Aprendizagem estatística em altas dimensões

Florencia Leonardi

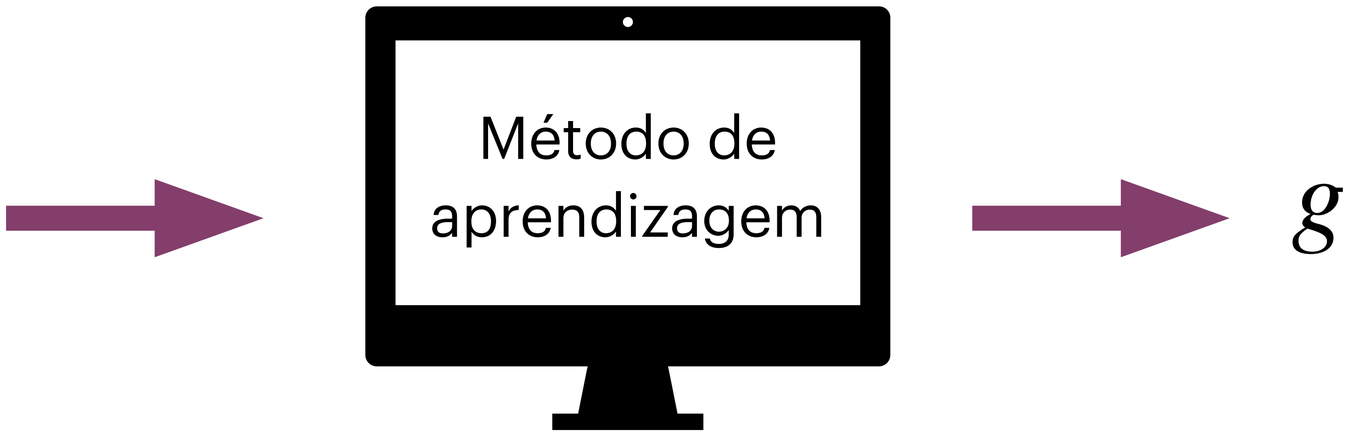
Conteúdo

- Aprendizagem estatística não supervisionada
- Análise de componentes principais
- Métodos de agrupamento - K-médias
- Modelos gráficos discretos e contínuos

Aprendizagem estatística não supervisionada

Aprendizagem

$$\mathscr{D} = \{x_1, \ldots, x_n\}$$



$$g$$

Predição

Novos dados: x → $g(x)$

Análise de componentes principais

- A análise de componentes principais pode ser visto como um método não supervisionado, já que ele não utiliza nenhuma informação de variáveis resposta associadas com as variáveis preditoras
- Ele também é um método de redução de dimensão, e pode ser utilizado como pre-processamento dos dados antes da utilização de métodos de aprendizagem supervisionada
- O método consiste em *projetar* os vetores de variáveis preditoras num sub-espaço de dimensão menor, de forma a maximizar a variância em cada nova coordenada

Análise de componentes principais

$$\mathscr{D} = \{x_1, \ldots, x_n\}$$

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}, \qquad i = 1, \ldots, n$$

$$\underbrace{\begin{pmatrix} x_{12} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{pmatrix}}_{\mathbf{X} \in \mathbb{R}^{n \times p}} \underbrace{\begin{pmatrix} \phi_{11} \\ \phi_{21} \\ \vdots \\ \phi_{p1} \end{pmatrix}}_{\phi \in \mathbb{R}^p} = \underbrace{\begin{pmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{n1} \end{pmatrix}}_{z \in \mathbb{R}^n}$$

Com esta transformação, *resumimos* a informação das p variáveis preditoras a uma única variável $z \in \mathbb{R}$ por observação

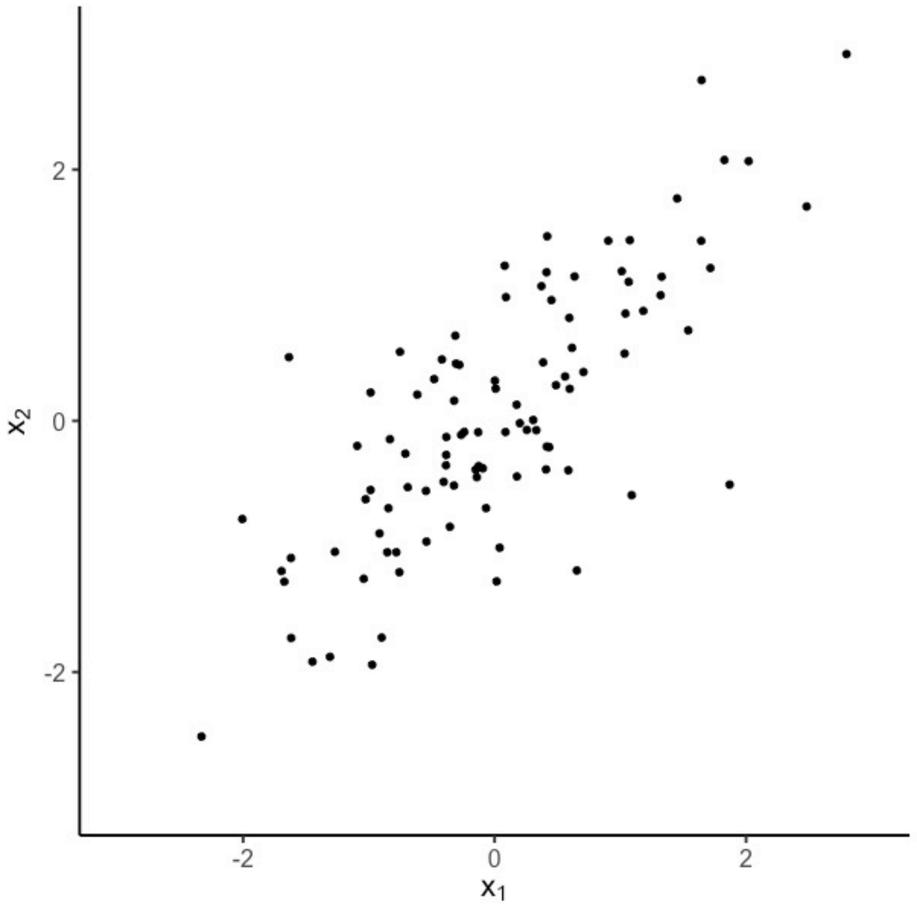
Análise de componentes principais

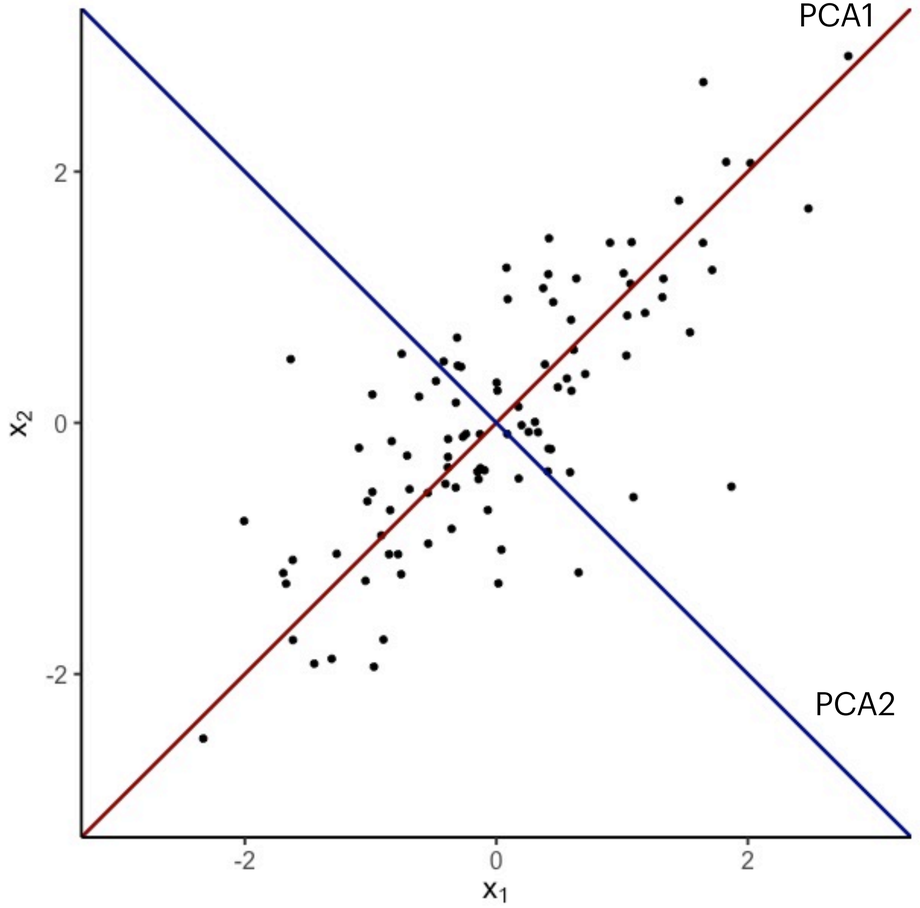
Qual seria o melhor vetor ϕ para fazer essa transformação?

Procuramos um vetor ϕ que tenha norma 1 e tal que o vetor transformado z tenha a maior variância possível, isto é procuramos

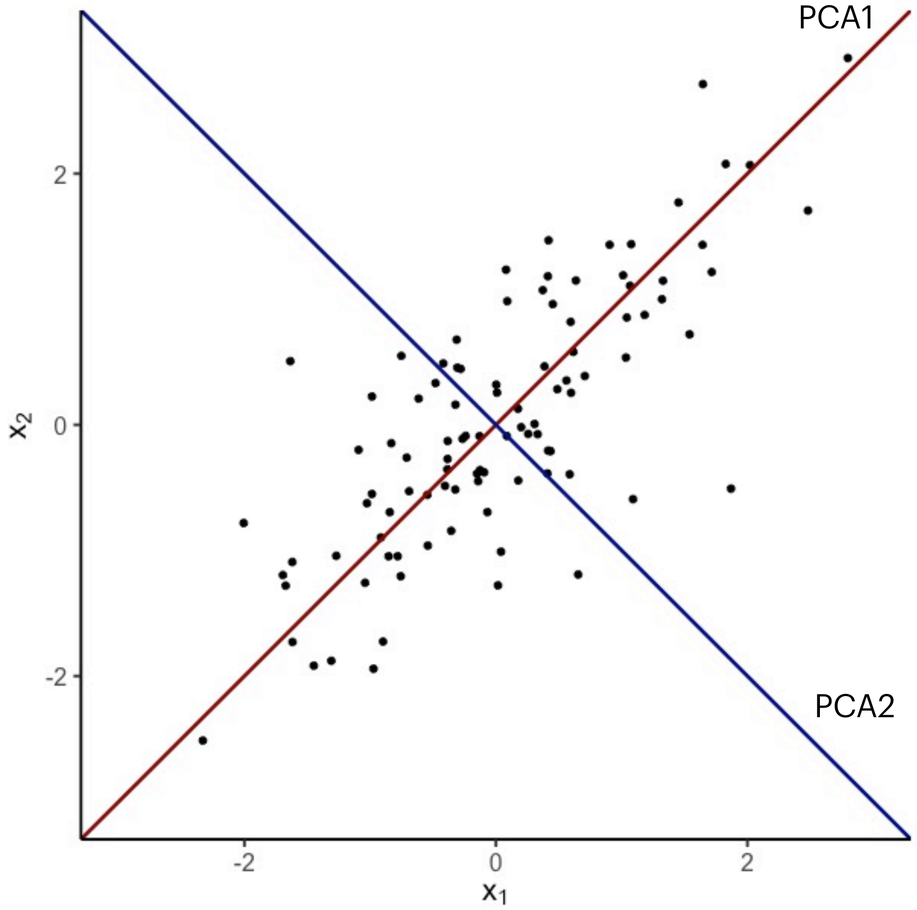
$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{Maximizar}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \quad \text{sujeito a} \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

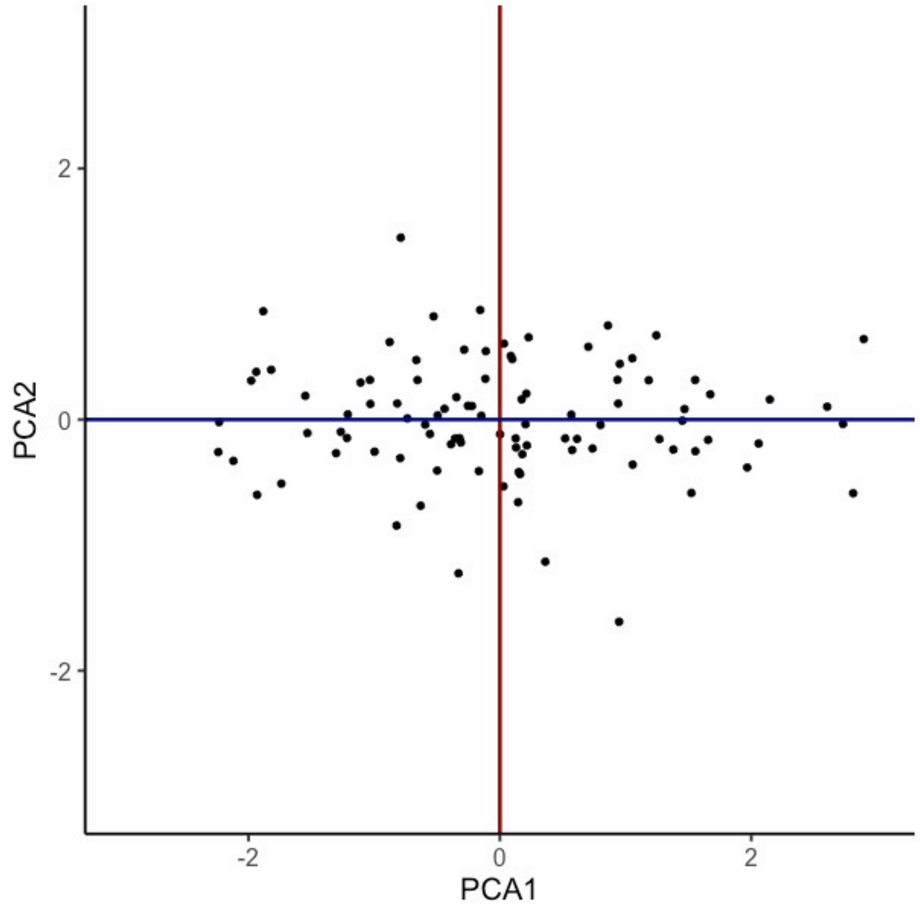
Análise de componentes principais



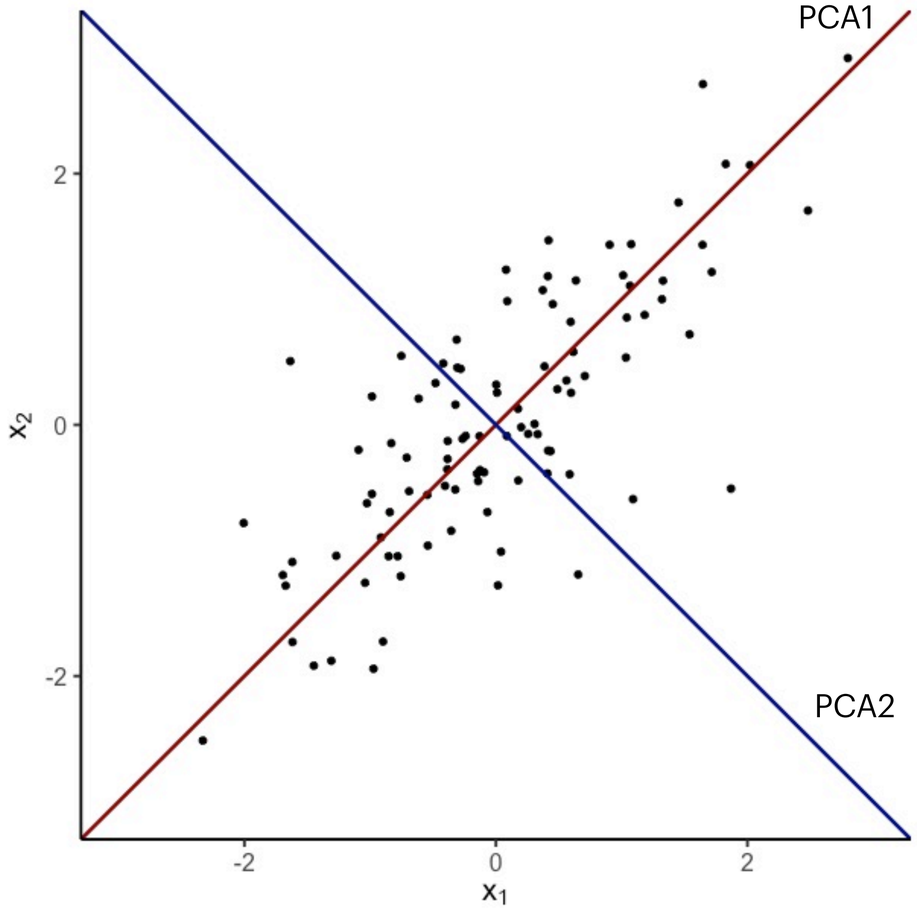


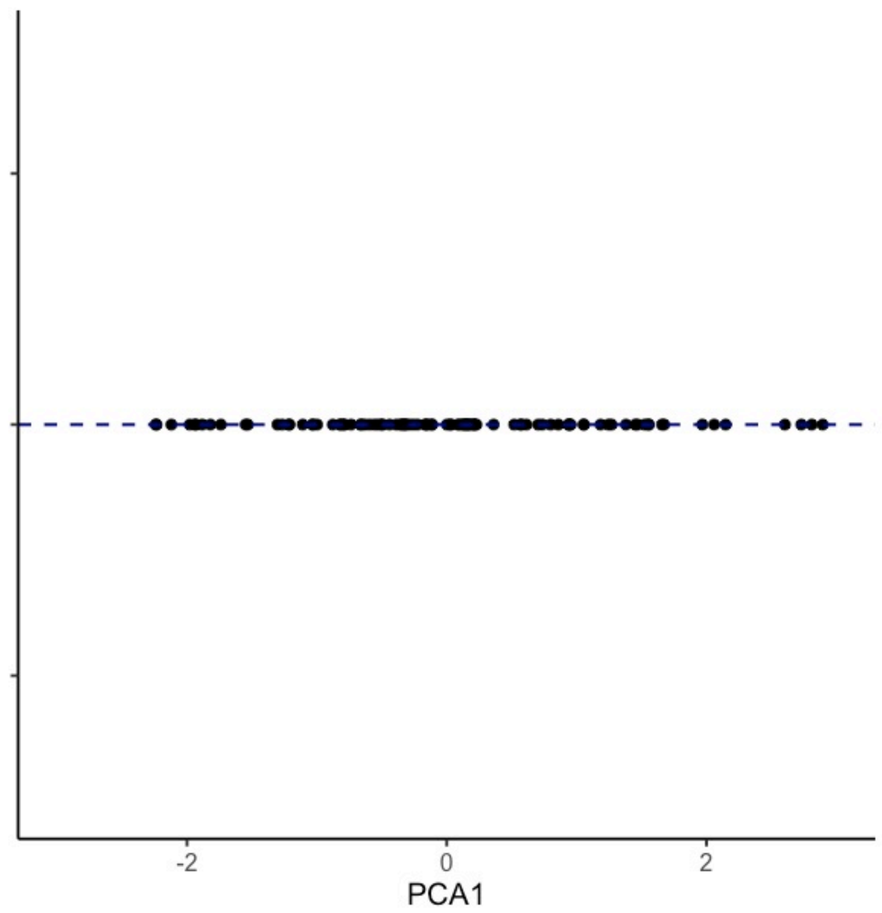
Análise de componentes principais



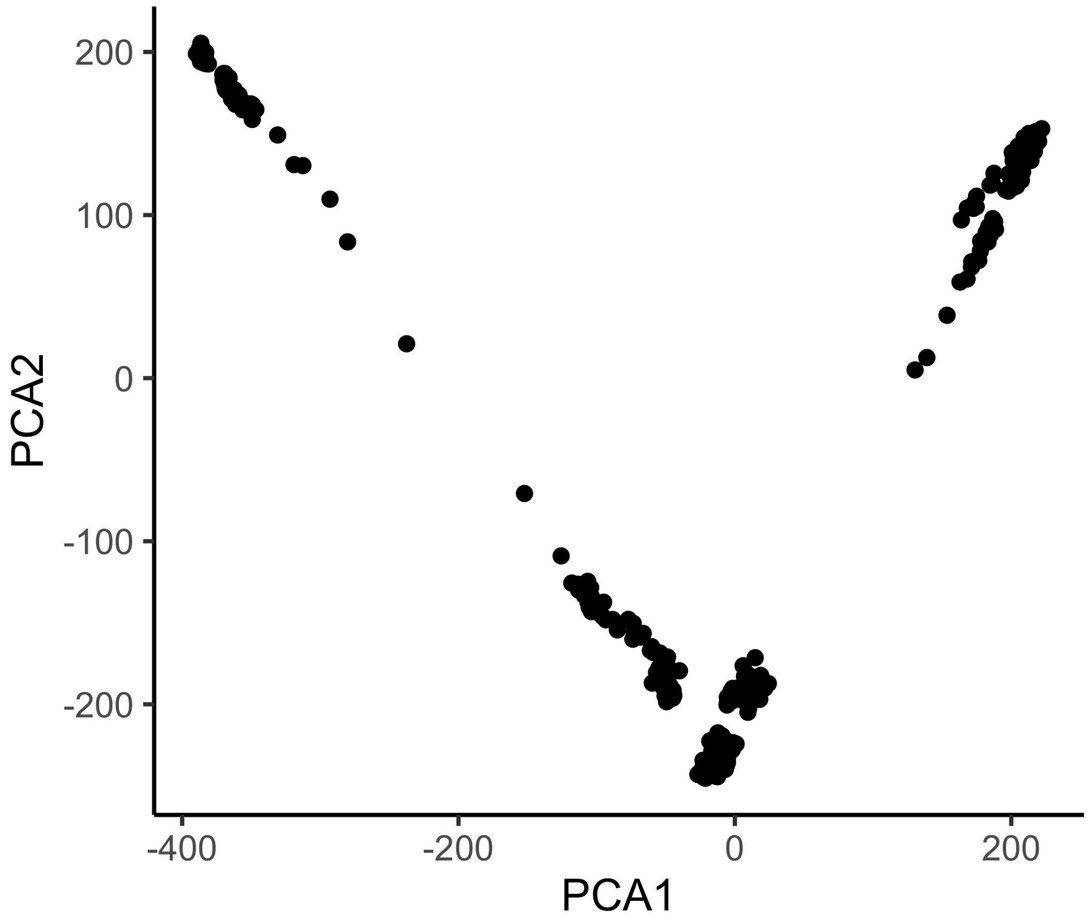


Análise de componentes principais





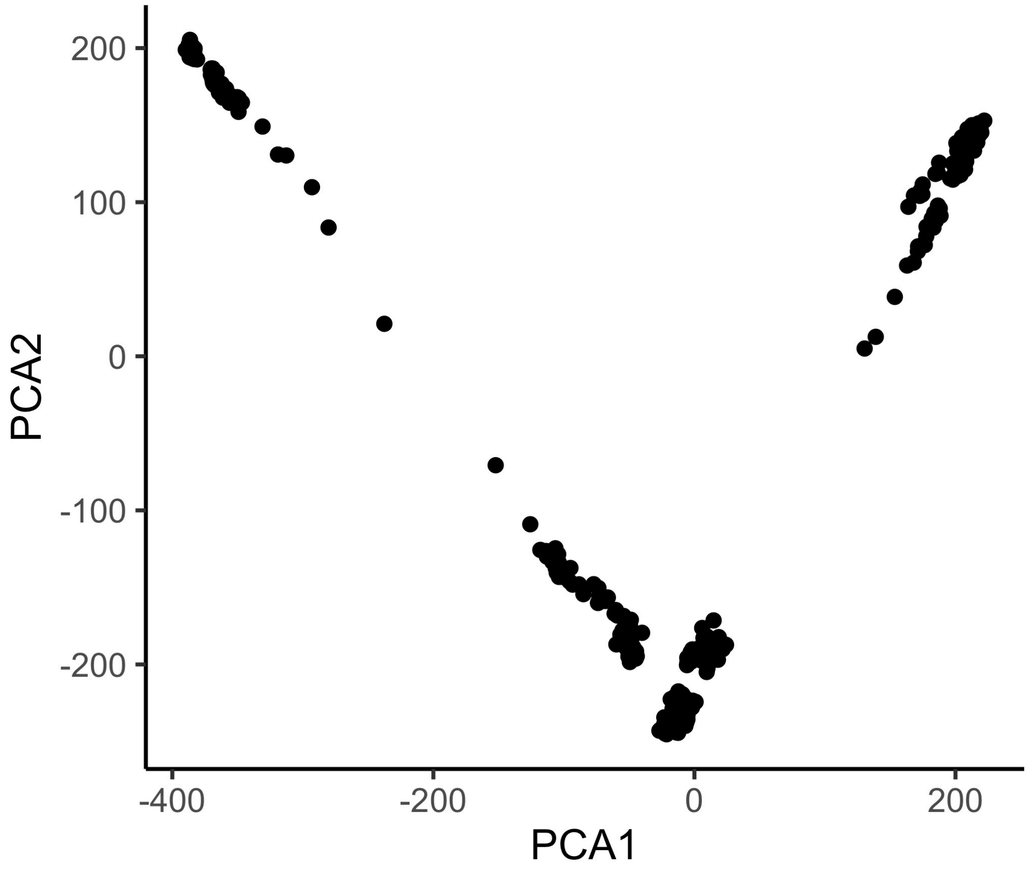
Exemplo: dados de SNPs

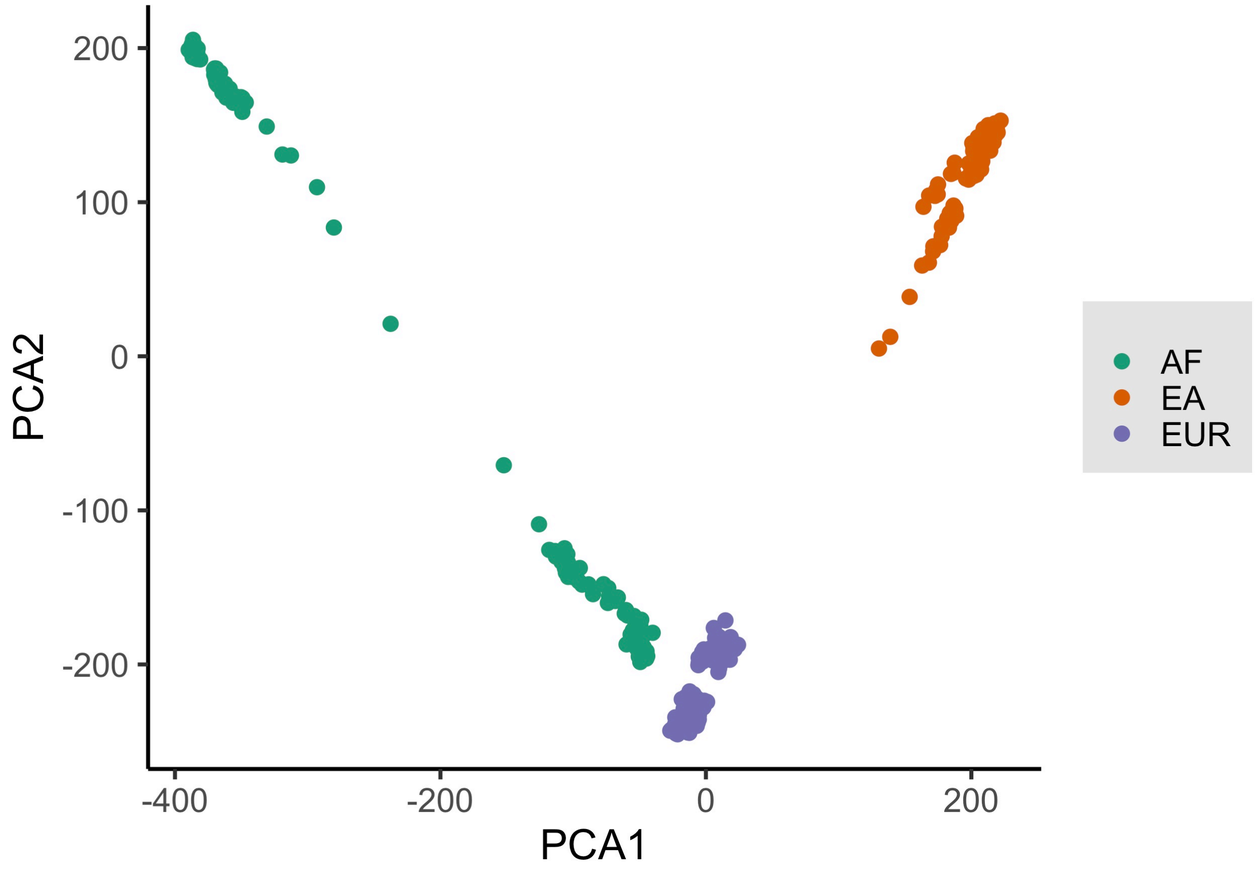


Exemplo de redução da dimensão por PCA para dados de SNPs (single nucleotide polymorphisms) de 551 indivíduos pertencentes a três populações (Africanos, Asiáticos do leste e Europeus)

Os dados consistem de 529.631 SNPs com valor no conjunto $\{0,1,2\}$ com a contagem do número de alelos raros em cada posição do genoma

Exemplo: dados de SNPs





Métodos de agrupamento (*clustering*)

- Agrupamento se refere a um conjunto amplo de técnicas para encontrar grupos num conjunto de dados
- Quando agrupamos as observações de um conjunto de dados, procuramos particioná-las em grupos distintos, de modo que as observações dentro de cada grupo sejam bastante semelhantes entre si, enquanto as observações em grupos diferentes são bastante diferentes umas das outras

Método de agrupamento *K*-médias

O método K-médias é um método bastante simples para particionar um conjunto de dados em K grupos diferentes e sem superposição.

Denotemos por $C_1, \ldots, C_K$ os conjuntos de índices das observações em cada grupo. Para ser um agrupamento, estes conjuntos devem satisfazer:

1. $C_1 \cup C_2 \cup \ldots C_K = \{1,\ldots,n\}$. Em outras palavras, cada observação deve pertencer ao menos a um grupo

2. $C_k \cap C_{k'} = \emptyset$ para todo $k \neq k'$. Em outras palavras, os grupos não têm interseção, nenhuma observação pertence a mais de um grupo

Método de agrupamento *K*-médias

A ideia por trás do método K-médias é minimizar a variabilidade dentro de cada grupo tanto quanto possível

A variabilidade dentro de cada grupo C_k é uma medida $W(C_k)$ definida de acordo com o problema

Então, o problema consiste em

$$\underset{C_1,\ldots,C_K}{\text{Minimizar}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Método de agrupamento *K*-médias

Existem várias formas diferentes de definir a medida $W(C_k)$, mas a mais comum

envolve a distância euclidiana dada por

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 ,$$

onde $|C_k|$ denota o número de observações no $k-$éssimo grupo. Logo o problema de otimização para

definir o agrupamento pelo método de K-médias é

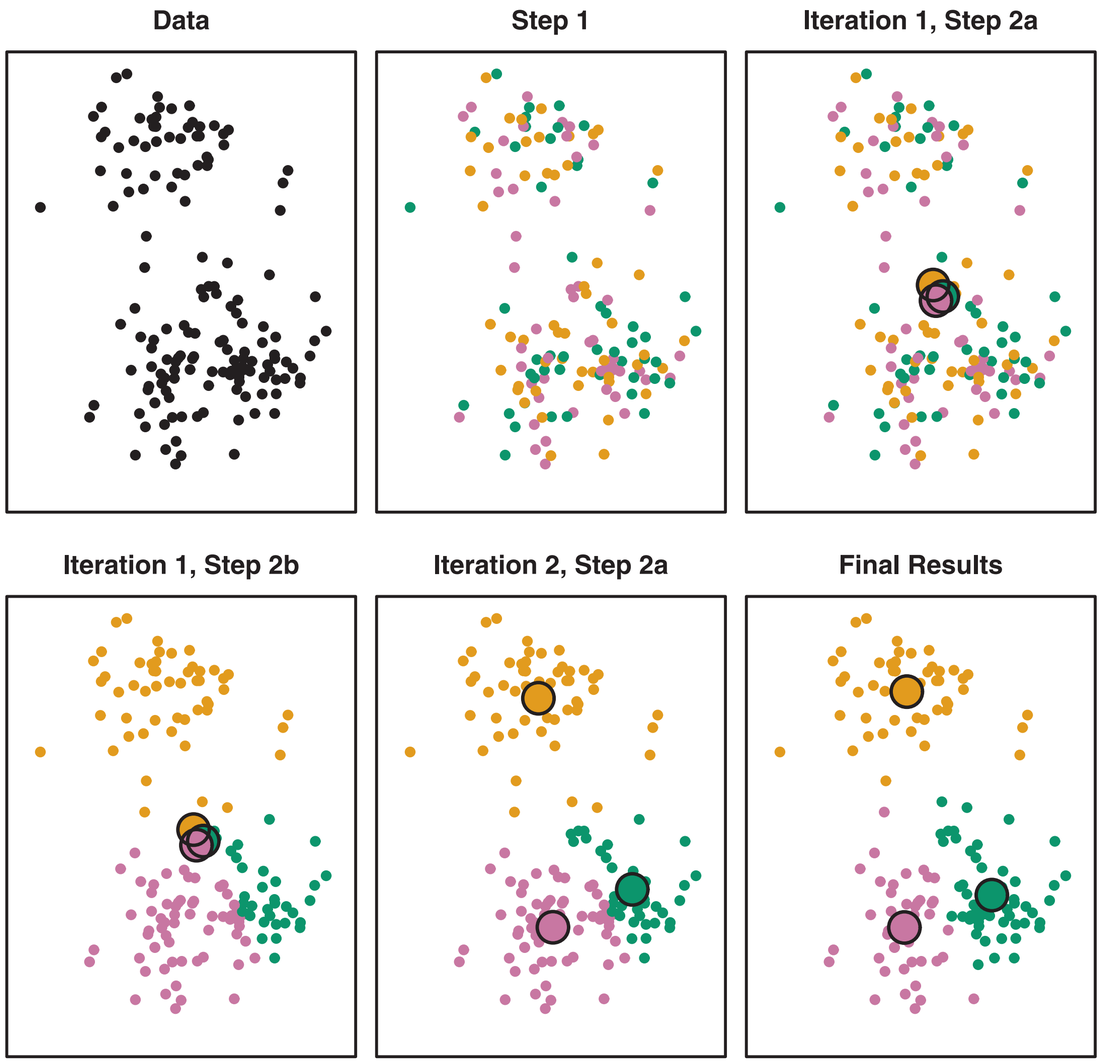
$$\underset{C_1,\ldots,C_K}{\text{Minimizar}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

Algoritmo: agrupamento por K-médias

1. Atribua um número, de 1 a K, aleatoriamente a cada uma das observações. Estes servem como inicialização dos grupos.
2. Itere os seguintes passos até que os grupos deixem de mudar:
    a. Para cada um dos K grupos, calcule o *centroide*. O *centroide* de cada grupo é o vetor médio das observações em cada grupo.

    b. Atribua cada observação ao grupo cujo *centroide* está mais próximo (onde a proximidade é definida em relação à distância euclidiana)

Método de agrupamento *K*-médias



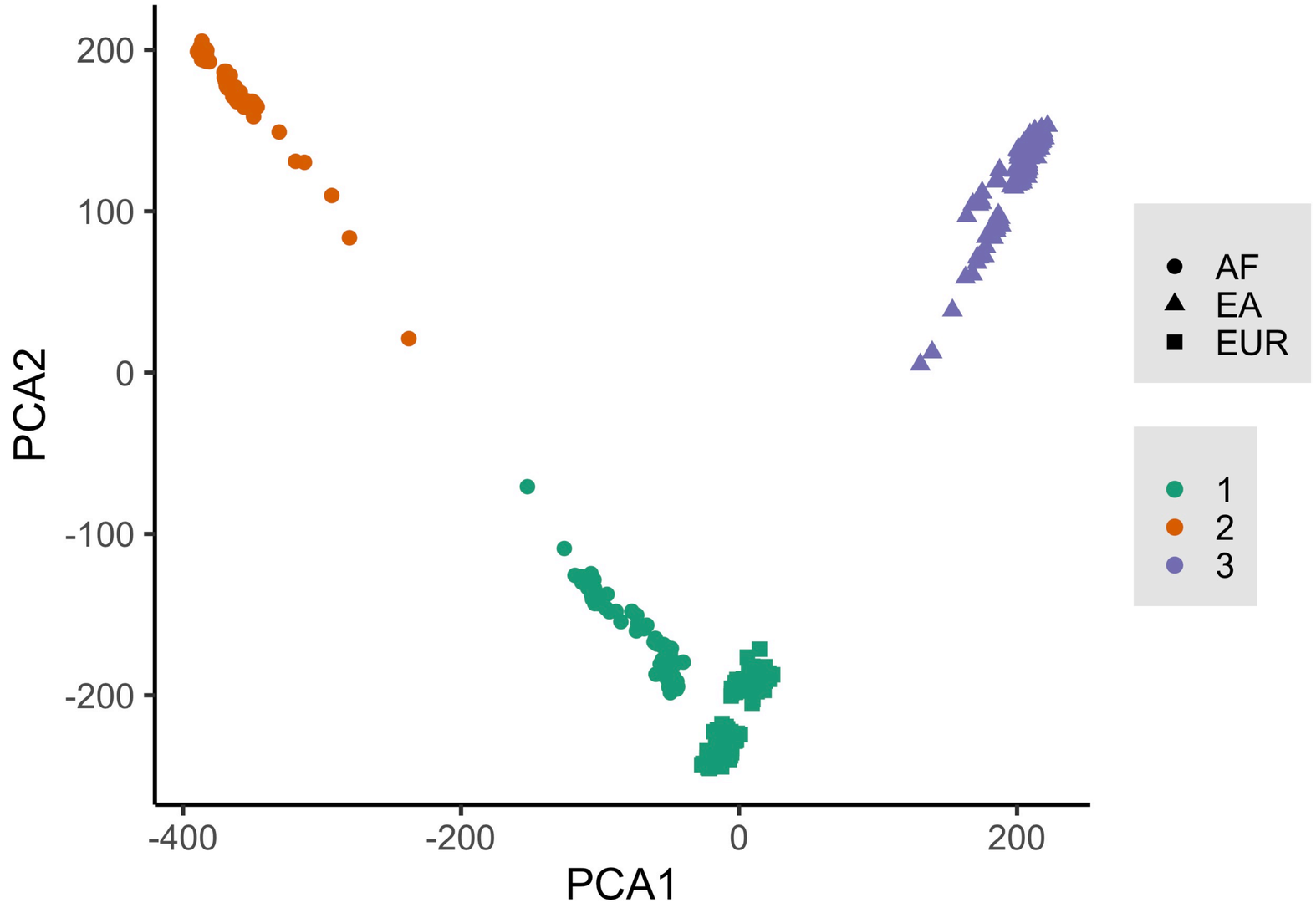
O algoritmo de K-médias encontra um mínimo local em vez do mínimo global. Portanto os resultados dependem do agrupamento inicial (aleatório)

Por este motivo, é importante rodar o algoritmo várias vezes a partir de diferentes agrupamentos.

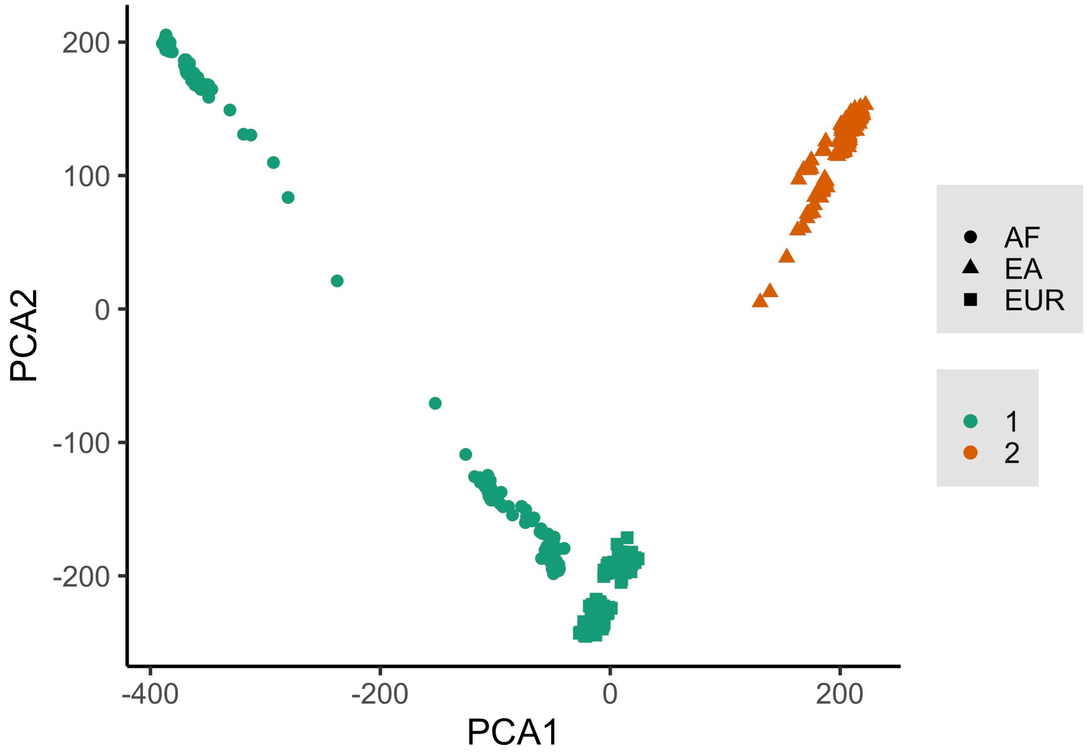
No final é escolhido o agrupamento ótimo, aquele que minimiza a função objetivo.

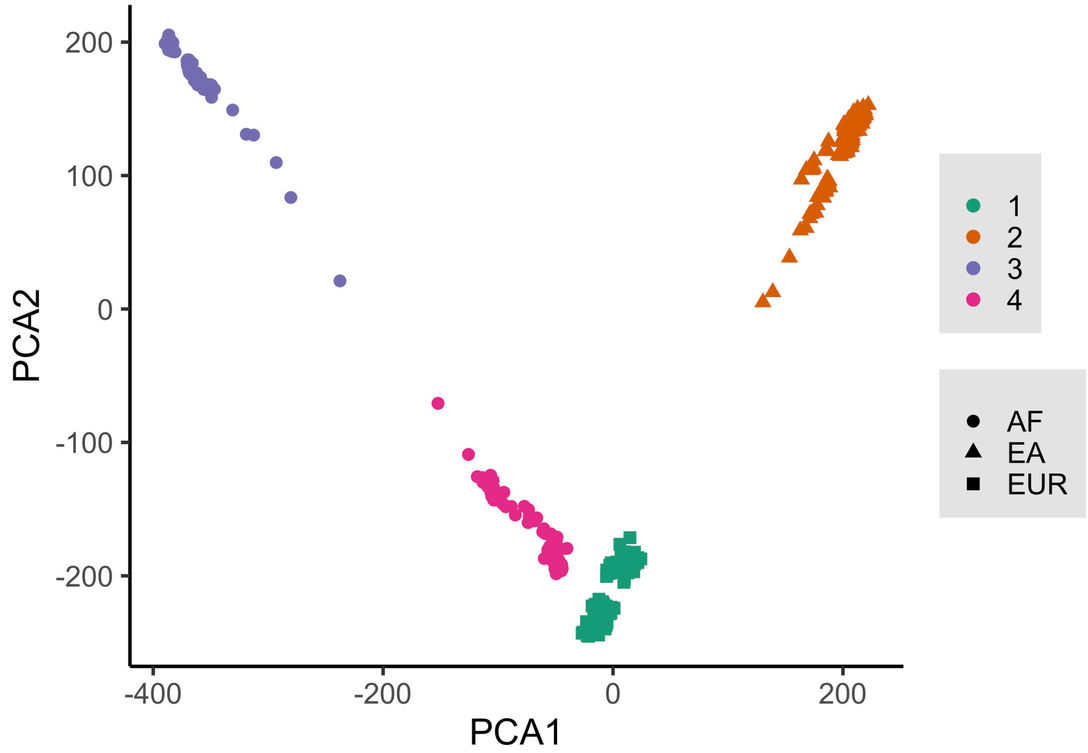
Figura do livro *An Introduction to Statistical Learning*

Exemplo: dados de SNPs



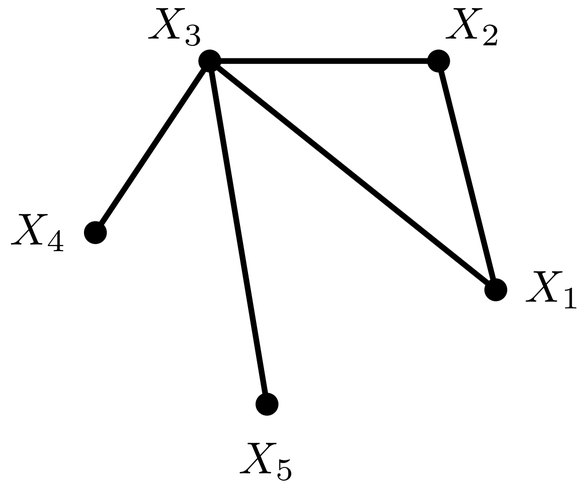
Exemplo: dados de SNPs





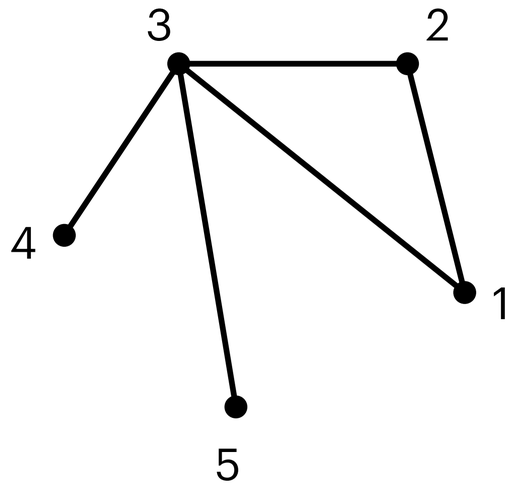
Modelos gráficos

- Os modelos gráficos são modelos de funções de distribuição conjunta de variáveis aleatórias, com certas relações de dependência condicional que são codificadas em *grafos*
- Um grafo consiste num conjunto de vértices e num conjunto de arestas (pares ordenados de vértices)
- Num modelo gráfico, cada vértice representa uma variável aleatória e as arestas dão uma representação visual para entender a distribuição conjunta das variáveis
- Aqui consideraremos modelos com grafos não direcionados, que também são conhecidos como *campos aleatórios Markovianos*



Modelos gráficos

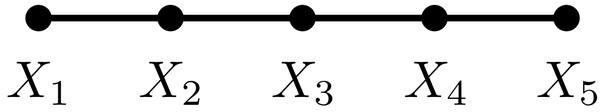
Formalmente, um grafo é um par ordenado $G = (V, E)$ onde V é um conjunto de vértices e E é um conjunto de arestas (definido como pares de vértices)

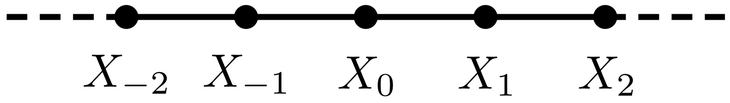


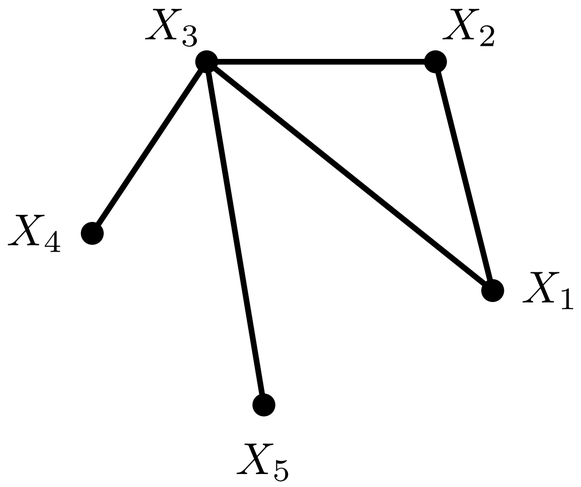
$$V = \{1, 2, 3, 4, 5\}$$

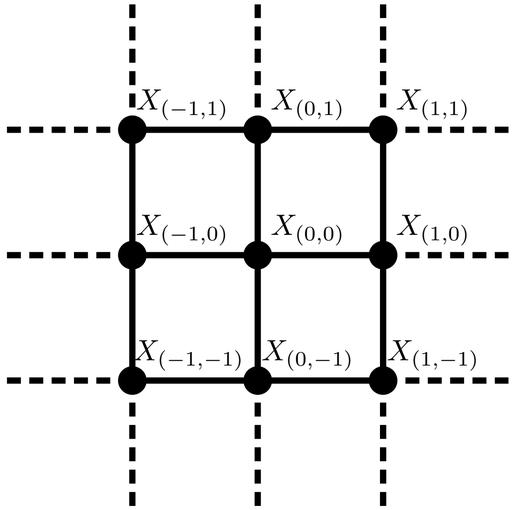
$$E = \{(1,2), (1,3), (2,3), (3,4), (3,5)\}$$

Modelos gráficos





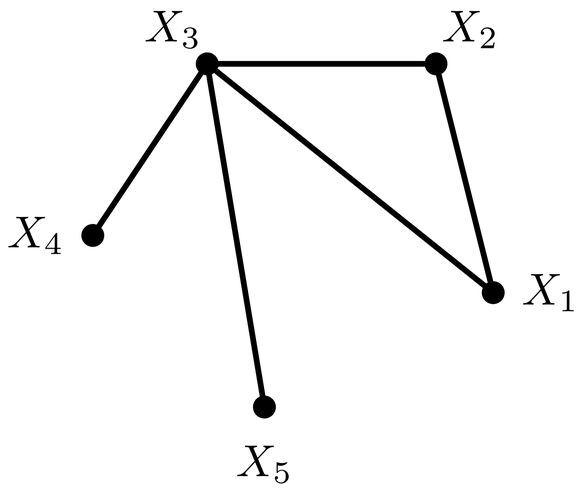




Nestes modelos, a ausência de uma aresta entre duas variáveis significa que as variáveis são condicionalmente independentes, dadas todas as outras variáveis

Modelos gráficos

A ausência de uma aresta significa que as variáveis aleatórias correspondentes são condicionalmente independentes dadas as outras variáveis



Neste exemplo temos que

$$X_1 \perp X_5 \mid X_2, X_3, X_4$$

Esta propriedade é conhecida como *propriedade de Markov por pares*

Modelos gráficos discretos

Suponhamos que observamos uma amostra i.i.d de tamanho n da distribuição

conjunta das variáveis aleatórias

X_1	X_2	X_3	X_4	X_5
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
		$\vdots$		
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}

Consideremos primeiramente o caso de variáveis aleatórias discretas

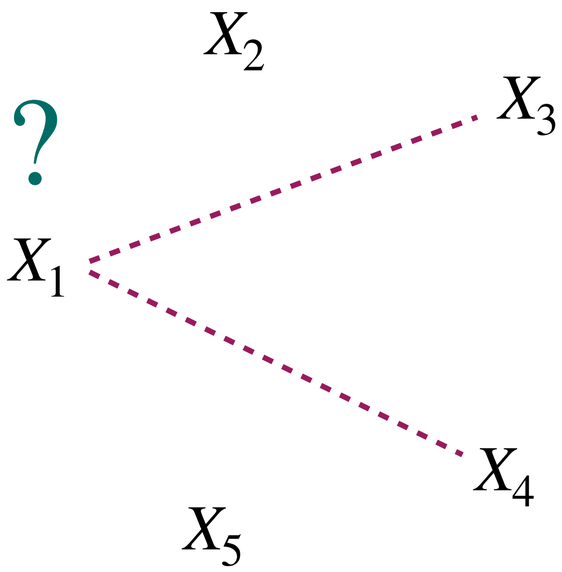
Modelos gráficos discretos

Neste caso, podemos estimar a vizinhança de um nó $v \in V$ a partir de um critério de máxima verossimilhança penalizada

$$\widehat{\mathsf{ne}}(v) = \underset{W \subset V \setminus \{v\}}{\arg\max} \left\{ \log \widehat{\mathbb{P}}(x_v^{(1:n)} \,|\, x_W^{(1:n)}) - c\,|A|^{|W|} \log n \right\}$$

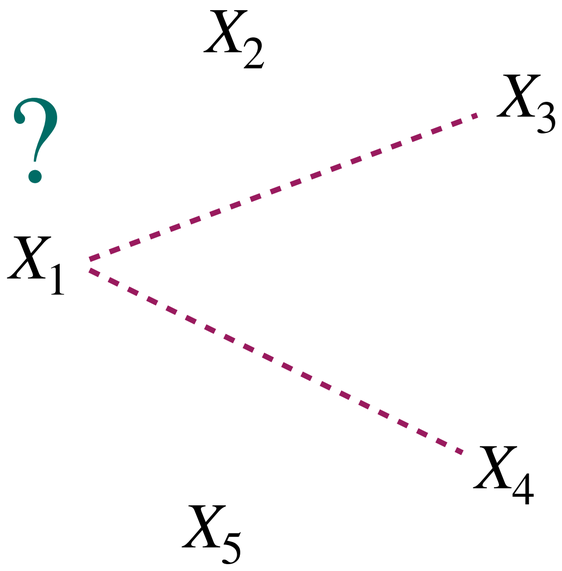
onde $\widehat{\mathbb{P}}(x_v^{(1:n)} \,|\, x_W^{(1:n)}) = \prod_{a_W \in A^W} \prod_{a_v \in A} \hat{p}(a_v \,|\, a_W)^{N(a_v, a_W)}$, com $N(a_v, a_W)$ o contador do número de ocorrências da configuração (a_v, a_W) na amostra e $\hat{p}(a_v \,|\, a_W)$ a probabilidade condicional estimada

Modelos gráficos discretos



X_1	X_2	X_3	X_4	X_5
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
		⋮		
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}

Modelos gráficos discretos

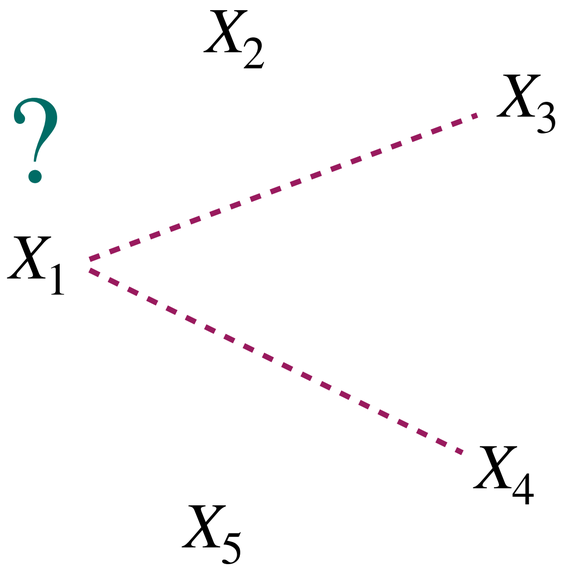


X_1	X_2	X_3	X_4	X_5
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
		⋮		
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}

$$\widehat{\mathbb{P}}(x_v^{(1:n)} \mid x_W^{(1:n)}) = \prod_{a_W \in A^W} \prod_{a_v \in A} \hat{p}(a_v \mid a_W)^{N(a_v, a_W)}$$

$$v = 1\,, \quad W = \{3,4\}$$

Modelos gráficos discretos



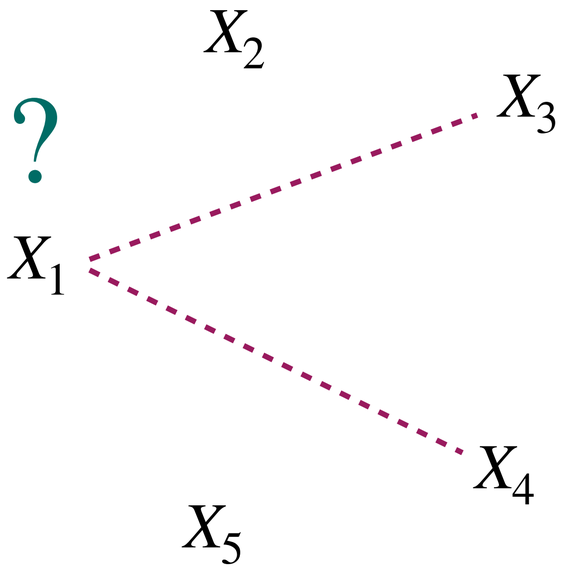
X_1	X_2	X_3	X_4	X_5
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
		⋮		
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}

$$\widehat{\mathbb{P}}(x_v^{(1:n)} \mid x_W^{(1:n)}) = \prod_{a_W \in A^W} \prod_{a_v \in A} \hat{p}(a_v \mid a_W)^{N(a_v, a_W)}$$

$$v = 1\,, \quad W = \{3,4\} \qquad a_v = 0\,, \quad a_W = \{0,1\}$$

$$N(a_v, a_W) = \sum_{i=1}^{n} \mathbf{1}\{x_{i1} = 0, x_{i3} = 0, x_{i4} = 1\}$$

Modelos gráficos discretos



X_1	X_2	X_3	X_4	X_5
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
		⋮		
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}

$$\widehat{\mathbb{P}}(x_v^{(1:n)} \mid x_W^{(1:n)}) = \prod_{a_W \in A^W} \prod_{a_v \in A} \hat{p}(a_v \mid a_W)^{N(a_v, a_W)}$$

$$v = 1\,, \quad W = \{3,4\} \qquad a_v = 0\,, \quad a_W = \{0,1\}$$

$$\hat{p}(0_1 \mid 0_3, 1_4) = \frac{N(0_1, 0_3, 1_4)}{N(0_3, 1_4)}$$

Modelos gráficos discretos

Uma vez estimada a vizinhança de cada nó, podemos estimar o grafo a partir da estimação do conjunto de arestas

$$\widehat{E} = \{(v, w) \in V \times V : v \in \widehat{\text{ne}}(w) \text{ ou } w \in \widehat{\text{ne}}(v)\}$$

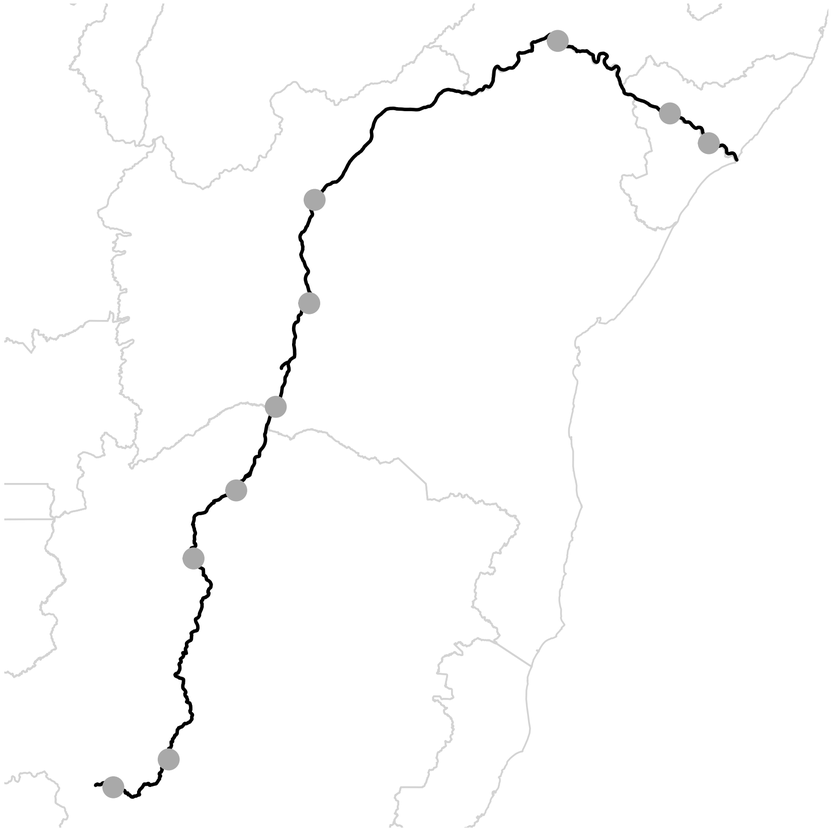
Pode-se demonstrar que $\widehat{\text{ne}}(v)$ converge para $\text{ne}(v)$ quando n vai para infinito, com probabilidade 1, e portanto também $\widehat{E}$ converge para E quando n cresce para infinito, ou seja o estimador do grafo é *consistente*

Exemplo: modelagem de índices de ações

- Para ilustrar a aplicação do estimador do grafo consideramos índices de ações correspondentes a 15 países (dados do site https://br.investing.com/indices/world-indices)
- A amostra consiste de $n = 530$ observações no tempo, onde cada variável corresponde à função indicadora de haver uma mudança positiva a partir do dia anterior, para cada um dos índices considerados
- Para reduzir a dependência, consideramos um intervalo de 4 dias entre as observações

Exemplo: modelagem de índices de ações

Exemplo: modelagem do fluxo no Rio São Francisco



Exemplo: modelagem do fluxo no Rio São Francisco

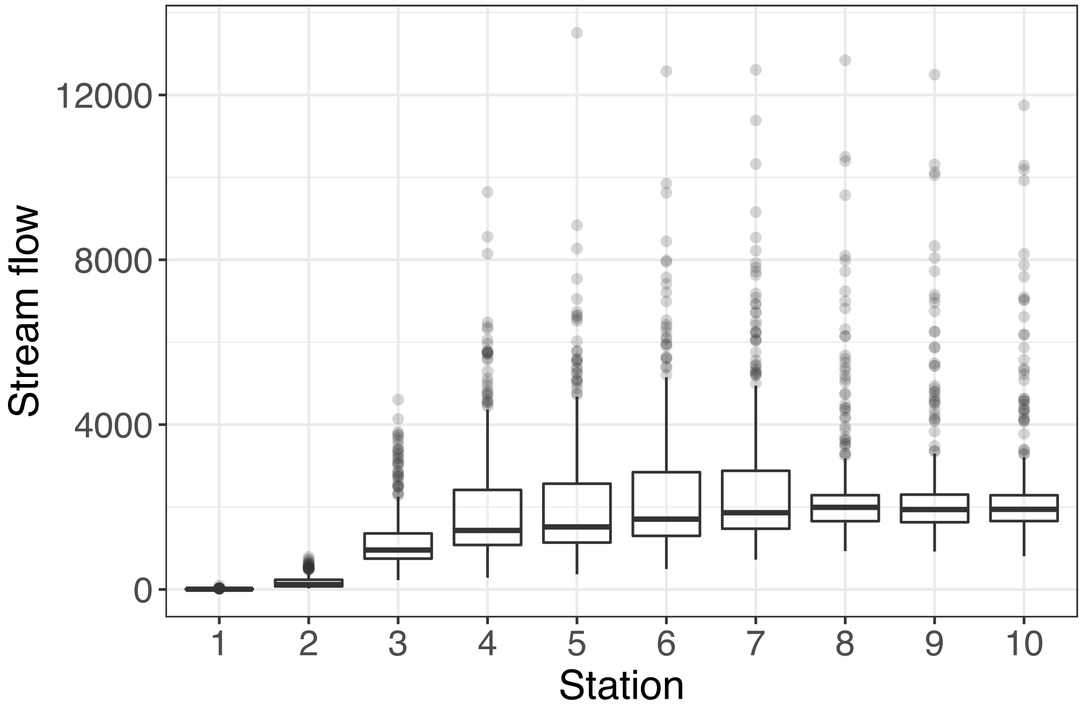
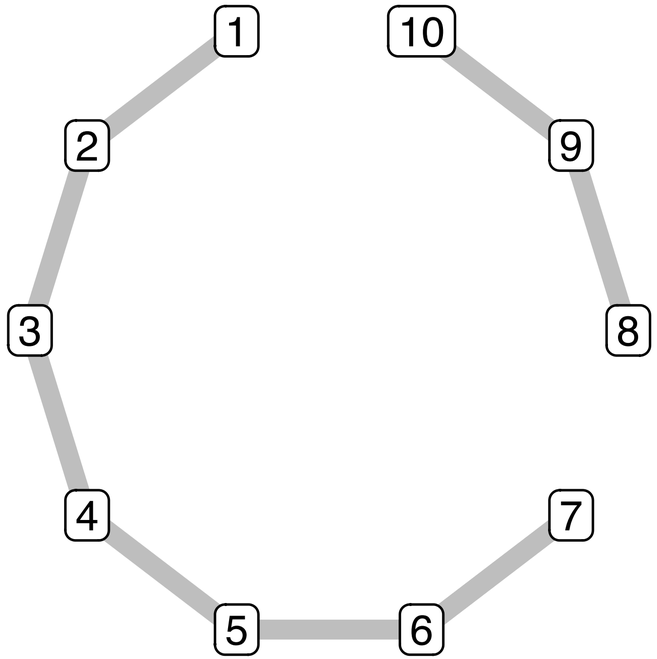


Figura do artigo Leonardi et al. (2021). *Independent block identification in multivariate time series*, Journal of Time Series Analysis, 42(1), 2021.



Grafo estimado

Modelos gráficos contínuos

- No caso contínuo, em geral é considerada uma distribuição Gaussiana multivariada com vetor de médias μ e matriz de covariância Σ
- Esta distribuição tem a propriedade que todas as distribuições marginais e condicionais também são Gaussianas
- A inversa da matriz de covariância Σ^{-1} contem informação sobre as covariâncias entre os vértices i e j, condicionalmente aos outros vértices. Observemos que no caso Gaussiano, a independência e a falta de correlação entre variáveis são propriedades equivalentes

Modelos gráficos contínuos

Em particular, se a entrada ij da matriz $\Theta = \Sigma^{-1}$ é igual a zero, então as variáveis i e j são condicionalmente independentes, dadas todas as outras variáveis.

Para estimar o modelo gráfico, utilizamos também um critério de máxima verossimilhança penalizada.

Neste caso, a função de verossimilhança está dada por

$$p_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{p/2}\det[\Sigma]^{1/2}} e^{\frac{1}{2}(x-\mu)^T\Sigma(x-\mu)}$$

Observemos que esta função não depende diretamente de Θ, que é a matriz que tem a informação que nos interessa

Modelos gráficos contínuos

Então usamos uma reparametrização da função densidade de probabilidade da Gaussiana multivariada dada por

$$p_{\gamma,\Theta}(x) = \exp\left\{ \sum_{s=1}^{p} \gamma_s x_s - \frac{1}{2} \sum_{s,t=1}^{p} \theta_{st} x_s x_t - A(\Theta) \right\}$$

onde $A(\Theta) = \frac{1}{2} \log \det[\Theta/(2\pi)]$, $\gamma \in \mathbb{R}^p$.

A matriz $\Theta = \Sigma^{-1}$ é chamada de matriz de precisão ou matriz de concentração.

Modelos gráficos contínuos

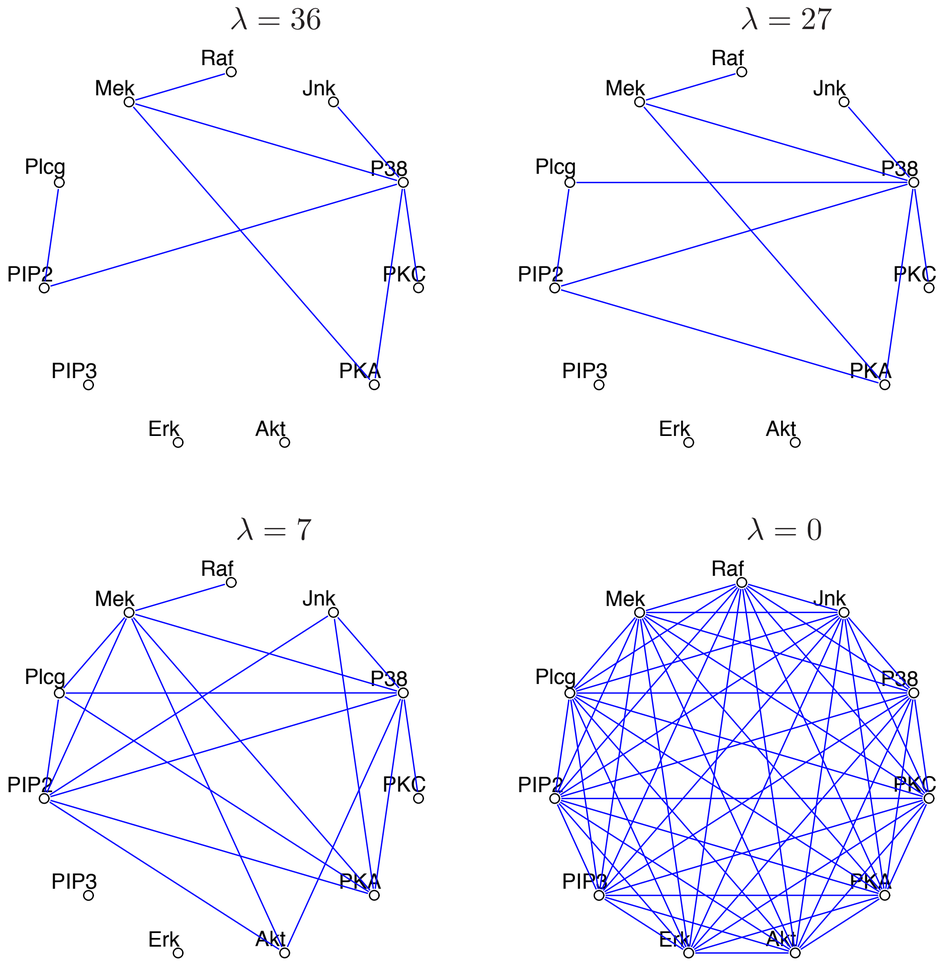
Para estimar a estrutura do grafo, é usual utilizar uma penalidade do tipo LASSO que leve algumas entradas da matriz Θ para zero (o que implica não ter aresta no grafo).

Este critério está dado por

$$\widehat{\Theta} = \arg\max_{\Theta \geq 0} \left\{ \log\det\Theta - \text{traço}(\mathbf{S}\Theta) - \lambda\rho_1(\Theta) \right\}$$

onde $\rho_1(\Theta) = \sum_{s \neq t} |\theta_{st}|$ e $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$ é a matriz de covariância empírica. A traço de uma matriz é definido como a soma dos elementos da diagonal da matriz.

Modelos gráficos contínuos



Exemplo de grafos não dirigidos estimados de um conjunto de dados de citometria de fluxo, para $p = 11$ proteínas medida em $n = 7466$ células. As estruturas dos grafos foram estimadas com diferentes valores da constante de penalização

Figura do livro The Elements of Statistical Learning