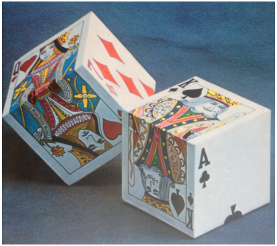
Fundamentos Estatísticos para Ciência dos Dados
Análise Fatorial

Renato Martins Assunção

DCC, UFMG - 2020



Análise Fatorial

- $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$, vetor p_dim
- Imagine que existem DUAS variáveis F_1 e F_2 que NÃO SÃO diretamente observáveis.
- Dizemos que são LATENTES
- F_1 e F_2 são independentes.
- Cada uma das X_i em $\underset{\sim}{X}$ é aproximadamente uma combinação linear de F_1 e F_2.

Análise Fatorial

- Isto é,
  $X_1 \approx \mu_1 + \ell_{11} F_1 + \ell_{12} F_2$
  $X_2 \approx \mu_2 + \ell_{21} F_1 + \ell_{22} F_2$
  .
  .
  .
  $X_p \approx \mu_p + \ell_{p1} F_1 + \ell_{p2} F_2$
- ℓ_{ij} = constantes desconhecidas
  = carga dos fatores
  = factor loading

Análise Fatorial

- Exemplo:
  $\underset{\sim}{X}$ = Vetor com notas de um aluno em 15 assuntos

X_1 = Gramática
X_2 = Literatura
X_3 = Redação
X_4 = História
X_5 = Geografia
X_6 = Filosofia
X_7 = Álgebra
X_8 = Geometria
X_9 = Lógica
X_{10} = Química
X_{11} = Física
X_{12} = Biologia
X_{13} = Inglês
X_{14} = Sociologia
X_{15} = Espanhol

Análise Fatorial

- Desempenho individual nestes 15 assuntos é muito variado:
    - Alguns vão bem em todas as disciplinas;
    - Alguns vão bem apenas em algumas e tem um desempenho médio nas outras;
    - Alguns vão muito bem em algumas e muito mal em outras;
    - Alguns vão mal em todas.
- Como entender essa diversidade?

Análise Fatorial

- Psicólogos se perguntaram se esta variabilidade na capacidade cognitiva não poderia ser explicada pela existência de uns poucos traços latentes, não observados.
- Por exemplo,
  FV = Fator refletindo habilidade verbal
  FQ = Fator refletindo habilidade lógico-quantitativa
- Os fatores tem uma escala centrada em zero.

Análise Fatorial

- Se
  $F = 0 \Rightarrow$ Indivíduo tem habilidade média no fator.
  $F > 0 \Rightarrow$ Habilidade acima da média.
  $F \gg 0 \Rightarrow$ Habilidade muito acima da média.
  $F < 0 \Rightarrow$ Habilidade abaixo da média.

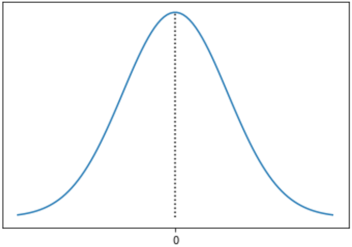


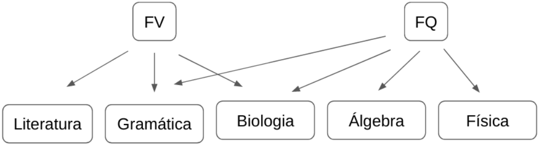
Figura: Densidade da distribuição de um fator na população.

Análise Fatorial

- Cada insivíduo recebe uma “dose” de FV e uma “dose” de FQ.
- Doses de FV e FQ são independentes.
  Por exemplo, alguns recebem muito de FV.
  Dentre estes,
  - Metade recebe $FQ+$
  - Metade recebe $FQ-$
- As notas nos 15 assuntos são reflexos e combinações desses dois fatores, além de um ruído causado por outros fatores que não levamos em conta.

Análise Fatorial

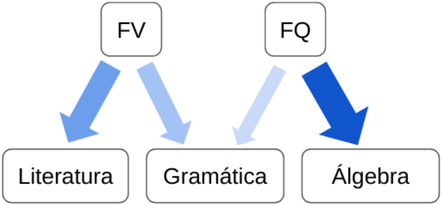
- Modelo hierárquico para a geraçáo das 15 notas de um mesmo indivíduo.
- Passo 1: Recebe doses independentes de FV e FQ.
  $FV-$ $\quad$ $FQ+$
- Passo 2: Estas doses afetam as notas dos 15 assuntos.



- Ausência de aresta $\Rightarrow$ sem influência do fator.

Análise Fatorial

- Peso ou importância da aresta representada por grossura da aresta.



- Peso da aresta é a carga do fator (factor loading).

Representação algébrica

- Gramática $= X_1 \approx \mu_{Gram} + \ell_{GV} * FV + \ell_{GQ} * FQ$
  Literatura $= X_2 \approx \mu_{Lit} + \boldsymbol{\ell_{LV}} * FV + \boldsymbol{\ell_{LQ}} * FQ$

  .

  .

  .

  Carga dos fatores.

Representação algébrica

- Gramática $= X_1 \approx \mu_{Gram} + \ell_{GV} * \boldsymbol{FV} + \ell_{GQ} * FQ$
  Literatura $= X_2 \approx \mu_{Lit} + \ell_{LV} * \boldsymbol{FV} + \ell_{LQ} * FQ$

  .

  .

  .

  Fator verbal do indivíduo. O mesmo para todos os assuntos.

Representação algébrica

- Mais formal:

$$X_1 = \mu_{Gram} + \ell_{GV} * FV + \ell_{GQ} * FQ + \varepsilon_{\boldsymbol{Gram}}$$
$$X_2 = \mu_{Lit} + \ell_{LV} * FV + \ell_{LQ} * FQ + \varepsilon_{\boldsymbol{Lit}}$$

.

.

.

$$X_{15} = \mu_{Esp} + \ell_{EV} * FV + \ell_{EQ} * FQ + \varepsilon_{\boldsymbol{Esp}}$$

Erros ou fatores não observados.

Representação algébrica

- Um exemplo esquemático.
  - Assuntos muito associados com <u>FV</u>:
    $X_1 = \mu_1 + (1.0) * FV + (0.1) * FQ + \varepsilon_1$
    $X_2 = \mu_2 + (0.8) * FV + (0.01) * FQ + \varepsilon_2$
  - Nenhuma das duas habilidades é muito relevante:
    $X_3 = \mu_3 + (0.1) * FV - (0.1) * FQ + \varepsilon_3$
  - Precisa ser bom em FQ <u>e</u> muito ruim em FV. :(
    $X_4 = \mu_4 - (0.3) * FV + (0.9) * FQ + \varepsilon_4$
  - Bom em FQ e pouco relevante em FV.
    $X_5 = \mu_5 - (0.1) * FV + (1.2) * FQ + \varepsilon_5$

Representação algébrica

- Representação matricial para um indivíduo:

$$\underset{\sim}{X} = \underset{\sim}{\mu} + \underset{pxm}{L} \cdot \underset{mx1}{\underset{\sim}{F}} + \underset{px1}{\underset{\sim}{\varepsilon}}$$

$L =$ matriz de carga dos fatores (loading).
$F =$ vetor dos fatores comuns (às p variáveis).
$\varepsilon =$ vetor dos erros ou fatores <u>específicos</u> (de cada variável).

Representação algébrica

- Vamos imaginar dois indivíduos com <u>seus dois vetores</u> instanciados:

$$\underset{\sim}{X}^{(1)} = \begin{pmatrix} X_1^{(1)} \\ \cdot \\ \cdot \\ X_{15}^{(1)} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \cdot \\ \cdot \\ \mu_{15} \end{pmatrix} + \begin{bmatrix} \ell_{11} & \ell_{12} \\ \cdot & \cdot \\ \cdot & \cdot \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} \begin{bmatrix} FV^{(1)} \\ FQ^{(1)} \end{bmatrix} + \begin{bmatrix} \varepsilon_1^{(1)} \\ \cdot \\ \cdot \\ \varepsilon_{15}^{(1)} \end{bmatrix}$$

$$\underset{\sim}{X}^{(2)} = \begin{pmatrix} X_1^{(2)} \\ \cdot \\ \cdot \\ X_{15}^{(2)} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \cdot \\ \cdot \\ \mu_{15} \end{pmatrix} + \begin{bmatrix} \ell_{11} & \ell_{12} \\ \cdot & \cdot \\ \cdot & \cdot \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} \begin{bmatrix} FV^{(2)} \\ FQ^{(2)} \end{bmatrix} + \begin{bmatrix} \varepsilon_1^{(2)} \\ \cdot \\ \cdot \\ \varepsilon_{15}^{(2)} \end{bmatrix}$$

Representação algébrica

- Vamos imaginar dois indivíduos com seus dois vetores instanciados:
$\underset{\sim}{X}^{(1)} = \underset{\sim}{\mu} + L \cdot \underset{\sim}{F}^{(1)} + \underset{\sim}{\varepsilon}^{(1)}$

$\underset{\sim}{X}^{(2)} = \underset{\sim}{\mu} + L \cdot \underset{\sim}{F}^{(2)} + \underset{\sim}{\varepsilon}^{(2)}$

$\underset{\sim}{\mu}$ e $L \Rightarrow$ iguais para todos os indivíduos.
$\underset{\sim}{F}^{(i)} = \begin{bmatrix} FV^{(i)} \\ FQ^{(i)} \end{bmatrix}$ = doses dos fatores recebidos pelos indivíduos.
$\underset{\sim}{\varepsilon}^{(i)}$ = erros do indivíduo (i)

Representação algébrica

- Vamos imaginar dois indivíduos com <u>seus dois vetores</u> instanciados:
  $\underset{\sim}{X}^{(1)} = \underset{\sim}{\boldsymbol{\mu}} + \boldsymbol{L} \cdot \underset{\sim}{\boldsymbol{F}}^{\boldsymbol{(1)}} + \underset{\sim}{\boldsymbol{\varepsilon}}^{\boldsymbol{(1)}}$

  $\underset{\sim}{X}^{(2)} = \underset{\sim}{\boldsymbol{\mu}} + \boldsymbol{L} \cdot \underset{\sim}{\boldsymbol{F}}^{\boldsymbol{(2)}} + \underset{\sim}{\boldsymbol{\varepsilon}}^{\boldsymbol{(2)}}$

  São iguais. Específicos do indivíduo.

Representação algébrica

- Só observamos o vetor $\underset{\sim}{X}$ em vários indivíduos.
- Queremos entender como $\underset{\sim}{X}$ varia.
- Basta entender como os fatores FV e FQ variam de indivíduo para indivíduo.
- As notas dos 15 assuntos apenas refletem e combinam estes fatores através da matriz L.
- Um pequeno ruído ε para cada disciplina é adicionado para levar em conta os demais fatores que estamos ignorando.
- Como podemos inferir L a partir dos dados?
- E os escores dos fatores $\begin{bmatrix} FV^{(i)} \\ FQ^{(i)} \end{bmatrix}$ de cada indivíduo, como obtemos?

Suposições do modelo fatorial

- Indivíduo (i) com $m = 2$ fatores.

$$\underset{px1}{\underset{\sim}{X}^{(i)}} = \underset{px1}{\underset{\sim}{\mu}} + \underset{px2}{L} \cdot \underset{2x1}{\underset{\sim}{F}^{(i)}} + \underset{px1}{\underset{\sim}{\varepsilon}^{(i)}}$$

- $\underset{\sim}{\mu}$ e L <u>são comuns</u> a todos os indivíduos, não aleatórios.
- $\underset{\sim}{F}^{(i)}$ e $\underset{\sim}{\varepsilon}^{(i)}$ variam de indivíduo para indivíduo, são aleatórios.
- $\mathbb{E}(\underset{2x1}{\underset{\sim}{F}^{(i)}}) = \mathbb{E}\begin{pmatrix} FV^{(i)} \\ FQ^{(i)} \end{pmatrix} = \begin{pmatrix} \mathbb{E}(FV^{(i)}) \\ \mathbb{E}(FQ^{(i)}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \underset{\sim}{0}$
- $Cov(\underset{2x2}{\underset{\sim}{F}^{(i)}}) = \begin{bmatrix} Var(FV^{(i)}) & Cov(FV^{(i)}, FQ^{(i)}) \\ Cov(FV^{(i)}, FQ^{(i)}) & Var(FQ^{(i)}) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Suposições do modelo fatorial

- A opção de tomar a variância de cada fator igual a 1 (e portanto, tomar o DP de cada fator = 1) é baseada no seguinte argumento:
  - Cada fator (FV ou FQ) terá a "mesma escala" indo de -2 a +2 aproximadamente ao variar dos menos habilidosos aos mais habilidosos.
  - Se o fator F_k afetar muito uma nota X_j isto será refletido numa carga ℓ_{jk} muito positiva (ou muito negativa).
  - Mas a escala de todos os fatores é a mesma (em DP's): vai de -2 a +2, aproximadamente.
- A covariância $Cov(FV^{(i)}, FQ^{(i)}) = 0$ pois estamos supondo fatores independentes.

Suposições do modelo fatorial

- Mais suposições, agora sobre $\underset{\sim}{\varepsilon}^{(i)}$:

$$\mathbb{E}(\underset{15x1}{\underset{\sim}{\varepsilon}^{(i)}}) = \mathbb{E}\begin{pmatrix} \varepsilon_1^{(i)} \\ \cdot \\ \cdot \\ \varepsilon_{15}^{(i)} \end{pmatrix} = \begin{pmatrix} \mathbb{E}(\varepsilon_1^{(i)}) \\ \cdot \\ \cdot \\ \mathbb{E}(\varepsilon_{15}^{(i)}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix} = \underset{\sim}{0}$$

Suposições do modelo fatorial

- Mais suposições, agora sobre $\underset{\sim}{\varepsilon}^{(i)}$:

$$\underset{15x15}{Cov(\underset{\sim}{\varepsilon})} = \begin{bmatrix} Var(\varepsilon_1^{(i)}) & Cov(\varepsilon_1^{(i)}, \varepsilon_2^{(i)}) & . & . & Cov(\varepsilon_1^{(i)}, \varepsilon_{15}^{(i)}) \\ Cov(\varepsilon_2^{(i)}, \varepsilon_1^{(i)}) & Var(\varepsilon_2^{(i)}) & . & . & Cov(\varepsilon_2^{(i)}, \varepsilon_{15}^{(i)}) \\ . & . & . & . & . \\ . & . & . & . & . \\ Cov(\varepsilon_{15}^{(i)}, \varepsilon_1^{(i)}) & Cov(\varepsilon_{15}^{(i)}, \varepsilon_2^{(i)}) & . & . & Var(\varepsilon_{15}^{(i)}) \end{bmatrix}$$

$$= \begin{bmatrix} \psi_1 & & & & & \\ & \psi_2 & & & & \\ & & . & & & \\ & & & . & & \\ & & & & . & \\ & & & & & \psi_{15} \end{bmatrix} = diag(\psi_1, ..., \psi_{15}) = \psi$$

Suposições do modelo fatorial

- É razoável deixar que $Var(\varepsilon_j^{(i)})$ varie.
- Podemos ter $\psi_j \neq \psi_k$.
- A razão é que a nota de redação, digamos, pode ter muito mais variabilidade que a nota de matemática devido a fatores não relacionados com FV ou FQ.
- A subjetividade do corretor da redação, a variação da qualidade da redação como fruto do conhecimento do aluno sobre o tema, entre outras causas, pode gerar mais variação na nota da redação do que a variação induzida pela diversidade de FV e FQ.

Suposições do modelo fatorial

- A covariância entre os erros de assuntos distintos, $Cov(\varepsilon_j^{(i)}, (\varepsilon_k^{(i)})$, provavelmente <u>não é</u> zero, mas deve ser pequena. Por isto façamos todas iguais a <u>zero</u> no modelo.
- Assim, adotamos $Cov(\underset{\sim}{\varepsilon}) = \begin{pmatrix} \psi_1 & & & & \\ & \cdot & & \underset{\sim}{0} & \\ & & \cdot & & \\ & \underset{\sim}{0} & & \cdot & \\ & & & & \psi_{15} \end{pmatrix} =$ diagonal.

Suposições do modelo fatorial

- Uma última suposição: $\underset{\sim}{F}$ e $\underset{\sim}{\varepsilon}$ são independentes.
- Isso implica que:

$$\underset{15x2}{Cov(\underset{\sim}{\varepsilon}, \underset{\sim}{F})} = \begin{bmatrix} Cov(\varepsilon_1, F_1) & Cov(\varepsilon_1, F_2) \\ Cov(\varepsilon_2, F_1) & Cov(\varepsilon_2, F_2) \\ . & . \\ . & . \\ Cov(\varepsilon_{15}, F_1) & Cov(\varepsilon_{15}, F_2) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ . & . \\ . & . \\ 0 & 0 \end{bmatrix}$$

- Note que, como $\underset{15x1}{\mathbb{E}(\underset{\sim}{\varepsilon})} = \underset{\sim}{0}$ e $\underset{2x1}{\mathbb{E}(\underset{\sim}{F})} = \underset{\sim}{0}$, temos $Cov(\underset{\sim}{\varepsilon}, \underset{\sim}{F}) = \mathbb{E}(\underset{\sim}{\varepsilon}\underset{\sim}{F}') = \underset{\sim}{0}$ e também $\mathbb{E}(\underset{\sim}{F}\underset{\sim}{\varepsilon}') = \underset{\sim}{0}$.
- Podemos agora obter a estrutura de convariância das observações $\underset{\sim}{X}$.

Suposições do modelo fatorial

- $\mathbb{E}(\underset{\sim}{X}) = \mathbb{E}(\underset{\sim}{\mu} + L\underset{\sim}{F} + \underset{\sim}{\varepsilon})$
  $= \underset{\sim}{\mu} + \mathbb{E}(L\underset{\sim}{F}) + \mathbb{E}(\underset{\sim}{\varepsilon})$
  $= \underset{\sim}{\mu} + L\mathbb{E}(\underset{\sim}{F}) + \mathbb{E}(\underset{\sim}{\varepsilon})$
  $= \underset{\sim}{\mu} + L\underset{\sim}{0} + \underset{\sim}{0}$
  $= \underset{\sim}{\mu}$
- É isto mesmo, o valor esperado das notas é o vetor $\underset{\sim}{\mu}$, que representa a média da população de interesse.
- $\underset{15x15}{Cov(\underset{\sim}{X})} = \Sigma = \mathbb{E}((\underset{\sim}{X} - \underset{\sim}{\mu})(\underset{\sim}{X} - \underset{\sim}{\mu})') = \mathbb{E}((L\underset{\sim}{F} + \underset{\sim}{\varepsilon})(L\underset{\sim}{F} + \underset{\sim}{\varepsilon})')$
  $= \mathbb{E}((L\underset{\sim}{F})(L\underset{\sim}{F})' + \underset{\sim}{\varepsilon}(L\underset{\sim}{F})' + (L\underset{\sim}{F})\underset{\sim}{\varepsilon}' + \underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}')$
  $= \mathbb{E}(L\underset{\sim}{F}\underset{\sim}{F}'L') + \mathbb{E}(\underset{\sim}{\varepsilon}(L\underset{\sim}{F})') + \mathbb{E}((L\underset{\sim}{F})\underset{\sim}{\varepsilon}') + \mathbb{E}(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}')$
  $= L\mathbb{E}(\underset{\sim}{F}\underset{\sim}{F}')L' + \underset{\sim}{0} + \underset{\sim}{0} + \psi$
  $= L\mathbb{I}_2 L' + \psi$
  $= LL' + \psi$
  Isto é, $Cov(\underset{\sim}{X}) = \Sigma = LL' + \psi$

Suposições do modelo fatorial

- No nosso exemplo com 15 notas e dois fatores:

$$\underset{15x15}{\Sigma} = \begin{bmatrix} l_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \\ . & . \\ . & . \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} \begin{bmatrix} l_{11} & \ell_{21} & . & . & \ell_{15,1} \\ \ell_{12} & \ell_{22} & . & . & \ell_{15,2} \end{bmatrix} + \begin{bmatrix} \psi_1 & & & \\ & . & & \\ & & . & \\ & & & . & \\ & & & & \psi_{15} \end{bmatrix} =$$

$$\begin{bmatrix} l_{11}^2 + \ell_{12}^2 + \psi_1 & \ell_{11}\ell_{12} + l_{12}l_{22} & \ell_{11}\ell_{31} + l_{12}l_{32} & . & . & \ell_{11}\ell_{15,1} + \ell_{12}\ell_{15,2} \\ \ell_{11}\ell_{21} + \ell_{12}\ell_{22} & \ell_{21}^2 + \ell_{22}^2 + \psi_2 & \ell_{21}\ell_{31} + \ell_{22}\ell_{32} & . & . & \ell_{21}\ell_{15,1} + \ell_{22}\ell_{15,2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & \ell_{15,1}^2 + \ell_{15,2}^2 + \psi_{15} \end{bmatrix}$$

Suposições do modelo fatorial

- Outra maneira de pensar sobre Σ é perceber que, se olharmos a matriz $\underset{15x2}{L}$ como um conjunto de 15 vetores-cargas,

$$L = \begin{bmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \\ . & . \\ . & . \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\ell}_1 \\ \underset{\sim}{\ell}_2 \\ . \\ . \\ \underset{\sim}{\ell}_{15} \end{bmatrix}$$

onde $\underset{\sim}{\ell}_j$ = cargas da disciplina j.

Suposições do modelo fatorial

- Então,

$$
\underset{15x15}{\Sigma} =
$$

$$
\begin{bmatrix}
\langle \underset{\sim}{\ell}_1, \underset{\sim}{\ell}_1 \rangle + \psi_1 & \langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_1 \rangle & \langle \underset{\sim}{\ell}_3, \underset{\sim}{\ell}_1 \rangle & \cdot & \cdot & \langle \underset{\sim}{\ell}_{15}, \underset{\sim}{\ell}_1 \rangle \\
\langle \underset{\sim}{\ell}_1, \underset{\sim}{\ell}_2 \rangle & \langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_2 \rangle + \psi_2 & \langle \underset{\sim}{\ell}_3, \underset{\sim}{\ell}_2 \rangle & \cdot & \cdot & \langle \underset{\sim}{\ell}_{15} \underset{\sim}{\ell}_2 \rangle \\
\langle \underset{\sim}{\ell}_1, \underset{\sim}{\ell}_3 \rangle & \langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_3 \rangle & \langle \underset{\sim}{\ell}_3, \underset{\sim}{\ell}_3 \rangle + \psi_3 & \cdot & \cdot & \langle \underset{\sim}{\ell}_{15} \underset{\sim}{\ell}_3 \rangle \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\langle \underset{\sim}{\ell}_1, \underset{\sim}{\ell}_{15} \rangle & \langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_{15} \rangle & \langle \underset{\sim}{\ell}_3, \underset{\sim}{\ell}_{15} \rangle & \cdot & \cdot & \langle \underset{\sim}{\ell}_{15}, \underset{\sim}{\ell}_{15} \rangle + \psi_{15}
\end{bmatrix}
$$

$$
= \begin{bmatrix}
\|\underset{\sim}{\ell}_1\|^2 + \psi_1 & \langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_1 \rangle & \cdot & \cdot & \langle \underset{\sim}{\ell}_{15}, \underset{\sim}{\ell}_1 \rangle \\
\langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_1 \rangle & \|\underset{\sim}{\ell}_2\|^2 + \psi_2 & \cdot & \cdot & \langle \underset{\sim}{\ell}_{15} \underset{\sim}{\ell}_2 \rangle \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\langle \underset{\sim}{\ell}_{15}, \underset{\sim}{\ell}_1 \rangle & \langle \underset{\sim}{\ell}_{15}, \underset{\sim}{\ell}_2 \rangle & \cdot & \cdot & \|\underset{\sim}{\ell}_{15}\|^2 + \psi_{15}
\end{bmatrix}
$$

Suposições do modelo fatorial

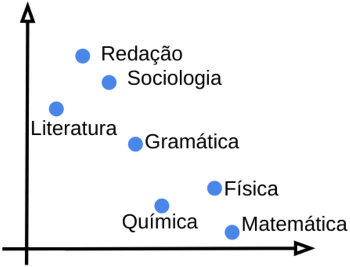
- Assim,
$$Var(X_i) = \Sigma_{ii} = \|\underset{\sim}{\ell}_i\|^2 + \psi_i = \ell_{i1}^2 + \ell_{i2}^2 + \psi_i$$
$\|\underset{\sim}{\ell}_i\|^2 \Rightarrow$ comumalidade
$\psi_i \Rightarrow$ variância específica
- Se os dois fatores latentes não possuem impacto na disciplina i (por exemplo, se a disciplina for educação física), então $\ell_{i1}^2 + \ell_{i2}^2 \approx 0$ e toda a variância da nota é devida aos fatores específicos ε_i e diferentes dos fatores latentes.
- Suponhamos que a disciplina X_i tenha uma carga grande do fator verbal ($\ell_{i1}^2 \gg 0$), mas uma carga pequena do fator quantitativo ($l_{i2}^2 \approx 0$). Então $Var(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + \psi_i \approx \ell_{i1}^2 + \psi_i$.
- Toda a variabilidade das notas entre os alunos é devida às diferenças do fator verbal.
- Alunos apenas com o fator quantitativo FQ muito diferentes não terão notas muito distintas.

Interpretando as cargas dos fatores

$$L = \begin{bmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \\ . & . \\ . & . \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\ell}_1 \\ \underset{\sim}{\ell}_2 \\ . \\ . \\ \underset{\sim}{\ell}_{15} \end{bmatrix}$$

- Podemos plotar as linhas de L num gráfico planar.
- A primeira coordenada (fator 1) no eixo horizontal e a segunda coordenada no eixo vertical (fator 2).

Interpretando as cargas dos fatores



- Esta representação mostra que as disciplinas Física, Química, Matemática ⇒ possuem cargas altas no Fator 1 e cargas baixas no Fator 2.
  Redação, Literatura, Sociologia ⇒ pouca carga do Fator 1 e muita cargado Fator 2.

Interpretando as cargas dos fatores

- Isto implica que, nesta disposição de colunas da matriz L, a primeira coluna (ou a primeira coordenada das linhas) representa o fator quantitativo.
- A segunda coluna de L representa o fator verbal.
- Observe que "Gramática" ficou a meio caminho, com carga mediana nos dois fatores. Para ter nota alta em "Gramática" é preciso ter "doses" razoáveis dos dois fatores OU uma "dose" bem grande de um dos fatores, qualquer um deles.

Métodos de estimação

1. **Máxima verossimilhança** (veremos mais tarde no curso)
2. **Componentes principais**

- Veremos apenas o segundo método.
- Pelo teorema espectral, $\Sigma = P \wedge P'$, onde $P = \begin{bmatrix} \underset{\sim}{v}_1 & \ldots & \underset{\sim}{v}_p \end{bmatrix}$ são os autovetores de Σ e $\wedge = \begin{pmatrix} \lambda_1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \lambda_p \end{pmatrix}$ é matriz diagonal com os autovalores.
- Manipulação matricial permite escrever

$$\Sigma = L^*(L^*)' = \begin{bmatrix} \sqrt{\lambda_1}\underset{\sim}{\vec{v}}_1 & \ldots & \sqrt{\lambda_p}\underset{\sim}{\vec{v}}_p \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\underset{\sim}{v}_1' \\ \cdot \\ \cdot \\ \sqrt{\lambda_p}\underset{\sim}{v}_p' \end{bmatrix}$$

Métodos de estimação

- Suponha que os últimos autovalores sejam ≈ 0.
- Isto implica que as últimas colunas de L^* são aproximadamente nulas e podem ser ignoradas.
- Mais formalmente, suponha que a soma dos k primeiros autovalores seja praticamente igual à soma de todos os p autovalores:

$$\frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_k + \lambda_{k+1} + ... + \lambda_p} \approx 1$$

- Ignorando as últimas colunas da matriz $\underset{pxp}{L^*}$ ficamos com uma matriz $\underset{pxk}{L}$:

$$\Sigma \approx LL' = \left[\sqrt{\lambda_1}\underset{\sim}{v}_1 \quad ... \quad \sqrt{\lambda_k}\underset{\sim}{v}_k\right] \begin{bmatrix} \sqrt{\lambda_1}\underset{\sim}{v}_1' \\ \cdot \\ \cdot \\ \sqrt{\lambda_k}\underset{\sim}{v}_k' \end{bmatrix}$$

Métodos de estimação

- Para completar o modelo fatorial, estimamos a matriz diagonal

$$\psi = \begin{pmatrix} \psi_1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \psi_p \end{pmatrix} = diag[\Sigma - LL']$$

Isto é, $\psi_i = \Sigma_{ii} - (LL')_{ii}$

Resumo prático

- Matriz de dados $\underset{nxp}{X}$
- Obtenha $S = Cov(X)$
- Alternativamente, se os s_{ii} forem muito distintos, usamos a matriz $R = cor(x)$, a matriz de correlação.
- Obtermos os autovalores ordenados e os autovetores de S.
- Calcule a soma acumulada $\frac{\lambda_1+\ldots+\lambda_k}{\lambda_1+\ldots+\lambda_k+\lambda_{k+1}+\ldots+\lambda_p} = a_k$
- Se $a_k \approx 1$ com k pequeno, então o modelo fatorial pode ser usado pois vai simplificar a estrutura dos dados.

Resumo prático

- Use os primeiro k autovetores (tal que $a_k \approx 1$) para criar a matriz de cargas

$$L = \left[\sqrt{\lambda_1}\underset{\sim}{v}_1 \quad \dots \quad \sqrt{\lambda_k}\underset{\sim}{v}_k\right] \text{ e } \psi = \begin{bmatrix} \psi_1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \psi_p \end{bmatrix} \text{ onde}$$

$$\psi_i = S_{ii} - (LL')_{ii}$$

- Um bom critério de escolha de k é verificar que a soma das entradas ao quadrado da matriz $(S - (LL' + \psi)) \leq \lambda_{k+1}^2 + \dots + \lambda_p^2$
  Assim, se $\lambda_{k+1}^2 + \dots + \lambda_p^2 \approx 0 \Rightarrow S \approx LL' + \psi$ e o modelo fatorial é um bom ajuste.

Exemplos de Johnson & Wichern

Example 9.3 (Factor analysis of consumer-preference data) In a consumer-preference study, a random sample of customers were asked to rate several attributes of a new product. The responses, on a 7-point semantic differential scale, were tabulated and the attribute correlation matrix constructed. The correlation matrix is presented next:

$$
\begin{array}{lc}
\textit{Attribute (Variable)} & \\
\text{Taste} & 1 \\
\text{Good buy for money} & 2 \\
\text{Flavor} & 3 \\
\text{Suitable for snack} & 4 \\
\text{Provides lots of energy} & 5
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \;\;1\;\; & \;\;2\;\; & \;\;3\;\; & \;\;4\;\; & \;\;5\;\; \end{array} \\
\begin{bmatrix}
1.00 & .02 & .96 & .42 & .01 \\
.02 & 1.00 & .13 & .71 & .85 \\
.96 & .13 & 1.00 & .50 & .11 \\
.42 & .71 & .50 & 1.00 & .79 \\
.01 & .85 & .11 & .79 & 1.00
\end{bmatrix}
\end{array}
$$

It is clear from the circled entries in the correlation matrix that variables 1 and 3 and variables 2 and 5 form groups. Variable 4 is "closer" to the (2, 5) group than the (1, 3) group. Given these results and the small number of variables, we might expect that the apparent linear relationships between the variables can be explained in terms of, at most, two or three common factors.

The first two eigenvalues, $\hat{\lambda}_1 = 2.85$ and $\hat{\lambda}_2 = 1.81$, of $\mathbf{R}$ are the only eigenvalues greater than unity. Moreover, $m = 2$ common factors will account for a cumulative proportion

$$\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p} = \frac{2.85 + 1.81}{5} = .93$$

of the *total* (standardized) sample variance. The estimated factor loadings, communalities, and specific variances, obtained using (9-15), (9-16), and (9-17), are given in Table 9.1.

Exemplos de Johnson & Wichern

Table 9.1

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\,\hat{e}_{ij}$ F_1	F_2	Communalities $\tilde{h}_i^2$	Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$
1. Taste	.56	.82	.98	.02
2. Good buy for money	.78	−.53	.88	.12
3. Flavor	.65	.75	.98	.02
4. Suitable for snack	.94	−.10	.89	.11
5. Provides lots of energy	.80	−.54	.93	.07
Eigenvalues	2.85	1.81		
Cumulative proportion of total (standardized) sample variance	.571	.932		

Exemplos de Johnson & Wichern

Now,

$$\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}' + \widetilde{\mathbf{\Psi}} = \begin{bmatrix} .56 & .82 \\ .78 & -.53 \\ .65 & .75 \\ .94 & -.10 \\ .80 & -.54 \end{bmatrix} \begin{bmatrix} .56 & .78 & .65 & .94 & .80 \\ .82 & -.53 & .75 & -.10 & -.54 \end{bmatrix}$$

$$+ \begin{bmatrix} .02 & 0 & 0 & 0 & 0 \\ 0 & .12 & 0 & 0 & 0 \\ 0 & 0 & .02 & 0 & 0 \\ 0 & 0 & 0 & .11 & 0 \\ 0 & 0 & 0 & 0 & .07 \end{bmatrix} = \begin{bmatrix} 1.00 & .01 & .97 & .44 & .00 \\ & 1.00 & .11 & .79 & .91 \\ & & 1.00 & .53 & .11 \\ & & & 1.00 & .81 \\ & & & & 1.00 \end{bmatrix}$$

nearly reproduces the correlation matrix **R**. Thus, on a purely descriptive basis, we would judge a two-factor model with the factor loadings displayed in Table 9.1 as providing a good fit to the data. The communalities (.98, .88, .98, .89, .93) indicate that the two factors account for a large percentage of the sample variance of each variable.

We shall not interpret the factors at this point. As we noted in Section 9.2, the factors (and loadings) are unique up to an orthogonal rotation. A rotation of the factors often reveals a simple structure and aids interpretation.

Exemplos de Johnson & Wichern

Example 9.4 (Factor analysis of stock-price data) Stock-price data consisting of $n = 103$ weekly rates of return on $p = 5$ stocks were introduced in Example 8.5. In that example, the first two sample principal components were obtained from $\mathbf{R}$. Taking $m = 1$ and $m = 2$, we can easily obtain principal component solutions to the orthogonal factor model. Specifically, the estimated factor loadings are the sample principal component coefficients (eigenvectors of $\mathbf{R}$), scaled by the square root of the corresponding eigenvalues. The estimated factor loadings, communalities, specific variances, and proportion of total (standardized) sample variance explained by each factor for the $m = 1$ and $m = 2$ factor solutions are available in Table 9.2. The communalities are given by (9-17). So, for example, with $m = 2$, $\tilde{h}_1^2 = \tilde{\ell}_{11}^2 + \tilde{\ell}_{12}^2 = (.732)^2 + (-.437)^2 = .73$.

Table 9.2

Variable	One-factor solution: Estimated factor loadings F_1	One-factor solution: Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$	Two-factor solution: Estimated factor loadings F_1	Two-factor solution: Estimated factor loadings F_2	Two-factor solution: Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$
1. J P Morgan	.732	.46	.732	−.437	.27
2. Citibank	.831	.31	.831	−.280	.23
3. Wells Fargo	.726	.47	.726	−.374	.33
4. Royal Dutch Shell	.605	.63	.605	.694	.15
5. ExxonMobil	.563	.68	.563	.719	.17
Cumulative proportion of total (standardized) sample variance explained	.487		.487	.769	

Exemplos de Johnson & Wichern

The residual matrix corresponding to the solution for $m = 2$ factors is

$$\mathbf{R} - \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}' - \widetilde{\boldsymbol{\Psi}} = \begin{bmatrix} 0 & -.099 & -.185 & -.025 & .056 \\ -.099 & 0 & -.134 & .014 & -.054 \\ -.185 & -.134 & 0 & .003 & .006 \\ -.025 & .014 & .003 & 0 & -.156 \\ .056 & -.054 & .006 & -.156 & 0 \end{bmatrix}$$

The proportion of the total variance explained by the two-factor solution is appreciably larger than that for the one-factor solution. However, for $m = 2$, $\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}'$ produces numbers that are, in general, larger than the sample correlations. This is particularly true for r_{13}.

It seems fairly clear that the first factor, F_1, represents general economic conditions and might be called a *market factor*. All of the stocks load highly on this factor, and the loadings are about equal. The second factor contrasts the banking stocks with the oil stocks. (The banks have relatively large negative loadings, and the oils have large positive loadings, on the factor.) Thus, F_2 seems to differentiate stocks in different industries and might be called an *industry factor*. To summarize, rates of return appear to be determined by general market conditions and activities that are unique to the different industries, as well as a residual or firm specific factor.

Métodos de estimação

- Existe um problema de identificabilidade na determinação do modelo fatorial.
- O problema é que a matriz de cargas L só pode ser conhecida a menos de uma rotação.
- Seja $\underset{2x2}{T}$ uma matriz ortogonal.

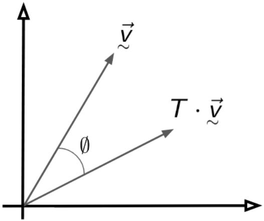
  Isto é, $TT' = T'T = \mathbb{I}_2 =$ identidade $= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- De álgebra de matrizes, sabemos que matrizes ortogonais correspondem a uma rotação ríida dos eixos coordenados.

Métodos de estimação

- Isto significa que uma matriz $\underset{2x2}{T}$ tal que $TT' = T'T = \mathbb{I}$ tem de ser da seguinte forma:
$$T = \begin{bmatrix} cos(\phi) & sin(\phi) \\ -sin(\phi) & cos(\phi) \end{bmatrix} \Rightarrow \text{rotação clockwise}$$
ou $$T = \begin{bmatrix} cos(\phi) & -sin(\phi) \\ sin(\phi) & cos(\phi) \end{bmatrix} \Rightarrow \text{counter-clockwise}$$
para $\phi \in [0, 2\pi]$
- Estas matrizes correspondem a rotações no plano.

Métodos de estimação

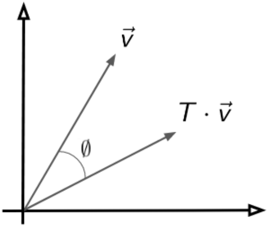
- Seja $\underset{2x1}{\underset{\sim}{\vec{V}}} = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2$ e $\underset{2x2}{T} = \begin{bmatrix} cos(\phi) & sin(\phi) \\ -sin(\phi) & cos(\phi) \end{bmatrix}$
- Então $\underset{2x2}{T}\underset{2x1}{\underset{\sim}{\vec{V}}}$ é um novo ponto no $\mathbb{R}^2$ obtido rotacionando $\underset{\sim}{\vec{V}}$ pelo ângulo ϕ na direção do relógio:



Métodos de estimação

- Considerando vetores-linha:

$$(T\vec{v})' = \vec{v}'T' = (x, y)\begin{bmatrix} cos(\phi) & -sin(\phi) \\ sin(\phi) & cos(\phi) \end{bmatrix}$$



A situação geométrica continua a mesma de antes, representar o vetor como linha ou coluna não altera o resultado.

- Vamos trabalhar com as linhas da matriz L.

Métodos de estimação

- Considere a matriz L das cargas

$$L = \begin{bmatrix} \underset{\sim}{\ell}'_1 \\ \cdot \\ \cdot \\ \underset{\sim}{\ell}'_{15} \end{bmatrix} = \begin{bmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \\ \cdot & \cdot \\ \cdot & \cdot \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} \text{ e } T = \text{matriz ortogonal.}$$

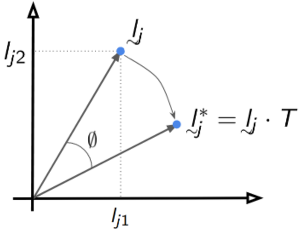
O produto $\underset{15x2}{\underbrace{\underset{15x2}{L} \cdot \underset{2x2}{T}}}$ pode ser pensado linha a linha

$$\underset{15x2}{L} \cdot \underset{2x2}{T} = \begin{bmatrix} \underset{\sim}{\ell}'_1 \\ \underset{\sim}{\ell}'_2 \\ \cdot \\ \cdot \\ \underset{\sim}{\ell}'_{15} \end{bmatrix} = \begin{bmatrix} \underset{1x2}{\underset{\sim}{\ell}'_1} \cdot \underset{2x2}{T} \\ \underset{\sim}{\ell}'_2 \cdot T \\ \cdot \\ \cdot \\ \underset{\sim}{\ell}'_{15} \cdot T \end{bmatrix} = L^*$$

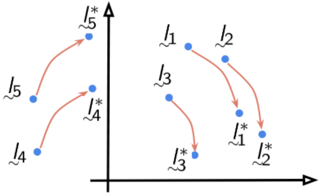
As linhas de L^* são as linhas de L rotacionadas de certo ângulo $\emptyset$ associado à matriz T.

Métodos de estimação

- Rodando uma linha de L



- Rodando cinco linhas de L



Métodos de estimação

- OK, o que tudo isto quer dizer?
- Suponha que o modelo fatorial é correto e que realmente podemos escrever ou decompor a matriz de covariância de $\underset{\sim}{X}$ como:

$$Var(\underset{\sim}{X}) = \underset{15x5}{\Sigma} = \underset{15x2}{L} \cdot \underset{2x15}{L'} + \underset{15x15}{\psi} \Rightarrow \text{diagonal}$$

- Seja $\underset{2x2}{T}$ $\underline{\underline{\text{qualquer}}}$ matriz ortogonal (de rotação no plano).

  Então podemos escrever

$$\Sigma = LL' + \psi = L\underbrace{TT'}_{\mathbb{I}_2}L' + \psi = \underbrace{(LT)}_{\underset{15x2}{L^*}}(LT)' + \psi = \underset{15x2}{L^*}\underset{2x15}{(L^*)'} + \psi$$

Métodos de estimação

- Isto significa que, se tivermos apenas Σ, teremos

$$LL' + \psi = \Sigma = L^*(L^*)' + \psi$$

onde $L^* = LT$ é diferente de L.

- As linhas de L^* são as linhas de L rotacionadas de um ângulo $\emptyset$.
- Como T é arbitrária (pode ser qualquer T) isto significa que podemos rodar L à vontade, com qualquer ângulo $\emptyset$, que sempre teremos uma representação de Σ da forma $\Sigma = L^*(L^*)' + \psi$.

Métodos de estimação

- Mas como interpretar os números que aparecem em L?
- Qual a L "correta"?
- Não é possível determinar uma única L tal que $\Sigma = LL' + \psi$
- Existem infinitos L com esta propriedade.
- Qualquer $L^* = LT$ (isto é, L rotacionada) terá a mesma propriedade.
- Todas as matrizes de carga L^* obtidas a partir de uma matriz L inicial terão a mesma capacidade de reproduzir a matriz de covariância Σ.

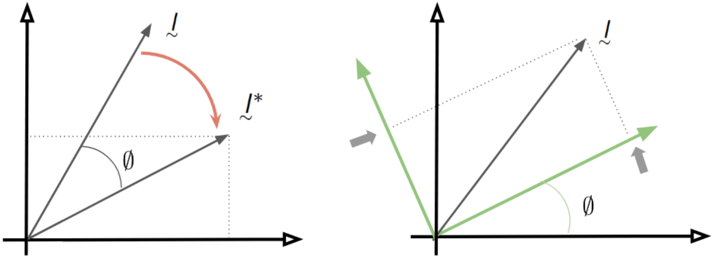
Métodos de estimação

- Ao invés disso se tornar um problema, transformamos os limões numa limonada.
- Caso uma matriz de cargas L inicialmente obtida por algum método de estimação não fornecer uma boa interpretação para os fatores, nós procuramos uma versão rotacionada $L^* = LT$ tal que as novas cargas sejam mais interpretáveis.
- É comum sermos capazes de terminar com uma estrutura mais simples que a matriz L inicial.
- Qual é esta estrutura mais simples?

Métodos de estimação

- <u>Idealmente</u>, nós gostaríamos de ver um padrão em que cada variável tenha uma carga alta num dos fatores e uma carga ≈ 0 nos demais.
- O objetivo é procurar uma rotação dos eixos de forma que as novas cargas fiquem o mais pròximo possível deste ideal.
- <u>OBS:</u> Se temos 15 pontos no plano (as cargas $\underset{\sim}{\ell_j}$) e <u>rodamos</u> todas elas de um ângulo $\emptyset$, isto é o mesmo que <u>rodar os dois eixos do plano</u> de $-\emptyset$ e deixar os "pontos intactos".

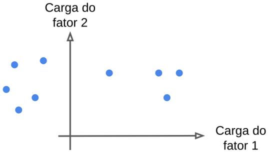
Métodos de estimação



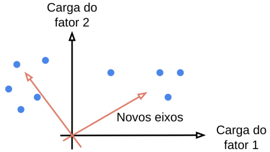
Coordenadas nos novos eixos são as mesmas de $\underset{\sim}{\ell}^*$nos eixos antigos.

Métodos de estimação

- Se nossas cargas $\underset{px2}{L} = \begin{bmatrix} . & . \\ . & . \\ . & . \end{bmatrix}$ são assim:



procuramos rodar os eixos até que as cargas sejam próximas do ideal



Nos novos eixos, as cargas são ≈ 0 exceto em um único fator.

Procedimento VARIMAX

- Defina $\tilde{\ell}^*_{ij} = \frac{\ell^*_{ij}}{h_i} = \frac{\ell^*_{ij}}{\ell^*_{i1}+\ell^*_{i2}+\ldots+\ell^*_{ip}} \in [0,1]$ as cargas dos fatores rotacionados e normalizados.
- Busque a rotação T tal que maximize

$$V = \frac{1}{p}\sum_{j=1}^{m}[\sum_{i=1}^{p}\tilde{\ell}^{*\,4}_{ij} - \frac{(\sum_{i=1}^{p}\tilde{\ell}^{*\,2}_{ij})^2}{p}]$$

$\propto \sum_{j=1}^{m}$(variância das (cargas)^2 normalizadas do fator j)

Maximizar V significa espalhar as (cargas)^2 o máximo possível, com valores altos em alguns fatores e valores ≈ 0 em outros.

- Tendo estimado a matriz de cargas, podemos estimar o valor dos fatores de cada indivíduo da amostra.

Procedimento VARIMAX

- Suponha que o i-ésimo indivíduo tenha o vetor $\underset{\sim}{X}_i$ e que tenhamos estimado $\underset{\sim}{\mu}$ (a média das variáveis sobre a amostra) e tenhamos também a matriz de cargas L (talvez rotacionada).
- O vetor $\underset{\sim}{F}_i$ <u>deste indivíduo</u> é estimado pela minimização da diferença entre $\underset{\sim}{X}_i$ e $\underset{\sim}{\mu} + L\underset{\sim}{F}$.

  Isto é, procuramos um vetor $\underset{\sim}{F}_i$ tal que ele minimize o comprimento

$$\|X_i - \mu - L\underset{\sim}{F}_i\|^2$$

- Veja a lista de exercícios (beer example) para um exemplo.