Reconhecimento de Padrões
Estimação de Densidades

Prof. George von Borries
Departamento de Estatística
Universidade de Brasília

1 - 2024

Ideia

A estimação de densidades para cada classe do espaço observado permite que seja utilizada a Regra de Bayes para produzir regras de discriminação.

- As probabilidades a posteriori $\mathrm{p}(\omega_j|\mathbf{x})$ podem ser expressas em função de $\mathrm{p}(\omega_j)$ e densidades condicionais as classes ω_j, $j \in \{1, \ldots, c\}$, através do Teorema de Bayes, i.e.

$$\mathrm{p}(\omega_j|\mathbf{x}) = \frac{\mathrm{p}(\mathbf{x}|\omega_j)p(\omega_j)}{\mathrm{p}(\mathbf{x})}$$

em que $\mathrm{p}(\mathbf{x}) = \sum_{i=1}^{c} \mathrm{p}(\mathbf{x}|\omega_i)\mathrm{p}(\omega_i)$.

- A **Regra de Bayes para Erro Mínimo** aloca $\mathbf{x}$ a classe ω_j se

$$\mathrm{p}(\mathbf{x}|\omega_j)\mathrm{p}(\omega_j) > \mathrm{p}(\mathbf{x}|\omega_k)\mathrm{p}(\omega_k)$$

para todo $k \neq j$ em $\{1, \ldots, c\}$.

Estimação Não Paramétrica de Densidades

Na estimação não paramétrica não existe informação sobre a forma da verdadeira densidade.

"Let the data speak for themselves."

1 Método do Histograma

- Seja $\mathbf{X} = \{X_1, \ldots, X_n\}$ uma amostra i.i.d. A ideia é agregar os dados em intervalos da forma $[x_0, x_0 + h)$ e usar a frequência relativa para aproximar a densidade $f(x)$ em $x \in [x_0, x_0 + h)$.
- Estimativa:

$$\mathrm{f}(x_0) = \mathrm{F}'(x_0) = \lim_{h \to 0^+} \frac{\mathrm{F}(x_0 + h) - \mathrm{F}(x_0)}{h} = \lim_{h \to 0^+} \frac{\mathrm{P}[x_0 < X < x_0 + h]}{h}$$

Estabelecida uma origem t_0 e um comprimento (*bandwith*) $h > 0$, o histograma constrói constantes nos intervalos $\{\mathrm{I}_k : [t_k\, t_{k+1}) : t_k = t_0 + hk, k \in \mathcal{Z}\}$, contanddo o número de pontos em cada I_k. Os intervalos são chamados de *bins*.

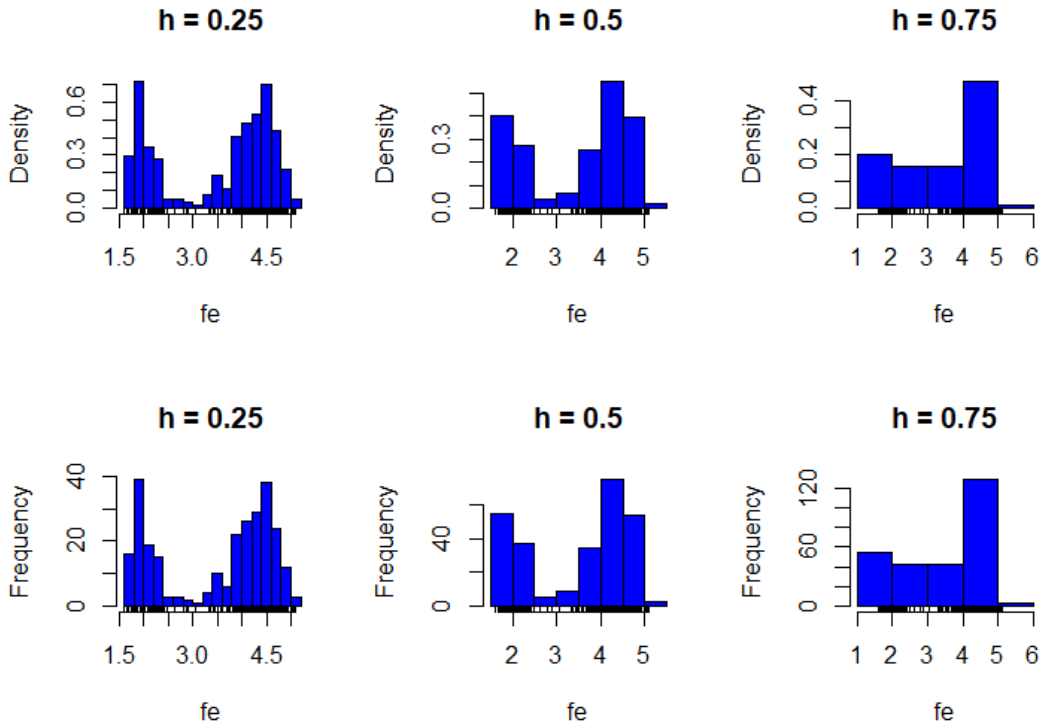
- O histograma no ponto x é definido como

$$\hat{f}_H(x; t_0, h) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}_{\{x_i \in I_k : x \in I_k\}}$$

ou

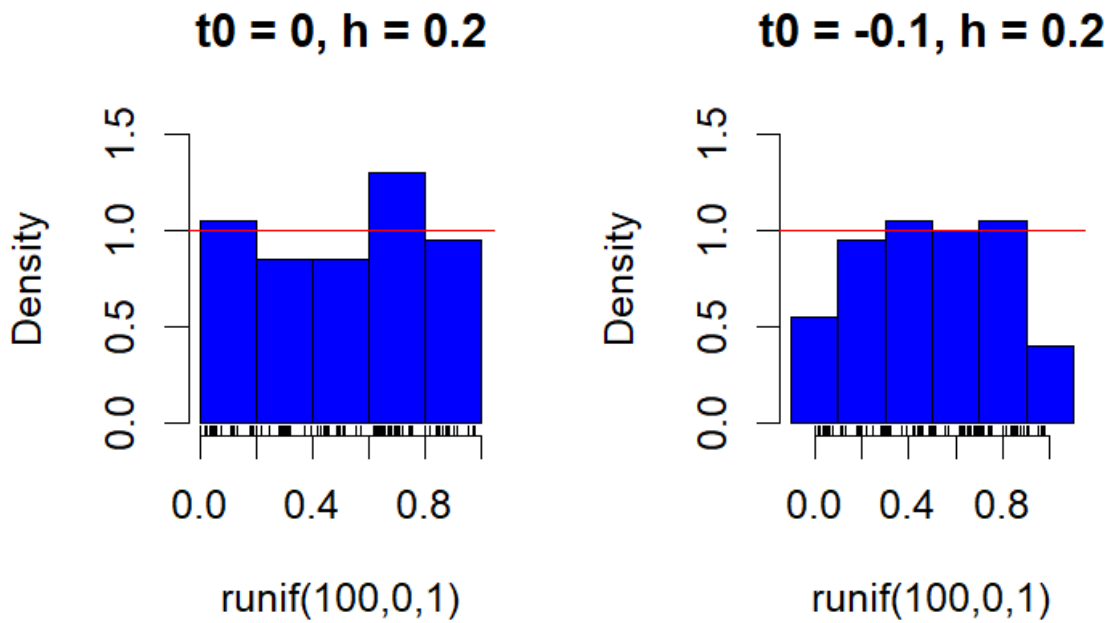
$$\hat{f}_H(x; t_0, h) = \frac{n_k}{nh} \quad \text{se} \quad x \in I_k \quad \text{para} \quad k \in \mathcal{Z}.$$

sendo n_k o número de pontos em I_k.



Histogramas (densidades e contagens) com diferentes valores de h.

Problema 1: dependência em relação a t_0.



Histogramas de 100 obs. da Uniforme(0,1), com diferentes valores de t_0.

Solução: Histograma móvel ou estimador de densidade "ingênuo" (*naive*).

- Utiliza intervalos da forma $(x-h, x+h)$ para agregar os dados e contar a frequência relativa, i.e., a estimava é obtida por

$$\mathrm{f}(x) = \mathrm{F}'(x) = \lim_{h \to 0^+} \frac{\mathrm{F}(x+h) - \mathrm{F}(x-h)}{2h} = \lim_{h \to 0^+} \frac{\mathrm{P}[x-h < X < x+h]}{2h}.$$

- O histograma no ponto x é definido como

$$\hat{\mathrm{f}}_{\mathrm{N}}(x; h) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}_{\{x-h < X_i < x+h\}}$$

Observar que $\sum_{i=1}^{n} \mathbf{1}_{\{x-h<X_i<x+h\}}$ tem distribuição $\mathrm{Binomial}(n, p_{x,h})$ em que $p_{x,h} = \mathrm{P}[x-h < X < x+h] = \mathrm{F}(x+h) - \mathrm{F}(x-h)$.

- Da Binomial, resulta que

$$\mathrm{E}\left[\hat{\mathrm{f}}_{\mathrm{N}}(x;h)\right] = \frac{\mathrm{F}(x+h) - \mathrm{F}(x-h)}{2h}$$

e

$$\mathrm{Var}\left[\hat{\mathrm{f}}_{\mathrm{N}}(x;h)\right] = \frac{\mathrm{F}(x+h) - \mathrm{F}(x-h)}{4nh^2} - \frac{(\mathrm{F}(x+h) - \mathrm{F}(x-h))^2}{4nh^2}$$

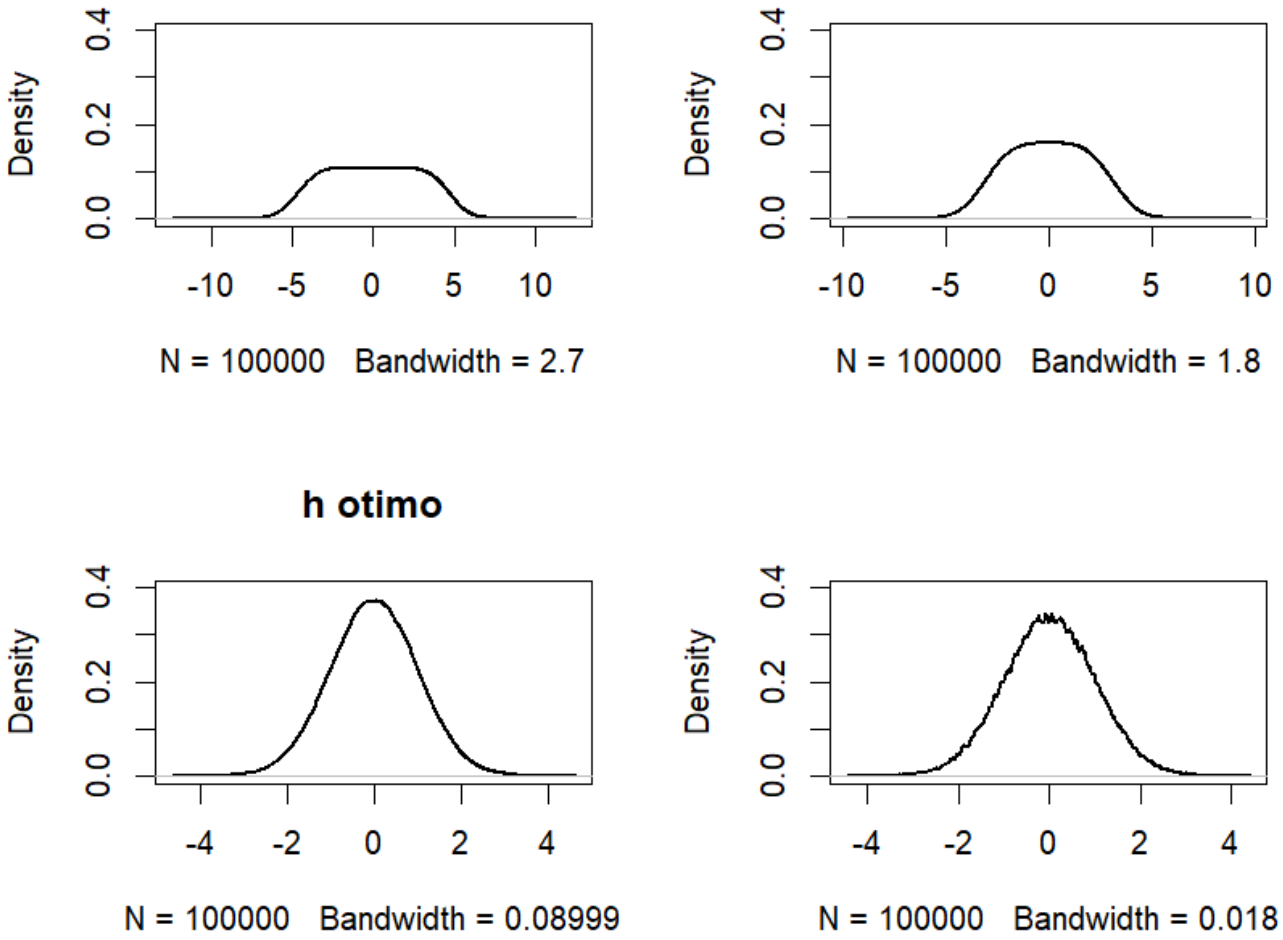
Permitindo observar que

(a) Se $h \to 0$: $\mathrm{E}\left[\hat{\mathrm{f}}_{\mathrm{N}}(x;h)\right] \to \mathrm{f}(x)$ (não viesado)
e $\mathrm{Var}\left[\hat{\mathrm{f}}_{\mathrm{N}}(x;h)\right] \to \infty$.

(b) Se $h \to \infty$: $\mathrm{E}\left[\hat{\mathrm{f}}_{\mathrm{N}}(x;h)\right] \to 0$ (viesado)
e $\mathrm{Var}\left[\hat{\mathrm{f}}_{\mathrm{N}}(x;h)\right] \to 0$.

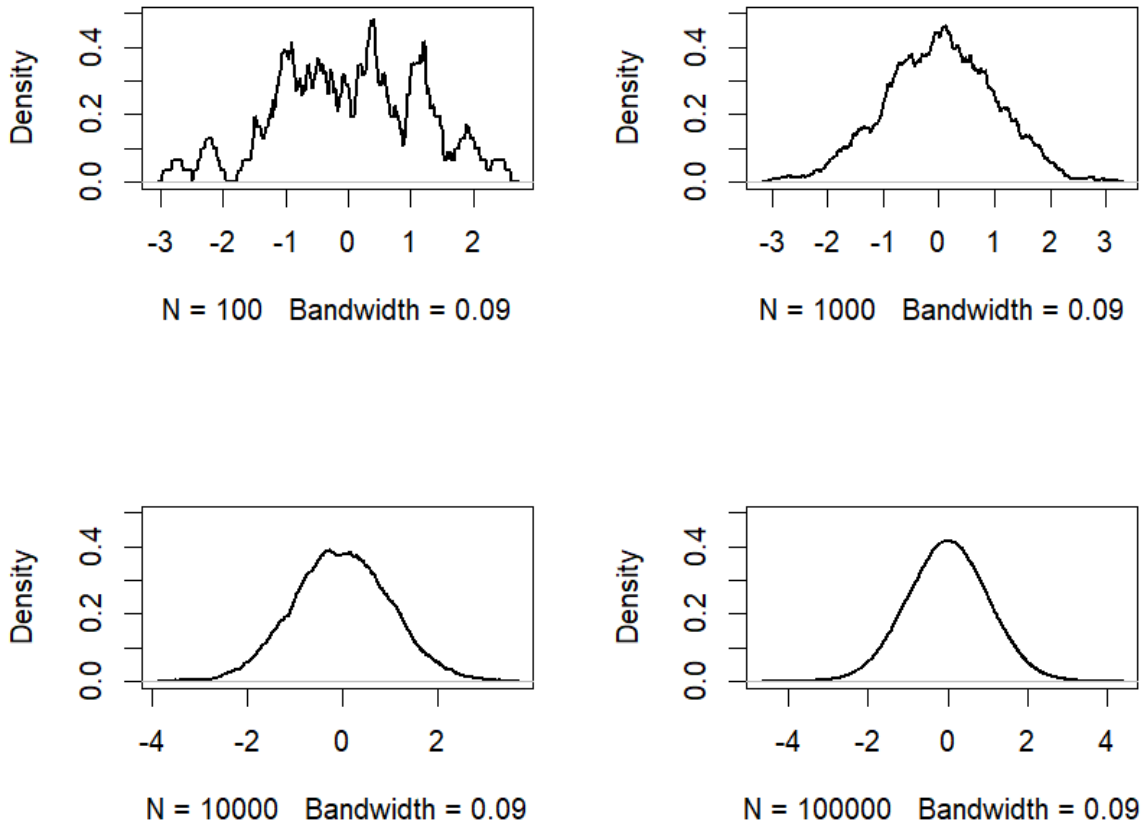
(c) Se $nh \to \infty$: $\mathrm{Var}\left[\hat{\mathrm{f}}_{\mathrm{N}}(x;h)\right] \to 0$.

(d) Se $nh \to \infty$ e $h \to 0$ (fazendo $n \to \infty$), viés $\to 0$.
O que seria a situação ideal.



Estimador de Densidade “Naive”.

Problema 2: Para boa aproximação é preciso obter intervalos pequenos e n grande!



Estimador de Densidade "Naive" para diferentes valores de n.

Solução: dar pesos maiores para observações mais próximas de x
$\Rightarrow$ Estimadores do Tipo Núcleo ("Kernel").

2 Estimação de Densidades baseada em Núcleos (ou Janelas de Parzen)

- Note que

$$\begin{aligned}\hat{\mathrm{f}}_{\mathrm{N}}(x;h) &= \frac{1}{2nh}\sum_{i=1}^{n}\mathbf{1}_{\{x-h<X_i<x+h\}} \\ &= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}\mathbf{1}_{\{-1<\frac{x-X_i}{h}<1\}} \\ &= \frac{1}{nh}\sum_{i=1}^{n}\mathrm{w}\left(\frac{x-X_i}{h}\right)\end{aligned}$$

  em que w é uma uniforme em $(-1,1)$, i.e., $\mathrm{w}(z)=\frac{1}{2}\mathbf{1}_{\{-1<z<1\}}$.

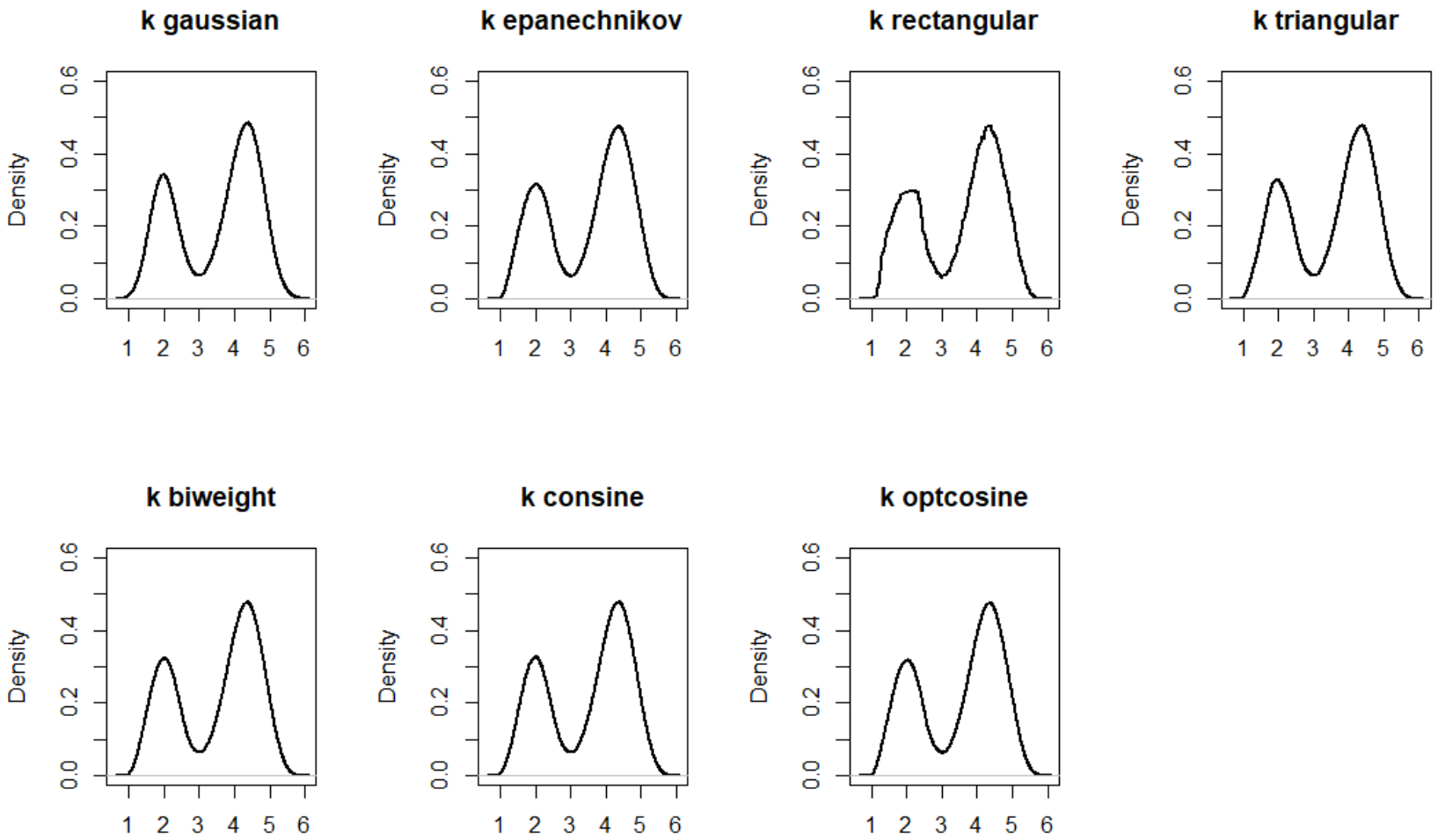
- $\mathrm{w}(.)$ atribui pesos iguais a todos os pontos $X_1,\ldots,X_n$.

- Substituindo $\mathrm{w}(.)$ por uma densidade arbitrária $\mathrm{k}(.)$ obtemos a função "kernel" que é simétrica e unimodal em 0.
- O estimador de densidade "Kernel" será então,

$$\hat{\mathrm{f}}(x;h) = \frac{1}{nh}\sum_{i=1}^{n}\mathrm{k}\left(\frac{x-X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{k}_h\left(x-X_i\right)$$

- A função `density` do `R` possui várias funções kernel: gaussian, rectangular, triangular, epanechnikov, biweight, cosine, optcosine. Webb (2011) cita as seguintes funções mais comuns

Kernel function	Analytic form, $K(x)$
Rectangular (or top hat)	$\frac{1}{2}$ for $\lvert x\rvert < 1$, 0 otherwise
Triangular	$1 - \lvert x\rvert$ for $\lvert x\rvert < 1$, 0 otherwise
Biweight (or Quartic)	$\frac{15}{16}(1-x^2)^2$ for $\lvert x\rvert < 1$, 0 otherwise
Normal (or Gaussian)	$\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$
Bartlett–Epanechnikov	$\frac{3}{4}(1-x^2/5)/\sqrt{5}$ for $\lvert x\rvert < \sqrt{5}$, 0 otherwise



Estimação de Densidade de `faithful$eruptions`[1] com diferentes funções Kernel.

[1] Ver faithful no Help do R.

Problema: Seleção do *bandwidth* h.

- O `R` e outras linguagens fazem esta escolha automaticamente, mas permitem a alteração.
- O cálculo de h é feito através de algumas medidas de erro.

ISE - Integrated Squared Error

$$\mathrm{ISE}[\hat{\mathrm{f}}(.,h)] = \int (\hat{\mathrm{f}}(x,h) - f(x))^2 dx$$

mede o erro total que é dependente da amostra.

MISE - Mean Integrated Squared Error

$$\mathrm{MISE}[\hat{\mathrm{f}}(.,h)] = \mathrm{E}[\mathrm{ISE}[\hat{\mathrm{f}}(.,h)]] = \int \mathrm{E}[(\hat{\mathrm{f}}(x,h) - f(x))^2] dx = \int \mathrm{E}[\mathrm{MSE}[\hat{\mathrm{f}}(x,h)]]$$

mede o erro médio que precisa ser minimizado.

AMISE - Asymptotic MISE

$$\mathrm{AMISE}[\hat{\mathrm{f}}(.,h)] = \frac{1}{4}\mu_2^2(\mathrm{k})\mathrm{R}(\mathrm{f}'')h^4 + \frac{\mathrm{R}(\mathrm{k})}{nh}$$

expansão assintótica de MISE, que permite obter h (por ser limitada), sendo $\mu_2^2(\mathrm{k}) = \int z^2\mathrm{k}(z)dz$ e $\mathrm{R}(g) = \int g(x)^2dx$.

- Assim,

$$h_{\mathrm{AMISE}} = \left[\frac{\mathrm{R}(\mathrm{k})}{\mu_2^2(\mathrm{k})\mathrm{R}(\mathrm{f}'')n}\right]^{1/5}$$

- Como as outras medidas MISE depende da curvatura da densidade desconhecida $f(x)$. A ideia é assumir $f \sim N(\mu, \sigma^2$ e obter estimativas de σ. Para o kernel normal,

$$\hat{h} = \left(\frac{4}{3}\right)^{1/5} n^{-1/5}\hat{\sigma} \approx 1.06n^{-1/5}\hat{\sigma}.$$

- O `R` possui a função *bw.nrd* para obter $\hat{h}$ e outras funções como `bw.ucv` para obtenção por validação cruzada (LOOCV) - ver `h_otimo` em `EDNP.R`.
- Mais detalhes podem ser obtidos em García-Português (2024) e DasGupta, A. (2008) Asymptotic Theory of Statistics and Probability. Springer, capítulo 22.

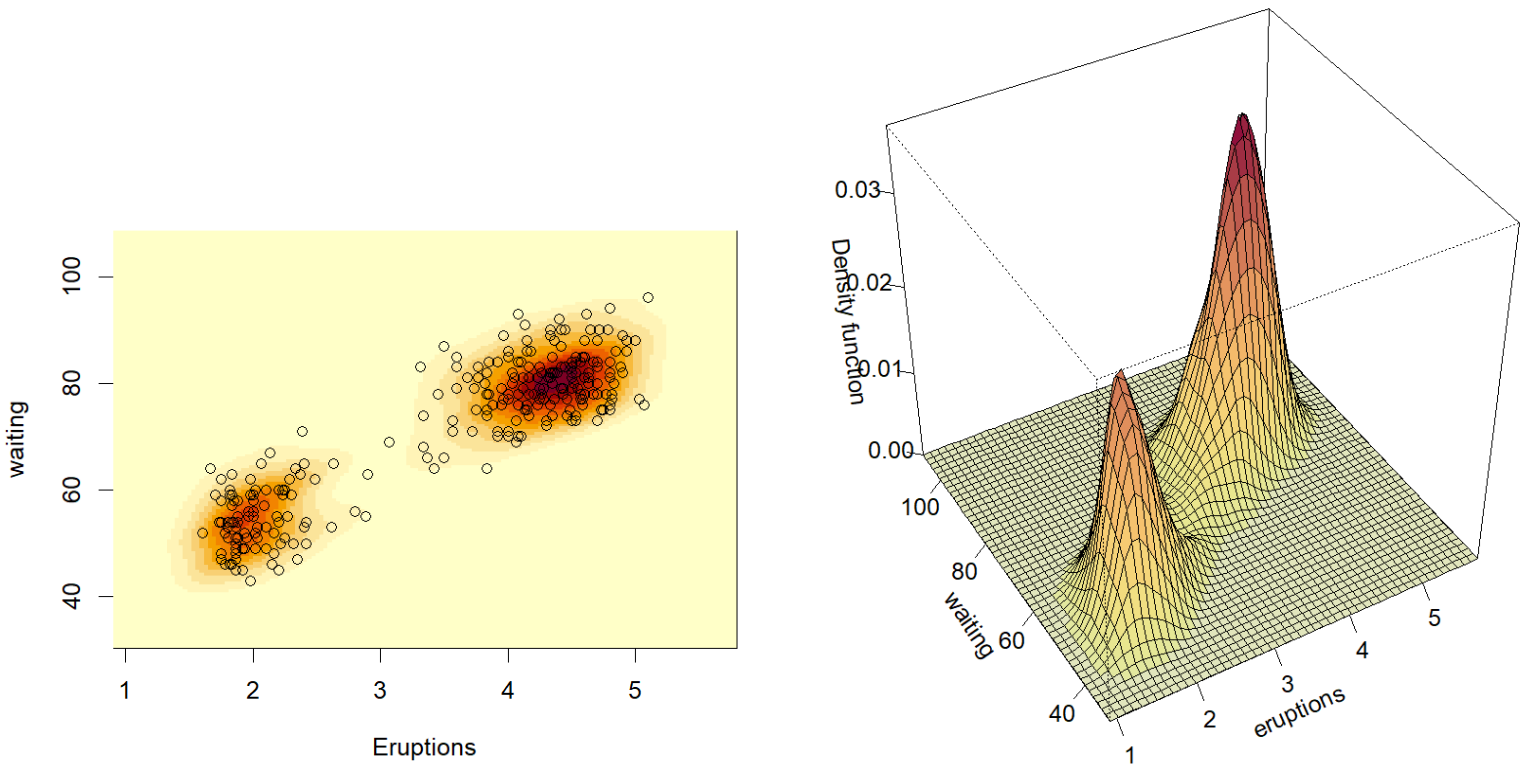
Observação: Estimação de Densidades Multivariadas

- Considere $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \in \Re^p$.

$$\hat{\mathrm{f}}(\mathbf{x}; \mathrm{H}) = \frac{1}{n|\mathrm{H}|^{1/2}} \sum_{i=1}^{n} \mathrm{k}(\mathrm{H}^{-1/2}(\mathbf{x} - \mathbf{X}_i))$$

  é geralmente simétrica e unimodal em $\mathbf{0}_{1\times p}$.
- H é a matriz de intervalos ("bandwith"), $p \times p$, simétrica e positiva definida.
- Se $p = 1$, $\mathrm{H} = h^2$.
- Um kernel multivariado comum é $\mathrm{k}(\mathbf{z}) = \phi(\mathbf{z}; \mathbf{0}, \mathbf{I}_p)$ da normal multivariada.
- O pacote `ks` do `R` apresenta funções para estimação de densidade multivariada.

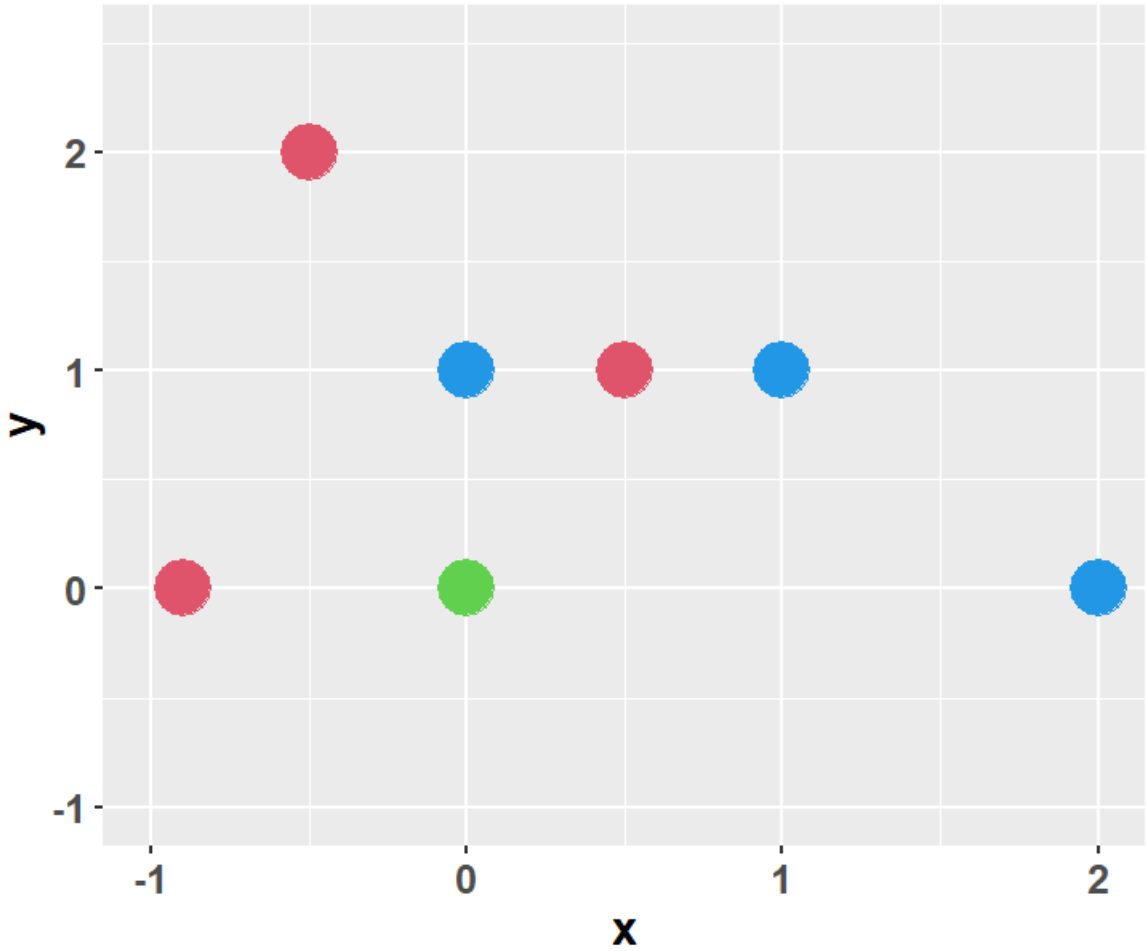
Faithful data Eruptions x Waiting Time



Estimação de Densidade Bivariada de `faithful` data.

3 Método dos k-Vizinhos mais Próximos (kNN)

Problema: Considere o seguinte conjunto de pontos, sendo $\mathbf{x} = (x, y) \in \omega_1$ se ponto azul e $\mathbf{x} = (x, y) \in \omega_2$ se ponto vermelho. Queremos classificar o ponto verde.



- Dado um ponto x_0, encontramos os k pontos de treinamento $x_{(r)}, r = 1, \ldots, k$ mais próximos de x_0 e classificamos usando a classe de maior proporção entre os k vizinhos.
- No caso de empate, a escolha é feita aleatoriamente.
- A medida de distância mais comum é a Distância Euclidiana, i.e.,

$$d_{(i)} = ||x_{(i)} - x_0||_2$$

- Geralmente as características (variáveis em análise) são padronizadas para evitar influência da escala de mensuração.

Exemplo 2.7, James et al. (2021)

Ver pag. 54 de James et al. (2021) e programa `EDNP`.

Exemplo 13.3.2, Hastie et al. (2017)

Ver seção 13.3 de Hastie et al. (2017) e programa `EDNP`.

> The STATLOG project (Michie et al., 1994) used part of a LANDSAT image as a benchmark for classification (82×100 pixels). Figure 13.6 shows four heat-map images, two in the visible spectrum and two in the infrared, for an area of agricultural land in Australia. Each pixel has a class label from the 7-element set $\mathcal{G} = \{$*red soil, cotton, vegetation stubble, mixture, gray soil, damp gray soil, very damp gray soil*$\}$, determined manually by research assistants surveying the area. The lower middle panel shows the actual land usage, shaded by different colors to indicate the classes. The objective is to classify the land usage at a pixel, based on the information in the four spectral bands.

Exemplo 13.3.2, Hastie et al. (2017) - continuação

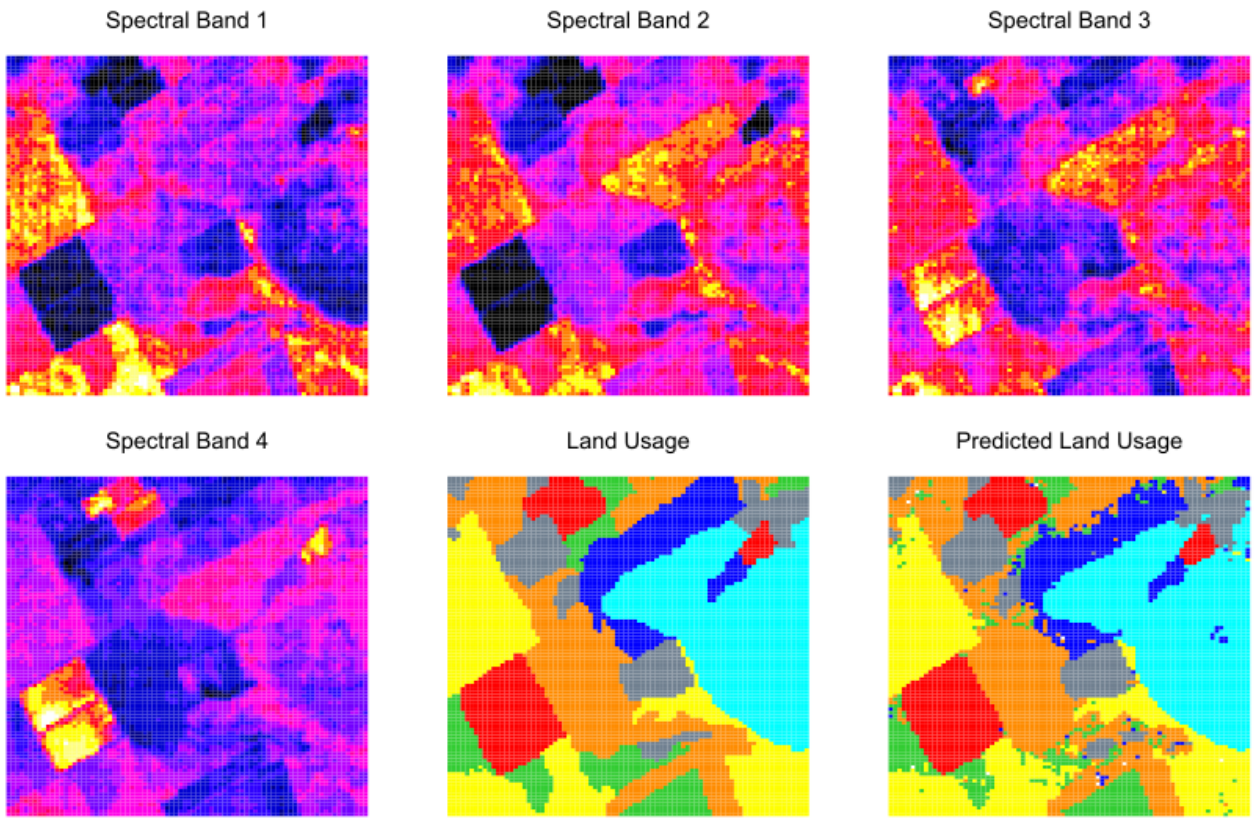


FIGURE 13.6. *The first four panels are LANDSAT images for an agricultural area in four spectral bands, depicted by heatmap shading. The remaining two panels give the actual land usage (color coded) and the predicted land usage using a five-nearest-neighbor rule described in the text.*

Validação

- **Índice Ajustado de Rand**: é um procedimento que permite comparar a classificação com o verdadeiro valor.
- Esta medida é utilizada também como validação em procedimentos da agrupamento, permitindo comparar o agrupamento com algum critério externo, i.e., um agrupamento prévio ou resultado padrão.

  Esta é uma medida de concordância entre dois critérios.
  Classificações/Agrupamentos com elementos nas mesmas classes/grupos tem valor esperado $\mathrm{ARI} \approx 1$ e agrupamentos em que todos os elementos mudam de grupos tem valor esperado $\mathrm{ARI} \approx 0$.

  O procedimento está disponível no pacote `MCLUST` do `R`.

Referências:

- Hubert, L. e Arabie, P. (1985) Comparing Partitions. *Journal of Classification*, Springer-Verlag.

- Yeung, K.Y. e Ruzzo, W.L. (2001) *The Adjusted Rand Index*. Texto técnico.

Iris Data

Iris é um género de plantas com flor, muito apreciado pelas suas diversas espécies, que ostentam flores de cores muito vivas. São, vulgarmente, designadas como lírios (Wikipédia).

Problema: classificar três tipos diferentes de flores Iris, chamadas setosa, versicolor e virginica. As características disponíveis são: largura e comprimento da sépala (parte da flor que dá sustentação a pétala) e pétala ($p = 4$).

(a) (b) (c)

Três tipos de flores Iris:

(a) setosa, (b) versicolor, (c) virginica. Fonte: Murphy, K.P. *Machine Learning*, 2012.

Ver `EDNP.R` para knn, incluindo validação cruzada.

Observações

- Pacotes `R`: `class` e `caret`.
- A técnica kNN adaptativa tenta corrigir a degradação de performance do kNN quando em altas dimensões, devido as grandes distâncias de alguns pontos. O algoritmo **DANN** (*Discriminant Adaptative Nearest-Neighbor* é descrito na Seção 13.4 de Hastie et. al (2017) e pode ser utilizado com o pacote `dann` do `R`.
- Os laboratórios da Seção 4.76 de James et al. (2021) e James et al. (2023) apresentam exemplos da técnica kNN utilizando `R` e `Python`, respectivamente.
- Para variáveis na escala ordinal, alterar a escala para

$$\frac{i-1/2}{M}, \quad i = 1, \ldots, M$$

e tratar como variáveis quantitativas.

Observações

- Para variáveis na escala nominal e binárias, utilizar medidas baseadas na respectiva tabela de contigências, i.e.,

	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	n

Medida	Similaridade	Intervalo
Jaccard	a/(a+b+c)	[0,1]
Sokal-Sneath	2(a+d)/(2(a+d)+b+c)	[0,1]
Rogers-Tanimoto	(a+d)/(a+2(b+c)+d)	[0,1]
Russell-Rao	a/n	[0,1]
Sokal-Michener	(a+d)/n	[0,1]
Dice	a/(2a+b+c)	[0,1/2]
Kulzinsky	a/(b+c)	[0,∞]

ver pacote `proxy` do `R`.

- Para dados em escala mista (cuidado), utilizar distâncias como a de Gower - ver pacote `gower`.

4 Naive Bayes (Bayes Ingênuo)

- Considere novamente o Teorema de Bayes,

$$\mathrm{p}(\omega_j|\mathbf{x}) = \frac{\mathrm{p}(\mathbf{x}|\omega_j)\mathrm{p}(\omega_j)}{\sum_{i=1}^{c} \mathrm{p}(\mathbf{x}|\omega_i)\mathrm{p}(\omega_i)}.$$

- $\mathrm{p}(\mathbf{x}|\omega_j)$ são as funções a priori e exigem algumas suposições para estimação, como no caso de LDA e QDA em que supõe-se normalidade.
- $\mathrm{p}(\omega_j)$ podem ser estimados pelas frequências relativas de cada classe, na amostra de treinamento.
- A técnica procura ajudar na solução de problemas encontrados em outros procedimentos mais complexos através da estimação das funções densidade a posteriori $\mathrm{p}(\omega_j|\mathbf{x})$ de forma simples ("ingênua").

Problemas: algumas situações podem complicar a análise discriminante.

- **Regressão logística** envolve a modelagem de $\mathrm{p}(w_j|\mathbf{x})$ através da função logística. Mas esta solução encontra problemas de estimação em situações como:
  - separação completa ou quase completa das classes.
  - pequenas amostras.
  - mais de duas classes exigindo n grande.

- **Discriminante Linear (LDA) e Quadrática (QDA)** assumem $\mathbf{X}_k \sim N_p$.

  A LDA assume $\mathbf{X}_k \sim \mathrm{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ e aloca $\mathbf{x}$ a classe em que

$$\mathbf{x}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log \mathrm{p}(\omega_k)$$

  é máximo.

  A QDA assume $\mathbf{X}_k \sim \mathrm{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ e aloca $\mathbf{x}$ a classe em que

$$-\frac{1}{2}\mathbf{x}^\mathsf{T}\boldsymbol{\Sigma}_k^{-1}\mathbf{x} + \mathbf{x}^\mathsf{T}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\mathsf{T}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log \mathrm{p}(\omega_k)$$

  é máximo.

- $\mathrm{p}(\mathbf{x}|\omega_j)$ é uma função densidade p-dimensional para uma observação $\mathbf{x}$ na classe j, $j = 1, \ldots, c$.
- Na técnica Naive Bayes, a suposição de normalidade multivariada é substituída por

$$\mathrm{p}(\mathbf{x}|\omega_j) = \mathrm{p}(x_1|\omega_j)\ \mathrm{p}(x_2|\omega_j)\ \ldots\ \mathrm{p}(x_p|\omega_j)$$

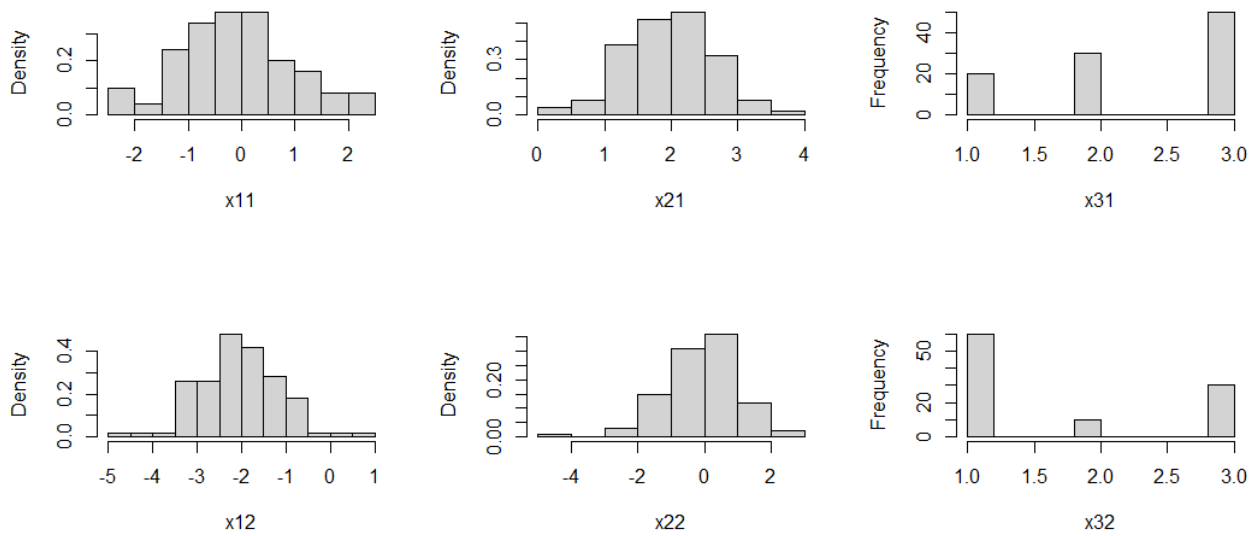
  em que os preditores são independentes em cada classe.
- Esta suposição é feita por conveniência e, apesar de não realistica, produz resultados satisfatórios em muitas ocasiões em que n não é muito maior que p.
- Naive Bayes resulta em aumento de viés e redução de variância dos estimadores, implicando em *MSE* razoável.
- O Teorema de Bayes será agora,

$$\mathrm{p}(\omega_j|\mathbf{x}) = \frac{\mathrm{p}(x_1|\omega_j)\ \mathrm{p}(x_2|\omega_j)\ \ldots\ \mathrm{p}(x_p|\omega_j)\ \mathrm{p}(\omega_j)}{\sum_{i=1}^{c} \mathrm{p}(x_1|\omega_j)\ \mathrm{p}(x_2|\omega_j)\ \ldots\ \mathrm{p}(x_p|\omega_j)\ \mathrm{p}(\omega_j)}$$

- Para estimar $\mathrm{p}(x_i|\omega_j)$ podem ser utilizados estimadores de densidade não paramétricos ou até a distribuição normal univariada.

Naive Bayes

Exemplo baseado em James et al. (págs. 156 e 157) em que $c = 2$ (ω_1, ω_2) e $p = 3$.
As variáveis apresentam o seguinte comportamento nas classes

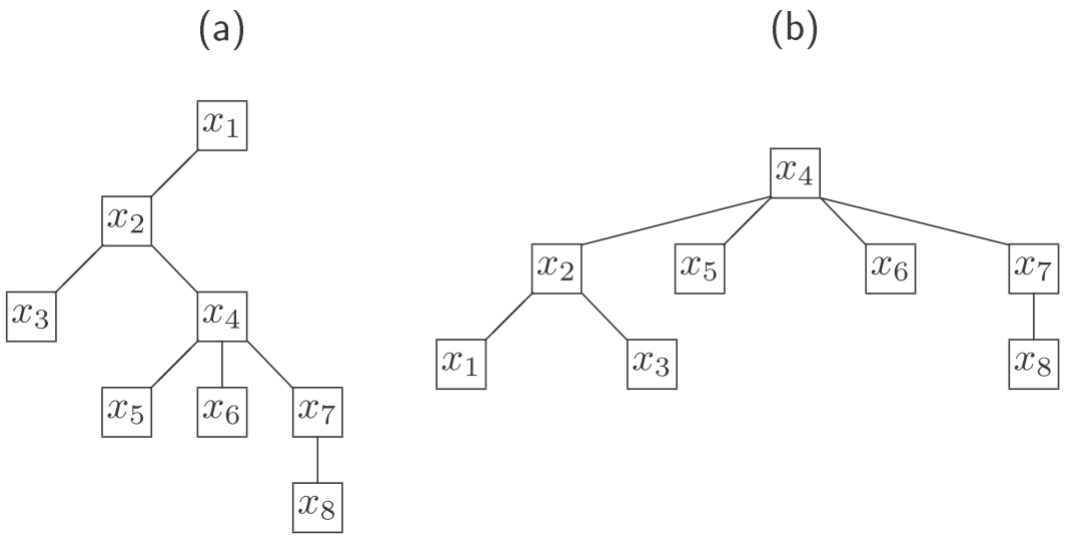


xij representa x_i na classe j.

O objetivo é classificar uma nova observação $x_0 = (0.4, 1.5, 1)^T$.
Ver `EDNP.R`.

Observação

- Em **Redes Bayesianas** são consideradas apenas as probabilidades condicionais que "podem" existir no modelo para estimar $\mathrm{p}(\omega_j|\mathbf{x})$.
- Um exemplo apresentado por Webb (2011) considera oito variáveis e grafo



Assim,

$$\mathrm{p}(\mathbf{x}) = \mathrm{p}(x_4)\ \mathrm{p}(x_1|x_2)\ \mathrm{p}(x_3|x_2)\ \mathrm{p}(x_2|x_4)\ \mathrm{p}(x_5|x_4)\ \mathrm{p}(x_6|x_4)\ \mathrm{p}(x_7|x_4)\ \mathrm{p}(x_8|x_7)$$

Comentários

- Um grafo é um par (V, E) de V vértices e E conexões entre vértices ("Edges").
- Um grafo direto é um grafo em que E são pares ordenados, i.e., se $(\alpha, \beta) \in \mathrm{E}$, então $(\beta, \alpha) \notin \mathrm{E}$ para os vértices α e β.
- Um grafo direto acíclico (DAG) é um grafo direto sem ciclos, i.e., não existe um caminho $\alpha_1 \to \alpha_2 \to \ldots \to \alpha_1$ para qualquer vértice α_1.
- Redes Bayesianas são DAG.

5 Densidades como Cópulas

- Referências:

  - Nagler, T. (2017) kdecopula: An R Package for the Kernel Estimation of Bivariate Copula Densities. Journal of Statistical Software, v. 84, Issue 7.

  - Provost, S.B. e Zang, Y. (2024) Nonparametric Copula Density Estimation Methodologies. Mathematics, 12, 398.

  http://doi.org/10.3390/math12030398

  Apresenta cópulas bivariadas e procedimentos de estimação.

- Cópulas permitem representar a distribuição conjuntade duas ou mais variáveis em termos de distribuições marginais e da estrutura de correlação.
- O efeito da dependência pode ser separado das contribuições de cada distribuição marginal.

- Suponha $\mathbf{u} = (u_1, \ldots, u_d)$ uma distribuição multivariada no espaço d-dimensional, em que a distribuição marginal de cada variável é $\mathrm{U}_i \sim \mathrm{U}(0,1), i = 1, \ldots, d$.
- Então a função cópula é

$$\mathrm{C}(\mathrm{U}_1, \ldots, \mathrm{U}_d) = \mathrm{p}(\mathrm{U}_1 \leq u_1, \ldots, \mathrm{U}_d \leq u_d).$$

- **Teorema de Skylar** (1959): Seja $\mathrm{H}(x_1, \ldots, x_d)$ a função distribuição conjunta acumulada das variáves aleatórias $x_1, \ldots, x_d$ com funções distribuição marginais (contínuas) $\mathrm{F}(x_1), \ldots, \mathrm{F}(x_d)$, respectivamente. Existe uma única cópula multivariada $\mathrm{C}(., \ldots, .) : \mathbf{1}^d \to \mathbf{1}$, tal que

$$\mathrm{H}(x_1, \ldots, x_d) = \mathrm{C}(\mathrm{F}(x_1), \ldots, \mathrm{F}(x_d))$$

em que $\mathrm{C}(., \ldots, .)$ é a distribuição conjunta acumulada com marginais uniformes.

- **Construção da Cópula**:
  $\mathrm{C}(x_1, \ldots, x_d) = \mathrm{H}(\mathrm{F}^{-1}(x_1), \ldots, \mathrm{F}^{-1}(x_d))$ é uma cópula d-dimensional com

$$\mathrm{F}^{-1}(u_i) = \inf\{x_i | \mathrm{F}(x_i) \geq u_i\}, \ \forall u_i \in (0, 1).$$

  A função densidade (derivada de H) é

$$\mathrm{h}(x_1, \ldots, x_d) = \mathrm{C}(\mathrm{F}(x_1), \ldots, \mathrm{F}(x_d)) \prod_{i=1}^{d} \mathrm{f}_i(x_i)$$

  em que $\mathrm{f}_i(x_i)$ é a função densidade marginal da i-ésima variável.
  A densidade da cópula (derivada de C) é

$$\mathrm{c}(u_1, \ldots, u_d) = \frac{\partial \mathrm{C}(\mathrm{U}_1, \ldots, \mathrm{U}_d)}{\partial u_1 \ldots \partial u_d}.$$

  e sua estimativa também pode ser utilizada para discriminação pela regra de Bayes.
- Pacotes `R`: `ks` com função `kcopula`, `np` com função `npcopula` e `kdecopula`.

Estimação Paramétrica de Densidades

No caso paramétrico assumimos que a densidade condicional $\mathrm{p}(\mathbf{x}|\omega_j)$ tem forma conhecida, mas com parâmetros $(\boldsymbol{\theta}_j)$ desconhecidos.

Abordagens

(a) Estimação frequentista:

$$\mathrm{p}(\mathbf{x}|\omega_j) = \mathrm{p}(\mathbf{x}|\hat{\boldsymbol{\theta}}_j)$$

com base na amostra observada $\mathrm{S}_j = \{\mathbf{x}_1^j, \ldots, \mathbf{x}_{n_j}^j\}, \mathbf{x}_i^j \in \Re^p$ para ω_j. Um exemplo seria EMV de $\boldsymbol{\theta}_j$.

(b) Estimação Bayesiana.

$$\mathrm{p}(\mathbf{x}|\omega_j) = \int \mathrm{p}(\mathbf{x}|\boldsymbol{\theta}_j)\ \mathrm{p}(\boldsymbol{\theta}_j|\mathrm{S}_j)\ d\boldsymbol{\theta}_j$$

em que $\mathrm{p}(\boldsymbol{\theta}_j|\mathrm{S}_j)$ é a densidade a posteriori de $\boldsymbol{\theta}_j$ baseada na priori $\mathrm{p}(\boldsymbol{\theta}_j)$ e nos dados S_j.

5 Classificadores Gaussianos

- Considere

$$\mathrm{p}(\mathbf{x}|\omega_j) = \mathrm{N}(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

sendo $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathsf{T}}$, $-\infty < x_i < \infty$, $i = 1, \ldots, p$ e $\boldsymbol{\Sigma}_j$ positiva definida, $j = 1, \ldots, c$.

Então,

$$\log[\mathrm{p}(\mathbf{x}|\omega_j)] = \log\left[\frac{\mathrm{p}(\mathbf{x}|\omega_j)\mathrm{p}(\omega_j)}{\mathrm{p}(\mathbf{x})}\right] = \log[\mathrm{p}(\mathbf{x}|\omega_j)] + \log[\mathrm{p}(\omega_j)] - \log[\mathrm{p}(\mathbf{x})]$$

- A LDA assume $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$ e aloca $\mathbf{x}$ a classe em que

$$\mathbf{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \log \mathrm{p}(\omega_j)$$

é máximo.

- A QDA assume $\boldsymbol{\Sigma}_k$ para a classe k e aloca $\mathbf{x}$ a classe em que

$$-\frac{1}{2}\mathbf{x}^{\mathsf{T}}\boldsymbol{\Sigma}_j^{-1}\mathbf{x} + \mathbf{x}^{\mathsf{T}}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^{\mathsf{T}}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\log|\boldsymbol{\Sigma}_j| + \log \mathrm{p}(\omega_j)$$

é máximo.

- Os estimadores de Máxima Verossilhança de $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}_j$ são
  $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ e $\hat{\boldsymbol{\Sigma}}_j = \frac{n-1}{n}\,\mathbf{S}_j$, com

$$\mathbf{S}_j = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^{\mathsf{T}}.$$

  As estatísticas $\bar{\mathbf{x}}$ e $\mathbf{S}_j$ são suficientes (Johnson e Wichern, 2007)

Problema: Calculo de $\hat{\boldsymbol{\Sigma}}_j^{-1}$ se $\hat{\boldsymbol{\Sigma}}_j$ singular!

Soluções:

(a) A suposição em LDA, $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$, simplifica o problema. Neste caso, $\hat{\boldsymbol{\Sigma}}$ e uma estimativa ponderada de $\hat{\boldsymbol{\Sigma}}_j$.

(b) Estimador Naive Bayes
  $\hat{\boldsymbol{\Sigma}}_j \Rightarrow$ dados independentes ou

$$\mathrm{p}(\mathbf{x}|\omega_j) = \prod_{\ell=1}^{p}\mathrm{N}(\mathrm{x}_\ell; \hat{\mu}_{j\ell}, \hat{\sigma}^2_{j\ell})$$

  em que $\hat{\sigma}^2_{j\ell} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \hat{\mu}_{j\ell})^2$.

(c) Realizar a discriminante numa projeção (PCA) em que $\hat{\boldsymbol{\Sigma}}_j$ não seja singular.

Soluções:

(d) Regularização da função discriminante.

λ : parâmetro de complexidade.

γ : parâmetro de regularização ("shrinkage")

$$\hat{\mathbf{\Sigma}}_j^{\lambda} = \frac{(1-\lambda)n_j\hat{\mathbf{\Sigma}}_j + \lambda n\hat{\mathbf{\Sigma}}}{(1-\lambda)n_j + \lambda n}$$

com $0 \leq \lambda \leq 1$. Note que

se $\lambda = 0 \Rightarrow \hat{\mathbf{\Sigma}}_j^{\lambda} = \hat{\mathbf{\Sigma}}_j$ (discriminante quadrático)

se $\lambda = 1 \Rightarrow \hat{\mathbf{\Sigma}}_j^{\lambda} = \hat{\mathbf{\Sigma}}$ (discriminante linear)

Acrescentando γ, temos

$$\hat{\mathbf{\Sigma}}_j^{\lambda,\gamma} = (1-\gamma)\hat{\mathbf{\Sigma}}_j^{\lambda} + \gamma c_j(\lambda)\mathbf{I}_p$$

sendo $\mathbf{I}_p$ a matriz identidade e

$$c_j(\lambda) = \frac{\mathrm{tr}(\hat{\mathbf{\Sigma}}_j^{\lambda})}{p} \quad \text{(média dos autovalores de } \hat{\mathbf{\Sigma}}_j^{\lambda}\text{)}$$

Obervação: A escolha de λ e γ se dá por validação cruzada.

Soluções:

(e) Estimativa robusta de $\boldsymbol{\Sigma}_j^{\lambda}$, quando as amostras são pequenas.

$$\tilde{\boldsymbol{\Sigma}}_j^{\lambda} = \frac{(1-\lambda)\tilde{\boldsymbol{\Sigma}}_j + \lambda \sum_{\ell_1}^{c} \tilde{\boldsymbol{\Sigma}}_\ell}{(1-\lambda) W_j + \lambda \sum_{\ell=1}^{c} W_\ell}$$

em que

$$\tilde{\boldsymbol{\Sigma}}_j = \sum_{i=1}^{n} z_{ji} w_i (\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_j)(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_j)^{\mathsf{T}}$$

$$\tilde{\boldsymbol{\mu}}_j = \sum_{i=1}^{n} z_{ji} w_i \frac{\mathbf{x}_i}{W_j} \quad \text{e} \quad W_j = \sum_{i=1}^{n} z_{ji} w_i$$

sendo $0 \leq w_i \leq 1$ os pesos associados a cada observação e $z_{ji} = 1$ se $\mathbf{x}_i \in w_j$ e $z_{ji} = 0$ caso contrário.

Observação: Se $w_i = 1$ para todo i, $\tilde{\boldsymbol{\Sigma}}_j = \hat{\boldsymbol{\Sigma}}_j$.

Referências:

- Scrucca, L.; Fraley, C.; Murphy, T.B.; Raftery, A.E. (2023) **Model-Based Clustering, Classification, and Density Estimation Using mclust in R**. CRC Press.
- Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). **Model-Based Clustering and Classification for Data Science: With Applications in R**. Cambridge University Press.
- McLachlan, G. J. and Peel, D. (2000). **Finite Mixture Models**. Wiley.
- Bishop, C. (2006). **Pattern Recognition and Machine Learning**. Springer. (Capítulo 9)
- Titterington, D. M., Smith, A. F., and Makov, U. E. (1985). Statistical Analysis of Finite Mixture Distributions. John Wiley & Sons.
- McLachlan, G. J. and Basford, K. E. (1988). Mixture Models: Inference and Applications to Clustering. Marcel Dekker.
- Frühwirth-Schnatter, S. (2006). Finite Mixture and Markov Switching Models. Springer.
- McNicholas, P. D. (2016). Mixture Model-Based Classification. CRC Press.

- Uma mistura de distribuições é uma distribuição de probabilidades obtida de uma combinação linear convexa de funções densidade de probabilidade (componentes da mistura).
- São algoritmos fuzzy, i.e., baseados em alocação probabilística hierárquica. - A forma geral de um Modelo de Mistura Finita para uma vetor aleatórico **x** p-dimensional é

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g) \tag{1}$$

onde, G é o número de componentes da mistura, π_g é a probabilidade de uma observação ter sido gerada pela g-ésima componente e $f_g(\dot{)}$ é a densidade da g-ésima componente da mistura.

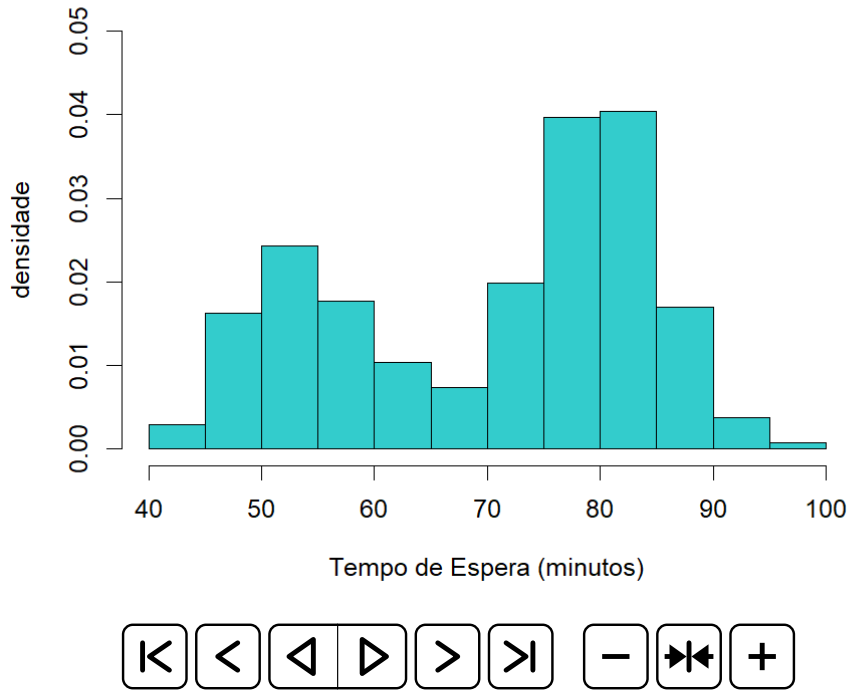
- Note que $\pi_g > 0$ e $\sum_{g=1}^{G} \pi_g = 1$, implicando que o espaço de parâmetros a ser estimado inclui $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{G-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$.

- Utilizados quando um única densidade não é capaz de representar os dados em estudo.
- Estes modelos são úteis em problemas superdimensionados e com amostras pequenas[2].

[2] von Borries, G. (2008) Partition Clustering of High Dimensional Low Sample Size Data Based on P-Values. PhD thesis, Kansas State University.

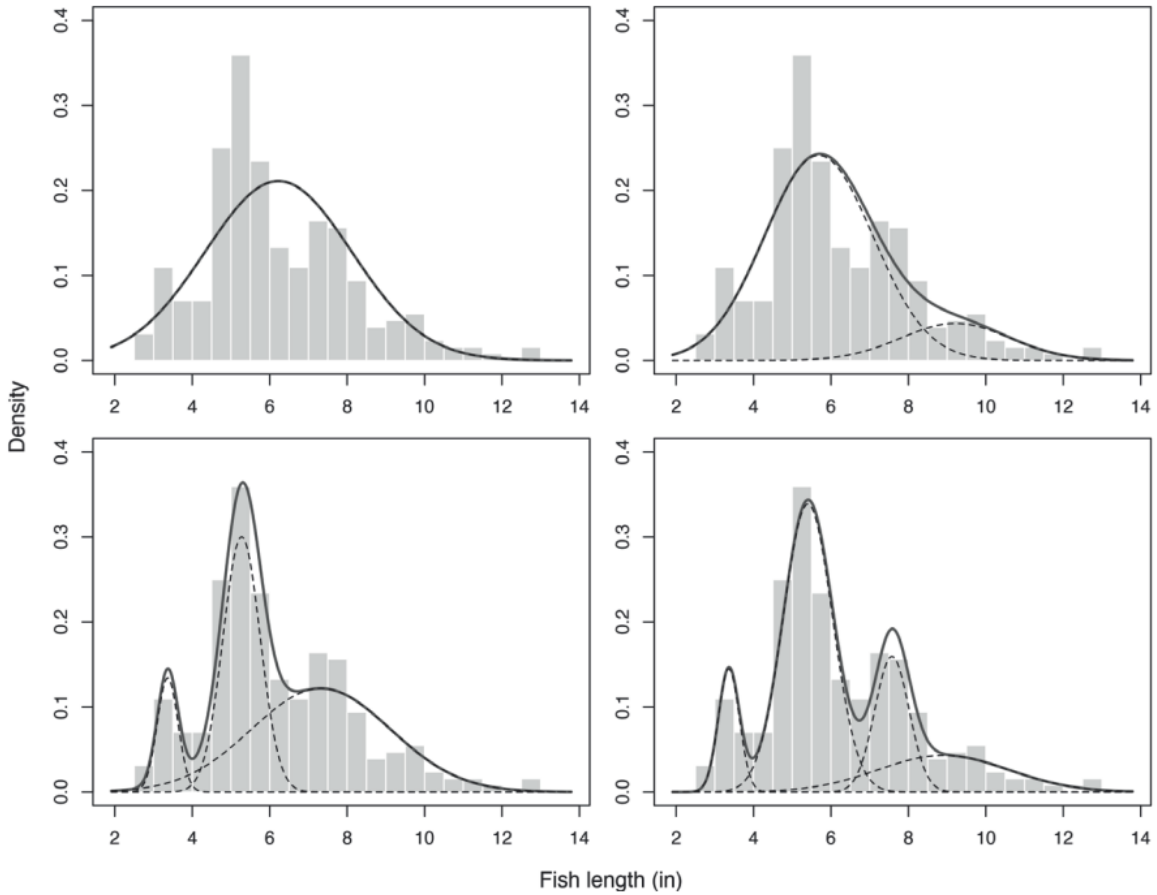
Justificativa

Considere os dados de tempo de espera do exemplo *Old Faithful Geyser Eruptions*. Qual modelo utilizar para estimar a densidade destes dados?



Exemplo: Comprimento de Peixes

Scrucca et al. (2023)



Estimação por Máxima Verossimilhança (MLE)

- Considere uma amostra aleatória $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

  A verossimilhança da mistura é

$$\mathcal{L}(\boldsymbol{\Psi}) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g) \right\}$$

  e a respectiva função log-verossimilhança

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g) \right\}$$

  com solução de máximo local obtida pela solução da equação de verossimilhança

$$\frac{\partial \ell(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}} = \mathbf{0}$$

- Problema: difícil solução, mesmo numericamente.
- Abordagem: reformulação do problema (dados incompletos) e aplicação do algoritmo EM (Expectation-Maximization).

Reformução do Problema

- Considere a realização de observações multivariadas $(\mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$, com $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})^{\mathsf{T}}$ variáveis latentes tais que

$$z_{ik} = \begin{cases} 1 & \text{se } \mathbf{x}_i \text{ pertence ao } g\text{-ésimo componente da mistura} \\ 0 & \text{caso contrário} \end{cases}$$

- Sob a suposição i.i.d. para $(\mathbf{x}_i, \mathbf{z}_i)$ a verossimilhança da mistura será

$$\mathcal{L}_c(\boldsymbol{\Psi}) = \prod_{i=1}^{n} f(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\Psi}) = \prod_{i=1}^{n} \mathrm{p}(\mathbf{z}_i) f(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\Psi}).$$

- Assumindo ainda $\mathbf{z}_i$ i.i.d. da distribuição multinomial com probabilidades $(\pi_1, \ldots, \pi_G)$, tem-se

$$\mathbf{z}_i \propto \prod_{g=1}^{G} \pi_g^{z_{ig}} \qquad \text{e} \qquad f(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\Psi}) = \prod_{g=1}^{G} f_g(\mathbf{x}_i|\boldsymbol{\theta}_g)^{z_{ig}}$$

resultando em

$$\ell_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left\{ \log \pi_g + \log f_g(\mathbf{x}_i|\boldsymbol{\theta}_g) \right\}.$$

Algoritmo EM

Alterna Expectation e Maximization através dos seguintes passos:

1. Iniciar com $\mathrm{s} = 0$ e escolher valores iniciais para os parâmetros em $\boldsymbol{\Psi}^{(0)}$,

$$\mathrm{Q}(\boldsymbol{\Psi}^{(0)}) = \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}^{(s)} \left\{ \log \pi_g + \log f_g(\mathbf{x}_i; \boldsymbol{\theta}_g) \right\}$$

2. Expectation (E): estimar a componente latente $\hat{z}_{ig}^{(s)}$,

$$\hat{z}_{ig}^{(s)} = \mathrm{p}\left(z_{ig} = 1 | \mathbf{x}_i, \hat{\boldsymbol{\Psi}}^{(t)} \right) = \frac{\pi_g^{(t)} f_g(\mathbf{x}_i | \boldsymbol{\theta}_g^{(t)})}{\sum_{g=1}^{G} \pi_g^{(t)} f_g(\mathbf{x}_i | \boldsymbol{\theta}_g^{(t)})}$$

3. Maximization (M): atualizar as estimativas dos parâmetros

$$\boldsymbol{\Psi}^{(\mathrm{s}+1)} = \arg\max_{\boldsymbol{\Psi}} \mathrm{Q}(\boldsymbol{\Psi}^{(\mathrm{s})})$$

   No caso de modelos finitos, $\pi_g^{\mathrm{s}+1} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}^{(\mathrm{s})}}{n}$.

4. Se o critério de convergência não for atendido, fazer $\mathrm{s} = \mathrm{s} + 1$ e retornar ao passo (2).

Propriedades do Algoritmo EM

Vantagens:

- EM tende a converge para um máximo local/global. O log da verossimilhança cresce, a menos que um ponto estacionário tenha sido atingido.
- Em geral, o custo operacional e computacional é menor que maximização por métodos convencionais.
- Em modelos de mistura, é garantido que as probabilidades fiquem no intervalo $[0, 1]$.
- Em GMMs as matrizes de covariância não possuem autovalores negativos.

Desvantagens:

- Estimativas dos parâmetros são dependentes de valores iniciais.
- A convergência pode ser demorada.
- Problemas numéricos de implementação podem degradar algumas vantagens do procedimento.
- Não produz estimativas de variância-covariância das estimativas dos parâmetros. Métodos adicionais, como bootstrap, são necessários para corrigir esta deficiência.

Modelos de Mistura de Normais (GMM)

- Geralmente, as densidades $f_g(\mathbf{x}|\boldsymbol{\theta}_g)$ pertencem a mesma família de distribuições.
- Existe a possibilidade de $f_g(\mathbf{x}|\boldsymbol{\theta}_g)$ pertencer a diferentes famílias, como em modelos inflados de zeros.
- O uso de modelos gaussianos, em que $f_g(\mathbf{x}|\boldsymbol{\theta}_g) = \phi_g(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ é a mistura de normais multivariadas, funciona para muitas aplicações.
- O maior problema neste caso é a estimação das matrizes de variância-covariância $\boldsymbol{\Sigma}_g$.
- Diferentes características do modelo GMM (Gaussian Mixture Models) são controladas pela SVD (*Singular Value Decomposition*) de $\boldsymbol{\Sigma}_g$, em

$$\boldsymbol{\Sigma}_g = \lambda_g \mathbf{U}_g \mathbf{D}_g \mathbf{U}_g^{\mathsf{T}},$$

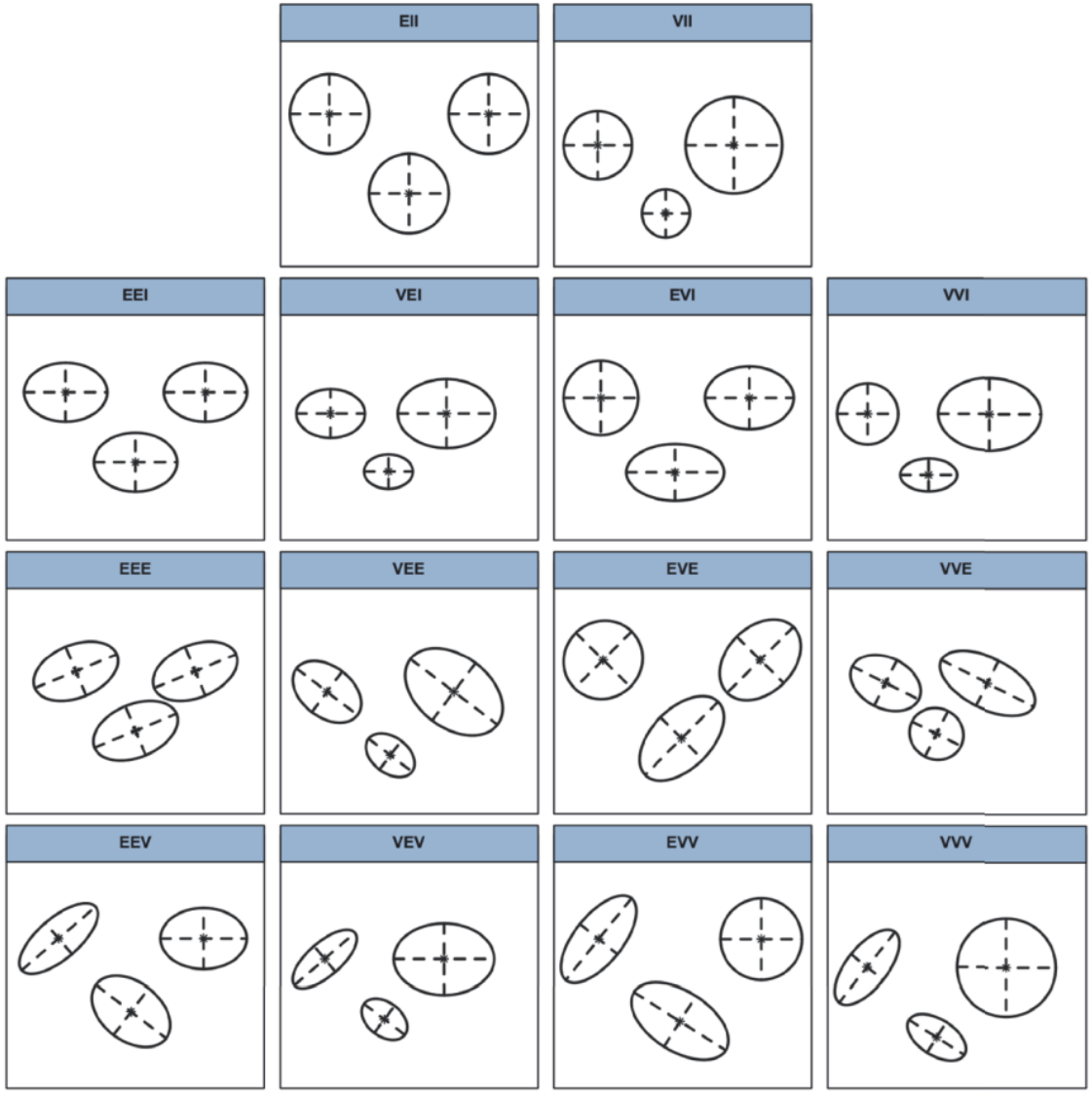
em que $\lambda_g = |\boldsymbol{\Sigma}_g|^{1/p}$ é um escalar de controle do volume, $\mathbf{D}_g$ é uma matriz diagonal de autovalores de $\boldsymbol{\Sigma}_g$, tal que $|\mathbf{D}_g| = 1$, e $\mathbf{U}_g$ uma matriz ortogonal de autovetores de $\boldsymbol{\Sigma}_g$.

Parametrizações de $\boldsymbol{\Sigma}_g$ (Scrucca et al., 2023)

Label	Model	Distribution	Volume	Shape	Orientation
EII	$\lambda \boldsymbol{I}$	Spherical	Equal	Equal	—
VII	$\lambda_k \boldsymbol{I}$	Spherical	Variable	Equal	—
EEI	$\lambda \boldsymbol{\Delta}$	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k \boldsymbol{\Delta}$	Diagonal	Variable	Equal	Coordinate axes
EVI	$\lambda \boldsymbol{\Delta}_k$	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k \boldsymbol{\Delta}_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda \boldsymbol{U} \boldsymbol{\Delta} \boldsymbol{U}^\top$	Ellipsoidal	Equal	Equal	Equal
VEE	$\lambda_k \boldsymbol{U} \boldsymbol{\Delta} \boldsymbol{U}^\top$	Ellipsoidal	Variable	Equal	Equal
EVE	$\lambda \boldsymbol{U} \boldsymbol{\Delta}_k \boldsymbol{U}^\top$	Ellipsoidal	Equal	Variable	Equal
VVE	$\lambda_k \boldsymbol{U} \boldsymbol{\Delta}_k \boldsymbol{U}^\top$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda \boldsymbol{U}_k \boldsymbol{\Delta} \boldsymbol{U}_k^\top$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k \boldsymbol{U}_k \boldsymbol{\Delta} \boldsymbol{U}_k^\top$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda \boldsymbol{U}_k \boldsymbol{\Delta}_k \boldsymbol{U}_k^\top$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k \boldsymbol{U}_k \boldsymbol{\Delta}_k \boldsymbol{U}_k^\top$	Ellipsoidal	Variable	Variable	Variable

- Nomeclatura do pacote `mclust` do `R`.
  `E` = igualdade, `V` = variável, `I` = matriz identidade.

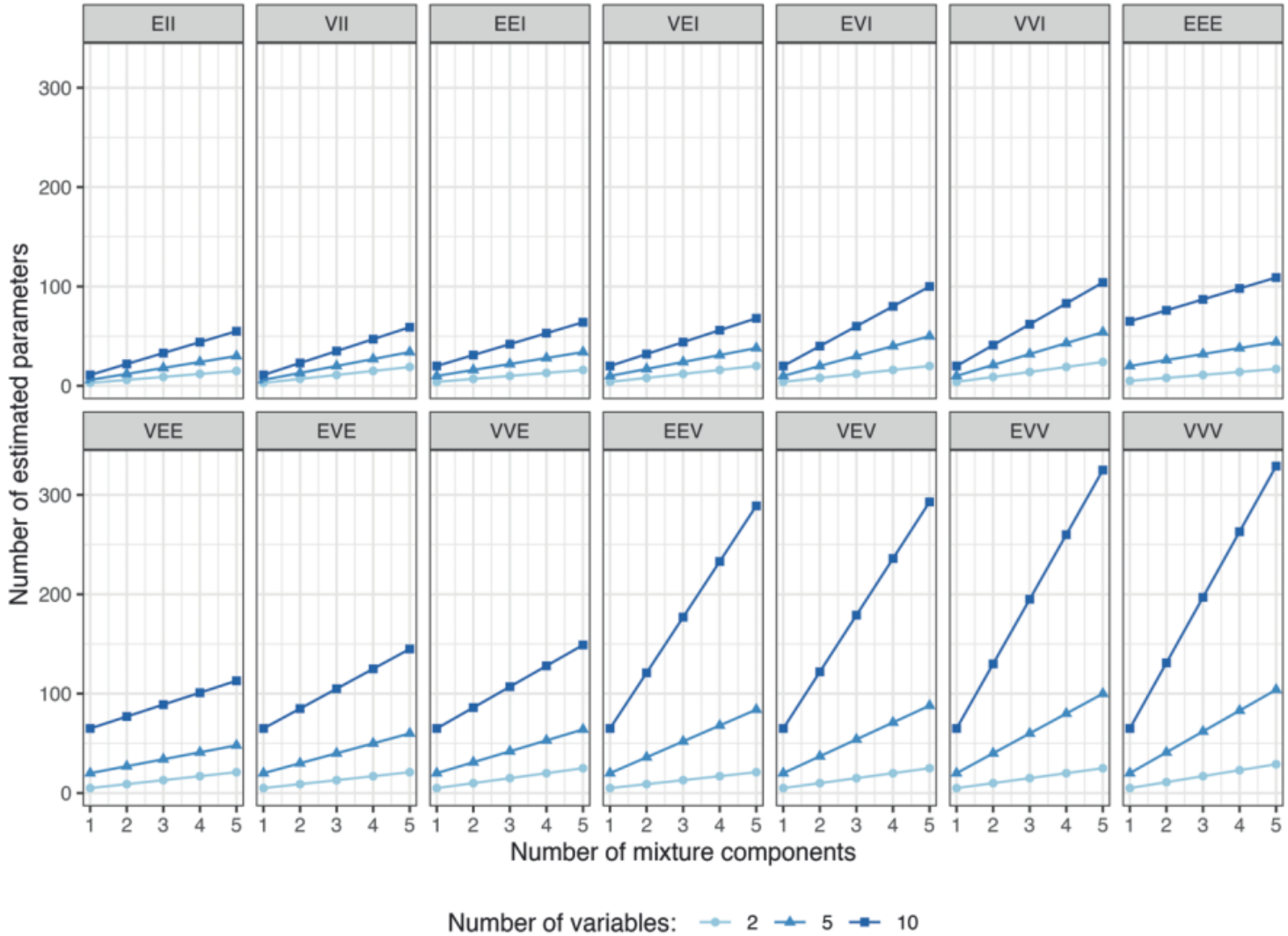
Representações das Parametrizações de $\mathbf{\Sigma}_g$ (Scrucca et al., 2023)



Número de parâmetros estimados para cada parametrização e indicação de procedimento em forma fechada (CF) ou iterativo (IP) - Scrucca et al., 2023.

Label	Model	Number of parameters	M-step
EII	$\lambda \boldsymbol{I}$	$(G-1)+Gd+1$	CF
VII	$\lambda_k \boldsymbol{I}$	$(G-1)+Gd+G$	CF
EEI	$\lambda \boldsymbol{\Delta}$	$(G-1)+Gd+d$	CF
VEI	$\lambda_k \boldsymbol{\Delta}$	$(G-1)+Gd+G+(d-1)$	IP
EVI	$\lambda \boldsymbol{\Delta}_k$	$(G-1)+Gd+1+G(d-1)$	CF
VVI	$\lambda_k \boldsymbol{\Delta}_k$	$(G-1)+Gd+G+G(d-1)$	CF
EEE	$\lambda \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^\top$	$(G-1)+Gd+1+(d-1)+d(d-1)/2$	CF
VEE	$\lambda_k \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^\top$	$(G-1)+Gd+G+(d-1)+d(d-1)/2$	IP
EVE	$\lambda \boldsymbol{U}\boldsymbol{\Delta}_k\boldsymbol{U}^\top$	$(G-1)+Gd+1+G(d-1)+d(d-1)/2$	IP
VVE	$\lambda_k \boldsymbol{U}\boldsymbol{\Delta}_k\boldsymbol{U}^\top$	$(G-1)+Gd+G+G(d-1)+d(d-1)/2$	IP
EEV	$\lambda \boldsymbol{U}_k\boldsymbol{\Delta}\boldsymbol{U}_k^\top$	$(G-1)+Gd+1+(d-1)+Gd(d-1)/2$	CF
VEV	$\lambda_k \boldsymbol{U}_k\boldsymbol{\Delta}\boldsymbol{U}_k^\top$	$(G-1)+Gd+G+(d-1)+Gd(d-1)/2$	IP
EVV	$\lambda \boldsymbol{U}_k\boldsymbol{\Delta}_k\boldsymbol{U}_k^\top$	$(G-1)+Gd+1+G(d-1)+Gd(d-1)/2$	CF
VVV	$\lambda_k \boldsymbol{U}_k\boldsymbol{\Delta}_k\boldsymbol{U}_k^\top$	$(G-1)+Gd+G+G(d-1)+Gd(d-1)/2$	CF

Número de parâmetros estimados como função do número de componentes - Scrucca et al., 2023.



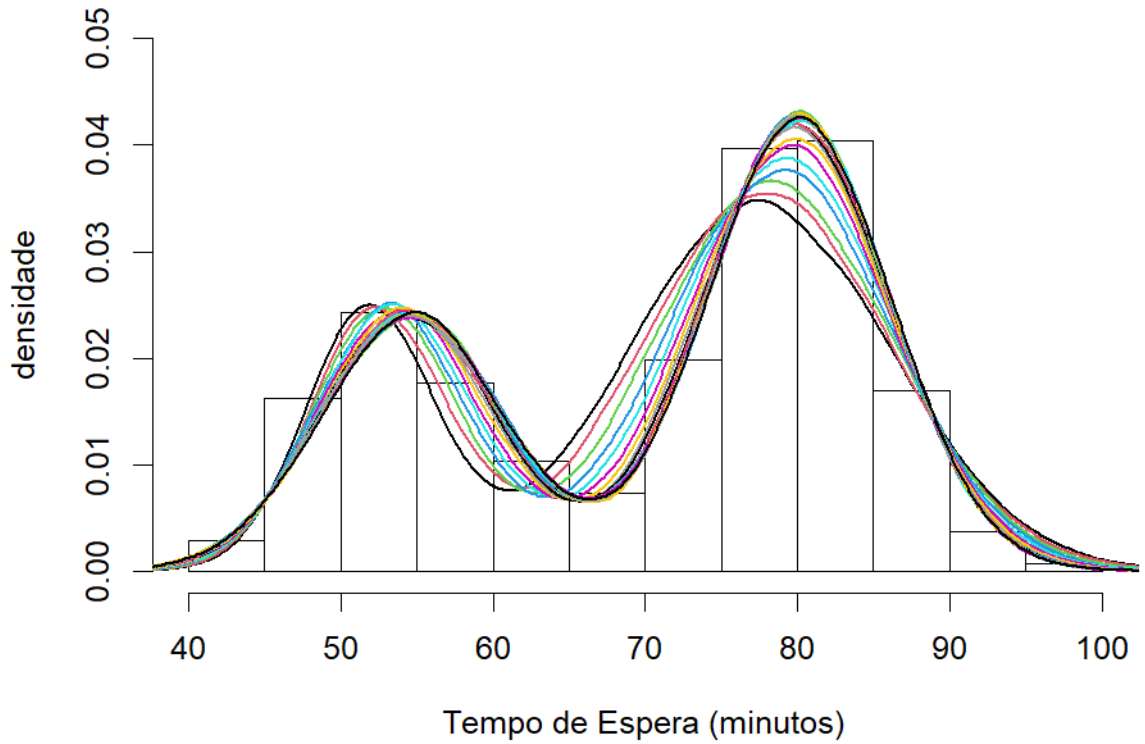
Tempo de Espera - Continuação

Programa R parcial considerando mistura de normais:

```
p <- 0.25;  mu1 <- 50;  sig1 <- 4;  mu2 <- 75;  sig2 <- 7

for (i in 1:25){
  deny <- p*dnorm(y,mu1,sig1) + (1-p)*dnorm(y,mu2,sig2)
  p1 <-  p*dnorm(y,mu1,sig1)/deny
  p2 <-  1- p1

  # updates:
  p <- sum(p1)/n
  mu1 <- sum(p1*y)/sum(p1)
  mu2 <- sum(p2*y)/sum(p2)
  sig11 <- sum(p1 *(y-mu1)^2)/ sum(p1)
  sig1<- sqrt(sig11)
  sig22 <- sum(p2 *(y-mu2)^2)/ sum(p2)
  sig2 <- sqrt(sig22)
  res <- c(i,round(c(p,mu1,sig1,mu2,sig2),3))
}

> res[2:6]
[1]  0.361 54.614  5.871 80.091  5.868
```

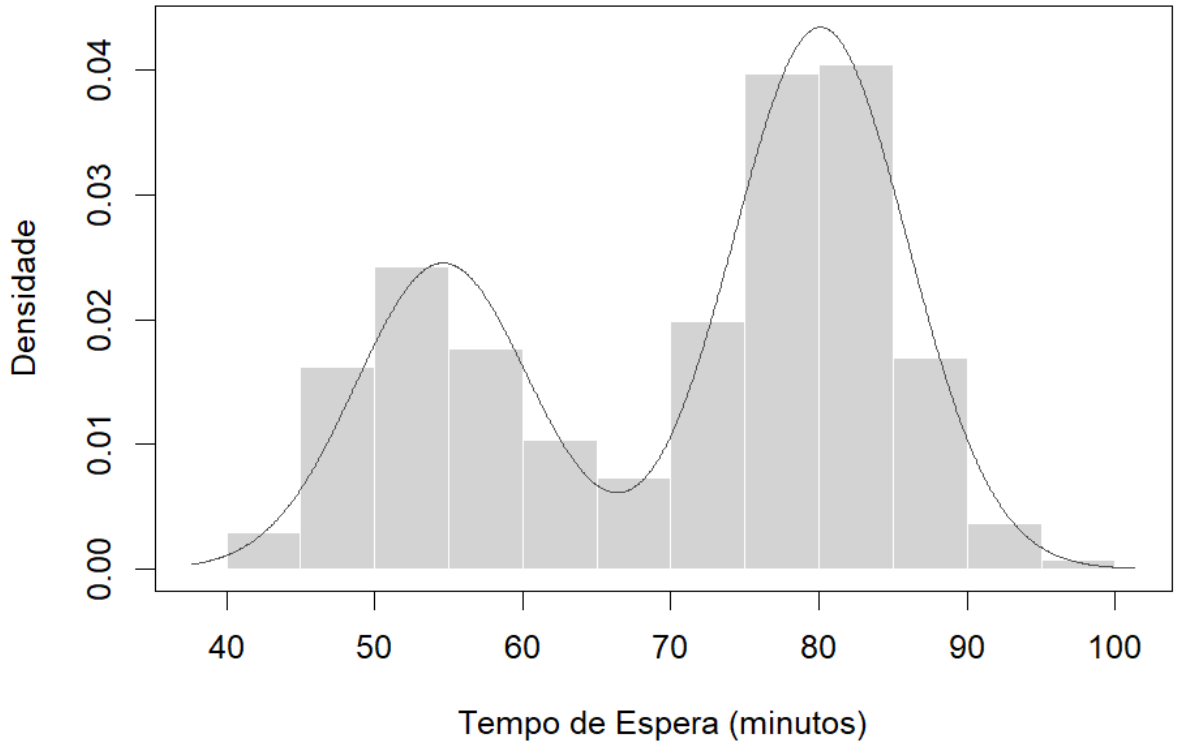
Tempo de Espera - Continuação

Imagem das densidades estimadas em 25 iterações.



Tempo de Espera - Continuação

Estimativa utilizando função `densityMclust` do pacote `mclust`.



Para detalhes ver programa `EDP.R` e Scrucca et al., 2023.

Seleção de Modelos

- **BIC - Bayesian Information Criteria**

$$\mathrm{BIC}_{\mathcal{M},G} = 2\ell_{\mathcal{M},G}(\hat{\boldsymbol{\Psi}}|\mathbf{x}) - \nu_{\mathcal{M},G}\log(n),$$

em que $\nu_{\mathcal{M},G}$ indica o número de parâmetros a estimar no modelo $\mathcal{M}$ com G classes/grupos.

BIC tende a enfatizar o número de componentes da mistura, em vez de selecionar o número de grupos.

- **ICL - Integrated Complete (data) Likelihood Criteria**

$$\mathrm{ICL}_{\mathcal{M},G} = \mathrm{BIC}_{\mathcal{M},G} + 2\sum_{i=1}^{n}\sum_{g=1}^{G} \mathrm{c}_{ig}\log(\hat{z}_{ig}),$$

em que c_{ig} é uma indicadora de $\mathbf{x}_i$ no grupo g e $\hat{z}_{ig}$ a probabilidade condicional de $\mathbf{x}_i$ pertencer a g-ésima mistura.

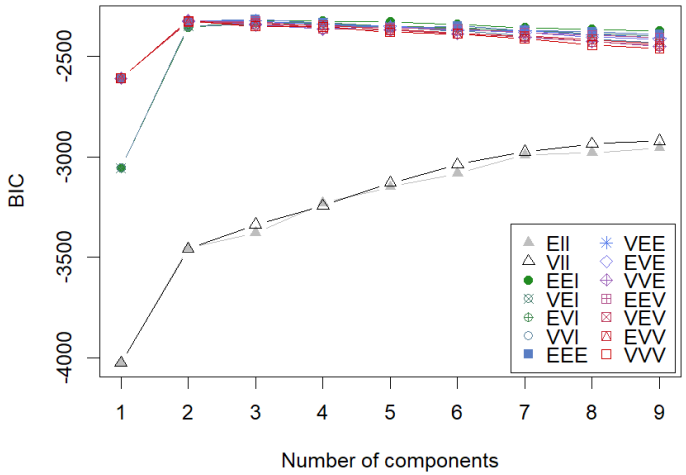
ICL tende a enfatizar o número de grupos.

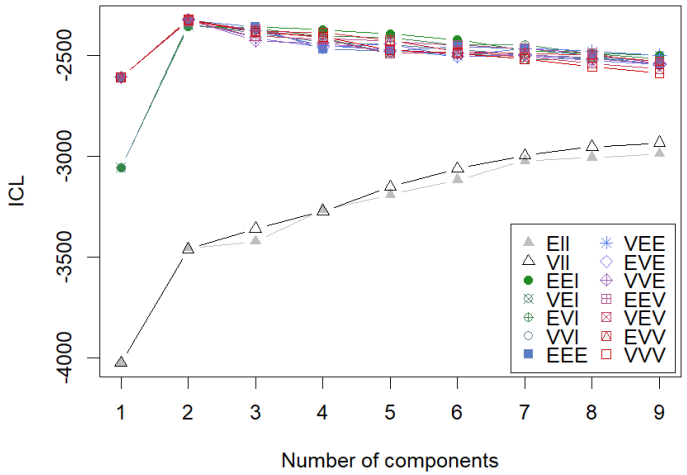
$\Rightarrow$ Maior BIC ou ICL, melhor ajuste.

- **Teste da Razão da Verossimilhança** utiliza a técnica boostrap para comparar se um modelo com mais grupos é significativamente melhor que um modelo com mais simples, porém com mesma estrutura de variância-covariância (ver Scrucca et al., 2023, pág. 23).

Tempo de Espera - Continuação

Seleção de modelos, `mclustBIC` e `mclustICL` do pacote `mclust`.





Para detalhes ver programa `EDP.R` e Scrucca et al., 2023.

Classificação

- **MAP - Maximum a Posteriori Classification** uma vez ajustado o GMM, alocar $\mathbf{x}_i$ a classe C_{g*} com $g* = \arg\max_g \hat{z}_{ig}$ tal que

$$\hat{z}_{ig} = \frac{\pi_g \ \phi(\mathbf{x}_i|\hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)}{\sum_{g=1}^{G} \pi_g \ \phi(\mathbf{x}_i|\hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)}.$$

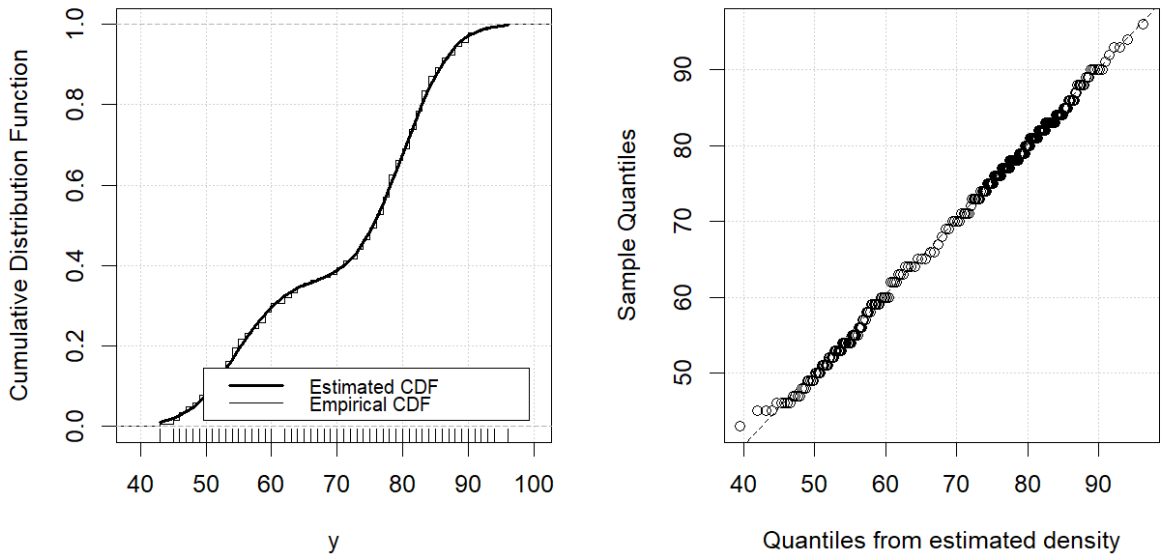
- A medida $u_i = 1 - \max_g \hat{z}_{ig}$, fornece o grau de incerteza sobre a alocação no grupo $g*$.
  Quanto menor u_i, menor a incerteza na alocação ($u_i \in [0, 1]$).

Diagnóstico visual

- Função distribuição acumulada vs função distribuição empírica (cdf).
- Quantis amostrais vs quantis da densidade estimada (qq).

Tempo de Espera - Continuação

Gráficos CDF e Q-Q obtidos com o `mclust`.



Para detalhes ver programa `EDP.R` e Scrucca et al., 2023.

Vantagens e Desvantagens do GMM

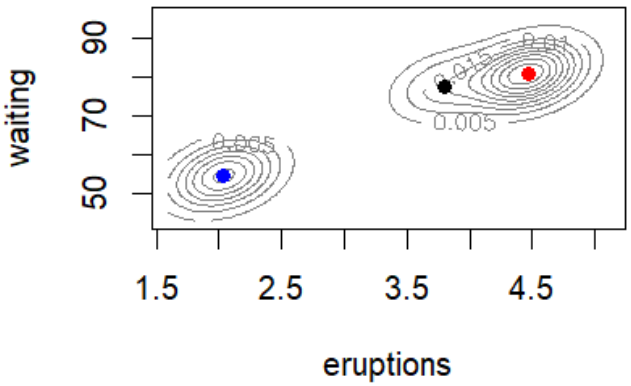
- Não precisa determinar o número de intervalos (ex. histogramas) ou comprimento de banda h (ex. kernel).
- Eficiente para grandes dimensões.
- EMV podem ser calculados.
- Algoritmo EM depende de bons valores iniciais e pode ser lento.
- Muitos componentes aumentam a variância da estimativa e modelos mais simples aumentam o viés.
- Número de componentes é previamente desconhecido.
- Necessitam de mais grupos (g) para ajustar os dados do que em misturas de outras densidades (ex. caudas pesadas) em certas aplicações (McLachan e Peel, 2000).

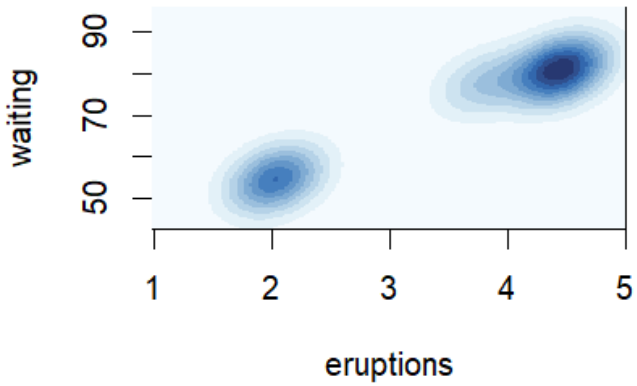
Observação

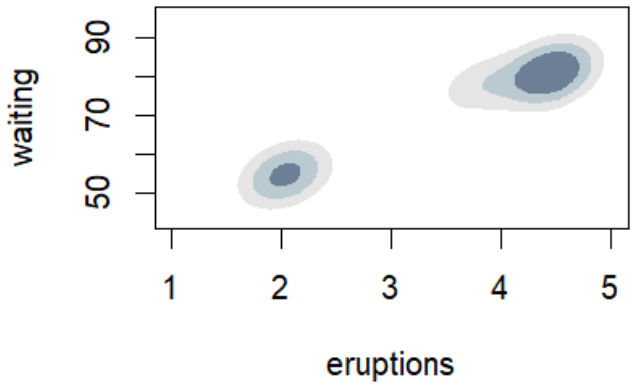
- No caso de dados com limite inferior e/ou superior é possível realizar uma transformação que força valores nulos para a densidade estimada fora dos limites possíveis de observação.
- Exemplos: tempo, em dias, de internação em UTI para pacientes diagnósticados com câncer do pulmão.
- Função `densityMclustBounded( , lbound = , ubound = )` do pacote `mclustAddons`.

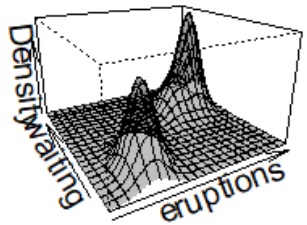
GMM com dados bidimensionais

Estimativa para base de dados `faithful` com $G = 3$ (seleção por BIC).





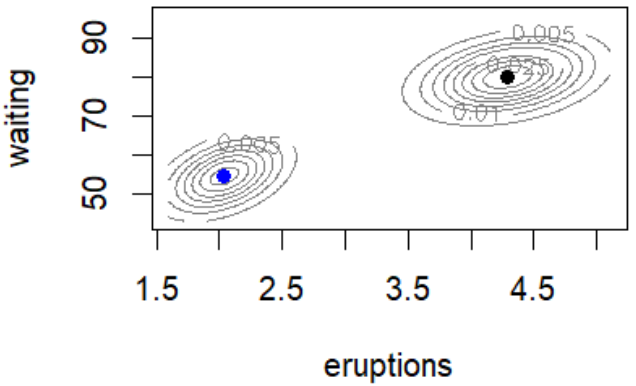


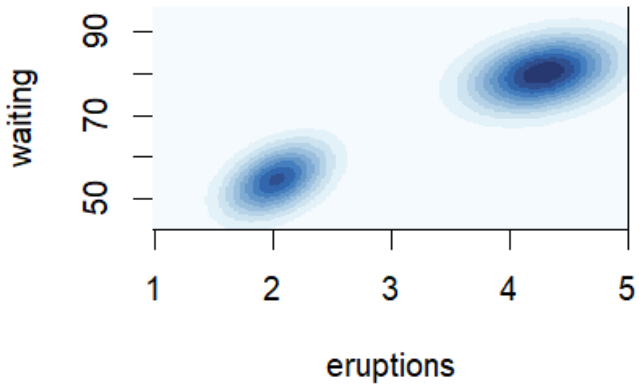


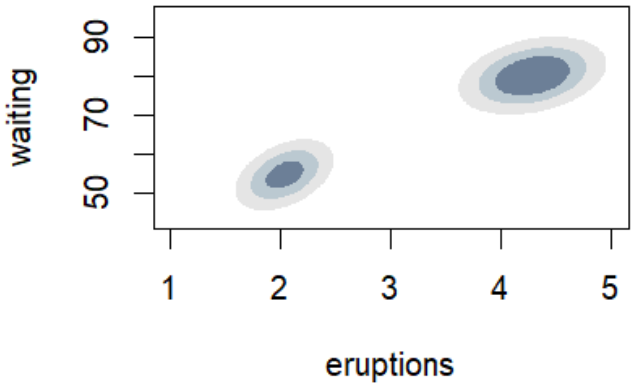
Para detalhes ver programa `EDP.R` e Scrucca et al., 2023.

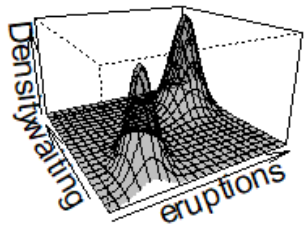
GMM com dados bidimensionais

Estimativa para base de dados `faithful` com $G = 2$.









Para detalhes ver programa `EDP.R` e Scrucca et al., 2023.

Aplicação em Agrupamento de Dados (Não Supervisionado)

1. von Borries, G.; Wang, H. (2009) Partition clustering of high dimensional low sample size data based on p-values. *Computational Statistics and Data Analysis*, vol. 53. Artigo

2. Lins, R. (2019) Implementação computacional de algoritmos para agrupamento de dados HDLSS e HDLLSS. Apresentação na 64a RBRAS, Cuibá-MT. Monografia, Programa R.

3. dos Santos, L.B. (2021) Agrupamento por misturas finitas normais com aplicação a dados de expressão gênica. Trabalho de conclusão do Bacharelado em Estatística, UnB. Monografia.