ANÁLISE DE COMPONENTES PRINCIPAIS

E

ANÁLISE FATORIAL

Especialização *Data Science*

Prof. Adriana Kroenke, Dra.

2020

Análise de Componentes Principais

"A Análise de Componentes Principais (ACP) é uma técnica exploratória multivariada que transforma um conjunto de variáveis correlacionadas num conjunto menor de variáveis independentes".

Maroco, 2014, p. 441

Análise de Componentes Principais

Tem como objetivo analisar quais variáveis (conjunto) explicam a maior parte da variabilidade total, revelando que tipo de relacionamento existe entre elas.

Com *p*-variáveis originais é possível obter *p* componentes principais.

Maroco, 2014, p. 441

Quadro de Observações e Variáveis (QOV)

Cada coluna refere-se a uma variável e as linhas, aos casos/observações.

Observações	Variáveis $\mathbf{X_1}$ $\mathbf{X_2}$ … $\mathbf{X_n}$
1 2 ... m	x_{11} x_{12} … x_{1n} x_{21} x_{22} … x_{2n} ..... x_{m1} x_{m2} x_{mn}

Representação Matricial dos Quadro de Observações e Variáveis

O quadro de observações e variáveis fornece a matriz **X**

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

Cada coluna de **X** é um vetor que representa um ponto-variável do espaço IR^m.

$$\mathbf{x}_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \dots \\ x_{m1} \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ \dots \\ x_{m2} \end{bmatrix} \bullet\bullet\bullet \quad \mathbf{x}_n = \begin{bmatrix} x_{1n} \\ x_{2n} \\ \dots \\ x_{mn} \end{bmatrix}$$

tem-se a representação da matriz **X** através de suas colunas:

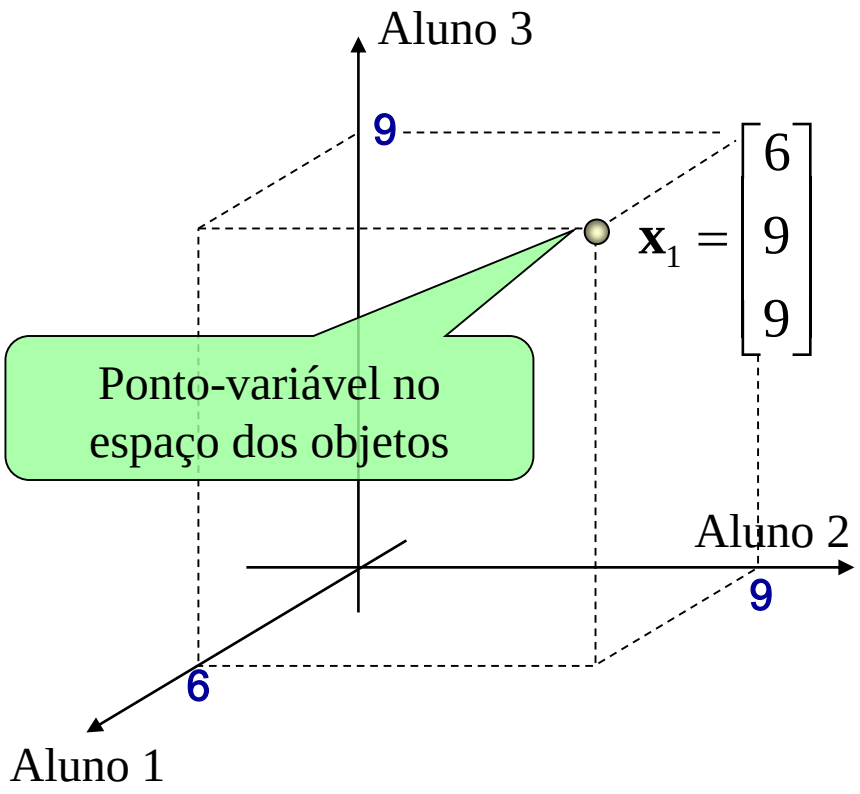
$$\mathbf{X} = [\mathbf{x}_1 \ \ \mathbf{x}_2 \ \ \dots \mathbf{x}_m \ ]$$

Pontos-Variáveis no IR^3

QOV	X1 (Port)	X2 (Cienc)	X3 (Matem)
Aluno 1	6	8	10
Aluno 2	9	4	5
Aluno 3	9	8	7

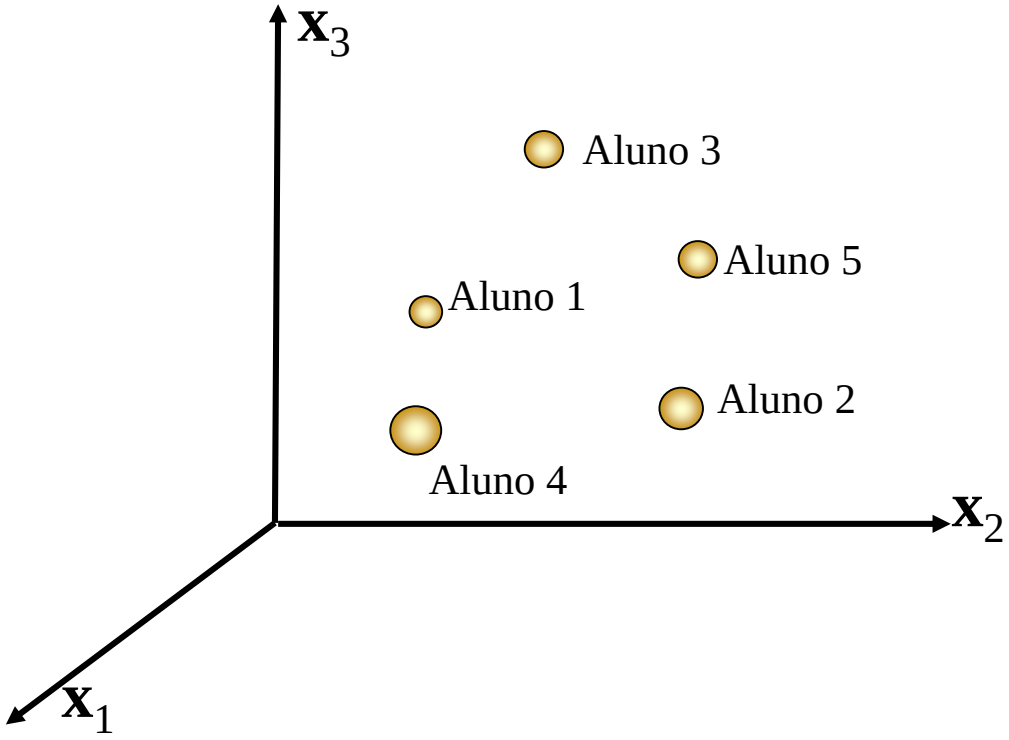
O QOV acima fornece a representação dos dados na matriz abaixo

$$\mathbf{X} = \begin{bmatrix} 6 & 8 & 10 \\ 9 & 4 & 5 \\ 9 & 8 & 7 \end{bmatrix}$$



Nuvem de Pontos-Objetos

Imagine que 5 alunos tenham notas em 3 disciplinas. Os alunos constituem uma nuvem de pontos.



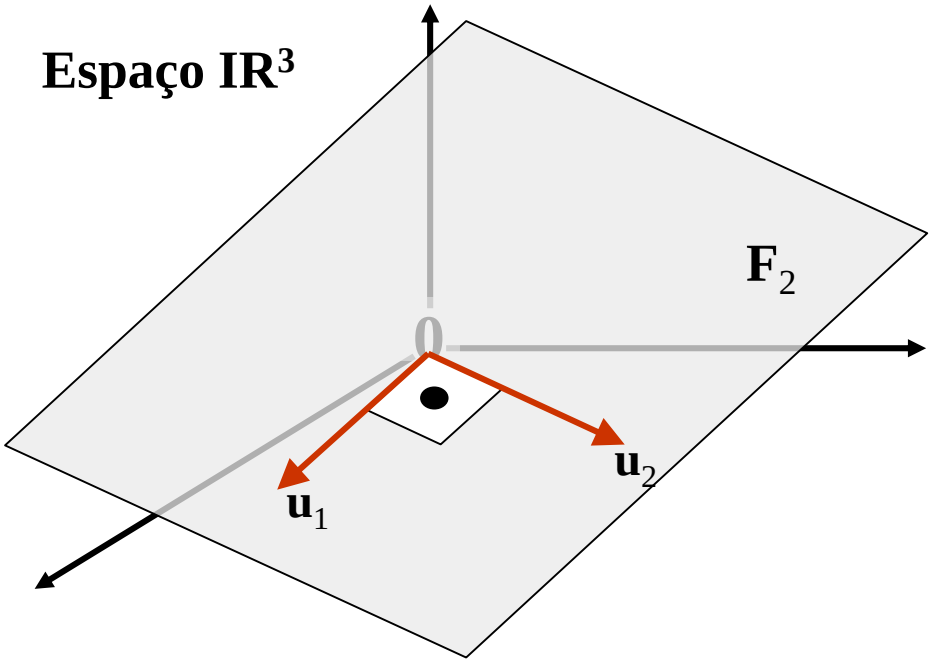
Subespaço de Duas Dimensões

Parte-se de um espaço vetorial com seu respectivo sistema de coordenadas.

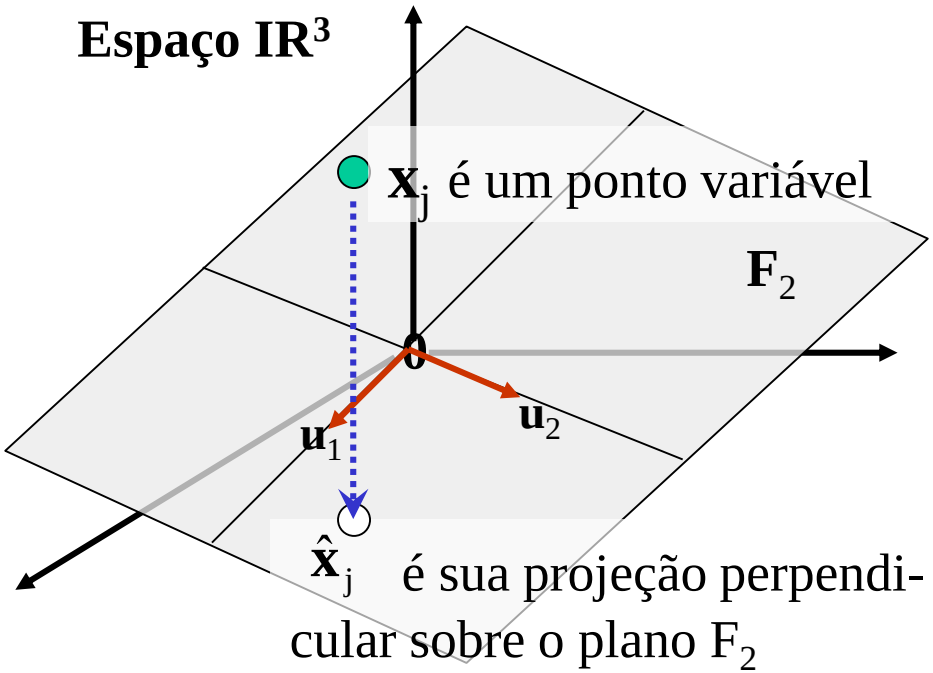
Dados dois vetores $\mathbf{u}_1$, $\mathbf{u}_2$ que

- são perpendiculares entre si;
- possuem comprimento unitário;

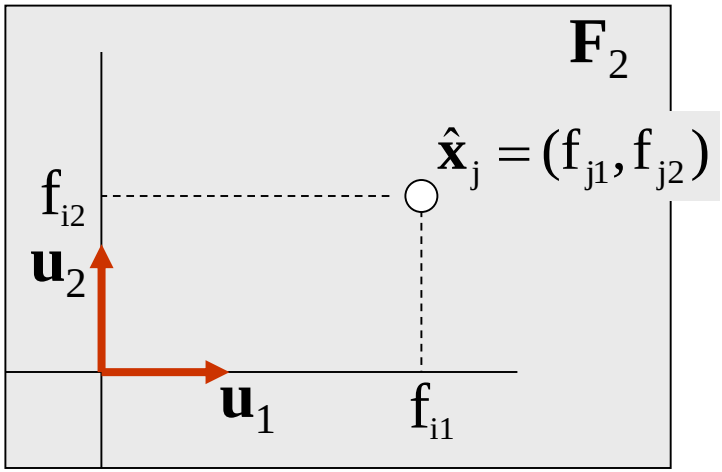
Estes vetores determinam um plano F_2 único que contém $\mathbf{u}_1$ e $\mathbf{u}_2$
Os vetores $\mathbf{u}_1$, $\mathbf{u}_2$ também são chamados de ***direções*** que determinam o plano F_2.



Projeção de um Ponto-Variável no Plano F_2



Observando o plano F_2, vê-se apenas a projeção

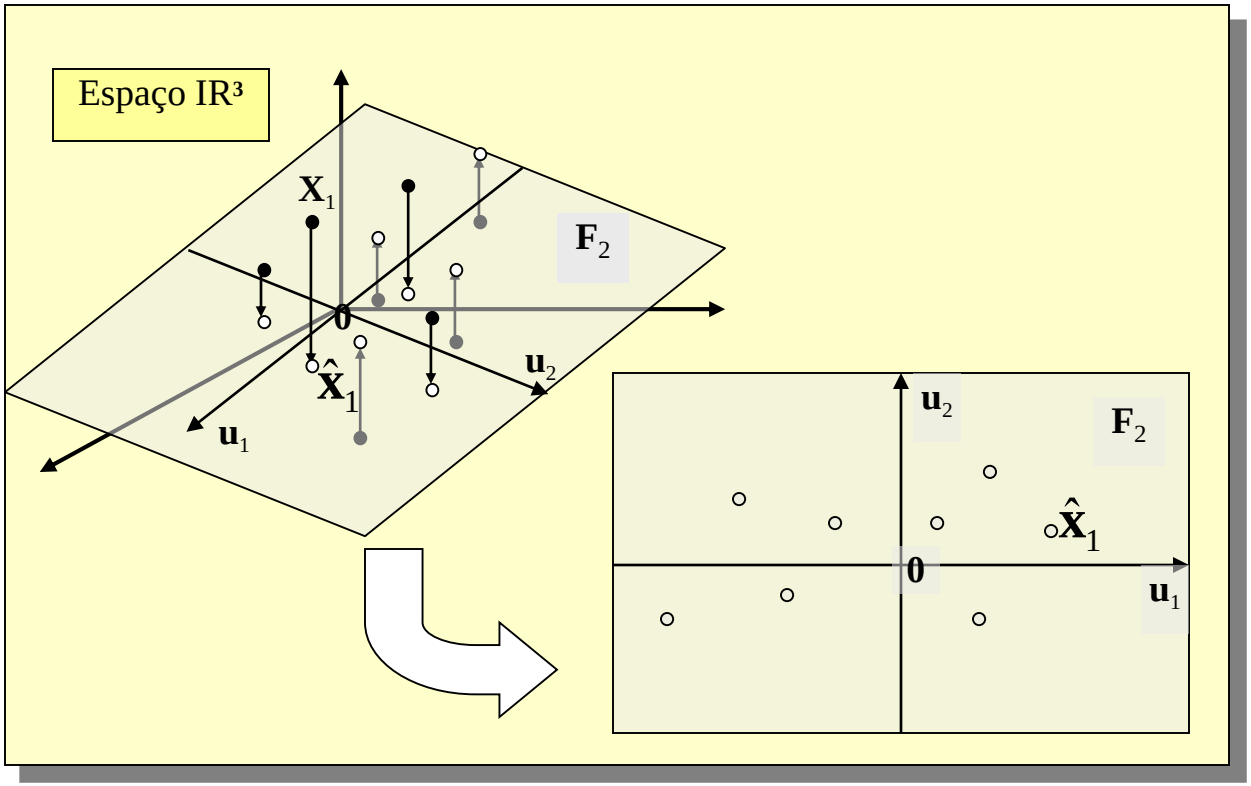


Para representar o ponto projetado no plano F_2 é necessário conhecer suas coordenadas f_{j1} e f_{j2}, as quais são calculadas pelos seguintes produtos escalares:

$$f_{j1} = \mathbf{u}_1'\mathbf{x}_j \text{ e } f_{j2} = \mathbf{u}_2'\mathbf{x}_j$$

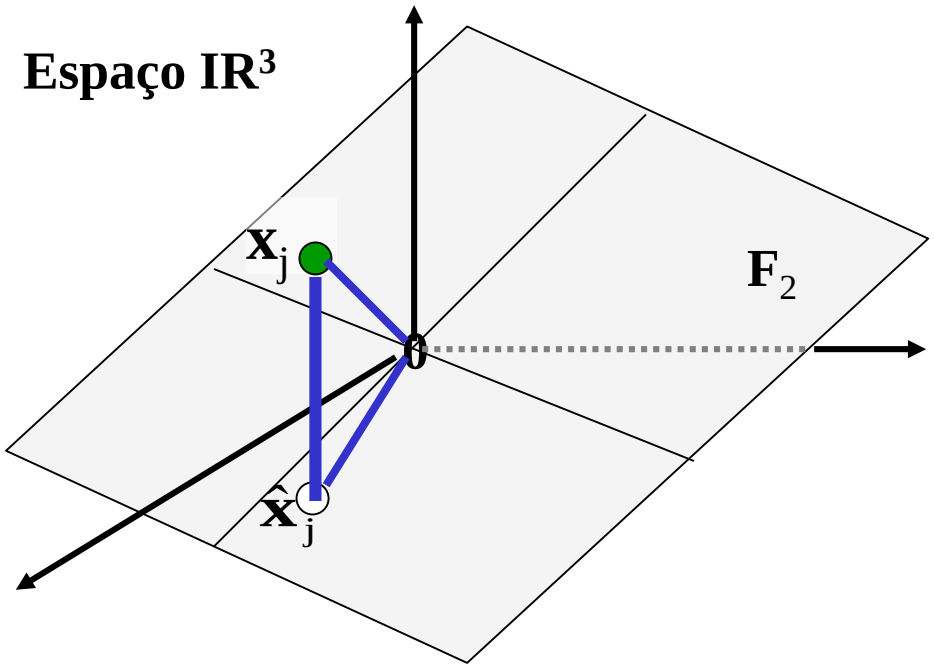
Projeção da Nuvem de Pontos-Variáveis

Nuvem de pontos-variáveis $\{\mathbf{x}_1 \quad \mathbf{x}_2 \quad \ldots \quad \mathbf{x}_n\}$

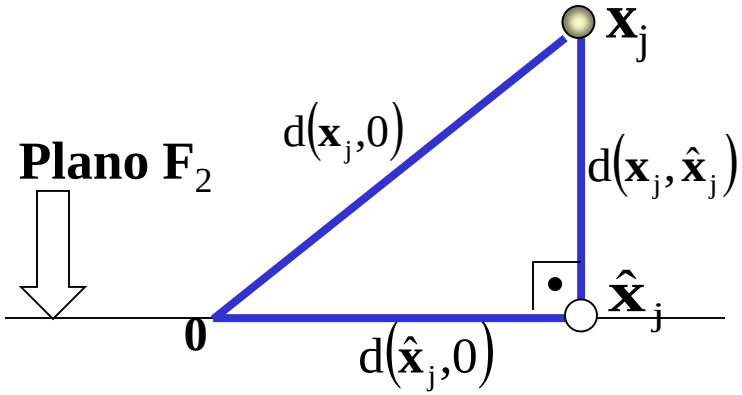


No plano F_2 observa-se as projeções da nuvem de pontos.

Inércias



Visto de outro ângulo ...



Teorema de Pitágoras: $d(\mathbf{x}_j,0)^2 = d(\hat{\mathbf{x}}_j,0)^2 + d(\mathbf{x}_j,\hat{\mathbf{x}}_j)^2$

Inércias

Teorema de Pitágoras: $d(\mathbf{x}_j,0)^2 = d(\hat{\mathbf{x}}_j,0)^2 + d(\mathbf{x}_j,\hat{\mathbf{x}}_j)^2$

aplica-se a todos os *n* pontos da nuvem. Então, somando-se tudo:

$$\sum_{j=1}^{n} d(\mathbf{x}_j,0)^2 = \sum_{j=1}^{n} d(\hat{\mathbf{x}}_j,0)^2 + \sum_{j=1}^{n} d(\mathbf{x}_j,\hat{\mathbf{x}}_j)^2$$

Inércia total I($\mathbf{X}$,0)

Inércia explicada I($\hat{\mathbf{X}}$,0)

Inércia não-explicada I($\mathbf{X}$,$\hat{\mathbf{X}}$)

Inércias

Tem-se, assim:

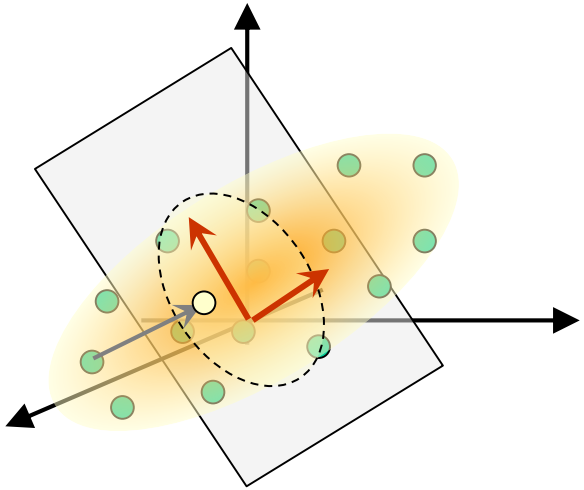
Inércia total = ***Inércia explicada*** + ***Inércia não-explicada***

$$I(\mathbf{X},0) = I(\hat{\mathbf{X}},0) + I(\mathbf{X},\hat{\mathbf{X}})$$

• A ***inércia total*** depende apenas da nuvem de pontos **X**: não depende da escolha do plano F_2, pois é calculada apenas sobre a distância entre os pontos no espaço e a origem.

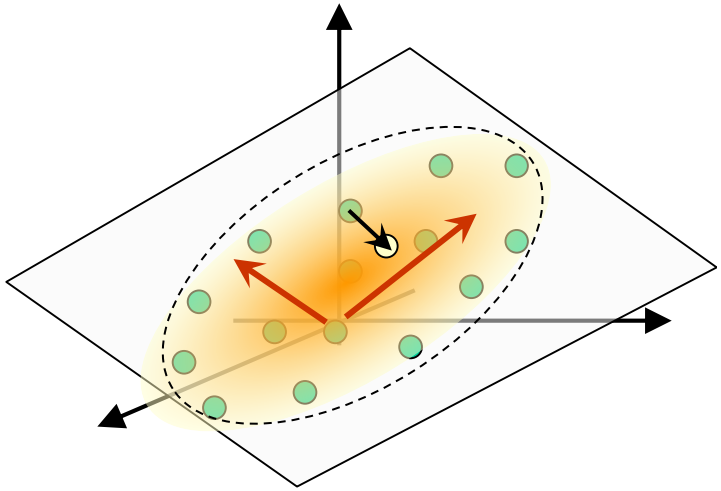
• As inércias ***explicada*** e ***não-explicada*** dependem da escolha do plano F_2 de projeção: quando uma aumenta a outra diminui.

A mesma nuvem do pontos (verdes) é vista de dois distintos planos sobre os quais os pontos da nuvem são perpendicularmente projetados. A elipse pontilhada representa a região de espalhamento do pontos projetados.



Pouca dispersão entre as projeções dos pontos da nuvem => inércia explicada baixa;

pontos distantes do plano => inércia não-explicada elevada.



Maior dispersão entre as projeções dos pontos da nuvem => inércia explicada elevada;

pontos próximos do plano de projeção => inércia não-explicada mais baixa.

Problema da ACP

O problema da ACP consiste em encontrar as direções principais $\boldsymbol{u_1}$ e $\boldsymbol{u_2}$ que determinam o plano de projeção da nuvem de pontos de modo a maximizar a inércia explicada, ou, o que é equivalente, a minimizar a inércia não-explicada.

Obtenção das Direções Principais

Ao considerar os pontos-variáveis no seu espaço de representação, prova-se que as direções principais, que maximizam a inércia explicada são obtidas por encontrar os ***autovetores*** e ***autovalores*** da matriz de inércia.

Ou seja: devem-se encontrar os valores do escalares λ e vetores $\boldsymbol{u}$ que satisfazem a relação

Obtenção das Direções Principais

Existem exatamente *m* autovalores λ_1, λ_2,...,λ_m que satisfazem a relação.

A soma dos autovalores é igual ao traço da matriz, de forma que se tem

$$\lambda_1 + \lambda_2 + ... + \lambda_m = \text{tr}(\mathbf{X}.\mathbf{X}') = \mathbf{I(X,0)} \text{ (}\textit{inércia total}\text{)}$$

Para cada autovalor existem autovetores associados, de forma que é possível encontrar-se *m* autovetores unitários correspondentes $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_m$ (*unitário* significa que o comprimento do vetor é 1, isto é, $|\mathbf{u}_j| = 1$).

Como a matriz de inércia $\mathbf{XX}'$ é simétrica, os autovalores λ_1, λ_2,...,λ_m são positivos e os autovetores $\mathbf{u}_1$, $\mathbf{u}_2$,...,$\mathbf{u}_m$ são vetores perpendiculares entre si.

Obtenção das Direções Principais

Os autovalores serão dispostos na ordem $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. Escolhem-se os maiores autovalores e seus autovetores associados. Sejam λ_1, λ_2 os dois maiores autovalores e $\mathbf{u}_1$, $\mathbf{u}_2$ e seus correspondentes autovetores. Estes são os dois autovetores que determinam as direções principais, pois a soma dos autovalores correspondentes $\lambda_1+\lambda_2$ fornece a inércia explicada máxima, ou seja,

$$\underbrace{I(X,0) = tr(X.X')}_{\text{Inércia total}} = \underbrace{\lambda_1 + \lambda_2}_{\text{Inércia explicada}} + \underbrace{\lambda_3 + \ldots + \lambda_m}_{\text{Inércia não-explicada}}$$

Obtenção das Direções Principais

$$\underbrace{I(X,0) = tr(X.X')}_{\text{Inércia total}} = \underbrace{\lambda_1 + \lambda_2}_{\text{Inércia explicada}} + \underbrace{\lambda_3 + \ldots + \lambda_m}_{\text{Inércia não-explicada}}$$

Exemplo numérico para m = 5 : digamos que foram encontrados $\lambda_1 = 6$; $\lambda_2 = 2$; $\lambda_3 = 1$, $\lambda_4 = 0{,}7$; $\lambda_5 = 0{,}3$. Então:

Inércia total = $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 = 6 + 2 + 1 + 0{,}7 + 0{,}3 = 10$;
Inércia explicada = $\lambda_1 + \lambda_2 = 6 + 2 = 8$;
Inércia não-explicada = $\lambda_3 + \lambda_4 + \lambda_5 = 1 + 0{,}7 + 0{,}3 = 2$.

Neste caso, 80% da inércia total será explicada pelos pontos projetados no plano fatorial.

Correlações de Variáveis

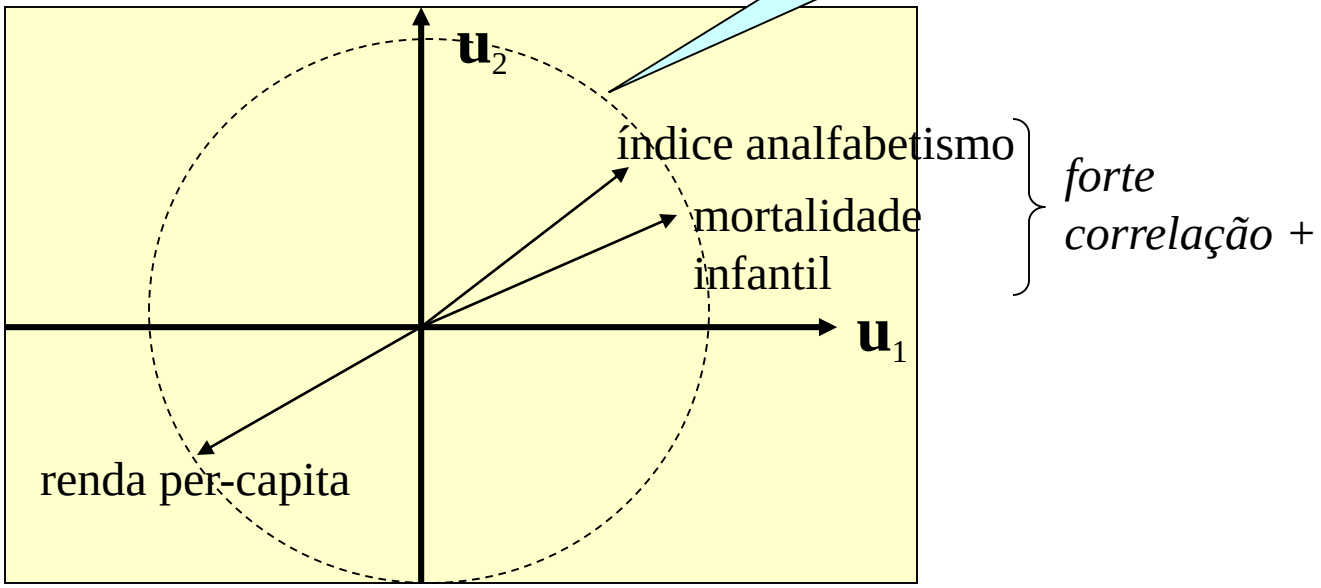
Grau de correlação de acordo com o ângulo entre eles formado:

perto de 0 graus: *forte correlação positiva*;

perto de 180 graus: *forte correlação negativa*;

perto de 90 graus: *correlação inexistente*.

círculo de correlações



Análise de Componentes Principais

As Componentes Principais podem ser obtidas por meio da matriz de covariâncias ou matriz de correlações.

KMO	Análise de componentes principais
1,00 - 0,90	Muito boa
0,80 - ,090	Boa
0,70 - 0,80	Média
0,60 - 0,70	Razoável
0,50 - 0,60	Má
< 0,50	Inaceitável

Fonte: Pereira (2004, p. 99)

Análise de Componentes Principais

Exemplo

Vamos supor um estudo etológico sobre relações sociais em primatas. Um pesquisador analisou dois grupos de chimpanzés quanto aos comportamentos sociais de cada elemento do grupo e classificou cada um dos elementos e diversas tipologias sociais.

Variáveis: intensidade da interação, frequência da interação, proximidade física, formalidade e sentimento de pertença.

Análise de Componentes Principais

Exemplo

Resolvendo...

Análise de Componentes Principais

Exercício.

Determinar a melhor marca de coxinha usando ACP.

Análise de Componentes Principais

Exercício.

Considere dados de 12 empresas (Arquivo AVA – ACP Exercício 2) no que se refere a 3 variáveis (unidades monetárias): Ganho bruto, ganho líquido e patrimônio acumulado. Determine as 3 componentes principais.

Referências

FÁVERO, et al. **Análise de dados: modelagem multivariada para tomada de decisões**. Rio de Janeiro: Elsevier, 2009.

HAIR, Joseph F. et al. **Análise multivariada de dados**. Bookman, 2009, 688 p.

MAROCO, João. **Análise estatística com utilização do SPSS**. Lisboa: Sílabo, 2007.

MINGOTI, Sueli Aparecida. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: UFMG, 2007.

PEREIRA, Alexandre. **Guia prático de utilização**: análise de dados para ciências sociais e psicologia. 5. ed. Lisboa: Sílabo, 2004.

Análise Fatorial

Análise Fatorial

É uma das técnicas multivariadas mais conhecidas e tem sido muito utilizada em áreas como química, educação, geologia, marketing, entre outras.

Exemplos:

- identificação do perfil dos consumidores ou os fatores que os levam a comprar;
- análise do posicionamento de produtos e serviços perante os concorrentes de mercado;
- elaboração de índices diferenciados de qualidade.

Análise Fatorial

“A Análise Fatorial” é uma técnica estatística que busca, através da avaliação de um conjunto de variáveis, a identificação de dimensões de variabilidade comuns existentes em um conjunto de fenômenos”.

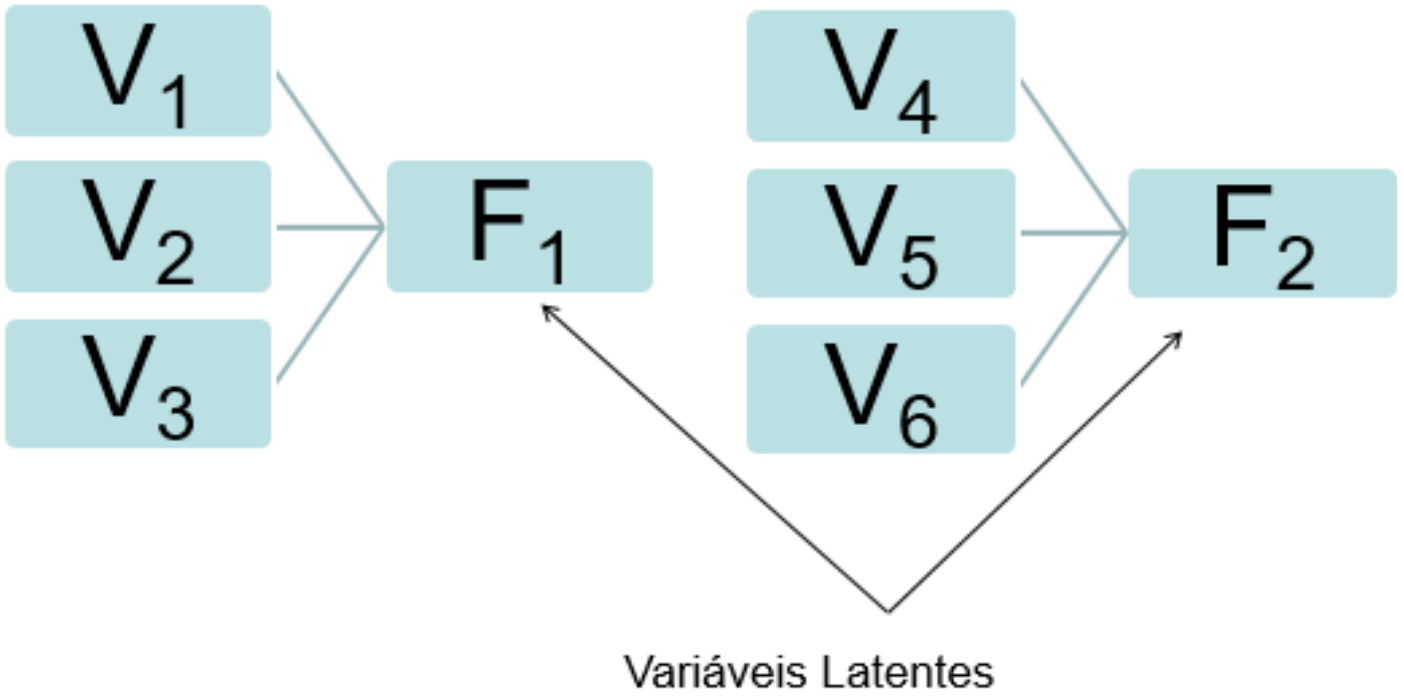
“O intuito é desvendar estruturas existentes, mas que não observáveis diretamente. Cada uma dessas dimensões de variabilidade comum recebe o nome de FATOR”.

(CORRAR; PAULO; DIAS FILHO, 2009, p. 74)

Análise Fatorial

- Por exemplo, o que leva um cliente a voltar a utilizar os serviços de uma empresa?
- Sua fidelidade está ligada ao preço?
- Há qualidade na prestação do serviço?
- Fidelidade está ligada ao atendimento?

Análise Fatorial



Análise Fatorial Exploratória e Confirmatória

- A modalidade de AF mais utilizada é a exploratória.

- Não exige do pesquisador o conhecimento prévio da relação de dependência entre as variáveis.

- O pesquisador não tem certeza de que as variáveis possuem uma estrutura de relacionamento, nem se essa estrutura pode ser interpretada de forma coerente.

Análise Fatorial Exploratória e Confirmatória

- Na AF confirmatória o pesquisador já parte de uma hipótese de relacionamento preconcebida entre um conjunto de variáveis e alguns fatores latentes.

- A AFC pretende confirmar se a teoria que sustenta a hipótese de relacionamento está correta ou não.

(CORRAR; PAULO; DIAS FILHO, 2009)

Preparação para Análise Fatorial

Antes de utilizar a AF, o pesquisador deve fazer algumas escolhas, que serão influenciadas pelo tipo de pesquisa que está sendo implementada. São elas:

a) Qual o método de extração dos fatores a ser utilizado?

b) Que tipo de análise será realizada?

c) Como será feita a escolha dos fatores?

d) Como aumentar o poder de explicação da AF?

Preparação para Análise Fatorial

Método de Extração:

Componentes Principais

É o método mais comum, pelo qual se procura uma combinação linear entre as variáveis, para que o máximo de variância seja explicado por essa combinação.

Esse procedimento resulta em fatores ortogonais, não correlacionados entre si.

Escolha dos Fatores

- A escolha do número de fatores é fundamental na AF;
- Como os fatores têm por objetivo a sumarização ou substituição do conjunto de variáveis, é natural que o número de fatores seja inferior ao número de variáveis;
- No entanto, ao preferir os fatores, ao invés de trabalhar com todas as variáveis, o pesquisador opta em não tratar 100% da variância observada, mas apenas uma parcela da variação total que consegue ser explicada pelos fatores.

(CORRAR; PAULO; DIAS FILHO, 2009)

Escolha dos Fatores

- A escolha do número de fatores determinará a capacidade de extrapolação das inferências que serão realizadas pela análise dos fatores.
- Limitar demais o número de fatores pode prejudicar as inferências, dada a explicação de uma parcela muito pequena da variância total dos dados pelos fatores.
- Por outro lado, um número grande de fatores pode eliminar o benefício da sumarização ou criar um problema por tratar um número muito grande de informação.

(CORRAR; PAULO; DIAS FILHO, 2009)

Escolha dos Fatores

Existem diversas técnicas para definição do número de fatores.

As principais são:

a) Critério do autovalor;
b) Critério do gráfico de declive ou *scree plot;*
c) Porcentagem da variância explicada

(CORRAR; PAULO; DIAS FILHO, 2009)

Critério do Autovalor

Apenas os fatores com autovalores acima de 1,0 são considerados.

O autovalor (*eigenvalue*) corresponde a quanto o fator consegue explicar da variância, ou seja, quanto da variância total dos dados pode ser associada ao fator.

(CORRAR; PAULO; DIAS FILHO, 2009)

Critério do Autovalor

Como se trabalha com dados padronizados, cada variável tem média zero e variância igual a 1,0.

Isso significa que fatores com autovalores abaixo de 1,0 são menos significativos do que uma variável original.

Esse critério também é denominado de critério da raiz latente ou critério *Kaiser* (*Kaiser test*).

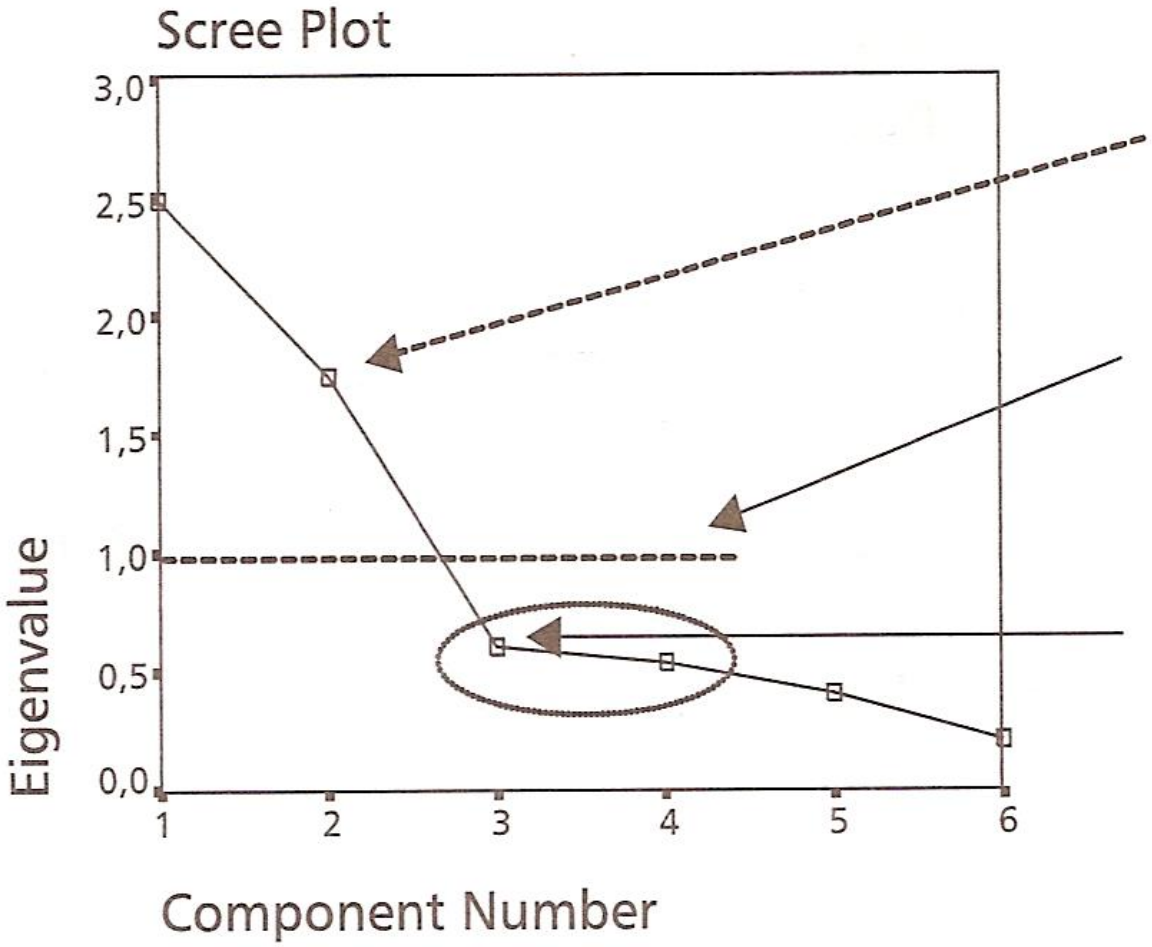
Scree Plot

- Por este critério, procura-se no gráfico um “ponto de salto”, que estaria representando um decréscimo de importância em relação à variância total (MINGOTI, 2009).

- Essa forma segue o raciocínio de que grande parcela da variância será explicada pelos primeiros fatores e que entre eles haverá sempre uma diferença significativa.

(CORRAR; PAULO; DIAS FILHO, 2009)

Scree Plot



Pelo *Kaiser test* seriam dois fatores

Limite para o *Kaiser test*

Pelo *scree plot* seriam três fatores

Porcentagem da Variância Explicada

- Considera-se o percentual de explicação da variância.

- O número de fatores a ser extraído é aquele que explica um percentual de variância considerado adequado pelo pesquisador.

- Se o pesquisador acredita que seu trabalho deve ser realizado com no mínimo 80% de variância explicada, então o número de fatores a serem escolhidos será aquele que permita explicar esse percentual de variação.

(CORRAR; PAULO; DIAS FILHO, 2009)

Como aumentar o Poder de Explicação da AF

- Busca-se soluções que expliquem o mesmo grau de variância total, mas que gerem resultados melhores em relação à sua interpretação.

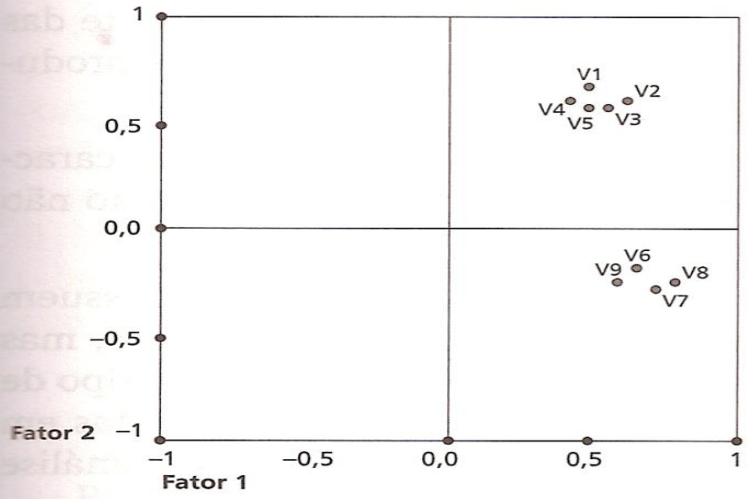
- Para isso usa-se rotação dos fatores.

- Métodos: Varimax, Quartimax, Equimax, Promax, etc.

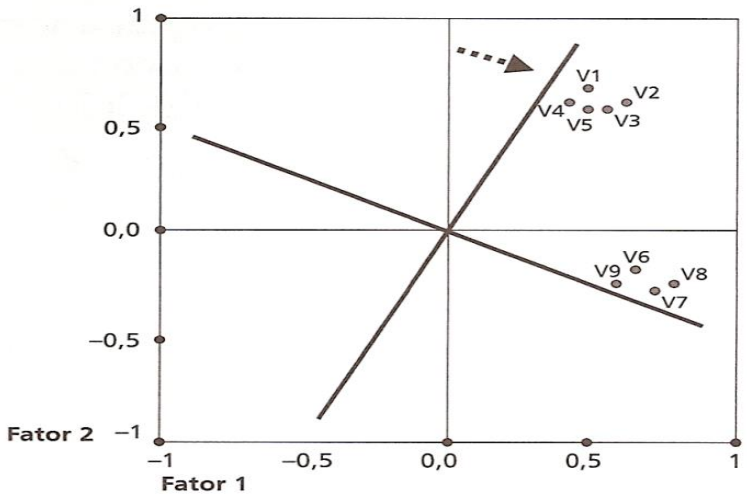
(CORRAR; PAULO; DIAS FILHO, 2009)

Como aumentar o Poder de Explicação da AF

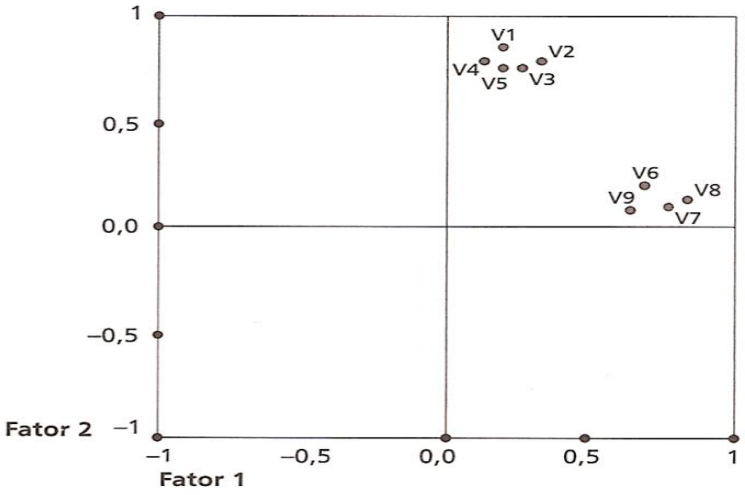
a) Fatores não rotacionados



(b) Rotação dos eixos



(c) Rotação realizada



Rotação Varimax

É um tipo de rotação ortogonal (mantém os fatores perpendiculares entre si, ou seja, sem correlação entre eles).

É o tipo de rotação mais utilizado e tem como característica minimizar a ocorrência de uma variável possuir altas cargas fatoriais para diferentes fatores, permitindo que uma variável seja facilmente identificada com um único fator.

(CORRAR; PAULO; DIAS FILHO, 2009)

Rotação Quartimax

Rotação ortogonal que minimiza o número de fatores necessários para explicar cada variável.

Tende a concentrar grande parte das variáveis em um único fator. Em função disso, esse método pode produzir estruturas de difícil interpretação.

(CORRAR; PAULO; DIAS FILHO, 2009)

Rotação Equimax

Rotação também ortogonal que tenta agregar tanto as características da rotação Varimax quanto da Quartimax.

Sua utilização não é comum.

Ainda, a Direct Oblimin e Promax que são rotações oblíquas e aumentam a complexidade.

(CORRAR; PAULO; DIAS FILHO, 2009)

Comunalidades

Outro conceito importante nos resultados produzidos pela AF são as comunalidades.

Elas representam o percentual de explicação que uma variável obteve pela AF, ou seja, quanto todos os fatores juntos são capazes de explicar uma variável.

Quanto mais próximo de 1 estiverem as comunalidades, maior é o poder de explicação dos fatores.

(CORRAR; PAULO; DIAS FILHO, 2009)

Passos para AF

1. Cálculo da matriz de correlação: é avaliado o grau de relacionamento entre as variáveis e a conveniência da aplicação da AF;
2. Extração dos fatores: determinação do método para cálculo dos fatores e definição do número de fatores a serem extraídos. Nesta etapa busca-se descobrir o quanto o modelo escolhido é adequado para representar os dados;
3. Rotação dos fatores: etapa na qual se busca dar maior capacidade de interpretação dos fatores.
4. Cálculo dos escores: os escores resultantes desta fase podem ser utilizados em diversas outras análises (discriminante, *cluster*, regressão logística, etc.)

(CORRAR; PAULO; DIAS FILHO, 2009)

Passos para AF

KMO	Análise de componentes principais
1,00 - 0,90	Muito boa
0,80 - ,090	Boa
0,70 - 0,80	Média
0,60 - 0,70	Razoável
0,50 - 0,60	Má
< 0,50	Inaceitável

Fonte: Pereira (2004, p. 99)

Teste de Esfericidade de Bartlett

Exemplo

Vamos a um exemplo no SPSS

Um analista de mercado quer estudar as relações estruturais entre 4 indicadores de 45 empresas.

- Prazo médio de recebimento de Vendas (PMRV, em dias)
- Endividamento (em %)
- Vendas (em R$ x mil)
- Margem líquida das vendas (em %)

AF *versus* ACP

Ambas são técnicas exploratórias multivariadas e ambas permitem a representação das variáveis originais num número mais reduzido de componentes/fatores.

A ACP não é um método de AF.

O objetivo da ACP é resumir a informação presente nas variáveis originais num número reduzido de índices que explicam o máximo possível de variância das variáveis originais.

A AF tem como objetivo identificar os fatores latentes que explicam as intercorrelações observadas nas variáveis originais.

MAROCO (2003, P. 291-292)

AF *versus* ACP

- As CP são combinações lineares ponderadas das variáveis originais enquanto os Fatores são variáveis não diretamente observáveis que hipoteticamente explicam as correlações observadas entre as variáveis originais.

- A ênfase da ACP é o "resumo" da variância dos dados originais.

- A ênfase da AF é sobre a explicação da covariância/correlação entre variáveis.

MAROCO (2003, P. 291-292)

Exercício

Referências

FÁVERO, et al. **Análise de dados: modelagem multivariada para tomada de decisões**. Rio de Janeiro: Elsevier, 2009.

HAIR, Joseph F. et al. **Análise multivariada de dados**. Bookman, 2009, 688 p.

MAROCO, João. **Análise estatística com utilização do SPSS**. Lisboa: Sílabo, 2007.

MINGOTI, Sueli Aparecida. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: UFMG, 2007.

PEREIRA, Alexandre. **Guia prático de utilização**: análise de dados para ciências sociais e psicologia. 5. ed. Lisboa: Sílabo, 2004.