

# 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 15

7 de junho de 2023

# Sumário

- 1 Análise de Componentes Independentes
- 2 Decomposição de matrizes

# Preliminares

- Vimos, no estudo de componentes principais, que essas são não correlacionadas, ou ortogonais. No caso da análise fatorial, há a suposição adicional de normalidade. Na análise de componentes independentes (ICA, em Inglês), a ideia é obter componentes independentes e não gaussianas.
- A ICA é relativamente recente, sendo introduzida na década de 1980 no contexto de redes neurais e encontra aplicações em processamento de sinais biomédicos, separação de sinais de áudio, séries temporais financeiras etc.
- As referências principais nessa área são Hyvärinen e Oja (1997), Hyvärinen (1999), Hyvärinen et al. (1999), Stone (2004) e Gegembauer (2010).
- Vamos supor que os dados consistem de um número  $p$  de variáveis,  $\mathbf{X} = (X_1, \dots, X_p)^\top$ , e consideramos a transformação

$$Y_i = \sum_{j=1}^p w_{ij} X_j(t), \text{ para } i = 1, \dots, q, \quad (1)$$

onde  $q < p$  e os  $w_{ij}$  são alguns coeficientes que definem a representação.

- Reunindo os coeficientes  $w_{ij}$  em uma matrix  $\mathbf{W}$ , de ordem  $q \times p$ , e os  $Y_j$  num vetor  $\mathbf{Y}$ , de ordem  $q \times 1$ , obtemos

$$\mathbf{Y} = \mathbf{W}\mathbf{X}$$

## Separação cega de fontes

- A diferença com CP e AF é que escolhemos  $\mathbf{W}$  de modo que as variáveis  $Y_i$  sejam independentes e não gaussianas. ICA é um caso particular da **Separação Cega de Fontes** (*blind source separation*).
- Considere a situação onde existem pessoas conversando na mesma sala, assim emitindo sinais de fala; ou telefones celulares emitindo suas ondas de rádio. Suponha que haja vários sensores ou receptores que estão em diferentes posições, sendo assim cada mistura das fontes originais é gravada com pesos um pouco diferentes.
- Com o objetivo de simplificar a exposição, digamos que existem três fontes que dão origem aos sinais, e também três sinais observados, denotados por  $X_1(t)$ ,  $X_2(t)$  e  $X_3(t)$ , os quais são as amplitudes dos sinais gravados no tempo  $t$ , e por  $S_1(t)$ ,  $S_2(t)$  e  $S_3(t)$  os sinais originais. Os  $X_i(t)$  são uma combinação linear dos  $S_i(t)$  com coeficientes constantes  $a_{ij}$ , os quais dão as misturas dos pesos e que dependem da distância entre as fontes e os sensores e se supõe que sejam desconhecidos:

$$X_1(t) = a_{11}S_1(t) + a_{12}S_2(t) + a_{13}S_3(t), \quad (3)$$

$$X_2(t) = a_{21}S_1(t) + a_{22}S_2(t) + a_{23}S_3(t),$$

$$X_3(t) = a_{31}S_1(t) + a_{32}S_2(t) + a_{33}S_3(t).$$

# Separação cega de fontes

- Em geral, não conhecemos os valores  $a_{ij}$  e a fonte do sinal  $S_i(t)$  também é desconhecida. O que gostaríamos de fazer é encontrar os sinais originais a partir das misturas  $X_i(t)$ . Isso se chama **Separação Cega de Fontes** (*Blind Source Separation*).
- O termo **cega** significa que temos pouca ou nenhuma informação a respeito das fontes.
- Supomos que os coeficientes de mistura  $a_{ij}$  sejam diferentes o suficiente, para tornar a matriz que formam invertível. Então, existe uma matriz **W** com coeficiente  $w_{ij}$ , tal que podemos escrever os  $S_i$  como:

$$S_i(t) = w_{i1}X_1(t) + w_{i2}X_2(t) + w_{i3}X_3(t), \quad i = 1, 2, 3, \quad (4)$$

- Tal matriz **W** pode ser encontrada como a inversa da matriz que consiste dos coeficientes de mistura  $a_{ij}$  em (3) se conhecermos os coeficientes  $a_{ij}$ .

## Separação cega de fontes

- A questão agora é: como podemos estimar os coeficientes  $w_{ij}$  em (4) ? Observamos os sinais  $X_1$ ,  $X_2$  e  $X_3$  e queremos achar uma matriz  $\mathbf{W}$  de modo que a representação seja dada pelos sinais da fonte original  $S_1$ ,  $S_2$  e  $S_3$ .
- Uma solução simples ao problema pode ser encontrada, considerando independência estatística dos sinais. De fato, se os sinais são não gaussianos, isto é suficiente para determinar os coeficientes  $w_{ij}$  dado que os sinais

$$Y_i(t) = w_{i1}X_1(t) + w_{i2}X_2(t) + w_{i3}X_3(t), \quad i = 1, 2, 3, \quad (5)$$

sejam estatisticamente independentes. Se os sinais  $Y_i$ , são independentes, então eles são iguais aos sinais originais  $S_i$ ,  $i = 1, 2, 3$  (a menos da multiplicação por um escalar). Usando apenas esta informação de independência estatística, podemos de fato estimar os coeficientes da matriz  $\mathbf{W}$ .

## ICA – metodologia

- O que vimos acima nos leva a seguinte definição da ICA. Dado um vetor de variáveis aleatórias  $\mathbf{X} = (X_1, X_2, \dots, X_p)^\top$ , supõe-se que estas sejam geradas como uma mistura linear de componentes independentes  $\mathbf{S} = (S_1, S_2, \dots, S_p)^\top$ :

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = \mathbf{A} \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_p \end{pmatrix}, \quad (6)$$

ou

$$\mathbf{X} = \mathbf{AS}, \quad (7)$$

onde  $\mathbf{A}=[a_{ij}]$  é uma matriz desconhecida (chamada *mixing*).

- A análise de componentes independentes consiste agora em estimarmos tanto a matriz  $\mathbf{A}$  como os  $S_i$ , onde apenas observamos os  $X_i$ . Observe que supomos que o número de componentes independentes  $S_i$  é igual ao número de variáveis observadas; isso é uma simplificação que não é completamente necessária.

# ICA – metodologia

- Pode ser demonstrado que este problema é bem definido, isto é, o modelo em (6) pode ser estimado se e somente se os componentes  $S_i$  são não gaussianos.
- Essa é uma necessidade fundamental que também explica a principal diferença entre ICA e análise fatorial, na qual a normalidade dos dados é levada em conta.
- De fato, ICA pode ser considerada como uma análise fatorial não gaussiana, visto que nesta também modelamos os dados como uma mistura linear de alguns fatores latentes.
- Além da estimação, tem-se que encontrar um algoritmo para executar os cálculos necessários. Como o princípio de estimação usa funções não quadráticas, os cálculos necessários, usualmente, não podem ser expressos usando algebra linear simples, podendo ser extremamente complexos. Algoritmos numéricos são assim uma parte integral dos métodos de estimação ICA.



## ICA – metodologia

- Os métodos numéricos são tipicamente baseados na otimização de algumas funções objetivas. O método básico de otimização é o método do gradiente. De interesse particular tem-se o algoritmo de ponto fixo chamado **FastICA**, que tem sido usado para explorar uma particular estrutura dos problemas ICA. Por exemplo, podemos usar esses métodos para encontrar o máximo de não normalidade como medido pelo valor absoluto do excesso de curtose.
- Estimando-se a matriz  $\mathbf{A}$ , supondo-se que exista sua inversa  $\mathbf{A}^{-1} = \mathbf{W}$ , obtemos

$$\mathbf{S} = \mathbf{W}\mathbf{X}. \quad (8)$$

- Há duas ambiguidades no procedimento da ICA. A primeira é que não podemos determinar as variâncias das  $S_i$ , pois se as multiplicarmos por uma constante  $C$ , basta dividir as colunas correspondentes de  $\mathbf{A}$  por  $C$ . Dessa maneira, o que fazemos é centrar as componentes, isto é,  $E(S_i) = 0$ , para todo  $i$ , e fixar as magnitudes das componentes independentes (c.i.'s), de modo que  $E(S_i^2) = 1$ , para todo  $i$ .

## ICA – metodologia

- A segunda ambiguidade é que, contrariamente ao que ocorre com a ACP, não podemos determinar a ordem das c.i., pois como **S** e **A** são desconhecidas, podemos mudar livremente a ordem dos termos em (7).
- Por que variáveis gaussianas não são permitidas? Para responder, considere o caso particular (3), suponha que **A** seja **ortogonal** e que as  $S_i(t)$  sejam gaussianas. Então, as variáveis  $X_i(t)$  são gaussianas (combinações lineares de v.a.s independentes e gaussianas), não correlacionadas (por que?) e de variância unitária. Sua densidade conjunta é dada por

$$f(x_1, x_2, x_3) = \frac{1}{(2\pi)^{3/2}} e^{-(x_1^2 + x_2^2 + x_3^2)/2},$$

que é simétrica, logo não contém nenhuma informação sobre as direções das colunas de **A**. Logo, **A** não pode ser estimada, ou ainda, **A** não é identificada para componentes gaussianas independentes.

## ICA – metodologia

- Seja  $y = \mathbf{v}'\mathbf{X} = \sum_{i=1}^p v_i X_i$ , onde  $\mathbf{v}$  é um vetor  $p \times 1$ . Se  $\mathbf{v}$  fosse uma das linhas de  $\mathbf{A}^{-1} = \mathbf{W}$ , essa combinação linear seria uma das c.i.'s  $S_i$ . O problema reduz-se, então, a determinar um vetor  $\mathbf{v}$  nessas condições.
- Pode-se provar que maximizando-se a não gaussianidade de  $\mathbf{v}'\mathbf{X}$  nos dá uma das componentes de  $\mathbf{S}$ . Para tanto, teremos que obter  $2p$  máximos locais, 2 para cada c.i. (correspondendo a  $S_i$  e  $-S_i$ ). Como as c.i.'s são não correlacionadas, podemos restringir a busca no espaço que fornece estimativas não correlacionadas com as anteriores e isso corresponde a **branquear** (*whitening*) as observações.
- Como obter componentes independentes? Podemos escolher duas formas como *proxy* de independência, que por sua vez determinam a forma do algoritmo ICA a usar:
  1. **Minimização da informação mútua** (MIM);
  2. **Maximização da não gaussianidade** (MNG).
- A família de algoritmos que usa MIM utiliza medidas como a **Divergência de Kullback-Leibler e Máxima Entropia**. A família que aborda a não gaussianidade, usa **curtose e negentropia**.

## ICA – Pré-processamento

- Como passos de pré-processamento, os algoritmos usam centragem (subtrair a média para obter um sinal de média zero), branqueamento e redução da dimensionalidade, usualmente via ACP e decomposição em valores singulares. O branqueamento assegura que todas as dimensões são tratadas igualmente antes que o algoritmo seja aplicado.
- Se a  $Cov(X, Y) = E(XY) - E(X)E(Y) = 0$  dizemos que  $X$  e  $Y$  são não correlacionadas. Se  $X$  e  $Y$  são independentes, então são não correlacionadas. Porém, não correlação não implica em independência, ou seja covariância zero não implica necessariamente em independência, a não ser que  $(X, Y)$  tenham distribuição gaussiana bivariada.
- Uma propriedade um pouco mais forte que não correlação é **brancura** (*whiteness*). Branqueamento de um vetor de média zero,  $\mathbf{y}$ , significa que seus componentes são não correlacionados e suas variâncias são unitárias. Em outras palavras, a matriz de covariância (assim como a matriz de correlação) de  $\mathbf{y}$  é igual a matriz identidade:

$$E[\mathbf{y}\mathbf{y}^T] = \mathbf{I}. \quad (9)$$

## ICA – Pré-processamento

- Consequentemente, branqueamento significa que transformamos linearmente o vetor de dados observados  $\mathbf{x}$  através de uma multiplicação linear com alguma matriz  $\mathbf{V}$

$$\mathbf{z} = \mathbf{V}\mathbf{x}, \quad (10)$$

obtendo então um novo vetor  $\mathbf{z}$  que é **branco**. O branqueamento é sempre possível. Um método bastante popular é usando a **decomposição em valores singulares** (*singular value decomposition-SVD*) da matriz de covariâncias da variável centrada, digamos  $\tilde{\mathbf{X}}$ :

$$E[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top] = \mathbf{E}\mathbf{D}\mathbf{E}^\top, \quad (11)$$

onde  $\mathbf{E}$  é a matriz ortogonal de autovetores de  $E[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top]$  e  $\mathbf{D}$  é a matriz diagonal de seus autovalores,  $\mathbf{D} = \text{diag}(d_1, \dots, d_p)$ .

- Branqueamento pode agora ser feito pela matriz de branqueamento

$$\mathbf{V} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^\top. \quad (12)$$

## ICA – Pré-processamento

- O branqueamento transforma  $\mathbf{A}$  em  $\tilde{\mathbf{A}}$ , tal que

$$\tilde{\mathbf{X}} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^\top \mathbf{A} \mathbf{S} = \tilde{\mathbf{A}}\mathbf{S}.$$

- O procedimento torna  $\tilde{\mathbf{A}}$  ortogonal, pois

$$E[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top] = \tilde{\mathbf{A}}E(\mathbf{S}\mathbf{S}^\top)\tilde{\mathbf{A}}^\top = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top = \mathbf{I}.$$

- Ainda, o processo de branqueamento reduz o número de parâmetros a estimar, de  $p^2$  para  $p(p-1)/2$ , devido à ortogonalidade da matriz  $\tilde{\mathbf{A}}$ .

# ICA – Algoritmos

- Há vários algoritmos para a análise de componentes independentes de um conjunto de dados: Algoritmo do Gradiente usando a Curtose, Algoritmo de Ponto Fixo usando Curtose, Algoritmo do Gradiente usando Negentropia, Algoritmo de Ponto Fixo usando Negentropia, Algoritmos para Estimação de Máxima Verossimilhança. Veja Gegembauer (2010) para detalhes.
- Algoritmos tradicionalmente utilizados para a ICA incluem **infomax**, **FastICA** e **JADE**.
- O pacote **ica**, no R, contempla esses algoritmos.
- Outro pacote: **ProDenICA**, este desenvolvido por T. Hastie e R. Tibshirani, que apresenta a particularidade de estimar a densidade de cada componente.

# ICA – Algoritmos

- Um número  $\xi$  diz-se um **ponto fixo** de uma função  $\varphi$  se  $\varphi(\xi) = \xi$ .
- Exemplo:  $\varphi(x) = x^2 - 2$  tem dois pontos fixos,  $-1$  e  $2$
- Encontrar pontos fixos e raízes de polinômios algébricos são problemas equivalentes.
- Se existir um ponto fixo de uma função  $\varphi(x)$  num intervalo  $[a, b]$ , ele pode ser encontrado pelo **algoritmo do ponto fixo**:

$$x^{(k+1)} = \varphi(x^{(k)}), \quad k \geq 0. \quad (13)$$

O algoritmo pode não convergir.

- O método de Newton-Raphson é um algoritmo de ponto fixo.
- **Teorema de convergência do ponto fixo**: Se  $\varphi$  for contínua e  $x^{(k)}$  gerada por (13) e se existir  $\xi$  tal que  $\lim_{k \rightarrow \infty} x^{(k)} = \xi$ , então  $\xi$  é um ponto fixo de  $\varphi$ .



# ICA – Algoritmos

- O pacote **FastICA**, desenvolvido por Hyvärinen and Oja (1997), usa uma aproximação da negentropia e o algoritmo é baseado em iterações de ponto fixo. É mais robusto que métodos baseados na curtose.
- **JADE**: Joint approximate diagonalization of eigenmatrices. Usa métodos matriciais.
- **InfoMax**: pacote que minimiza a **informação mútua**:

$$I(X, Y) = H(X) - H(X|Y), \quad (14)$$

onde  $H(X|Y)$  é a entropia condicional ( $H(X, Y) - H(Y)$ ) e  $H(X)$  é a entropia de  $X$ .

- No caso contínuo,

$$I(X, Y) = \int f(x, y) \log \frac{f(x, y)}{f_X(x)f_Y(y)} dx dy. \quad (15)$$

- $I(X, Y) = 0$  se, e somente se,  $X$  e  $Y$  independentes.
- Amari et al. (1996): algoritmo InfoMax.

# ICA – Algoritmos

- Uma maneira de extrair as componentes é forçá-las a serem as mais não normais possíveis.
- **Negentropia** pode ser usada para estimar não-normalidade; é uma medida de distância da normalidade:

$$N(X) = H(X_{\text{normal}}) - H(X). \quad (16)$$

- Para uma dada matriz de covariância, a distribuição que tem a maior entropia é a gaussiana. Assim, a negentropia é uma medida estritamente positiva da não-gaussianidade.
- Como é difícil calcular a negentropia pela definição (16), usa-se a aproximação de Hyvärinen and Oja.

# ICA – Exemplo 1

- **Exemplo 1.** Vamos retomar o exemplo visto na seção de Análise Fatorial, sobre poluição, com 9 variáveis.
- Usaremos o pacote `ica` do R, que tem notação diferente para as diferentes matrizes. Em cada caso faremos a correspondência entre as notações.
- A matriz **M** (*mixing*) estimada, de ordem  $9 \times 9$ , correspondente à matriz **A** do texto, é

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
[1,]	-139.3790190	-149.7021007	-325.3324853	1077.925146	155.3171887	68.243758	111.7946050	-274.782659
[2,]	-0.4866963	-0.2949352	-0.5650103	1.190811	0.3524588	0.355581	0.4229097	-0.879921
[3,]	-12.4332654	2.4281842	-25.8117720	42.551213	13.2787874	13.166554	16.9792887	-77.566097
[4,]	-42.4416295	-15.1698381	-86.3634728	405.249890	356.4777419	14.649072	-24.5665422	-99.761388
[5,]	-264.2307634	-730.1007230	224.9821643	-230.382230	38.0537401	123.889785	9.4797710	87.720764
[6,]	-109.9753576	-35.9892578	-184.6838704	661.491067	263.0009705	33.477923	308.4517253	-40.708621
[7,]	-121.4505476	21.0283925	-136.7716779	-19.803719	5.3052447	19.311117	15.3286940	2.762610
[8,]	-17.9760476	-19.9296505	2.3521578	10.763812	14.0445497	232.900590	-0.7320548	4.559599
[9,]	-7362.9405123	3856.0539441	3007.1595557	-1949.521311	-1347.9118353	1263.932257	-977.7231474	-677.758695
	[,9]							
[1,]	15.571397							
[2,]	-1.274516							
[3,]	-8.172252							
[4,]	-1.558269							
[5,]	227.872428							
[6,]	57.382165							
[7,]	14.380424							
[8,]	6.372196							
[9,]	-609.643712							

## ICA – Exemplo 1

- Temos, por exemplo,

$$X_1 = -139,379S_1 - 149,702S_2 + \dots + 15,571S_9.$$

- O programa também fornece as matrizes :

**S**, de ordem  $193 \times 9$ , matriz estimada das componentes independentes, obtida por

$$\mathbf{S} = \mathbf{W}\mathbf{X} = \mathbf{Y}\mathbf{R};$$

**W**, de ordem  $9 \times 9$ , matriz *un-mixing*, tal que  $\mathbf{X}\mathbf{Y}\mathbf{W} = \mathbf{S}$ . **W** é escolhida de modo a maximizar a negentropia, com a restrição de que **W** seja ortogonal;

**Y**, de ordem  $193 \times 9$ , matriz pré-branqueada;

**Q**, de ordem  $9 \times 15$ , matriz de pré-branqueamento, tal que  $\mathbf{Y} = \mathbf{Q}\mathbf{X}$ ;

**R**, de ordem  $9 \times 9$ ; matriz de rotação ortogonal, tal que  $\mathbf{S} = \mathbf{Y}\mathbf{R}$ .

# ICA – Exemplo 1

- Fornece, também:

vafs : variância explicada por cada c.i.

alg : algoritmo usado (parallel ou deflation)

fun : função contraste (para aproximar a negentropia)

alpha : tuning parameter

iter : número de iterações do algoritmo

- As variâncias explicadas pelas c.i's são dadas por

0.600985581 0.170645639 0.102434106 0.062162898 0.022555357

0.018513146 0.011778296 0.006196794 0.004728182

- Vemos que as três primeiras componente explicam cerca de 88% da variância dos dados. Na Figura 1 temos os gráficos das primeiras duas componentes independentes.

# ICA – Exemplo 1

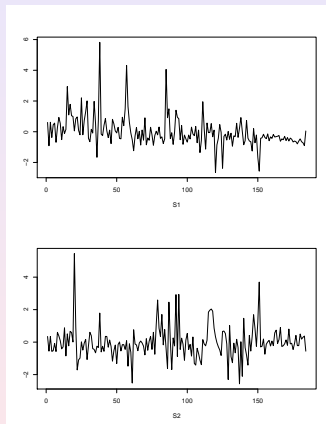


Figura: Gráficos das duas primeiras c.i.'s para o Exemplo 13.3.

# ICA – Exemplo 1

A matriz **W** estimada é dada por

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
[1,]	-1.750287e-04	-0.074821293	0.0018890899	-1.219544e-04	-4.394832e-04	-9.771709e-05	-1.951470e-03	9.258294e-04
[2,]	-3.688197e-04	-0.193146765	0.0014816133	1.790061e-04	-1.018530e-03	5.109103e-04	-8.073011e-05	5.379454e-04
[3,]	-2.673837e-04	-0.024630937	0.0003486251	1.736989e-05	2.822484e-04	6.088192e-04	-5.353500e-03	-1.075091e-04
[4,]	1.176365e-03	-0.020637504	-0.0041025331	-1.933696e-04	-1.798993e-04	-3.679419e-05	-1.609738e-03	5.375905e-05
[5,]	-1.511512e-03	0.035410750	0.0012706791	2.872724e-03	1.468919e-04	5.799778e-04	5.234842e-04	1.298423e-05
[6,]	-2.234726e-05	-0.004478221	0.0004852519	-1.619936e-04	-1.281089e-04	-1.639626e-05	-1.038226e-04	4.386960e-03
[7,]	-1.548268e-03	0.105156093	0.0053742429	-2.024956e-03	9.977071e-05	3.208043e-03	-9.018445e-04	-2.629357e-04
[8,]	1.119943e-04	0.110425897	-0.0148241124	-5.879721e-05	-1.557664e-04	5.604000e-04	1.003537e-03	5.855453e-04
[9,]	3.545273e-04	-0.752508234	0.0074526851	-6.531333e-05	2.558401e-04	4.482189e-04	7.568601e-04	3.558750e-04
[,9]								
[1,]	-8.288335e-05							
[2,]	4.515382e-05							
[3,]	7.534598e-05							
[4,]	2.056474e-05							
[5,]	-1.503345e-05							
[6,]	-3.322196e-06							
[7,]	-1.102442e-05							
[8,]	-4.811093e-06							
[9,]	1.592838e-06							

Assim, por exemplo,

$$S_1 = -0,000175X_1 - 0,0748X_2 + \dots + 0,0000829X_9.$$

Vemos que somente as variáveis  $X_3=Zn$ , e  $X_4=Ba$  são as mais importantes para explicar as componentes independentes  $S_i$ ,  $i = 1, \dots, 9$ .

## ICA – Exemplo 1

O pacote ProDenICA do R, fornece as densidades estimadas das c.i.'s, mostradas na Figura 2.

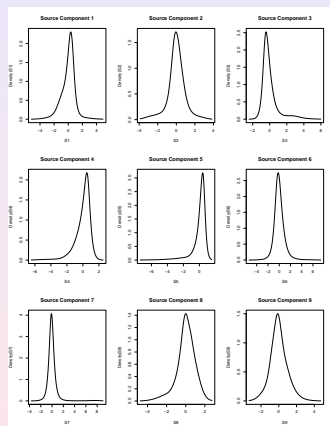


Figura: Gráficos das densidades das c.i.'s para o Exemplo 1.



# ICA – Exemplo 1

Vamos considerar, agora, três componentes independentes. A matriz **A** agora fica

	[,1]	[,2]	[,3]
[1,]	130.0588599	-1172.278514	-50.1191462
[2,]	0.4506871	-1.538268	-0.2837152
[3,]	17.7456231	-64.094123	-17.9385466
[4,]	46.0218183	-489.157909	-30.7345628
[5,]	215.8328265	141.834601	846.2085520
[6,]	85.0325691	-771.625038	-24.1875542
[7,]	76.2477919	-26.575128	-18.3258463
[8,]	61.9279843	-24.349130	49.9983656
[9,]	8353.6861764	3325.603526	-2506.6567435

## ICA – Exemplo 1

- A matriz **W** fica

```

      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]
[1,] 0.0002104413 2.494246e-07 7.925442e-06 8.257729e-05 3.307134e-04 1.416700e-04 1.255409e-05 3.473013e-05
[2,] -0.0005069541 -7.063526e-07 -2.903874e-05 -2.097610e-04 -8.165741e-06 -3.338968e-04 -2.364408e-05 -2.260694e-05
[3,] 0.0000287935 -1.201862e-07 -1.273793e-05 -2.839277e-06 1.091458e-03 2.931156e-05 7.179379e-06 8.385744e-05
      [,9]
[1,] 1.056008e-04
[2,] 1.310325e-05
[3,] -2.959094e-05

```

- A variância explicada por cada c.i. é mostrada abaixo:  
[1] 0.77283297 0.14705258 0.07750351
- Vemos que as duas primeiras c.i.'s explicam 92% da variância dos dados. Usando-se a função **fastICA** do pacote **fastICA** do R, podemos fazer alguns gráficos.
- A Figura 3 mostra os dados pré-processados, as componentes principais estimadas e as componentes independentes estimadas. Na Figura 4 temos os gráficos das três componentes.

# ICA – Exemplo 1

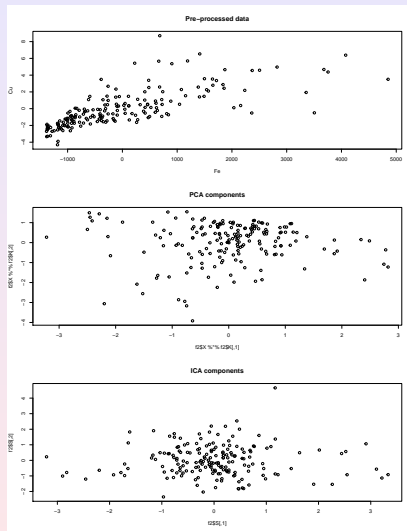


Figura: Gráficos dos dados, CP's e CI's para o Exemplo 1.

# ICA – Exemplo 1

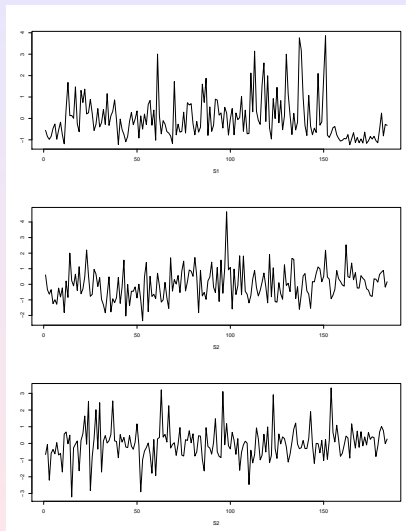


Figura: Gráficos das três componentes independentes.

# Decomposição espectral

- Seja  $\mathbf{A}$  uma matriz quadrada de ordem  $m$ . Segue-se que  $|\mathbf{A} - \lambda \mathbf{I}|$  é um polinômio de ordem  $m$  em  $\lambda$  e terá  $m$  raízes complexas  $\lambda_1, \dots, \lambda_m$ . Essas raízes são chamadas *raízes características* ou *autovalores* de  $\mathbf{A}$ .
- Como  $\mathbf{A} - \lambda_j \mathbf{I}$  é singular,  $j = 1, \dots, m$ , existe um vetor  $\mathbf{a}_j$ , cujas coordenadas não são todas nulas, tal que  $(\mathbf{A} - \lambda_j \mathbf{I})\mathbf{a}_j = \mathbf{0}$ , ou seja,  $\mathbf{A}\mathbf{a}_j = \lambda_j \mathbf{a}_j$ ,  $j = 1, \dots, m$ . Os vetores  $\mathbf{a}_1, \dots, \mathbf{a}_m$  são chamados *vetores característicos* ou *autovetores* de  $\mathbf{A}$ .
- Os seguintes resultados são válidos.
  - (1) O posto de  $\mathbf{A}$ ,  $\rho(\mathbf{A})$ , dá o número de autovalores de  $\mathbf{A}$  não nulos.
  - (2)  $\text{tr}(\mathbf{A}) = \sum_{j=1}^m \lambda_j$ .
  - (3)  $|\mathbf{A}| = \prod_{j=1}^m \lambda_j$ .
  - (4) Se  $\mathbf{A}$  é uma matriz simétrica, real, todos os seus autovalores são reais, e para cada autovalor real existe um autovetor real.
  - (5) Se  $\mathbf{A}$  é simétrica, real, os autovetores correspondentes a autovalores distintos são ortogonais.
  - (6) Se  $\mathbf{A}$  é não negativa definida, então  $\lambda_j \geq 0$ ,  $j = 1, \dots, m$ .
  - (7) Se  $\mathbf{A}$  é simétrica, de ordem  $m \times m$ , existe uma matriz ortogonal  $\mathbf{X}$ , tal que

$$\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{\Lambda} = \text{diag}\{\lambda_1, \dots, \lambda_m\},$$

ou  $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}'$ , onde os  $\lambda_j$  são os autovalores de  $\mathbf{A}$  e as colunas de  $\mathbf{X}$  são os correspondentes autovetores.

# Decomposição espectral

- O resultado (7) é chamado o *teorema espectral* para matrizes simétricas. Segue-se que a *decomposição espectral* de  $\mathbf{A}$  é dada por

$$\mathbf{A} = \sum_{j=1}^m \lambda_j \mathbf{x}_j \mathbf{x}_j',$$

onde  $\mathbf{x}_j$  é o autovetor correspondente a  $\lambda_j$ .

- Se  $\mathbf{A}$  é uma matriz quadrada de ordem  $m$ , positiva definida, existe uma matriz triangular inferior  $\mathbf{T}$ , com elementos da diagonal principal positivos, tal que

$$\mathbf{T}^{-1} \mathbf{A} (\mathbf{T}')^{-1} = \mathbf{I}_m, \quad \text{ou} \quad \mathbf{A} = \mathbf{T} \mathbf{T}'.$$

- Esta decomposição de  $\mathbf{A}$  é chamada *decomposição de Cholesky*.

# Decomposição em valores singulares - DVS

- DVS é uma fatoração de uma matriz qualquer em outras três matrizes com características importantes.
- Dada uma matriz  $\mathbf{A}$ , de ordem  $m \times n$ ,  $m \geq n$ , não necessariamente de posto máximo, uma DVS de  $\mathbf{A}$  é uma fatoração da forma  $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ , onde:
  - $\mathbf{U}$ , de ordem  $m \times n$ , é ortogonal;
  - $\mathbf{V}$ , de ordem  $m \times n$ , é ortogonal;
  - $\mathbf{\Sigma}$  é diagonal se  $m = n$ , caso contrário adicionam-se  $m - n$  linhas de zeros em  $\mathbf{\Sigma}$ .
  - Os valores  $\sigma_1, \dots, \sigma_n$  da diagonal principal de  $\mathbf{\Sigma}$  são chamados **valores singulares** de  $\mathbf{\Sigma}$  e  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n > 0$ .
  - As colunas de  $\mathbf{U}$ ,  $u_1, \dots, u_m$  são chamados **vetores singulares à esquerda** de  $\mathbf{A}$ , enquanto as colunas de  $\mathbf{V}$ ,  $v_1, \dots, v_n$  são chamados **vetores singulares à direita** de  $\mathbf{A}$ , de tal forma que  $\mathbf{A}v_j = \sigma_j u_j$ .

# Decomposição em valores singulares - DVS

- Decomposição por Autovalores  $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$ ;
  - usa a mesma base  $\mathbf{X}$  para os espaços linha e coluna;
  - geralmente não usa base ortonormal;
  - é definida somente para matrizes quadradas.
- DVS
  - usa bases diferentes  $\mathbf{V}$  ,  $\mathbf{U}$  para os espaços linha e coluna;
  - usa bases ortonormais;
  - existe para todas as matrizes.
- Entretanto, para matrizes simétricas positivas definidas, a decomposição por autovalores e a DVS são idênticas.



# Aplicações da DVS

- Cálculo de propriedades de matrizes;
- Aproximação de  $\mathbf{A}$  por matrizes de posto baixo;
- Processamento de Sinais e Imagens
  - Compressão de Imagens;
  - Eliminação de ruídos (ruídos normalmente possuem  $\sigma_j$  pequenos);
  - Recuperação de Informações.

## Curtose

- 1) Seja  $X$  uma variável aleatória qualquer, com média  $\mu$  e variância  $\sigma^2$ . Então, a *assimetria* de  $X$  é definida por

$$A(X) = E \left( \frac{(X - \mu)^3}{\sigma^3} \right), \quad (17)$$

enquanto a *curtose* de  $X$  é definida por

$$K(X) = E \left( \frac{(X - \mu)^4}{\sigma^4} \right). \quad (18)$$

- 2) Para uma distribuição normal,  $A = 0$  e  $K = 3$ , donde a quantidade  $e(X) = K(X) - 3$  ser chamada *excesso de curtose*. Distribuições com caudas pesadas têm curtose maior do que 3, e esta pode mesmo ser infinita.

## Curtose

- 3) Com uma amostra  $X_1, \dots, X_T$  de  $X$ , considere o  $r$ -ésimo momento amostral

$$m_r = \frac{1}{T} \sum_{t=1}^T (X_t - \bar{X})^r,$$

onde  $\hat{\mu} = \bar{X}$ . Substituindo os momentos verdadeiros de  $X$  pelos respectivos momentos amostrais, obtemos os estimadores

$$\hat{A}(X) = \frac{m_3}{m_2^{3/2}} = \frac{1}{T} \sum_{t=1}^T \left( \frac{X_t - \bar{X}}{\hat{\sigma}} \right)^3, \quad (19)$$

$$\hat{K}(X) = \frac{m_4}{m_2^2} = \frac{1}{T} \sum_{t=1}^T \left( \frac{X_t - \bar{X}}{\hat{\sigma}} \right)^4, \quad (20)$$

respectivamente, onde  $\hat{\sigma}^2 = \sum_{t=1}^T (X_t - \bar{X})^2 / T$ . Segue-se que  $\hat{e}(X) = \hat{K}(X) - 3$ .

- 4) Pode-se provar que, se tivermos uma amostra de uma distribuição normal, e  $T$  for grande, então

$$\hat{A} \sim \mathcal{N}(0, 6/T), \quad \hat{K} \sim \mathcal{N}(3, 24/T). \quad (21)$$

Esses fatos podem ser utilizados para testar a normalidade de uma série (Jarque-Bera).

## Referências

Amari, S., Cichocki, A. and Yang, H. H. (1996). A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, **8**, 757–763.

Gegembauer, H. V. (2010). *Análise de Componentes Independentes com Aplicações em Séries Temporais Financeiras*. Dissertação de Mestrado, IME-USP.

Hyvärinen, A. and Oja, E. (1997). A fast fixed point algorithm for independent component analysis. *Neural Computation*, **9**, 1483–1492.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithm for independent component analysis. *IEEE Transactions on Neural Network*, **10**, 626–634.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Alorithms and applications. *Neural Networks*, **13**, 411–430.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001). *Independent Component Analysis*. New York: Wiley.