Aula 8: Estimação Densidade de Probabilidade

Prof. Dr. Eder Angelo Milani

22/05/2023

Estimação de Densidades pelo Método Kernel

A estimação de densidade pelo método Kernel generaliza a ideia da estimação de densidades pelo histograma.

Considere a amostra $X_1, \ldots, X_n$, então a estimativa da densidade obtida pelo histograma é

$$\hat{f}(x) = \frac{1}{2hn} \times k,$$

sendo que k é o número de pontos amostrais no intervalo $(x-h; x+h)$. Este estimador pode ser escrito como

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} w\Big(\frac{x - X_i}{h}\Big),$$

sendo que $w(t) = \frac{1}{2} I(|t| < 1)$ é uma função peso. O estimador da densidade $\hat{f}(x)$ acima com $w(t) = \frac{1}{2} I(|t| < 1)$ é chamado de estimador de densidade ingênuo. Esta função peso tem a propriedade

$$\int_{-1}^{1} w(t)dt = 1,$$

e $w(t) \geq 0$, assim $w(t)$ é uma densidade de probabilidade com suporte no intervalo $[-1; 1]$.

Estimação de densidade pelo kernel substitui a função peso $w(t)$ no estimador ingênuo por uma função $K(.)$ chamada de função kernel, tal que

$$\int_{-\infty}^{\infty} K(t)dt = 1$$

Em estimação de densidade, $K(.)$ é usualmente uma função densidade de probabilidade simétrica. A função peso $w(t) = \frac{1}{2} I(|t| < 1)$ é chamada de kernel retangular. A densidade estimada em x é a soma dos retângulos localizados até uma distância de h unidade de x.

Suponha que $K(.)$ é um outra densidade de probabilidade simétrica centrada na origem, e defina por

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\Big(\frac{x - X_i}{h}\Big),$$

então $\hat{f}$ é uma função densidade de probabilidade. Por exemplo, $K(x)$ pode ser a densidade triangular em $[-1; 1]$ (kernel triangular) ou a densidade normal padrão (kernel Gaussiano). O estimador kernel triangular corresponde a soma de áreas de triangulos, em vez de retângulo. O estimador kernel Gaussiano centraliza uma densidade normal em cada valor observado.

A partir da definição do estimador de densidade do kernel acima, segue que certas propriedades de cotinuidade e diferenciabilidade de $K(x)$ também vale para $\hat{f}_K(x)$. Se $K(x)$ é uma densidade de probabilidade, então $\hat{f}_K(x)$ é contínua em x se $K(x)$ é contínua em x, e $\hat{f}_K(x)$ tem derivada de r-ésima ordem em x se $K^{(r)}(x)$ existe. Em particular, se $K(x)$ é um kernel Gaussiano, então $\hat{f}$ é contínua e tem derivada de todas as ordem.

O estimador de densidade histograma corresponde ao estimador de densidade kernel retangular. A largura da classe h é um parâmetro de suavização; pequenos valores de h revelam características locais da densidade, enquanto que grandes valores de h produzem uma estimativa de densidade mais uniforme. Na estimativa de densidade kernel, h é chamado de largura da banda, parâmetro de suavização ou largura da janela.

O efeito da variação da largura da banda é ilustrado na figura abaixo. Os $n = 10$ pontos amostrais na Figura são

$$-0,77; -0,60; -0,25; 0,14; 0,45; 0,64; 0,65; 1,19; 1,71; 1,74$$

sendo gerados a partir da distribuição normal padrão. À medida que a largura da janela h diminui, a estimativa de densidade se torna mais grosseira, enquanto que para h maior corresponde a estimativas de densidades mais suaves. (Este exemplo é apresentado simplesmente para ilustrar graficamente o método kernel - a estimativa da densidade não é muito útil para tal tamanho amostral).

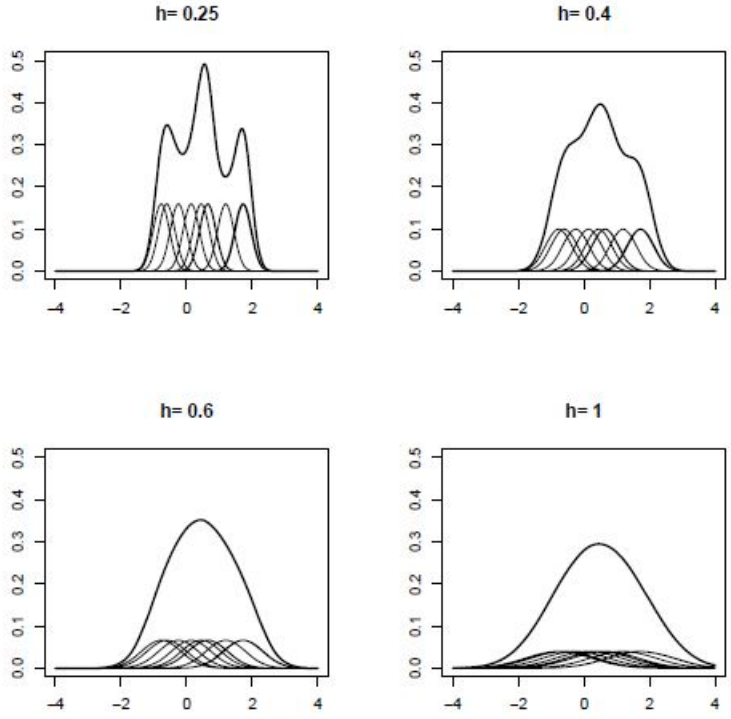


Figure 1: Estimação da densidade utilizando o kernel Gaussiano para diferentes valores de h

Os principais kernels são apresentados na Tabela a seguir. Destaca-se também o suporte de cada kernel.

Kernel	$K(t)$	Suporte	σ_k^2
Gaussiano	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right)$	$\mathbb{R}$	1
Epanechnikov	$\frac{3}{4}(1-t^2)$	$\lvert t\rvert < 1$	1/5
Retangular	$\frac{1}{2}$	$\lvert t\rvert < 1$	1/3
Triangular	$1-\lvert t\rvert$	$\lvert t\rvert < 1$	1/6
Biweight	$\frac{15}{16}(1-t^2)^2$	$\lvert t\rvert < 1$	1/7
Cosseno	$\frac{\pi}{4}\cos(\frac{\pi}{2}t)$	$\lvert t\rvert < 1$	1-8/π^2

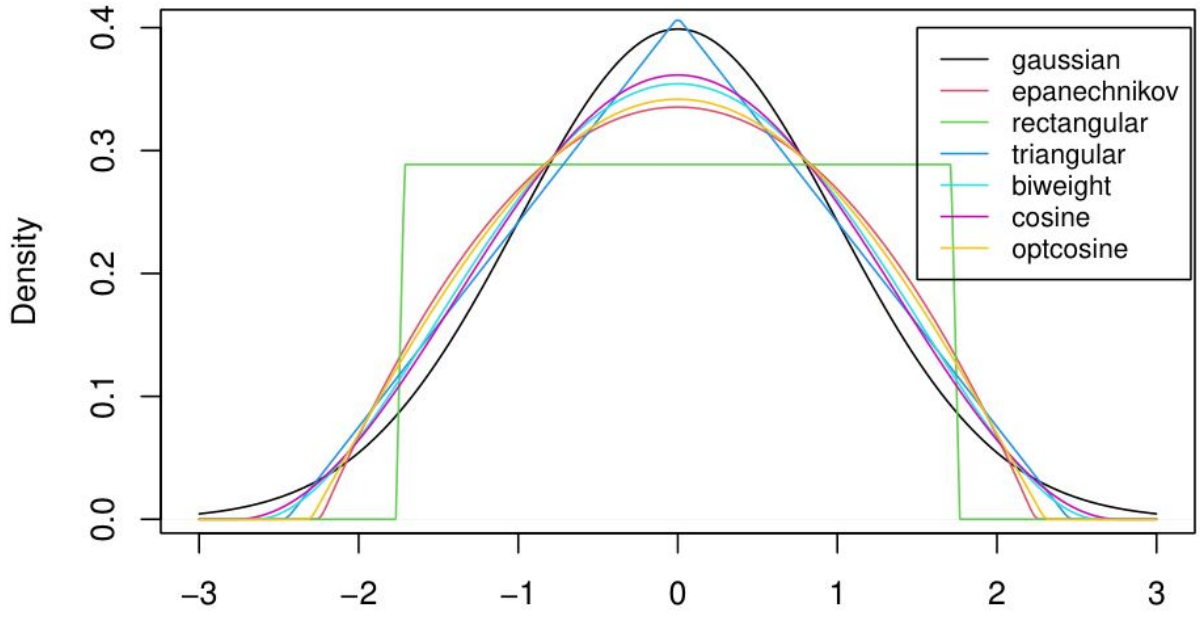


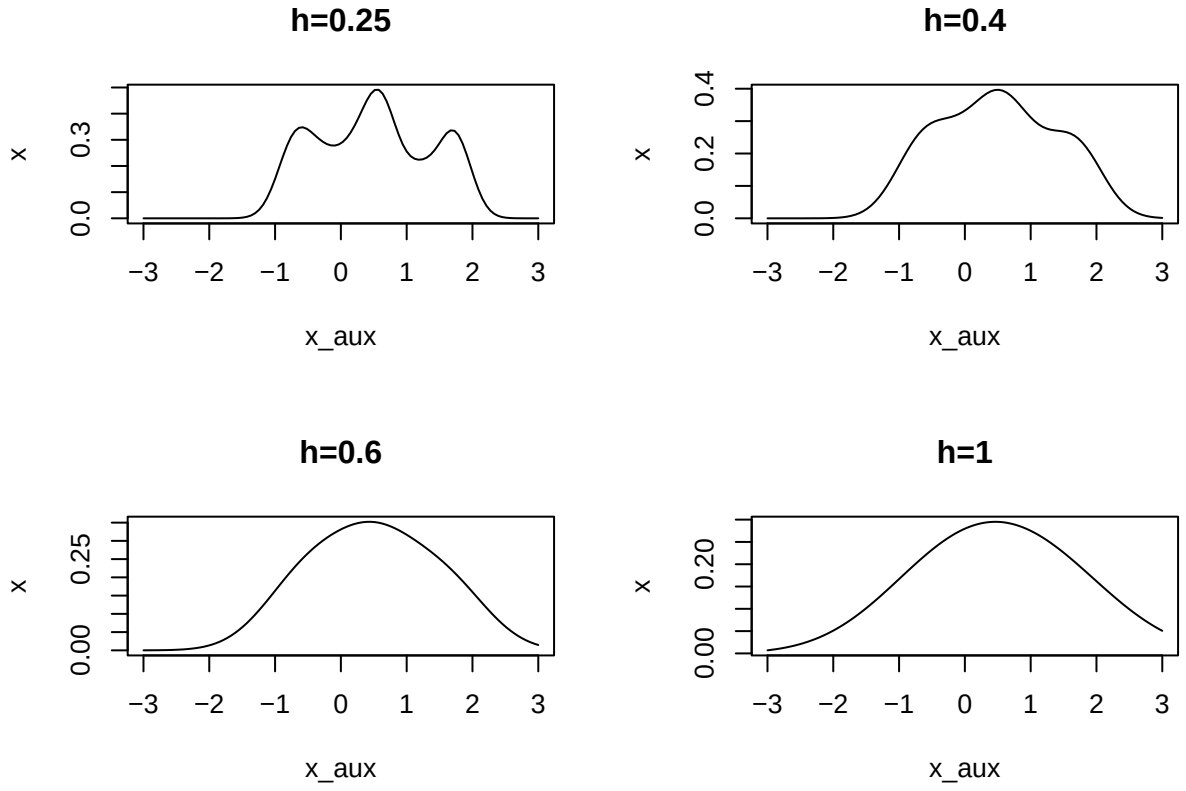
Figure 2: Ilustração dos tipos de kernels

Exemplo

A partir dos 10 valores amostrais dados anteriormente, apresentar a estimativa da densidade utilizando o kernel gaussiano.

```
X=c(-0.77, -0.60, -0.25, 0.14, 0.45, 0.64, 0.65, 1.19, 1.71, 1.74)

kernel_gau=function(X,h,x){
  n=length(X)
  t=(x-X)/h
  w=(2*pi)^(-0.5)*exp(-0.5*t^2)
  f=(1/n)*(1/h)*sum(w)
  return(f)
}

x_aux=seq(-3,3,length.out = 100)
```

```
x=numeric(100)
for( i in 1:100){
  x[i]=kernel_gau(X=X,h=0.25,x_aux[i])
}

par(mfrow=c(2,2))


plot(x_aux,x,type="l", main="h=0.25")

x=numeric(100)
for( i in 1:100){
  x[i]=kernel_gau(X=X,h=0.4,x_aux[i])
}

plot(x_aux,x,type="l", main="h=0.4")


x=numeric(100)
for( i in 1:100){
  x[i]=kernel_gau(X=X,h=0.6,x_aux[i])
}

plot(x_aux,x,type="l", main="h=0.6")


x=numeric(100)
for( i in 1:100){
  x[i]=kernel_gau(X=X,h=1,x_aux[i])
}

plot(x_aux,x,type="l", main="h=1")
```



Observação: Defini-se o erro quadrático médio integrado (EQMI) como

$$EQMI = \int E(\hat{f}(x) - f(x))^2 dx$$

1. Para o kernel Gaussiano, a largura da banda h que otimiza o EQMI é

$$h = (4/3)^{1/5}\sigma n^{-1/5} = 1.06\sigma n^{-1/5},$$

mas essa escolha é ideal quando a distribuição é normal. Se a densidade verdadeira não for unimodal, o resultado acima tenderá a ser excessivamente suave. Alternativamente, podemos usar uma estimativa mais robusta, sendo dada por

$$\hat{\sigma} = min(S, IQR/1.34)$$

sendo que S é o desvio padrão e IQR é o intervalo interquartil da amostra. Silverman (1978), indica que uma escolha melhor ainda para o kernel Gaussiano é dada por

$$h = 0.9\,\min(S, IQR/1.34)n^{-1/5},$$

sendo um bom ponto de partida para uma ampla gama de distribuições que não são necessariamente normais, unimodais ou simétricas.

Exercícios

1. Criar uma função no R que calcula o Kernel Epanechnikov
2. Criar uma função no R que calcula o Kernel Retangular
3. Criar uma função no R que calcula o Kernel Triangular
4. Criar uma função no R que calcula o Kernel biweight
5. Criar uma função no R que calcula o Kernel cosseno

A partir da função, encontre a estimativa da densidade para os seguintes cenários

- amostra de tamanho 100 da distribuição N(1,3)
- amostra de tamanho 200 da distribuição Exp(1)

6. Utilizar as funções criadas anteriormente e estimar a densidade das duas variáveis do dataset `geyser` do pacote `MASS`.