Análise Multivariada de Dados - Aula 06

Análise de Componentes Principais

Kaique Matias de Andrade Roberto

Ciências Atuariais

HECSA - Escola de Negócios

FIAM-FAAM-FMU

Conteúdo

Conceitos que aprendemos em Aulas anteriores

Conceitos que aprendemos em Aulas anteriores

Nas últimas Aulas focamos em aspectos teóricos e práticos dos Esquemas de Aglomeração (ou Análise de Cluster).

Conceitos que aprendemos em Aulas anteriores

Na última aula nós lidamos com alguns aspectos da teoria de matrizes:

- recapitulamos o que é uma matriz;
- relembramos algumas operações com matrizes;
- calculamos determinantes e inversas de matrizes;
- tivemos uma introdução à diagonalização.

Definição

Definição

A técnica de análise de componentes principais foi inicialmente descrita por Karl Pearson (1901) no contexto do estudo de problemas biométricos.

Definição

Uma descrição de métodos computacionais práticos veio muito mais tarde de Hotelling (1933). Mesmo então, os cálculos eram extremamente amedrontadores para mais do que poucas variáveis porque tinham que ser feitos à mão.

Definição

Somente após os computadores eletrônicos terem se tornado disponíveis generalizadamente é que a técnica de componentes principais alcançou amplo uso.

Definição

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1n}X_n$$

O objetivo da análise de componentes principais é, ao tomar p variáveis $X_1, ..., X_p$ e encontrar combinações lineares $Z_1, ..., Z_p$ que sejam não-correlacionados na ordem de sua importância, e que descreva a variação dos dados.

Definição

A falta de correlação significa que os índices estão medindo diferentes "dimensões"dos dados, e a ordem é tal que $\text{Var}(Z_1) > \text{Var}(Z_2) > ... > \text{Var}(Z_p)$.

Definição

Os índices Z_j's são os **componentes principais**.

Definição

Ao fazer uma análise de componentes principais, há sempre a esperança de que as variâncias da maioria dos índices serão tão baixas a ponto de serem desprezíveis.

Definição

Neste caso, a maior parte da variação no conjunto de dados completos pode ser descrita adequadamente pelas poucas variáveis Z com variâncias que não são desprezíveis, e algum grau de economia é então alcançado.

Definição

A análise de componentes principais nem sempre funciona, no sentido de que um grande número de variáveis originais são reduzidas a um pequeno número de variáveis transformadas.

Definição

De fato, se as variáveis originais são não correlacionadas, então a análise não chega a nada.

Definição

Os melhores resultados são obtidos quando as variáveis originais são altamente correlacionadas, positivamente ou negativamente.

Definição

Se este é o caso, então é bastante concebível que 20 ou mais variáveis originais possam ser adequadamente representadas por duas ou três componentes principais.

Definição

Se este estado desejável de relações de fato ocorre, então os componentes principais importantes serão de algum interesse como medidas das dimensões subjacente aos dados.

Definição

Será também de valor saber que há uma boa quantidade de redundância nas variáveis originais, com a maioria delas medindo coisas semelhantes.

Procedimento para uma análise de componentes principais

Procedimento para uma análise de componentes principais

Uma análise de componentes principais começa com dados de p variáveis e n indivíduos:

Procedimento para uma análise de componentes principais

Caso	X_1	X_2	...	X_p
1	X_{11}	X_{12}	...	X_{1p}
2	X_{21}	X_{22}	...	X_{2p}
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
n	X_{n1}	X_{n2}	...	X_{np}

Procedimento para uma análise de componentes principais

O primeiro componente principal é então uma combinação linear de $X_1, ..., X_p$

$$Z_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$

que varia tanto quanto possível para os indivíduos, sujeito à condição de que

$$a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = 1.$$

Procedimento para uma análise de componentes principais

Assim $\text{Var}(Z_1)$ é tão grande quanto possível dada esta restrição sobre as constantes a_{1j}. A restrição é introduzida porque se isto não é feito, então $\text{Var}(Z_1)$ pode ser aumentada fazendo simplesmente crescer qualquer um dos valores a_{1j}.

Procedimento para uma análise de componentes principais

O segundo componente principal

$$Z_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$

é escolhido de modo que $\text{Var}(Z_2)$ seja tão grande quanto possível sujeito à restrição de que

$$a_{21}^2 + a_{22}^2 + ... + a_{2p}^2 = 1$$

e também à condição de que Z_1 e Z_2 tenham correlação zero para os dados.

Procedimento para uma análise de componentes principais

O terceiro componente principal

$$Z_3 = a_{31}X_1 + a_{32}X_2 + ... + a_{3p}X_p$$

é tal que Var(Z_3) seja tão grande quanto possível sujeito à restrição de que

$$a_{31}^2 + a_{32}^2 + ... + a_{3p}^2 = 1$$

e também que Z_3 seja não correlacionada com ambas Z_1 e Z_2.

Procedimento para uma análise de componentes principais

Os próximos componentes principais são construídos da mesma maneira. Se existirem p variáveis, então existirão no máximo p componentes principais.

Procedimento para uma análise de componentes principais

Para se usar os resultados de uma análise de componentes principais, não é necessário saber como as equações, para os componentes principais, são obtidas.

Procedimento para uma análise de componentes principais

Entretanto, é útil entender a natureza das equações. De fato, uma análise de componentes principais envolve encontrar os autovalores de uma matriz de covariâncias amostrais.

Procedimento para uma análise de componentes principais

A matriz de covariâncias é simétrica e tem forma

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{pmatrix}$$

onde $c_{ii} = \text{Var}(X_i)$ e $c_{ij} = \text{Var}(X_i, X_j)$ se $i \neq j$.

Teorema Espectral: Se

$A \in M_n(\mathbb{R})$ é simétrica ($A = A^t$)

então A é diagonalizável.

Procedimento para uma análise de componentes principais

As variâncias dos componentes principais são os autovalores da matriz C. Existem p destes autovalores, alguns dos quais podem ser zero. Autovalores negativos não são possíveis para uma matriz de covariâncias.

Procedimento para uma análise de componentes principais

Assumindo que os autovalores estão ordenados como $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, então λ_i corresponde ao i-ésimo componente principal

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_p.$$

Procedimento para uma análise de componentes principais

Em particular, $\text{Var}(Z_i) = \lambda_i$, e as constantes $a_{i1}, a_{i2}, ..., a_{ip}$ são os elementos do correspondente autovetor, escalonado de modo que

$$a_{i1}^2 + a_{i2}^2 + ... + a_{ip}^2 = 1.$$

Procedimento para uma análise de componentes principais

Uma propriedade importante dos autovalores é que a soma deles é igual à soma dos elementos da diagonal (o traço) da matriz C. Isto é,

$$\lambda_1 + ... + \lambda_p = c_{11} + ... + c_{pp},$$

ou seja, a soma das variâncias dos componentes principais é igual á soma das variâncias das variáveis originais.

Procedimento para uma análise de componentes principais

Portanto, em certo sentido, os componentes principais contam com toda a variação nos dados originais.

Procedimento para uma análise de componentes principais

A fim de evitar uma ou duas variáveis tendo uma indevida influência nos componentes principais, é usual codificar as variáveis $X_1, X_2, ..., X_p$ para terem médias zero e variâncias um no início de uma análise.

Procedimento para uma análise de componentes principais

A matriz C então toma a forma

$$C = \begin{pmatrix} 1 & c_{12} & \cdots & c_{1p} \\ c_{21} & 1 & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & 1 \end{pmatrix}$$

onde $c_{ij} = \text{Correl}(X_i, X_j)$.

Procedimento para uma análise de componentes principais

Em outras palavras, a análise de componentes principais é feita sobre a matriz de correlação. Neste caso, a soma dos termos da diagonal, e, portanto, a soma dos autovalores, é igual a p, o número de variáveis X.

Procedimento para uma análise de componentes principais

Os passos em uma análise de componentes principais podem agora ser estabelecidos:

Procedimento para uma análise de componentes principais

Passo 1

Comece padronizando as variáveis $X_1, X_2, ..., X_p$ para terem médias zero e variâncias unitárias. Isto é usual, mas é omitido em alguns casos em que se assume que a importância das variáveis é refletida em suas variâncias.

Procedimento para uma análise de componentes principais

Passo 2

Calcule a matriz de covariâncias C. Esta é uma matriz de correlações se o passo 1 foi feito.

Procedimento para uma análise de componentes principais

Passo 3

Encontre os autovalores $\lambda_1, ..., \lambda_p$ e os correspondentes autovetores $v_1, v_2, ..., v_p$. Os coeficientes do i-ésimo componente principal são então os coeficientes de v_i, enquanto que λ_i é sua variância.

Procedimento para uma análise de componentes principais

Passo 4

Descarte quaisquer componentes que explicam somente uma pequena proporção da variação nos dados. Por exemplo, começando com 20 variáveis, pode ser obtido que os primeiros três componentes expliquem 90% da variância total. Com base nisto, os outros 17 componentes podem ser razoavelmente ignorados.

Exemplos

Exemplos

$$Z_1 = 0.452 X_1 + 0.462 X_2 + \ldots + 0.398 X_5$$

Componente	Autovalor	Autovetores (coeficientes para os componentes principais)				
		X_1	X_2	X_3	X_4	X_5
1	3,616	0,452	0,462	0,451	0,471	0,398
2	0,532	–0,051	0,300	0,325	0,185	–0,877
3	0,386	0,691	0,341	–0,455	–0,411	–0,179
4	0,302	–0,420	0,548	–0,606	0,388	0,069
5	0,165	0,374	–0,530	–0,343	0,652	–0,192

Nota: Os autovalores são as variâncias dos componentes principais. Os autovetores dão os coeficientes das variáveis X padronizadas usadas para calcular os componentes principais.

$$Var(Z_1) = 3.616$$

Exemplos

Os pássaros sendo considerados foram pegos após uma forte tempestade. Os primeiros 21 deles se recuperaram, os outros 28 morreram. Uma questão de interesse é, portanto, se os sobreviventes e não-sobreviventes mostram alguma diferença.

Exemplos

A situação será agora considerada em termos das componentes principais.

Comentários Finais

Comentários Finais

Em resumo, na aula de hoje nós:

- aprendemos o que é a Análise de Componentes Principais (ACP);
- executamos no Excel uma ACP em um exemplo proveniente do livro-texto.

Comentários Finais

Na próxima aula nós lidaremos com a Análise Fatorial.

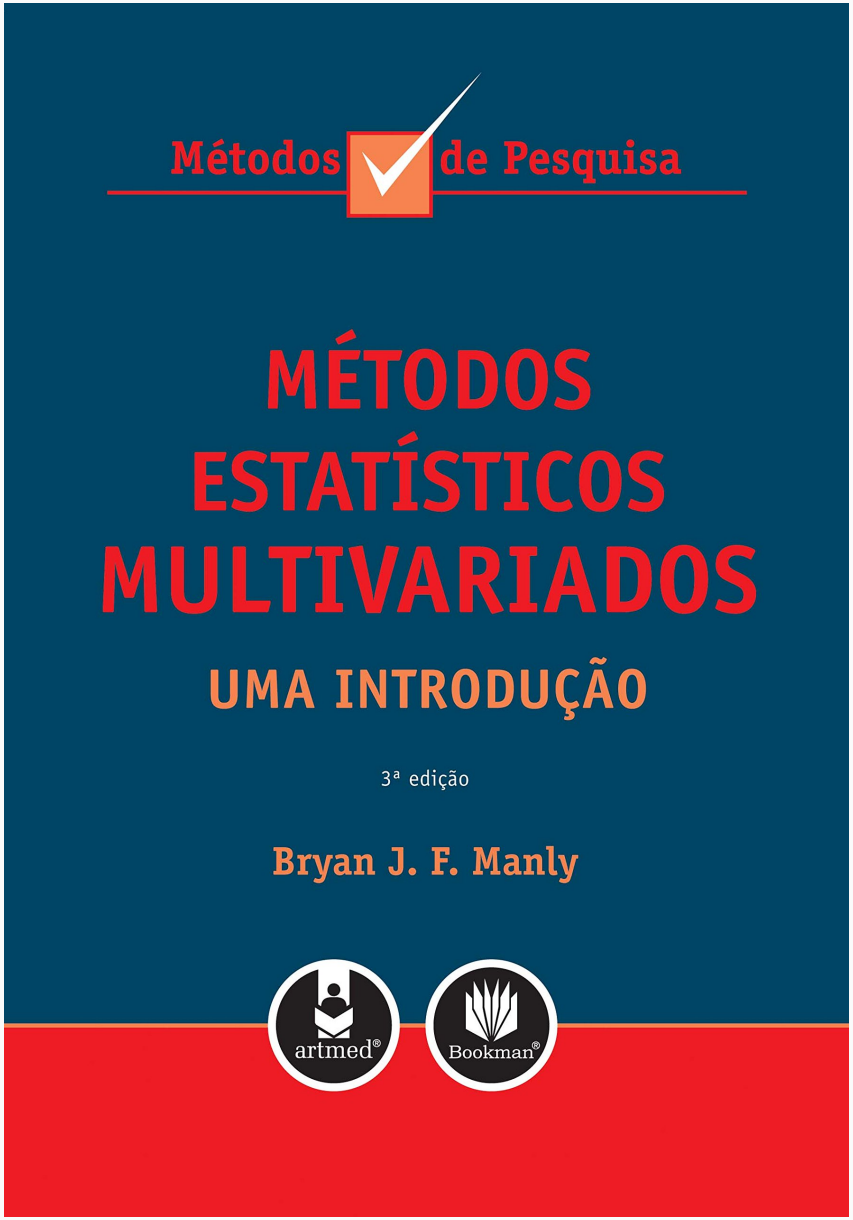
Comentários Finais

EXERCÍCIOS PARA APS (E PREPARAÇÃO PARA A N2)

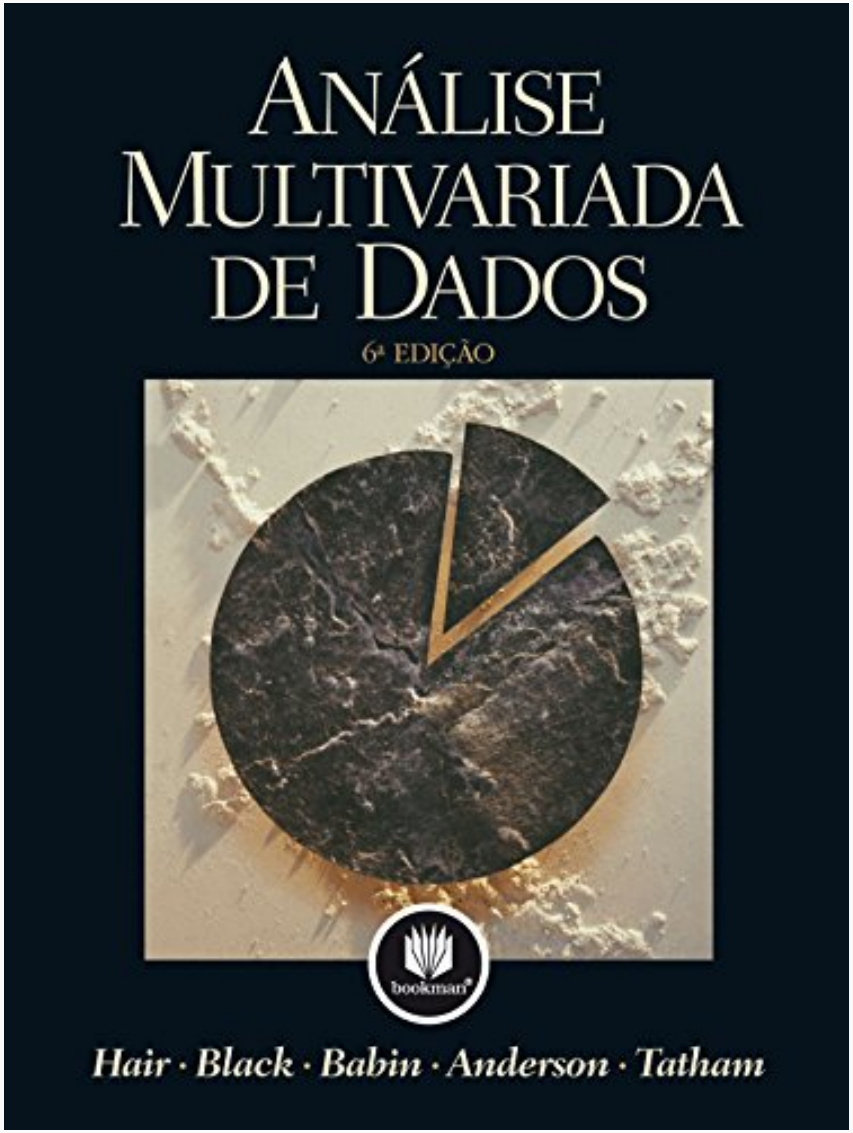
resolva os Exercícios 6.1-6.3.

Referências

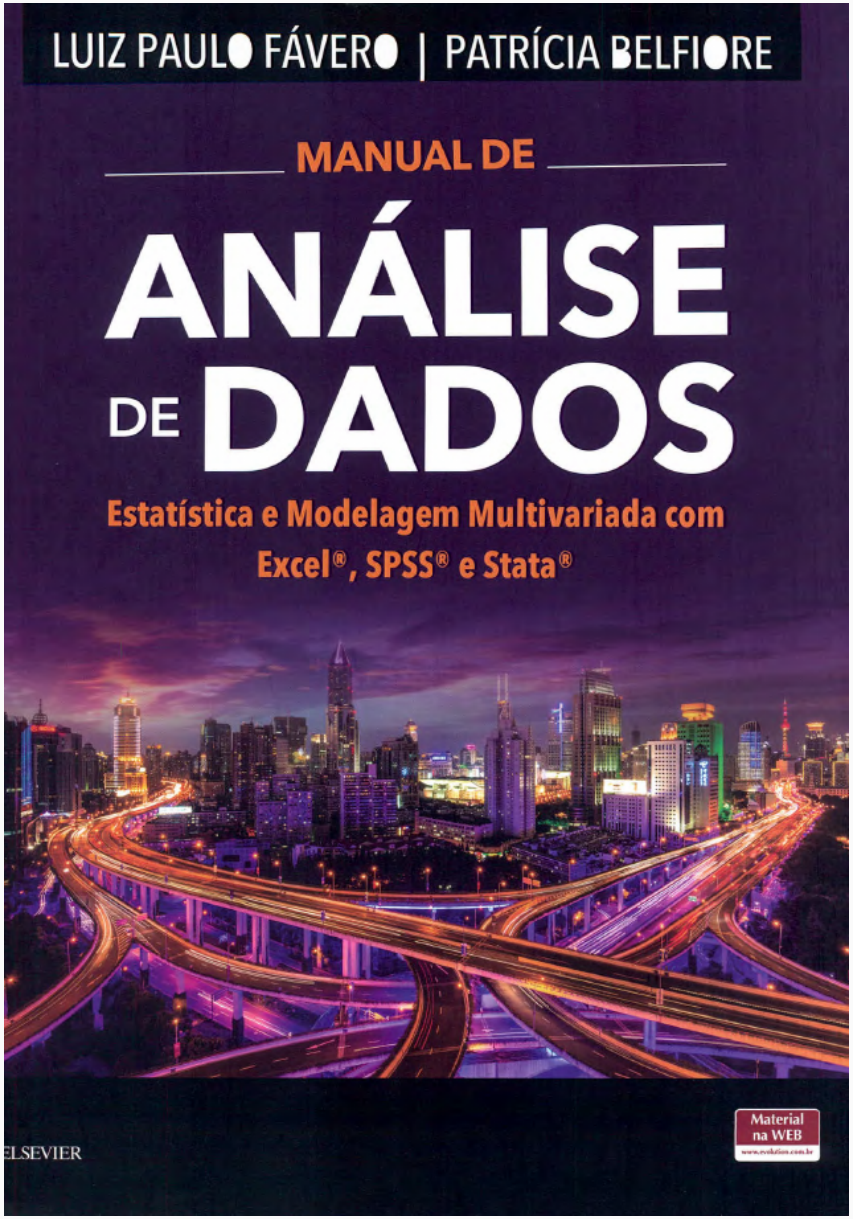
Referências



Referências



Referências



Bons Estudos!