Análise Exploratória de Dados e Estimação de Densidade

Mineração de Dados

Universidade Federal do ABC

Iniciando a análise

- Ao recebermos um conjunto de dados, normalmente, recebemos junto algumas informações (metadados por exemplo)
- A primeira coisa a se fazer é explorar os dados, suas características e identificar possíveis problemas.

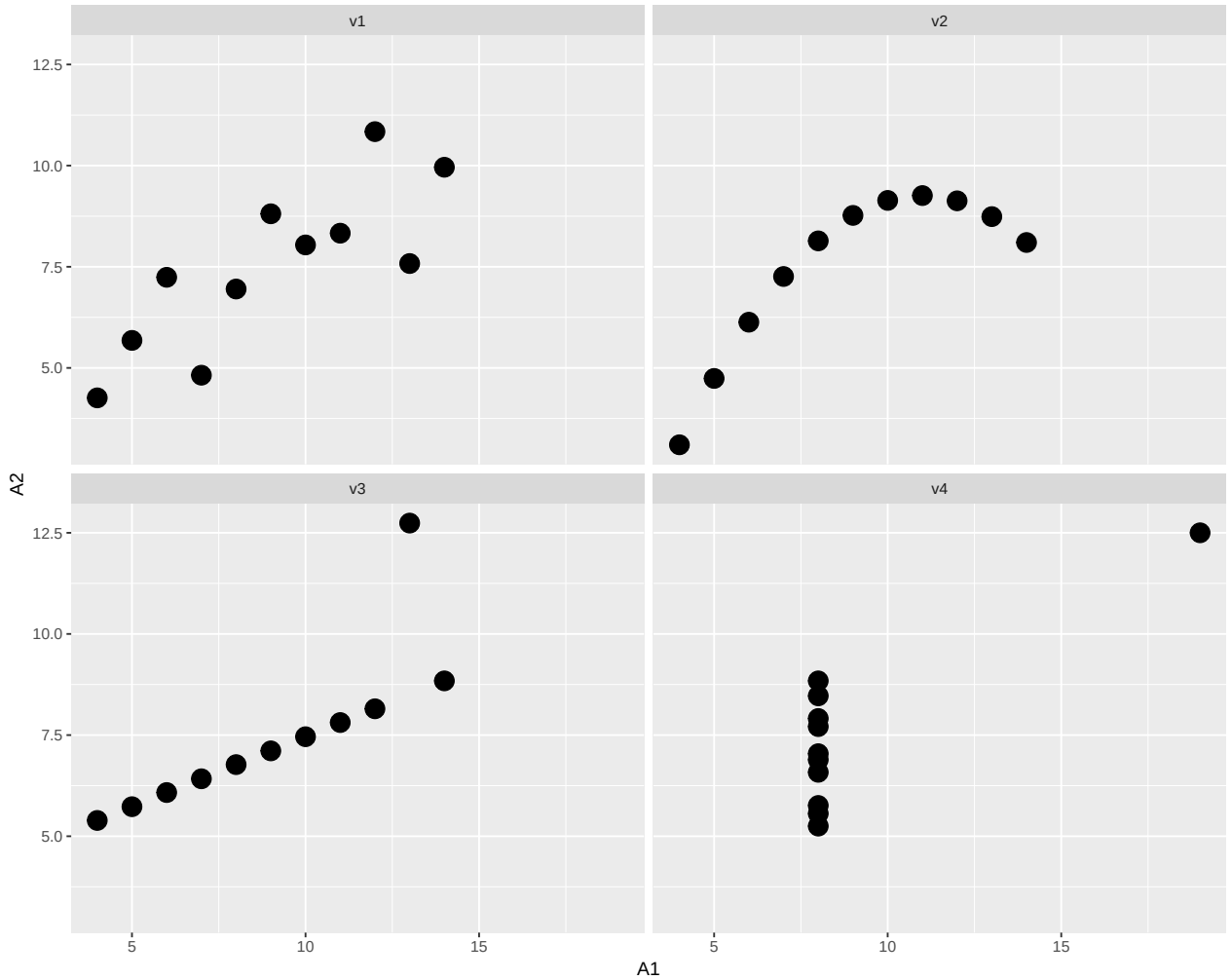
Quarteto de Anscombe

- Vamos começar por 4 bases de dados simples, cada uma com 11 objetos e 2 atributos. Essas bases de dados são chamadas de *Anscombe's Quartet'*.

	V1	V2	V3	V4
média(A1)	9,000	9,000	9,000	9,000
média(A2)	7,501	7,501	7,500	7,501
variância(A1)	11,000	11,000	11,000	11,000
variância(A2)	4,127	4,128	4,123	4,123
correlação(A1,A2)	0,816	0,816	0,816	0,817

Quarteto de Anscombe

- Estas bases de dados possuem características básicas idênticas. No entanto...



Estatísticas descritivas

- Outras opções além das estatísticas mais comuns (média, variância, mediana etc)
- Quartil
- Amplitude Interquartil (Interquartile Range)
- Estes são usualmente sumarizados em um *Boxplot*
  - Limites superior e inferior podem ser derivados do IQR (existem variações):
    - LI = 1º Quartil - 1.5 * IQR
    - LS = 3º Quartil + 1.5 * IQR

Boxplot

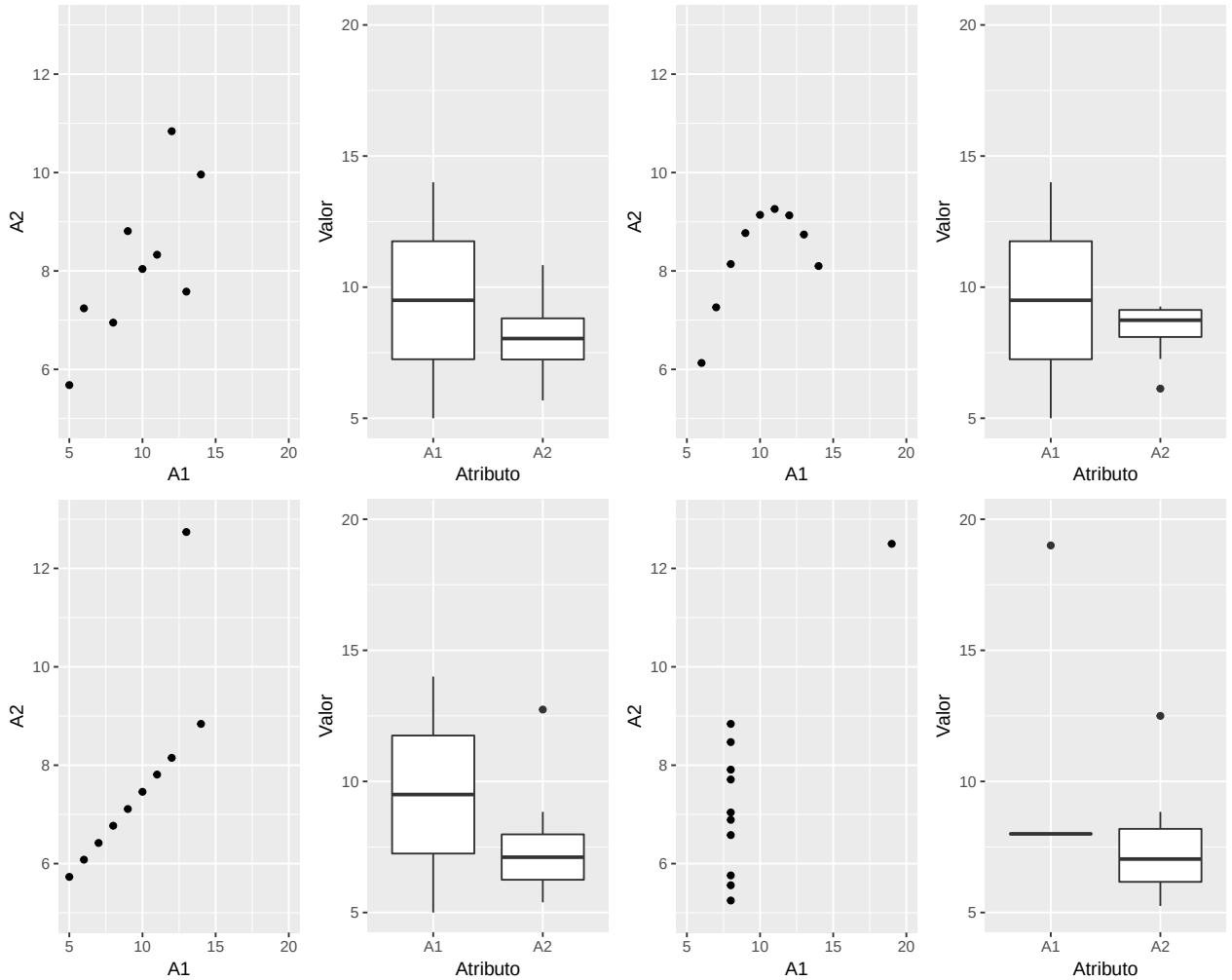
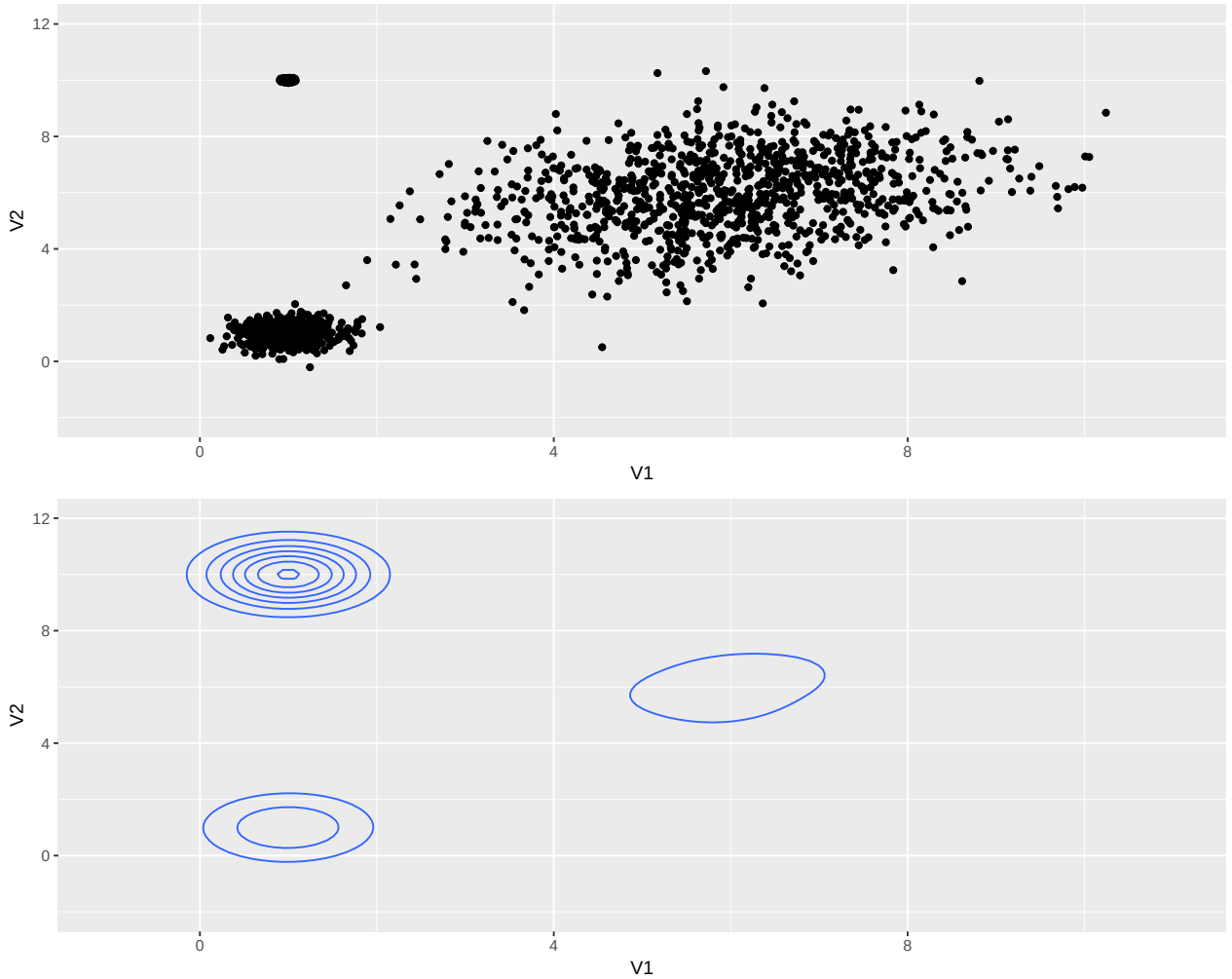


Gráfico de contorno

- Gráficos de dispersão são ótimas ferramentas para visualizar dados em 2-D, mas sobreposição de pontos pode dificultar a visualização
- Gráficos de contorno são uma excelente ferramenta para tratar estes casos
  - Relacionam uma terceira variável aos eixos x e y
  - No caso de sobreposição de pontos a terceira variável pode ser a densidade

Gráfico de contorno



Estimando densidade

- Estimar densidade é necessário em diversos algoritmos utilizados em MD.
- Conforme veremos durante o curso, boa parte dos algoritmos buscam aproximar a função $p(Y|\mathcal{X})$ em que Y é a saída desejada e $\mathcal{X}$ são os dados.
- Por enquanto, focaremos em estimar a função $p(x)$.

Estimando densidade

- Existem duas abordagens principais para se estimar densidade:
  - Paramétrica: assume uma determinada forma (distribuição) para a variável
  - Não-paramétrica: não é baseada em uma premissa sobre o formato da distribuição
    - Flexibilidade tem um custo
    - Pode se tornar inviável quando se tem muitos atributos (voltarei nesse assunto)

Estimando densidade - Abordagem Paramétrica

- Necessário saber *a priori* a distribuição dos dados (sua forma)
  - os parâmetros são estimados baseando-se nos dados observados
- Vamos cobrir nessa aula apenas a distribuição Normal, mas os conceitos são aplicáveis às demais
  - quase sempre assume-se normalidade, embora nem sempre os dados suportem essa premissa

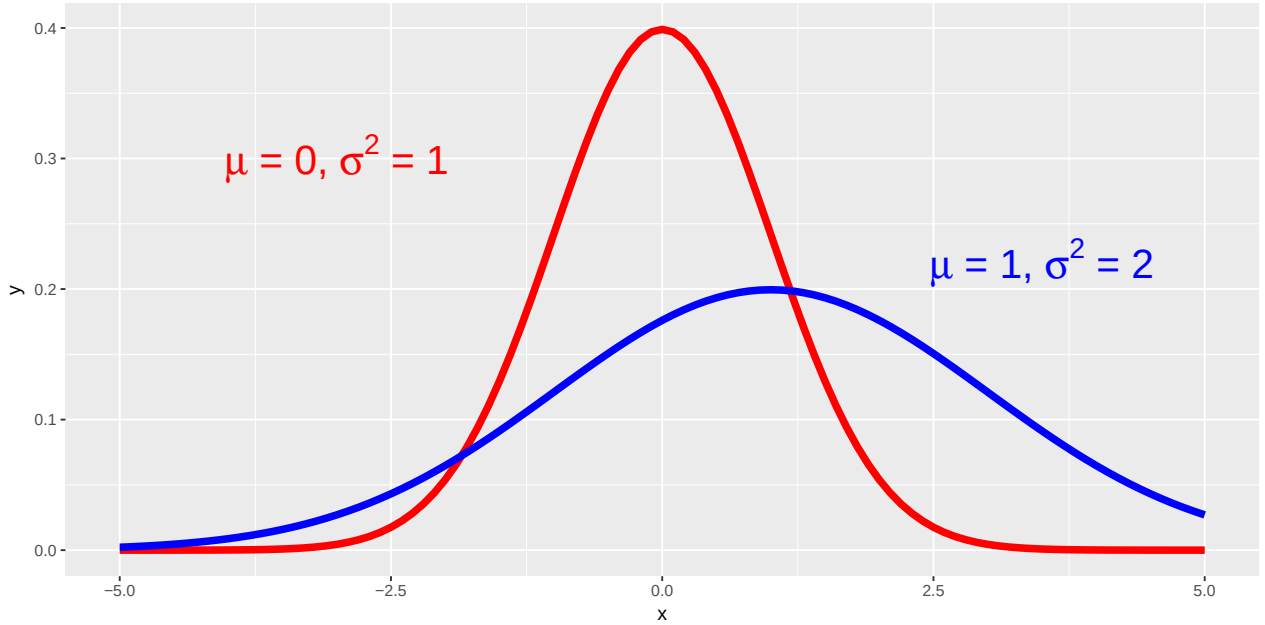
Estimando densidade - Abordagem Paramétrica

- Lembrando a equação que define a distribuição Normal

$$f(x\ ;\ \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Dois parâmetros definem a distribuição:
  - média (μ): define centro
  - variância (σ^2): define concentração (~68% dos valores estão a 1 desvio padrão (σ) da média)

Estimando densidade - Abordagem Paramétrica



Estimando densidade - Abordagem Paramétrica

- Temos um conjunto de pontos $\{x_i\}_{i=1}^{N}$. Como encontrar os valores de μ e σ^2 que melhor se ajustam aos dados?
- **Estimador de Máxima Verossimilhança** (*Maximum Likelihood Estimate*)
  - Normalmente mais simples que outras alternativas
  - Boas propriedades de convergência

Estimando densidade - Abordagem Paramétrica

- Princípios gerais
  - Definir a função conjunta de probabilidades em relação aos parâmetros da distribuição (θ):

$$p(x_1, x_2, ..., x_N; \theta)$$

  - Normalmente assume-se que as amostras sao *independentes e identicamente distribuídas* (i.i.d)

$$p(x_1, x_2, ..., x_N; \theta) = \prod_{n=1}^{N} p(x_n; \theta)$$

Estimando densidade - Abordagem Paramétrica

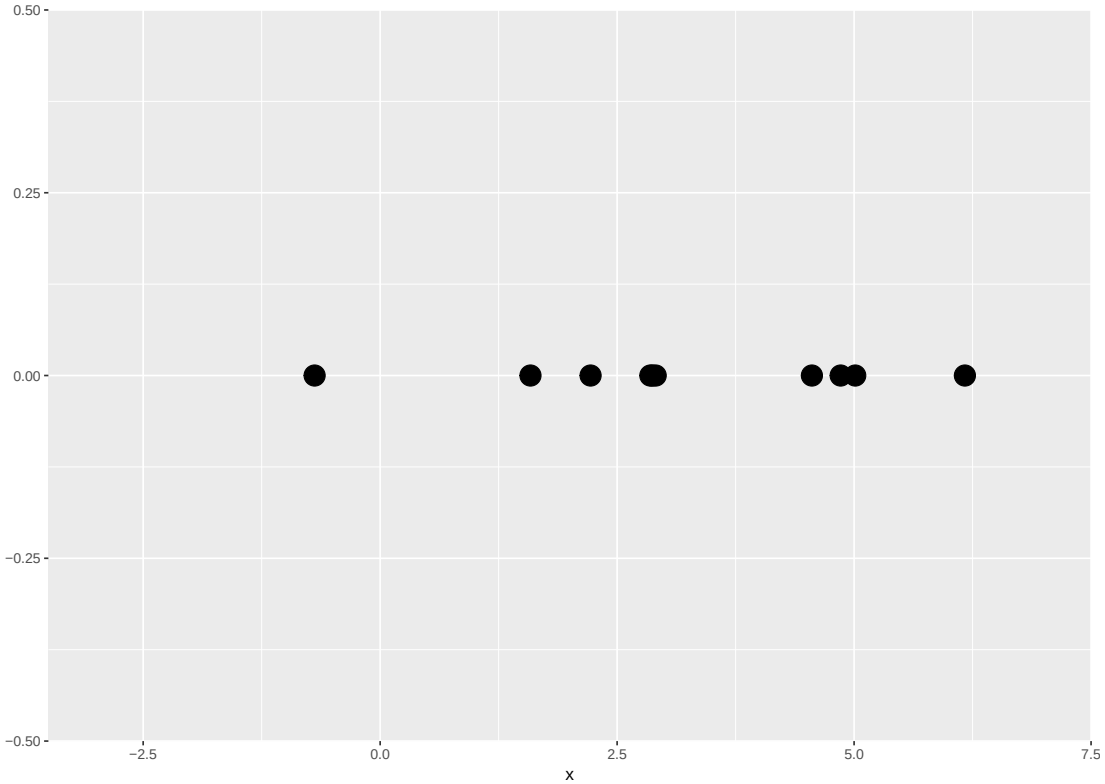
- Princípios gerais
  - Seja $L(\theta)$ a função de verossimilhança (*likelihood*)

$$L(\theta) = \prod_{n=1}^{N} p(x_n; \theta)$$

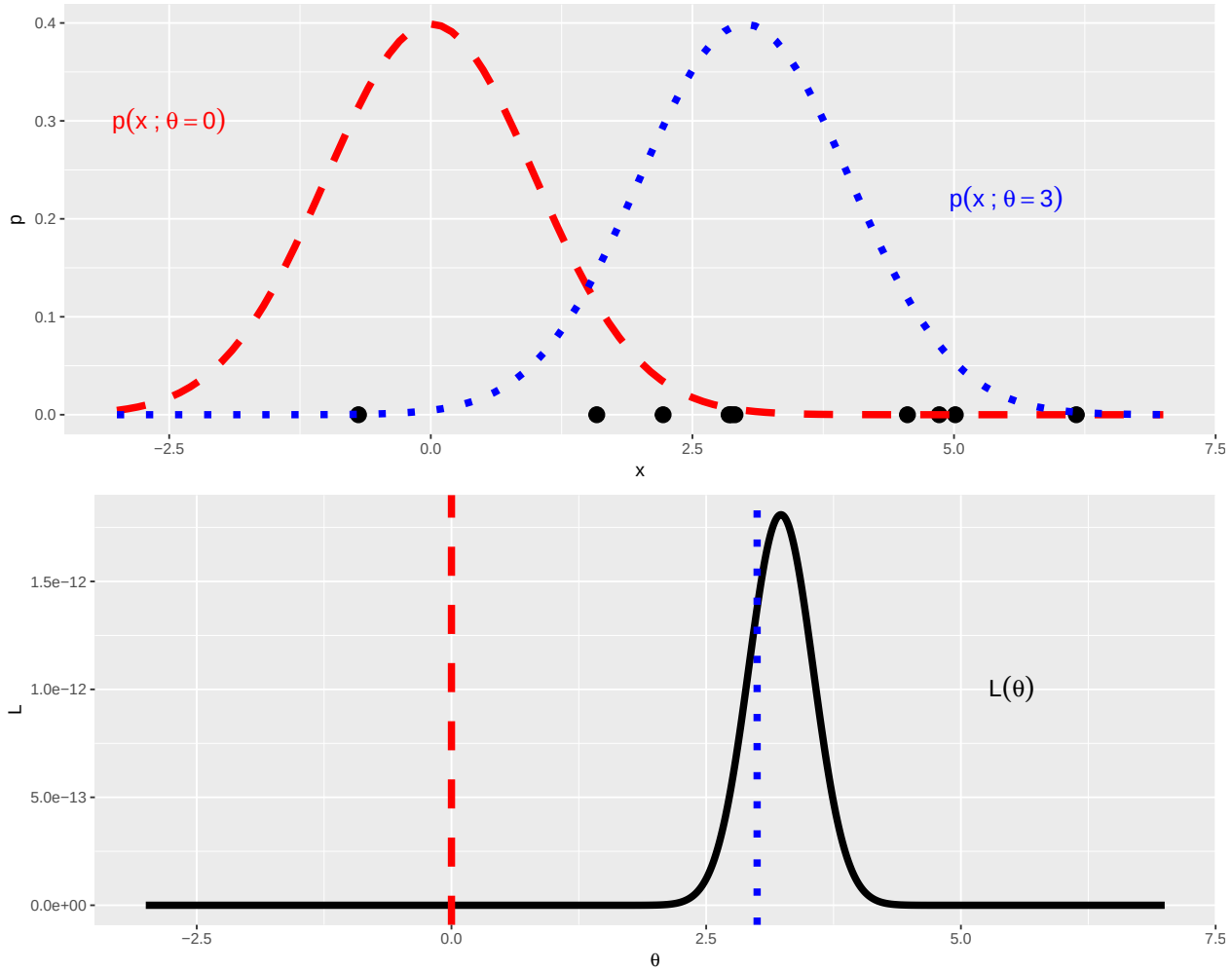
  - L não é uma função de densidade de probabilidades, e sim uma função de θ em relação às amostras

Estimando densidade - Abordagem Paramétrica

- Suponha o seguinte conjunto de pontos
  - Vamos assumir que sabemos a variância (σ^2) mas não a média
  - Neste caso, $\theta = \{\mu\}$



Estimando densidade - Abordagem Paramétrica

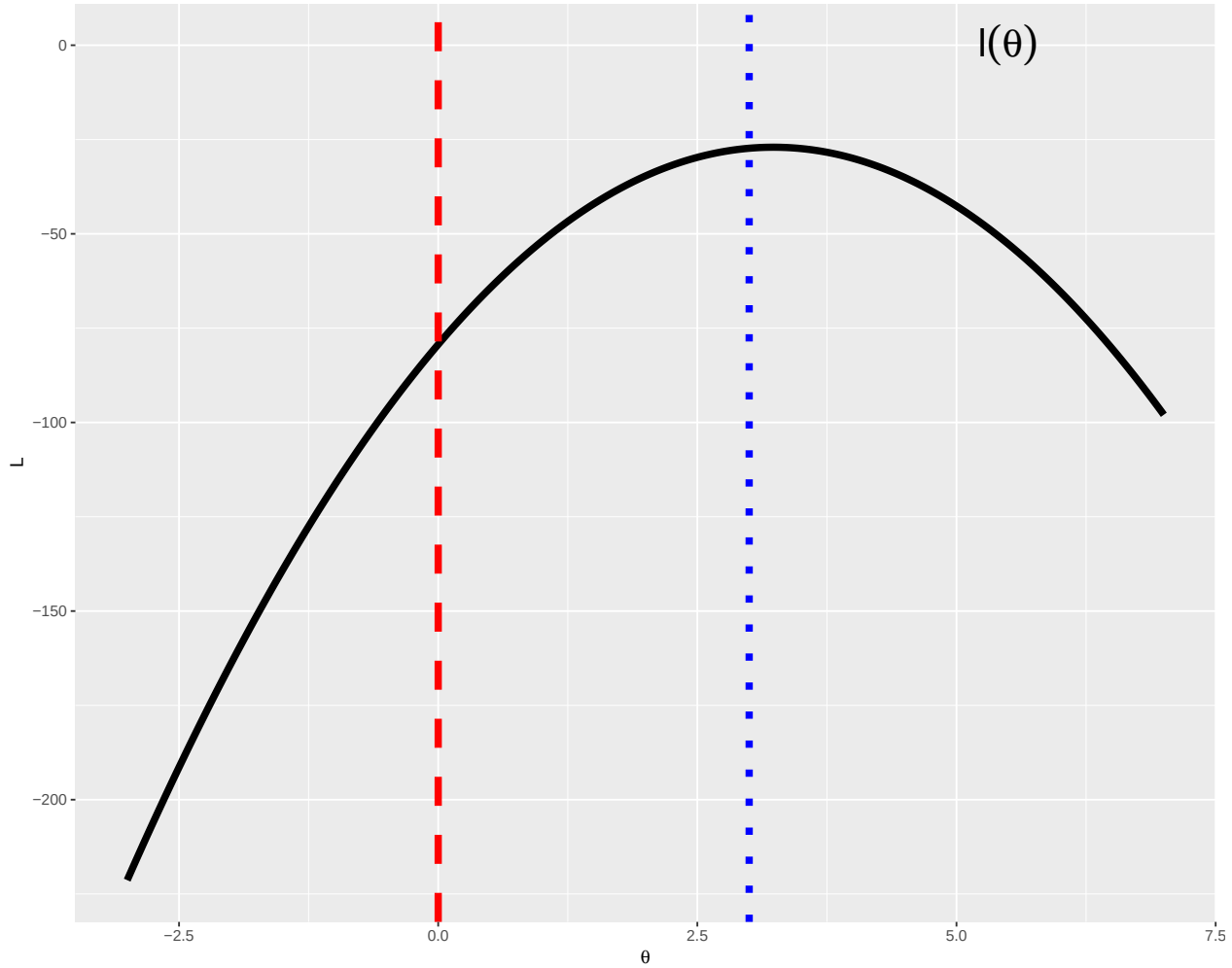


Estimando densidade - Abordagem Paramétrica

- Normalmente é mais fácil trabalhar com o log da função de verossimilhança
  - Log é uma função monotônica (preserva a ordem), logo o problema de maximização é equivalente

$$\log L(\theta) = l(\theta) = \log \prod_{n=1}^{N} p(x_n; \theta) = \sum_{n=1}^{N} \log p(x_n; \theta)$$

Estimando densidade - Abordagem Paramétrica



Estimando densidade - Abordagem Paramétrica

- O estimador de máxima verossimilhança corresponde ao parâmetro θ que maximiza a função de verossimilhança (ou de log verossimilhança)

$$\theta_{ML} = \underset{\theta}{\text{argmax}} l(\theta)$$

Estimando densidade - Abordagem Paramétrica

- Para encontrar θ_{ML} resolvemos $l'(\theta) = 0$
  - Condições de segunda ordem também devem ser verificadas ($l''(\theta) < 0$)
  - Existem alguns cuidados em relação aos limites do espaço de parâmetros
  - Lembre-se que é uma *estimativa*, garantias apenas no limite ao infinito de número de amostras

Estimando densidade - Abordagem Paramétrica

- Vamos ver o processo na distribuição normal ($\theta = \{\mu, \sigma^2\}$):

$$p(x\ ;\ \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\ln p(x\ ;\ \mu, \sigma^2) = \ln(1) - \frac{1}{2}\ln\left(2\pi\sigma^2\right)\ - \frac{1}{2\sigma^2}(x-\mu)^2$$

$$\ln p(x\ ;\ \mu, \sigma^2) = -\frac{1}{2}\ln\left(2\pi\sigma^2\right)\ - \frac{1}{2\sigma^2}(x-\mu)^2$$

Estimando densidade - Abordagem Paramétrica

- Substituindo na equação do likelihood:

$$l(\theta) = \sum_{n=1}^{N} \ln p(x_n; \theta) = \sum_{n=1}^{N} -\frac{1}{2} \ln \left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

$$l(\theta) = -\frac{N}{2} \ln (2\pi) - \frac{N}{2} \ln \left(\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2$$

Estimando densidade - Abordagem Paramétrica

- Temos que derivar em relação à média:

$$\frac{dl(\theta)}{d\mu} = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\frac{d\left[(x_n-\mu)^2\right]}{d\mu} \quad [\text{regra da cadeia:} h'(g(x))g'(x)]$$

$$\frac{dl(\theta)}{d\mu} = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}2(x_n-\mu)\cdot -1 = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)$$

Estimando densidade - Abordagem Paramétrica

- Para encontrar a estimativa de máxima verossimilhança devemos igualar a derivada a 0:

$$\frac{dl(\theta)}{d\mu} = 0 \Rightarrow \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) = 0$$

$$\frac{1}{\sigma^2}\left[\left(\sum_{n=1}^{N} x_n\right) - N\mu\right] = 0, \text{igual a zero se } \left(\sum_{n=1}^{N} x_n\right) - N\mu = 0$$

$$\sum_{n=1}^{N} x_n = N\mu \Rightarrow \mu = \frac{\sum_{n=1}^{N} x_n}{N}$$

Estimando densidade - Abordagem Paramétrica

- Falta a estimativa para a variância da distribuição
  - Para simplificar considere $\theta_1 = \sigma^2$

$$l(\theta) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\theta_1) - \frac{1}{2\theta_1}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$\frac{dl(\theta)}{d\theta_1} = -\frac{N}{2\theta_1} + \frac{1}{2(\theta_1)^2}\sum_{n=1}^{N}(x_n - \mu)^2 \quad \left[\frac{d\ln(x)}{dx} = \frac{1}{x}\right]$$

$$\frac{dl(\theta)}{d\theta_1} = \frac{1}{2\theta_1}\left[-N + \frac{1}{\theta_1}\sum_{n=1}^{N}(x_n - \mu)^2\right]$$

Estimando densidade - Abordagem Paramétrica

- Para encontrar a estimativa de máxima verossimilhança devemos igualar a derivada a 0 (lembrando que $\theta_1 = \sigma^2 \wedge \sigma^2 > 0$):

$$\frac{dl(\theta)}{d\theta_1} = 0 \Rightarrow \frac{1}{2\theta_1}\left[-N + \frac{1}{\theta_1}\sum_{n=1}^{N}(x_n - \mu)^2\right] = 0$$

- Igual a zero se:

$$-N + \frac{1}{\theta_1}\sum_{n=1}^{N}(x_n - \mu)^2 = 0 \Rightarrow \sigma^2 = \theta_1 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

Estimando densidade - Abordagem Paramétrica

- Note que μ nesse caso corresponde à estimativa da média obtida (estamos maximizando em relação às duas quantidades)
- Essa abordagem não é perfeita
  - o estimador da variância é *enviesado* (o valor do parâmetro é subestimado)
  - o estimador sem viés corresponde a:

$$\sigma^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu)^2$$

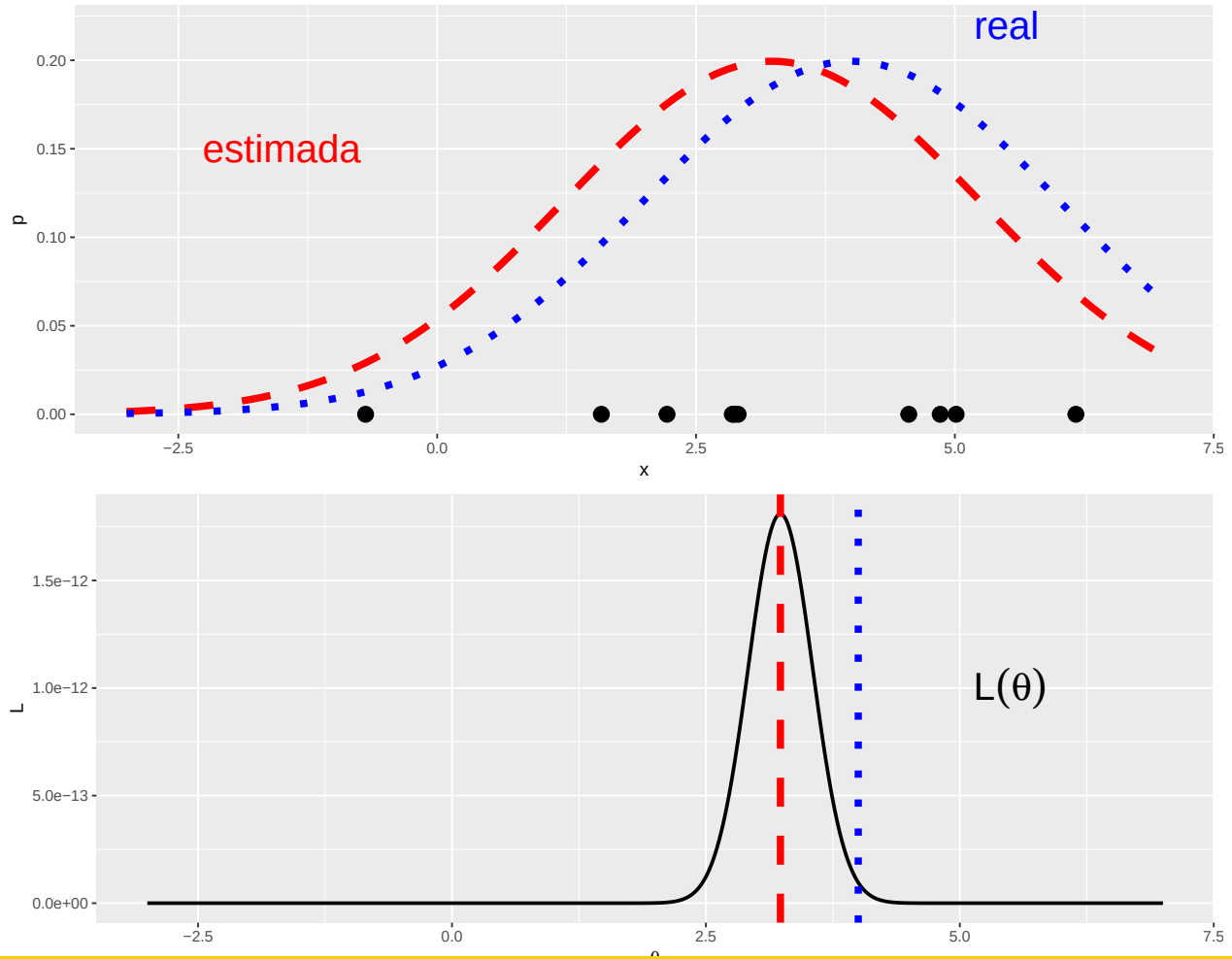
  - neste caso específico, o problema não é preocupante
    - assumindo que o número de amostas (N) é grande

Estimando densidade - Abordagem Paramétrica

- Voltando ao nosso exemplo, temos:
  - $\hat{\mu} = 3.23$
  - $\hat{\sigma}^2_{\text{biased}} = 3.57$
  - $\hat{\sigma}^2_{\text{unbiased}} = 3.97$
  - μ_{real} 4.00
  - σ^2_{real} 2.00

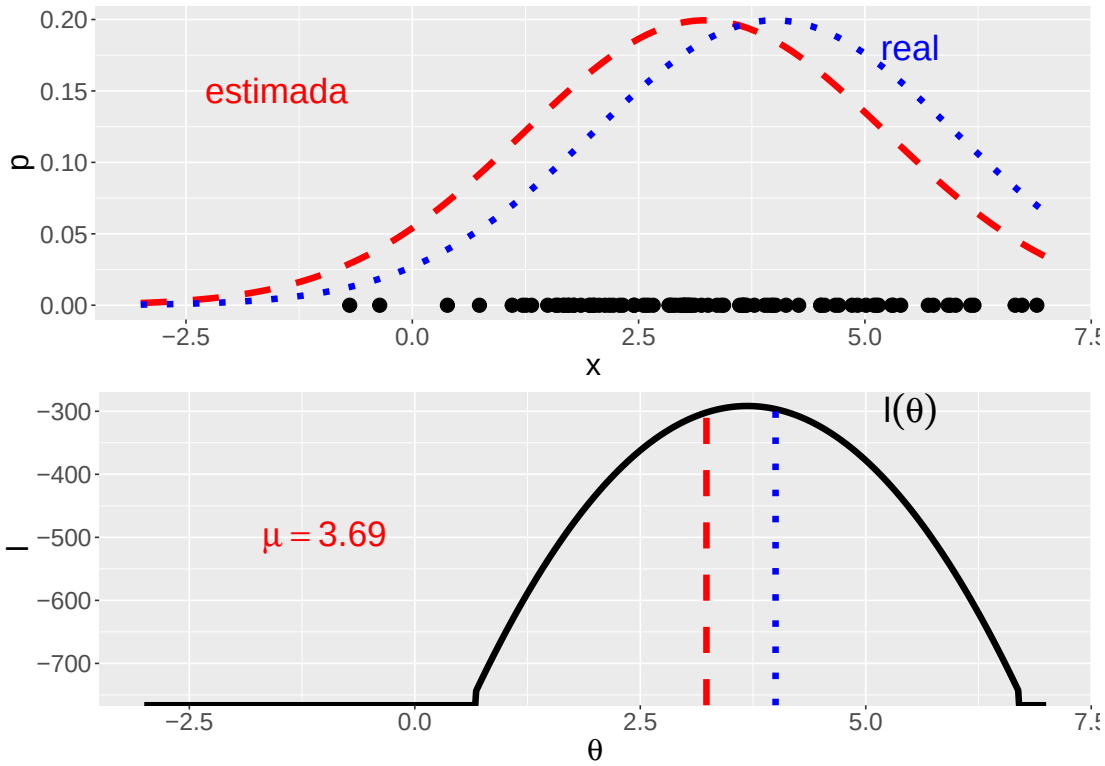
Estimando densidade - Abordagem Paramétrica

- Para simplificar a visualização, voltamos a assumir que a variância é conhecida



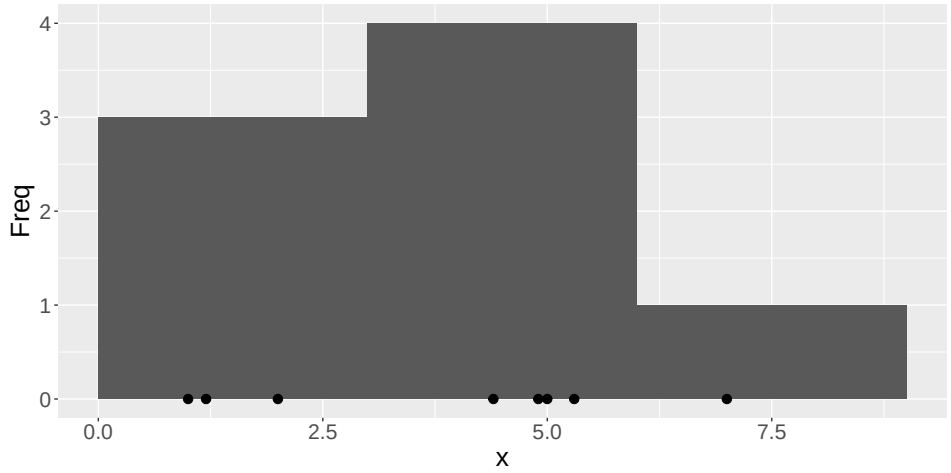
Estimando densidade - Abordagem Paramétrica

- Observe como devido ao pequeno número de amostras os parâmetros corretos possuem valor baixo de verossimilhança, veja o que acontece com 100 amostras ao invés de 10



Estimando densidade - Abordagem Não-paramétrica

- Quando não conhecemos a distribuição geradora dos dados ou não temos um bom palpite
  - Se os dados não suportam o palpite, as estimativas de densidade podem ser muito ruins
  - Lembre-se, o primeiro passo deve ser sempre conhecer melhor os dados que serão analisados
- Lembram dos histogramas? Eles serão o ponto inicial desse tópico



Estimando densidade - Abordagem Não-paramétrica

- Podemos extrair uma estimativa de densidade a partir do histograma

$$\hat{p}(x) = \frac{\text{número de objetos na barra}}{Nh}$$

  - N: número total de objetos
  - h: largura da barra (volume)
- Desvantagens:
  - Descontinuidades
  - Densidade igual por toda a barra, independente da disposição dos objetos

Estimando densidade - Abordagem Não-paramétrica

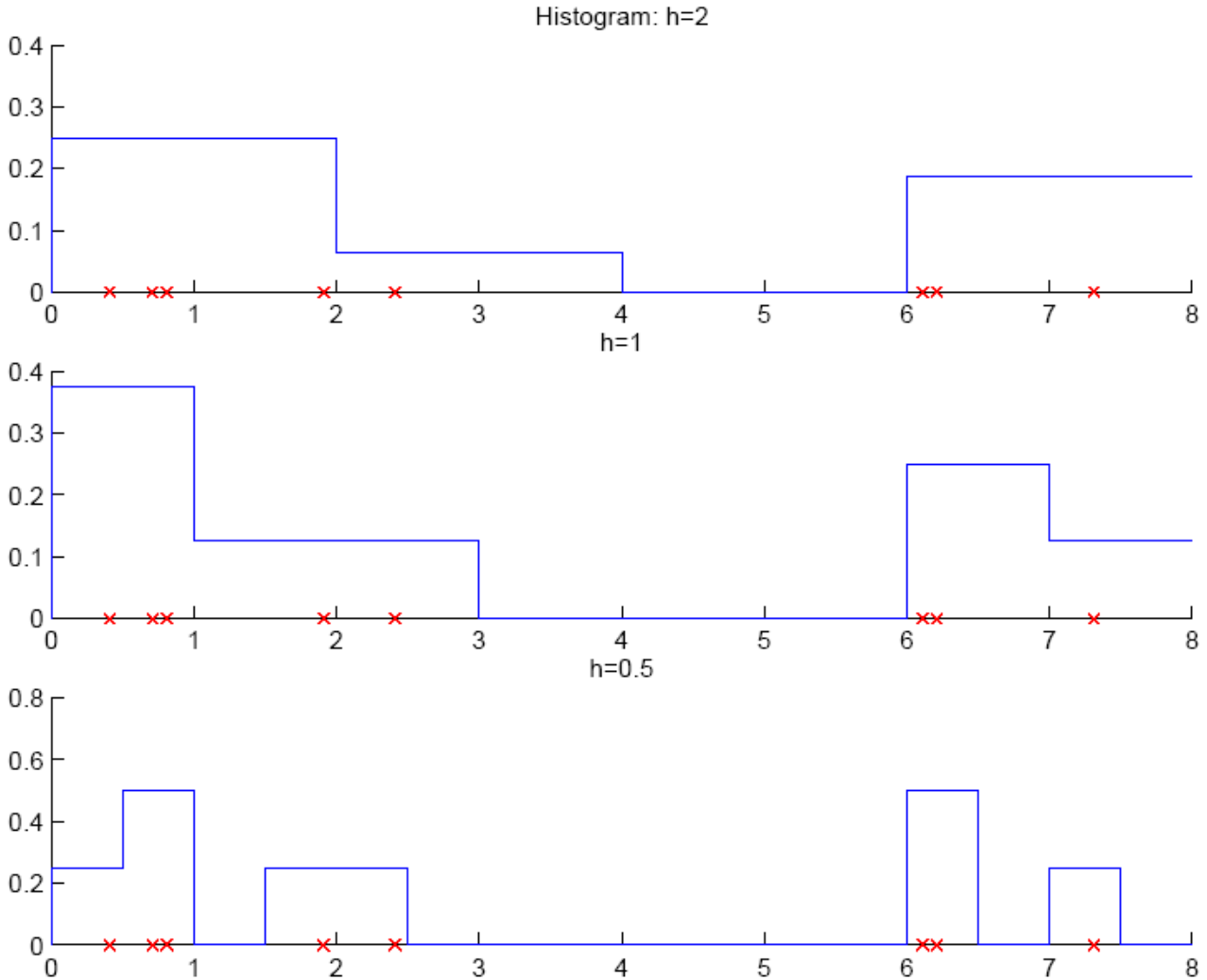


Figura 1: Estimativa histograma

Estimando densidade - Abordagem Não-paramétrica

- Podemos melhorar a estimativa considerando regiões de vizinhança
  - Abordagem também conhecida como *Parzen Window*
  - Necessário definir uma noção de distância
  - Inclusão de uma função de *kernel*

$$\hat{p}(x) = \frac{1}{Nh} \sum_{n=1}^{N} K\left(\frac{x - x_n}{h}\right)$$

- *Kernel* hiper-cubo unitário centrado na origem

$$K(u) = \begin{cases} 1, \text{se } |u| < 1/2 \\ 0, \text{caso contrário} \end{cases}$$

Estimando densidade - Abordagem Não-paramétrica

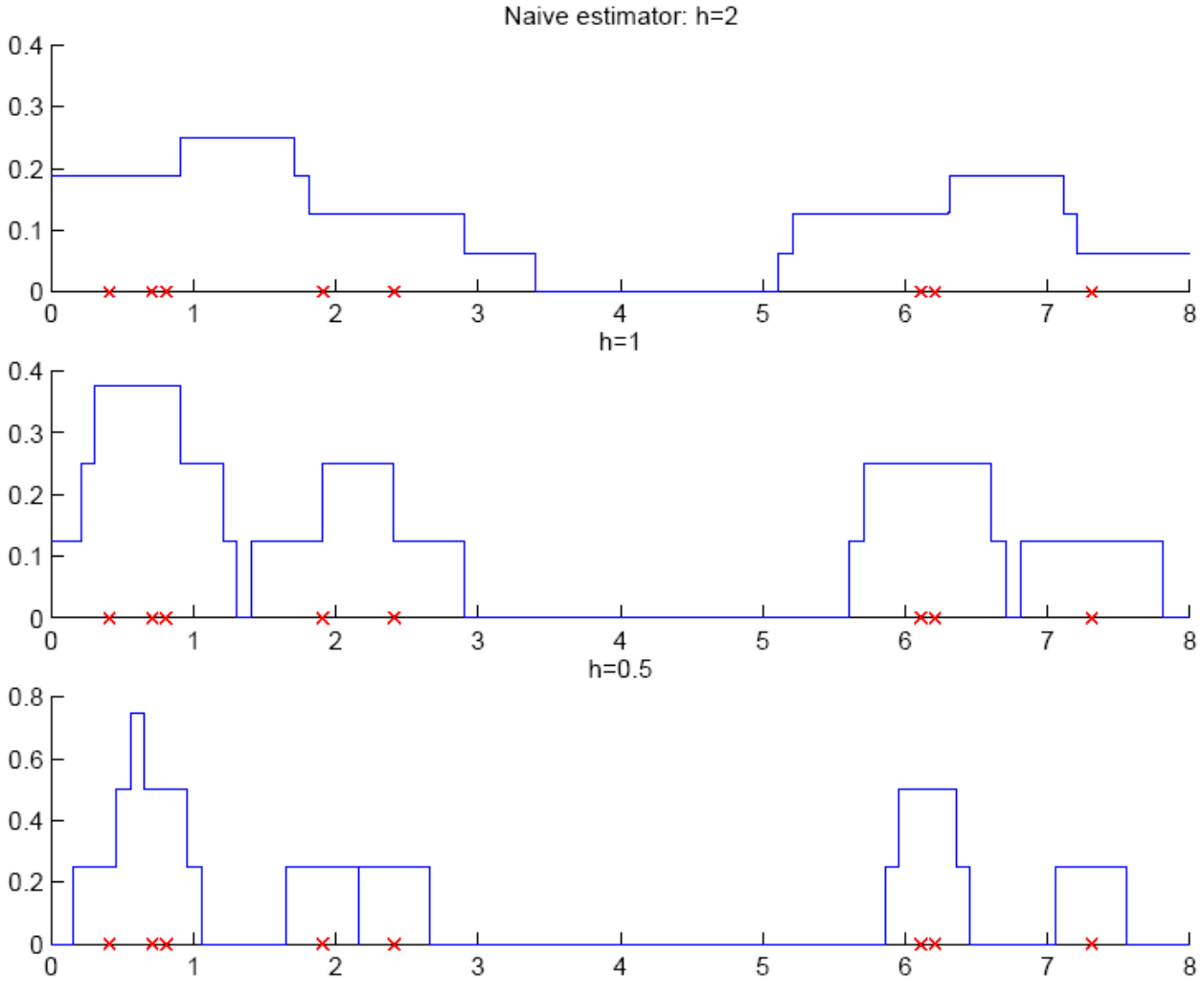


Figura 2: Estimativa hiper-cubo unitário

Estimando densidade - Abordagem Não-paramétrica

- Utilizando um *kernel* suave obtemos estimativas mais apropriadas, o mais comum é o *kernel* gaussiano
  - Note que isso não significa que estamos assumindo que os dados foram gerados por uma distribuição Normal

$$K(u) = \frac{1}{\sqrt{2\pi}} exp\left[-\frac{u^2}{2}\right]$$

  - Podemos simplificar adotando um corte:
    - $K(\cdot) = 0$ se $|x - x_n| > 3h$

Estimando densidade - Abordagem Não-paramétrica

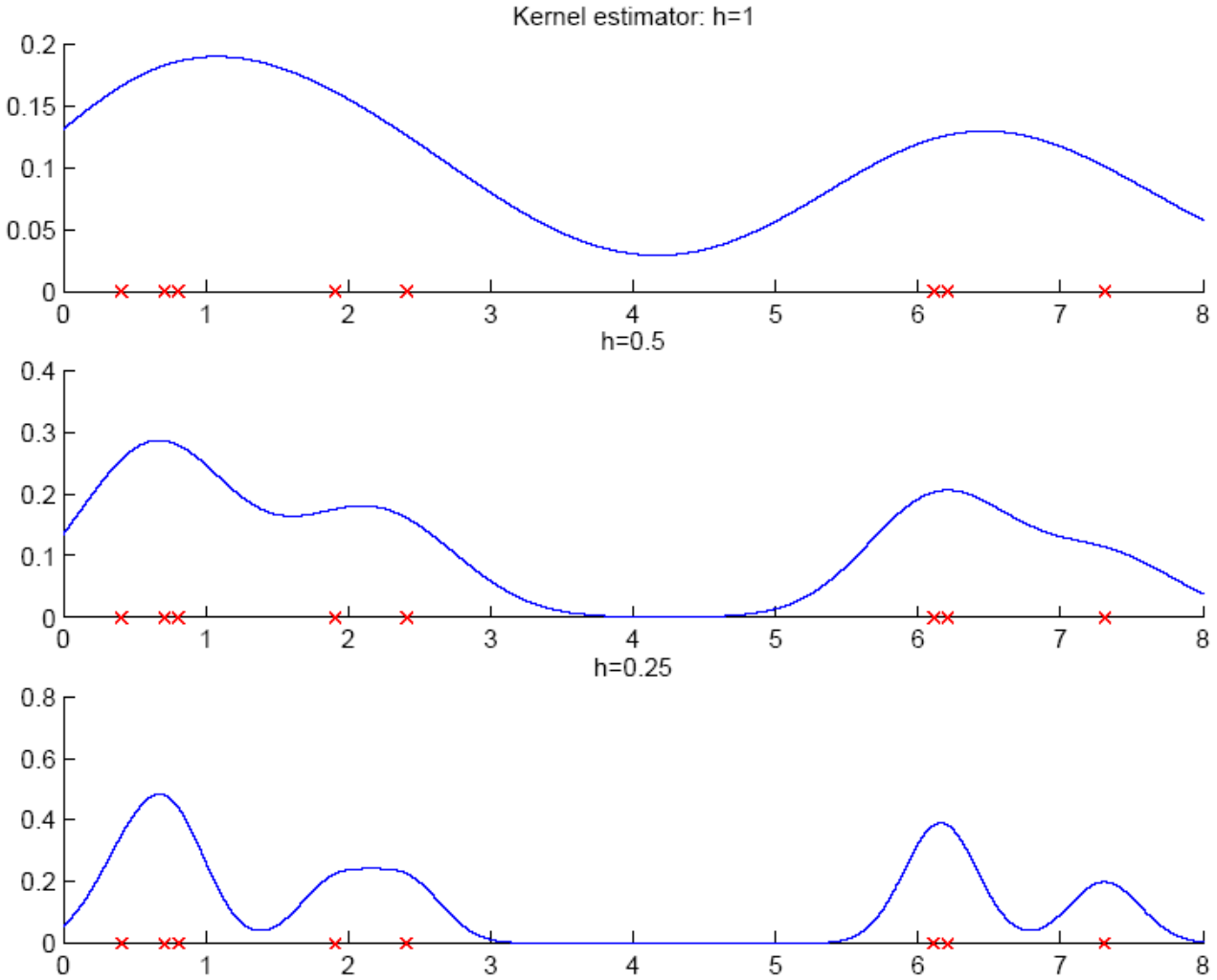


Figura 3: Estimativa kernel gaussiano

Estimando densidade - Abordagem Não-paramétrica

- Outros *kernels* podem ser utilizados, contanto que o pico seja em $u = 0$ e diminua conforme $|u|$ aumenta.
  - Para ser uma função de densidade legítima deve satisfazer:

$$\int K(u)du = 1$$

$$K(u) \geq 0$$

- Ainda temos que definir um parâmetro de largura (h) apropriado
  - h pequeno estimativa fica suscetível a ruído nos dados
  - h grande estimativa fica artificialmente suavizada

Estimando densidade - Abordagem Não-paramétrica

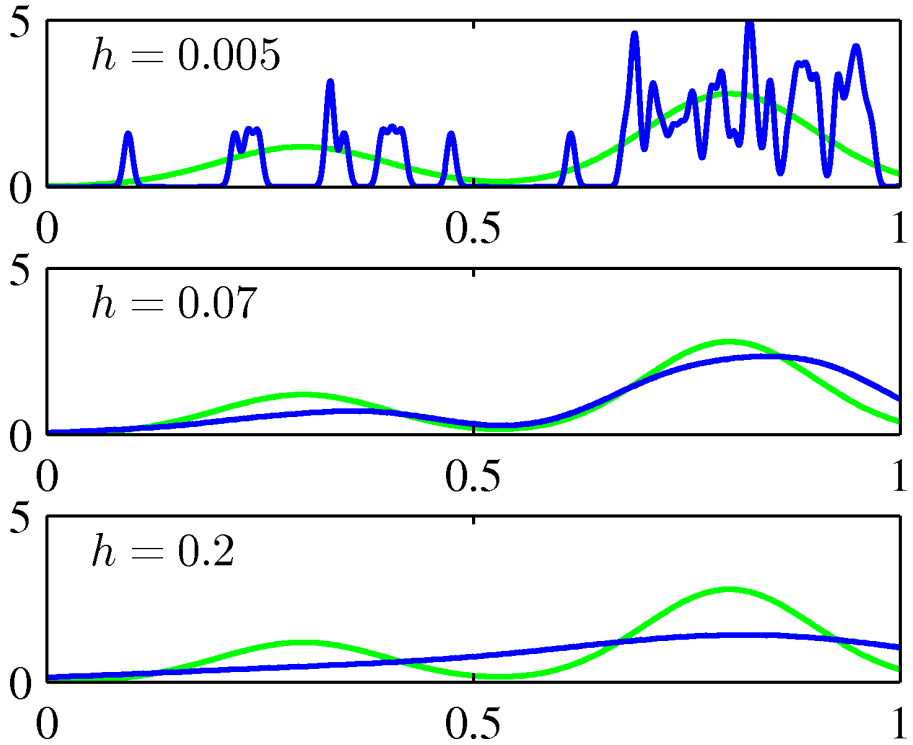


Figura 4: Influência parâmetro h

- Curva verde corresponde ao modelo gerador dos dados
- Curva azul corresponde à estimativa de densidade obtida

Estimando densidade - Abordagem Não-paramétrica

- Como definir h?
  - Existem heurísticas adotadas em pacotes estatísticos (Ex: R), baseadas em premissas sobre os dados
  - Funcionam bem se as premissas (Ex: Normalidade) sob as quais foram desenvolvidadas forem satisfeitas
  - No caso geral, esse se torna um hiper-parâmetro a ser otimizado

Estimando densidade - Abordagem Não-paramétrica

- Com número de atributos > 1, pode-se utilizar uma distribuição Normal multivariada
  - No entanto, o número de parâmetros a serem estimados é quadrático em relação ao número de atributos
- Nos casos em que o número de amostras é pequeno, pode-se tratar os atributos de forma independente:

$$\hat{p}(\mathbf{x}_0) = \frac{1}{N} \sum_{n=1}^{N} \left\{ \prod_{m=1}^{M} \frac{1}{h_m} K\left(\frac{x_{0m} - x_{nm}}{h_m} \right) \right\}$$

- Dessa forma, o número de parâmetros é linear em relação ao número de atributos

Estimando densidade - Abordagem Não-paramétrica

- Em uma abordagem diferente, ao invés de limitarmos o volume (via h) tornamos o volume grande o suficiente para conter um mínimo de amostras
  - Regiões de alta densidade obtém h pequeno evitando a suavização artificial
  - Regiões de baixa densidade obtém h grande evitando ruído
  - Esta abordagem é chamada de *k-Nearest Neighbor Estimator*

$$\hat{p}(x) = \frac{1}{N d_k(x)} \sum_{n=1}^{N} K\left(\frac{x - x_n}{d_k(x)}\right)$$

$d_k(x)$ é a distância do k-ésimo vizinho mais próximo de x

Maldição da dimensionalidade

- Devemos levar em consideração a *maldição da dimensionalidade* quando o número de atributos for alto
  - número de regiões cresce exponencialmente com o número de dimensões
  - para estimar a densidade precisaríamos que o número de amostras acompanhasse esse crescimento

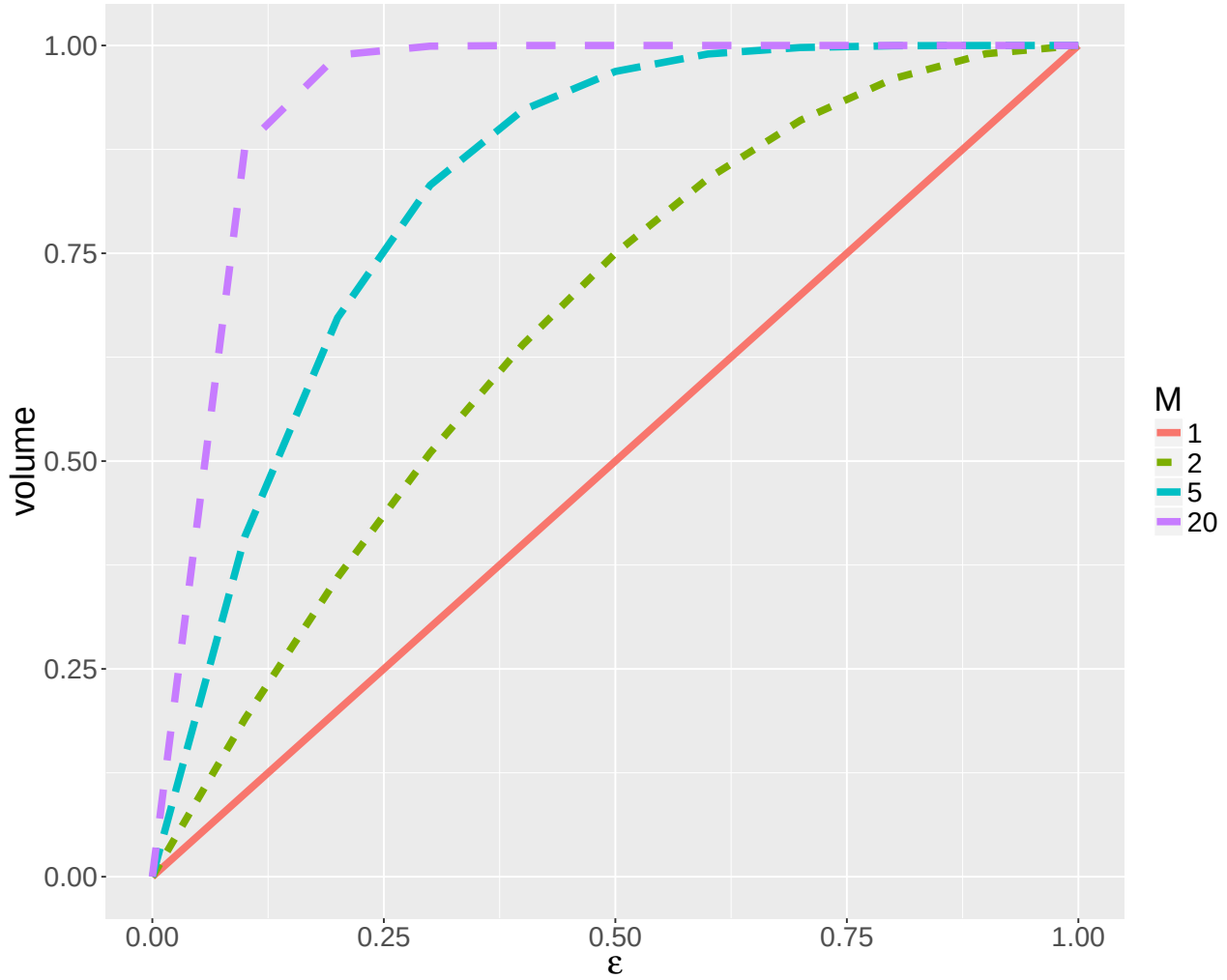
Maldição da dimensionalidade

- Nossa intuição pode nos enganar, por exemplo:
  - Considere uma esfera com raio r em M dimensões
  - Qual seria a fração do volume da esfera na área entre $r = 1 - \epsilon$ e $r = 1$
  - Volume de uma esfera com raio r em M dimensões é proporcional a r^M:

$V_M(r) = K_M r^M$, K_M é uma constante relacionada apenas a M

$$\frac{V_M(1) - V_M(1-\epsilon)}{V_M(1)} = 1 - (1-\epsilon)^M$$

Maldição da dimensionalidade



Maldição da dimensionalidade

- Portanto, a maior parte do volume em alta dimensionalidade vai estar concentrado próximo a superfície
- Discutiremos sobre como tentar evitar (amenizar) esse problema na aula sobre seleção de atributos

Referências

P. Tan, M. Steinbach e V. Kumar, Introduction to Data Mining. **Capítulo 3**

D. Hand, H. Manilla e P. Smith. Principles of Data Mining. **Capítulo 3**

E. Alpaydin, Introduction to Machine Learning. **Seção 8.2**

C. Bishop. Pattern Recognition and Machine Learning. **Seções 1.2.4 e 2.5.1**

R. Duda, P. Hart e D. Stork. Pattern Classification. **Seção 3.2**