Aula 5: k-means

Paulo C. Marques F.

Aula ministrada no Insper

1 de Abril de 2016

Aprendizagem não supervisionada

- Encerramos, por hora, o capítulo sobre aprendizagem supervisonada e iniciamos o estudo dos casos de "aprendizagem não supervisionada".
- Situações inferenciais em que as observações não estão relacionadas a uma resposta y_i fornecida por um "supervisor".
- A partir de agora, os dados não possuem mais rótulos que os classifiquem, ou respostas quantitativas (como nos problemas de regressão).
- É uma área muito menos desenvolvida em "Machine Learning".

Análise de clusters (1)

- O problema exemplar em aprendizagem não supervisionada é a análise de clusters (conglomerados).
- O objetivo é definir classes de equivalência tais que os dados dentro de uma mesma classe sejam "similares" segundo alguma perspectiva.
- Queremos encontrar estruturas nos dados.
- O resultado de uma análise de clusters pode ser utilizado para verificar as decisões de um eventual supervisor.

Análise de clusters (2)

- Para cada unidade amostral $i = 1, \ldots, n$, conhecemos apenas o vetor $x_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$.
- Formalmente, um cluster C_i é o conjunto dos índices dos dados que pertencem a ele.
- Queremos construir $k \geq 1$ clusters $C_1, \ldots, C_k$ tais que

$$\cup_{i=1}^{k} C_i = \{1, \ldots, n\},$$

e $C_i \cap C_j = \emptyset$ (hard clustering), quando $i \neq j$, de modo a minimizar a dispersão intra-clusters

$$W = \frac{1}{2} \sum_{r=1}^{k} \frac{1}{n_r} \sum_{i \in C_r} \sum_{j \in C_r} \|x_i - x_j\|^2,$$

em que n_r é o número de observações no cluster C_r e $\|x_i - x_j\|^2 = \sum_{\ell=1}^{p} (x_{i\ell} - x_{j\ell})^2$ é o quadrado da distância euclidiana entre x_i e x_j.

Análise de clusters (3)

- De quantas maneiras $A(n, k)$ podemos formar os k clusters a partir das n observações?
- Suponha que você é a n-ésima unidade amostral.
- Você pode fazer duas escolhas, mutuamente exclusivas.
- Você pode decidir criar um cluster apenas para você e as demais $n-1$ pessoas se agruparão em $k-1$ clusters de $A(n-1, k-1)$ maneiras.
- Ou você pode escolher entrar em um de k clusters e as demais $n-1$ pessoas se agruparão nestes k clusters de $A(n-1, k)$ maneiras.
- Deste modo, obtemos a relação de recorrência

$$A(n, k) = A(n-1, k-1) + k \cdot A(n-1, k),$$

com as condições $A(n, 1) = A(n, n) = 1$.

Análise de clusters (4)

```
A <- function(n, k) {
    if (k == 1 || k == n) {
        return(1)
    }
    return(A(n-1, k-1) + k*A(n-1, k))
}
```

- Há casos factíveis: $A(10, 4) = 34\,105$, por exemplo.
- No entanto, $A(30, 4) \approx 10^{16}$ e o problema se torna computacionalmente intratável.
- Os $A(n, k)$ são conhecidos na literatura de combinatória como números de Stirling de segunda espécie.
- De fato, é possível (vide Polya ou Knuth) resolver a relação de recorrência e encontrar

$$A(n, k) = \frac{1}{k!} \sum_{r=1}^{k} (-1)^{k-r} \binom{k}{r} r^n.$$

Em busca de uma aproximação (1)

Lema 1. Vale a identidade

$$\sum_{i \in C_r} \sum_{j \in C_r} \|x_i - x_j\|^2 = 2\, n_r \sum_{i \in C_r} \|x_i - \bar{x}_r\|^2,$$

em que $\bar{x}_r = \sum_{i \in C_r} x_i / n_r$ é a média das observações pertencentes ao cluster C_r.

- A idéia da demonstração é usar que

$$\|u - v\|^2 = \langle u - v, u - v \rangle = \|u\|^2 - 2 \langle u, v \rangle + \|v\|^2$$

  e "somar zero" no lugar adequado.

Em busca de uma aproximação (2)

Demonstração

$$\sum_{i \in C_r} \sum_{j \in C_r} \|x_i - x_j\|^2 = \sum_{i \in C_r} \sum_{j \in C_r} \|(x_i - \bar{x}_r) - (x_j - \bar{x}_r)\|^2$$

$$= \sum_{i \in C_r} \sum_{j \in C_r} \left(\|x_i - \bar{x}_r\|^2 - 2 \langle x_i - \bar{x}_r, x_j - \bar{x}_r \rangle + \|x_j - \bar{x}_r\|^2 \right)$$

$$= \sum_{i \in C_r} \left(n_r \|x_i - \bar{x}_r\|^2 - 2 \left\langle x_i - \bar{x}_r, \sum_{j \in C_r} (x_j - \bar{x}_r) \right\rangle + \sum_{j \in C_r} \|x_j - \bar{x}_r\|^2 \right)$$

$$= 2\, n_r \sum_{i \in C_r} \|x_i - \bar{x}_r\|^2.$$

Em busca de uma aproximação (3)

- Portanto, pelo Lema 1, o problema original

$$\arg \min_{C_1,\ldots,C_k} \frac{1}{2} \sum_{r=1}^{k} \frac{1}{n_r} \sum_{i \in C_r} \sum_{j \in C_r} \| x_i - x_j \|^2$$

é equivalente a

$$\arg \min_{C_1,\ldots,C_k} \sum_{r=1}^{k} \sum_{i \in C_r} \| x_i - \bar{x}_r \|^2.$$

Em busca de uma aproximação (4)

Lema 2. Para os vetores $u_1, \ldots, u_m \in \mathbb{R}^p$, a quantidade

$$\sum_{i=1}^{m} \|u_i - c\|^2$$

é minimizada escolhendo-se o vetor $c = \bar{u} = \sum_{i=1}^{m} u_i/m$.

Demonstração

$$\begin{aligned}
\sum_{i=1}^{m} \|u_i - c\|^2 &= \sum_{i=1}^{m} \|(u_i - \bar{u}) - (c - \bar{u})\|^2 \\
&= \sum_{i=1}^{m} \left(\|u_i - \bar{u}\|^2 - 2\langle u_i - \bar{u}, c - \bar{u} \rangle + \|c - \bar{u}\|^2 \right) \\
&= \sum_{i=1}^{m} \|u_i - \bar{u}\|^2 + m\|c - \bar{u}\|^2.
\end{aligned}$$

Em busca de uma aproximação (5)

- Usando o Lema 2, o problema original equivale a minimizar o "custo estendido"

$$\arg \min_{\substack{C_1,\ldots,C_k \\ m_1,\ldots,m_k}} \sum_{r=1}^{k} \sum_{i \in C_r} \|x_i - m_r\|^2.$$

- Esta representação do problema sugere uma solução iterativa em que primeiramente fixamos os m_r's e minimizamos o custo estendido escolhendo os C_r's adequadamente, e posteriormente fixamos os C_r's e minimizamos o custo estendido escolhendo os m_r's como sendo as médias das observações nos respectivos clusters.

Algoritmo k-means (Lloyd (1957))

- Inicializamos arbitrariamente $m_1, \ldots m_k$.
- Alocamos a observação x_i no cluster C_r tal que

$$r = \arg \min_{1 \leq r \leq k} \|x_i - m_r\|,$$

  para $i = 1, \ldots, n$. Deste modo, determinamos $C_1, \ldots, C_k$.
- Fazemos $m_r = \bar{x}_r$, para $r = 1, \ldots, k$.
- Iteramos os dois passos anteriores até que o valor de W fique inalterado.

O algoritmo k-means converge

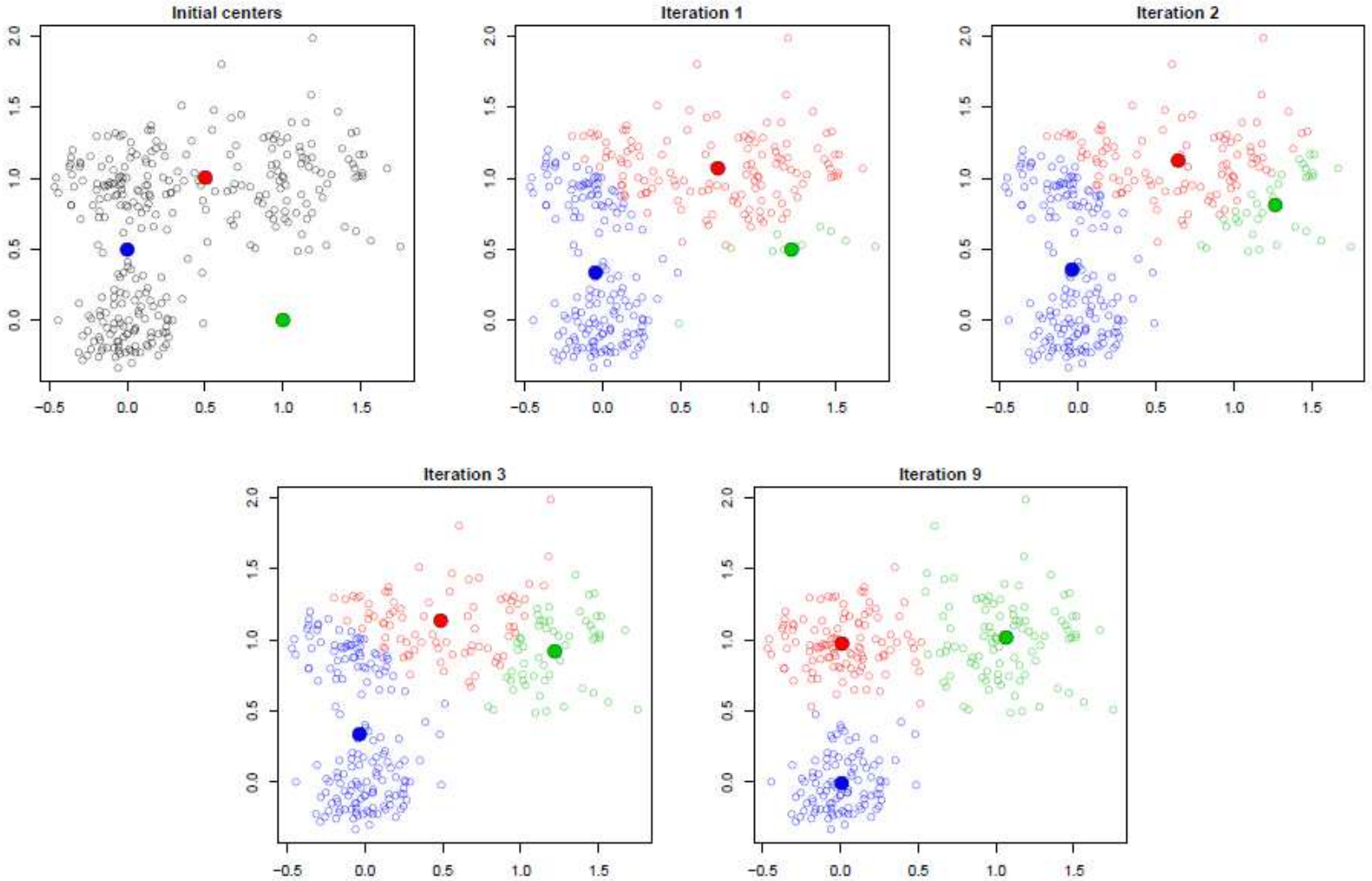
- Note que os dois passos iterativos do algoritmo k-means reduzem o valor do custo estendido

$$\sum_{r=1}^{k} \sum_{i \in C_r} \|x_i - m_r\|^2.$$

- Uma vez que o conjunto sobre o qual minimizamos é finito, o algoritmo k-means eventualmente converge.
- No entanto, não há nenhuma garantia de que encontraremos um mínimo global de W.
- De fato, o algoritmo k-means fornece uma configuração de clusters que produz um mínimo local para W.
- Por este motivo, os praticantes executam o algoritmo diversas vezes com inicializações distintas para os m_r's.
- O algoritmo k-means é implementado no `R` pela função `kmeans()`.

Iterações de exemplo

- Exemplo com $x_i \in \mathbb{R}^2$, $n = 300$ e $k = 3$ (Tibshirani).



Como escolher k?

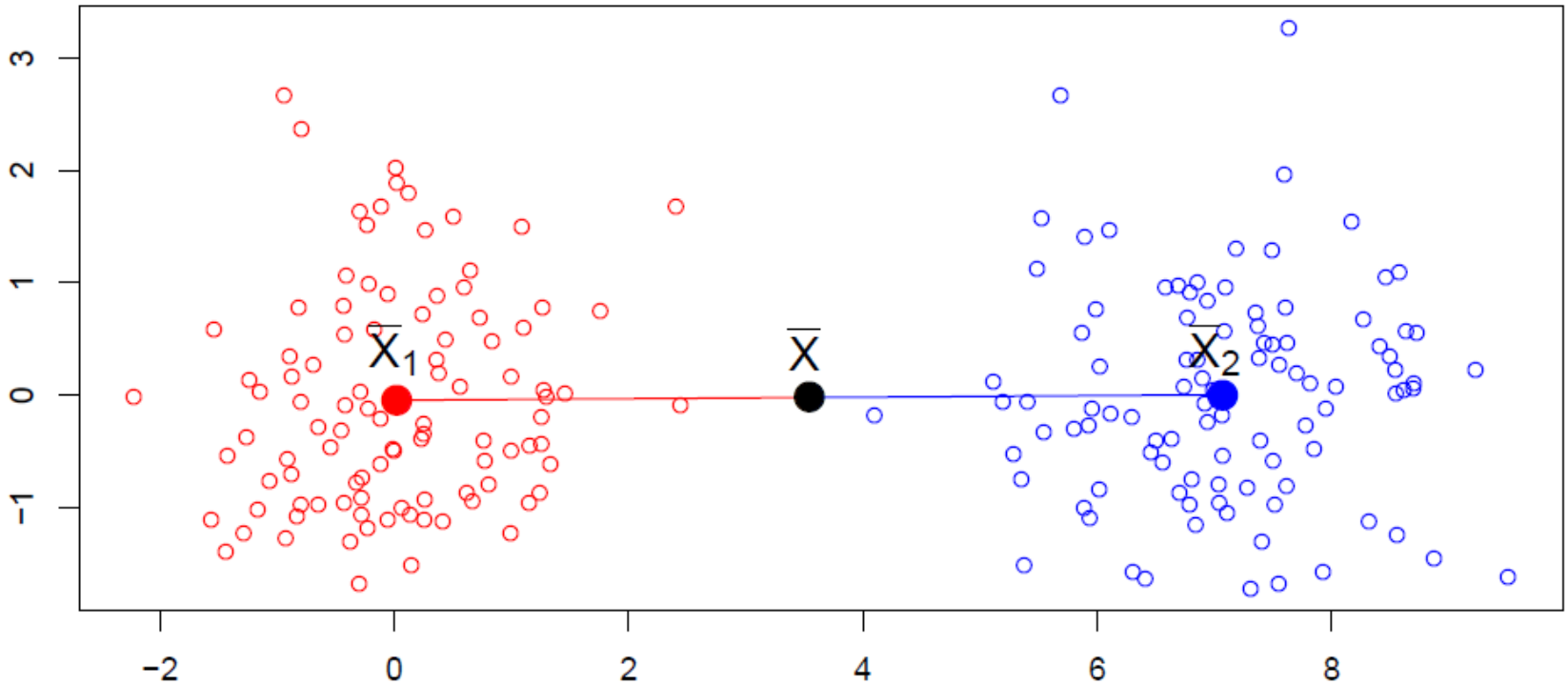
- Algumas vezes k é escolhido subjetivamente (no "olhômetro").
- Há aplicações em que k é determinado *a priori*: suponha que temos k especialistas que irão examinar as unidades amostrais e queremos distribuir dados "similares" para cada um deles.
- Uma alternativa seria obter $W(k)$ para diversos k's e pegar o menor deles.
- Definindo a dispersão entre-clusters

$$B(k) = \sum_{r=1}^{k} n_r \|\bar{x}_r - \bar{x}\|,$$

em que $\bar{x} = \sum_{i=1}^{n} x_i/n$ é a média de todas as observações, uma segunda alternativa seria escolher o k que maximiza $B(k)$.

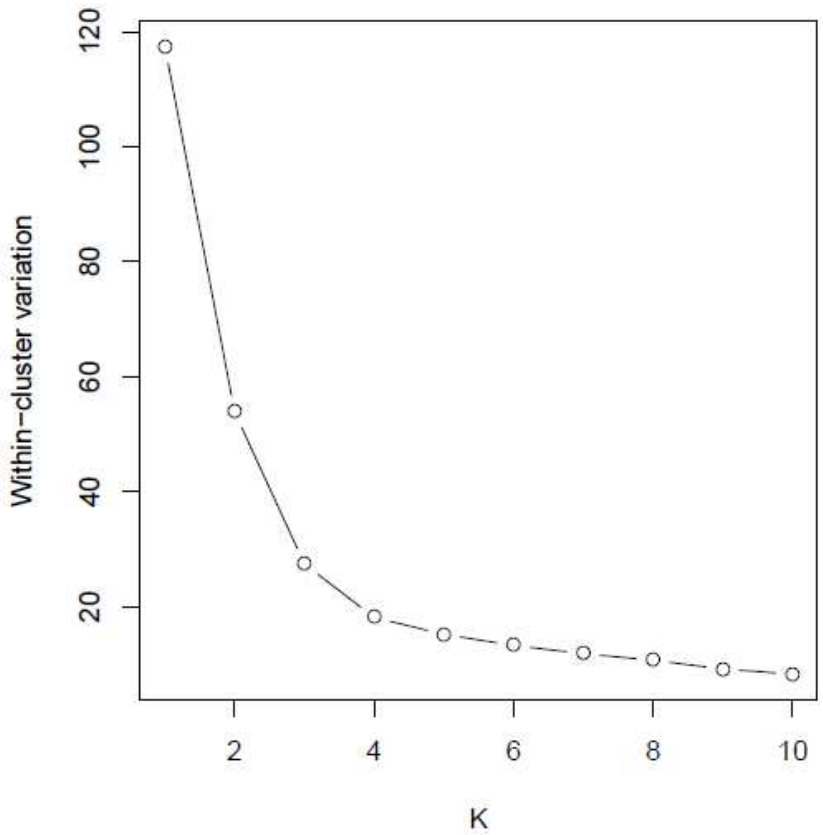
Dispersão entre-clusters

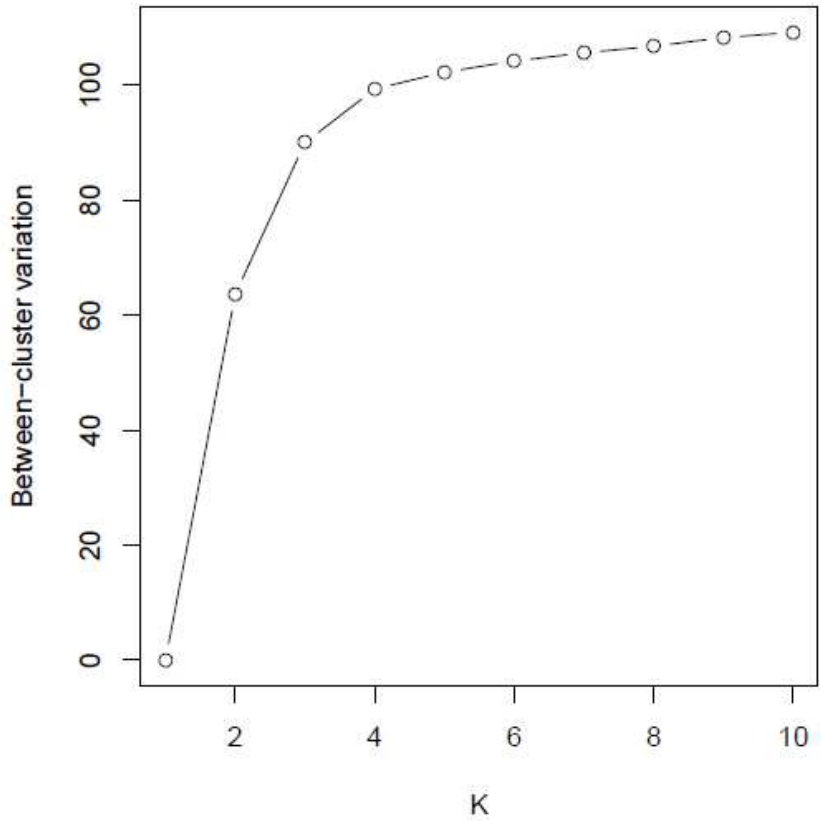
- Exemplo com $x_i \in \mathbb{R}^2$, $n = 100$ e $k = 2$ (Tibshirani).



Nenhuma das alternativas funciona

- Exemplo com $x_i \in \mathbb{R}^2$, $n = 250$ e $k = 1, \ldots, 10$ (Tibshirani).



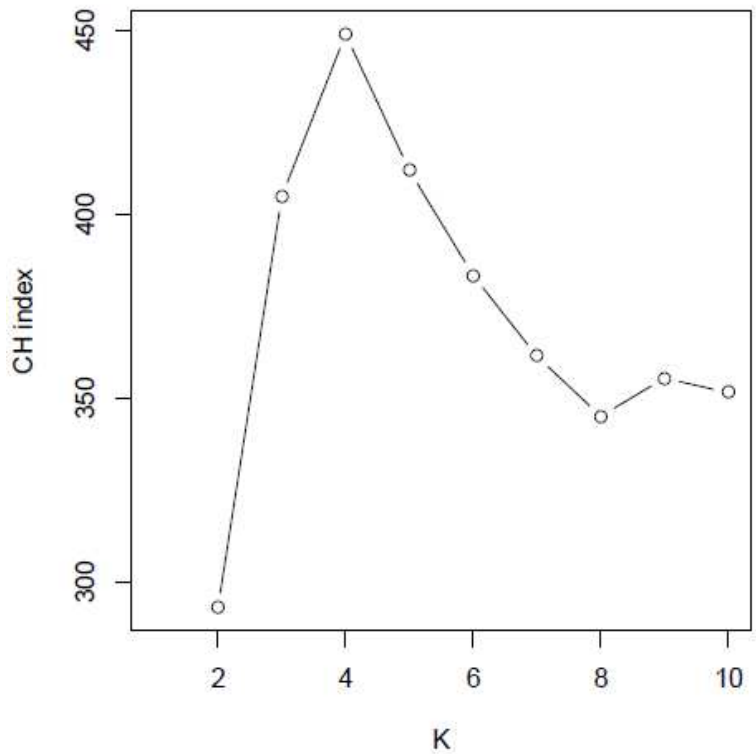


Índice CH

- A idéia de Calinski e Harabasz (1974) é obter simultaneamente um W pequeno e um B grande definindo o índice

$$\mathrm{CH}(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}; \qquad \hat{k} = \arg \max_{2 \leq k \leq k_{max}} \mathrm{CH}(k).$$

- Exemplo com $x_i \in \mathbb{R}^2$, $n = 250$ e $k = 2, \ldots, 10$ (Tibshirani).



Obrigado pela participação de todos!