Nonparametric Statistics

Ch.8 Density estimation

Motivation

- ❖ Density estimation plays a fundamental role to understand data characteristics.

- ❖ As we learned so far, both parametric and nonparametric approaches exist for density estimation.

- ❖ Parametric methods are sometimes very restrictive or inadequate. Nonparametric density estimation methods are alternatives to parametric ones and also can be used by an initial step to see if parametric modeling is appropriate.

Parametric density estimation

- ❖ In the case that one believes that the density function of a random variable X belongs to a parametric family :

Ex] $X \sim N(\mu, \sigma^2)$, density estimator :

$$\hat{f}_X(x) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{(x-\bar{x})^2}{s^2}\right)$$

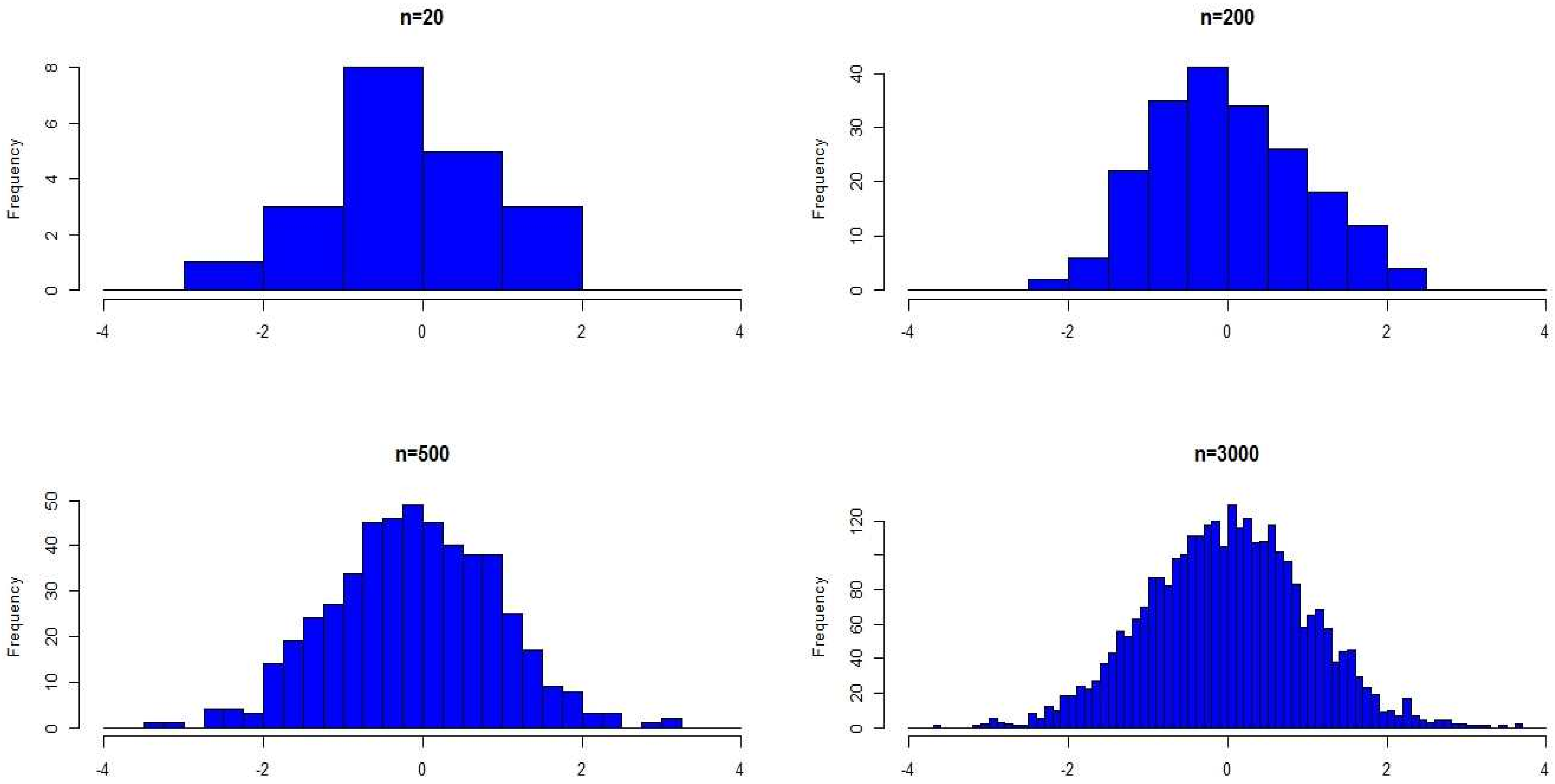
$X \sim \chi^2(r)$, density estimator :

$$\hat{f}_X(x) = \frac{1}{\Gamma\left(\frac{\bar{x}}{2}\right) 2^{\bar{x}/2}} x^{\frac{\bar{x}}{2}-1} \exp\left(-\frac{x}{2}\right)$$

- ❖ As noted previously, this is reliable only when $X \sim N(\mu, \sigma^2)$ or $X \sim \chi^2(r)$ is true. How can we assure?

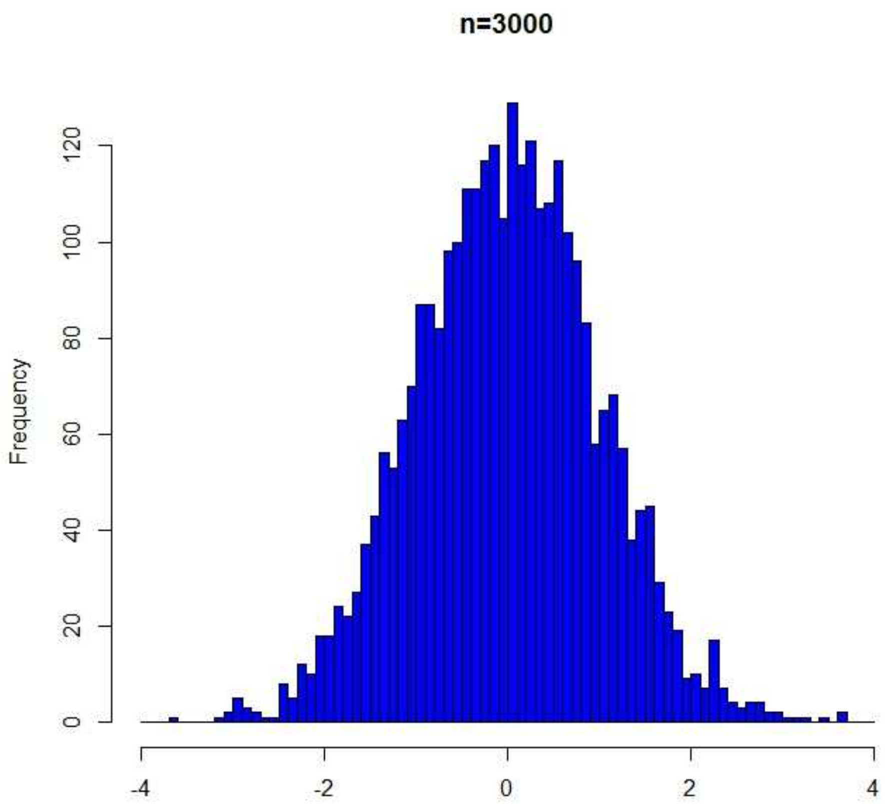
Histogram

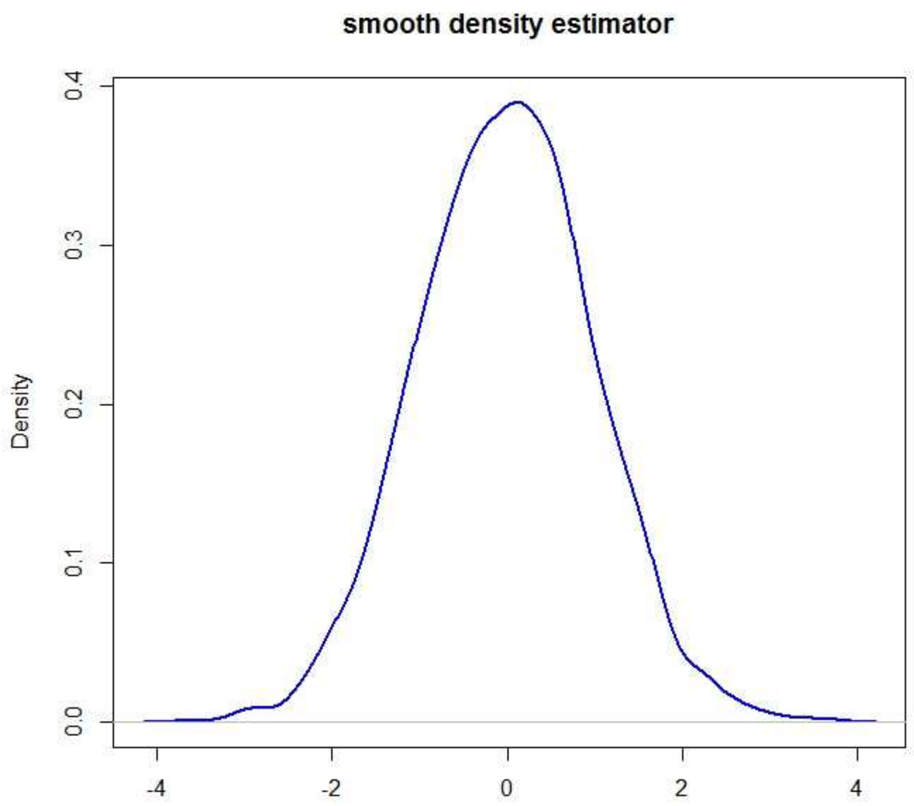
❖ The histogram is perhaps the simplest nonparametric density estimator.



Histogram

- ❖ Histogram and smoothed density estimator


n=3000


smooth density estimator

Histogram

❖ Mathematical representation of histogram : Let B_j , $j = 1, \dots, m$ be bins.

$$\hat{f}_H(x) = \frac{1}{nh} \left(\# \ of \ x_i's \ in \ the \ same \ bin \ as \ x\right) \ , \qquad h\text{: } bin \ width$$

Note that $\frac{(\# \ of \ x_i's \ in \ the \ same \ bin \ as \ x)}{n}$ is the relative frequency (R_j).

❖ Why do we divide it by h ?

- For any density function g, $\int g(x)dx = 1$.
- Note that the area of each bar corresponding each bin B_j equals to $height \ of \ bar \ (k \, R_j) \times bin \ width \ (h)$. Thus, $1 = \int \hat{f}_H(x)dx = \sum_{j=1}^{m} k \, R_j h = kh$. Therefore, $k = 1/h$.
- Dividing by h is for normalizing to 1.

Drawback of Histogram

- The histogram estimator is consistent, but has several disadvantages.

- Step function (constant over each bin)

- Results heavily depends on break points (selection of bins)

- Bin width choice

Kernel density estimator (KDE)

- Let $X_1, \dots . X_n$ be a sample from a population with pdf f. The kernel density estimator is defined as

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

where h is a bandwidth and K is a used-defined nonnegative and symmetric kernel function satisfying $K_h(u) = K(u/h)/h$ and $\int K(u)du = 1$.

- Note that

$$\int \hat{f}_h(x)dx = \frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{h}K\left(\frac{X_i - x}{h}\right)dx = \frac{1}{n}\sum_{i=1}^{n} 1 = 1$$

Kernel density estimator

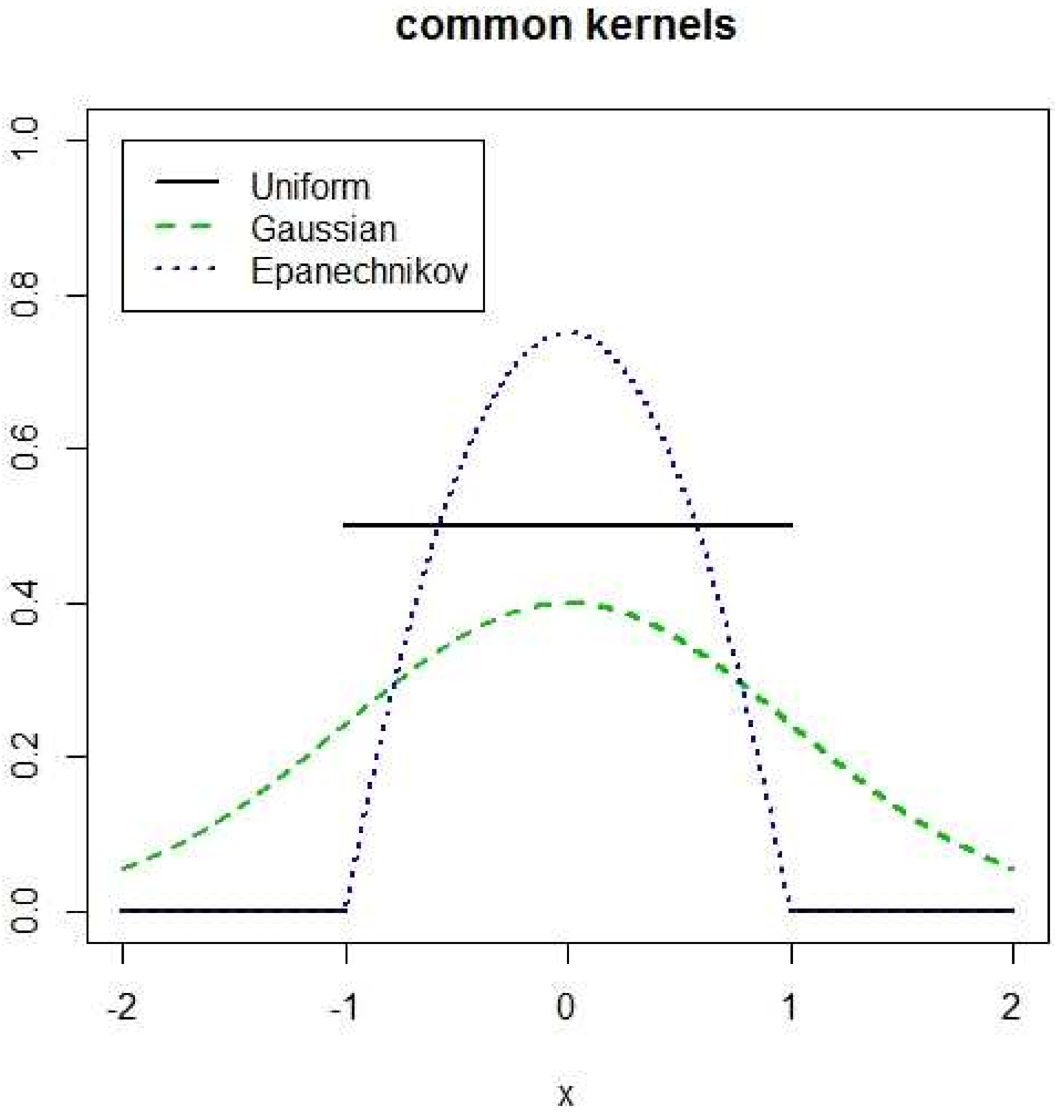
❖ Common kernel functions

- Gaussian kernel : $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$
- Epanechnikov kernel : $K(u) = 0.75(1 - u^2)I(|u| < 1)$
- Uniform kernel : $K(u) = 0.5\, I(|u| < 1)$ − produces discontinuous estimator

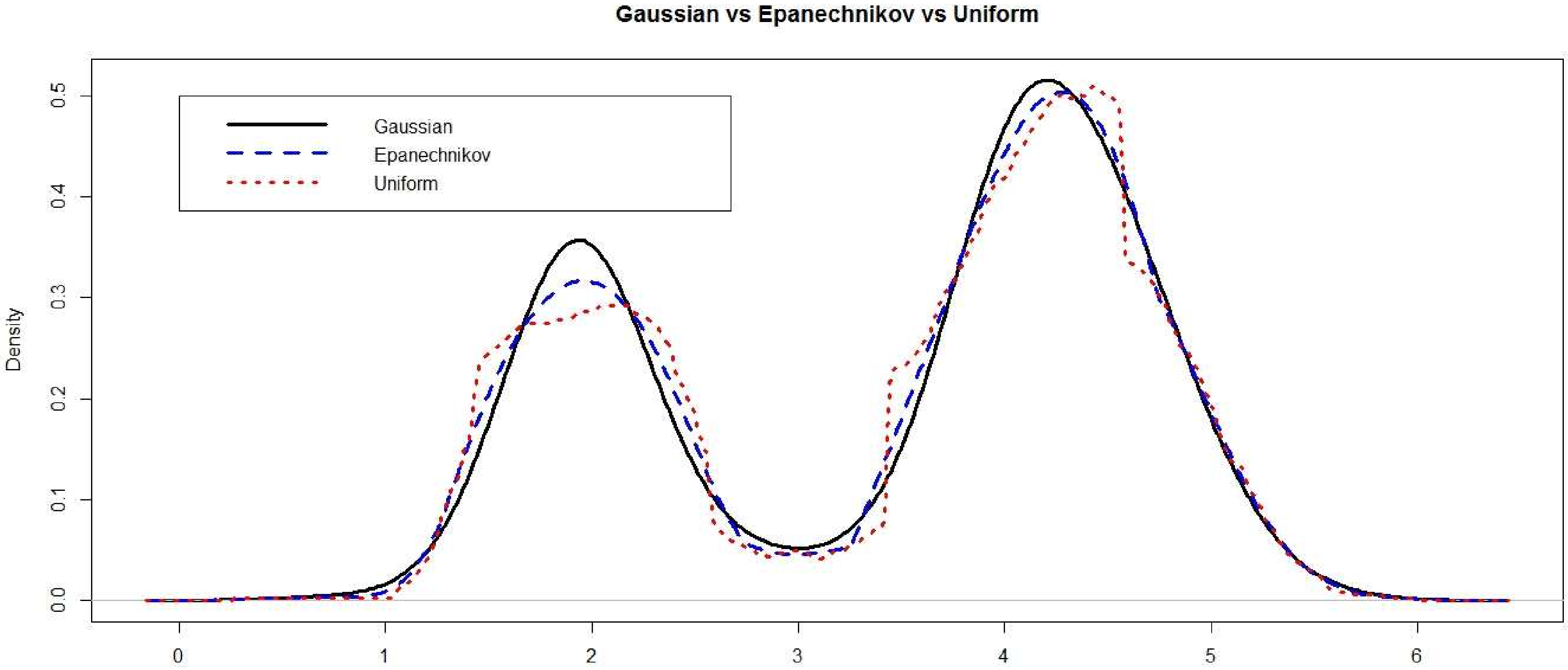
❖ It is widely accepted that the choice of kernel is not so important.

❖ Compactly supported kernels are sometimes preferred due to good statistical properties of resulting estimators.

Common kernels



Example : Geyser data (eruption times)



Theoretical properties

- Under some suitable conditions, the kernel density estimator is asymptotically unbiased provided that $h \to 0$ as $n \to \infty$.

- Consistency

  Under some suitable conditions, the kernel density estimator is consistent provided that $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

$$\hat{f}_h(x) \to f(x) \quad \text{in probability}$$

- Asymptotic Mean integrated Squared Error (AMISE) (or Risk) of KDE can be obtained. We omit the detailed formula.

$$AMISE(h) = R(h) = \int R(x;h)dx = \int E\left(\hat{f}_h(x) - f(x)\right)^2 dx$$

Bandwidth selection : important issue

- In the kernel density estimation, selecting h is very important since it directly affect the estimator. As described in the previous slide, the risk also depends on h.

- A large h produces a over-smoothed estimator, and a small h gives a rigid curve (undersmoothing).

- How to select h? There are two common approaches. The first one is based on Cross-Validation. The second one is based on minimizing the risk with respect to h.

Bandwidth selection : important issue

- ❖ It is known that the optimal h minimizing $R(h)$ is given as

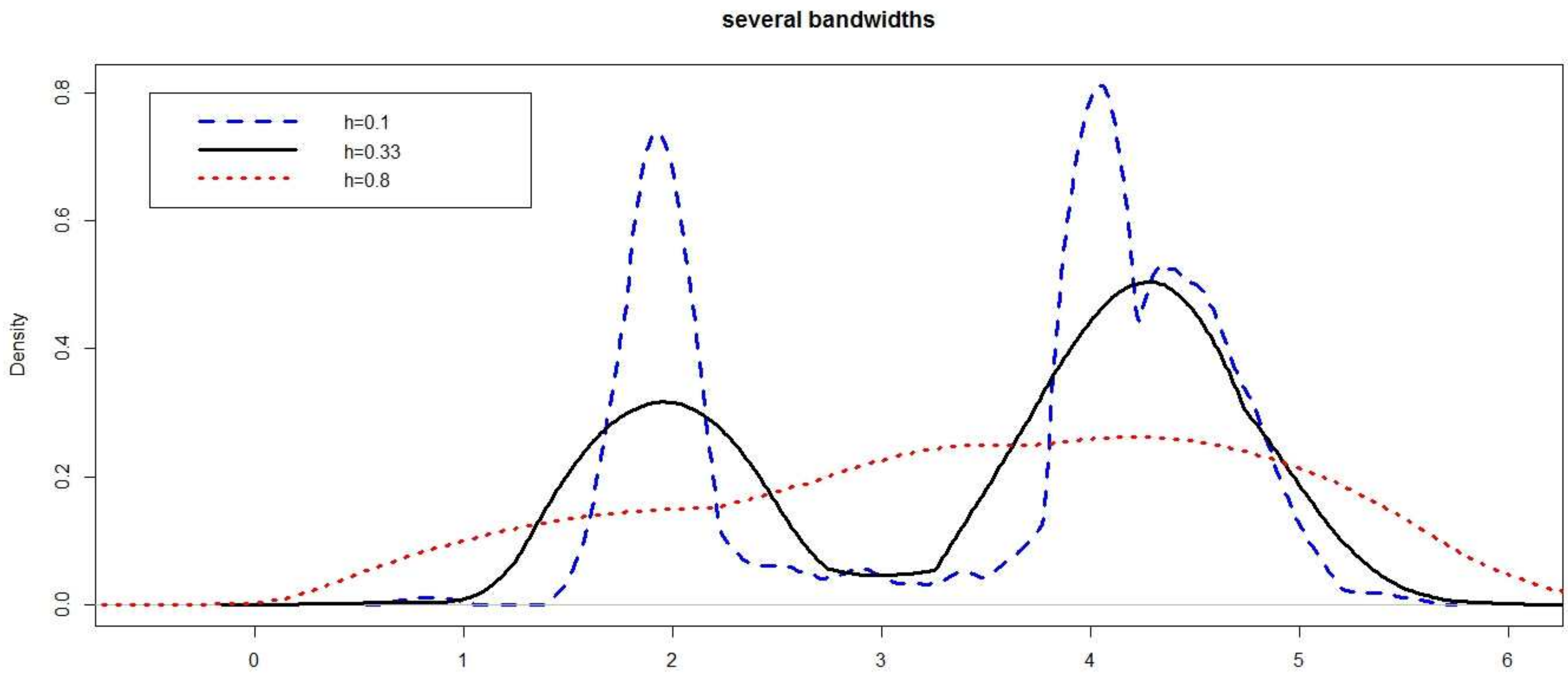
$$\hat{h}^{opt} = \left(\frac{\int K^2(u)du}{n\int u^2 K(u)du \int f'(x)^2 dx} \right)^{\frac{1}{5}}$$

- ❖ Note that $\hat{h}^{opt}$ still contain the unknown quantity $\int f'(x)^2 dx$ which needs to be estimated.

- ❖ A very simple and naïve approach to estimate $\int f'(x)^2 dx$ is to assume the normal distribution. This method is called “Rule-of-Thumb” bandwidth selection. This would work poor if the true distribution is far from normal.

Example : geyser data

❖ The epanechnikov kernel was used.



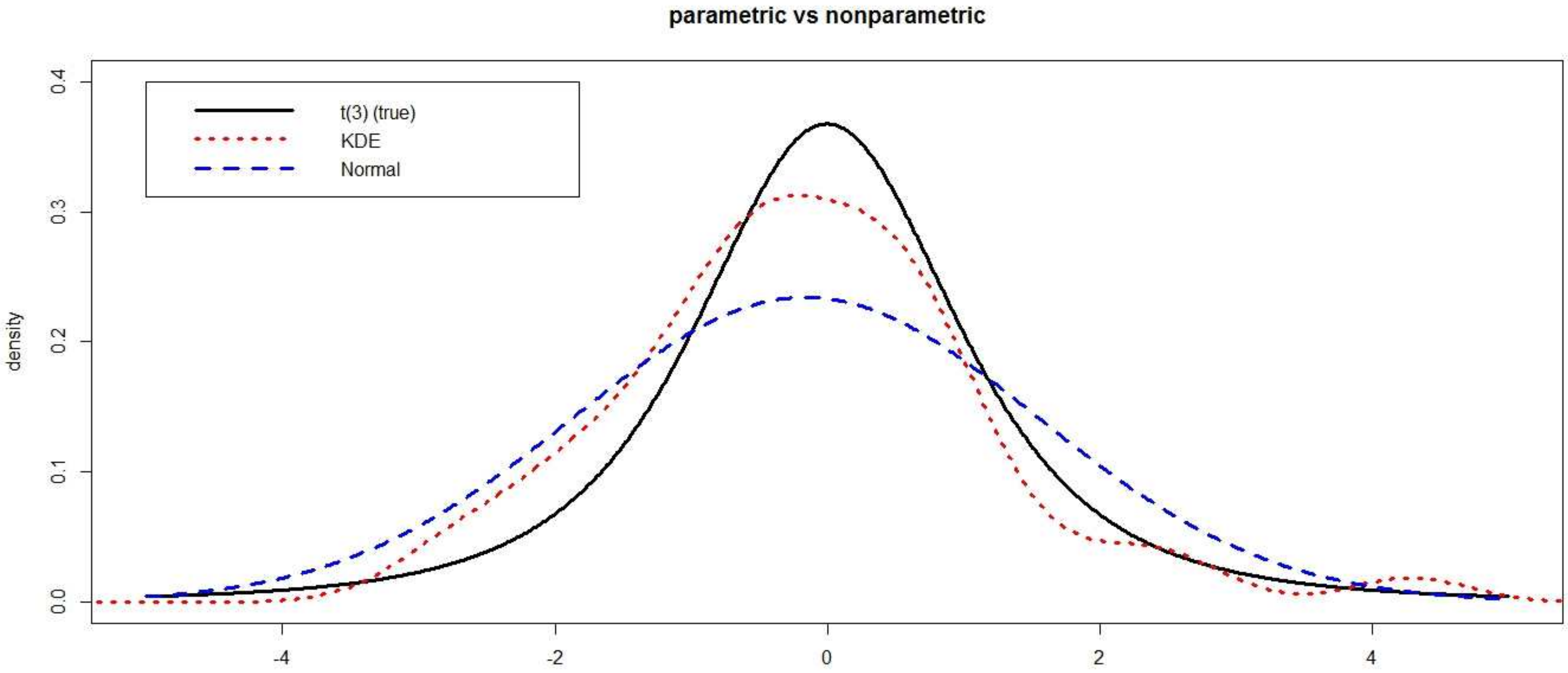
Optimal kernel

❖ The choice of kernel is not as important as that of bandwidth.

❖ When all other quantities are fixed, which kernel minimizes the risk? The answer is the “epanechnikov” kernel. So, this kernel is also called the optimal kernel.

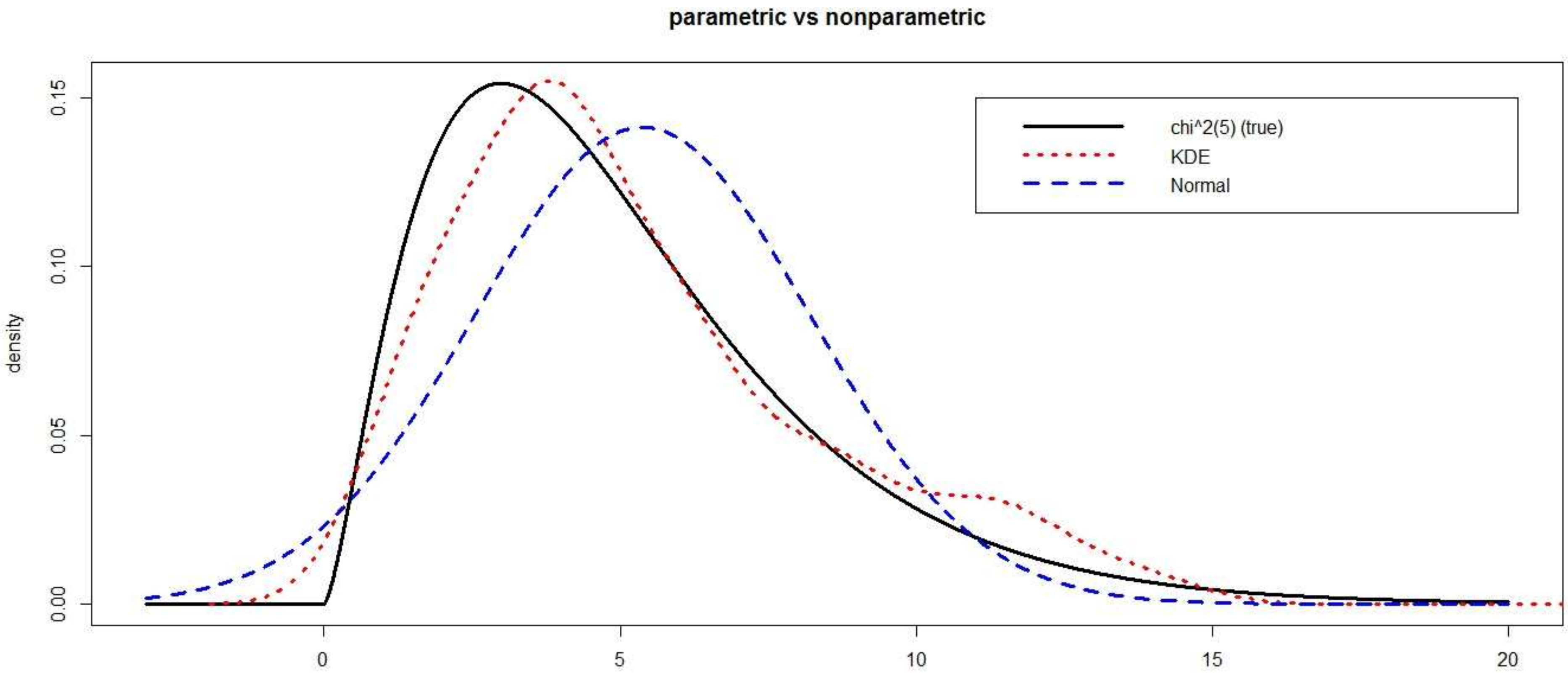
Example (Heavy tailed distribution)

- Suppose that there is a sample from $t(3)$ and we conjecture that the true distribution is normal.



Example (Asymmetric distribution)

- Suppose that there is a sample from $\chi^2(5)$ and we conjecture that the true distribution is normal.



Example

- Calculate KDE at $x = 2\ and\ \ 7/3$ with $h = 1\ and\ \ 2/3$.

$$1 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 4 \quad 4$$

$$\hat{f}_{h=1}(2) = \frac{1}{8 \cdot 1}\big(K(0) + K(0)\big) = \frac{3}{16}$$

$$\hat{f}_{h=1}\left(\frac{7}{3}\right) = \frac{1}{8 \cdot 1}\left(2 \cdot K\left(\frac{1}{3}\right) + 3 \cdot K\left(\frac{2}{3}\right)\right) = \frac{31}{96}$$

$$\hat{f}_{h=2/3}(2) = \frac{1}{8 \cdot 2/3}\big(K(0) + K(0)\big) = \frac{9}{32}$$

$$\hat{f}_{h=2/3}\left(\frac{7}{3}\right) = \frac{1}{8 \cdot 2/3}\left(2 \cdot K\left(\frac{1}{2}\right) + 3 \cdot K(1)\right) = \frac{27}{128}$$