrutura definida pelas outras variáveis e, portanto, merece eventualmente ser eliminada (MAROCO, 2007).

Com base na Tabela 7.8, podemos observar que o MSA foi superior a 0,50 para todas as variáveis, exceto para a variável Endividamento, que apresentou o resultado de 0,318. Entretanto, a decisão de eliminação desta variável não será tomada agora, uma vez que ela pode apresentar elevados valores de comunalidades e de cargas fatoriais, o que indicará que a variável poderá representar, sozinha, um único fator.

**Tabela 7.9:** Comunalidades

**Communalities**

| | Initial | Extraction |
|---|---|---|
| PMRV | 1,000 | ,769 |
| Endividamento | 1,000 | ,955 |
| Vendas | 1,000 | ,758 |
| Margem_líquida | 1,000 | ,831 |

Extraction Method: Principal Component Analysis.

A Tabela 7.9 apresenta as comunalidades, sendo esta representada pela variância total explicada pelos fatores em cada variável. As comunalidades iniciais são iguais a 1 e, após a extração, variam entre 0 e 1, sendo mais próximas de 0 quando os fatores comuns explicam baixa ou nenhuma variância da variável, e 1 quando toda a variância é explicada por todos os fatores. No nosso exemplo, observe que todas as variáveis possuem forte relação com os fatores retidos, pois têm elevadas comunalidades, conforme mostrado na coluna **Extraction.** Assim, a decisão de não excluirmos a variável Endividamento, quando da análise da estatística MSA, foi tomada de forma correta!

Os valores próprios (*eigenvalues* ou autovalores) para cada fator, bem como os respectivos percentuais de variância explicada, são apresentados na Tabela 7.10. Com base na regra de retenção de fatores com valores superiores a 1 (conforme apresentado na Figura 7.16), foram retidos dois fatores que conseguem explicar 82,832% da variância dos dados originais. A tabela ainda apresenta o percentual de variância explicada por fator antes e depois da rotação.

**Tabela 7.10:** *Eigenvalues* e Percentual de Variância Explicada pelos Fatores

**Total Variance Explained**

| | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| Component | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2,242 | 56,061 | 56,061 | 2,242 | 56,061 | 56,061 | 2,198 | 54,939 | 54,939 |
| 2 | 1,071 | 26,772 | 82,832 | 1,071 | 26,772 | 82,832 | 1,116 | 27,893 | 82,832 |
| 3 | ,376 | 9,393 | 92,225 | | | | | | |
| 4 | ,311 | 7,775 | 100,000 | | | | | | |

Extraction Method: Principal Component Analysis.

O gráfico Scree corrobora a retenção de apenas dois fatores, conforme a Figura 7.19.

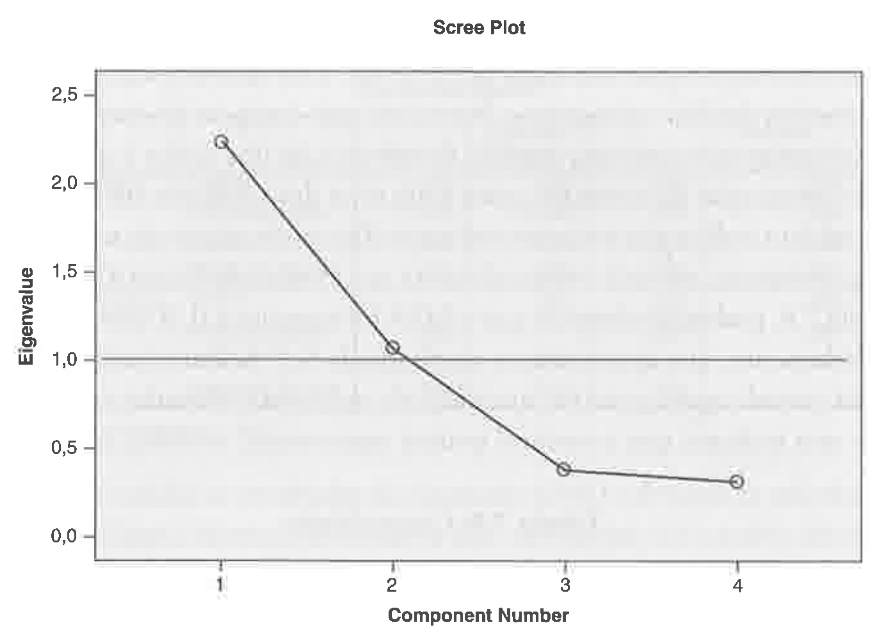


**Figura 7.19:** Gráfico Scree.

A matriz dos componentes (Tabela 7.11) apresenta as cargas (*loadings*) que correlacionam as variáveis com os fatores antes da rotação, ou seja, permite verificar qual fator melhor explica cada uma das variáveis originais.

Com base na Tabela 7.11, podemos afirmar que, para as variáveis PMRV, Vendas e Margem Líquida, há predomínio do Fator 1 (Componente 1) e, para a variável Endividamento, predomina o Fator 2.

Quanto ao nome dos fatores, podemos empiricamente chamar o Fator 1 de volume de negócios (faturamento) e o Fator 2 de estrutura de capital. Esses nomes são apenas sugestões, não sendo, por vezes, tão fácil a nomeação dos fatores, principalmente quando o pesquisador possuir uma base de dados com um número elevado de variáveis que extraem um pequeno número de fatores.



**Tabela 7.11:** Matriz dos Componentes ou Cargas Fatoriais

**Component Matrix[a]**

| | Component | |
|---|---|---|
| | 1 | 2 |
| PMRV | ,876 | ,044 |
| Endividamento | ,269 | ,940 |
| Vendas | ,868 | ,061 |
| Margem_líquida | ,806 | -,427 |

Extraction Method: Principal Component Analysis.
a. 2 components extracted.

De acordo com a Expressão (7.5) apresentada no início do capítulo e com base nas cargas fatoriais da Tabela 7.11, podemos representar cada variável em função dos dois fatores. Assim:

$$\text{PMRV} = 0{,}876 \times F_1 + 0{,}044 \times F_2$$
$$\text{Endividamento} = 0{,}269 \times F_1 + 0{,}940 \times F_2$$
$$\text{Vendas} = 0{,}868 \times F_1 + 0{,}061 \times F_2$$
$$\text{Margem Líquida} = 0{,}806 \times F_1 - 0{,}427 \times F_2$$

Vale destacar que a soma dos quadrados das cargas (*loadings*) das variáveis para cada fator (soma em coluna) é o valor próprio (*eigenvalue*) dos componentes (*vide* Tabela 7.10). Assim:

$$\text{Fator 1: } (0{,}876)^2 + (0{,}269)^2 + (0{,}868)^2 + (0{,}806)^2 = 2{,}242$$

$$\text{Fator 2: } (0{,}044)^2 + (0{,}940)^2 + (0{,}061)^2 + (-0{,}427)^2 = 1{,}071$$

Analogamente, a soma dos quadrados das cargas dos fatores para cada variável (soma em linha) representa as comunalidades (*vide* Tabela 7.9):

$$\text{PMRV} = (0{,}876)^2 + (0{,}044)^2 = 0{,}769$$
$$\text{Endividamento} = (0{,}269)^2 + (0{,}940)^2 = 0{,}955$$
$$\text{Vendas} = (0{,}868)^2 + (0{,}061)^2 = 0{,}758$$
$$\text{Margem Líquida} = (0{,}806)^2 + (-0{,}427)^2 = 0{,}831$$

A seguir, é apresentada a tabela que contém as cargas fatoriais após a rotação (Tabela 7.12). A rotação ortogonal tem por objetivo extremar os valores das cargas, de modo que cada variável se associe a um fator.

**Tabela 7.12:** Matriz dos Componentes após a Rotação ou Cargas Rotacionadas

**Rotated Component Matrix[a]**

| | Component | |
|---|---|---|
| | 1 | 2 |
| PMRV | ,850 | ,215 |
| Endividamento | ,079 | ,974 |
| Vendas | ,840 | ,230 |
| Margem_líquida | ,874 | -,261 |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

Podemos perceber que todas as variáveis, à exceção da variável Endividamento, possuem elevada carga no Fator 1. O Fator 2, conforme já era de se esperar, discrimina apenas a variável Endividamento.

A Figura 7.20 apresenta o gráfico dos componentes após a rotação, sendo possível verificar a proximidade das variáveis (nós ou sementes) com os respectivos fatores (eixos).

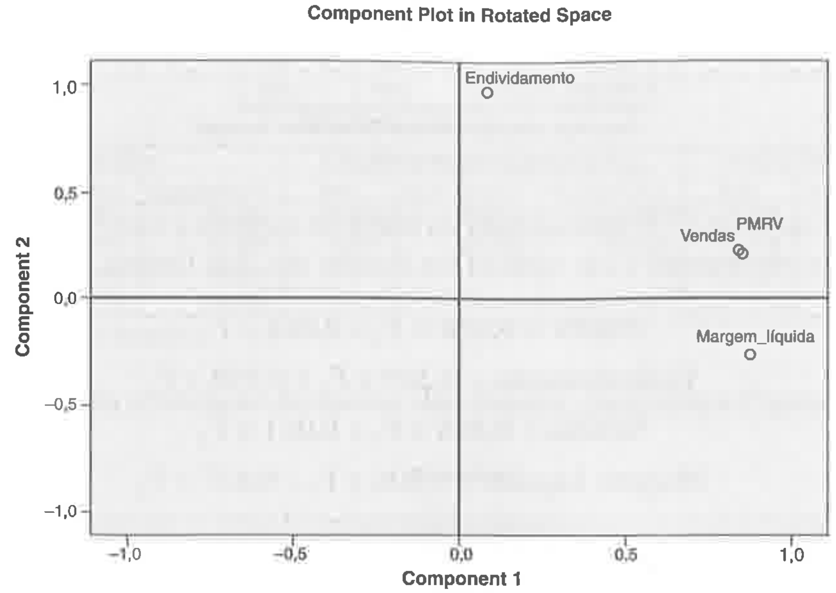


**Figura 7.20:** *Loading Plot* ou relação entre variáveis e fatores após a rotação.

A Tabela 7.13 fornece os escores fatoriais para cada fator.

**Tabela 7.13:** Matriz de Coeficientes dos Escores Fatoriais

**Component Score Coefficient Matrix**

| | Component | |
|---|---|---|
| | 1 | 2 |
| PMRV | ,375 | ,117 |
| Endividamento | -,054 | ,884 |
| Vendas | ,369 | ,131 |
| Margem_líquida | ,430 | -,320 |

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

Com base nos escores da Tabela 7.13 e na Expressão (7.6) apresentada no início do capítulo, podemos calcular os fatores para cada observação da amostra, levando-se em consideração as variáveis padronizadas. Assim:

$$\textbf{Fator 1} = 0{,}375 \times ZPMRV - 0{,}054 \times ZEndividamento + 0{,}369 \times ZVendas + 0{,}430 \times ZMargem\_líquida$$
$$\textbf{Fator 2} = 0{,}117 \times ZPMRV + 0{,}884 \times ZEndividamento + 0{,}131 \times ZVendas - 0{,}320 \times ZMargem\_líquida$$

Consideremos, por exemplo, a primeira observação da base de dados, que apresenta os seguintes valores para cada uma das variáveis padronizadas:

$$ZPMRV = -0{,}3754;$$
$$ZEndividamento = -0{,}4891;$$



$$ZVendas = 0{,}1593;$$
$$ZMargem\_líquida = 0{,}6257.$$

Assim, o cálculo do Fator 1 (Fator Principal) e do Fator 2, para a primeira observação, será:

$$\textbf{Fator 1} = 0{,}375 \times (-0{,}3754) - 0{,}054 \times (-0{,}4891) + 0{,}369 \times (0{,}1593) + 0{,}430 \times (0{,}6257) = \mathbf{0{,}2141}$$
$$\textbf{Fator 2} = 0{,}117 \times (-0{,}3754) + 0{,}884 \times (-0{,}4891) + 0{,}131 \times (0{,}1593) - 0{,}320 \times (0{,}6257) = \mathbf{-0{,}6557}$$

Quando selecionamos a opção para salvar os escores fatoriais, conforme apresentado na Figura 7.18 (menu **Scores**, opção **Save as variables**), o SPSS inseriu na base de dados duas novas variáveis (FAC1_1 e FAC2_1), que representam os fatores 1 e 2 e cujos valores da primeira observação são exatamente os que acabamos de calcular por meio da multiplicação dos escores fatoriais por cada um dos respectivos valores das variáveis padronizadas. Estes fatores criados pela AF podem ser utilizados em substituição às variáveis originais em outras técnicas multivariadas.

É importante ressaltar que a correlação de Pearson entre FAC1_1 e FAC1_2 é 0, uma vez que o método de rotação Varimax gera fatores ortogonais entre si. De acordo com as Figuras 7.21 e 7.22, ao clicarmos em **Analyze → Correlate → Bivariate** e inserirmos os dois fatores resultantes da AF em **Variables**, teremos, como resultado, esta comprovação (Tabela 7.14).

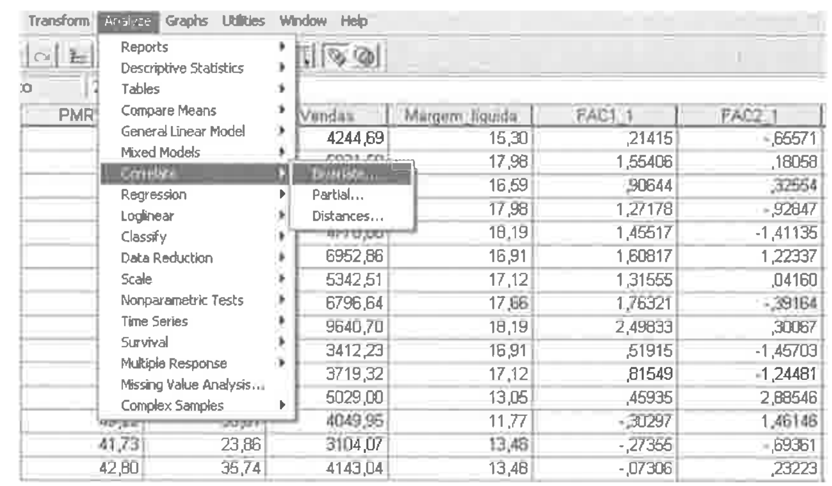


**Figura 7.21:** Correlação entre os dois fatores.


**Figura 7.22:** Elaboração da correlação de Pearson entre os fatores.

**Tabela 7.14:** Resultado da Correlação de Pearson entre os Fatores após a Rotação

**Correlations**

| | | REGR factor score 1 for analysis 2 | REGR factor score 2 for analysis 2 |
|---|---|---|---|
| REGR factor score 1 for analysis 2 | Pearson Correlation | 1 | ,000 |
| | Sig. (2-tailed) | | 1,000 |
| | N | 45 | 45 |
| REGR factor score 2 for analysis 2 | Pearson Correlation | ,000 | 1 |
| | Sig. (2-tailed) | 1,000 | |
| | N | 45 | 45 |

## 7.9. RELAÇÃO COM OUTRAS TÉCNICAS

De acordo com Hair, Anderson, Tatham e Black (2005), "se o objetivo do pesquisador é identificar variáveis apropriadas para a aplicação subsequente em outras técnicas estatísticas, então alguma forma de redução de dados será empregada. As opções incluem: (1) examinar a matriz fatorial e selecionar a variável com maior carga fatorial como representativa substituta para uma dimensão particular; ou (2) substituir o conjunto original de variáveis por um menor e inteiramente novo, criado a partir de escalas múltiplas ou escores fatoriais".

Ao permitir que um conjunto de variáveis correlacionadas seja transformado em um conjunto menor de variáveis (fatores) não correlacionadas, os escores fatoriais proporcionados pela AF podem ser utilizados, caso não haja perda significativa de variância explicada pelos fatores gerados, como *inputs* em diversas técnicas multivariadas, como, por exemplo, em regressões múltiplas (Capítulo 10), análises discriminantes (Capítulo 11) ou regressões logísticas (Capítulo 12), reduzindo assim o problema de multicolinearidade, quando os fatores forem gerados por meio de métodos ortogonais de rotação.

Vale destacar que a análise de conglomerados se assemelha à análise fatorial, pois ambas caracterizam-se como técnicas de agrupamento. Entretanto, quando o pesquisador estiver interessado no agrupamento de variáveis, em vez de observações, a AF mostra-se mais adequada, pela própria natureza dos testes a ela pertinentes. Ressalta-se que o Fator 1, também conhecido como Fator Principal, pode ser utilizado para a estratificação, por meio de quartis ou decis, de bases de dados que apresentam um número elevado de observações e que, portanto, por vezes impossibilita a aplicação direta da análise de conglomerados.

Por fim, como um fator pode representar um indicador de avaliação de determinado fenômeno que comporta um conjunto de variáveis originais, como, por exemplo, um indicador de inteligência de estudantes, de desempenho de empresas ou de nível sociodemográfico de municípios, os fatores gerados por meio da AF podem ser introduzidos como variáveis dependentes em modelos que utilizem técnicas com mais de uma variável dependente, como a MANOVA ou a correlação canônica, que serão estudadas nos Capítulos 13 e 14, respectivamente.

## 7.10. CONSIDERAÇÕES FINAIS

A análise fatorial permite ao pesquisador identificar um número relativamente pequeno de fatores comuns que podem ser utilizados para representar relações entre um grande número de variáveis inter-relacionadas. A partir das correlações observadas entre as variáveis originais, a AF estima os fatores comuns que são subjacentes às variáveis e não são diretamente observáveis.

As possibilidades de aplicação da AF são enormes, dada sua capacidade de simplificação ou redução de um grande número de dados por intermédio da determinação das dimensões latentes (fatores).



Porém, como já mencionado anteriormente, a AF é uma técnica exploratória, não cabendo generalizações além de seus resultados. Mas os fatores proporcionados podem ser utilizados como *inputs* em muitas outras técnicas.

Segundo Mingoti (2005), em muitas pesquisas, é comum que sejam encontradas variáveis com características e naturezas diferentes e, neste sentido, a AF pode auxiliar o pesquisador na determinação de índices que evidenciem essas diferenças.

Com aplicações em diversas áreas do conhecimento, a AF tem se tornado bastante popular e útil para a redução de dados, bem como para a determinação e identificação de índices ou parâmetros de comparação entre objetos, indivíduos, empresas, organizações, municípios, países ou quaisquer outros tipos de observações que apresentem variáveis que possam ser agrupadas para a criação de indicadores de desempenho não observáveis e que levam em consideração o comportamento daquelas variáveis originais observadas.

## 7.11. EXERCÍCIOS – APLICAÇÃO DE BANCOS DE DADOS

1. A seguir são apresentados os resultados do processamento de dados resultantes de uma análise fatorial aplicada em 12 variáveis de 45 municípios do Estado do Rio Grande do Sul relativas a investimentos na área de educação.

   Variáveis:

   - *INSTR*: população sem instrução/população com mais de 10 anos de idade;
   - *PROFF*: número de professores do ensino fundamental na rede pública municipal/população total;
   - *PROFM*: número de professores do ensino médio na rede pública municipal/população total;
   - *PROFS*: número de professores de ensino superior/população total;
   - *EMPIN*: número de empregos na indústria/população total;
   - *EMPCO*: número de empregos em comércio e serviços/população total;
   - *EMPAG*: número de empregos na agricultura/população total;
   - *RICMS*: receita de ICMS/população total;
   - *ESTD1*: número de pessoas com 1 a 6 anos de estudo /população com mais de 10 anos de idade;
   - *ESTD2*: número de pessoas com 7 a 10 anos de estudo/população com mais de 10 anos de idade;
   - *ESTD3*: número de pessoas com 11 a 14 anos de estudo/população com mais de 10 anos de idade;
   - *ESTD4*: número de pessoas com 15 a 18 anos de estudo/população com mais de 10 anos de idade.

| Variável | Fator 1 | Fator 2 | Comunalidades |
|---|---|---|---|
| INSTR | 0,6693 | –0,6019 | 0,8102461 |
| PROFF | –0,8402 | 0,1376 | 0,7248698 |
| PROFM | –0,8211 | 0,1811 | 0,70700242 |
| PROFS | –0,7292 | 0,5487 | 0,83280433 |
| EMPIN | –0,2074 | 0,7789 | 0,64969997 |
| EMPCO | –0,159 | 0,7903 | 0,64985509 |
| EMPAG | 0,0453 | 0,7457 | 0,55812058 |
| RICMS | –0,4275 | 0,6256 | 0,57413161 |
| ESTD1 | –0,7821 | 0,5042 | 0,86589805 |
| ESTD2 | –0,8705 | 0,2935 | 0,8439125 |
| ESTD3 | –0,8508 | 0,119 | 0,73802164 |
| ESTD4 | –0,8228 | 0,3267 | 0,78373273 |
| % Variância Total Explicada | 59,81 | 37,28 | |

Interprete:
a) cargas fatoriais na matriz rotacionada (importância dos fatores na composição de cada variável);
b) comunalidades (índice de variabilidade total explicada por todos os fatores para cada variável);
c) percentual da variância total explicada por meio de cada fator após a rotação.

2. O arquivo **Indicadores_financeiros.sav** contém dados relativos a 100 empresas listadas na *Revista Exame* (Melhores e Maiores) de 2005 (exercício social encerrado em 31/12/2004). As variáveis disponíveis referem-se a:

- Vendas, em US$ mil;
- Lucro Líquido (LL), em US$ mil;
- Patrimônio Líquido (PL), em US$ mil;
- Rentabilidade, em %;
- Capital Circulante Líquido (CCL), em US$ mil;
- Liquidez geral, em %;
- Endividamento geral, em %.

Pede-se:
a) Aplique a análise fatorial, considerando o método de Componentes Principais e Rotação Varimax para a obtenção dos escores fatoriais, com patamar mínimo para *eigenvalues* igual a 1,0.
b) Aplique a análise fatorial, considerando o método de Componentes Principais e Rotação Varimax para a obtenção dos escores fatoriais, com patamar mínimo para *eigenvalues* igual a 0,5.
c) Interprete e compare os resultados obtidos em cada uma das opções anteriores.
d) Considerando o *eigenvalue* de "corte" igual a 1,0 (item (a)), estratifique as empresas em dois grupos, de acordo com o seguinte critério para o Fator Principal:

| Critério para FAC1_1 | Grupo |
|---|---|
| FAC1_1 < 2,00 | A |
| FAC1_1 > 2,00 | B |

e) Os resultados dos grupos gerados no item (d) são muito diferentes daqueles que seriam obtidos por meio de uma análise de conglomerados hierárquicos com método *Between Groups* e distância quadrática euclidiana?
f) Ainda considerando o *eigenvalue* de "corte" igual a 1,0, se o critério de estratificação agora fosse dado pela mediana, ou seja:

| Critério para FAC1_1 | Grupo |
|---|---|
| FAC1_1 < mediana | A |
| FAC1_1 > mediana | B |

As empresas pertencentes a cada grupo seriam as mesmas em relação à solução obtida por meio da análise de conglomerados? Justifique sua resposta.

3) A tabela a seguir fornece dados a respeito de 17 supermercados, com as seguintes variáveis já padronizadas: Número de Lojas, Índice de Liquidez, Número de empregados, Faturamento anual, Patrimônio líquido (PL) e Anos de atuação no mercado. Pede-se:
a) Elabore uma análise fatorial, considerando o método de Componentes Principais e Rotação Varimax para a obtenção dos escores fatoriais, com patamar mínimo para *eigenvalues* igual a 1,0.
b) Elabore uma análise fatorial, considerando o método de Componentes Principais e Rotação Varimax para a obtenção dos escores fatoriais, com patamar mínimo para *eigenvalues* igual a 0,5.



c) Aplique uma análise de conglomerados hierárquicos com método *Between Groups* e distância quadrática euclidiana e sugira um número adequado de *clusters* para esta base de dados.
d) De acordo com o resultado sugerido no item (c), proponha uma estratificação do Fator Principal para que a formação dos conglomerados seja similar.

| Empresa | Lojas | Liquidez | Empregados | Faturamento | PL | Anos_mercado |
|---|---|---|---|---|---|---|
| A | – 0,29 | – 0,38 | – 0,51 | – 0,54 | – 0,50 | – 0,28 |
| B | 3,68 | 3,78 | – 0,58 | – 0,63 | – 0,91 | 3,92 |
| C | – 0,45 | – 0,40 | – 0,58 | – 0,54 | – 0,50 | – 0,28 |
| D | – 0,27 | 0,39 | – 0,58 | – 0,54 | – 0,62 | – 0,28 |
| E | – 0,29 | – 0,40 | – 0,58 | – 0,63 | – 0,50 | – 0,28 |
| F | 0,73 | 0,71 | – 0,63 | – 0,64 | – 0,96 | – 0,28 |
| G | – 0,49 | – 0,41 | – 0,39 | – 0,54 | – 0,50 | – 0,28 |
| H | – 0,29 | – 0,24 | – 0,02 | – 0,15 | – 0,81 | – 0,28 |
| I | – 0,56 | – 0,36 | – 0,14 | – 0,15 | – 0,62 | – 0,28 |
| J | – 0,51 | – 0,41 | – 0,14 | 0,17 | 1,79 | – 0,28 |
| K | – 0,52 | – 0,41 | 1,38 | 2,59 | 1,13 | – 0,28 |
| L | – 0,53 | – 0,41 | 0,38 | 0,17 | 1,78 | – 0,28 |
| M | – 0,53 | – 0,41 | 3,41 | 2,59 | 1,13 | – 0,28 |
| N | – 0,54 | – 0,13 | – 0,14 | – 0,24 | 1,39 | – 0,28 |
| O | – 0,41 | – 0,36 | – 0,58 | – 0,56 | – 0,62 | – 0,28 |
| P | – 0,06 | – 0,29 | – 0,58 | – 0,60 | – 0,87 | – 0,27 |
| Q | 0,40 | – 0,28 | – 0,48 | – 0,52 | – 0,92 | – 0,28 |

4) O arquivo **Mun_SP.sav** contém dados sociodemográficos dos 96 distritos do Município de São Paulo (fonte: http://atlasambiental.prefeitura.sp.gov.br/, acesso em 17/09/2007). As variáveis referem-se a:

| | |
|---|---|
| Renda Familiar | Renda média familiar em reais de outubro/97 (Pesquisa Origem/Destino/Metrô, 1990) |
| Quota Residencial | Área construída residencial (m²)/habitante (PMSP TPCL, 1999) |
| Nível de Escolaridade | Média de anos de estudo da população de 4 anos e mais (IBGE, 1996) |
| Perfil Etário | Em anos (IBGE, 1996) |
| Mortalidade Infantil | Taxa de mortalidade infantil por 1.000 nascidos vivos (Seade, 1998) |
| Crescimento Populacional | Taxa anual de crescimento populacional (IBGE 91/96) |
| Índice de Criminalidade | Taxa de mortalidade por causas externas por 100.000 habitantes (PMSP, PRO-AIM, 1998) |
| População Favelada | Porcentagem de população favelada em relação à população total do distrito – favelas com mais de 50 barracos (IBGE, 1996) |
| Densidade Populacional | Densidade populacional bruta – habitantes/hectare (IBGE, 1996) |

*Fonte*: Site http://atlasambiental.prefeitura.sp.gov.br/, acesso em 17/09/2007.

Pede-se:
a) Elabore uma análise fatorial (método de Componentes Principais e Rotação Varimax) e interprete todos os *outputs*.
b) Com base no Fator Principal extraído no item (a), estratifique os 96 distritos do Município de São Paulo, de acordo com o seguinte critério:

| Grupo Distrital | Fator Principal |
|---|---|
| GRUPO E | FAC1_1 < –1,01 |
| GRUPO D | -1,00 < FAC1_1 < –0,51 |
| GRUPO C | -0,50 < FAC1_1 < –0,01 |
| GRUPO B | 0,00 < FAC1_1 < 0,99 |
| GRUPO A | FAC1_1 > 1,00 |

5) O arquivo **Avaliação_Desempenho.sav** contém dados relativos à pontualidade, produtividade, perfil de liderança e relacionamento interpessoal (todos em forma de notas de 0 a 10) de 40 profissionais de uma empresa. Esses dados foram coletados pela área de recursos humanos para determinar a nota média de cada funcionário em relação a esses quesitos e, para tanto, foram elaboradas entrevistas com profissionais de mesma hierarquia, bem como provenientes de hierarquias superiores e inferiores dentro da empresa. Pede-se:
   a) Por meio de uma análise fatorial (método de Componentes Principais e Rotação Varimax), determine o número de fatores que compõem as quatro variáveis originais.
   b) A estatística KMO e o teste de esfericidade de Bartlett indicam que a aplicação da AF foi adequada?
   c) Qual o percentual de variância total explicada pelo(s) fator(es) gerado(s)?
   d) Qual variável mais contribui positivamente para a formação de cada fator?
   e) Como este(s) fator(es) poderia(m) ser nomeado(s)?
   f) Elabore um *ranking* com o Fator Principal gerado. Quais os três profissionais mais bem avaliados? E os três com pior desempenho?

## 7.12. RESUMO

A análise fatorial é uma técnica estatística multivariada de interdependência que visa à redução dos dados e à criação de indicadores que representam variáveis originais. Neste sentido, transforma variáveis correlacionadas em fatores não correlacionados (ortogonais), de maneira a reduzir a complexidade e facilitar a interpretação dos dados. Esta técnica busca determinar quantos fatores podem ser obtidos por meio da modelagem e o que eles representam, embora nomeá-los não seja uma tarefa objetiva.

Seu uso pressupõe que as variáveis originais sejam correlacionadas e compartilhem um ou mais componentes, cuja relação é explicada por fatores. Uma forma de verificar essa premissa é por meio da análise da matriz de correlações, do teste de esfericidade de Bartlett e da estatística Kaiser-Meyer-Olkin (KMO).

A matriz anti-imagem, que apresenta as Medidas de Adequação da Amostra (MSA), representa outra forma de obtenção de indícios sobre a necessidade de eliminação de uma ou mais variáveis. Se uma variável apresentar baixo valor na diagonal principal e alto valor fora dela, talvez haja necessidade de excluí-la do modelo.

Dentre as formas de se determinar a quantidade de fatores, destacam-se o critério da raiz latente (critério de Kaiser), o critério *a priori*, o critério de percentagem de variância e o critério do gráfico Scree.

É conhecida por comunalidade a proporção da variância de cada variável explicada pelo conjunto dos fatores selecionados, sendo calculada pela somatória dos quadrados da carga de fator na variável.

A matriz de componentes apresenta as cargas fatoriais, ou *loadings coefficients*, que indicam a importância de cada fator em cada variável. A matriz de coeficientes dos escores fatoriais apresenta a importância, ou peso, de càda variável em cada fator.

As rotações não alteram a correlação entre as variáveis, mas tornam os fatores mais facilmente interpretáveis. O objetivo da rotação é fazer com que cada variável apresente altas cargas em apenas um fator e baixas cargas nos demais fatores, facilitando sua interpretação (matriz dos componentes após a rotação) por parte do pesquisador. As principais formas de rotação ortogonal são a Varimax, a Quartimax e a Equamax, sendo a primeira a mais usual por maximizar a variância entre as cargas de cada variável em cada fator. Além disso, merecem destaque, como métodos de rotação oblíqua, as formas conhecidas por Oblimin e Promax, que preservam as comunalidades, porém geram fatores mais fortemente correlacionados.



## 7.13. QUESTÕES COMPLEMENTARES

a) Defina análise fatorial.
b) Qual é o objetivo da análise fatorial?
c) Indique três situações de pesquisa em que poderia ser utilizada a técnica de análise fatorial.
d) Discorra brevemente sobre os seguintes conceitos: comunalidade, carga fatorial, *eigenvalue* e escore fatorial.
e) Qual o objetivo do teste de esfericidade de Bartlett e da estatística KMO? Quais são suas hipóteses subjacentes?
f) Qual a finalidade da rotação de fatores?
g) Quais as diferenças entre a rotação ortogonal e a oblíqua?
h) Como o pesquisador pode identificar e decidir o número de fatores a serem retidos pela análise fatorial?