Universidade Federal do Paraná

Laboratório de Estatística e Geoinformação - LEG

Aprendizado não supervisionado

Prof.: Eduardo Vargas Ferreira

Aprendizado supervisionado versus não supervisionado

- Na maior parte deste curso tratamos de **aprendizado supervisionado**, p. ex., regressão e classificação;
- Nestes casos, observamos os preditores $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$, bem como a variável resposta $\mathbf{Y}$;
- E o objetivo era prever uma nova observação Y utilizando $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$;
- Em **aprendizado não supervisionado**, não temos acesso às respostas Y (apenas aos $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$);
- É uma abordagem mais abstrata, e o objetivo é descobrir características interessantes sobre as medidas (por exemplo, subgrupos entre as variáveis, padrões etc).

Métodos de aprendizagem não supervisionada

- Vamos discutir dois métodos:
  - ⋆ **Análise de componentes principais**: o objetivo é explicar a estrutura de covariância através de combinações lineares das variáveis $X_1, X_2, \ldots, X_p$;
  - ⋆ **Clustering**: se trata de uma ampla classe de métodos para descobrir agrupamentos nos dados.
- Exemplos:
  - ⋆ Pacientes com câncer agrupados de acordo com similaridades nas expressões gênicas;
  - ⋆ Grupos de consumidores caracterizados pelos seus históricos de navegação e compras;
  - ⋆ Filmes agrupados pelas notas dos espectadores.

Análise de componentes principais (ACP)

- Suponha que temos $X_1, X_2, \ldots, X_p$. Então, temos p variáveis para reproduzir a variabilidade geral do sistema;
- Porém, é possível que parte dessa variabilidade seja explicada por um número mínimo $k < p$ de variáveis;
- **Componentes principais** são as sequências de combinações lineares de $X_1, X_2, \ldots, X_p$ que maximizam a sua variância;
- Genericamente, representa um novo sistema de coordenadas, rotacionando o original com $X_1, X_2, \ldots, X_p$ como coordenadas;
- E os novos eixos representam as direções de **máxima variabilidade**, com uma estrutura mais simples (parcimoniosa) para descrever a estrutura de covariância;

Análise de componentes principais (ACP)

- A **primeira componente principal** de um conjunto de características $X_1, X_2, \ldots, X_p$ é a combinação linear normalizada ($\sum_{j=1}^{p} \phi_{j1}^2 = 1$)

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \ldots + \phi_{p1} X_p,$$

em que $\phi_1 = [\phi_{11}, \ldots, \phi_{p1}]^t$ são as cargas da 1ª c.p.;

- Elas maximizam a $Var(Z_1) = \phi_1^t \Sigma \phi_1$, em que Σ é a matriz de covariância de $\mathbf{X} = [X_1, \ldots, X_p]^t$;

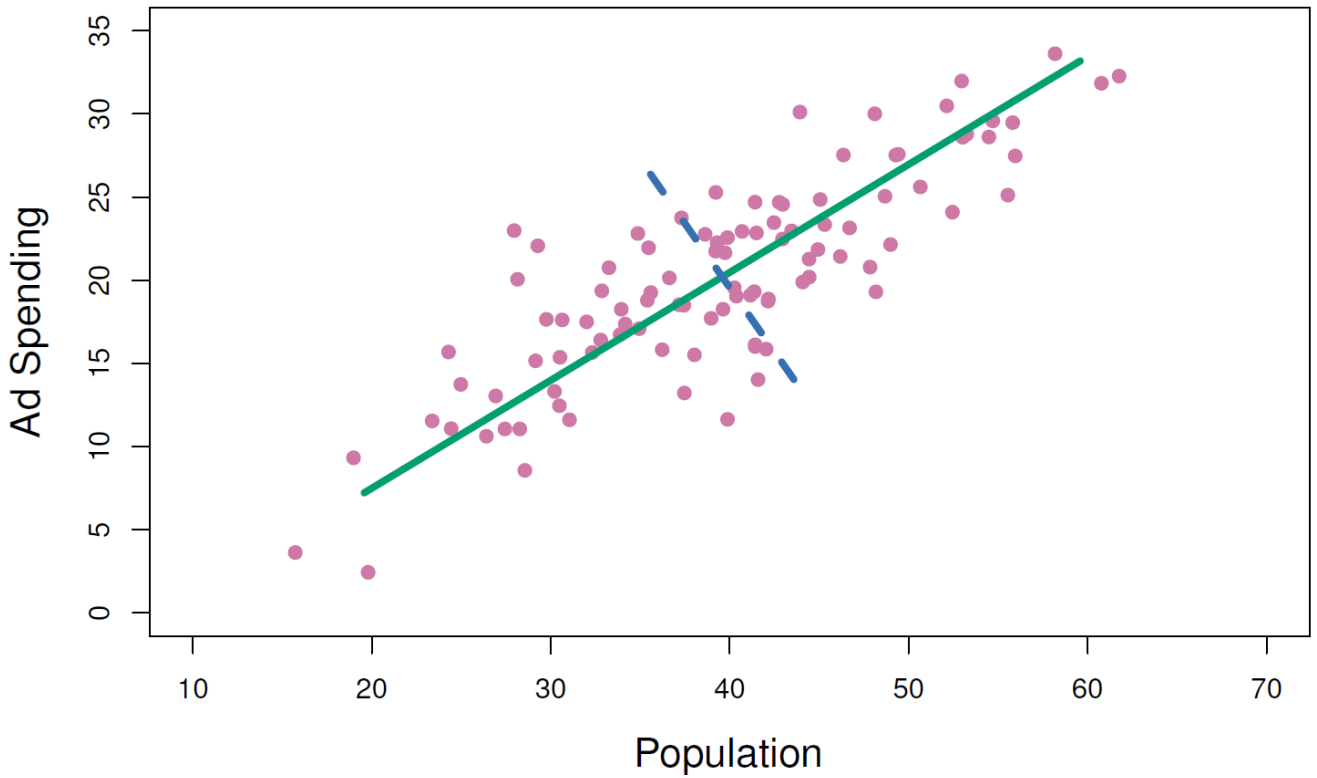
- A normalização deve-se ao fato da $Var(Z_1)$ aumentar a medida que ϕ_1 aumenta, assim eliminamos esta inconveniência;

- A **segunda componente principal** $\phi_2^t \mathbf{X}$ maximiza $Var(\phi_2^t \mathbf{X})$, sujeito a restrição $\phi_2^t \phi_2 = 1$ e $Cov(\phi_1^t \mathbf{X}, \phi_2^t \mathbf{X}) = 0$;

- A ***i*-ésima componente principal** $\phi_i^t \mathbf{X}$ maximiza $Var(\phi_i^t \mathbf{X})$, sujeito a restrição $\phi_i^t \phi_i = 1$ e $Cov(\phi_k^t \mathbf{X}, \phi_i^t \mathbf{X}) = 0, \forall k < i$.

ACP: exemplo

- O gráfico abaixo retrata o tamanho da população (pop) versus gasto com publicidade (ad) em 100 diferentes cidades;



- A linha verde representa a direção da primeira componente principal e a azul da segunda.

Cálculo das componentes principais

- Suponha que temos uma matriz $\mathbf{X}$, $n \times p$.

- Desde que temos interesse somente na variabilidade, assumimos que os vetores coluna de $\mathbf{X}$ têm média zero;

- E, considerando deste fato, encontramos a combinação linear das características $\mathbf{X}_i$ da forma

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \ldots + \phi_{p1} x_{ip},$$

para $i = 1, \ldots, n$, tal que apresente a maior variabilidade, sujeito a restrição $\sum_{j=1}^{p} \phi_{j1}^2 = 1$;

- Como os $\mathbf{x}_{.j}$'s tem média zero, os $\mathbf{z}_{.j}$'s também terão. Assim, a variância de $\mathbf{z}_{.j} = \dfrac{\sum_{i=1}^{n} z_{ij}^2}{n}$.

Cálculo das componentes principais

- Dessa forma, a primeira componente principal resolve o problema de otimização

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

Resultado: Seja $\mathbf{\Sigma}$ a matriz de covariância associada ao vetor aleatório $\mathbf{X}^t = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$, e $(\lambda_j, \mathbf{e}_j)$ o autopar, tal que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.Então

$$Z_j = \mathbf{e}_j^t\mathbf{X} = e_{1j}X_1 + e_{2j}X_2 + \ldots + e_{pj}X_p,$$

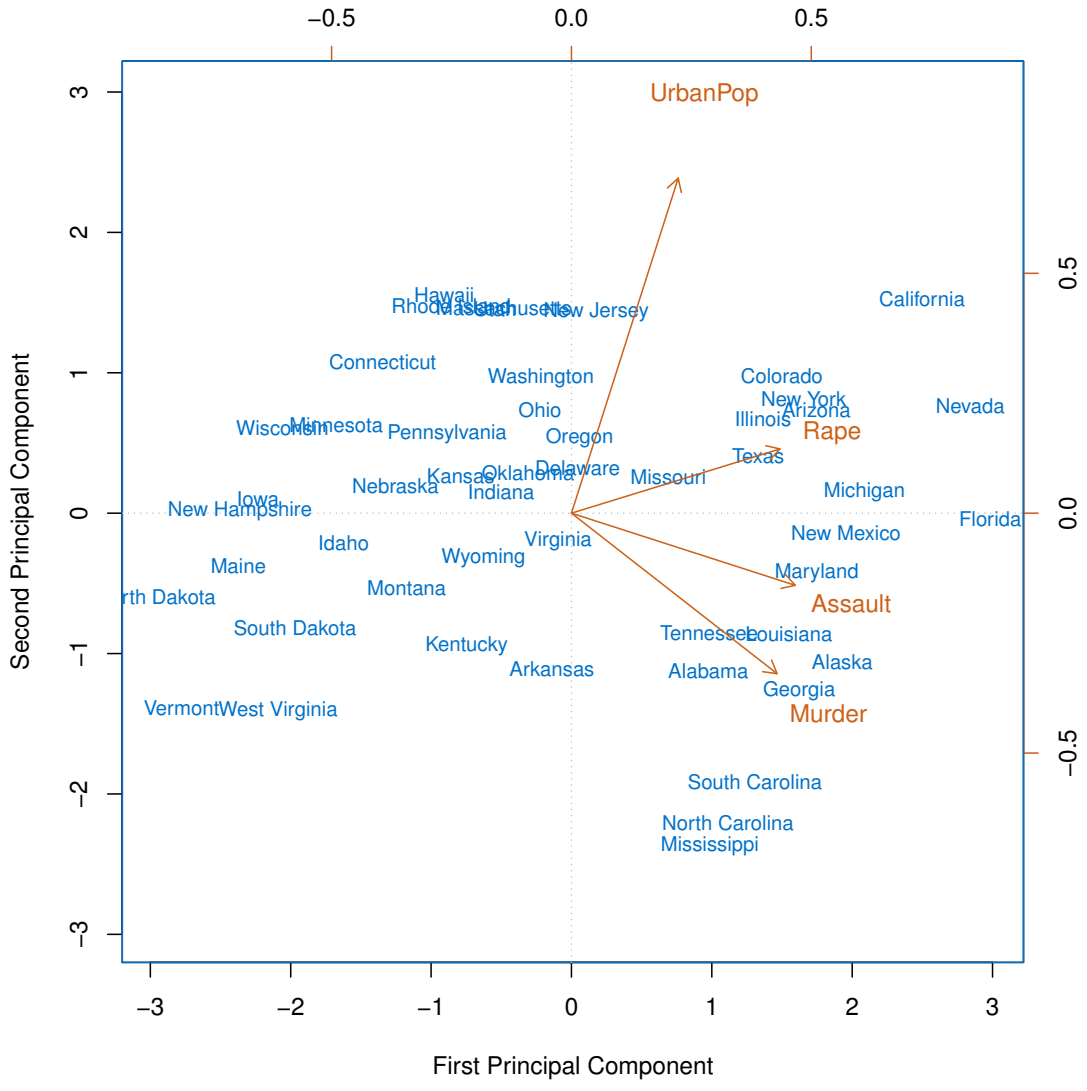
com $Var(Z_j) = \mathbf{e}_j^t\Sigma\mathbf{e}_j = \lambda_j$ e $Cov(Z_j, Z_k) = \mathbf{e}_j^t\Sigma\mathbf{e}_k$.

Exemplo: `USAarrests` data

- Os dados contém o número de prisões por 100.000 residentes nos 50 estados dos Estados Unidos;
- As prisões decorrem de um dos três crimes: assalto (`assault`), assassinato (`murder`) ou estupro (`rape`);
- Registrou-se, também, o percentual da população vivendo em área urbana de cada estado (`UrbanPop`);
- Dessa forma, temos $n = 50$ e $p = 4$.
- No gráfico a seguir são apresentadas as duas primeiras componentes principais para o dados em questão.

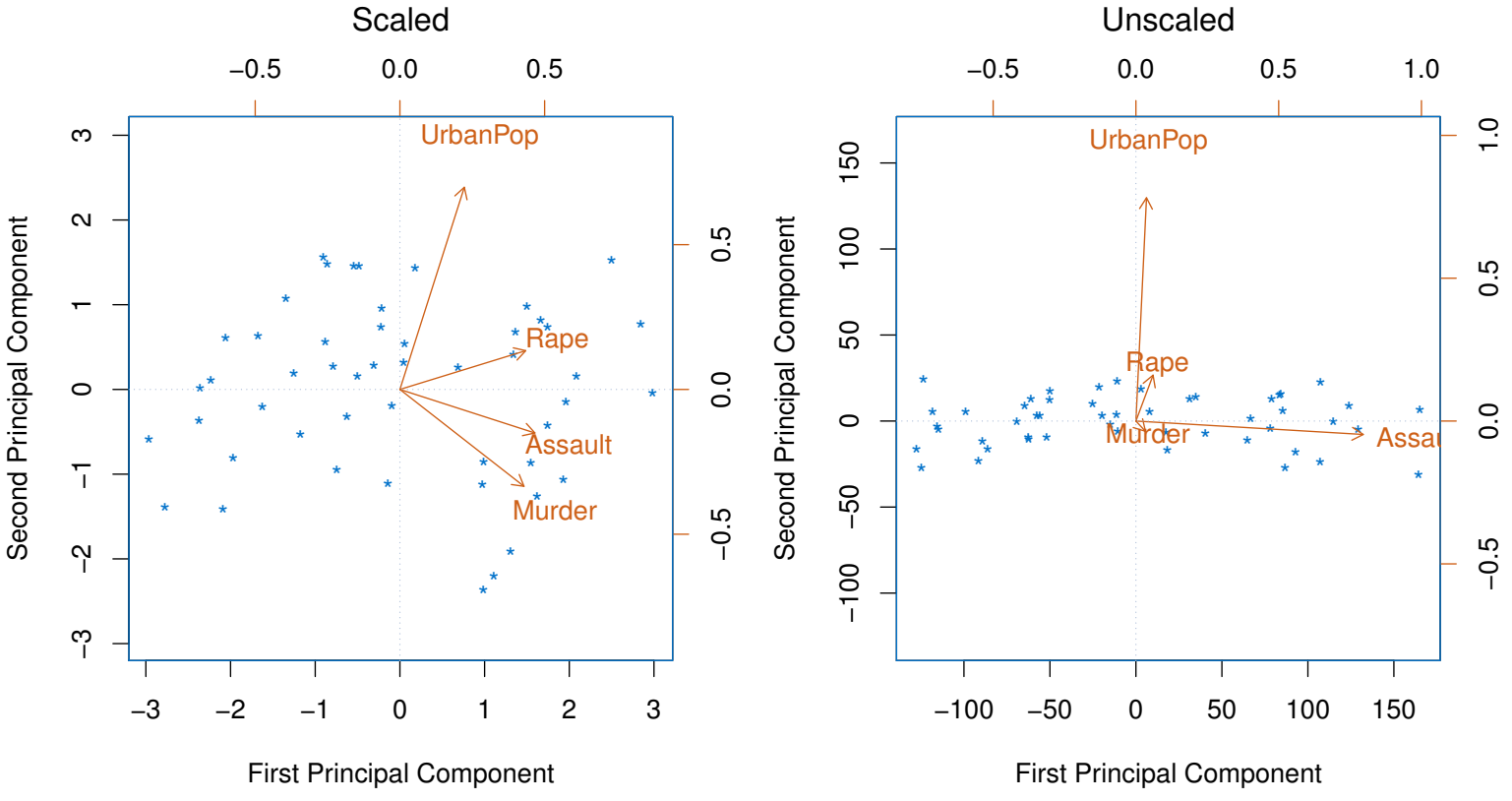
Exemplo: `USAarrests` data

- Os nomes em azul representam o escore para as duas c.p.;



Escalando as variáveis

- Se as variáveis estão em diferentes unidades é recomendável escalar cada uma para se ter um desvio padrão igual a 1.
- Se estiverem em uma mesma métrica, pode-se ou não escalá-las;



Proporção da variância explicada

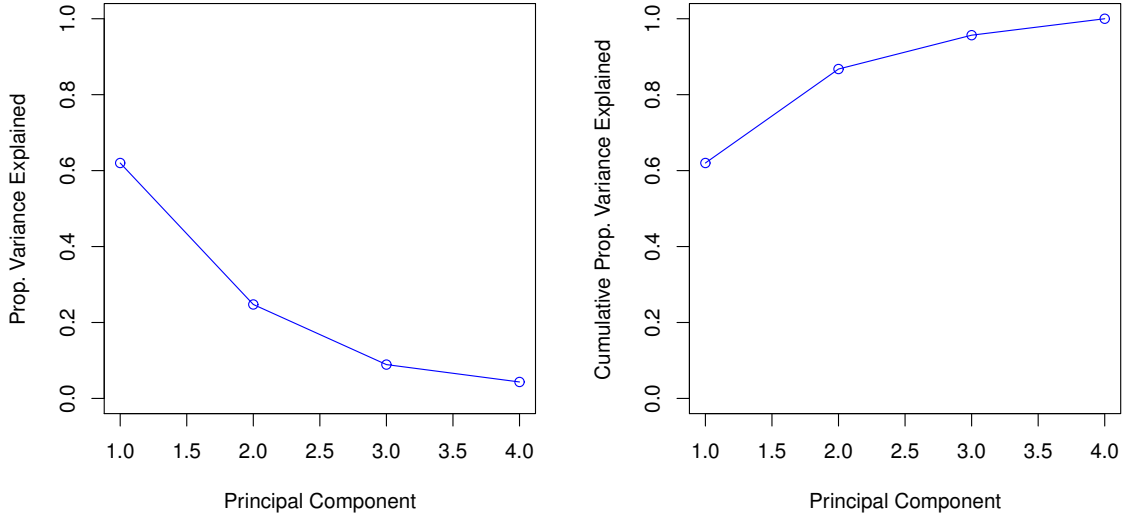
- A variância total presente nos dados é definida como

$$\sum_{j=1}^{p} Var(\mathbf{X}_j) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{j=1}^{p} Var(\mathbf{Z}_j)$$

- Assim, a proporção da variância explicada pela j-ésima componente principal é dada por

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}$$

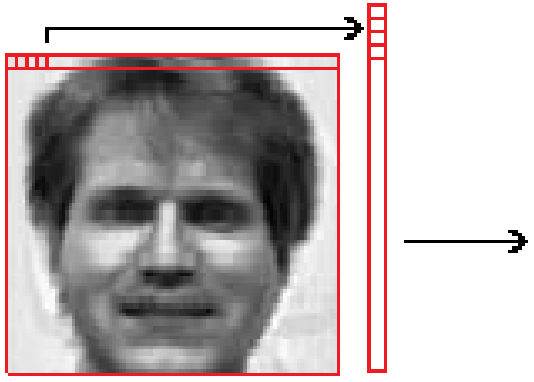
- Abaixo, a proporção da variância explicada por cada uma das quatro c.p. referente ao `USAarrests` data;

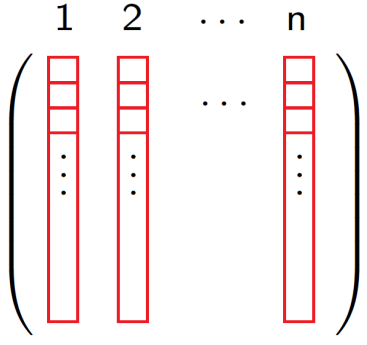


Exemplo: Eigen Faces

- Considere uma base com n imagens de faces;

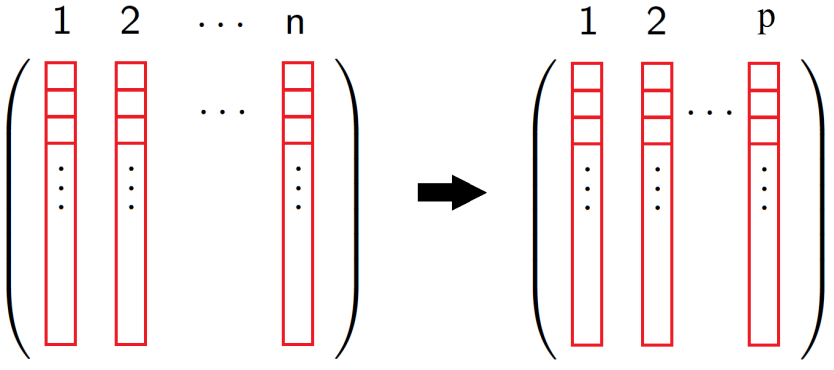
- Cada face é uma matriz $k \times k$, que transformamos em um vetor k^2;
- Em seguida agrupamos em uma matriz $k^2 \times n$.

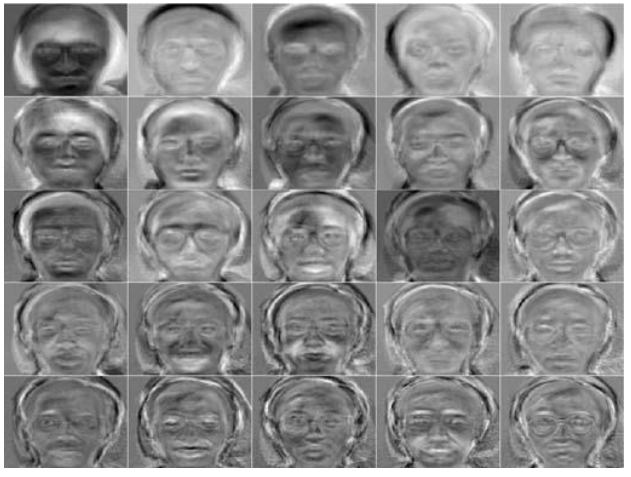




Exemplo: Eigen Faces

- Extraindo as componentes principais chegamos a uma matriz $k^2 \times p$;





  - Para uma específica face temos:

  - **Pergunta:** Quantas componentes precisamos?

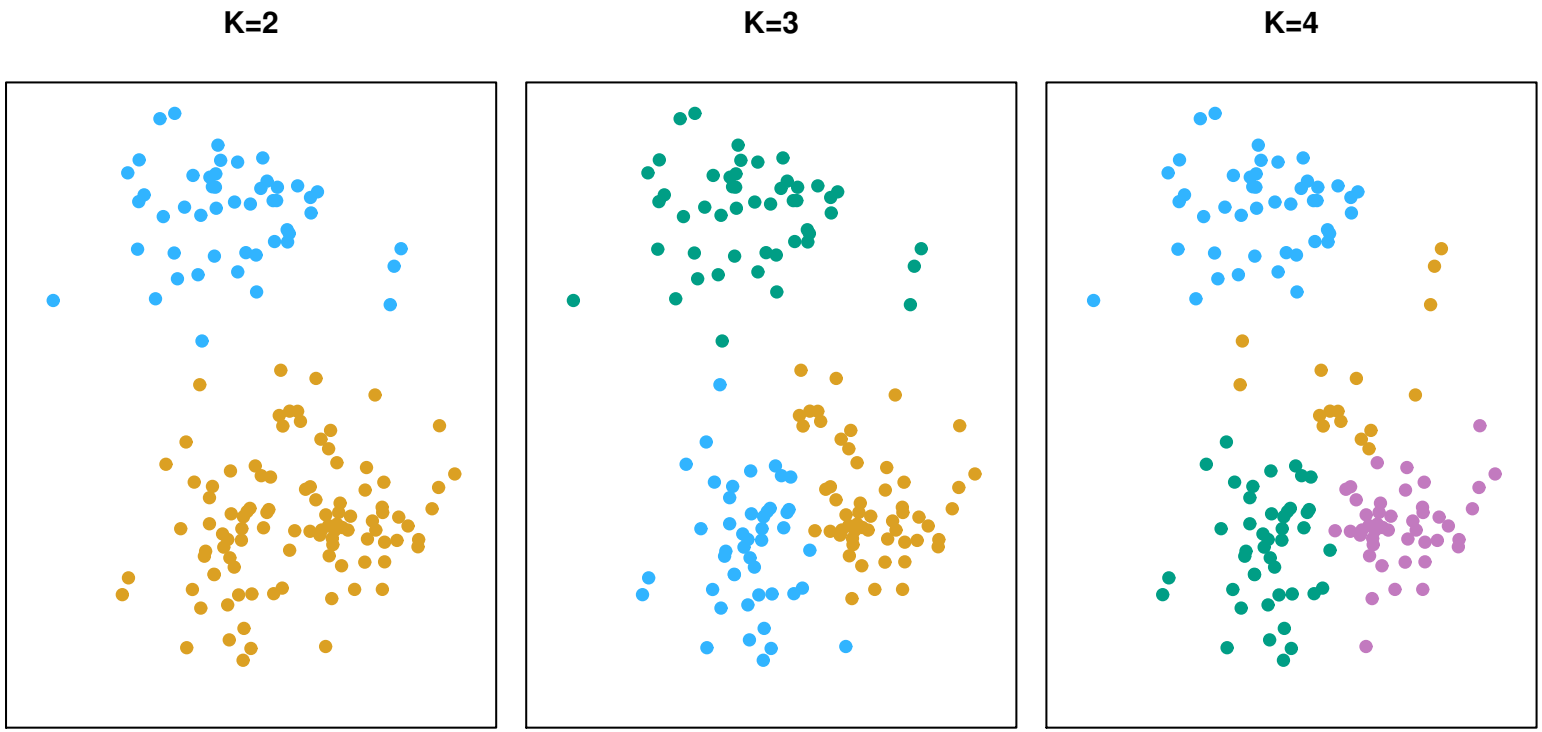
Clustering

- **Clustering** refere-se ao conjunto de técnicas para encontrar subgrupos (ou *clusters*) a partir dos dados;
- Buscamos partições em grupos distintos, tal que observações dentro de cada grupo sejam similares entre si e diferentes dos demais;
- Ou seja, subgrupos homogêneos segundo algum critério de similaridade;
- Diferente dos componentes principais, que busca explicar o máximo da variância em baixas dimensões.
- Veremos dois métodos:
  - ⋆ ***K*-means clustering**: procuramos partições dos dados em um número pré-determinado de clusters;
  - ⋆ **Hierarchical clustering**: não sabemos de antemão quantos clusters utilizaremos. Isso será feito através de uma representação visual.

Exemplo: K-means

- Os dados simulados consistem em 150 observações no espaço 2-dimensional;

- Os painéis representam os resultados de K-means para diferentes valores de K (que representa o número de cluster);



Detalhes do algoritmo

- Seja $C_1, C_2, \ldots, C_K$ respectivos grupos contendo os índices das observações, satisfazendo as seguintes propriedades:

  - $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. Em outras palavras, cada observação pertence à pelo menos um grupo;

  - $C_k \cap C_{k'} = \emptyset$. Ou seja, as observações não pertencem a mais de um grupo ao mesmo tempo.

- P. ex., se a i-ésima observação está no k-ésimo cluster, então $i \in C_k$;

- A ideia do K-means é buscar agrupamentos tal que a variação dentro de cada cluster seja tão pequena quanto possível;

Detalhes do algoritmo

- A variação dentro do cluster C_k (*within-cluster variation*) é medida pelo $WCV(C_k)$. É a quantidade pela qual as observações dentro do cluster diferem entre si;

- Tipicamente utilizamos a distância Euclideana

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

em que $|C_k|$ denota o número de observações no k-ésimo cluster;

- Sendo assim, queremos resolver o seguinte problema

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

- I.e., particionar as observação em K clusters, tal que o total da variação dentro de cada agrupamento (somado para todo k) seja o menor possível.

Detalhes do algoritmo

- Note que,

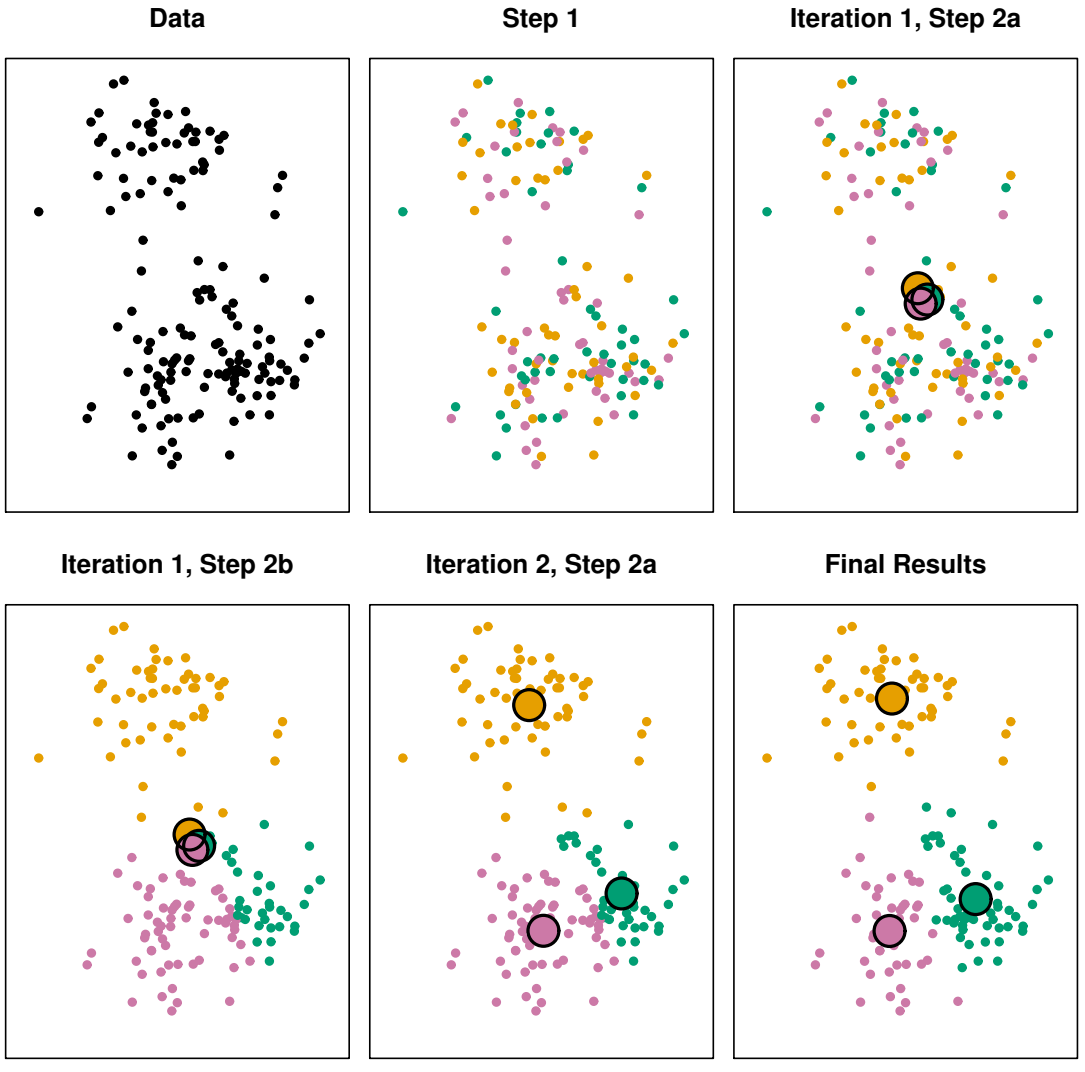
$$\underset{C_1,\ldots,C_K}{\textit{argmin}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} = \underset{C_1,\ldots,C_K}{\textit{argmin}} \left\{ \sum_{k=1}^{K} 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \right\}$$

- $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$ é a média da característica j no cluster C_k.

Algoritmo

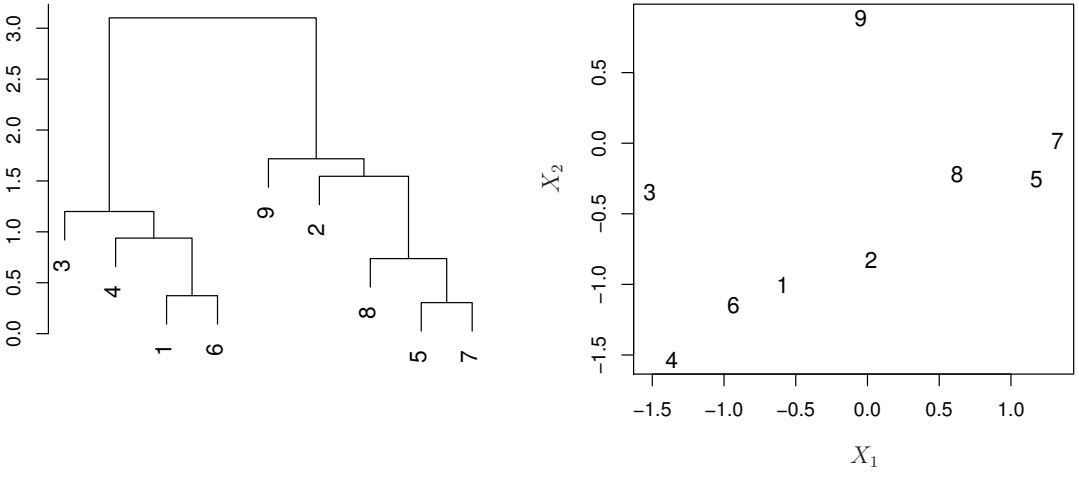
- ⋆ **Step 1:** Atribua, aleatoriamente, cada observação em um dos K clusters (este é o chute inicial);
- ⋆ **Step 2:** Itere até que os clusters se estabilizem:
  - (a) Para cada K cluster, calcule seu centroide;
  - (b) Atribua cada observação ao cluster mais próximo (menor distância Euclideana).

Exemplo

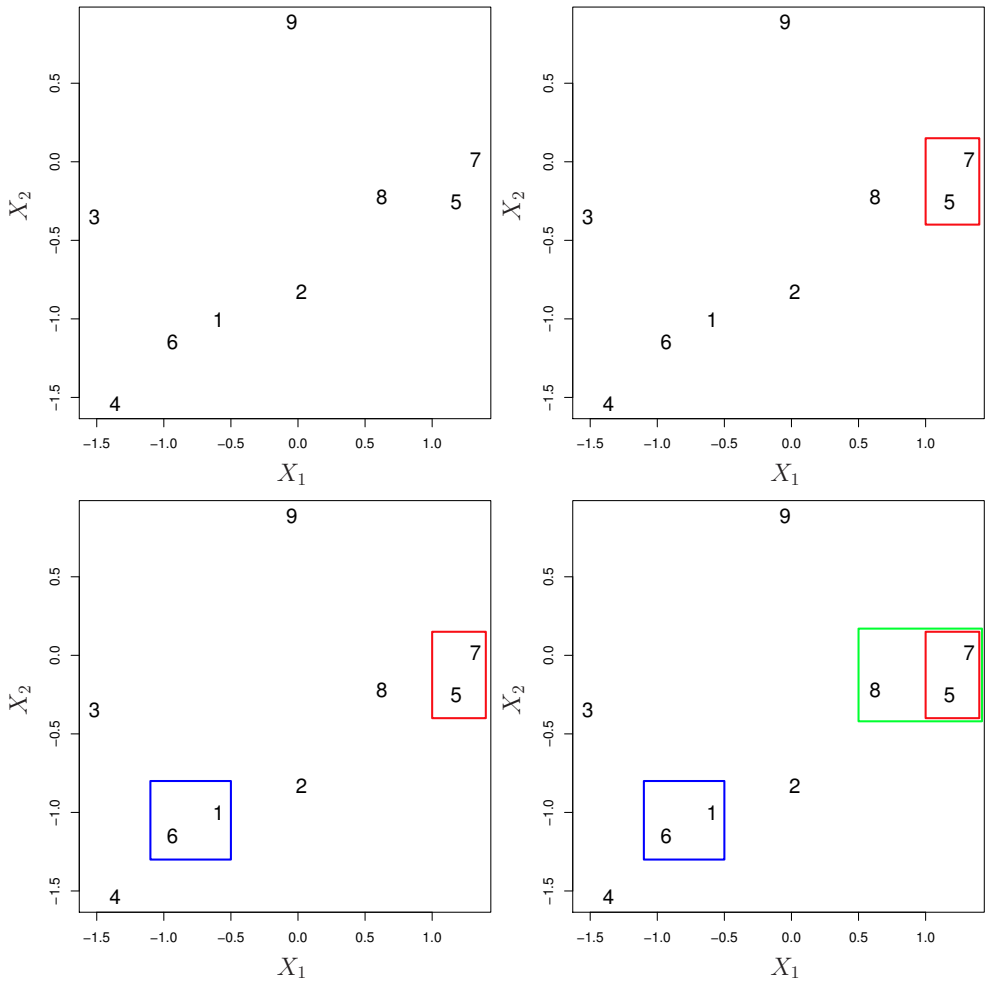

Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Hierarchical Clustering

- Como vimos, k-means exige que preespecifiquemos o número de clusters K. O que pode ser uma desvantagem;
- **Hierarchical clustering** é uma abordagem alternativa que não exige comprometimento com a escolha de K;
- Descreveremos a abordagem por aglomeração (*agglomerative clustering*).
- A ideia é construir um dendrograma com folhas que se agrupam até chegar ao tronco;



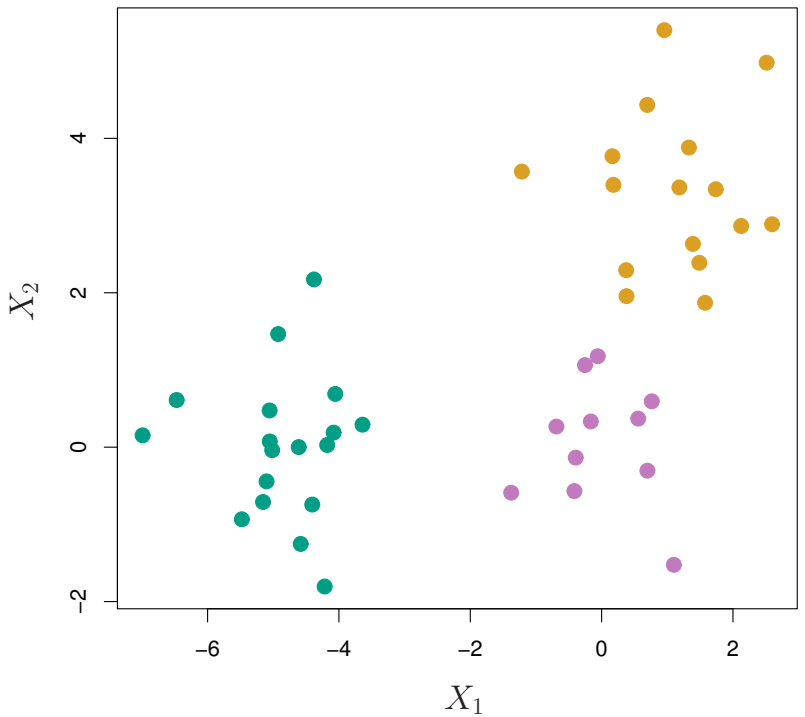
Ideia do algoritmo



- Iniciamos com cada ponto sendo seu próprio cluster;
- Identificamos os dois clusters mais próximos e os agrupamos;
- Repetimos este processo até restar um cluster.

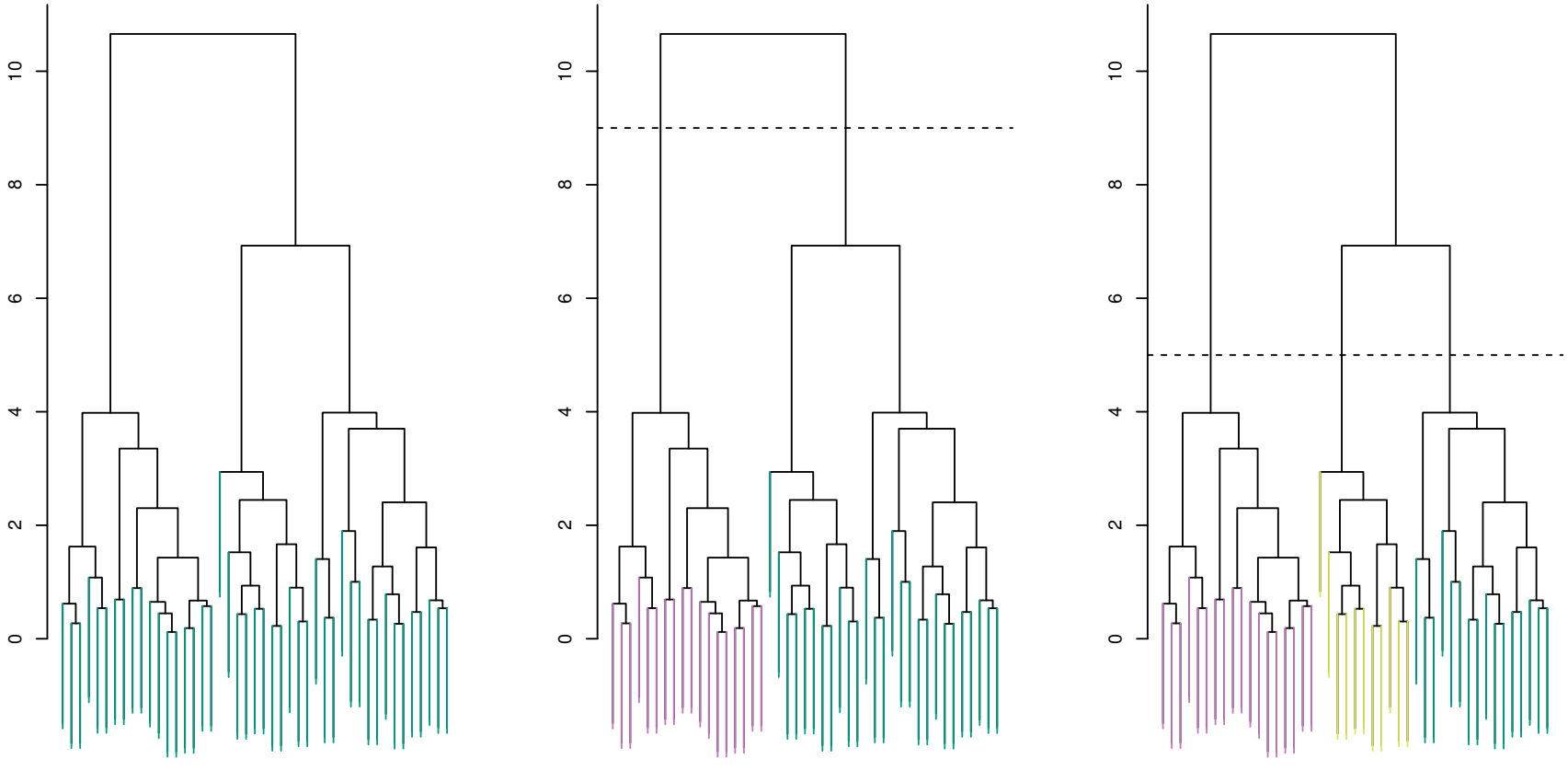
Exemplo

- Temos 45 observações geradas no espaço 2-dimensional;
- De fato, temos 3 classes distintas (separadas por cores);
- Todavia, trataremos esses rótulos como desconhecidos, e agruparemos as observações a fim de descobrir suas classes.



Exemplo

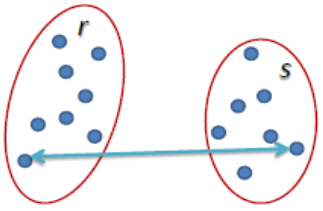
- Abaixo, três dendrogramas com diferentes alturas de corte (que resulta em clusters distintos);



Tipo de Linkage

Complete

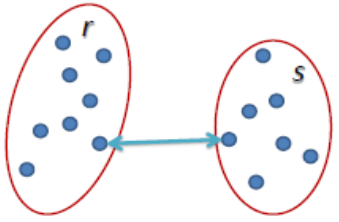
- Calculamos a máxima dissimilaridade entre os clusters.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Single

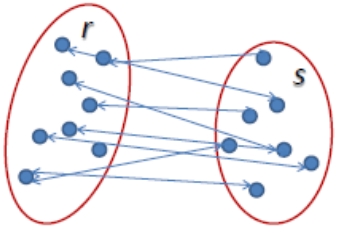
- Calculamos a mínima dissimilaridade entre os clusters.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Average

- Calculamos a dissimilaridade média entre os clusters.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Exemplo

- Em geral, **average** e **complete** linkage tendem a produzir agrupamentos mais equilibrados.