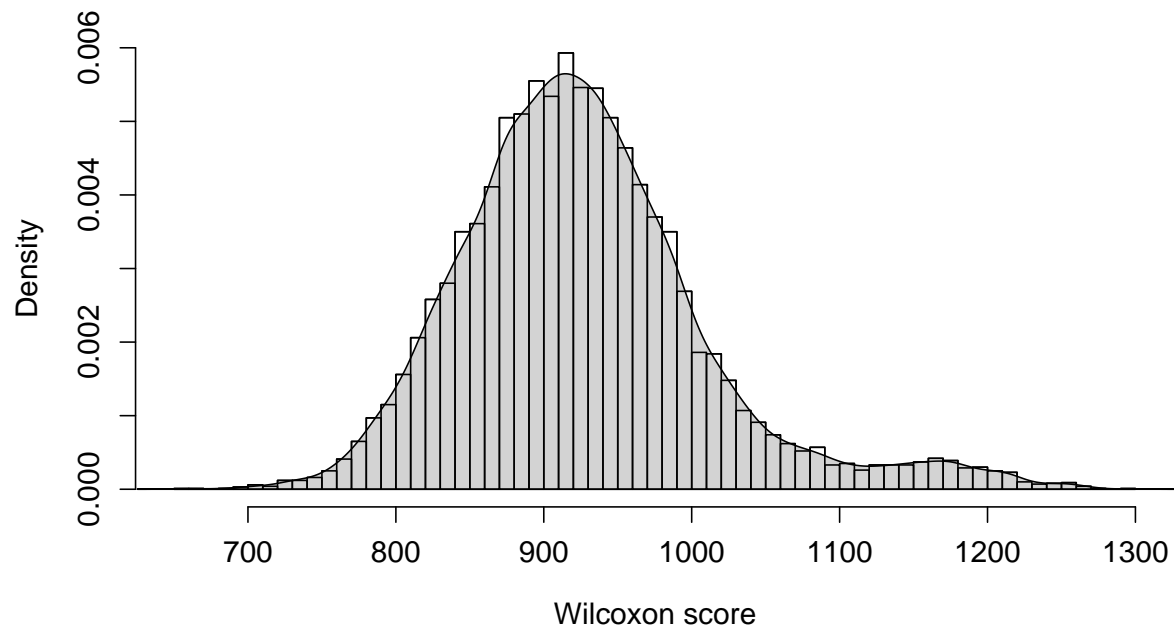


Introduction to Curve Estimation



Michael E. Tarter & Micheal D. Lock

Model-Free Curve Estimation

Monographs on Statistics and Applied Probability 56

Chapman & Hall, 1993.

Chapters 1–4.

Outline

1. Generalized representation
2. Short review on Fourier series
3. Fourier series density estimation
4. Kernel density estimation
5. Optimizing density estimates

Generalized representation

Estimation versus Specification



We are familiar
with its theory
and application.



How can we be
sure about the
underlying distribution?

Usual density representation:

- composed of elementary functions
- usually in closed form
- finite and rather small number of “personalized” parameters

Generalized representation:

- infinite number of parameters
- usually: representation as infinite sum of elementary functions
- Fourier series density estimation
- Kernel density estimation

Complex Fourier series

$$f(x) = \sum_{k=-\infty}^{\infty} B_k \exp\{2\pi i k x\}$$

- $x \in [0, 1]$.
- $\{B_k\}$ are called Fourier coefficients.
- Why can we represent any function in such a way?

Some useful features:

$\Psi_k = \exp\{2\pi i k x\}$, $\{\Psi_k\}$ forms an **orthonormal** sequence, that is

$$\int_0^1 \exp\{2\pi i(k-l)x\} dx = \begin{cases} 1 & k = l \\ 0 & k \neq l \end{cases}$$

$\{\Psi_k\}$ is **complete**, that is

$$\lim_{m \rightarrow \infty} \int_0^1 \left(f(x) - \sum_{k=-m}^m B_k \exp\{2\pi i k x\} \right)^2 dx = 0$$

Therefore, we can expand every function $f(x)$, $x \in [0, 1]$, in space L_2 with Fourier series.

L_2 function assumes that $\|f\|^2 = \int |f(x)|^2 dx < \infty$, which holds for most of the curves we are interested in.

Fourier series density estimation

Given an iid sample $\{X_j\}$, $j = 1, \dots, n$, with support on $[0, 1]$ (otherwise rescale).

Representation of true density:

$$f(x) = \sum_{k=-\infty}^{\infty} B_k \exp\{2\pi i k x\} \quad \text{with} \quad B_k = \int_0^1 f(x) \exp\{-2\pi i k x\} dx$$

Estimator:

$$\hat{f}(x) = \sum_{k=-\infty}^{\infty} b_k \hat{B}_k \exp\{2\pi i k x\} \quad \text{with} \quad \hat{B}_k = \frac{1}{n} \sum_{j=1}^n \exp\{-2\pi i k X_j\}$$

$\{b_k\}$ are called multipliers.

Estimator:

$$\hat{f}(x) = \sum_{k=-\infty}^{\infty} b_k \hat{B}_k \exp\{2\pi i k x\} \quad \text{with} \quad \hat{B}_k = \frac{1}{n} \sum_{j=1}^n \exp\{-2\pi i k X_j\}$$

$\{b_k\}$ are called multipliers.

Easy computation:

Use $\exp\{-2\pi i k X_j\} = \cos(2\pi k X_j) - i \sin(2\pi k X_j)$ and $\hat{B}_{-k} = \hat{B}_k^*$
(complex conjugate). $\hat{B}_0 \equiv 1$.

Therefore, computation only needed for positive k .

\hat{B}_k is **unbiased** estimator for B_k .

However, \hat{f} is usually **biased** because number of terms is either infinite or unknown.

Another advantage of sample coefficients $\{\hat{B}_k\}$: Same set leads to **variety of other estimates**.

That's where multipliers come into play!

Fourier multipliers

“Raw” density estimator:

$$b_k = \begin{cases} 1 & |k| \leq m \\ 0 & |k| > m \end{cases} \Rightarrow \hat{f}(x) = \sum_{k=-m}^m \hat{B}_k \exp\{2\pi i k x\}$$

Evaluate $\hat{f}(x)$ in equally spaced points $x \in [0, 1]$.

Estimating the expectation

$$\hat{\mu} = \int_0^1 x \hat{f}(x) dx = \dots = \frac{1}{2} + \sum_{\substack{k=-m \\ k \neq 0}}^m \frac{1}{2\pi i k} \hat{B}_k$$

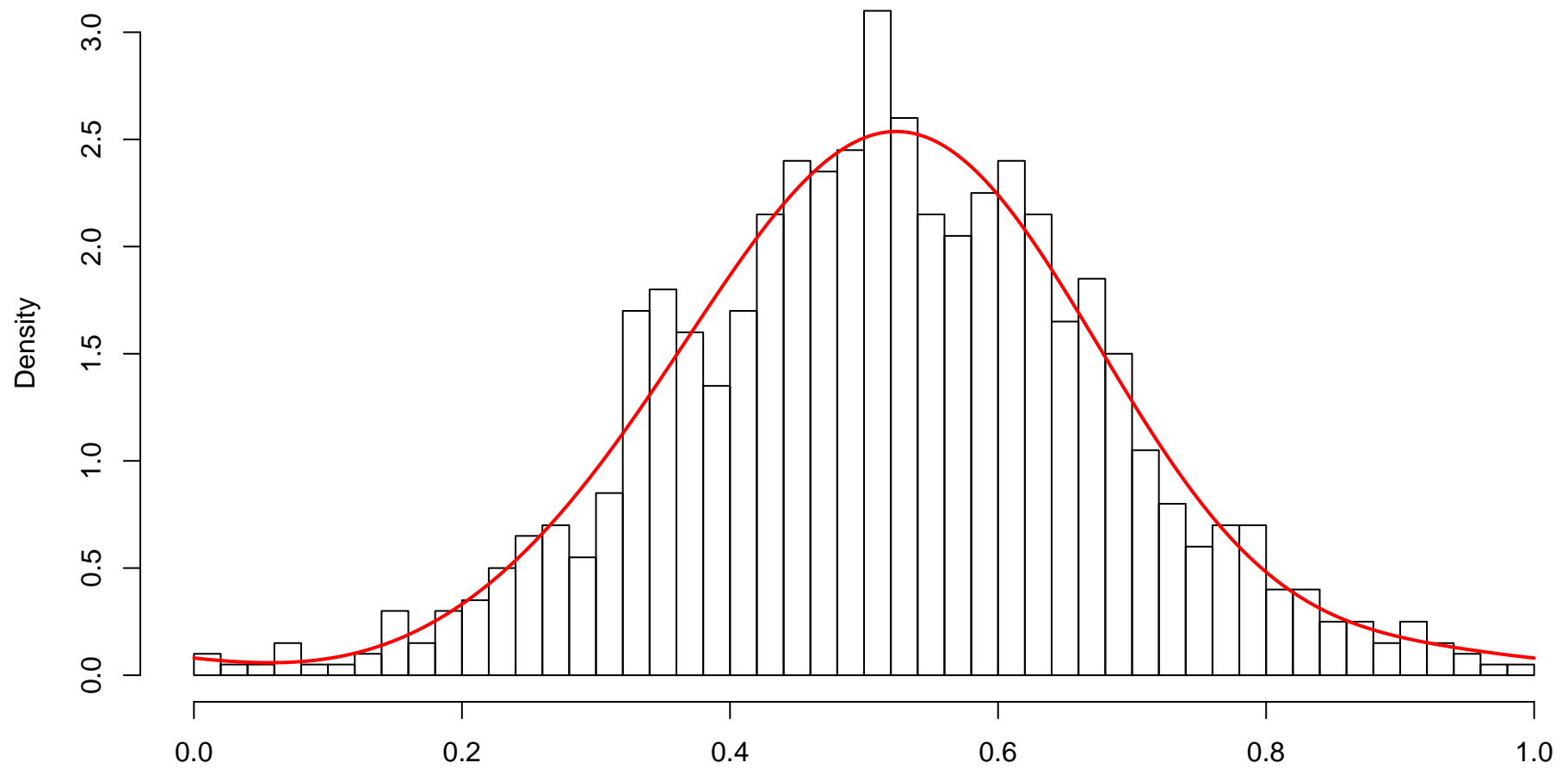
$$b_k = \begin{cases} (2\pi i k)^{-1} & |k| \leq m, k \neq 0 \\ 0 & |k| > m \end{cases}, \text{ evaluate at } x = 0 \text{ and add } \frac{1}{2}.$$

Advantages of multipliers

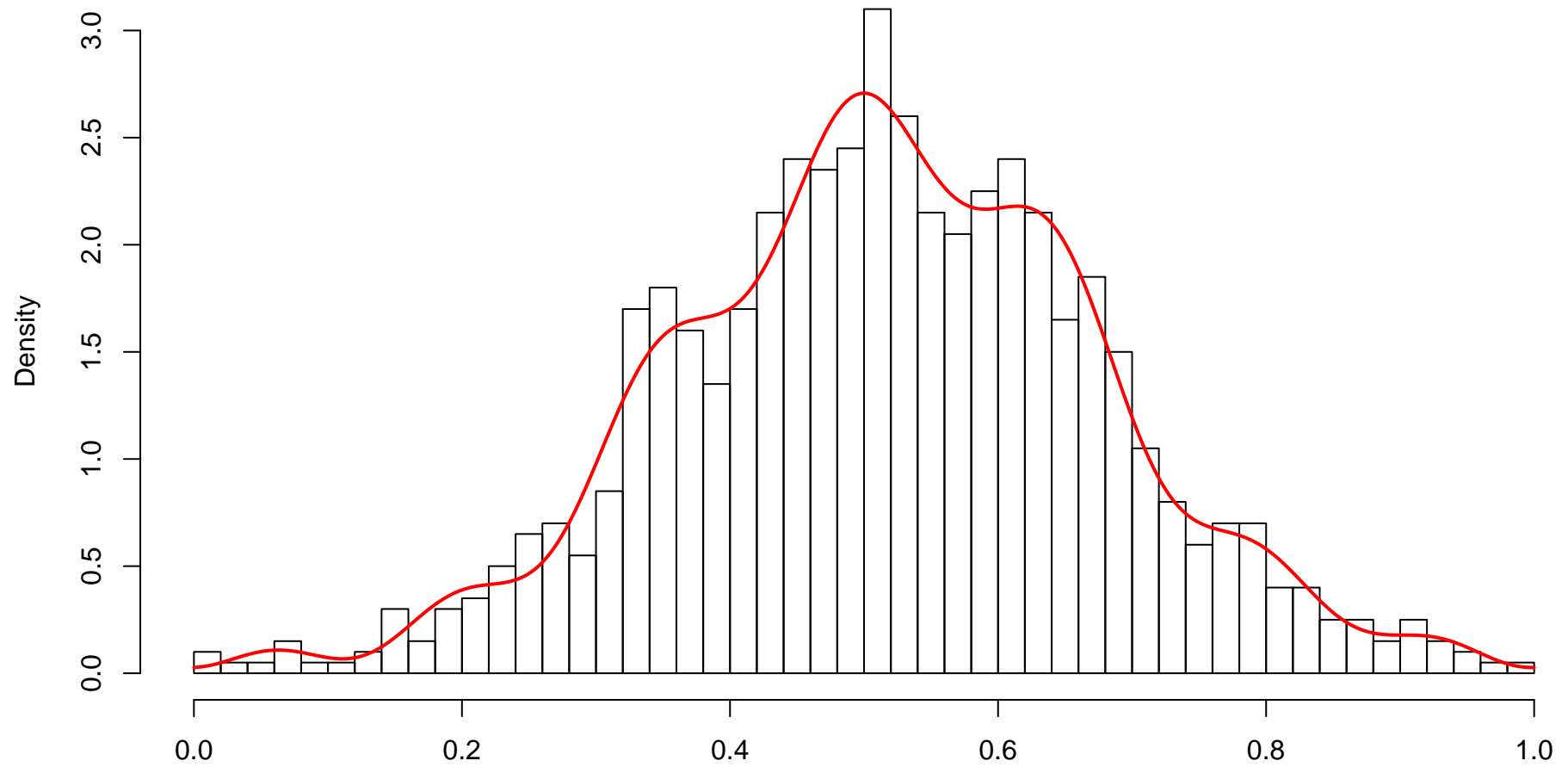
- Examination of various distributional features without recomputing sample coefficients.
- Optimize the estimation procedure.
- Smoothing of estimated curve vs. higher contrast.

Some examples . . .

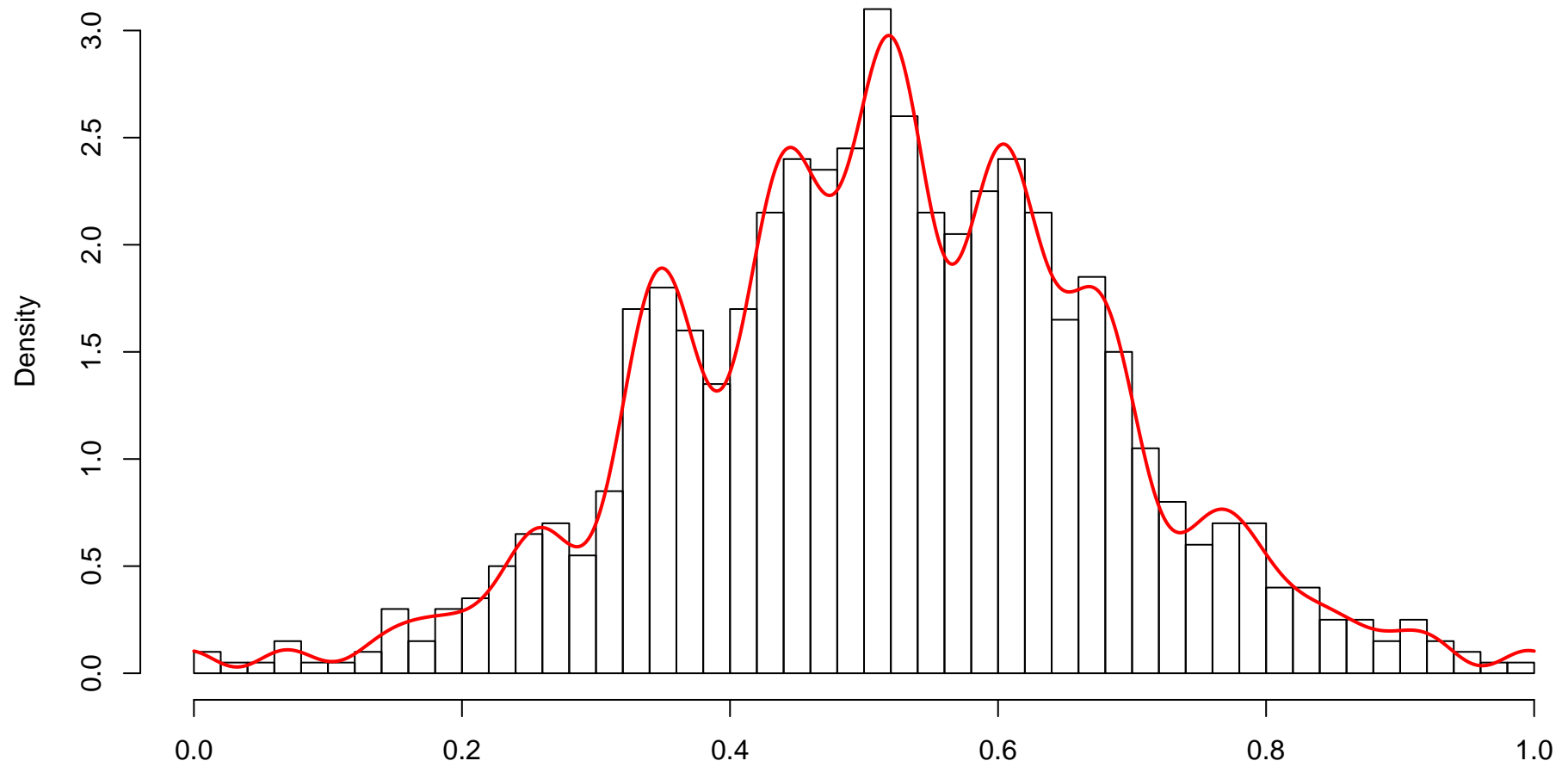
Raw Fourier series density estimator with $m = 3$



Raw Fourier series density estimator with $m = 7$



Raw Fourier series density estimator with $m = 15$



Kernel density estimation

Histograms are crude kernel density estimators where the kernel is a block (rectangular shape) somehow positioned over a data point.

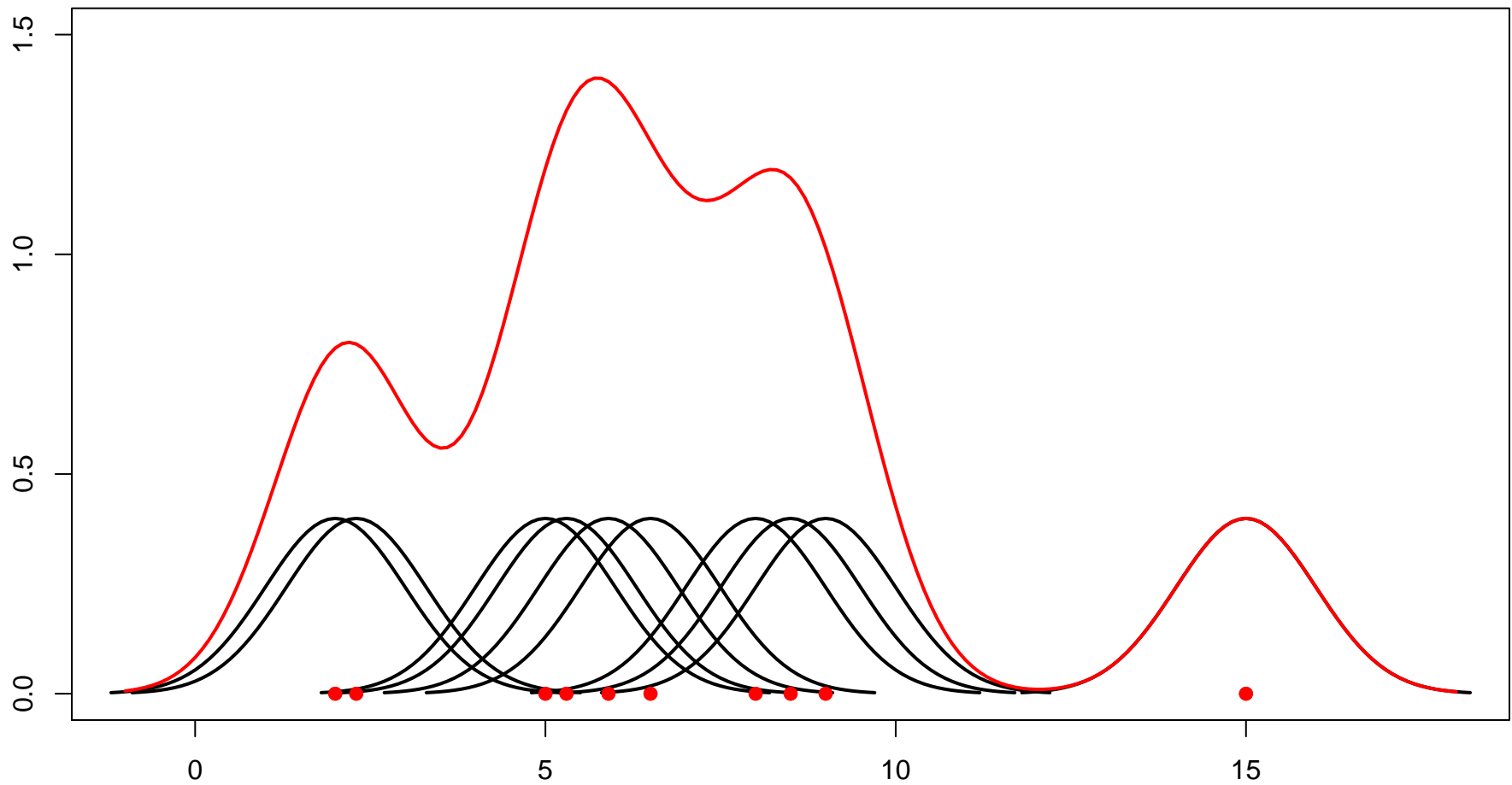
Kernel density estimation

Histograms are crude kernel density estimators where the kernel is a block (rectangular shape) somehow positioned over a data point.

Kernel estimators:

- use **various shapes** as kernels
 - place the **center** of a kernel right over the data point
 - spread the influence of one point with **varying kernel width**
- ⇒ contribution from each kernel is summed to overall estimate

Gaussian kernel density estimate



Kernel estimator

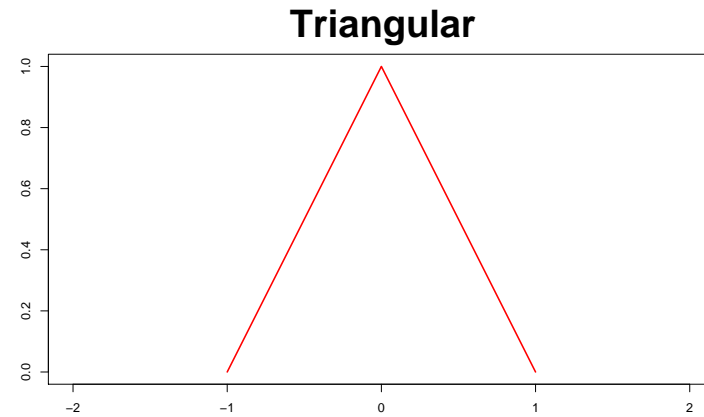
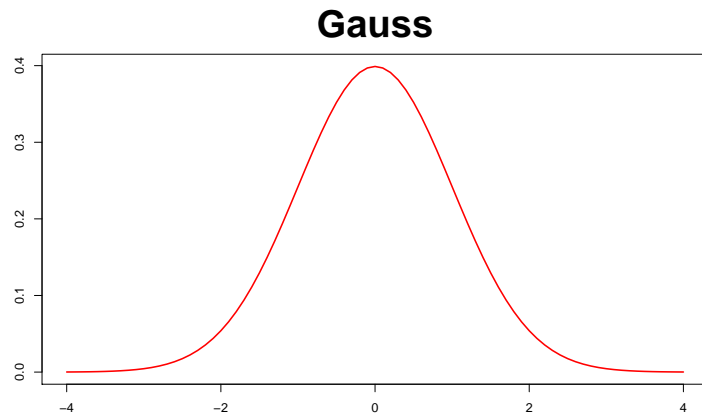
$$\hat{f}(x) = \frac{1}{nh} \sum_{j=1}^n K\left(\frac{x - X_j}{h}\right)$$

- h is called **bandwidth** or **smoothing parameter**.
- K is the **kernel function**: nonnegative and symmetric such that

$$\int K(x)dx = 1 \quad \text{and} \quad \int xK(x)dx = 0.$$

- Under mild conditions (h must decrease with increasing n) the kernel estimate converges in probability to the true density.
- Choice of kernel function usually depends on computational criteria.
- Choice of bandwidth is more important (see literature on “Kernel Smoothing”).

Some kernel functions



$$K(y) = \frac{3(1 - y^2/5)}{4\sqrt{5}}, \quad |y| \leq \sqrt{5}$$

Duality of Fourier series and kernel methodology

$$\begin{aligned}\hat{f}(x) &= \sum_k b_k \hat{B}_k \exp\{2\pi i k x\} \\ &= \frac{1}{n} \sum_{j=1}^n \sum_k b_k \exp\{2\pi i k (x - X_j)\}\end{aligned}$$

With $h = 1$:

$$K(x) = \sum_k b_k \exp\{2\pi i k x\}$$

The Dirichlet kernel

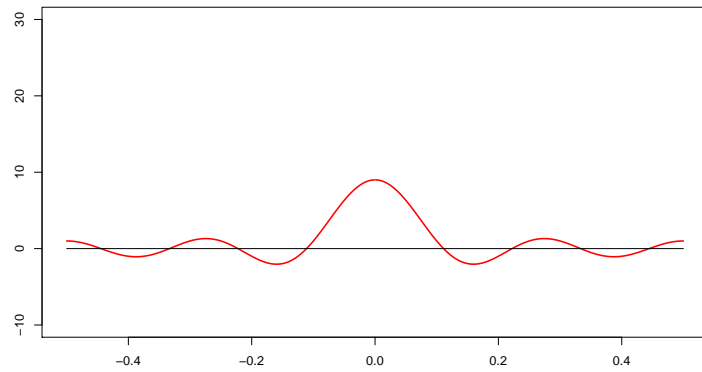
The raw density estimator has kernel K_D :

$$K_D(x) = \sum_{k=-m}^m \exp\{2\pi i k x\} = \dots = \frac{\sin(\pi(2m+1)x)}{\sin(\pi x)}$$

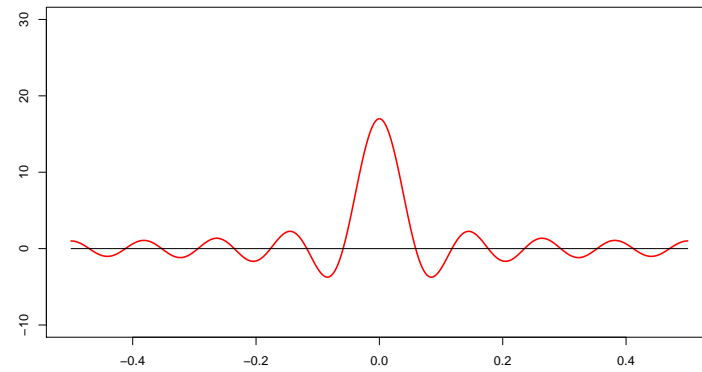
where $\lim_{x \rightarrow 0} K_D(x) = 2m + 1$.

Dirichlet kernels

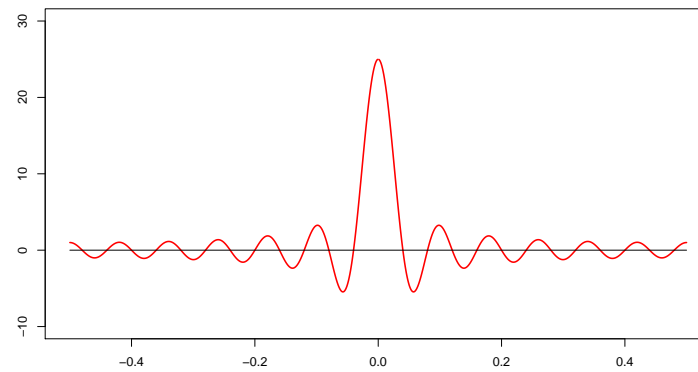
Dirichlet with $m = 4$



Dirichlet with $m = 8$



Dirichlet with $m = 12$



Differences between kernel and Fourier representation

- Fourier estimates are restricted to finite intervals while some kernels are not.
- As Dirichlet kernel shows, kernel estimates can result in negative values if the kernel function takes on negative values.

- **Kernel:** K controls shape, h controls spread of kernel.
Two-step strategy: Select kernel function and choose data-dependent smoothing parameter.
- **Fourier:** m controls both shape and spread.
Goodness-of-fit can be governed by entire multiplier sequence.

Optimizing density estimates

Optimization with regard to **weighted mean integrated square error (MISE)**:

$$J(\hat{f}, f, w) = \mathbb{E} \int_0^1 \left(f(x) - \hat{f}(x) \right)^2 w(x) dx.$$

$w(x)$ is nonnegative weight function to emphasize estimation over subregions. First consider optimization with $w(x) \equiv 1$.

The raw density estimator again

$$J(\hat{f}, f) = 2 \sum_{k=1}^m \frac{1}{n} (1 - |B_k|^2) + 2 \sum_{k=m+1}^{\infty} |B_k|^2$$

The raw density estimator again

$$J(\hat{f}, f) = 2 \sum_{k=1}^m \frac{1}{n} (1 - |B_k|^2) + 2 \sum_{k=m+1}^{\infty} |B_k|^2$$



Variance component



Bias component

Single term stopping rule

- Estimate $\Delta J_s = J(\hat{f}_s, f) - J(\hat{f}_{s-1}, f)$, gain of including s th Fourier coefficient. MISE is decreased if ΔJ_s is negative.
- Include terms only if their inclusion results in negative difference.
Multiple testing problem!
- Inclusion of higher-order terms results in rough estimate.
- Suggestions: Stop after t successive nonnegative inclusions. Choice of t is data/curve dependent.

Other stopping rules

- Different considerations about estimating MISE lead to various optimization concepts.
- Not at all generally superior to single term rule. Depends on curve features.

Multiplier sequences

- So far: “raw” estimate with $b_k = 1$ or $b_k = 0$.
- Now allow $\{b_k\}$ to be **sequence tending to zero** with increasing k .
- Concepts depend again on considerations about MISE.
- Question of advisable stopping rule remains.

Two-step strategy with multiplier sequence

1. Estimate with raw estimator and one of former stopping rules.
2. Applying a multiplier sequence to the remaining terms will always improve the estimate.

Weighted MISE

$$J(\hat{f}, f, w) = \mathbb{E} \int_0^1 \left(f(x) - \hat{f}(x) \right)^2 w(x) dx.$$

- Weight functions $w(x)$ emphasize **subregions** of support interval (e.g. left or right tails).
 - Turns out that unweighted MISE leads to great accuracy in regions with high density.
- ⇒ **Weighting will improve estimate** when other regions are of interest.

Data transformation

- Data needs rescaling to $[0, 1]$. Always possible: $\frac{X_j - \min(X)}{\max(X) - \min(X)}$
- Next approach: Transform data in nonlinear manner to emphasize subregions.

Data transformation

- Data needs rescaling to $[0, 1]$. Always possible: $\frac{X_j - \min(X)}{\max(X) - \min(X)}$
 - Next approach: Transform data in nonlinear manner to emphasize subregions.
 - Let $G : [a, b] \rightarrow [0, 1]$ be strictly increasing one-to-one function with $g(x) = \frac{dG(x)}{dx}$.
- $\Rightarrow \Psi_k(G(x)) = \exp\{2\pi i k G(x)\}$ is orthonormal on $[a, b]$ with respect to weight $g(x)$.

Transformation and optimization

- Data transformation with $G(x)$ is equivalent to weighted MISE with $w(x) = 1/g(x)$.
 - Only difference to unweighted MISE: Computation of Fourier coefficients involves application of $G(x)$.
- ⇒ Strategy: Transform data, optimize with unweighed procedures, retransform.
- Most efficient: Transform data to unimodal symmetric distribution.

Application to gene expression data

- **Problem:** Fitting two distributions to another by removing a minimal number of data points.
- **Idea:** Estimate the two densities in an optimal manner. Remove points until goodness-of-fit is high with regard to modified MISE.