Universidade de São Paulo
B R A S I L

Análise estatística multivariada aplicada a processos químicos

- Escola Politécnica
- Departamento de Engenharia Química

- Prof. Dr. Cláudio Augusto Oller do Nascimento
- Prof. Dr. Roberto Guardani

- 2007

Parte 4. ANÁLISE DE FATORES

Introdução

Análise de Fatores é uma técnica voltada à identificação de ***fatores***, que são **entidades não observáveis**, que podem afetar uma ou mais variáveis em um conjunto de dados que contém um grande número de variáveis mensuráveis. Os **fatores**, ou **fatores comuns**, não observáveis, podem corresponder a processos implícitos, relacionados com a variabilidade dos dados.

Análise de fatores foi usada pela primeira vez por Spearman (1904), para explicar a relação entre inteligência e desempenho de estudantes em provas de diferentes disciplinas. O uso da técnica causou polêmica na época.

Sob o aspecto matemático, a análise de fatores (**AF**) tem muita semelhança com **PCA,** pois ambas as técnicas visam extrair informações a partir da matriz de covariância, Σ (*pxp*). No entanto, conceitualmente são técnicas diferentes: **PCA** consiste na obtenção de combinações das variáveis mensuráveis originais, combinações essas que possibilitam descrever o sistema utilizando menor número dessas novas variáveis obtidas, que são os componentes principais;. **AF** é uma técnica exploratória, que consiste em identificar, a partir da covariância entre variáveis mensuráveis, fatores não mensuráveis, implícitos, que podem explicar a estrutura da matriz de covariância. A influência dos fatores é quantificável.

A hipótese adotada em **AF** é que medidas de uma variável qualquer, x_j, podem ser expressas como o resultado da soma dos efeitos de um conjunto de fatores comuns, F_k, não mensuráveis e independentes entre si, mais um fator específico da variável ε_j, da seguinte forma:

$$x_{i,j} = \sum_{k=1}^{f} \ell_{jk} F_k + \varepsilon_j \qquad (4.1)$$

Os fatores comuns afetam todas as variáveis x_j, $j=1,\ldots p$, de forma diferente, dependendo dos pesos, ℓ_{jk} .

Exemplo de aplicação:

Foram medidas as notas obtidas por um grupo de alunos em diferentes disciplinas, representadas pelas seguintes variáveis:

E – inglês, F – física,
H – história, M – matemática,
P – português Q = química

Matriz de correlação obtida:

	E	F	H	M	P	Q
E	1					
F	0,380	1				
H	0,686	0,351	1			
M	0,320	0,620	0,284	1		
P	0,730	0,430	0,735	0,370	1	
Q	0,360	0,510	0,336	0,540	0,405	1

Supondo que as variáveis medidas (as notas dos alunos) podem ser influenciadas por dois fatores comuns, mais um fator específico, pode-se obter o seguinte conjunto de equações para cada variável medida:

$$E = 0{,}20F_1 + 0{,}80F_2 + \varepsilon_E \qquad F = 0{,}70F_1 + 0{,}30F_2 + \varepsilon_F$$
$$H = 0{,}15F_1 + 0{,}82F_2 + \varepsilon_H \qquad M = 0{,}80F_1 + 0{,}20F_2 + \varepsilon_M$$
$$P = 0{,}25F_1 + 0{,}85F_2 + \varepsilon_P \qquad Q = 0{,}60F_1 + 0{,}30F_2 + \varepsilon_Q$$

Observa-se que as correlações entre os pesos de quaisquer duas notas correspondem aos coeficientes de correlação da matriz de correlação (ex.: correlação entre as notas de matemática e história: 0,80x0,15 + 0,20x0,82 = 0,284). No modelo acima, os fatores comuns considerados podem ser relacionados a características como:

F_1 = habilidade quantitativa de cada aluno;
F_2 = habilidade comunicativa de cada aluno,

e aos fatores ε_j, específicos de cada disciplina.

Um resultado adicional do tratamento é a quantificação do efeito dos fatores comuns e dos específicos, feita a partir dos conceitos de comunalidade e de variância específica de cada variável. A *comunalidade* é definida como a fração da variância total do sistema abrangida pelos *fatores comuns*. A *variância específica* representa a fração não abrangida pelos fatores comuns e é atribuída aos *fatores específicos*. A seguinte tabela representa esses resultados:

Variável	Comunalidade F_1	Comunalidade F_2	Comunalidade Total	Variância Específica
E	0,040	0,640	0,680	0,320
F	0,490	0,090	0,580	0,420
H	0,023	0,672	0,695	0,305
M	0,640	0,040	0,680	0,320
P	0,063	0,723	0,786	0,214
Q	0,360	0,090	0,450	0,550
Total	1,616	2,255	3,871	2,219

Formulação matemática

Dado um conjunto de p variáveis centradas na média, formando um vetor X (px1), com matriz de covariância Σ (pxp), o modelo de fatores supõe que X pode ser expresso como uma combinação linear de alguns fatores comuns F_1, F_2,…, F_m, e p fontes adicionais de variação, ε_1, ε_2,…, ε_p, que são os fatores específicos, ou erros. O modelo de fatores é expresso como:

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

Ou, em notação matricial:

$$
\underset{(p\times 1)}{\mathbf{X}} = \underset{(p\times m)}{\mathbf{L}}\,\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}} \tag{2}
$$

Os coeficientes ℓ_{jk} são os pesos do fator k sobre a variável j. A matriz **L** (pxm) é chamada “matriz de pesos”. Cada fator específico, ε_j está associado à variável X_j. Portanto, trata-se de um modelo com p + m variáveis aleatórias, não observáveis (os fatores comuns e os fatores específicos). A solução desse sistema pode ser obtida com as seguintes hipóteses adicionais. Primeiro,supõe-se que os fatores F_k e ε_j são centrados nos valores médios e independentes entre si (os fatores são ortogonais):

$$
E(\mathbf{F}) = \underset{(m\times 1)}{\mathbf{0}}, \qquad E(\boldsymbol{\varepsilon}) = \underset{(p\times 1)}{\mathbf{0}} \qquad Cov(\mathbf{F}) = E\left(\mathbf{F}\mathbf{F}^T\right) = \underset{(m\times m)}{\mathbf{I}}
$$

$$
Cov(\boldsymbol{\varepsilon}) = E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right) = \underset{(p\times p)}{\boldsymbol{\Psi}} = \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Psi_p \end{bmatrix} \tag{3}
$$

Admite-se que F e ε são independentes, ou seja:

$$Cov(\boldsymbol{\varepsilon}, \mathbf{F}) = E(\boldsymbol{\varepsilon}\mathbf{F}^T) = \underset{(p \times m)}{\mathbf{0}} \tag{4}$$

As Eqs. 1 a 4 constituem o <u>modelo de fatores ortogonais</u>. A solução implica resolver um modelo de covariância de X:

$$\begin{aligned}
\boldsymbol{\Sigma} = Cov(\mathbf{X}) &= E(\mathbf{X}\mathbf{X}^T) = E\left((\mathbf{LF} + \boldsymbol{\varepsilon})(\mathbf{LF} + \boldsymbol{\varepsilon})^T\right) \\
&= E\left(\mathbf{LF}(\mathbf{LF})^T + \boldsymbol{\varepsilon}(\mathbf{LF})^T + \mathbf{LF}\boldsymbol{\varepsilon}^T + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right) \\
&= \mathbf{L}E(\mathbf{F}\mathbf{F}^T)\mathbf{L}^T + E(\boldsymbol{\varepsilon}\mathbf{F}^T)\mathbf{L}^T + \mathbf{L}E(\mathbf{F}\boldsymbol{\varepsilon}^T) + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)
\end{aligned}$$

Aplicando-se as hipóteses das Eqs. 3 e 4, chega-se a:

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} \tag{5}$$

Os elementos da diagonal principal de Σ, representando a variância de X_i, correspondem a:

$$Var(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2 + \psi_i \tag{5-1}$$

As hipóteses da Eq. 3 também implicam a seguinte relação para a covariância entre X e F:

$$\begin{aligned}
Cov(\mathbf{X}, \mathbf{F}) &= E\left((\mathbf{LF} + \boldsymbol{\varepsilon})\mathbf{F}^T\right) = E\left(\mathbf{LF}\mathbf{F}^T + \boldsymbol{\varepsilon}\mathbf{F}^T\right) \\
&= \mathbf{L}E(\mathbf{F}\mathbf{F}^T) + E(\boldsymbol{\varepsilon}\mathbf{F}^T) = \mathbf{L} \\
Cov(X_i, F_j) &= \ell_{ij}
\end{aligned} \tag{5-2}$$

Componentes da variância de X:

$$Var(X_i) = \underbrace{\ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2}_{\text{Comunalidade}} + \underbrace{\psi_i}_{\substack{\text{Variância} \\ \text{Específica}}} \tag{6}$$

Portanto, comunalidade é a fração da variância total de cada variável X_i representada pelos fatores comuns, F_j (j =1,…,m). Corresponde à soma dos quadrados dos coeficientes dos fatores comuns, na Eq. 1. O restante da variância de X_i é representado pela variância específica, ψ_i.

Inexistência de solução única para o problema

O modelo de covariância apresentado na Eq. 5 não tem solução única, ou seja, é possível haver diferentes matrizes de pesos, **L**, correspondentes a uma mesma matriz de covariância, Σ. Para ilustrar isso, seja uma matriz ortogonal **T** (*m x m*). Portanto:

$$\mathbf{T}\mathbf{T}^T = \mathbf{T}^T\mathbf{T} = \mathbf{I}$$

A Eq. 2 pode ser escrita:

$$\mathbf{X} = \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}\mathbf{T}\mathbf{T}^T\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\varepsilon} \tag{7}$$

em que:

$$\mathbf{L}^* = \mathbf{L}\mathbf{T} \quad \text{e} \quad \mathbf{F}^* = \mathbf{T}^T\mathbf{F}$$

como

$$E(\mathbf{F}^*) = \mathbf{T}^T E(\mathbf{F}) = \mathbf{0}$$

e

$$Cov(\mathbf{F}^*) = E(\mathbf{F}^*\mathbf{F}^{*T}) = E(\mathbf{T}^T\mathbf{F}\mathbf{F}^T\mathbf{T}) = E(\mathbf{T}^T\mathbf{T}) = \mathbf{I}$$

então, ambos os vetores **F** e **F*** têm as mesmas propriedades estatísticas. A Eq. 5 é satisfeita tanto com a matriz **L** quanto com **L***:

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi} = \mathbf{L}\mathbf{T}\mathbf{T}^T\mathbf{L}^T + \boldsymbol{\Psi} = \mathbf{L}^*\mathbf{L}^{*T} + \boldsymbol{\Psi} \tag{8}$$

Objetivos específicos da análise de fatores

1. Identificar o menor número de Fatores Comuns, que melhor explicam as correlações entre as variáveis mensuráveis.
2. Identificar, via rotações, a solução mais plausível.
3. Estimar os pesos, comunalidades e variâncias específicas das variáveis mensuráveis.
4. Prover uma interpretação para os fatores comuns.
5. Se necessário, estimar os valores dos fatores (“factor scores”).

Principais técnicas de análise

Fatoração por Componentes Principais ("Principal Component Factoring").

Consiste na mesma metodologia usada para estimar os componentes principais, baseada no cálculo dos autovalores e autovetores da matriz de covariância, ou de correlação (para dados padronizados).

Exemplo:
A seguinte matriz de correlação refere-se a dados horários em um reator catalítico de desidrogenação de parafinas (200 horas). Lista de variáveis medidas:

1) Vazão de carga para o reator;
2) Concentração de produto na saída do reator;
3) Pressão na entrada do reator;
4) Temperatura de alimentação do fluido refrigerante;
5) Temperatura de saída do fluido refrigerante;
6) Vazão de fluido refrigerante.

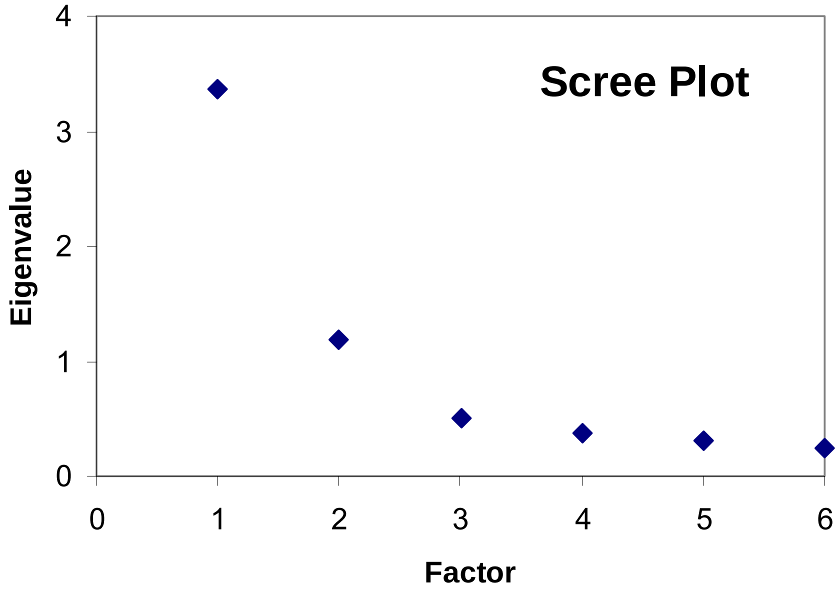
Variáveis	**1**	**2**	**3**	**4**	**5**	**6**
1	1.000	0.620	0.540	0.320	0.284	0.370
2	0.620	1.000	0.510	0.380	0.351	0.430
3	0.540	0.510	1.000	0.360	0.336	0.405
4	0.320	0.380	0.360	1.000	0.686	0.730
5	0.284	0.351	0.336	0.686	1.000	0.735
6	0.370	0.430	0.405	0.730	0.735	1.000

Resultado da aplicação da técnica de PCA, com 6 fatores considerados.

				Fatores		
Variáveis	**1**	**2**	**3**	**4**	**5**	**6**
1	-0.675	0.557	-0.190	0.444	0.027	-0.021
2	-0.718	0.447	-0.346	-0.405	-0.003	-0.019
3	-0.683	0.418	0.592	-0.093	-0.002	-0.012
4	-0.793	-0.410	-0.015	0.040	-0.416	-0.169
5	-0.774	-0.461	-0.002	0.007	0.372	-0.223
6	-0.838	-0.359	-0.016	0.021	0.032	0.409
Autovalores	**3.367**	**1.194**	**0.507**	**0.372**	**0.313**	**0.247**

Na tabela, a Variável 1, por exemplo, é representada por:

$$X_1 = -0.675F_1 + 0.557F_2 - 0.190F_3 + 0.444F_4 + 0.027F_5 - 0.021F_6$$



Aplicando-se o critério de mínimo autovalor = 1, extraem-se 2 fatores.

O resultado da análise é:

	Pesos dos fatores							
Variáveis	**Fator 1**	**Fator 2**	**Comunalidade**	**Fator 3**	**Fator 4**	**Fator 5**	**Fator 6**	**Var. Específica**
1	-0.675	0.557	0.766	-0.190	0.444	0.027	-0.021	0.234
2	-0.718	0.447	0.716	-0.346	-0.405	-0.003	-0.019	0.284
3	-0.683	0.418	0.641	0.592	-0.093	-0.002	-0.012	0.359
4	-0.793	-0.410	0.797	-0.015	0.040	-0.416	-0.169	0.203
5	-0.774	-0.461	0.812	-0.002	0.007	0.372	-0.223	0.188
6	-0.838	-0.359	0.831	-0.016	0.021	0.032	0.409	0.169
Autovalores	**3.367**	**1.194**	**4.561**	**0.507**	**0.372**	**0.313**	**0.247**	**1.439**

Variância total: 6 (comunalidade + específica); Variância dos 2 fatores: 76% da variância total (=4,561/6).

Considerando-se os 2 fatores extraídos, podem-se expressar esses mesmos fatores a partir da inversão da matriz de pesos, **L** (6x6), obtendo-se a matriz $\mathbf{L}^{-1}$. Retendo-se apenas dois fatores, obtem-se a matriz com os coeficientes para estimar os fatores comuns a partir de diferentes valores dos indicadores:

Variable	**Factor1**	**Factor2**
Var 1	-0.200	0.467
Var 2	-0.213	0.374
Var 3	-0.203	0.350
Var 4	-0.236	-0.343
Var 5	-0.230	-0.386
Var 6	-0.249	-0.301

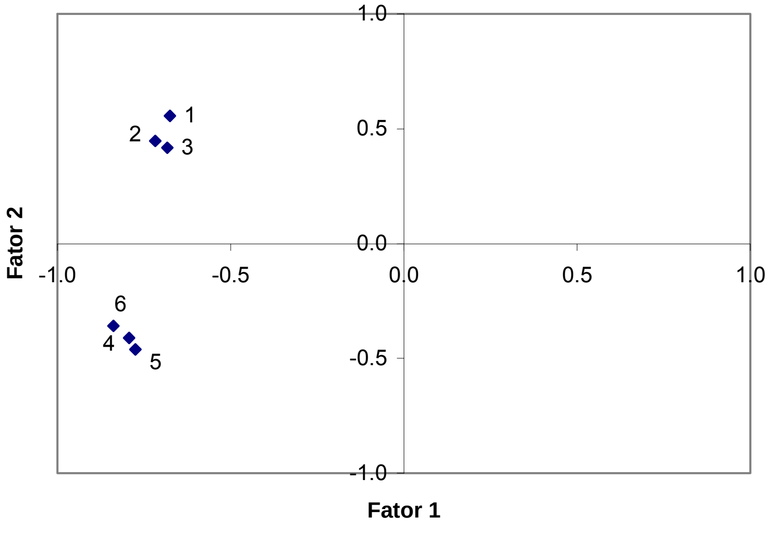
A parte da variância não explicada pelos 2 fatores pode ser calculada da seguinte forma:

$$\underset{(p\times p)}{\mathbf{R}} = \mathbf{\Sigma} - \mathbf{L}\mathbf{L}^T$$

obtendo-se a matriz de resíduos, **R**:

Variáveis	**1**	**2**	**3**	**4**	**5**	**6**
1	0.234	-0.114	-0.154	0.013	0.018	0.005
2	-0.114	0.284	-0.167	-0.006	0.001	-0.011
3	-0.154	-0.167	0.359	-0.010	0.000	-0.017
4	0.013	-0.006	-0.010	0.203	-0.117	-0.082
5	0.018	0.001	0.000	-0.117	0.188	-0.079
6	0.005	-0.011	-0.017	-0.082	-0.079	0.169

Um gráfico de dispersão dos pesos dos dois fatores para as 6 variáveis mensuráveis é:



Uma vez que não existe solução única para a Eq. 5 (ver Eq. 8), podem-se tentar outras combinações lineares dos fatores, para facilitar a interpretação.

Técnicas de rotação.

A rotação dos fatores constitui-se em uma transformação ortogonal dos pesos (e conseqüentemente dos fatores) visando compatibilizar os pesos com interpretações plausíveis dos fatores. Seja uma matriz ortogonal **T** (*m* x *m)*. Portanto:

$$\mathbf{T}\mathbf{T}^T = \mathbf{T}^T\mathbf{T} = \mathbf{I}$$

Conforme visto, podem-se ter matrizes de fatores $\mathbf{L}^* = \mathbf{L}\mathbf{T}$, que satisfazem a equação $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi}$, pois:

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi} = \mathbf{L}\mathbf{T}\mathbf{T}^T\mathbf{L}^T + \mathbf{\Psi} = \mathbf{L}^*\mathbf{L}^{*T} + \mathbf{\Psi} \quad (8)$$

Podem-se escolher, então, matrizes de coeficientes $\mathbf{L}^*$ de modo a facilitar a interpretação dos pesos em função do sistema em estudo.

<u>Rotação Quartimax</u>

Rotação ortogonal que objetiva tornar os pesos de todos os indicadores altos em relação a *um dos fatores*. Além disso, cada indicador deve ter peso alto em relação a um dos fatores e peso baixo em relação aos demais fatores. Isto é obtido maximizando-se a variância dos pesos <u>em relação aos fatores</u>. Para qualquer variável X_i define-se:

$$Q_i = \frac{\sum_{j=1}^{m}\left(\ell_{ij}^2 - \lambda_i^2\right)^2}{m} \quad (9)$$

Em que Q_i representa a variância das comunalidades (quadrado dos pesos) da variável *i*, ℓ_{ij} é o peso do fator *j* para a variável *i* e λ^2_i é o valor médio do quadrado dos pesos para a variável *i*.

O método consiste em selecionar a matriz $\mathbf{L}^*$ tal que a função $Q = \sum_{i=1}^{p} Q_i$ seja maximizada.

<u>Varimax</u>

A rotação ortogonal Varimax visa fazer com que cada variável tenha máximo peso correspondendo a um e apenas um dos fatores. Assim, deve ser obtida a matriz **L*** tal que maximize a variância V_j do quadrado dos pesos <u>em relação às variáveis</u>. Para qualquer fator F_j define-se:

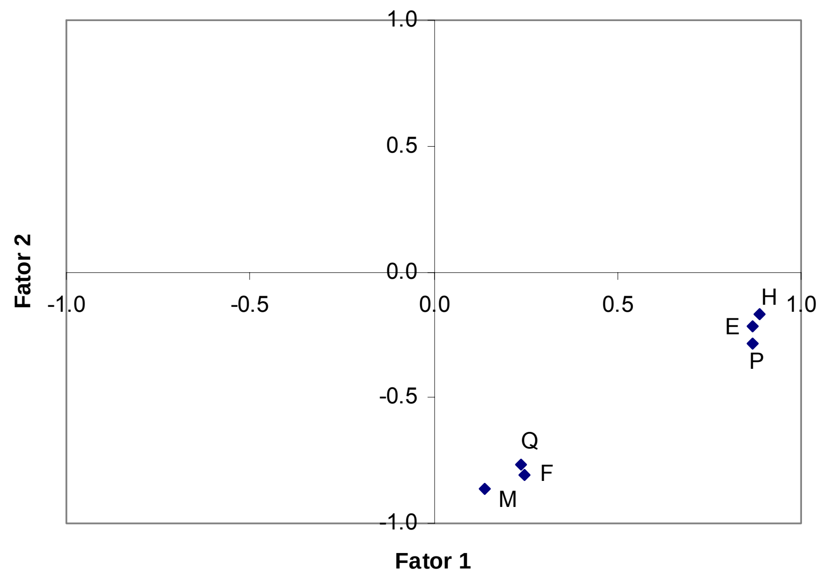
$$V_j = \frac{\sum_{j=1}^{p}\left(\ell_{ij}^2 - \lambda_j^2\right)^2}{p} \qquad (10)$$

em que λ^2_j é o valor médio do quadrado dos pesos do fator j. O método consiste em selecionar a matriz **L*** tal que a função $V = \sum_{j=1}^{m} V_j$ seja maximizada.

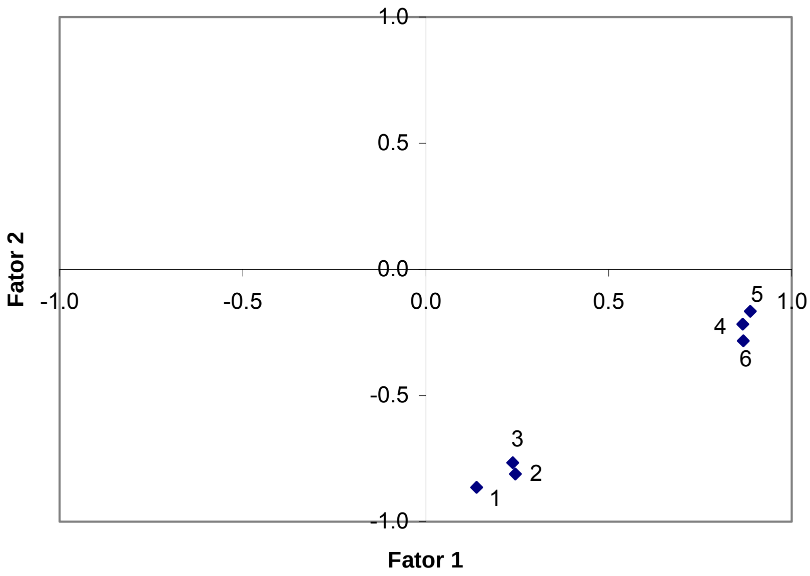
Em ambas as rotações a estrutura de fatores (e, portanto, a comunalidade) não é alterada.

Há outras rotações ortogonais propostas na literatura, assim como há casos de rotações <u>oblíquas</u>.

Resultado de Rotação Quartimax



Resultado de Rotação Varimax



No caso, ambas as rotações (Varimax e Quartimax) resultaram nos mesmos valores para os pesos. O resultado depende do sistema em estudo.

Fatoração pelo Eixo Principal ("Principal Axis Factoring").

Trata-se de método iterativo, baseado nos componentes principais. Parte-se da hipótese de que as comunalidades (na diagonal principal da matriz de correlação) são todas iguais a 1. Na primeira etapa, são calculados os pesos dos fatores da mesma que no método de componentes principais. A partir da segunda etapa, a matriz de correlação é modificada a cada iteração a partir dos pesos dos fatores extraídos, inserindo-se as novas comunalidades na diagonal principal. O procedimento de cálculo dos componentes principais é repetido até que não haja mais variação nas comunalidades das variáveis. No exemplo, 9 iterações foram necessárias para a variação da comunalidade passar a ser menor que a tolerância adotada (0,001):

Communalities

Iteration	Variation	1	2	3	4	5	6
1	0.359	0.766	0.714	0.641	0.797	0.812	0.829
2	0.128	0.698	0.626	0.513	0.725	0.744	0.784
3	0.042	0.679	0.598	0.471	0.698	0.719	0.774
4	0.014	0.675	0.588	0.457	0.688	0.708	0.774
5	0.005	0.674	0.585	0.453	0.684	0.703	0.776
6	0.003	0.675	0.583	0.451	0.682	0.700	0.779
7	0.002	0.676	0.582	0.451	0.681	0.698	0.781
8	0.001	0.677	0.582	0.451	0.681	0.697	0.782
9	0.001	0.677	0.581	0.450	0.680	0.697	0.783

Comentários finais

A Análise de Fatores pode ser considerada, do ponto de vista de procedimentos de cálculo, como uma extensão da técnica de Análise de Componentes Principais, pois ambas as técnicas consistem em extrair informações da matriz de covariância (ou de correlação, para variáveis padronizadas). No entanto, a Análise de Fatores é mais elaborada, por envolver manipulações adicionais visando facilitar a interpretação dos resultados.

Embora relacionadas sob o aspecto da técnica, ambas as análises são conceitualmente distintas:

Em "PCA" o objetivo é a redução do número de variáveis do modelo, buscando-se combinações lineares que representam o comportamento das variáveis originais.

Na Análise de Fatores busca-se explicar as causas para as correlações entre as variáveis, ou associar essas correlações com processos não mensuráveis, ou mesmo confirmar hipóteses referentes a esses processos (que correspondem aos fatores comuns).

Ambas as técnicas baseiam-se na covariância entre as variáveis. Portanto, se houver associações entre variáveis que não tenham efeito sobre a covariância, tanto a técnica de PCA quanto a Análise de Fatores ficam prejudicadas.