

Técnicas Multivariadas em Saúde

Lupércio França Bessegato
Dep. Estatística/UFJF

Técnicas Multivariadas em Saúde - 2015

Roteiro

1. Introdução
2. Distribuições de Probabilidade Multivariadas
3. Representação de Dados Multivariados
4. Testes de Significância $c/$ Dados Multivariados
5. Análise de Componentes Principais
6. Análise Fatorial
7. Análise de Correlação Canônica
8. Análise de Conglomerados
9. Análise Discriminante
10. Análise de Correspondência
11. Referências

Técnicas Multivariadas em Saúde - 2015

Análise Fatorial

Técnicas Multivariadas em Saúde - 2015

Análise Fatorial

- **Objetivo:**
 - √ Descrever as relações de covariância entre muitas variáveis em termos de poucas quantidades aleatórias subjacentes e não observáveis
- **Motivação:**
 - √ Variáveis de um grupo altamente correlacionadas entre si, mas com pequenas correlações de outros grupos
 - √ É concebível que cada grupo de variáveis represente um fator (ou construto) que seja o responsável pelas correlações observadas

Técnicas Multivariadas em Saúde - 2015

- **Análise fatorial:**

- √ Pode ser considerada uma extensão da Análise de Componentes Principais

- Ambas são tentativas de aproximar S .
- A aproximação baseada em Análise Fatorial é mais elaborada

- √ **Questão principal:**

- Dados são consistentes com a estrutura prescrita?

Técnicas Multivariadas em Saúde - 2015

- **Análise Fatorial Exploratória:**

- √ Busca encontrar os fatores subjacentes às variáveis originais amostradas

- √ Em geral, efetuada quando não se tem noção clara da quantidade de fatores do modelo e nem do que representam

- **Análise Fatorial Confirmatória:**

- √ Tem-se em mãos um modelo fatorial pré-especificado (modelo hipotético) e deseja-se verificar se é aplicável ou consistente com os dados amostrais de que dispõe

Técnicas Multivariadas em Saúde - 2015

Modelo Fatorial Ortogonal via Matriz de Correlações

- Seja o vetor aleatório

$$\mathbf{X}' = [X_1, X_2, \dots, X_p].$$

com vetor de médias $\boldsymbol{\mu}$, matriz de covariâncias é $\boldsymbol{\Sigma}$, e matriz de correlações \mathbf{P} .

- Sejam as variáveis originais padronizadas:

$$Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}$$

- √ \mathbf{P} é a matriz de covariâncias do vetor aleatório \mathbf{Z} , cujos componentes são as variáveis padronizadas

Técnicas Multivariadas em Saúde - 2015

- **Modelo Fatorial Ortogonal**

- √ Construído via a matriz de correlação populacional

- √ Relaciona linearmente as variáveis padronizadas e os m fatores comuns (que são desconhecidos)

- √ Fatores são variáveis independentes

Técnicas Multivariadas em Saúde - 2015

• Equações do modelo:

$$\begin{aligned} Z_1 &= l_{11}F_1 + l_{12}F_2 + \dots + l_{1m}F_m + \epsilon_1 \\ Z_2 &= l_{21}F_1 + l_{22}F_2 + \dots + l_{2m}F_m + \epsilon_2 \\ &\vdots \quad \vdots \quad \quad \quad \vdots \quad \quad \vdots \\ Z_p &= l_{p1}F_1 + l_{p2}F_2 + \dots + l_{pm}F_m + \epsilon_p \end{aligned}$$

√ Em notação matricial:

$$\mathbf{V}(\mathbf{X} - \boldsymbol{\mu}) = \mathbf{LF} + \boldsymbol{\epsilon}.$$

\mathbf{V} = diagonal $[\sigma_1, \sigma_2, \dots, \sigma_p]$.

$$\mathbf{L}_{p \times m} = \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1m} \\ l_{21} & l_{22} & \dots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \dots & l_{pm} \end{bmatrix} \cdot \mathbf{F}_{m \times 1} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} \cdot \boldsymbol{\epsilon}_{p \times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix}.$$

Técnicas Multivariadas em Saúde - 2015

• Modelo fatorial:

$$\mathbf{V}(\mathbf{X} - \boldsymbol{\mu}) = \mathbf{LF} + \boldsymbol{\epsilon}.$$

√ \mathbf{F} : vetor aleatório contendo m fatores

– Essas variáveis latentes precisam ser identificadas

√ $\boldsymbol{\epsilon}$: vetor dos erros aleatórios

– Erros de medida e variação de Z_i que não é explicada pelos fatores comuns

√ \mathbf{L} : matriz de loadings fatoriais

– l_{ij} : representa o grau de relacionamento entre Z_i e F_j .

√ O modelo de análise fatorial assume que as variáveis Z_i estão relacionadas linearmente com os fatores

– Variáveis originais padronizadas são representadas por p+m variáveis não observáveis

Técnicas Multivariadas em Saúde - 2015

Modelo de Fatores Ortogonais

• Suposições:

- i. Todos os fatores tem média zero $E[\mathbf{F}] = \mathbf{0}$.
- ii. Todos os fatores são não correlacionados e tem variância um. $\text{Cov}[\mathbf{F}] = \mathbf{I}_m$.
- iii. Todos os erros tem média igual a zero $E[\boldsymbol{\epsilon}] = \mathbf{0}$.
- iv. Erros são não correlacionados entre si e não necessariamente tem a mesma variância

$$\text{Cov}[\boldsymbol{\epsilon}] = \text{diagonal}(\psi_1, \psi_2, \dots, \psi_p). \quad \text{Var}[\epsilon_j] = \psi_j \\ \text{Cov}(\epsilon_i, \epsilon_j) = 0, \quad \forall i \neq j.$$

Técnicas Multivariadas em Saúde - 2015

v. Os vetores $\boldsymbol{\epsilon}$ e \mathbf{F} são independentes

$$\text{Cov}(\boldsymbol{\epsilon}_{p \times 1}, \mathbf{F}_{m \times 1}) = E[\boldsymbol{\epsilon}\mathbf{F}'] = \mathbf{0}.$$

√ \mathbf{F} e $\boldsymbol{\epsilon}$ são duas fontes de variação distintas, relacionadas às variáveis padronizadas Z_i , não havendo qualquer relacionamento entre estas fontes de informação.

• Assumido o modelo, \mathbf{P} pode ser reparametrizada

$$\mathbf{P}_{p \times p} = \mathbf{LL}' + \boldsymbol{\Psi}.$$

√ O objetivo é encontrar as matrizes $\mathbf{L}_{p \times m}$ e $\boldsymbol{\Psi}_{p \times p}$ que possam representar a matriz $\mathbf{P}_{p \times p}$.

– Há matrizes de correlação que não podem ser decompostas na forma do modelo

Técnicas Multivariadas em Saúde - 2015

• Consequências da decomposição fatorial de **P**:

√ Variância de Z_i é decomposta em duas partes:

$$\text{Var}[Z_i] = h_i^2 + \psi_i$$

$$\text{onde } h_i^2 = l_{i1}^2 + l_{i2}^2 + \dots + l_{im}^2.$$

- h_i^2 : comunalidade
 - variabilidade explicada pelos m fatores que é uma fonte comum de variação de Z_i .
- ψ_i : variância específica
 - Parte da variabilidade de Z_i associada apenas ao erro aleatório

Técnicas Multivariadas em Saúde - 2015

√ Covariâncias entre variáveis e fatores

$$\text{Cov}(Z_i, Z_k) = l_{i1}l_{k1} + l_{i2}l_{k2} + \dots + l_{im}l_{km}, \quad i, k = 1, 2, \dots, p, \quad i \neq k.$$

$$\text{Cov}(Z_i, F_j) = \text{Corr}(Z_i, F_j) = l_{ij}, \quad i = 1, 2, \dots, p \text{ e } j = 1, 2, \dots, m.$$

√ Proporção da variância total explicada pelo fator F_j :

$$\text{Proporção explicada}_{F_j} = \frac{\sum_{i=1}^p l_{ij}^2}{p}.$$

Técnicas Multivariadas em Saúde - 2015

Escolha do Número de Fatores m

• Critério 1:

√ Análise da proporção de variância total relacionada com cada autovalor

√ Permanecem aqueles autovalores que representam maiores proporções de variância total

√ m = quantidade de autovalores retidos

Técnicas Multivariadas em Saúde - 2015

• Critério 2:

√ Permanecem os autovalores maiores que 1

√ m = quantidade de autovalores retidos

√ Ideia básica do critério:

- Mantem no sistema nas novas dimensões pelo menos a informação da variância de uma variável original

Técnicas Multivariadas em Saúde - 2015

- Critério 3:
 - √ Observação do gráfico scree plot
 - √ Valor de m é igual ao número de autovalores anteriores ao 'ponto de salto'.
- Importante:
 - √ Uma escolha adequada do valor de m deve levar em consideração a interpretabilidade dos fatores
 - √ Deve-se observar também o princípio da parcimônia
 - Descrição da estrutura de variabilidade do vetor aleatório \mathbf{Z} com um número pequeno de fatores

Técnicas Multivariadas em Saúde - 2015

- Importante:
 - √ Modelo fatorial ortogonal só pode ser aplicado quando as variáveis originais são correlacionadas entre si
 - Caso contrário, cada fator ficará relacionado com apenas uma variável original

Técnicas Multivariadas em Saúde - 2015

Métodos de Estimação de \mathbf{L} e $\boldsymbol{\psi}$

- Escolhe-se o valor de m
- Métodos de estimação das matrizes \mathbf{L} e $\boldsymbol{\psi}$:
 - √ Método de componentes principais
 - Em geral, utilizado como um análise exploratória dos dados, em termos dos fatores subjacentes
 - √ Método de fatores principais
 - Refinamento do método das componentes principais
 - √ Método da máxima verossimilhança
 - Indicado apenas quando \mathbf{Z} tem distribuição normal

Técnicas Multivariadas em Saúde - 2015

Método das Componentes Principais

- Matrizes \mathbf{L} e $\boldsymbol{\psi}$ serão estimadas por:

$$\hat{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1} \hat{\mathbf{e}}_1, \sqrt{\hat{\lambda}_2} \hat{\mathbf{e}}_2, \dots, \sqrt{\hat{\lambda}_m} \hat{\mathbf{e}}_m \right].$$

$$\hat{\boldsymbol{\psi}} = \text{diagonal} \left(\mathbf{R}_{p \times p} - \hat{\mathbf{L}}_{p \times m} \hat{\mathbf{L}}'_{p \times m} \right).$$

- √ Aproximação de \mathbf{R}

$$\mathbf{R}_{p \times p} \approx \hat{\mathbf{L}}_{p \times m} \hat{\mathbf{L}}'_{p \times m} + \hat{\boldsymbol{\psi}}.$$

Técnicas Multivariadas em Saúde - 2015

- Matriz residual:

$$MRes = R_{p \times p} - (\hat{L}_{p \times m} \hat{L}'_{p \times m} + \hat{\psi})$$

- √ Pode servir como critério de avaliação do modelo
 - Seus valores deveriam ser próximos de zero
 - Matriz é nula somente quando o valor de m é igual a p
- √ Os elementos da diagonal da matriz R são reproduzidos exatamente pela reprodução do modelo
 - O mesmo não ocorre para os outros elementos da matriz R (covariâncias das variáveis Z_i e Z_j)

Técnicas Multivariadas em Saúde - 2015

- Método das componentes principais na estimação de LL' e ψ .

$$\text{Proporção explicada}_{F_j} = \frac{\sum_{i=1}^p l_{ij}^2}{p}$$

- √ Representa o quanto cada fator consegue captar da variabilidade original das variáveis Z_i .

Técnicas Multivariadas em Saúde - 2015

Exemplo 9.3 – Preferência de Consumidor

- Amostra aleatória de consumidores pontuando atributos de novo produto:
 - √ Respostas em escala semântica de 7 valores
 - √ X_1 : Gosto
 - √ X_2 : Preço
 - √ X_3 : Aroma
 - √ X_4 : Adequado para lanche
 - √ X_5 : Fornece muita energia
 - √ Dados: *BD_multivariada.xls/preferencia_consumidor*

Técnicas Multivariadas em Saúde - 2015

- Solução por Componentes Principais:

Variável	Loadings Estimados				Comunalidade	Variância Específica		
	F1	F2	F1	F2		h^2	ψ_i	
Gosto	0,331	0,560	0,607	0,816	$(0,560)^2 + (0,816)^2 =$	0,979	$1 - 0,979 =$	0,021
Preço	0,460	0,777	-0,390	-0,524	$(0,777)^2 + (-0,524)^2 =$	0,879	$1 - 0,879 =$	0,121
Aroma	0,382	0,645	0,557	0,748	$(0,645)^2 + (0,748)^2 =$	0,976	$1 - 0,976 =$	0,024
Adequado lanche	0,556	0,939	-0,078	-0,105	$(0,939)^2 + (-0,105)^2 =$	0,893	$1 - 0,893 =$	0,107
Energético	0,473	0,798	-0,404	-0,543	$(0,798)^2 + (-0,543)^2 =$	0,932	$1 - 0,932 =$	0,068
Autovalor	2,853		1,806					
	57,1%		33,2%					

- √ Em geral, uma rotação pode mostrar uma estrutura simples (interpretação simples)

Técnicas Multivariadas em Saúde - 2015

• Análise Fatorial - Minitab

Stat > Multivariate > Factor Analysis →

Factor Analysis

Variables: C1 X1, C2 X2, C3 X3, C4 X4, C5 X5, C7, C8, C9 h², C10 Psi, C11, C12

Number of factors to extract: 2

Method of Extraction: Principal components Maximum likelihood

Type of Rotation: None Equimax Varimax Quartimax Orthocon with gamma: _____

Factor Analysis - Options

Matrix to Factor: Correlation Covariance

Source of Matrix: Compute from variables Use matrix: M1

Storage

Loadings: C6-C7

Coefficients: _____

Scores: _____

Rotation matrix: _____

Residual matrix: M2

Eigenvalues: C8

Eigenvector matrix: M3

Técnicas Multivariadas em Saúde - 2015

• Análise Fatorial da Matriz de Correlações:

Factor Analysis: M1

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Communality
Var 1	0,560	0,816	0,979
Var 2	0,777	-0,524	0,879
Var 3	0,645	0,748	0,976
Var 4	0,939	-0,105	0,893
Var 5	0,798	-0,543	0,932
Variance	2,8531	1,8063	4,6594
% Var	0,571	0,361	0,932

$LL' + \Psi$ reproduz aproximadamente R

Técnicas Multivariadas em Saúde - 2015

• Cálculo Comunalidades e Resíduos – Minitab:

```

Let C9 = C6*C6 + C7*C7 # Comunalidades
Let C10 = 1 - C9 # Variâncias Específicas
Diagonal C10 M4 # Matriz ΨPsi
Subtract M4 M2 M5 # Matriz de resíduos
Print M5
    
```

• Matriz de Resíduos:

Data Display

Matrix M8

0,0000000	0,0126425	-0,0116968	-0,0201455	0,0064418
0,0126425	0,0000000	0,0204813	-0,0749273	-0,0551752
-0,0116968	0,0204813	0,0000000	-0,0275656	0,0011935
-0,0201455	-0,0749273	-0,0275656	0,0000000	-0,0165955
0,0064418	-0,0551752	0,0011935	-0,0165955	0,0000000

Técnicas Multivariadas em Saúde - 2015

Exemplo 9.4 – Ações New York

• Taxas de retorno de 5 ações negociadas na Bolsa de New York

- √ Período: jan/75 a Dez/76
 - Observadas 100 semanas
- √ Ações:
 - Allied Chemical
 - du Pont
 - Union Carbide
 - Exxon
 - Texaco
- √ Dados: *BD_multivariada.xls/acoes_NY*

Técnicas Multivariadas em Saúde - 2015

- Vetor de médias amostral (\bar{x})

```

Descriptive Statistics: allied_chemical; du_pont; union_carbide; Exxon; texaco
Variable      Mean
allied_chemical 0,00543
du_pont        0,00483
union_carbide  0,00565
Exxon         0,00629
texaco        0,00371
    
```

- Matriz de correlações amostral (S)

```

Correlations: allied_chemical; du_pont; union_carbide; Exxon; texaco
du_pont      allied_chemical  du_pont  union_carbide  Exxon
union_carbide 0,577          0,509    0,598
Exxon        0,387          0,390    0,436
texaco       0,462          0,322    0,426          0,524
    
```

Técnicas Multivariadas em Saúde - 2015

- Análise Fatorial - Solução por Componentes Principais

$$\sqrt{m} = 1$$

```

Factor Analysis: allied_chemical; du_pont; union_carbide; Exxon; texaco
Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities
Variable      Factor1  Communality
allied_chemical 0,783    0,614
du_pont        0,773    0,597
union_carbide  0,794    0,631
Exxon         0,713    0,508
texaco        0,712    0,507

Variance      2,8565   2,8565
% Var         0,571    0,571
    
```

$$\sqrt{m} = 2$$

```

Principal Component Factor Analysis of the Correlation Matrix

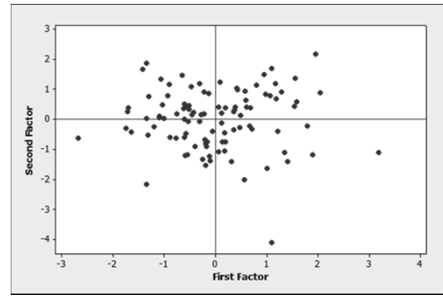
Unrotated Factor Loadings and Communalities
Variable      Factor1  Factor2  Communality
allied_chemical 0,783    0,217    0,661
du_pont        0,773    0,458    0,806
union_carbide  0,794    0,234    0,686
Exxon         0,713    -0,492    0,731
texaco        0,712    -0,524    0,781

Variance      2,8565   0,8091   3,6656
% Var         0,571    0,162    0,733
    
```

Técnicas Multivariadas em Saúde - 2015

- Scores dos fatores
- Plot dos fatores

F1	F2
0,145	-0,752
-0,121	-1,229
3,188	-1,110
1,182	0,677
-0,463	1,082
1,899	-1,182
-1,091	0,109
1,096	1,682
-0,581	0,406
-0,892	1,166
0,598	0,637
-1,345	1,876
2,043	0,876
1,581	0,589
0,127	-0,130
1,560	1,365
0,182	-1,064
-0,883	-0,613
-0,613	0,359
1,102	-4,108
-1,335	0,017
0,573	-2,011
0,987	0,834
-0,126	0,865
-1,644	-0,471



Técnicas Multivariadas em Saúde - 2015

- Comparação soluções:

Variável	Fatores Comuns				Solução			
	F1	F2	F1	F2	1 fator	2 fatores	1 fator	2 fatores
Allied Chemical	0,464	0,783	0,241	0,217	0,615	0,385	0,661	0,339
Du Pont	0,457	0,773	0,509	0,458	0,596	0,404	0,806	0,194
Union Carbide	0,470	0,794	0,261	0,234	0,630	0,370	0,685	0,315
Exxon	0,422	0,713	-0,525	-0,472	0,507	0,493	0,731	0,269
Texaco	0,421	0,712	-0,582	-0,524	0,507	0,493	0,780	0,220
Autovalor	2,856		0,809		57,1%		73,3%	

√ Componentes obtidos a partir de R

√ % acumulada da solução a dois fatores é bem maior que a da solução a um fator

Técnicas Multivariadas em Saúde - 2015

- Matriz de Correlações

Correlation				
1,00000	0,57692	0,50866	0,38672	0,46218
0,57692	1,00000	0,59838	0,38952	0,32195
0,50866	0,59838	1,00000	0,43610	0,42563
0,38672	0,38952	0,43610	1,00000	0,52353
0,46218	0,32195	0,42563	0,52353	1,00000

- Matriz de Resíduos: $\hat{R} - \hat{L}\hat{L}' - \hat{\Psi}$

√ $m = 1$

0	-0,028	-0,114	-0,172	-0,096
-0,028	0	-0,015	-0,161	-0,228
-0,114	-0,015	0	-0,130	-0,140
-0,172	-0,161	-0,130	0	0,016
-0,096	-0,228	-0,140	0,016	0

√ $m = 2$

0	-0,128	-0,164	-0,069	0,018
-0,128	0	-0,123	0,055	0,012
-0,164	-0,123	0	-0,019	0,017
-0,069	0,055	-0,019	0	-0,231
0,018	0,012	-0,017	-0,231	0

√ Para $m=2$ $\hat{L}\hat{L}'$ produz números maiores (principalmente r_{45})

Técnicas Multivariadas em Saúde - 2015

- Interpretação:

Principal Component Factor Analysis of the Correlation Matrix			
Unrotated Factor Loadings and Communalities			
Variable	Factor1	Factor2	Communality
allied_chemical	0,783	0,217	0,661
du_pont	0,773	0,458	0,806
union_carbide	0,794	0,234	0,686
exxon	0,713	-0,472	0,731
texaco	0,712	-0,524	0,781
Variance	2,8565	0,8091	3,6656
% Var	0,571	0,162	0,733

√ F_1 : fator de mercado
(condições econômicas gerais)

√ F_2 : fator industrial
– contrasta ações de indústrias químicas e de óleo e gás
(diferencia setores)

- Em essência, mesma conclusão de ACP (ex. 8.5)

Técnicas Multivariadas em Saúde - 2015

Método do Fatores Principais

- Também chamado Método de Componentes Principais Iterativo
- Idéia básica:
 - √ Refinar as estimativas de L_{pxm} e Ψ_{pxp} .

Técnicas Multivariadas em Saúde - 2015

- Procedimento
 - √ Estimativas iniciais pelo método das componentes principais
 - √ Troca dos elementos da diagonal de R pelas comunalidades estimadas
 - √ Novas estimações a partir da matriz R^*
 - √ Substituição dos elementos da diagonal principal pelas comunalidades estimadas
 - √ Procedimento é repetido até que as diferenças entre as comunalidades estimadas sejam desprezíveis

Técnicas Multivariadas em Saúde - 2015

Método da Máxima Verossimilhança

- Só pode ser utilizado quando a forma da distribuição de probabilidades é conhecida
- Suposição:
 - √ Vetor aleatório \mathbf{X} tem distribuição normal p-variada
 - √ Consequência:
 - Vetor das variáveis padronizadas é normal p-variado
 - Fatores tem distribuição normal multivariada com vetor de médias zero e matriz de covariâncias \mathbf{I}_m
 - Erros tem distribuição normal p-variada com vetor de médias zero e matriz de covariâncias $\boldsymbol{\psi}$.

Técnicas Multivariadas em Saúde - 2015

- A função de verossimilhança é expressa como:

$$L(\mathbf{0}, \mathbf{P}) = \frac{1}{(2\pi)^{np/2} |\mathbf{L}\mathbf{L}' + \boldsymbol{\psi}|^{n/2}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^n \mathbf{z}'_j (\mathbf{L}\mathbf{L}' + \boldsymbol{\psi})^{-1} \mathbf{z}_j \right\}.$$

- √ A função de verossimilhança depende da matrizes \mathbf{L} e $\boldsymbol{\psi}$, através da matriz de correlação \mathbf{P} .
- √ As estimativas de máxima verossimilhança de $\hat{\mathbf{L}}$ e $\hat{\boldsymbol{\psi}}$ são as matrizes \mathbf{L} e $\boldsymbol{\psi}$ que maximizam a função de verossimilhança.
- √ Maximização é feita por métodos numéricos
- √ Método mais sofisticado que os métodos de componentes e fatores principais
 - Produz estimativas mais precisas

Técnicas Multivariadas em Saúde - 2015

- Cuidados:

- √ Está fundamentado na suposição de normalidade multivariada dos vetores \mathbf{Z} , \mathbf{F} e $\boldsymbol{\epsilon}$.
 - Apenas a normalidade do vetor \mathbf{Z} pode ser investigada a priori a partir dos dados amostrais
 - Fatores e erros são variáveis aleatórias não observáveis

Técnicas Multivariadas em Saúde - 2015

- Valor de m:

- √ Método de máxima verossimilhança
 - Mudança de valor de m altera as estimativas dos loadings
- √ Método de componentes principais
 - Aumento no valor de m não altera os loadings para os fatores obtidos anteriormente
- √ Quando os dados provêm de distribuição normal multivariada
 - Usar método de componentes principais como análise exploratória dos fatores e estimação do valor provável de m
 - Posteriormente, qualidade da solução inicial poderá ser melhorada pelo uso do método de máxima verossimilhança

Técnicas Multivariadas em Saúde - 2015

- Dados omissos:
 - √ São considerados apenas os elementos amostrais com observações completas
(Análise de componentes principais e análise fatorial)
 - √ Caso haja muitas observações com dados omissos em algumas variáveis, deve-se avaliar até que ponto as análises são válidas.

Técnicas Multivariadas em Saúde - 2015

Exemplo 9.4 – Ações New York Solução Máxima Verossimilhança

- Taxas de retorno de 5 ações negociadas na Bolsa de New York
 - √ Período: jan/75 a Dez/76
 - Observadas 100 semanas
 - √ Ações:
 - Allied Chemical
 - du Pont
 - Union Carbide
 - Exxon
 - Texaco
 - √ Dados: *BD_multivariada.xls/acoes_NY*

Técnicas Multivariadas em Saúde - 2015

√ Solução por Máxima Verossimilhança

Factor Analysis: allied_chemical; du_pont; union_carbide; exxon; texaco

Maximum Likelihood Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Communality
allied_chemical	0,687	-0,176	0,503
du_pont	0,704	-0,506	0,751
union_carbide	0,685	-0,234	0,525
exxon	0,620	0,086	0,392
texaco	0,782	0,452	0,816
Variance	2,4338	0,5538	2,9876
% Var	0,487	0,111	0,598

√ Solução por Componentes Principais

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Communality
allied_chemical	0,783	0,217	0,661
du_pont	0,773	0,458	0,806
union_carbide	0,794	0,234	0,686
exxon	0,713	-0,472	0,731
texaco	0,712	-0,524	0,781
Variance	2,8565	0,8091	3,6656
% Var	0,571	0,162	0,733

Técnicas Multivariadas em Saúde - 2015

Variável	Máxima Verossimilhança				Componentes Principais			
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	h^2	Ψ_1	$\hat{\lambda}_1$	$\hat{\lambda}_2$	h^2	Ψ_1
Allied Chemical	0,687	-0,176	0,503	0,497	0,783	0,217	0,661	0,339
Du Pont	0,704	-0,506	0,751	0,249	0,773	0,458	0,806	0,194
Union Carbide	0,685	-0,234	0,525	0,475	0,794	0,234	0,685	0,315
Exxon	0,620	0,086	0,392	0,608	0,713	-0,472	0,731	0,269
Texaco	0,782	0,452	0,816	0,184	0,712	-0,524	0,780	0,220
Variância	2,434	0,554	2,988		2,856	0,809	3,663	
% Variância Total	48,7%	11,1%	59,8%		57,1%	16,2%	73,3%	

- √ Proporção da variância total amostral padronizada explicada é maior para a fatoração por componentes principais que por máxima verossimilhança
 - Componentes principais têm a propriedade de otimizar a variância
- √ F_1 : loadings positivos e grandes
 - não tanto quanto por componentes principais
- √ F_2 : sinais consistentes com contraste, mas magnitudes em alguns casos são menores
 - comparação entre Du Pont e Texaco

Técnicas Multivariadas em Saúde - 2015

- Matriz de Resíduos: $R - \hat{L}\hat{L}' - \hat{\Psi}$

√ Máxima Verossimilhança

m=2					
	0	-0,004	0,004	0,024	-0,004
	-0,004	0	0,003	0,004	0,000
	0,004	0,003	0	-0,031	0,005
	0,024	0,004	-0,031	0	0,000
	-0,004	0,000	0,005	0,000	0

√ Componentes Principais

	0	-0,128	-0,164	-0,069	0,018
	-0,128	0	-0,123	0,055	0,012
	-0,164	-0,123	0	-0,019	-0,017
	-0,069	0,055	-0,019	0	-0,231
	0,018	0,012	-0,017	-0,231	0

√ Elementos da matriz de resíduos são bem menores que aqueles obtidos pela análise fatorial por componentes principais

√ Escolha:

- Solução por máxima verossimilhança

Técnicas Multivariadas em Saúde - 2015

- Importante:

√ Os padrões dos *loadings* fatoriais iniciais estão restritos pela condição de unicidade da estimativa de $\hat{L}' \hat{\Psi} \hat{L} = \Delta$

√ Padrões fatoriais úteis frequentemente não são revelados até que os fatores sejam rotacionados

Técnicas Multivariadas em Saúde - 2015

Exemplo 9.6 – Declato Olímpico

- Estudo de escores olímpicos de decatlo (Linden, 1977)
- √ 160 observações multivariadas (139 atletas)
- √ Período: 1948 a 1976
- √ Escores padronizados para cada um dos 10 eventos
- √ Análise Fatorial de R por componentes principais e por máxima verossimilhança
- √ Dados: *BD_multivariada.xls/decatlo*

Técnicas Multivariadas em Saúde - 2015

√ A distribuição dos scores padronizados são normais ou aproximadamente normais para cada um dos 10 eventos (Linden, 1977)

√ Variáveis:

- X₁: 100 m rasos
- X₂: Salto em distância
- X₃: Arremesso de peso
- X₄: Salto em altura
- X₅: 400 m rasos
- X₆: 100 m com barreiras
- X₇: Lançamento de disco
- X₈: Salto com vara
- X₉: Lançamento de dardo
- X₁₀: 1.500 m

Técnicas Multivariadas em Saúde - 2015

• Matriz de Correlações

Matrix M1									
1,00	0,59	0,35	0,34	0,63	0,40	0,28	0,20	0,11	-0,07
0,59	1,00	0,42	0,51	0,49	0,52	0,31	0,36	0,21	0,09
0,35	0,42	1,00	0,38	0,39	0,36	0,73	0,24	0,44	-0,08
0,34	0,51	0,38	1,00	0,29	0,46	0,27	0,39	0,17	0,18
0,63	0,49	0,19	0,29	1,00	0,34	0,17	0,23	0,13	0,39
0,40	0,52	0,36	0,46	0,34	1,00	0,32	0,33	0,18	0,00
0,28	0,31	0,73	0,27	0,17	0,32	1,00	0,24	0,34	-0,02
0,20	0,36	0,24	0,39	0,23	0,33	0,24	1,00	0,24	0,17
0,11	0,21	0,44	0,17	0,13	0,18	0,34	0,24	1,00	0,00
-0,07	0,09	-0,08	0,18	0,39	0,00	-0,02	0,17	0,00	1,00

√ Há correlação potencial entre scores sucessivos de atletas que concluíram a prova em mais de uma Olimpíada

- Efetuada análise usando 139 escores representativo de cada atleta
- Escolheu-se aleatoriamente um dos escores dos atletas que participaram de mais de uma Olimpíada

Técnicas Multivariadas em Saúde - 2015

√ Solução por Componentes Principais

Factor Analysis: M1					
Principal Component Factor Analysis of the Correlation Matrix					
Unrotated Factor Loadings and Communalities					
Variable	Factor1	Factor2	Factor3	Factor4	Communality
Var 1	0,691	-0,217	0,520	-0,206	0,837
Var 2	0,789	-0,184	0,193	0,092	0,701
Var 3	0,702	0,535	-0,047	-0,175	0,811
Var 4	0,676	-0,134	-0,139	0,296	0,648
Var 5	0,620	-0,551	0,084	-0,419	0,870
Var 6	0,687	-0,042	0,161	0,245	0,618
Var 7	0,621	0,521	-0,109	-0,234	0,724
Var 8	0,538	-0,087	-0,411	0,440	0,660
Var 9	0,434	0,439	-0,372	-0,235	0,574
Var 10	0,147	-0,596	-0,658	-0,279	0,888
Variance	3,786	1,517	1,114	0,913	7,3317
% Var	0,379	0,152	0,111	0,091	0,728

√ Solução por Máxima Verossimilhança

Maximum Likelihood Factor Analysis of the Correlation Matrix					
Unrotated Factor Loadings and Communalities					
Variable	Factor1	Factor2	Factor3	Factor4	Communality
Var 1	0,493	-0,225	0,700	0,179	0,816
Var 2	0,464	-0,321	0,455	-0,240	0,625
Var 3	0,876	-0,334	-0,235	0,040	0,936
Var 4	0,348	-0,407	0,204	-0,437	0,519
Var 5	0,144	-0,563	0,551	0,166	0,669
Var 6	0,418	-0,247	0,307	-0,345	0,449
Var 7	0,666	-0,301	-0,182	0,041	0,569
Var 8	0,211	-0,322	0,127	-0,360	0,294
Var 9	0,392	-0,194	-0,138	-0,058	0,214
Var 10	-0,440	-0,864	-0,070	0,016	0,946
Variance	2,3788	1,8296	1,2648	0,5643	6,0374
% Var	0,238	0,183	0,126	0,056	0,604

Técnicas Multivariadas em Saúde - 2015

Variável	Componentes Principais						Máxima Verossimilhança					
	ℓ_1	ℓ_2	ℓ_3	ℓ_4	h^2	Ψ^2	ℓ_1	ℓ_2	ℓ_3	ℓ_4	h^2	Ψ^2
100 m rasos	0,691	-0,217	0,520	-0,206	0,837	0,163	0,493	-0,225	0,700	0,179	0,816	0,184
Salto em distância	0,789	-0,184	0,193	0,092	0,701	0,299	0,464	-0,381	0,455	-0,240	0,625	0,375
Arremesso de peso	0,702	0,535	-0,047	-0,175	0,811	0,189	0,876	-0,334	-0,235	0,040	0,936	0,064
Salto em altura	0,674	-0,134	-0,139	-0,396	0,648	0,352	0,348	-0,407	0,204	-0,437	0,519	0,481
400 m rasos	0,620	-0,551	0,084	-0,419	0,870	0,130	0,144	-0,563	0,551	0,166	0,669	0,331
100 m barreiras	0,687	-0,042	0,161	0,345	0,618	0,382	0,418	-0,247	0,307	-0,345	0,449	0,551
Lançamento disco	0,621	0,521	-0,109	-0,234	0,724	0,276	0,666	-0,301	-0,182	0,041	0,569	0,431
Salto com vara	0,538	-0,087	-0,411	0,440	0,660	0,340	0,211	-0,322	0,127	-0,360	0,294	0,706
Lançamento dardo	0,434	0,439	-0,372	-0,235	0,574	0,426	0,392	-0,194	-0,138	-0,058	0,214	0,786
1500 m	0,147	-0,596	-0,658	-0,279	0,888	0,112	-0,440	-0,864	-0,070	0,016	0,946	0,054
Variância	3,787	1,517	1,114	0,913	7,332		2,379	1,830	1,265	0,564	6,037	
% Variância Total	37,9%	15,2%	11,1%	9,1%	73,3%		23,8%	18,3%	12,6%	5,6%	60,4%	

Técnicas Multivariadas em Saúde - 2015

• As soluções dos dois métodos são bem diferentes

√ Solução por Componentes Principais

Factor Analysis: M1					
Principal Component Factor Analysis of the Correlation Matrix					
Unrotated Factor Loadings and Communalities					
Variable	Factor1	Factor2	Factor3	Factor4	Communality
Var 1	0,691	-0,217	0,520	-0,206	0,837
Var 2	0,789	-0,184	0,193	0,092	0,701
Var 3	0,702	0,535	-0,047	-0,175	0,811
Var 4	0,676	-0,134	-0,139	0,296	0,648
Var 5	0,620	-0,551	0,084	-0,419	0,870
Var 6	0,687	-0,042	0,161	0,245	0,618
Var 7	0,621	0,521	-0,109	-0,234	0,724
Var 8	0,538	-0,087	-0,411	0,440	0,660
Var 9	0,434	0,439	-0,372	-0,235	0,574
Var 10	0,147	-0,596	-0,658	-0,279	0,888

- F₁: Todos os loadings positivos e grandes (exceto X₁₀)
Habilidade atlética geral
- F₂: Contraste entre habilidade de corrida e de arremesso
- F₃: Contraste resistência (1500 m) com velocidade (100 m)
 - Embora haja loading relativamente alto para salto com vara
- F₄: É um mistério neste ponto

Técnicas Multivariadas em Saúde - 2015

✓ Solução por Máxima Verossimilhança

Maximum Likelihood Factor Analysis of the Correlation Matrix					
Unrotated Factor Loadings and Communalities					
Variable	Factor1	Factor2	Factor3	Factor4	Communality
Var 1	0,493	-0,225	0,700	0,179	0,816
Var 2	0,464	-0,351	0,455	-0,240	0,625
Var 3	0,876	-0,334	-0,235	0,040	0,936
Var 4	0,348	-0,407	0,204	-0,437	0,519
Var 5	0,144	-0,563	0,551	0,166	0,669
Var 6	0,418	-0,247	0,307	-0,345	0,449
Var 7	0,666	-0,301	-0,182	0,041	0,569
Var 8	0,211	-0,322	0,127	-0,360	0,294
Var 9	0,392	-0,194	-0,138	-0,058	0,214
Var 10	-0,440	-0,864	-0,070	0,016	0,946

- F₁: Loadings de arremesso de disco e de peso são altos
Fator de força
- F₂: Corrida 1500 m única variável com loading alto
Fator resistência
- F₃: Loadings de corrida 100 e 400 m são altos
Fator velocidade
- F₄: Não é facilmente identificável
(pode ter algo com habilidade de salto e força nas pernas)

Técnicas Multivariadas em Saúde - 2015

• Matriz de Resíduos: $R - \hat{L}\hat{L}' - \hat{\Psi}$

✓ Componentes Principais

Componentes Principais									
0	-0,075	-0,080	-0,001	-0,047	-0,096	-0,027	0,114	0,051	-0,016
-0,075	0	-0,010	-0,056	-0,077	-0,092	-0,041	-0,042	0,042	0,017
-0,030	-0,010	0	0,042	-0,020	-0,032	-0,031	-0,034	-0,158	0,056
-0,001	-0,056	0,042	0	-0,024	-0,122	-0,001	-0,215	-0,022	0,020
-0,047	-0,077	-0,020	-0,024	0	0,022	-0,017	0,067	0,036	-0,091
-0,096	-0,092	-0,032	-0,122	0,022	0	0,014	-0,129	0,041	0,076
-0,027	-0,041	-0,031	-0,001	-0,017	0,014	0	0,009	-0,254	0,062
0,114	-0,042	-0,034	-0,215	0,067	-0,129	0,009	0	-0,005	-0,109
0,051	0,042	-0,158	-0,022	0,036	0,041	-0,254	-0,005	0	-0,112
-0,016	0,017	0,056	0,020	-0,091	0,076	0,062	-0,109	-0,112	0

✓ Máxima Verossimilhança

Máxima Verossimilhança									
0	0,001	0,000	0,013	0,017	-0,014	0,004	-0,001	-0,020	-0,001
0,001	0	0,002	-0,004	-0,002	0,009	-0,021	-0,005	0,003	0,000
0,000	0,002	0	0,005	-0,002	-0,004	0,001	-0,008	0,001	0,000
0,013	-0,004	0,005	0	-0,029	0,001	-0,029	0,002	-0,042	0,002
0,017	-0,002	-0,002	-0,029	0	0,029	-0,002	0,008	0,050	0,003
-0,014	0,009	-0,004	0,001	0,029	0	0,037	-0,001	-0,010	-0,002
0,004	-0,021	0,001	-0,029	-0,002	0,037	0	0,041	-0,003	0,001
-0,001	-0,005	-0,008	0,002	0,008	-0,001	0,041	0	0,092	-0,001
-0,020	0,003	0,001	-0,042	0,050	-0,010	-0,003	0,092	0	-0,004
-0,001	0,000	0,000	0,002	0,003	-0,002	0,001	-0,001	-0,004	0

Técnicas Multivariadas em Saúde - 2015

Rotação dos Fatores

- A matriz de covariância Σ é reproduzida pelos loadings fatoriais obtidos por transformação ortogonal, da mesma maneira que os loadings iniciais.

✓ Matriz de covariâncias estimada

$$\hat{L}\hat{L}' + \hat{\Psi} = \hat{L}T T' \hat{L}' + \hat{\Psi} = \hat{L}^* \hat{L}^{*'} + \hat{\Psi}$$

✓ $T T' = T' T = I$

✓ \hat{L}^* : matriz de loadings rotacionados

✓ A matriz de resíduos permanece a mesma (\hat{h}_i^2 e $\hat{\Psi}_i$)

$$S_n - \hat{L}\hat{L}' - \hat{\Psi} = S_n - \hat{L}^* \hat{L}^{*'} - \hat{\Psi}$$

✓ Do ponto de vista estatístico é irrelevante obter \hat{L} ou \hat{L}^*

Técnicas Multivariadas em Saúde - 2015

• Comentários:

✓ Com a rotação, busca-se uma estrutura mais simples
- loadings originais podem não ter fácil interpretação

✓ Ideal: encontrar um padrão de loadings tais que cada variável carregue-se fortemente em um único fator
(com loadings moderados nos outros fatores)

✓ Nem sempre é possível obter esta estrutura mais simples

Técnicas Multivariadas em Saúde - 2015

Exemplo 9.8 – Examination Scores

- √ Lawley & Maxwell (1971)
- √ Avaliações de 220 estudantes do sexo masculino
- √ $p = 6$
- √ Dados: *BD_multivariada.xls/examination*
- √ Matriz de correlações:

Gaélico	1	0,439	0,410	0,288	0,329	0,248
Inglês	0,439	1	0,351	0,354	0,320	0,329
História	0,410	0,351	1	0,164	0,190	0,181
Aritmética	0,288	0,354	0,164	1	0,595	0,470
Álgebra	0,329	0,320	0,190	0,595	1	0,464
Geometria	0,248	0,329	0,181	0,470	0,464	1

Técnicas Multivariadas em Saúde - 2015

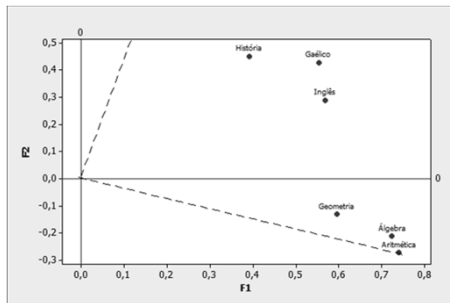
- Solução por Máxima Verossimilhança:

Factor Analysis: M1			
Maximum Likelihood Factor Analysis of the Correlation Matrix			
Unrotated Factor Loadings and Communalities			
Variable	Factor1	Factor2	Communality
Var 1	0,553	0,428	0,490
Var 2	0,568	0,288	0,406
Var 3	0,392	0,450	0,356
Var 4	0,740	-0,272	0,622
Var 5	0,724	-0,212	0,569
Var 6	0,595	-0,132	0,372
Variance	2,2095	0,6056	2,8151
% Var	0,368	0,101	0,469

- √ F_1 : reflete a resposta global dos estudantes à instrução
fator de inteligência geral
- √ F_2 : não é facilmente identificável
 - Fator “Math – nonmath”
 - metade positiva, metade negativa (fator bipolar)

Técnicas Multivariadas em Saúde - 2015

- Plot dos loadings fatoriais:



- √ Todos os pontos caem no primeiro quadrante
- √ Revelam-se mais claramente 2 clusters das variáveis

Técnicas Multivariadas em Saúde - 2015

- Rotação horária de 20° : $T = \begin{bmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{bmatrix} = \begin{bmatrix} 0,9397 & 0,3420 \\ -0,3420 & 0,9397 \end{bmatrix}$
- Rotação dos loadings:

Unrotated Factor Loadings and Communalities			
Variable	Factor1	Factor2	Communality
Var 1	0,373	0,594	0,490
Var 2	0,435	0,467	0,406
Var 3	0,214	0,558	0,356
Var 4	0,789	0,001	0,622
Var 5	0,752	0,054	0,569
Var 6	0,604	0,083	0,372
Variance	1,9295	0,8856	2,8151
% Var	0,322	0,148	0,469

- √ F_1^* : variáveis matemáticas do teste com loading alto
- (desprezíveis em F_2^*)
- Fator de habilidade matemática
- √ F_2^* : variáveis de habilidade verbal com loadings altos
- Fator de habilidade verbal

Técnicas Multivariadas em Saúde - 2015

- **Comentários:**

- √ O fator de inteligência geral está submergido dos fatores F_1^* e F_2^* .
- √ As comunalidades não se modificam (fatores com e sem rotação)

Técnicas Multivariadas em Saúde - 2015

Crítérios de Rotação

- **Ideal:**

- √ Transformação que fizesse os loadings de cada Z_i ter valor grande em apenas um dos fatores e valores pequenos (ou moderados) nos outros
 - Para facilitar a interpretação dos fatores

- **Alguns critérios para encontrar matriz ortogonal:**

- √ Varimax
- √ Quartimax
- √ Orthomax

Técnicas Multivariadas em Saúde - 2015

- **Qualidade de ajuste**

- √ A rotação não acrescenta nenhuma melhoria em relação ao ajuste original
 - Matriz residual original não é alterada pela transformação ortogonal
 - Valores estimados de comunalidade e variâncias específicas permanecem inalterados

- **Interpretação:**

- √ Novos fatores podem ser de mais fácil interpretação
- Quando a solução sem rotação já é de boa qualidade, não se recomenda rotação
 - √ Solução rotacionada pode ser pior

Técnicas Multivariadas em Saúde - 2015

- **Critério Varimax:**

- √ É um dos mais utilizados
- √ Em geral, produz soluções mais simples

- **Critério Quartimax**

- √ Tem tendência de gerar fatores, onde todas as variáveis têm loadings elevados

- **Critério Orthomax**

- √ É uma média ponderada dos dois outros métodos

Técnicas Multivariadas em Saúde - 2015

Exemplo 9.9 – Preferência Consumidor

- (continuação exemplo 9.3)

- $\sqrt{X_1}$: Gosto
- $\sqrt{X_2}$: Preço
- $\sqrt{X_3}$: Aroma
- $\sqrt{X_4}$: Adequado para lanche
- $\sqrt{X_5}$: fornece muita energia
- $\sqrt{\text{Dados}}$: *BD_multivariada.xls/preferencia_consumidor*

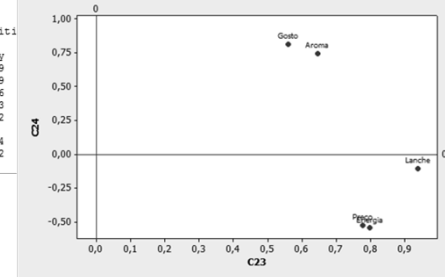
Técnicas Multivariadas em Saúde - 2015

- Solução por Componentes Principais:

Principal Component Factor Analysis of th

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Communality
Var 1	0,560	0,816	0,979
Var 2	0,777	-0,524	0,879
Var 3	0,645	0,748	0,976
Var 4	0,939	-0,105	0,893
Var 5	0,798	-0,543	0,932
Variance	2,8531	1,8063	4,6594
% Var	0,571	0,361	0,932



Técnicas Multivariadas em Saúde - 2015

- Rotação Varimax:

Rotated Factor Loadings and Communalities
Varimax Rotation

Variable	Factor1	Factor2	Communality
Var 1	0,020	0,989	0,979
Var 2	0,937	-0,011	0,879
Var 3	0,129	0,979	0,976
Var 4	0,842	0,428	0,893
Var 5	0,965	-0,016	0,932
Variance	2,5374	2,1220	4,6594
% Var	0,507	0,424	0,932

- $\sqrt{F_1}$ *: Fator nutricional
Variável 4 está mais alinhada com F_1 *
- $\sqrt{F_2}$ *: Fator sabor

Técnicas Multivariadas em Saúde - 2015

- Comparação

Variável	Loadings Estimados				h_i^2
	Originais		Rotação		
	L_{11}	L_{12}	L_{21}	L_{22}	
Gosto	0,560	0,816	0,020	0,989481	0,979
Preço	0,777	-0,524	0,937	-0,01123	0,879
Aroma	0,645	0,748	0,129	0,979467	0,976
Adequado lanche	0,939	-0,105	0,842	0,428052	0,893
Energético	0,798	-0,543	0,965	-0,01563	0,932
Variância	2,853	1,806	2,537	2,122	4,6594
% variabilidade	57,1%	36,1%	50,7%	42,4%	93,2%

Técnicas Multivariadas em Saúde - 2015

Exemplo 9.10 – Ações New York Rotação

Continuação Exemplo 9.4

- Taxas de retorno de 5 ações negociadas na Bolsa de New York

√ Período: jan/75 a Dez/76

– Observadas 100 semanas

√ Ações:

- Allied Chemical
- du Pont
- Union Carbide
- Exxon
- Texaco

Técnicas Multivariadas em Saúde - 2015

√ Solução por Máxima Verossimilhança

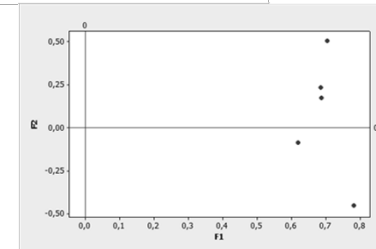
Factor Analysis: allied_chemical; du_pont; union_carbide; exxon; texaco

Maximum Likelihood Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Communality
allied_chemical	0,687	-0,176	0,503
du_pont	0,704	-0,506	0,751
union_carbide	0,685	-0,234	0,525
exxon	0,620	0,086	0,392
texaco	0,782	0,452	0,816
Variance	2,4338	0,5538	2,9876
% Var	0,487	0,111	0,598

√ Plot dos loadings:



Técnicas Multivariadas em Saúde - 2015

- Rotação Varimax:

Rotated Factor Loadings and Communalities
Varimax Rotation

Variable	Factor1	Factor2	Communality
allied_chemical	0,600	0,379	0,503
du_pont	0,851	0,164	0,751
union_carbide	0,641	0,338	0,525
exxon	0,363	0,510	0,392
texaco	0,208	0,879	0,816

- Matriz de Resíduos:

m=2

0	-0,004	0,004	0,024	-0,004
-0,004	0	0,003	0,004	0,000
0,004	0,003	0	-0,031	0,005
0,024	0,004	-0,031	0	0,000
-0,004	0,000	0,005	0,000	0

Técnicas Multivariadas em Saúde - 2015

Loadings Estimados

Variável	Originais		Rotação		h_i^2
	l_{11}	l_{12}	l_{11}	l_{12}	
Allied Chemical	0,687	-0,176	0,600	0,379	0,503
Du Pont	0,704	-0,506	0,851	0,164	0,751
Union Carbide	0,685	-0,234	0,641	0,338	0,525
Exxon	0,620	0,086	0,363	0,510	0,392
Texaco	0,782	0,452	0,208	0,879	0,816
Variância	2,434	0,554	1,670	1,318	2,988
% variabilidade	48,7%	11,1%	33,4%	26,4%	59,8%

√ F_1^* : Indústrias químicas carregam fortemente

– Representam condições econômicas que afetam essas ações

√ F_2^* : Ações de óleo & gás carregam fortemente

– Representam condições econômicas que afetam essas ações

√ Rotação tende a destruir um fator geral

Técnicas Multivariadas em Saúde - 2015

Exemplo 9.11 – Declato Olímpico

- Continuação exemplo 9.6
 - √ 160 observações multivariadas (139 atletas)
 - √ Período: 1948 a 1976
 - √ Escores padronizados para cada um dos 10 eventos
 - √ Análise Fatorial de R por componentes principais e por máxima verossimilhança
 - √ Dados: *BD_multivariada.xls/decatlo*

Técnicas Multivariadas em Saúde - 2015

√ A distribuição dos scores padronizados são normais ou aproximadamente normais para cada um dos 10 eventos (Linden, 1977)

√ Variáveis:

- X₁: 100 m rasos
- X₂: Salto em distância
- X₃: Arremesso de peso
- X₄: Salto em altura
- X₅: 400 m rasos
- X₆: 100 m com barreiras
- X₇: Lançamento de disco
- X₈: Salto com vara
- X₉: Lançamento de dardo
- X₁₀: 1.500 m

Técnicas Multivariadas em Saúde - 2015

√ Solução por Componentes Principais

Factor Analysis: M1

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Factor4	Communality
Var 1	0,691	-0,227	0,820	-0,206	0,937
Var 2	0,789	-0,184	0,193	0,092	0,701
Var 3	0,702	0,535	-0,047	-0,178	0,811
Var 4	0,874	-0,134	-0,139	0,396	0,648
Var 5	0,620	-0,951	0,084	-0,419	0,870
Var 6	0,687	-0,042	0,161	0,345	0,618
Var 7	0,821	0,821	-0,109	-0,234	0,724
Var 8	0,328	-0,087	-0,411	0,440	0,460
Var 9	0,434	0,439	-0,372	-0,235	0,574
Var 10	0,147	-0,596	-0,688	-0,279	0,888
Variance	2,7866	1,5179	1,1164	0,9194	7,3317
% Var	0,379	0,182	0,111	0,091	0,732

√ Solução por Máxima Verossimilhança

Maximum Likelihood Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Factor4	Communality
Var 1	0,493	-0,225	0,700	0,179	0,816
Var 2	0,464	-0,381	0,455	-0,240	0,625
Var 3	0,876	-0,334	-0,235	0,040	0,936
Var 4	0,348	-0,407	0,204	-0,437	0,519
Var 5	0,144	-0,563	0,551	0,166	0,669
Var 6	0,418	-0,247	0,307	-0,345	0,449
Var 7	0,666	-0,301	-0,182	0,041	0,569
Var 8	0,211	-0,322	0,127	-0,360	0,294
Var 9	0,392	-0,194	-0,138	-0,058	0,214
Var 10	-0,440	-0,864	-0,070	0,016	0,946
Variance	2,3788	1,8296	1,2648	0,5643	6,0374
% Var	0,238	0,183	0,126	0,056	0,604

Técnicas Multivariadas em Saúde - 2015

√ Rotação Componentes Principais

Rotated Factor Loadings and Communalities

Varimax Rotation

Variable	Factor1	Factor2	Factor3	Factor4	Communality
Var 1	0,884	0,136	0,156	0,113	0,837
Var 2	0,631	0,194	0,515	0,006	0,701
Var 3	0,245	0,825	0,223	0,148	0,811
Var 4	0,239	0,151	0,750	-0,077	0,648
Var 5	0,797	0,075	0,102	-0,468	0,870
Var 6	0,404	0,153	0,635	0,170	0,618
Var 7	0,186	0,814	0,147	0,079	0,724
Var 8	-0,036	0,176	0,762	-0,217	0,460
Var 9	-0,048	0,735	0,110	-0,141	0,574
Var 10	0,045	-0,041	0,112	-0,934	0,888
Variance	2,1345	2,0230	1,9407	1,2335	7,3317
% Var	0,213	0,202	0,194	0,123	0,733

√ Rotação Máxima Verossimilhança

Maximum Likelihood Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Factor4	Communality
Var 1	0,493	-0,225	0,700	0,179	0,816
Var 2	0,464	-0,381	0,455	-0,240	0,625
Var 3	0,876	-0,334	-0,235	0,040	0,936
Var 4	0,348	-0,407	0,204	-0,437	0,519
Var 5	0,144	-0,563	0,551	0,166	0,669
Var 6	0,418	-0,247	0,307	-0,345	0,449
Var 7	0,666	-0,301	-0,182	0,041	0,569
Var 8	0,211	-0,322	0,127	-0,360	0,294
Var 9	0,392	-0,194	-0,138	-0,058	0,214
Var 10	-0,440	-0,864	-0,070	0,016	0,946
Variance	2,3788	1,8296	1,2648	0,5643	6,0374
% Var	0,238	0,183	0,126	0,056	0,604

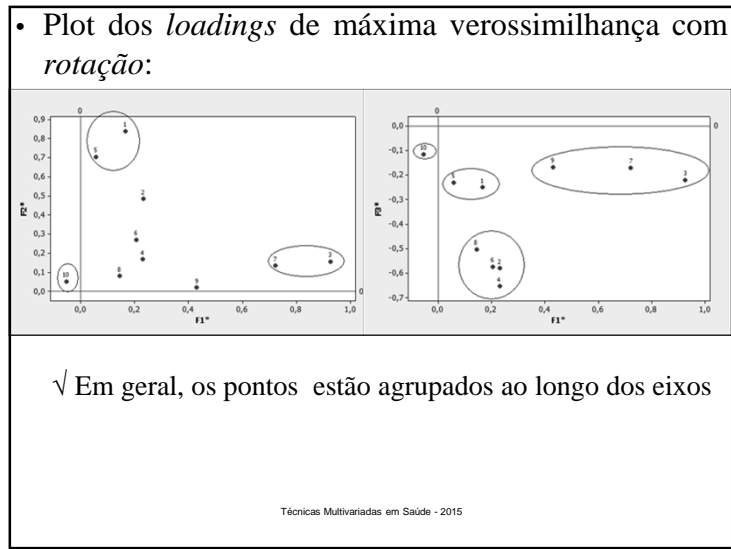
Técnicas Multivariadas em Saúde - 2015

Variável	Loadings Estimados									
	Componentes Principais					Máxima Verossimilhança				
	L_{11}	L_{12}	L_{13}	L_{14}	h_1^2	L_{11}	L_{12}	L_{13}	L_{14}	h_1^2
100m rasos	0,884	0,136	0,156	0,113	0,837	0,166	0,841	0,25022	0,137	0,816
Salto em distância	0,631	0,194	0,515	0,006	0,701	0,232	0,484	0,57987	-0,011	0,625
Arremesso de peso	0,245	0,825	0,223	0,148	0,811	0,927	0,157	0,7187	0,066	0,936
Salto em altura	0,239	0,151	0,750	-0,077	0,648	0,230	0,168	0,65275	-0,111	0,519
400m rasos	0,797	0,075	0,107	-0,468	0,870	0,058	0,705	0,72959	-0,340	0,669
100m com barreiras	0,404	0,153	0,635	0,170	0,618	0,206	0,270	0,57339	0,069	0,449
Lançamento de disco	0,186	0,814	0,147	0,079	0,724	0,723	0,133	0,17076	0,009	0,569
Salto com vara	-0,036	0,176	0,762	-0,217	0,660	-0,144	0,081	0,50208	-0,121	0,294
Lançamento de dardo	-0,048	0,735	0,110	-0,141	0,574	0,431	0,021	-0,16697	-0,007	0,214
1.500 m	0,045	-0,041	0,112	-0,934	0,888	-0,054	0,052	-0,11492	-0,963	0,946
Variancia	2,134	2,023	1,941	1,233	7,332	1,771	1,592	1,577	1,097	6,037
% variabilidade	21,3%	20,2%	19,4%	12,3%	73,3%	17,7%	15,9%	15,8%	11,0%	60,4%

√ Ambos os métodos
 – Loadings apontam para mesmo atributos subjacentes
 – Fatores 1 e 2 não estão na mesma ordem

Técnicas Multivariadas em Saúde - 2015

- Solução de máxima verossimilhança -Interpretação:
 - √ F1*: Arremesso de peso, lançamento de disco e lançamento de dardo.
 - Linden (1977): *explosive arm strength*
 - √ F2*: 100 m rasos e 400 m rasos (salto em distância)
 - Linden (1977): *running speed*
 - √ F3*: Salto em altura, 110 m com barreiras, salto com vara e salto em distância
 - Linden (1977): *explosive leg strength*
 - √ F4*: 1.500 m e 400 m rasos
 - Linden (1977): *running endurance*
- Técnicas Multivariadas em Saúde - 2015



Estimação dos Escores dos Fatores

- Modelo fatorial:

$$Z_1 = l_{11}F_1 + l_{12}F_2 + \dots + l_{1m}F_m + \epsilon_1$$

$$Z_2 = l_{21}F_1 + l_{22}F_2 + \dots + l_{2m}F_m + \epsilon_2$$

$$\vdots$$

$$Z_p = l_{p1}F_1 + l_{p2}F_2 + \dots + l_{pm}F_m + \epsilon_p$$
- √ Para cada elemento amostral k , seu escore no fator F_j é calculado como:

$$\hat{F}_{jk} = w_{j1}Z_{1k} + w_{j2}Z_{2k} + \dots + w_{jp}Z_{pk}$$
 - w_{jk} : peso de ponderação de cada variável no Fator F_j

Técnicas Multivariadas em Saúde - 2015

- Os escores podem ser utilizados para construir:
 - √ Gráficos
 - √ Mapas de percepção
 - √ Variável reposta ou explicativa em outros métodos
- Métodos de estimação dos scores:
 - √ Método de mínimos quadrados ponderados
 - √ Método de regressão
 - √ Método ad hoc

Técnicas Multivariadas em Saúde - 2015

Exemplo 9.12 – Cálculo dos Scores Fatoriais

- Ações New York
 - √ Solução por Máxima Verossimilhança de R

Rotated Factor Loadings

0,599557	0,379038
0,851127	0,164465
0,640928	0,337589
0,363370	0,509725
0,207976	0,879263

Matrix \Psi

0,496861	0,000000	0,000000	0,000000	0,000000
0,000000	0,248534	0,000000	0,000000	0,000000
0,000000	0,000000	0,475245	0,000000	0,000000
0,000000	0,000000	0,000000	0,608143	0,000000
0,000000	0,000000	0,000000	0,000000	0,183642

√ Regressão: $\hat{f} = \hat{L}'R^{-1}z$

√ Mínimos quadrados ponderados: $\hat{f} = (\hat{L}'\hat{\Psi}^{-1}\hat{L}')^{-1}\hat{L}'\hat{\Psi}^{-1}z$

Técnicas Multivariadas em Saúde - 2015

- Para o vetor de observações padronizadas:
 - √ $z' = [0,5; -1,40; -0,20; -0,70; 1,40]$
 - √ scores por regressão

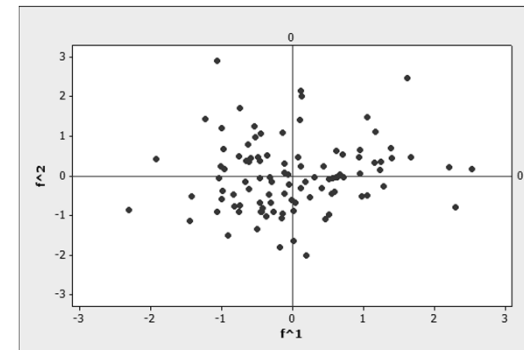
$$\hat{f} = \hat{L}'Rz = \begin{bmatrix} 0,1838 & 0,6626 & 0,2188 & 0,0479 & -0,2085 \\ 0,0402 & -0,1877 & 0,0164 & 0,1123 & 0,8553 \end{bmatrix} \begin{bmatrix} 0,50 \\ -1,40 \\ -0,20 \\ -0,70 \\ 1,40 \end{bmatrix} = \begin{bmatrix} -1,205 \\ 1,398 \end{bmatrix}$$

- √ Para cálculo da matriz de dados

$$\hat{f}' = z'R^{-1}\hat{L}'$$

Técnicas Multivariadas em Saúde - 2015

- Plot scores fatoriais:



Técnicas Multivariadas em Saúde - 2015

Critérios para Determinação do Valor de m

- Teste de hipótese para auxiliar na decisão do número de fatores (m) que são suficientes para o modelo de análise fatorial
- Suposições do teste:
 - √ Os vetores aleatórios \mathbf{Z} e \mathbf{F} têm distribuição normal multivariada
 - √ Amostras de tamanho grande

Técnicas Multivariadas em Saúde - 2015

• Teste

- √ H_0 : m fatores são suficientes vs.
- H_1 : necessários mais que m fatores

√ Estatística de teste (Bartlett, 1954)

$$T = \left[n - 1 - \frac{2p + 4m + 5}{6} \right] \ln \left(\frac{|\hat{\mathbf{L}}_z \hat{\mathbf{L}}_z' + \hat{\Psi}_z|}{|\mathbf{R}|} \right)$$

√ Sob H_0 : $T \sim \chi^2_{gl}$, com $gl = \frac{1}{2} [(p - m)^2 - p - m]$

√ Para o teste ser válido $gl > 0$

- Se $p = 5, m \leq 2$
- Se $p = 10, m \leq 5$
- Se $p = 20, m \leq 14$

√ Para n grande e m pequeno em relação a p , o teste tende a rejeitar H_0 (indica aumento de m)

Técnicas Multivariadas em Saúde - 2015

• Comentários:

- √ Teste somente é válido para dados provenientes de distribuição normal p -variada
- √ Para n grande e m pequeno em relação a p , o teste tende a rejeitar H_0 (indicando o aumento de m)
 - Baseando-se apenas na indicação do teste, a tendência será reter no sistema um número muito grande de fatores sem necessidade

Técnicas Multivariadas em Saúde - 2015

Exemplo 9.7 – Teste para valor de m

• Ações Bolsa New York ($m = 2$)

$$\hat{\mathbf{L}}_z \hat{\mathbf{L}}_z' + \hat{\Psi} = \begin{bmatrix} 1,00000 & 0,57264 & 0,51223 & 0,41107 & 0,45797 \\ 0,57264 & 1,00000 & 0,60103 & 0,39311 & 0,32162 \\ 0,51223 & 0,60103 & 1,00000 & 0,40497 & 0,43013 \\ 0,41107 & 0,39311 & 0,40497 & 1,00000 & 0,52375 \\ 0,45797 & 0,32162 & 0,43013 & 0,52375 & 1,00000 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1,00000 & 0,57692 & 0,50866 & 0,38672 & 0,46218 \\ 0,57692 & 1,00000 & 0,59838 & 0,38952 & 0,32195 \\ 0,50866 & 0,59838 & 1,00000 & 0,43610 & 0,42563 \\ 0,38672 & 0,38952 & 0,43610 & 1,00000 & 0,52353 \\ 0,46218 & 0,32195 & 0,42563 & 0,52353 & 1,00000 \end{bmatrix}$$

$$|\hat{\mathbf{L}}_z \hat{\mathbf{L}}_z' + \hat{\Psi}| = 0,194403 \text{ e } |\mathbf{R}| = 0,193234$$

$$T = \left[100 - 1 - \frac{2(5) + 4(2) + 5}{6} \right] \ln \left(\frac{0,194403}{0,193234} \right) = 0,574$$

√ Valor crítico: $gl = \frac{1}{2} [(5 - 2)^2 - 5 - m] = 1$

$$\chi^2_1(0,05) = 3,84 \rightarrow \text{Não se rejeita } H_0$$

$$p\text{-valor: } P\{\chi^2_1 > 0,574\} = 0,448674$$

- H_0 não deveria ser rejeitada em qualquer nível razoável

Técnicas Multivariadas em Saúde - 2015

Crítério de Akaike

Akaike (1974, 1987)

- Suposições:
 - √ Dados provenientes de normal multivariada
 - √ Envolve método de máxima verossimilhança
- Critério: escolhe-se o valor de m que minimize a função AIC

$$AIC = [-\ln(\max\{L(\mu, \Sigma)\})]$$

- √ Se dois métodos tem a mesma verossimilhança, o procedimento vai privilegiar o modelo com menor número de fatores

Técnicas Multivariadas em Saúde - 2015

Crítério Bayesiano de Schwarz

- Suposições:
 - √ Dados provenientes de normal multivariada
 - √ Envolve método de máxima verossimilhança
- Critério: escolhe-se o valor de m que minimize a função SBC

$$SBC = [-\ln(\max\{L(\mu, \Sigma)\})] + \frac{m}{2} \ln(n)$$

- √ Os critérios AIC e SBC devem ser usados com cautela
 - Em geral, indicam quantidade de fatores maior que a necessária
- √ Em geral, o critério SBC resulta em melhores soluções que o método Akaike

Técnicas Multivariadas em Saúde - 2015

Matriz de Resíduos

- A observação da matriz de resíduos:
 - √ Muitas vezes, pode indicar quando o número de fatores está superdimensionado
 - √ Ex.:
 - Se m não for muito pequeno e a matriz de resíduos estiver próxima de zero, recomenda-se testar outras soluções para m menores que o valor já especificado

Técnicas Multivariadas em Saúde - 2015

- Importante:
 - √ Análise fatorial deve ser utilizada apenas se utilizada em situações em que as variáveis originais são correlacionadas
 - √ Consequência:
 - Evitar soluções com m elevado tal que determinados fatores fiquem relacionados com uma única variável original
 - √ Em situações em que aparecem fatores relacionados a uma única variável Z_i é recomendável retirar a variável Z_i e reestimar o modelo de análise fatorial

Técnicas Multivariadas em Saúde - 2015

Validação do Modelo

- Análise Fatorial está fundamentada em suposições que não podem ser verificadas a priori:
 - √ Linearidade e independência dos fatores
 - √ Interpretação centrada na informação contida na matriz L (estimada a partir da escolha prévia de m)
- É importante avaliar até que ponto a matriz \hat{L} está representando corretamente a relação existente entre as variáveis originais e os fatores do modelo

Técnicas Multivariadas em Saúde - 2015

Estratégia para Análise Fatorial

Johnson & Wichern (2002)

- Decisões em qualquer Análise Fatorial
 - √ Escolha de m , o número de fatores comuns
 - Há muitos testes de adequação assintóticos que são apropriados apenas os dados que são aproximadamente normais
 - Os teste provavelmente rejeitarão o modelo para m pequeno se o número de variáveis e de observações for alto
 - Em geral a escolha é baseada em alguma combinação de:
 - proporção de variância amostral explicada
 - conhecimento do assunto
 - razoabilidade dos resultados

Técnicas Multivariadas em Saúde - 2015

- √ Escolha do método de solução e do tipo de rotação
 - São decisões menos cruciais
 - A maioria de análises fatoriais satisfatórias são aquelas em que:
 - são tentados mais de um método de rotação
 - os resultados confirmam substancialmente a mesma estrutura

Técnicas Multivariadas em Saúde - 2015

Roteiro

1. Execute uma Análise Fatorial por componentes principais
 - √ Este método é particularmente apropriado para uma primeira passagem pelos dados
 - √ Procure observações suspeitas plotando os escores fatoriais
 - Calcule os escores padronizados e as distâncias quadráticas para cada observação
 - √ Tente rotação Varimax

Técnicas Multivariadas em Saúde - 2015

2. Execute Análise Fatorial de Máxima Verossimilhança,
 - √ (incluir uma rotação Varimax)
3. Compare as soluções obtidas pelas duas análises fatoriais
 - √ Os loadings se agrupam da mesma maneira?
 - √ Plote os escores fatoriais obtidos por componentes principais vs. os obtido pela solução de máxima verossimilhança
4. Repita passos anteriores para outros valores de m
 - √ Os fatores extras contribuem necessariamente para a compreensão e interpretação dos dados?

Técnicas Multivariadas em Saúde - 2015

5. Para grandes conjuntos de dados, divida-os pela metade e execute uma Análise Fatorial em cada parte
 - √ Compare os dois resultados, com aquele obtido do conjunto de dados completo
 - √ Verifique a estabilidade da solução
 - √ A divisão pode ser aleatória ou determinística

Técnicas Multivariadas em Saúde - 2015

Exemplo 9.14

- Medidas de ossos e crânios de frangos White Leghorn
 - √ Dados originais: Dunn (1928)
 - √ Análise fatorial elaborada por Wright (1954)
 - √ Variáveis:

Crânio	Pernas	Asas
√ X_1 : comprimento	√ X_3 : fêmur (comp.)	√ X_5 : úmero (comp.)
√ X_2 : amplitude	√ X_4 : tíbia (comp.)	√ X_6 : cúbito (comp.)
 - √ Dados: *BD_multivariada.xls/frangos*

Técnicas Multivariadas em Saúde - 2015

• Matriz de Correlações amostral:

R (conjunto completo)						
1,000	0,505	0,569	0,602	0,621	0,603	
0,505	1,000	0,422	0,467	0,482	0,450	
0,569	0,422	1,000	0,926	0,877	0,878	
0,602	0,467	0,926	1,000	0,874	0,894	
0,621	0,482	0,877	0,874	1,000	0,937	
0,603	0,450	0,878	0,894	0,937	1,000	

R_1 ($n_1 = 137$)						R_2 ($n_2 = 139$)					
1,000	0,696	0,588	0,639	0,694	0,660	1,000	0,366	0,572	0,587	0,587	0,598
0,696	1,000	0,540	0,575	0,606	0,584	0,366	1,000	0,352	0,406	0,420	0,386
0,588	0,540	1,000	0,901	0,844	0,866	0,572	0,352	1,000	0,950	0,909	0,894
0,639	0,575	0,901	1,000	0,835	0,863	0,587	0,406	0,950	1,000	0,911	0,927
0,694	0,606	0,844	0,835	1,000	0,931	0,587	0,420	0,909	0,911	1,000	0,940
0,660	0,584	0,866	0,863	0,931	1,000	0,598	0,386	0,894	0,927	0,940	1,000

Técnicas Multivariadas em Saúde - 2015

• Solução por Componentes Principais

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Communality
Var 1	0,741	0,350	0,573	0,999
Var 2	0,604	0,721	-0,340	1,000
Var 3	0,929	-0,233	-0,075	0,923
Var 4	0,943	-0,174	-0,067	0,925
Var 5	0,948	-0,143	-0,045	0,920
Var 6	0,945	-0,189	-0,047	0,930
Variance	4,4564	0,7824	0,4584	5,6973
% Var	0,743	0,130	0,076	0,950

Rotated Factor Loadings and Communalities
Varimax Rotation

Variable	Factor1	Factor2	Factor3	Communality
Var 1	0,355	0,902	0,243	0,999
Var 2	0,235	0,211	0,949	1,000
Var 3	0,921	0,218	0,165	0,923
Var 4	0,904	0,251	0,212	0,925
Var 5	0,888	0,284	0,228	0,920
Var 6	0,908	0,264	0,191	0,930
Variance	3,4578	1,1198	1,1197	5,6973
% Var	0,576	0,187	0,187	0,950

Técnicas Multivariadas em Saúde - 2015

• Solução por Máxima Verossimilhança

Factor Analysis: M1

Maximum Likelihood Factor Analysis of the Correlation Matrix

* NOTE * Heywood case

Rotated Factor Loadings and Communalities
Varimax Rotation

Variable	Factor1	Factor2	Factor3	Communality
Var 1	0,468	-0,491	0,154	0,484
Var 2	0,209	-0,808	0,066	0,701
Var 3	0,887	-0,281	0,126	0,881
Var 4	0,941	-0,336	-0,026	1,000
Var 5	0,816	-0,349	0,444	0,985
Var 6	0,846	-0,314	0,310	0,910
Variance	3,3169	1,3061	0,3375	4,9605
%				

√ Heywood Case: variância específica da tíbia = 0
 – Replicar o resultado, usando a opção Heywood (SAS ou similar)

Técnicas Multivariadas em Saúde - 2015

• Solução por Componentes Principais

Componentes Principais

Variável	Loadings Estimados			Loadings Estimados (Rotação)			h_i^2	Ψ_i
	λ_{11}	λ_{12}	λ_{13}	λ_{21}	λ_{22}	λ_{23}		
Comprimento	0,741	0,350	0,573	0,355	0,902	0,243	0,999	0,001
Amplitude	0,604	0,721	-0,340	0,235	0,211	0,949	1,000	0,000
Fêmur	0,929	-0,233	-0,075	0,921	0,218	0,165	0,923	0,077
Tíbia	0,943	-0,174	-0,067	0,904	0,251	0,212	0,925	0,075
Úmero	0,948	-0,143	-0,045	0,888	0,284	0,228	0,920	0,080
Cúbito	0,945	-0,189	-0,047	0,908	0,264	0,191	0,930	0,070
Variância	4,456	0,782	0,458	3,458	1,120	1,120	5,697	
% variabilidade	74,3%	13,0%	7,6%	57,6%	18,7%	18,7%	95,0%	

• Matriz de Resíduos

Matrix M8

-0,0000000	0,0002522	0,0055712	0,0028457	-0,0051820	-0,0036794
0,0002522	-0,0000000	0,0031813	-0,0000630	-0,0027691	-0,0006024
0,0055712	0,0031813	-0,0000000	0,0041433	-0,0397322	-0,0468809
0,0028457	-0,0000630	0,0041433	-0,0000000	-0,0478565	-0,0332595
-0,0051820	-0,0027691	-0,0397322	-0,0478565	-0,0000000	0,0127770
-0,0036794	-0,0006024	-0,0468809	-0,0332595	0,0127770	-0,0000000

Técnicas Multivariadas em Saúde - 2015

• Solução por Máxima Verossimilhança

Máxima Verossimilhança

Variável	Loadings Estimados			Loadings Estimados (Rotação)			h_i^2	Ψ_i
	λ_{11}	λ_{12}	λ_{13}	λ_{21}	λ_{22}	λ_{23}		
Comprimento	0,602	0,279	0,209	0,468	0,491	0,154	0,484	0,516
Amplitude	0,467	0,674	0,172	0,209	0,808	0,066	0,701	0,299
Fêmur	0,926	-0,054	0,143	0,887	0,281	0,126	0,881	0,119
Tíbia	1,000	0,000	0,000	0,941	0,336	-0,026	1,000	0,000
Úmero	0,874	-0,010	0,470	0,816	0,349	0,444	0,985	0,015
Cúbito	0,894	-0,034	0,331	0,846	0,314	0,310	0,910	0,090
Variância	4,001	0,536	0,424	3,317	1,306	0,338	4,961	
% variabilidade	66,7%	8,9%	7,1%	55,3%	21,8%	5,6%	82,7%	

• Matriz de Resíduos

Matrix M9

0,0000000	0,0000290	-0,0033400	-0,0000000	-0,0005029	0,0052478
0,0000290	0,0000000	0,0013411	-0,0000000	0,0000741	-0,0010782
-0,0033400	0,0013411	0,0000000	-0,0000000	-0,0001735	0,0008743
-0,0000000	-0,0000000	-0,0000000	0,0000000	-0,0000000	-0,0000000
-0,0005029	0,0000741	-0,0001735	-0,0000000	0,0000000	-0,0001086
0,0052478	-0,0010782	0,0008743	-0,0000000	-0,0001086	0,0000000

Técnicas Multivariadas em Saúde - 2015

- **Comentários:**
 - √ Após a rotação, os dois métodos de solução parecem fornecer resultados diferentes
 - √ Componentes principais:
 - F_1 : Todos os fatores, exceto X_1 e X_2 .
 - F_2 e F_3 : cada um com uma única variável
 - Solução por três fatores comuns parece estar garantida (F_3 explica uma quantidade significativa da variância)
 - √ Máxima Verossimilhança:
 - F_1 : Todos os fatores exceto X_1 e X_2
 - F_2 : Dimensão cabeça (X_1 e X_2)
 - F_3 : não é claro (provavelmente não é necessário)
 - Matriz de resíduos com elementos bastante pequenos (retenção de 3 ou menos fatores)

Técnicas Multivariadas em Saúde - 2015

• **Plot dos escores de F_1 e F_2**

- √ Escores por Mínimos Quadrados Ponderados de estimavas de máxima verossimilhança com rotação

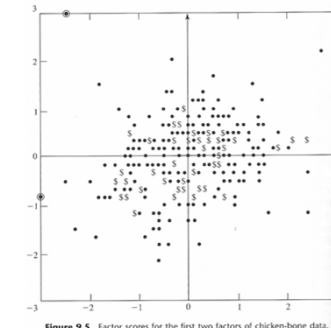


Figure 9.5 Factor scores for the first two factors of chicken-bone data.

- Gráficos deste tipo permitem identificar observações que não são consistentes com o restante das observações

Técnicas Multivariadas em Saúde - 2015

• **Plot dos escores fatoriais obtidos por componentes principais e por máxima verossimilhança**

- √ Se os loadings de um particular fator concordam entre si, os pares de escores deveriam se agrupar próximos à identidade
- √ Conjuntos de loadings que não concordam produzirão escores fatoriais que se desviam deste padrão
 - Usualmente, associado com o último fator, podendo sugerir que o número de fatores é muito grande (últimos fatores não significativos)

Técnicas Multivariadas em Saúde - 2015

√ F_1 – Componentes Principais vs. Máxima Verossimilhança

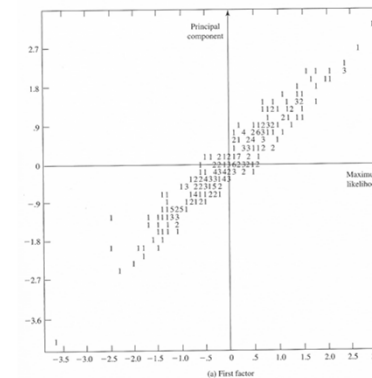
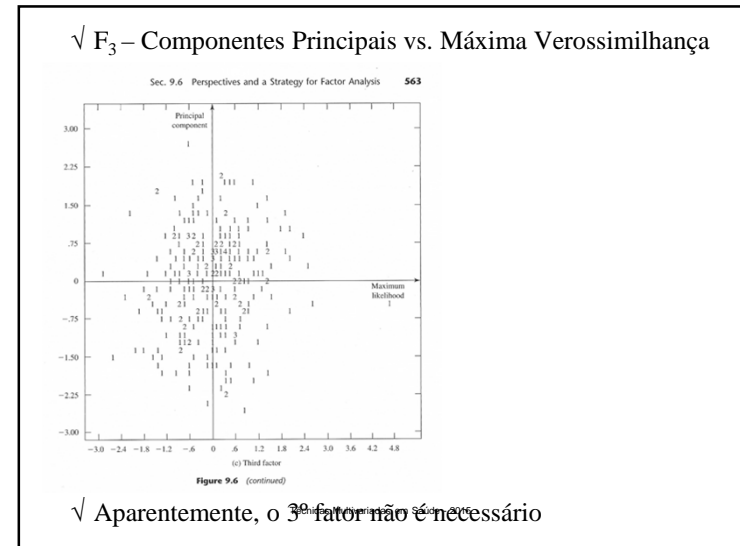
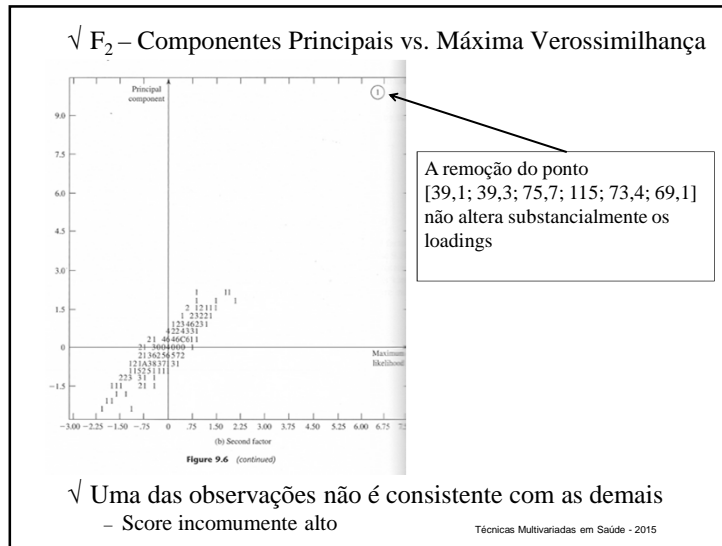


Figure 9.6 Pairs of factor scores for the chicken-bone data. (Loadings are estimated by principal component and maximum likelihood methods.)

Técnicas Multivariadas em Saúde - 2015



- Divisão do conjunto de dados em duas partes iguais (o conjunto de dados é grande)
 - √ Matrizes de correlação
- Técnicas Multivariadas em Saúde - 2015

√ Estimativas dos loadings por Componentes Principais com rotação ($m=3$)

Variável	Componentes Principais - Loadings c/ Rotação									
	1º Conjunto ($n_1 = 137$ obs.)					2º Conjunto ($n_2 = 139$ obs.)				
	L_{11}	L_{12}	L_{13}	h_1^2	Ψ_1	L_{21}	L_{22}	L_{23}	h_2^2	Ψ_2
Comprimento	0,360	0,361	0,853	0,988	0,012	0,352	0,921	0,167	1,000	0,000
Amplitude	0,303	0,899	0,312	0,997	0,003	0,203	0,145	0,968	0,999	0,001
Fêmur	0,914	0,238	0,175	0,923	0,077	0,930	0,239	0,130	0,939	0,061
Tibia	0,877	0,270	0,242	0,901	0,099	0,925	0,248	0,187	0,952	0,048
Úmero	0,830	0,247	0,395	0,906	0,094	0,912	0,252	0,208	0,939	0,061
Cúbito	0,871	0,231	0,332	0,922	0,078	0,914	0,272	0,168	0,938	0,062
Variância	3,273	1,182	1,180	5,636		3,553	1,125	1,088	5,766	
% variabilidade	54,6%	19,7%	19,7%	93,9%		59,2%	18,8%	18,1%	96,1%	

– Os resultados das duas metades são bastante similares

Técnicas Multivariadas em Saúde - 2015

- Conclusões

- √ Fatores F_2^* e F_3^* trocam de posição, mas coletivamente representam as dimensões da cabeça
- √ Fator F_1^* aparenta ser dimensões do corpo (pernas e asas)
- √ A solução é estável
 - É a mesma interpretação dada pelo conjunto completo
- √ Parece que modelo com um ou dois fatores é suficiente para ajustar os dados

Técnicas Multivariadas em Saúde - 2015

Outras Medidas de Ajuste do Modelo

- √ Critério de Kaiser-Meyer-Olkin (KMO)
- √ Teste de Esfericidade de Bartlett para R

Técnicas Multivariadas em Saúde - 2015

Critério de Kaiser-Meyer-Olkin

- √ Rencher (2002) sugere que para um modelo de Análise Fatorial possa ser ajustado adequadamente aos dados é necessário que R^{-1} seja próxima da matriz diagonal

- O coeficiente KMO baseia-se nesse princípio

$$KMO = \frac{\sum_{i \neq j} R_{ij}^2}{\sum_{i \neq j} R_{ij}^2 + \sum_{i \neq j} Q_{ij}^2}$$

- √ $R_{ij} = \text{Corr}(X_i, X_j)$
- √ $Q_{ij} =$ Correlação parcial entre 2 variáveis quando todas as outras variáveis são consideradas constantes
- √ $Q_{ij} \approx \text{zero} \rightarrow KMO \approx 1 \rightarrow R^{-1} \approx \text{diagonal}$

Técnicas Multivariadas em Saúde - 2015

- Adequabilidade do ajuste de um modelo de Análise Fatorial (Rice, 1977)

- √ modelo adequado: $KMO \geq 0,8$
- √ modelo excelente: $KMO \geq 0,9$
- √ modelo péssimo: $KMO \leq 0,5$

Técnicas Multivariadas em Saúde - 2015

Teste de Bartlett

- Teste de Esfericidade da matriz de correlação
 - √ Suposições:
 - variáveis provenientes de distribuição normal multivariada
 - modelo de Análise Fatorial pressupõe que as variáveis respostas são correlacionadas entre si
 - √ Teste de hipótese para verificar se a matriz de correlação populacional é próxima ou não da identidade

Técnicas Multivariadas em Saúde - 2015

- Hipóteses:
 - √ $H_0: \boldsymbol{\rho} = \mathbf{I}$ vs. $H_1: \boldsymbol{\rho} \neq \mathbf{I}$
- Estatística de teste:

$$T = - \left[n - \frac{1}{6}(2p + 1) \right] \sum_{i=1}^p \ln \hat{\lambda}_i$$

- √ $\hat{\lambda}_i$: autovalores da matriz de correlação amostral R
- Sob H_0 : $T \sim \chi^2_{gl}$, com $gl = \frac{1}{2} p(p - 1)$
 - √ Para se ajustar o modelo de Análise Fatorial, o teste de Bartlett deve rejeitar H_0 .

Técnicas Multivariadas em Saúde - 2015

Comentários

- Análise fatorial permanece muito subjetiva
 - √ Exemplos em que o modelo oferece explicações razoáveis em termos de poucos fatores interpretáveis
 - √ Infelizmente o critério para julgar a qualidade de qualquer análise fatorial não têm sido bem quantificado
- A qualidade do ajuste parece depender do critério Huau (wow)
 - √ Huau, eu compreendi estes fatores

Técnicas Multivariadas em Saúde - 2015

Referências

Técnicas Multivariadas em Saúde - 2015

Bibliografia Recomendada

- MANLY, B. J. F. *Métodos Estatísticos Multivariados: uma Introdução*. Bookman, 2008.
- JOHNSON, R. A.; WINCHERN, D. W. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2007
- MINGOTI, D.C. *Análise de Dados através de Métodos de Estatística Multivariada*. Ed. UFMG, 2005.
- EVERITT, B.; HOTHORN, T. *An Introduction to Applied Multivariate Analysis with R*. Springer, 2011.

Técnicas Multivariadas em Saúde - 2015