# MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

## Aula 14

23 de junho de 2021

# Sumário

# Preliminares

- Primeiramente observemos que para explicar as relações entre $p$ variáveis são necessárias $p$ componentes principais; por esse motivo, o modelo adotado não é o ideal.
- O fato de as componentes principais não serem correlacionadas e ordenadas com variâncias decrescentes e o fato de que a técnica corresponde a uma fatoração da matriz de covariâncias da variáveis originais fazem com que a aproximação obtida quando consideramos apenas as primeiras componentes principais seja razoável.
- No entanto, essa técnica pode introduzir um erro sistemático na reprodução das correlações originais, pois podem existir uma ou mais dessas variáveis que sejam muito correlacionadas com as componentes principais desprezadas do que com aquelas retidas na análise.
- Outra observação importante, é que essa técnica utiliza toda a informação sobre cada uma das variáveis originais, embora seja razoável imaginar que uma parcela de sua variabilidade seja específica, nada tendo a ver com as demais variáveis do conjunto sob investigação.
- Além disso, pode-se suspeitar que os "verdadeiros fatores" responsáveis pela geração das observações tenham todos a mesma importância ou que sejam correlacionados entre si.

# Preliminares

- Alguns desses problemas podem ser solucionados por meio da técnica de Análise Fatorial. A ideia que a fundamenta está baseada na partição da variância de cada variável do sistema multivariado numa variância comum e numa variância específica. Além disso, supõe-se que as correlações entre as $p$ variáveis são geradas por um número $m < p$ de variáveis latentes (ou fatores).
- A vantagem dessa técnica relativamente àquela de componentes principais está na habilidade de reprodução da estrutura de correlações originais por meio de um pequeno número de fatores sem os erros sistemáticos que podem ocorrer quando simplesmente desprezamos algumas componentes principais.
- As desvantagens da Análise Fatorial estão na maior dificuldade de cálculo dos escores fatoriais e na existência de múltiplas soluções. Na realidade a estrutura de correlações das variáveis originais pode ser igualmente reproduzida por qualquer outro conjunto de variáveis latentes de mesma dimensão. A não ser que se imponham restrições adicionais, infinitas soluções equivalentes sempre existirão.

# AF – metodologia

- Consideremos um vetor $\mathbf{X} = (X_1, \ldots, X_p)^\top$ com média $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ e matriz de covariâncias $\boldsymbol{\Sigma}$ com elementos $\sigma_{ij}$, $i, j = 1, \ldots, p$.
- O modelo utilizado para Análise Fatorial de dados provenientes da observação das $p$ variáveis agrupadas $X_1, \ldots, X_p$ é

$$X_i - \mu_i = \lambda_{i1} F_1 + \ldots + \lambda_{im} F_m + e_i = \sum_{j=1}^{m} \lambda_{ij} F_j + e_i, \; i = 1, \ldots, p \qquad (1)$$

  em que $F_j$ é o $j$-ésimo fator comum a todas as variáveis, $\lambda_{ij}$ é o parâmetro (chamado de carga fatorial) que indica a importância desse fator na composição da $i$-ésima variável e $e_j$ é um fator específico para essa variável.
- Os coeficientes dos fatores, $\lambda_{ij}$, identificam o peso relativo de cada variável no componente. Quanto maior for o valor absoluto do coeficiente, mais importante é a variável correspondente ao estimar o fator. Os escores fatoriais são os valores estimados dos fatores.

# AF – metodologia

- Como dito acima, as cargas fatoriais indicam o quanto um fator explica uma variável e variam de -1 a $+1$.
- Cargas próximas de -1 ou $+1$ indicam que o fator explica fortemente a variável. Cargas próximas de zero indicam que o fator tem pouca influência sobre a variável.
- Cargas fatoriais são difíceis de interpretar quando não houver rotação. Esta simplifica a estrutura das cargas e torna os fatores mais claramente distinguíveis e fáceis de interpretar.
- Há diversos métodos de rotação (ver abaixo) e devemos escolher aquele que proporciona melhor interpretação.

# AF – metodologia

- Em notação matricial, o modelo pode ser escrito como

$$\mathbf{X} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{f} + \mathbf{e} \tag{2}$$

em que $\boldsymbol{\Lambda}$ é a matriz com dimensão $p \times m$ de cargas fatoriais, $\mathbf{f} = (F_1, \ldots, F_m)^\top$ é o vetor cujos elementos são os fatores comuns e $\mathbf{e} = (e_1, \ldots, e_p)$ é um vetor cujos elementos são os fatores específicos.

- Adicionalmente, supomos que $E(\mathbf{f}) = \mathbf{0}$, $Cov(\mathbf{f}) = \mathbf{I}_m$, $E(\mathbf{e}) = \mathbf{0}$ e $Cov(\mathbf{e}) = \boldsymbol{\psi} = \text{diag}(\psi_1, \ldots, \psi_p)$ e que $Cov(\mathbf{f}, \mathbf{e}) = \mathbf{0}$. Os elementos de $\boldsymbol{\psi}$ são as variâncias específicas.

# AF – metodologia

- Para avaliar a relação entre a estrutura de covariâncias de $\mathbf{X}$ e os fatores, observemos que

$$\begin{aligned} Cov(X_i, X_k) &= Cov\left(\sum_{j=1}^{m} \lambda_{ij} F_j, \sum_{\ell=1}^{m} \lambda_{k\ell} F_\ell\right) \\ &= \sum_{j=1}^{m}\sum_{\ell=1}^{m} \lambda_{ij}\lambda_{k\ell} E(F_j F_\ell) + E(e_i e_j) \qquad (3) \\ &= \sum_{j=1}^{m}\sum_{\ell=1}^{m} \lambda_{ij}\lambda_{k\ell} + E(e_i e_j) \end{aligned}$$

- Consequentemente, $Cov(X_i, X_k) = \sigma_{ik} = \sum_{j=1}^{m} \lambda_{ij}\lambda_{kj}$ se $i \neq k$ e $Cov(X_i, X_i) = \sigma_{ii} = \sum_{j=1}^{m} \lambda_{ij}^2$. O termo $\sum_{j=1}^{m} \lambda_{ij}^2$ é conhecido por comunalidade da $i$-ésima variável.
- Em notação matricial, podemos escrever

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \boldsymbol{\psi}$$

e o objetivo é estimar os elementos de $\mathbf{\Lambda}$ e $\boldsymbol{\psi}$.

# AF – metodologia

- A comunalidade de cada variável é a proporção da variabilidade explicada pelos fatores.
- Quanto mais próxima de 1, melhor é a explicação da variável pelo fator. Decidimos acrescentar um fator se ele contribuir significativamente ao ajuste de algumas variáveis.
- A variância de um fator fornece a variabilidade nos dados explicada pelo fator. Se usarmos CP para extrair fatores, e não usarmos rotação, a variância de cada fator é igual ao seu autovalor.
- Rotação muda a distribuição da proporção da variabilidade explicada por cada fator. Mas a variação total explicada por todos os fatores se mantém.
- Quanto maior a variância de um fator, mais ele explica a variabilidade nos dados. A porcentagem da variância explixada por cada fator varia de zero a 1.

# AF – metodologia

- Em Análise de Componentes Principais, consideramos o modelo linear $\mathbf{Y} = \mathbf{BX}$ em que $\mathbf{Y}$ é o vetor cujos elementos são as componentes principais e $\mathbf{B} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top)^\top$ é a matriz cuja $i$-ésima linha contém os coeficientes da $i$-ésima componente principal.
- A matriz de covariâncias de $\mathbf{X}$ é fatorada como $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$. Em Análise Fatorial, consideramos o modelo linear $\mathbf{X} = \boldsymbol{\Lambda}\mathbf{f} + \mathbf{e}$ e a matriz de covariâncias de $\mathbf{X}$ é fatorada como $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\psi}$.
- Uma diferença entre os dois enfoques é que enquanto a fatoração de $\boldsymbol{\Sigma}$ é única em Análise de Componentes Principais, ela não o é em Análise Fatorial, pois se $\mathbf{T}$ for uma matriz ortogonal (*i.e.*, $\mathbf{T}\mathbf{T}^\top = \mathbf{I}_m$), obteremos

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\psi} = \boldsymbol{\Lambda}\mathbf{T}\mathbf{T}^\top\boldsymbol{\Lambda}^\top + \boldsymbol{\psi} = \boldsymbol{\Lambda}\mathbf{T}(\boldsymbol{\Lambda}\mathbf{T})^\top + \boldsymbol{\psi}$$

  e embora as cargas fatoriais $\boldsymbol{\Lambda}\mathbf{T}$ sejam diferentes das cargas fatoriais $\boldsymbol{\Lambda}$, a habilidade de reproduzir a matriz de covariâncias $\boldsymbol{\Sigma}$ não se altera. Escolhendo matrizes ortogonais diferentes, podemos determinar cargas fatoriais diferentes. A escolha de uma transformação conveniente será discutida posteriormente.

# AF – metodologia

Uma análise fatorial consiste dos seguintes passos:

a) Estimação dos parâmetros do modelo ($\lambda_{ij}$ e $\psi_i$) a partir de um conjunto de observações das variáveis $X_1, \ldots, X_p$.

b) Interpretação dos fatores determinados a partir das cargas fatoriais obtidas em a). Com esse objetivo considera-se a rotação dos fatores por meio de transformações ortogonais.

c) Estimação dos valores dos fatores comuns, chamados escores fatoriais para cada unidade amostral a partir dos valores das cargas fatoriais e das variáveis observadas.

# AF – metodologia

- Existem duas classes de métodos para estimação dos parâmetros do modelo fatorial. Na primeira classe consideramos o método de máxima verossimilhança e na segunda, métodos heurísticos como o método do fator principal ou o método do centroide.
- Para o método de máxima verossimilhança, supomos adicionalmente que as variáveis $X_1, \ldots, X_p$ seguem uma distribuição normal (multivariada) e que o número de fatores $m$ é conhecido. Os estimadores são obtidos por meio da solução do sistema de equações (ver Nota de Capítulo 3)

$$\begin{aligned} \mathbf{S}\psi^{-1}\mathbf{\Lambda} &= \mathbf{\Lambda}(\mathbf{I}_m + \mathbf{\Lambda}^\top \psi^{-1}\mathbf{\Lambda}) \\ \text{diag}(\mathbf{S}) &= \text{diag}(\mathbf{\Lambda}\mathbf{\Lambda}^\top + \psi) \end{aligned} \tag{4}$$

  que deve ser resolvido por meio de métodos iterativos. Detalhes podem ser encontrados em Morrison (1972).
- Uma das vantagens desse método é que as mudanças de escala das variáveis originais alteram os estimadores apenas por uma mudança de escala.

# AF – metodologia

- Se uma das variáveis $X_1, \ldots, X_p$ for multiplicada por uma constante, os estimadores das cargas fatoriais correspondentes ficam multiplicados pela mesma constante e o estimador da variância específica associada fica multiplicado pelo quadrado da constante. Dessa forma, podemos fazer os cálculos com as variáveis padronizadas (substituindo a matriz de covariâncias amostral **S** pela correspondente matriz de correlações amostrais **R** e posteriormente escrever os resultados em termos das unidades de medida originais).
- O método do fator principal está intimamente relacionado com a técnica utilizada na análise de componentes principais. Segundo esse método, os fatores são escolhidos obedecendo à ordem decrescente de sua contribuição à comunalidade total do sistema multivariado.
- Nesse contexto, o processo tem início com a determinação de um fator $F_1$ cuja contribuição à comunalidade total é a maior possível; em seguida, um segundo fator não correlacionado com $F_1$ e tal que maximize a comunalidade residual é obtido. O processo continua até que a comunalidade total tenha sido exaurida.

# AF – metodologia

Na prática, as comunalidades e as variâncias específicas devem ser estimadas com base nos dados amostrais. Embora existam vários métodos idealizados para essa finalidade, nenhum se mostra superior aos demais. Dentre os estimadores mais comuns para a comunalidade de uma variável $X_i$, destacamos:

i) o quadrado do **coeficiente de correlação múltipla** entre a variável $X_i$ e as demais;

ii) o maior valor absoluto dos elementos de $i$-ésima linha da matriz de correlações amostrais;

iii) estimadores obtidos de análises preliminares por meio de processos iterativos.

# AF – metodologia

Outro problema prático é a determinação do número de fatores a incluir na análise. Os critérios mais utilizados para esse fim são:

i) determinação do número de fatores por meio de algum conhecimento *a priori* sobre a estrutura dos dados;

ii) número de componentes principais correspondentes a autovalores da matriz $\mathbf{R}$ maiores que 1;

iii) explicação de certa proporção (escolhida arbitrariamente) da comunalidade ou da variância total.

# AF – metodologia

Um algoritmo comumente utilizado para a obtenção das cargas fatoriais e das variâncias específicas é

i) Obter as $p$ componentes principais com base na matriz de correlações amostrais $\mathbf{R}$.

ii) Escolher $m$ fatores segundo um dos critérios mencionados.

iii) Substituir os elementos da diagonal principal de $\mathbf{R}$ por estimadores das comunalidades correspondentes por meio de um dos métodos descritos acima, obtendo a chamada matriz de correlações reduzida, $\mathbf{R}^*$.

iv) Extrair $m$ fatores da matriz $\mathbf{R}^*$, obtendo novos estimadores das comunalidades que vão substituir aqueles obtidos anteriormente na diagonal principal.

v) Repetir o processo dos itens ii) - iv) até que a diferença entre dois conjuntos sucessivos de estimadores das comunalidades seja desprezável.

# AF – metodologia

- O método do centroide foi desenvolvido por Thurstone (1947) para simplificar os cálculos mas não é muito utilizado em virtude das recentes facilidades computacionais; os resultados obtidos por intermédio desse método não diferem muito daqueles obtidos pelo método do fator principal.
- Como a interpretação dos fatores numa análise fatorial é uma característica importante em aplicações práticas, pode-se utilizar a técnica de rotação dos fatores para obter resultados mais palatáveis.
- Consideremos um exemplo em que cinco variáveis $A, B, C, D$ e $E$ são representadas num espaço fatorial bidimensional conforme a representação da Figura 1.

# AF – metodologia

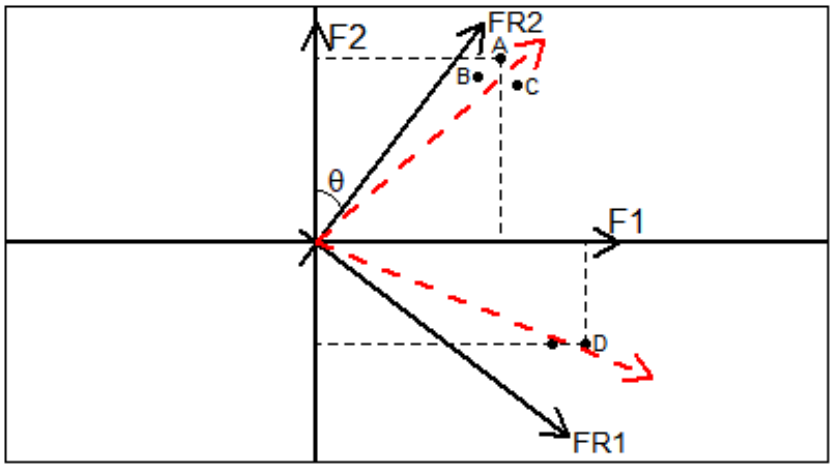


Figura: Representação de cinco variáveis num espaço vetorial bidimensional.

# AF – metodologia

Como ilustrado na Tabela 1, as cargas fatoriais relativas ao fator $F_1$ são altas e positivas para todas as variáveis. Por outro lado, apenas as variáveis $A, B$ e $C$ têm cargas positivas no fator $F_2$; as cargas das variáveis $D$ e $E$ são negativas nesse fator.

Tabela: Cargas fatoriais para as variáveis $A, B, C, D$ e $E$

| | Fatores iniciais | | Fatores rotacionados | |
|---|---|---|---|---|
| Variável | $F_1$ | $F_2$ | $FR_1$ | $FR_2$ |
| A | 0,75 | 0,63 | 0,14 | 0,95 |
| B | 0,69 | 0,57 | 0,14 | 0,90 |
| C | 0,80 | 0,49 | 0,18 | 0,92 |
| D | 0,85 | -0,42 | 0,94 | 0,09 |
| E | 0,76 | -0,42 | 0,92 | 0,07 |

# AF – metodologia

- Dois aglomerados de variáveis podem ser identificados na Figura 1: um formado pelas variáveis $A, B$ e $C$ e o outro pelas variáveis $D$ e $E$.
- Apesar disso, esses aglomerados não são evidentes nas cargas fatoriais da Tabela 1. Uma rotação do fatores (com os eixos rotulados $FR_1$ e $FR_2$) como aquela indicada na figura juntamente com as novas cargas fatoriais apresentadas na Tabela 1 ressaltam a separação entre os dois conjuntos de variáveis.
- Na solução inicial, cada variável é explicada por dois fatores enquanto que na solução obtida com a rotação dos fatores, apenas um deles é suficiente para explicar a correspondente estrutura de covariância.
- Em princípio, também podemos considerar rotações oblíquas, que são bem mais flexíveis, pois os fatores não precisam ser necessariamente ortogonais. Essa característica pode até ser considerada mais realista, pois a ortogonalidade não é determinante da relação entre os fatores. Os eixos realçados em vermelho na Figura 1 correspondem a uma dessas rotações oblíquas.

# AF – metodologia

O objetivo de qualquer rotação é obter fatores interpretáveis e com a estrutura mais simples possível. Nesse sentido, Thurstone (1947) sugere condições para se obter uma estrutura mais simples, nomeadamente:

i) Cada linha da matriz de cargas fatoriais $\Lambda$ deve conter pelo menos um valor nulo.

ii) Cada coluna da matriz de cargas fatoriais deveria ter pelo menos tantos valores nulos quantas forem as colunas.

iii) Para cada par de colunas deve haver algumas variáveis com cargas fatoriais pequenas numa delas e altas na outra.

iv) Para cada par de colunas uma grande porcentagem das variáveis deve ter cargas fatoriais não nulas em ambas.

v) Para cada par de colunas deve haver somente um pequeno número de variáveis com cargas fatoriais altas em ambas.

# AF – metodologia

Como consequência dessas sugestões,

i) Muitas variáveis (representadas como vetores no espaço dos fatores) devem ficar próximas dos eixos.

ii) Muitas variáveis devem ficar próximas da origem quando o número de fatores for grande.

iii) Somente um pequeno número de variáveis ficam longe dos eixos.

# AF – metodologia

A principal crítica às sugestões de Thurstone é que na prática poucas são as situações que admitem uma simplificação tão grande. O que se procura fazer é simplificar as linhas e colunas da matriz de cargas fatoriais e os métodos mais comumente empregados com essa finalidade são:

- Método Varimax em que se procura simplificar a complexidade fatorial, tentando-se obter fatores com poucos valores grandes e muitos valores nulos ou pequenos na respectiva coluna da matriz de cargas fatoriais. Após uma rotação Varimax, cada variável original tende a estar associada com poucos (preferencialmente, um) fatores e cada fator tende a se associar com poucas variáveis. Esse é o método mais utilizado na prática.
- Método Quartimax em que se procura maximizar o número de fatores necessários para explicar cada variável. Em geral, esse método produz um fator associado em que muitas variáveis têm cargas altas ou médias, o que nem sempre é conveniente para a interpretação.
- Método Equimax, uma mistura dos métodos Varimax e Quartimax.
- Método Promax, utilizado para rotações oblíquas.

# AF – estimação

- Um dos objetivos tanto da Análise de Componentes Principais quanto da Análise Fatorial, é substituir as $p$ variáveis originais $X_1, \ldots, X_p$ por um número menor, digamos, $m$ em análises subsequentes.
- No caso de componentes principais, podem-se utilizar as estimativas $\widehat{Y}_{ik} = \widehat{\boldsymbol{\beta}}_i \mathbf{x}_k, \ i = 1, \ldots, m$ para substituir os valores $\mathbf{x}_k$ observados para a $k$-ésima unidade amostral.
- Esse processo é mais complicado quando lidamos com a obtenção dos valores dos fatores $F_1, \ldots, F_m$ (denominados escores fatoriais) em Análise Fatorial, que não podem ser estimados no sentido estatístico usual, pois os fatores não são observáveis.
- Com esse objetivo, o método de Bartlett (1937) consiste em considerar (2) como um modelo de regressão heterocedático em que se supõe que as matrizes de cargas fatoriais, $\boldsymbol{\Lambda}$ e de variâncias específicas $\boldsymbol{\psi}$, são conhecidas e se considera o termo $\mathbf{e}$ como um vetor de erros.

# AF – estimação

Minimizando

$$Q(\mathbf{f}) = \mathbf{e}^\top \psi^{-1} \mathbf{e} = (\mathbf{x} - \boldsymbol{\mu} - \boldsymbol{\Lambda}\mathbf{f})^\top \psi^{-1} (\mathbf{x} - \boldsymbol{\mu} - \boldsymbol{\Lambda}\mathbf{f})$$

obtemos

$$\widehat{\mathbf{f}} = [\boldsymbol{\Lambda}^\top \psi^{-1} \boldsymbol{\Lambda}]^{-1} \boldsymbol{\Lambda}^\top \psi^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

e substituindo $\boldsymbol{\Lambda}$, $\psi$ e $\boldsymbol{\mu}$, respectivamente, por estimativas $\widehat{\boldsymbol{\Lambda}}$, $\widehat{\psi}$ e $\overline{\mathbf{x}}$, podemos construir os escores fatoriais para a $k$-ésima unidade amostral como

$$\widehat{\mathbf{f}}_k = [\widehat{\boldsymbol{\Lambda}}^\top \widehat{\psi}^{-1} \widehat{\boldsymbol{\Lambda}}]^{-1} \widehat{\boldsymbol{\Lambda}}^\top \widehat{\psi}^{-1} (\mathbf{x}_k - \overline{\mathbf{x}}).$$

# AF – estimação

- Alternativamente, no método de regressão, supõe-se que os fatores comuns, $\mathbf{f}$ e específicos $\mathbf{e}$ são independentes e têm distribuições normais multivariadas com dimensões $m$ e $p$, respectivamente, de forma que o par $(\mathbf{X} - \boldsymbol{\mu}, \mathbf{f})$ também tem uma distribuição normal multivariada de dimensão $p + m$ com matriz de covariâncias

$$\begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\psi} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}^\top & \mathbf{I}_m \end{bmatrix}.$$

- Utilizando propriedades da distribuição normal multivariada, segue que a distribuição condicional de $\mathbf{f}$ dado $\mathbf{X} - \boldsymbol{\mu}$ também é normal multivariada com vetor de médias

$$E(\mathbf{f}|\mathbf{X} - \boldsymbol{\mu}) = \boldsymbol{\Lambda}^\top[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\psi}]^{-1}(\mathbf{X} - \boldsymbol{\mu})$$

e matriz de covariâncias

$$Cov(\mathbf{f}|\mathbf{X} - \boldsymbol{\mu}) = \mathbf{I}_m - \boldsymbol{\Lambda}^\top[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\psi}]^{-1}\boldsymbol{\Lambda}.$$

# AF – Número de fatores

- O termo $\boldsymbol{\Lambda}^\top[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\psi}]^{-1}$ corresponde aos coeficientes de uma regressão multivariada tendo os fatores como variáveis respostas e $\mathbf{X} - \boldsymbol{\mu}$ como variáveis explicativas. Utilizando as estimativas $\widehat{\boldsymbol{\Lambda}}$, $\widehat{\boldsymbol{\psi}}$, podemos calcular os escores fatoriais para a $k$-ésima unidade amostral (com valores das variáveis originais $\mathbf{x}_k$) por meio de

$$\widehat{\mathbf{f}}_k = \widehat{\boldsymbol{\Lambda}}^\top[\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Lambda}}^\top + \widehat{\boldsymbol{\psi}}]^{-1}(\mathbf{x}_k - \bar{\mathbf{x}}).$$

  Pacotes do R que podem ser utilizados para a AF são o psych, o GPArotation e o robustfa.
- Para determinar o número de fatores a considerar, podemos usar vários critérios:
  [1]Teste Scree: devido a Cattel (1966), consiste do gráfico dos autovalores contra o número de fatores (ou CPs, de uma ACP). Procura-se o ponto (elbow) onde a inclinação muda drasticamente.

  [2] Regra de Kaiser–Guttman: devida a Guttman(1954) e Kaiser (1960), é similar ao teste scree, mas consideram-se componentes ou fatores com autovalores maiores do que 1. Esse é o gráfico que temos usado.

  [3] Análise paralela: devida a Horn (1965), propõe reter somente autovalores que sejam superiores ou iguais à média dos autovalores obtidos de $k$ matrizes de correlações calculadas com $n$ observações aleatórias.

# AF – Número de fatores

A estratégia da Análise Paralela é:

(a) Gere $n$ v.a.s de acordo com uma $N(0,1)$ independentemente para $p$ variáveis;

(b) calcule a matriz de correlações de Pearson;

(c) calcule os autovalores dessa matriz;

(d) repita passos (a)-(c) $k$ vezes;

(e) calcule uma medida de localização, ML, para os $k$ vetores de autovalores: média, mediana, $p$-quantil etc;

(f) Substitua o valor 1 da regra de Guttman–Kaiser, ou seja, conte o número de autovalores maiores que ML.

Há, também, soluções não gráficas a esses testes.

- Ótima coordenada:

$$n_{oc} = \#\{(\lambda_i \geq 1) \text{ e } (\lambda_i \geq \lambda \text{ previsto pelo teste scree K-G}\}$$

ou

$$n_{oc} = \#\{(\lambda_i \geq ML \text{ e } (\lambda_i \geq \lambda \text{ previsto pelo teste scree da AP}\}$$

- Fator de aceleração: coloca ênfase na coordenada onde a inclinação da curva muda abruptamente

# AF – Exemplo 1

- **Exemplo 1.** Num estudo planejado para avaliar o nível de poluição atmosférica por meio de medidas de elementos químicos depositados em cascas de árvores, obtiveram-se observações da concentração de Al, Ba, Cu, Fe, Zn, P, Cl, Sr e Ca entre outros elementos em 193 unidades da espécie *Tipuana tipu* na cidade de São Paulo.
- Esses dados constituem um subconjunto daqueles disponíveis em `http://www.ime.usp.br/~jmsinger/MorettinSinger/arvores.xls` O objetivo aqui é obter um conjunto de fatores que permitam identificar características comuns a essas variáveis. Os resultados provenientes de uma análise de componentes principais estão dispostos na Tabela 2.

# AF – Exemplo 1

Tabela: Coeficientes de componentes principais (CP) para os dados do Exemplo 1

| | CP1 | CP2 | CP3 | CP4 | CP5 | CP6 | CP7 | CP8 | CP9 |
|---|---|---|---|---|---|---|---|---|---|
| Al | 0.90 | -0.11 | 0.09 | 0.21 | -0.16 | 0.19 | -0.08 | -0.17 | 0.17 |
| Ba | 0.88 | -0.16 | 0.09 | 0.10 | -0.10 | 0.27 | 0.09 | 0.31 | -0.01 |
| Cu | 0.82 | 0.18 | -0.05 | -0.23 | 0.31 | -0.18 | -0.31 | 0.08 | 0.01 |
| Fe | 0.95 | -0.10 | 0.07 | 0.10 | -0.03 | 0.10 | 0.00 | -0.18 | -0.19 |
| Zn | 0.83 | 0.16 | -0.13 | -0.22 | 0.29 | -0.21 | 0.31 | -0.05 | 0.05 |
| P | 0.25 | 0.69 | -0.25 | 0.53 | -0.20 | -0.28 | 0.00 | 0.05 | -0.01 |
| Cl | 0.17 | 0.53 | 0.60 | -0.42 | -0.39 | -0.07 | 0.01 | 0.00 | 0.00 |
| Mg | -0.24 | 0.22 | 0.78 | 0.35 | 0.40 | 0.05 | 0.02 | 0.00 | 0.00 |
| Ca | -0.20 | 0.77 | -0.33 | -0.12 | 0.15 | 0.47 | 0.00 | -0.04 | 0.00 |
| % Var | 0.45 | 0.17 | 0.13 | 0.08 | 0.07 | 0.06 | 0.02 | 0.02 | 0.01 |
| % Acum | 0.45 | 0.61 | 0.75 | 0.83 | 0.89 | 0.95 | 0.97 | 0.99 | 1.00 |

# AF – Exemplo 1

Uma análise da porcentagem da variância explicada pelas componentes principais juntamente com um exame do gráfico da escarpa sedimentar (scree plot) correspondente, apresentado na Figura 2 sugere que três fatores, que explicam 75% da variância total do sistema de variáveis originais poderiam contemplar uma representação adequada.

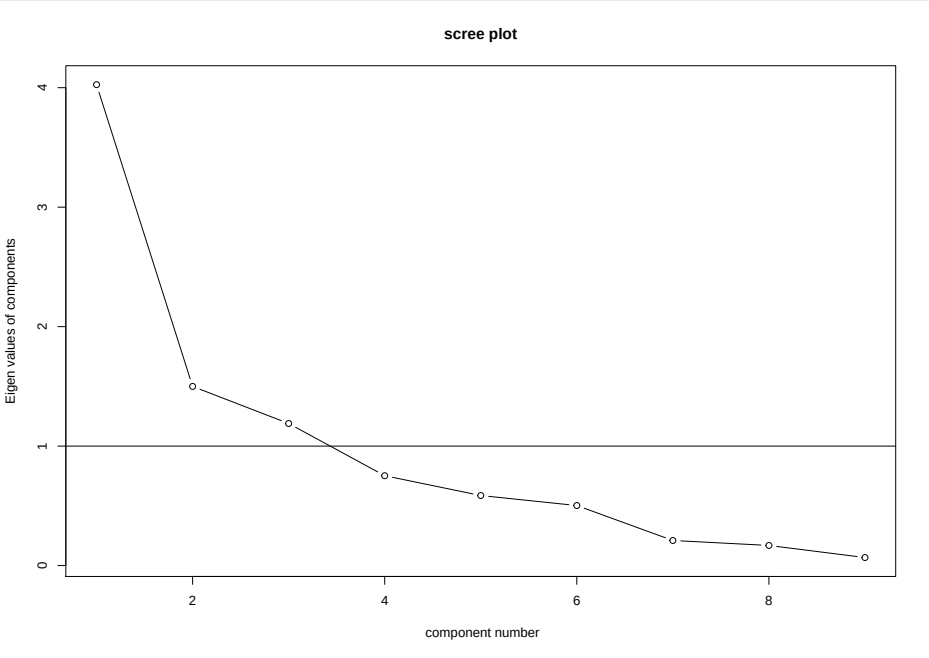


Figura: Gráfico da escarpa sedimentar para os dados do Exemplo 1.

# AF – Exemplo 1

As cargas fatoriais correspondentes três fatores rotacionados obliquamente juntamente com as comunalidades e especificidades correspondentes estão dispostos na Tabela 3.

Tabela: Cargas fatoriais, comunalidades e especificidades correspondentes a uma análise fatorial para os dados do Exemplo 1

| | Fator 1 | Fator 2 | Fator 3 | Comunalidade | Especificidade |
|---|---|---|---|---|---|
| Al | 0.91 | -0.11 | 0.03 | 0.82 | 0.18 |
| Ba | 0.86 | -0.13 | 0.00 | 0.75 | 0.25 |
| Cu | 0.75 | 0.27 | -0.04 | 0.66 | 0.34 |
| Fe | 0.97 | -0.08 | 0.01 | 0.95 | 0.05 |
| Zn | 0.74 | 0.27 | -0.13 | 0.68 | 0.32 |
| P | 0.20 | 0.47 | 0.05 | 0.25 | 0.75 |
| Cl | 0.22 | 0.27 | 0.39 | 0.22 | 0.78 |
| Mg | -0.04 | 0.02 | 0.73 | 0.54 | 0.46 |
| Ca | -0.21 | 0.67 | 0.01 | 0.49 | 0.51 |

# AF – Exemplo 1

Os gráficos das Figuras 3 e 4 também podem ser utilizados para identificar os fatores.

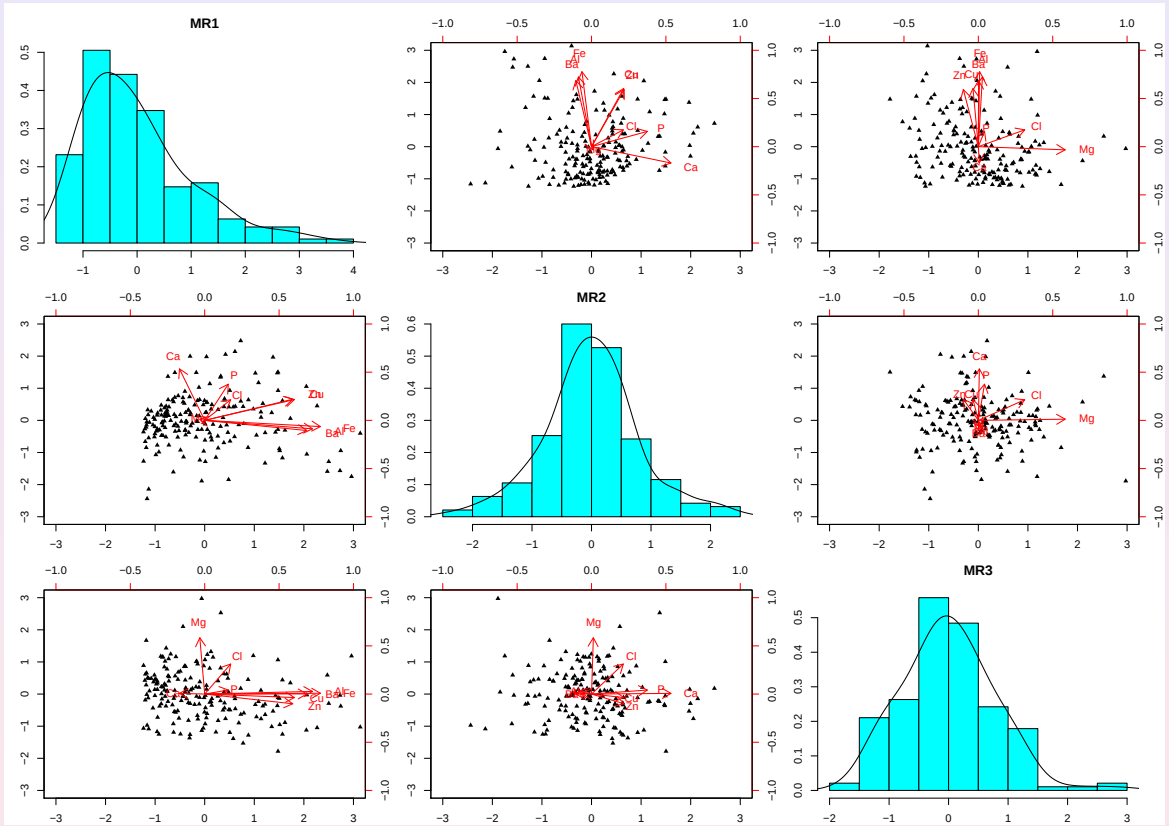


Figura: Gráfico *biplot* correspondente aos três fatores considerados para os dados do Exemplo 1.

# AF – Exemplo 1

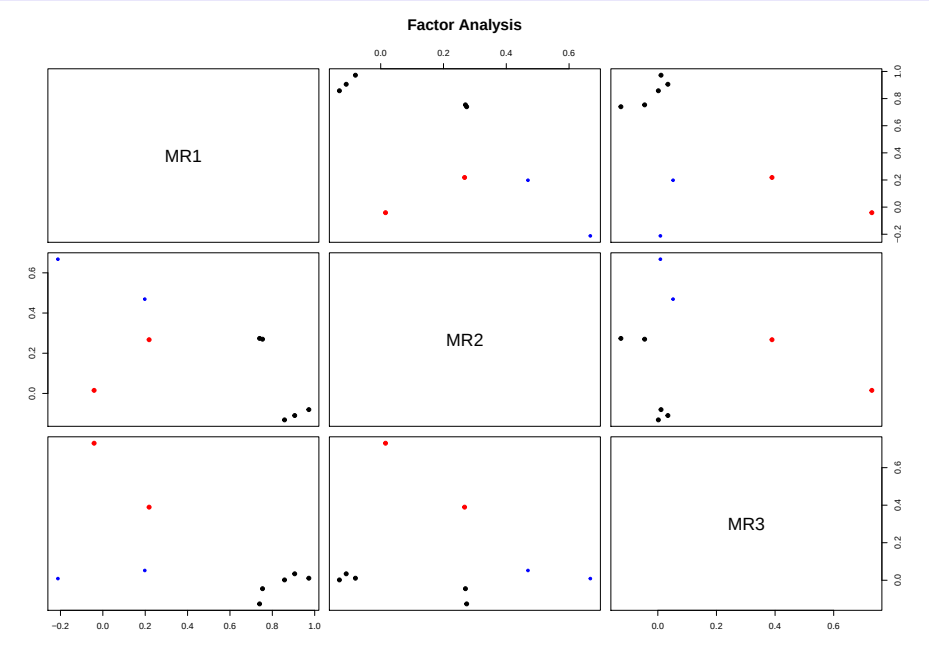


Figura: Gráfico cartesiano correspondente aos três fatores considerados para os dados do Exemplo 1.

O Fator 1 tem cargas fatoriais concentradas nos elementos Al, Ba, Cu, Fe e Zn, que estão associados à poluição de origem veicular gerada por desgaste de freios e ressuspensão de poeira; o Fator 2 está tem cargas fatoriais concentradas em P e Ca, características da saúde arbórea e o Fator 3 tem as cargas fatoriais concentradas em Cl e Mg.

# AF – Exemplo 2

- **Exemplo 2**. Vamos retomar o Exemplo 1 da Aula 15, em que se pretendia avaliar os efeitos de variáveis climáticas na ocorrência de suicídios por enforcamento na cidade de São Paulo.
- Nesse exemplo obtivemos duas CPs que explicavam 82% da variância total.
- Vamos usar a função factanal() do pacote stats. Essa função usa MV numa matriz de covariâncias ou numa matriz de dados.
- Obtemos os resultados abaixo:

```
Call:
factanal(x = clima1, factors = 2, rotation = "varimax")

Uniquenesses:
tempmax tempmin tempmed  precip  nebmed
  0.061   0.020   0.005   0.890   0.482

Loadings:
        Factor1 Factor2
tempmax  0.932  -0.266
tempmin  0.883   0.447
tempmed  0.997
precip           0.323
nebmed  -0.142   0.706

               Factor1 Factor2
SS loadings      2.668   0.874
Proportion Var   0.534   0.175
Cumulative Var   0.534   0.708

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 11.03 on 1 degree of freedom.
The p-value is 0.000895
```

# AF – Exemplo 2

- Na saída, encontramos o termo uniqueness, ou ruído, corresponde à proporção da variabilidade que não pode ser explicada pela combinação linear dos fatores(variância específica). Valor alto para uma variável indica que o fator não contribui bem para a sua variância. É a diagonal da matriz $\psi$. Nesse caso, os fatores não contribuem para a variância de Precipitação e Nebulosidade.
- A matriz $\lambda$ das cargas fatoriais é dada a seguir:

```
> load<-fafit\$loadings[,1:2]
            Factor1     Factor2
tempmax  0.93205114 -0.26553163
tempmin  0.88305960  0.44690513
tempmed  0.99713621  0.02715943
precip   0.07422757  0.32313465
nebmed  -0.14166420  0.70579832
```

- A matriz $\psi$ é dada por

```
> Psi <-diag(fafit\$uniquenesses)

          [,1]       [,2]  [,3]   [,4]      [,5]
[1,] 0.0607741 0.00000000 0.000 0.0000 0.0000000
[2,] 0.0000000 0.02048162 0.000 0.0000 0.0000000
[3,] 0.0000000 0.00000000 0.005 0.0000 0.0000000
[4,] 0.0000000 0.00000000 0.000 0.8901 0.0000000
[5,] 0.0000000 0.00000000 0.000 0.0000 0.4817635
```

- A comunalidade é obtida tomando os quadrados das cargas:

```
> apply(fafit\$loadings^2,1,sum)
  tempmax   tempmin   tempmed    precip    nebmed
```

# AF – Exemplo 2

- Não foi possível considerar 3 fatores, pois o programa não aceita valor maior do que 2 para 5 variáveis.
- A proporção da variância explicada pelos fatores é $70,8\%$, menor do que as CPs.
- A estimativa da matriz $\Sigma$ é dada por

```
             tempmax   tempmin     tempmed      precip
tempmax   1.00000048 0.7043893  0.92217026 -0.01661858
tempmin   0.70438926 1.0000001  0.89266839  0.20995790
tempmed   0.92217026 0.8926684  1.00001826  0.08279115
precip   -0.01661858 0.2099579  0.08279115  1.00002576
nebmed   -0.31945006 0.1903270 -0.12208942  0.21755251
             nebmed
tempmax -0.3194501
tempmin  0.1903270
tempmed -0.1220894
precip   0.2175525
nebmed   0.9999835
```

# AF – Exemplo 2

- A matriz residual estimada é

```
          tempmax   tempmin   tempmed    precip    nebmed
tempmax  0.000000 -0.000390  0.000104 -0.015332  0.009215
tempmin -0.000390  0.000000  0.000035 -0.008720  0.000870
tempmed  0.000104  0.000035 -0.000018  0.003171 -0.000927
precip  -0.015332 -0.008720  0.003171 -0.000026  0.072404
nebmed   0.009215  0.000870 -0.000927  0.072404  0.000017
```

  e vemos que os valores são próximos de zero, indicando que o modelo fatorial está adequado.

- Para determinar o número de fatores podemos obter os autovalores da matriz de correlações estimada:

```
> ev<-eigen(cor(clima1))
eigen() decomposition
values
[1] 2.70399548 1.41461150 0.71677349 0.14576231 0.01885722
```

  mostrando que tomamos 2 fatores, correspondentes aos valores próprios maiores que 1.

# AF – Exemplo 2

A seguir temos alguns gráficos mencionados anteriormente sobre a determinação do número de fatores.

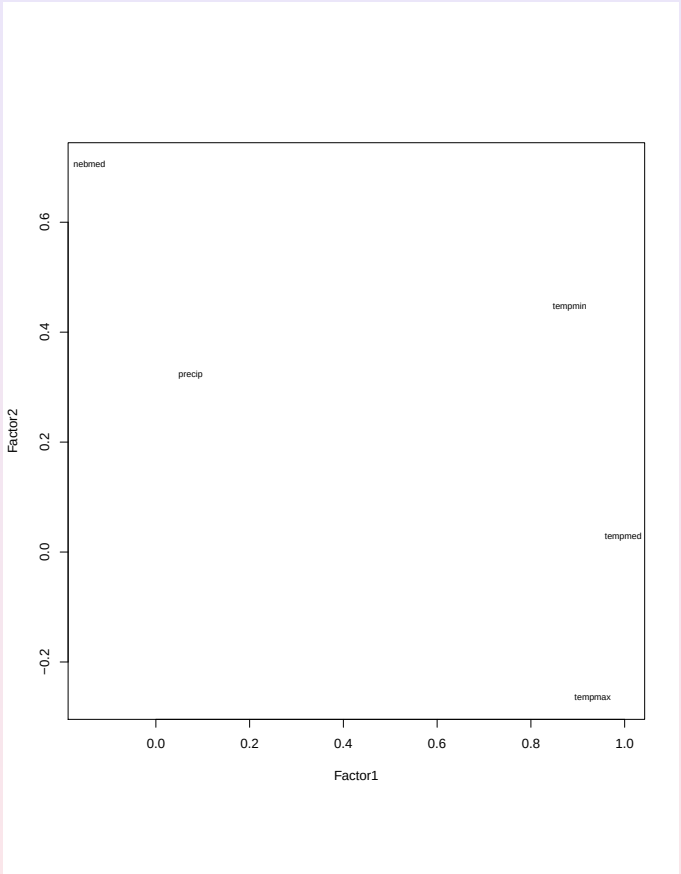


Figura: Gráfico do fator 1 vs fator 2 para o Exemplo2

# AF – Exemplo 2

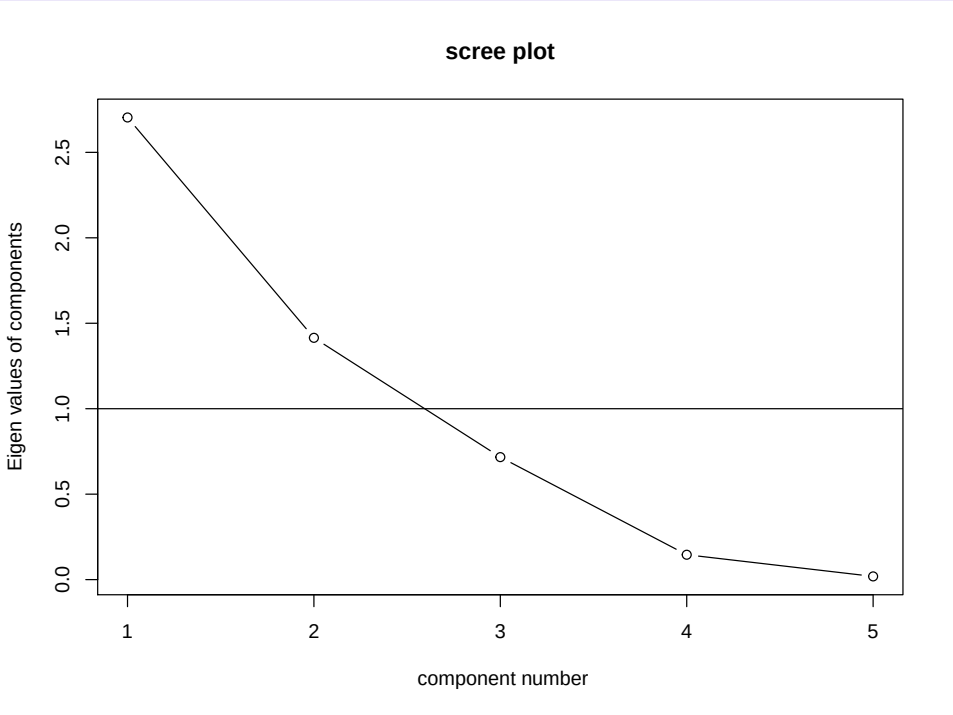


Figura: Scree plot para Exemplo 2.

# AF – Exemplo 2

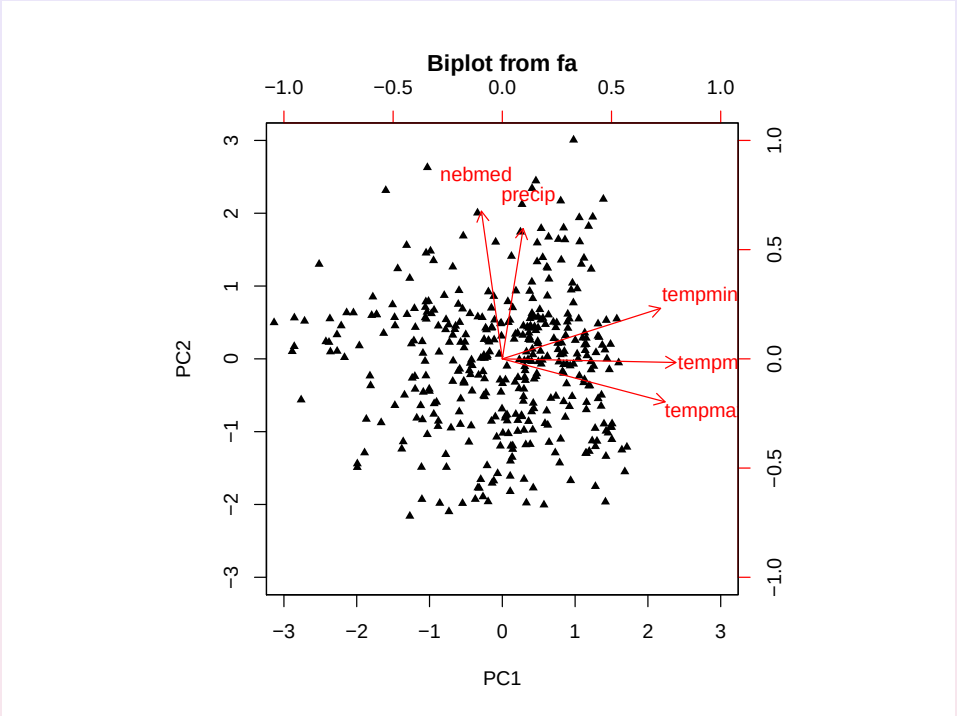


Figura: Biplot para Exemplo 2.

# AF – Exemplo 2

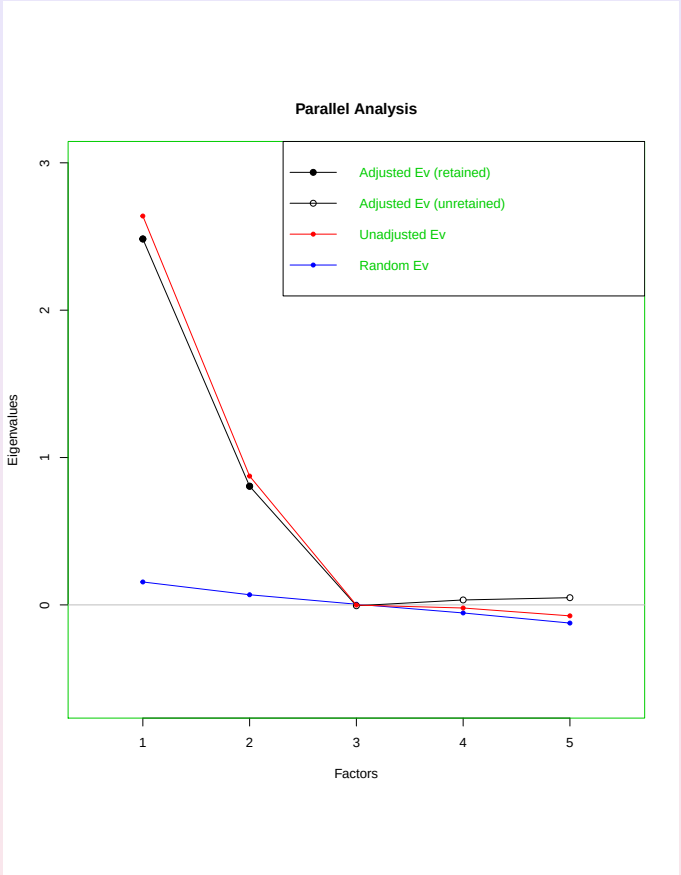


Figura: Análise Paralela para Exemplo 2.

# AF – Exemplo 2

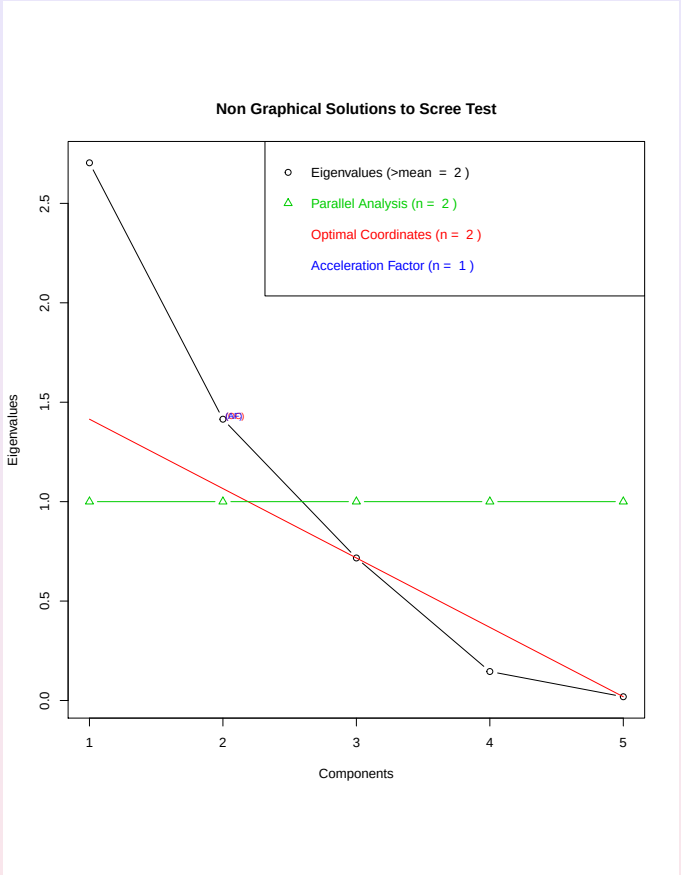


Figura: Soluções não gráficas para Exemplo 2.

# Referências

Bartlett, M.S. (1937). The statistical conception of mental factors. *British Journal of Psychology,* **28**, 97-104.

Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning*, 2nd Edition, Springer.

Härdle, W.K. and Simar, L. (2015). *Applied Multivariate Statistical Analysis.* Springer.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Morettin, P. A. e Singer, J. M. (2021). *Estatística e Ciência de Dados*. Texto Preliminar, IME-USP.

Morrison, D.F. (1976). *Multivariate Statistical Methods, 2nd Ed.* New York: McGraw-Hill.

Thurstone, L.L. (1947). *Multiple Factor Analysis: A development and expansion of vectors of the mind.*. Chicago: University of Chicago Press.