

Kernel Estimator and Bandwidth Selection for Density and its Derivatives

The `kedd` Package
Version 1.0.0

by Arsalane Chouaib Guidoum*

14 August 2013

1 Introduction

In statistics, the univariate kernel density estimation (KDE) is a non-parametric way to estimate the probability density function $f(x)$ of a random variable X , is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample. This techniques are widely used in various inference procedures such as signal processing, data mining and econometrics, see e.g., Silverman [1986], Wand and Jones [1995], Jeffrey [1996], Wolfgang et al [2004], Alexandre [2009]. The kernel estimator are standard in many books with applications and computer vision, see Wolfgang [1991], Scott [1992], Bowman and Azzalini [1997], Venables and Ripley [2002], for computational complexity and with implementation in `S`, for an overview. Estimation of the density derivatives also comes up in various other applications like estimation of modes and inflexion points of densities, a good list of applications which require the estimation of density derivatives can be found in Singh [1977].

There already exist a number of packages that can perform kernel density estimation in `R` (`density` in `R` base); see for example `KernSmooth` [Wand and Ripley, 2013], `sm` [Bowman and Azzalini, 2013], `np` [Tristen and Jeffrey, 2008] and `feature` [Duong and Matt, 2013], they exist also of functions for kernel density derivative estimation (KDDE), e.g., `kdde` in `ks` package [Duong, 2007]. We introduce in this vignette a new `R` package `kedd` [Guidoum, 2013] for use with the statistical programming environment `R Development Core Team` [2013], which implements smoothing techniques and computing bandwidth selectors of the r^{th} derivative of a probability density $f(x)$ for univariate data, using several kernels functions.

2 Convolutions and derivatives in kernels

In non-parametric statistics, a kernel is a weighting function used in non-parametric estimation techniques. Kernels are used in kernel density estimation to estimate random variables density functions $f(x)$, or in kernel regression to estimate the conditional expectation of a random variable, see e.g., Silverman [1986], Wand and Jones [1995]. In general any functions having the following assumptions can be used as a kernel:

(A1) $K(x) \geq 0$ and $\int_{\mathbb{R}} K(x)dx = 1$.

(A2) Symmetric about the origin, e.g., $\int_{\mathbb{R}} xK(x)dx = 0$.

*Department of Probabilities & Statistics.
Faculty of Mathematics.
University of Science and Technology Houari Boumediene.
BP 32 El-Alia, U.S.T.H.B, Algeria.
acguidoum@usthb.dz

(A3) Has finite second moment, e.g., $\mu_2(K) = \int_{\mathbb{R}} x^2 K(x) dx < \infty$. We denote $R(K) = \int_{\mathbb{R}} (K(x))^2 dx$. If $K(x)$ is a kernel, then so is the function $\bar{K}(x)$ defined by $\bar{K}(x) = \lambda K(\lambda x)$, where $\lambda > 0$, this can be used to select a scale that is appropriate for the data. The kernel function is very important to spreading a probability mass of $1/n$, the most widely used kernel is the Gaussian of zero mean and unit variance. Some classical of kernel function $K(x; r)$ (r is the maximum derivative of kernel) in **kedd** package are the following:

Kernel	$K(x; r)$	$R(K)$	$\mu_2(K)$
Gaussian	$K(x; \infty) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) 1_{]-\infty, +\infty[}$	$1/(2\sqrt{\pi})$	1
Epanechnikov	$K(x; 2) = \frac{3}{4} (1 - x^2) 1_{(x \leq 1)}$	3/5	1/5
Uniform	$K(x; 0) = \frac{1}{2} 1_{(x \leq 1)}$	1/2	1/3
Triangular	$K(x; 1) = (1 - x) 1_{(x \leq 1)}$	2/3	1/6
Triweight	$K(x; 6) = \frac{35}{32} (1 - x^2)^3 1_{(x \leq 1)}$	350/429	1/9
Tricube	$K(x; 9) = \frac{70}{81} (1 - x ^3)^3 1_{(x \leq 1)}$	175/247	35/243
Biweight	$K(x; 4) = \frac{15}{16} (1 - x^2)^2 1_{(x \leq 1)}$	5/7	1/7
Cosine	$K(x; \infty) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}x\right) 1_{(x \leq 1)}$	$\pi^2/16$	$(-8 + \pi^2)/\pi^2$

Table 1: Kernel functions in **kedd** package.

The r^{th} derivative of the kernel function $K(x)$ is written as:

$$K^{(r)}(x) = \frac{d^r}{dx^r} K(x) \quad (1)$$

and convolution of $K^{(r)}(x)$ is:

$$K^{(r)} * K^{(r)}(x) = \int_{\mathbb{R}} K^{(r)}(x) K^{(r)}(x - y) dy \quad (2)$$

for example the r^{th} derivative of the Gaussian kernel is given by:

$$K^{(r)}(x) = (-1)^r H_r(x) K(x)$$

and the r^{th} convolution can be written as:

$$K^{(r)} * K^{(r)}(x) = (-1)^{2r} \int_{\mathbb{R}} H_r(x) H_r(x - y) K(x) K(x - y) dy$$

where $H_r(x)$ is the r^{th} Hermite polynomial, see e.g., [Olver et al \[2010\]](#). We use **kernel.fun** for kernel derivative defined by (1), and **kernel.conv** for kernel convolution defined by (2).

For example the first derivative of the Gaussian kernel displayed on the left in Figure 1. On the right is the first convolution of the Gaussian kernel.

```
> library(kedd)
> kernel.fun(x = seq(-0.02, 0.02, by=0.01), deriv.order = 1, kernel = "gaussian")$kx
[1] 0.007977250 0.003989223 0.000000000 -0.003989223 -0.007977250
> kernel.conv(x = seq(-0.02, 0.02, by=0.01), deriv.order = 1, kernel = "gaussian")$kx
[1] -0.1410051 -0.1410368 -0.1410474 -0.1410368 -0.1410051
> plot(kernel.fun(deriv.order = 1, kernel = "gaussian"))
> plot(kernel.conv(deriv.order = 1, kernel = "gaussian"))
```

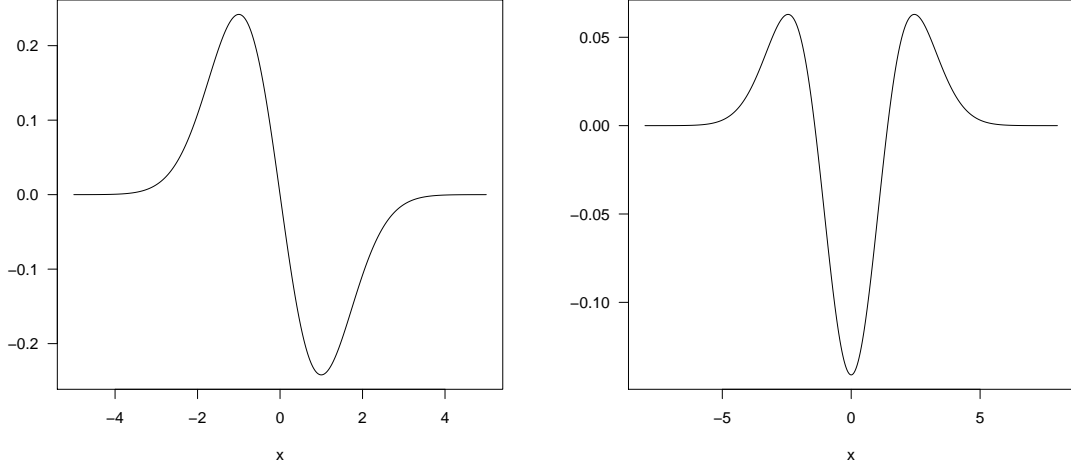


Figure 1: (Left) First derivative of the Gaussian kernel. (Right) Convolution of the first derivative Gaussian kernel.

3 Kernel density derivative estimator

Let (X_1, X_2, \dots, X_n) be a data sample, independent and identically distributed of a continuous random variable X , with density function $f(x)$. If the kernel K is differentiable r times then a natural estimator of the r^{th} derivative of $f(x)$ the r^{th} derivative of the kernel estimate [Bhattacharya, 1967, Schuster, 1969, Alekseev, 1972]:

$$\hat{f}_h^{(r)}(x) = \frac{d^r}{dx^r} \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) = \frac{1}{nh^{r+1}} \sum_{i=1}^n K^{(r)}\left(\frac{x - X_i}{h}\right) \quad (3)$$

where $K^{(r)}$ is r^{th} derivative of the kernel function K , which we take to be a symmetric probability density with at least r non zero derivatives when estimating $f^{(r)}(x)$, and h is the bandwidth, this parameter is very important that controls the degree of smoothing applied to the data.

The following assumptions on the density $f^{(r)}(x)$, the bandwidth h , and the kernel K :

- (A4) The $(r + 2)$ derivatives $f^{(r+2)}(x)$ is continuous, square integrable and ultimately monotone.
- (A5) In the asymptotic framework, as $\lim_{n \rightarrow \infty} h_n = 0$ and $\lim_{n \rightarrow \infty} nh_n^{2r+1} = \infty$, i.e., as the number of sample n is increased h approaches zero at a rate slower than $1/n^{2r+1}$.
- (A6) Assumptions about K are introduced in the previous section.

As seen in Equation (3), when working with a kernel estimator of the r^{th} derivative function two choices must be made: the kernel function K and the smoothing parameter or bandwidth h . The choice of K is a problem of less importance, because K is not very sensitive to the shape of estimator, and different functions that produce good results can be used. In practice, the choice of an efficient method for the computation of h , for an observed data sample is a crucial problem, because of the effect of the bandwidth on the shape of the corresponding estimator. If the bandwidth is small, we will obtain an under smoothed estimator, with high variability. On the contrary, if the value of h is big, the resulting estimator will be over smooth and farther from the function that we are trying to estimate.

An example is drawn in Figure 2 where we show in left four different kernel (Gaussian, biweight, triweight and tricube) estimators of the first derivative of a bimodal (separated) Gaussian density (Equation 5), and a given value of $h = 0.6$. On the right, using the Gaussian kernel and four different values for the bandwidth.

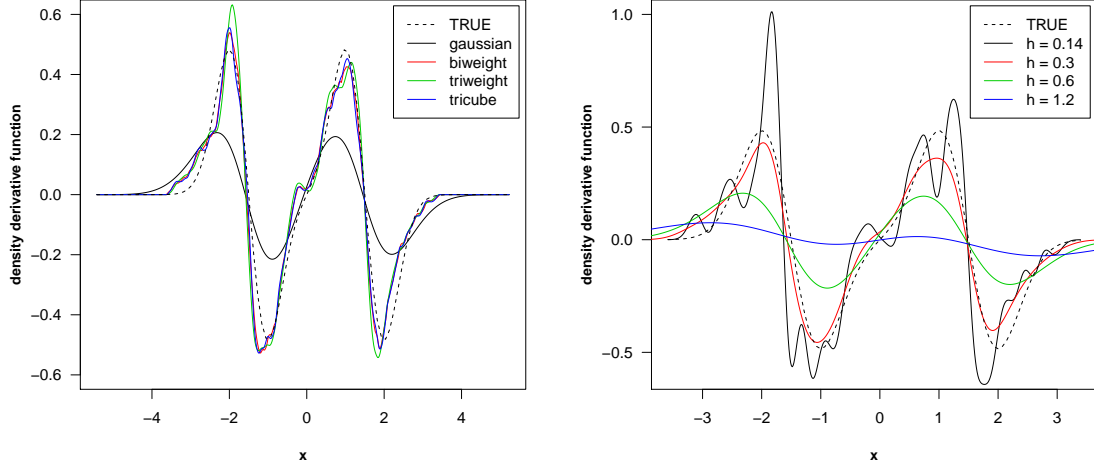


Figure 2: (Left) Different kernels for estimation, with $h = 0.6$. (Right) Effect of the bandwidth on the kernel estimator.

We have implemented in **R** the function `dkde` corresponds to the derivative of kernel density estimator (Equation 3). Eight possibilities are allowed for the kernel functions that are summarized in Table 1. We enumerate the arguments and results of this function in Table 2.

Arguments	Description
<code>x</code>	The data sample.
<code>y</code>	The points of the grid at which the density derivative is to be estimated. The default are $4h$ outside of $\text{range}(x)$.
<code>deriv.order</code>	Derivative order (scalar).
<code>h</code>	The smoothing bandwidth to be used. The default, "ucv" unbiased cross-validation.
<code>kernel</code>	The kernel function (see Table 1), by default "gaussian".
Results	Description
<code>eval.points</code>	The coordinates of the points where the density derivative is estimated.
<code>est.fx</code>	The estimated density derivative values (Equation 3).

Table 2: Summary of arguments and results of `dkde`.

Working with the dataset 'bimodal' correspond to data sample of 200 random numbers of a bimodality (separated) of a two-component Gaussian mixture density (Equation 4), with the following parameters: $-\mu_1 = \mu_2 = 3/2$ and $\sigma_1 = \sigma_2 = 1/2$. The `dkde` function enables to compute the r^{th} derivative of kernel density estimator over a grid of points, with a bandwidth selected by the user, but it also allows to estimate directly this parameter by the unbiased cross-validation method `h.ucv` (see following Section). We have chosen this method as the automatic one because it is the fastest in computation time terms. Now we estimate the first three derivatives of $f(x)$, can be written as:

$$f(x) = 0.5\phi(\mu_1, \sigma_1) + 0.5\phi(\mu_2, \sigma_2) \quad (4)$$

$$f^{(1)}(x) = 0.5(-4x - 6)\phi(\mu_1, \sigma_1) + 0.5(-4x + 6)\phi(\mu_2, \sigma_2) \quad (5)$$

$$f^{(2)}(x) = 0.5 \left((-4x - 6)^2 - 4 \right) \phi(\mu_1, \sigma_1) + 0.5 \left((-4x + 6)^2 - 4 \right) \phi(\mu_2, \sigma_2) \quad (6)$$

$$f^{(3)}(x) = 0.5(-4x - 6) \left((-4x - 6)^2 - 12 \right) \phi(\mu_1, \sigma_1) + 0.5(-4x + 6) \left((-4x + 6)^2 - 12 \right) \phi(\mu_2, \sigma_2) \quad (7)$$

where ϕ is a standard normal density.

```
> hatf <- dkde(bimodal, deriv.order = 0)
```

Data: bimodal (200 obs.); Kernel: gaussian

Derivative order: 0; Bandwidth 'h' = 0.2098

eval.points	est.fx
Min. :-3.86436	Min. :0.0000032
1st Qu.:-1.98016	1st Qu.:0.0147846
Median :-0.09595	Median :0.0737948
Mean :-0.09595	Mean :0.1324227
3rd Qu.: 1.78826	3rd Qu.:0.2326044
Max. : 3.67246	Max. :0.4374314

```
> hatf1 <- dkde(bimodal, deriv.order = 1)
```

Data: bimodal (200 obs.); Kernel: gaussian

Derivative order: 1; Bandwidth 'h' = 0.259

eval.points	est.fx
Min. :-4.06125	Min. :-0.4870865
1st Qu.:-2.07860	1st Qu.:-0.1521016
Median :-0.09595	Median : 0.0009041
Mean :-0.09595	Mean : 0.0000000
3rd Qu.: 1.88670	3rd Qu.: 0.1731795
Max. : 3.86935	Max. : 0.5038096

```
> hatf2 <- dkde(bimodal, deriv.order = 2)
```

Data: bimodal (200 obs.); Kernel: gaussian

Derivative order: 2; Bandwidth 'h' = 0.3017

eval.points	est.fx
Min. :-4.23200	Min. :-1.6800486
1st Qu.:-2.16398	1st Qu.: 0.0012798
Median :-0.09595	Median : 0.1421495
Mean :-0.09595	Mean :-0.0000073
3rd Qu.: 1.97208	3rd Qu.: 0.3389096
Max. : 4.04010	Max. : 0.7457487

```
> hatf3 <- dkde(bimodal, deriv.order = 3)
```

Data: bimodal (200 obs.); Kernel: gaussian

Derivative order: 3; Bandwidth 'h' = 0.3367

eval.points	est.fx
Min. :-4.37205	Min. :-4.353602
1st Qu.:-2.23400	1st Qu.:-0.472761
Median :-0.09595	Median : 0.001312
Mean :-0.09595	Mean :-0.000008
3rd Qu.: 2.04210	3rd Qu.: 0.388689
Max. : 4.18016	Max. : 3.614749

By default, the function `dkde` selects a grid of 512 points in the data range and used the Gaussian kernel. The output is a list containing the estimated values in the points of the grid, this last sequence and the bandwidth h (by default, using unbiased cross-validation method). In Figure 3 we show the first three derivatives estimators of $f(x)$ obtained with the code:

```
> fx <- function(x) 0.5 * dnorm(x,-1.5,0.5) + 0.5 * dnorm(x,1.5,0.5)
> fx1 <- function(x) 0.5 * (-4*x-6) * dnorm(x,-1.5,0.5) + 0.5 * (-4*x+6) *
+ dnorm(x,1.5,0.5)
> fx2 <- function(x) 0.5 * ((-4*x-6)^2 - 4) * dnorm(x,-1.5,0.5) + 0.5 *
+ ((-4*x+6)^2 - 4) * dnorm(x,1.5,0.5)
> fx3 <- function(x) 0.5 * (-4*x-6) * ((-4*x-6)^2 - 12) * dnorm(x,-1.5,0.5) +
+ 0.5 * (-4*x+6) * ((-4*x+6)^2 - 12) * dnorm(x,1.5,0.5)
> plot(hatf ,fx = fx)
> plot(hatf1,fx = fx1)
> plot(hatf2,fx = fx2)
> plot(hatf3,fx = fx3)
```

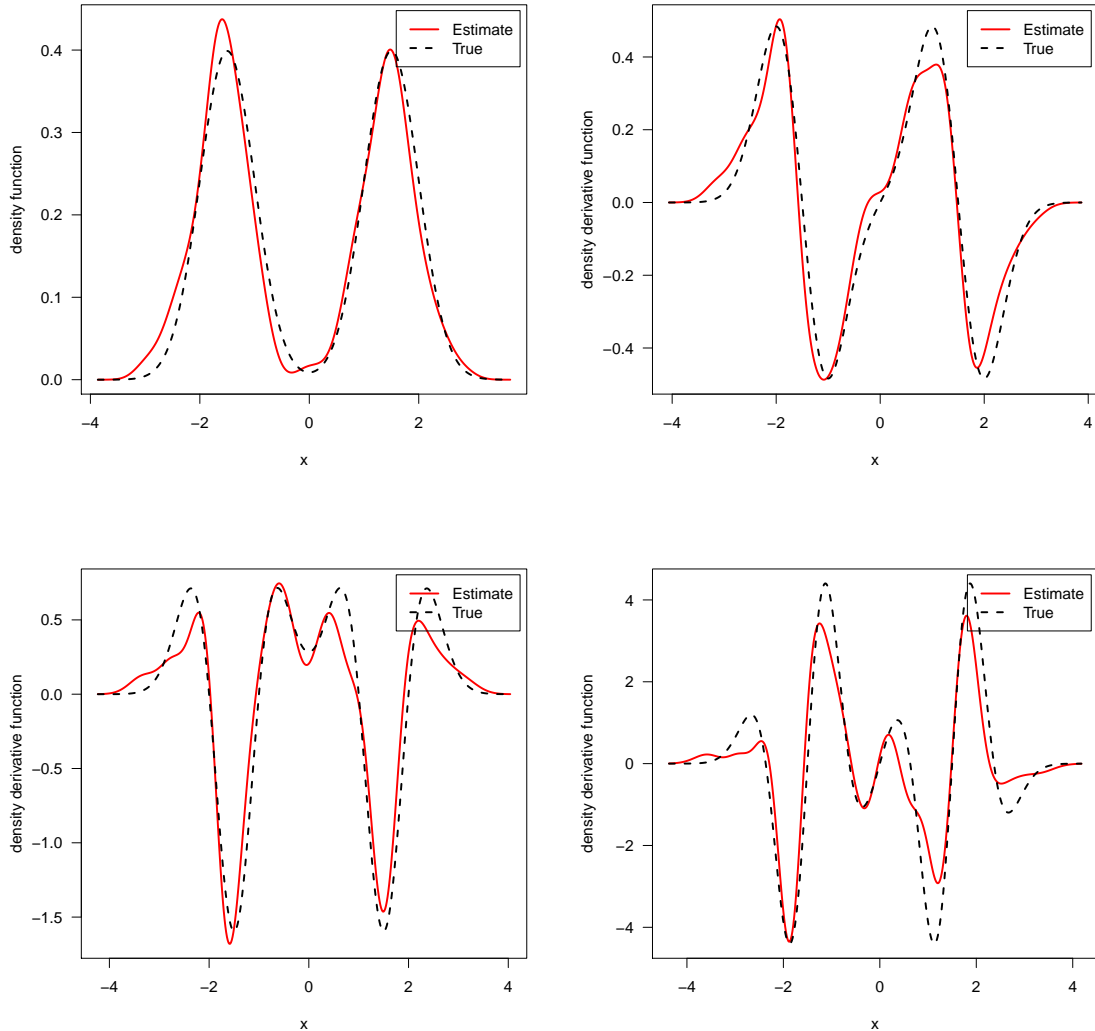


Figure 3: Kernel density derivative estimates obtained with the function `dkde`. (top left) density estimate $\hat{f}_h(x)$. (top right) first derivative $\hat{f}_h^{(1)}(x)$. (bottom left) second derivative $\hat{f}_h^{(2)}(x)$. (bottom right) third derivative $\hat{f}_h^{(3)}(x)$.

4 Bandwidth selections

Despite the great number of bandwidth selection techniques in kernel density estimator or regression estimation, as for example Rudemo [1982], Bowman [1984], Scott and George [1987], Sheather and Jones [1991], Chiu [1991a,b, 1992], Feluch and Koronacki [1992], Stute [1992], Jones et al [1996], Sheather [2004], Duong and Hazelton [2003, 2005], Heidenreich et al [2013], to the best of our knowledge, only few paper have been studied in the context of estimating the r^{th} derivative of a density $f(x)$, see Peter and Marron [1987], Wolfgang et al [1990], Jones and Kappenman [1991], Stoker [1993]. In this section we summarize the techniques of cross-validation methods for bandwidth choice in the kernel estimation of the derivatives of a probability density. The practicality of this methods is demonstrated by an example.

4.1 Optimal bandwidth

We Consider the following AMISE version of the r^{th} derivative of a probability density $f(x)$ [Scott, 1992, p. 131]:

$$\text{AMISE}(h, r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{1}{4}h^4\mu_2^2(K)R(f^{(r+2)}) \quad (8)$$

The optimal bandwidth minimizing (8) is:

$$h^* = \left[\frac{(2r+1)R(K^{(r)})}{\mu_2^2(K)R(f^{(r+2)})} \right]^{1/(2r+5)} n^{-1/(2r+5)} \quad (9)$$

whereof:

$$\text{AMISE}(h, r) = \frac{2r+5}{4}R(K^{(r)})^{\frac{4}{(2r+5)}} \left[\frac{\mu_2^2(K)R(f^{(r+2)})}{2r+1} \right]^{\frac{2r+1}{2r+5}} n^{-\frac{4}{2r+5}} \quad (10)$$

which is the smallest possible AMISE for estimation of $\hat{f}_h^{(r)}$. The function `h.amise` provides the optimal bandwidth under AMISE. The same possibilities for the kernel function as in the function `dkde` appear here. We enumerate the arguments and results of this function in Table 3.

Arguments	Description
<code>x</code>	The data sample.
<code>deriv.order</code>	Derivative order (scalar).
<code>lower, upper</code>	Range over which to minimize. The default is almost always satisfactory, <code>hos</code> (Over-smoothing) is calculated internally from an kernel.
<code>tol</code>	The convergence tolerance for optimize.
<code>kernel</code>	The kernel function (see Table 1), by default "gaussian".
Results	Description
<code>h</code>	Value of bandwidth (Equation 9).
<code>amise</code>	The AMISE value (Equation 10).

Table 3: Summary of arguments and results of `h.amise`.

The following example computes this bandwidth for a first three derivatives estimators of (4).

```
> h.amise(bimodal, deriv.order = 0)
```

Call: Asymptotic Mean Integrated Squared Error

Derivative order = 0

Data: bimodal (200 obs.); Kernel: gaussian

AMISE = 0.002602521; Bandwidth 'h' = 1.284843

```
> h.amise(bimodal, deriv.order = 1)
```

Call: Aymptotic Mean Integrated Squared Error

```
Derivative order = 1
Data: bimodal (200 obs.);      Kernel: gaussian
AMISE = 0.0009282042;          Bandwidth 'h' = 1.774593

> h.amise(bimodal, deriv.order = 2)
```

Call: Aymptotic Mean Integrated Squared Error

```
Derivative order = 2
Data: bimodal (200 obs.);      Kernel: gaussian
AMISE = 0.0003062873;          Bandwidth 'h' = 2.245869

> h.amise(bimodal, deriv.order = 3)
```

Call: Aymptotic Mean Integrated Squared Error

```
Derivative order = 3
Data: bimodal (200 obs.);      Kernel: gaussian
AMISE = 8.793292e-05;          Bandwidth 'h' = 2.690288
```

4.2 Maximum likelihood cross-validation

This method was proposed by [Habbema, Hermans and Van den Broek \[1974\]](#) and [Duin \[1976\]](#). They proposed to choose h so that the pseudo-likelihood $\prod_{i=1}^n \hat{f}_h(X_i)$ is maximized. However this has a trivial maximum at $h = 0$, so the cross-validation principle is invoked by replacing $\hat{f}_h(x)$ by the leave-one-out $\hat{f}_{h,i}(x)$, where:

$$\hat{f}_{h,i}(X_i) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{X_j - X_i}{h}\right)$$

Define that h as good which approaches the finite maximum of

$$h_{mlcv} = \operatorname{argmax}_{h>0} \text{MLCV}(h) \tag{11}$$

$$\text{MLCV}(h) = \left(n^{-1} \sum_{i=1}^n \log \left[\sum_{j \neq i} K\left(\frac{X_j - X_i}{h}\right) \right] - \log[(n-1)h] \right) \tag{12}$$

The function `h.mlcv` computed the maximum likelihood cross-validation for bandwidth selection. We enumerate the arguments and results of this function in [Table 4](#).

Arguments	Description
<code>x</code>	The data sample.
<code>lower, upper</code>	Range over which to minimize. The default is almost always satisfactory.
<code>tol</code>	The convergence tolerance for optimize.
<code>kernel</code>	The kernel function (see Table 1), by default <code>"gaussian"</code> .
Results	Description
<code>h</code>	Value of bandwidth (Equation 11).
<code>mlcv</code>	The MLCV value (Equation 12).

Table 4: Summary of arguments and results of `h.mlcv`.

The following example computes this bandwidth of bimodal Gaussian density ([Equation 4](#)), by different kernels.


```

> kernels <- eval(formals(h.mlcv.default)$kernel)
> hmlcv <- numeric()
> for(i in 1:length(kernels))
+       hmlcv[i] <- h.mlcv(bimodal, kernel = kernels[i])$h

> data.frame(kernels,hmlcv)

      kernels      hmlcv
1   gaussian 0.2302871
2 epanechnikov 0.4480106
3    uniform 0.3417343
4  triangular 0.4897095
5   triweight 0.6383113
6    tricube 0.5445540
7   biweight 0.5620119
8     cosine 0.4554530

```

The plot of the maximal likelihood cross validation function MLCV is shown in Figure 4 for Gaussian kernel in the left, and Epanechnikov kernel in the right, obtained with the code:

```

> plot(h.mlcv(bimodal, kernel = kernels[1]), seq.bws = seq(0.1,1,length=50))
> plot(h.mlcv(bimodal, kernel = kernels[2]), seq.bws = seq(0.1,1,length=50))

```

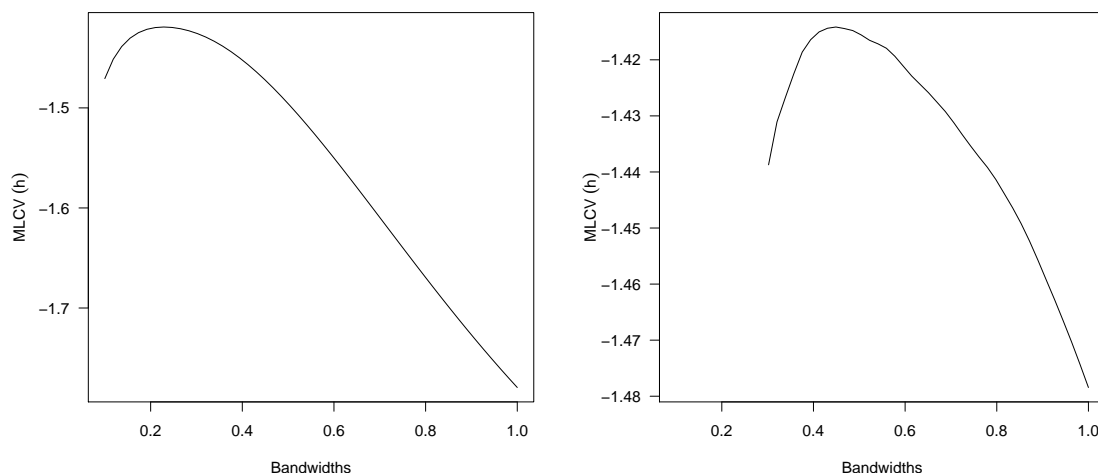


Figure 4: MLCV function obtained by `h.mlcv`, using Gaussian kernel (Left) and Epanechnikov kernel (Right).

4.3 Unbiased cross-validation

Rudemo [1982] and Bowman [1984] proposed a so-called unbiased (least-squares) cross-validation (UCV) in kernel density estimator, is probably the most popular and best studied one. An adaptation of unbiased cross-validation is proposed by Wolfgang et al [1990] for bandwidth choice in the r^{th} derivative of kernel density estimator. The essential idea of this methods, it aims to estimate h the minimizer of $ISE(h)$. The minimization criterion is defined by:

$$h_{ucv} = \operatorname{argmin}_{h>0} UCV(h, r) \quad (13)$$

$$\text{UCV}(h, r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(K^{(r)} * K^{(r)} - 2K^{(2r)} \right) \left(\frac{X_j - X_i}{h} \right) \quad (14)$$

In general, cross-validation functions in non-parametric bandwidth selection present several local minima. These minima are more likely to appear at too small values of the bandwidth [Peter and Marron, 1991]. The function `h.ucv` computes the unbiased cross-validation for bandwidth selection. We enumerate the arguments and results of this function in Table 5.

Arguments	Description
<code>x</code>	The data sample.
<code>deriv.order</code>	Derivative order (scalar).
<code>lower, upper</code>	Range over which to minimize. The default is almost always satisfactory, <code>hos</code> (Over-smoothing) is calculated internally from an kernel.
<code>tol</code>	The convergence tolerance for optimize.
<code>kernel</code>	The kernel function (see Table 1), by default "gaussian".
Results	Description
<code>h</code>	Value of bandwidth (Equation 13).
<code>min.ucv</code>	The minimal UCV value (Equation 14).

Table 5: Summary of arguments and results of `h.ucv`.

The following example computes the bandwidth h by this method for a first three derivatives estimators of (4).

```
> h.ucv(bimodal, deriv.order = 0)

Call:
  Unbiased Cross-Validation

Derivative order = 0
Data: bimodal (200 obs.);      Kernel: gaussian
Min UCV = -0.290735;          Bandwidth 'h' = 0.2098

> h.ucv(bimodal, deriv.order = 1)

Call:
  Unbiased Cross-Validation

Derivative order = 1
Data: bimodal (200 obs.);      Kernel: gaussian
Min UCV = -0.6523005;          Bandwidth 'h' = 0.2590212

> h.ucv(bimodal, deriv.order = 2)

Call:
  Unbiased Cross-Validation

Derivative order = 2
Data: bimodal (200 obs.);      Kernel: gaussian
Min UCV = -4.374522;          Bandwidth 'h' = 0.3017092

> h.ucv(bimodal, deriv.order = 3)

Call:
  Unbiased Cross-Validation

Derivative order = 3
Data: bimodal (200 obs.);      Kernel: gaussian
Min UCV = -46.74623;          Bandwidth 'h' = 0.3367229

The plot of UCV function obtained with the code (Figure 5):
> for (i in 0:3) plot(h.ucv(bimodal, deriv.order = i))
```

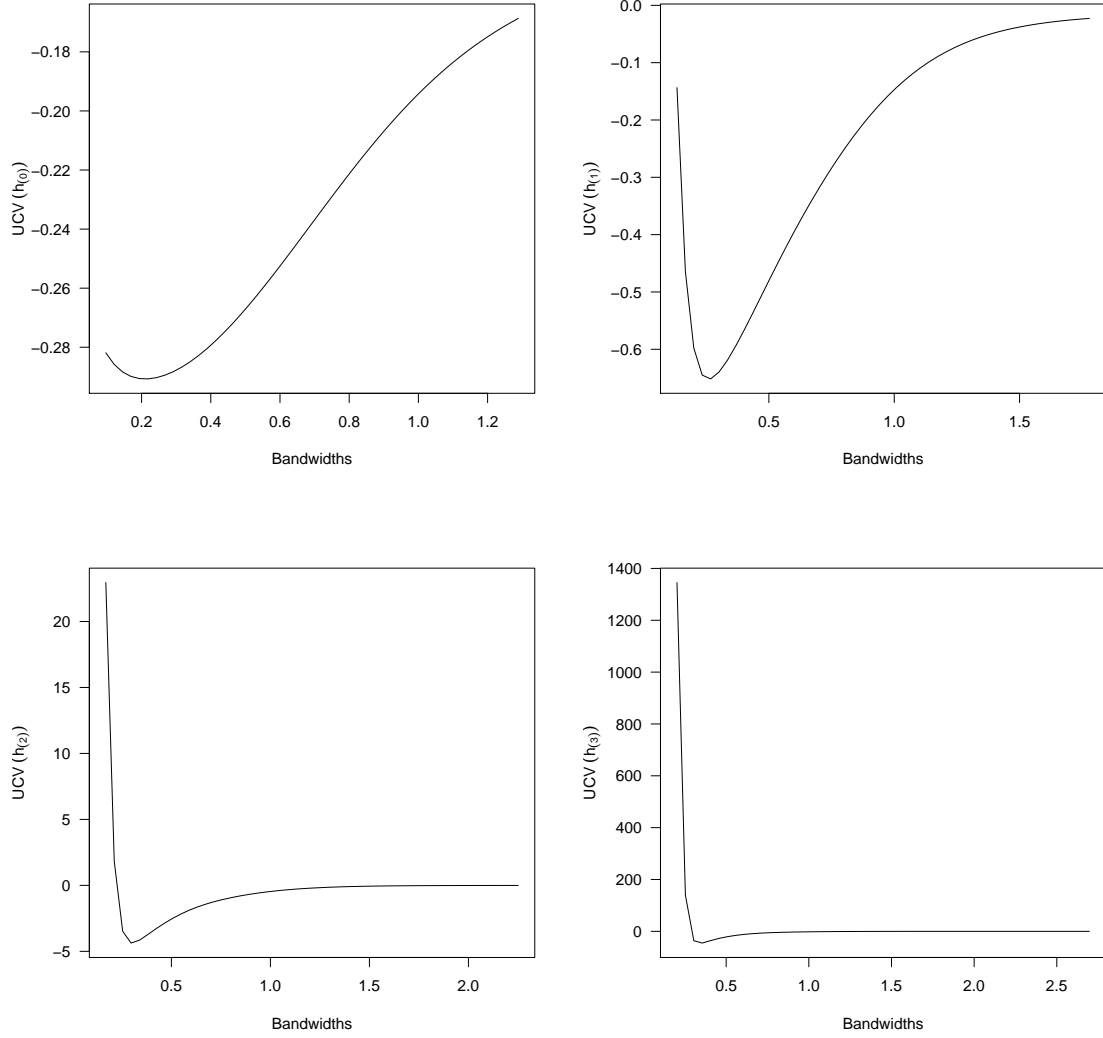


Figure 5: UCV function obtained by `h.ucv`. (top left) `deriv.order = 0`. (top right) `deriv.order = 1`. (bottom left) `deriv.order = 2`. (bottom right) `deriv.order = 3`.

4.4 Biased cross-validation

Biased cross-validation was proposed by [Scott and George \[1987\]](#), which has as its immediate target the AMISE (8). They proposed to estimate $R(f^{(r+2)})$ by!

$$\hat{R}(f^{(r+2)}) = R(\hat{f}_h^{(r+2)}) - \frac{R(K^{(r+2)})}{nh^{2r+5}}$$

There are two versions of BCV, depending on the estimator of $R(f^{(r+2)})$. We can use [\[Scott and George, 1987\]](#)

$$\hat{R}(f^{(r+2)}) = \frac{(-1)^{r+2}}{n(n-1)h^{2r+5}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K^{(r+2)} * K^{(r+2)} \left(\frac{X_j - X_i}{h} \right)$$

or we could use [\[Jones and Kappenman, 1991\]](#)

$$\hat{R}(f^{(r+2)}) = \frac{(-1)^{r+2}}{n(n-1)h^{2r+5}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K^{(2r+4)} \left(\frac{X_j - X_i}{h} \right)$$

From this we obtain respectively an adaptation of biased cross-validation for bandwidth choice in the r^{th} derivative of kernel density estimator, is given by:

$$\text{BCV}_1(h, r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{\mu_2^2(K)}{4} \frac{(-1)^{r+2}}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K^{(r+2)} * K^{(r+2)} \left(\frac{X_j - X_i}{h} \right) \quad (15)$$

$$\text{BCV}_2(h, r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{\mu_2^2(K)}{4} \frac{(-1)^{r+2}}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K^{(2r+4)} \left(\frac{X_j - X_i}{h} \right) \quad (16)$$

The BCV selectors h_{bcv_1} and h_{bcv_2} are the minimisers of the appropriate BCV function. The function `h.bcv` computes the biased cross-validation for bandwidth selection. We enumerate the arguments and results of this function in Table 6.

Arguments	Description
<code>x</code>	The data sample.
<code>whichbcv</code>	Method selected, 1 = BCV1 or 2 = BCV2, by default BCV1.
<code>deriv.order</code>	Derivative order (scalar).
<code>lower, upper</code>	Range over which to minimize. The default is almost always satisfactory, <code>hos</code> (Over-smoothing) is calculated internally from an kernel.
<code>tol</code>	The convergence tolerance for optimize.
<code>kernel</code>	The kernel function (see Table 1), by default "gaussian".
Results	Description
<code>h</code>	Value of bandwidth.
<code>min.bcv</code>	The minimal BCV value (Equation 15 or 16).

Table 6: Summary of arguments and results of `h.bcv`.

The following example computes the bandwidth parameter by this method for kernel estimator of Equation (4) and its first derivative estimators.

```
> h.bcv(bimodal, whichbcv = 1, deriv.order = 0)
```

```
Call:      Biased Cross-Validation 1
```

```
Derivative order = 0
```

```
Data: bimodal (200 obs.);      Kernel: gaussian
Min BCV = 0.008406978;      Bandwidth 'h' = 0.2239608
```

```
> h.bcv(bimodal, whichbcv = 2, deriv.order = 0)
```

```
Call:      Biased Cross-Validation 2
```

```
Derivative order = 0
```

```
Data: bimodal (200 obs.);      Kernel: gaussian
Min BCV = 0.007133677;      Bandwidth 'h' = 0.1748194
```

```
> h.bcv(bimodal, whichbcv = 1, deriv.order = 1, lower=0.1, upper=0.8)
```

```
Call:      Biased Cross-Validation 1
```

```
Derivative order = 1
```

```
Data: bimodal (200 obs.);      Kernel: gaussian
Min BCV = 0.06626954;      Bandwidth 'h' = 0.3366177
```

```
> h.bcv(bimodal, whichbcv = 2, deriv.order = 1, lower=0.1, upper=0.8)
```

Call: Biased Cross-Validation 2

```
Derivative order = 1
Data: bimodal (200 obs.);      Kernel: gaussian
Min BCV = -0.3470459;          Bandwidth 'h' = 0.1356763
```

The plot of BCV function obtained with the code `h.bcv` (Figure 6):

```
> ## deriv.order = 0
> plot(h.bcv(bimodal, whichbcv = 2, deriv.order = 0))
> lines(h.bcv(bimodal, whichbcv = 1, deriv.order = 0), col="red")
> legend("topright", c("BCV1", "BCV2"), lty=1, col=c("red", "black"),
+       inset = .015)
> ## deriv.order = 1
> plot(h.bcv(bimodal, whichbcv = 2, deriv.order = 1), seq.bws =
+       seq(0.1, 0.8, length=50))
> lines(h.bcv(bimodal, whichbcv = 1, deriv.order = 1), col="red")
> legend("topright", c("BCV1", "BCV2"), lty=1, col=c("red", "black"),
+       inset = .015)
```

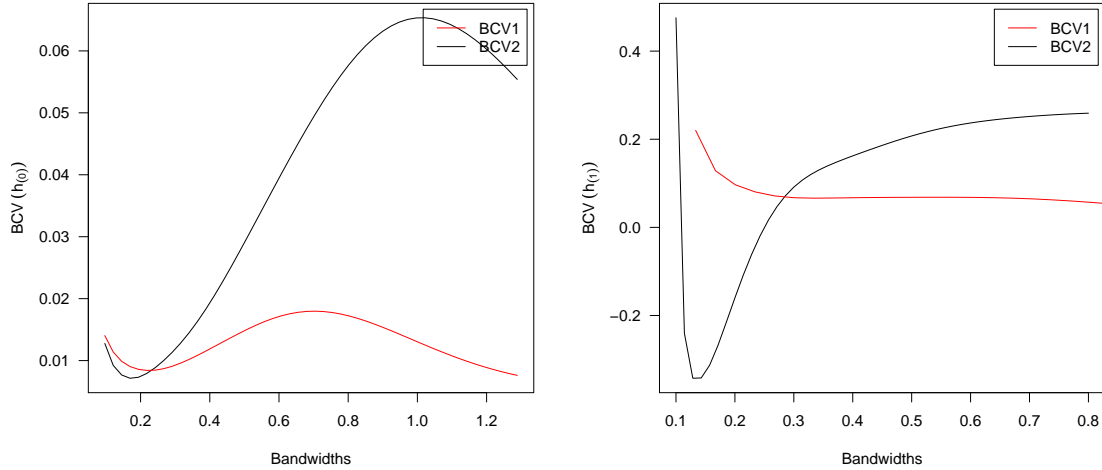


Figure 6: BCV function obtained by `h.bcv`. (Left) BCV_1 vs BCV_2 (`deriv.order = 0`). (Right) BCV_1 vs BCV_2 (`deriv.order = 1`).

4.5 Complete cross-validation

Jones and Kappenman [1991] proposed a so-called complete cross-validation (CCV) in kernel density estimator. This method can be extended to the estimation of derivative of the density, basing our estimate of integrated squared density derivative [Peter and Marron, 1987] we get the following. Thus, h_{ccv} , say, is the h that minimises:

$$CCV(h, r) = R\left(\hat{f}_h^{(r)}\right) - \bar{\theta}_r(h) + \frac{1}{2}\mu_2(K)h^2\bar{\theta}_{r+1}(h) + \frac{1}{24}(6\mu_2^2(K) - \delta(K))h^4\bar{\theta}_{r+2}(h) \quad (17)$$

where,

$$R\left(\hat{f}_h^{(r)}\right) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K^{(r)} * K^{(r)} \left(\frac{X_j - X_i}{h} \right)$$

and

$$\bar{\theta}_r(h) = \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K^{(2r)}\left(\frac{X_j - X_i}{h}\right)$$

with : $\delta(K) = \int_{\mathbb{R}} x^4 K(x) dx$.

The function `h.ccv` computes the complete cross-validation for bandwidth selection. We enumerate the arguments and results of this function in Table 7.

Arguments	Description
<code>x</code>	The data sample.
<code>deriv.order</code>	Derivative order (scalar).
<code>lower, upper</code>	Range over which to minimize. The default is almost always satisfactory, <code>hos</code> (Over-smoothing) is calculated internally from an kernel.
<code>tol</code>	The convergence tolerance for optimize.
<code>kernel</code>	The kernel function (see Table 1), by default "gaussian".
Results	Description
<code>h</code>	Value of bandwidth.
<code>min.ccv</code>	The minimal CCV value (Equation 17).

Table 7: Summary of arguments and results of `h.ccv`.

The following example computes the bandwidth h by this method for a first three derivatives estimators of (4). This time we set Over-smoothing in `upper = 0.5`.

```
> h.ccv(bimodal, deriv.order = 0, upper = 0.5)
```

Call: Complete Cross-Validation

```
Derivative order = 0
Data: bimodal (200 obs.);      Kernel: gaussian
Min CCV = 0.00764383;          Bandwidth 'h' = 0.1790795
```

```
> h.ccv(bimodal, deriv.order = 1, upper = 0.5)
```

Call: Complete Cross-Validation

```
Derivative order = 1
Data: bimodal (200 obs.);      Kernel: gaussian
Min CCV = -0.197061;          Bandwidth 'h' = 0.1427856
```

```
> h.ccv(bimodal, deriv.order = 2, upper = 0.5)
```

Call: Complete Cross-Validation

```
Derivative order = 2
Data: bimodal (200 obs.);      Kernel: gaussian
Min CCV = -127.6184;          Bandwidth 'h' = 0.1259258
```

```
> h.ccv(bimodal, deriv.order = 3, upper = 0.5)
```

Call: Complete Cross-Validation

```
Derivative order = 3
Data: bimodal (200 obs.);      Kernel: gaussian
Min CCV = -35547.32;          Bandwidth 'h' = 0.1426445
```

The plot of CCV function obtained with the code (Figure 7):

```
> for (i in 0:3)
+   plot(h.ccv(bimodal, deriv.order = i), seq.bws=seq(0.1,0.5,length=50))
```

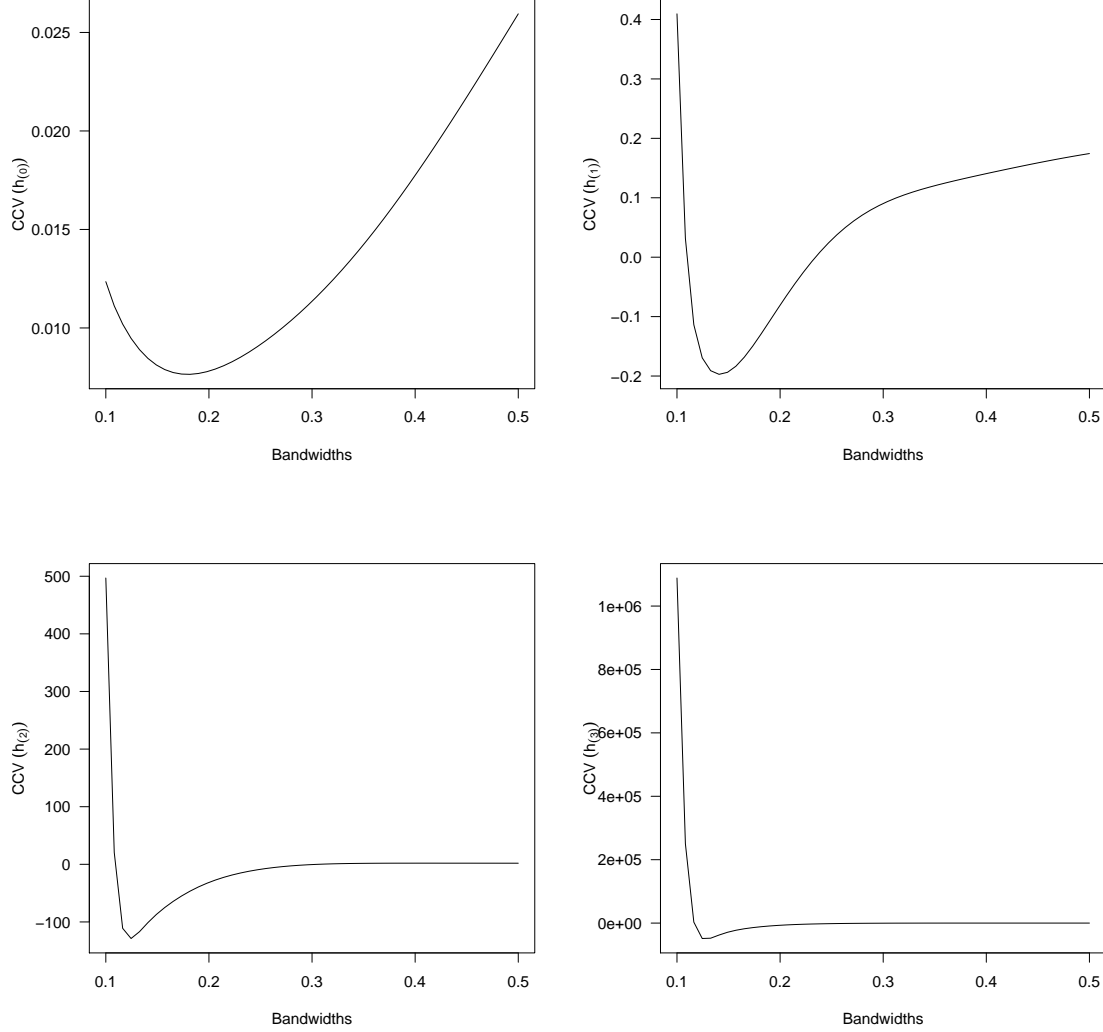


Figure 7: CCV function obtained by `h.ccv`. (top left) `deriv.order = 0`. (top right) `deriv.order = 1`. (bottom left) `deriv.order = 2`. (bottom right) `deriv.order = 3`.

4.6 Modified cross-validation

Stute [1992] proposed a so-called modified cross-validation (MCV) in kernel density estimator. This method can be extended to the estimation of derivative of a probability density, the essential idea based on approximated the problematic term by the aid of the Hajek projection. The minimization criterion is defined by:

$$\text{MCV}(h, r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \varphi^{(r)}\left(\frac{X_j - X_i}{h}\right) \quad (18)$$

where

$$\varphi^{(r)}(c) = \left(K^{(r)} * K^{(r)} - K^{(2r)} - \frac{\mu_2(K)}{2} K^{(2r+2)} \right)(c)$$

The function `h.mcv` computes the modified cross-validation for bandwidth selection. We enumerate the arguments and results of this function in Table 8.

Arguments	Description
<code>x</code>	The data sample.
<code>deriv.order</code>	Derivative order (scalar).
<code>lower,upper</code>	Range over which to minimize. The default is almost always satisfactory, <code>hos</code> (Over-smoothing) is calculated internally from an kernel.
<code>tol</code>	The convergence tolerance for optimize.
<code>kernel</code>	The kernel function (see Table 1), by default "gaussian".
Results	Description
<code>h</code>	Value of bandwidth.
<code>min.mcv</code>	The minimal MCV value (Equation 18).

Table 8: Summary of arguments and results of `h.mcv`.

The following example computes the bandwidth h by this method for a first three derivatives estimators of (4). We set Over-smoothing in `upper = 0.5`.

```
> h.mcv(bimodal, deriv.order = 0, upper = 0.5)
```

Call: Modified Cross-Validation

Derivative order = 0

Data: bimodal (200 obs.); Kernel: gaussian

Min MCV = 0.007370888; Bandwidth 'h' = 0.2402603

```
> h.mcv(bimodal, deriv.order = 1, upper = 0.5)
```

Call: Modified Cross-Validation

Derivative order = 1

Data: bimodal (200 obs.); Kernel: gaussian

Min MCV = 0.04308772; Bandwidth 'h' = 0.196102

```
> h.mcv(bimodal, deriv.order = 2, upper = 0.5)
```

Call: Modified Cross-Validation

Derivative order = 2

Data: bimodal (200 obs.); Kernel: gaussian

Min MCV = -19.17733; Bandwidth 'h' = 0.1304198

```
> h.mcv(bimodal, deriv.order = 3, upper = 0.5)
```

Call: Modified Cross-Validation

Derivative order = 3

Data: bimodal (200 obs.); Kernel: gaussian

Min MCV = -8391.447; Bandwidth 'h' = 0.1426445

The plot of MCV function obtained with the code (Figure 8):

```
> for (i in 0:3)
+   plot(h.mcv(bimodal, deriv.order = i), seq.bws=seq(0.1,0.5,length=50))
```

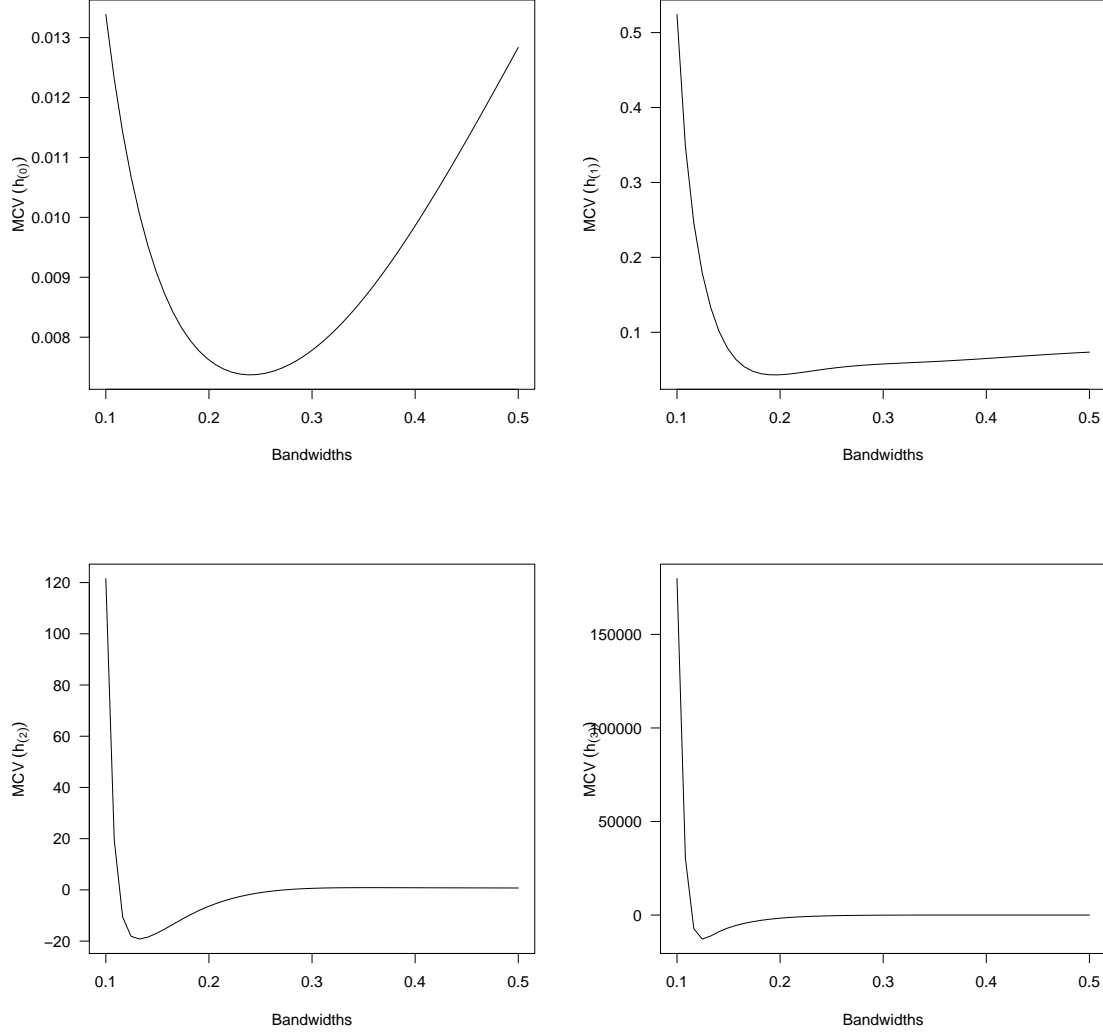



Figure 8: MCV function obtained by `h.mcv`. (top left) `deriv.order` = 0. (top right) `deriv.order` = 1. (bottom left) `deriv.order` = 2. (bottom right) `deriv.order` = 3.

4.7 Trimmed cross-validation

Feluch and Koronacki [1992] proposed a so-called trimmed cross-validation (TCV) in kernel density estimator, a simple modification of the unbiased (least-squares) cross-validation criterion (14). We consider the following "trimmed" version of "unbiased", to be minimized with respect to h :

$$\text{TCV}(h, r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \varphi^{(r)}\left(\frac{X_j - X_i}{h}\right) \quad (19)$$

where

$$\varphi^{(r)}(c) = \left[K^{(r)} * K^{(r)} - 2K^{(2r)}1\left(|c| > \frac{c_n}{h}\right) \right](c)$$

$1(\cdot)$ denotes the indicator function and c_n is a sequence of positive constants, as $\lim_{n \rightarrow \infty} c_n/h \rightarrow 0$, here we take $c_n = 1/n$, for assure the convergence. The function `h.tcv` computes the trimmed cross-validation for bandwidth selection. We enumerate the arguments and results of this function in Table 9.

The following example computes the bandwidth h by this method for a first three derivatives estimators of (4).

Arguments	Description
<code>x</code>	The data sample.
<code>deriv.order</code>	Derivative order (scalar).
<code>lower, upper</code>	Range over which to minimize. The default is almost always satisfactory, <code>hos</code> (Over-smoothing) is calculated internally from an kernel.
<code>tol</code>	The convergence tolerance for optimize.
<code>kernel</code>	The kernel function (see Table 1), by default "gaussian".
Results	Description
<code>h</code>	Value of bandwidth.
<code>min.tcv</code>	The minimal TCV value (Equation 19).

Table 9: Summary of arguments and results of `h.tcv`.

```
> h.tcv(bimodal, deriv.order = 0)

Call:
      Trimmed Cross-Validation

Derivative order = 0
Data: bimodal (200 obs.);      Kernel: gaussian
Min TCV = -0.2806542;      Bandwidth 'h' = 0.2570739

> h.tcv(bimodal, deriv.order = 1)

Call:
      Trimmed Cross-Validation

Derivative order = 1
Data: bimodal (200 obs.);      Kernel: gaussian
Min TCV = -0.3956968;      Bandwidth 'h' = 0.5095463

> h.tcv(bimodal, deriv.order = 2)

Call:
      Trimmed Cross-Validation

Derivative order = 2
Data: bimodal (200 obs.);      Kernel: gaussian
Min TCV = -1.153544;      Bandwidth 'h' = 0.65599

> h.tcv(bimodal, deriv.order = 3)

Call:
      Trimmed Cross-Validation

Derivative order = 3
Data: bimodal (200 obs.);      Kernel: gaussian
Min TCV = -3.961823;      Bandwidth 'h' = 0.7453813

The plot of TCV function obtained with the code (Figure 9):

> for (i in 0:3) plot(h.tcv(bimodal, deriv.order = i))
```

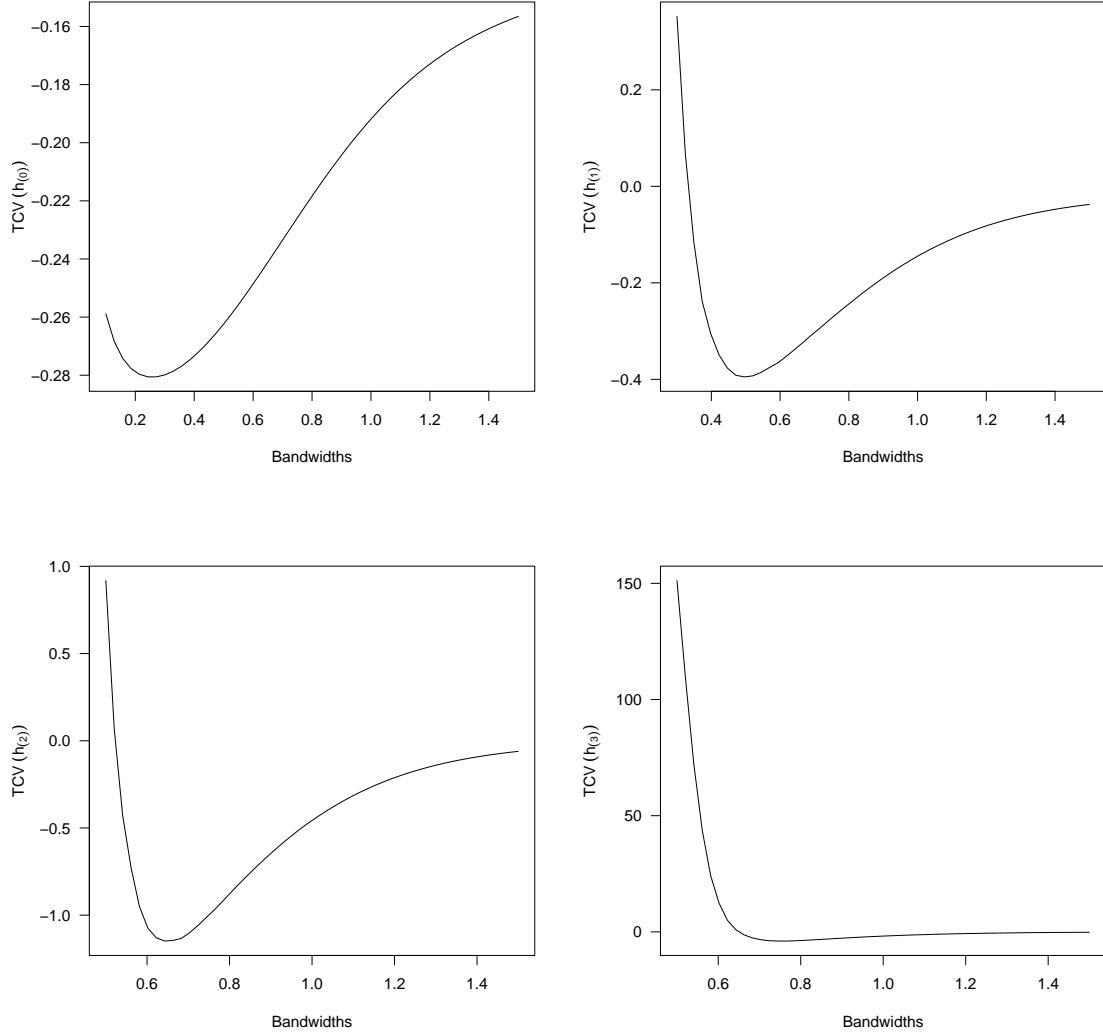


Figure 9: TCV function obtained by `h.tcv`. (top left) `deriv.order` = 0. (top right) `deriv.order` = 1. (bottom left) `deriv.order` = 2. (bottom right) `deriv.order` = 3.

5 Summary

We have implemented in **R** the estimators of the defined functions and the bandwidth selection procedures of the above sections. The package **kedd** contains seven functions, in Table 10 we can find a summary of the contents of the package.

The current feature set of the package can be split in four main categories: compute the convolutions and derivatives of a kernel function, compute the kernel estimators for a density of probability and its derivatives, computing the bandwidth selectors with different methods, displaying the kernel estimators and selection functions of the bandwidth. Moreover, the package follows the general **R** philosophy of working with model objects. This means that instead of merely returning, say, a kernel estimator of r^{th} derivative of a density, many functions will return an object containing, it's functions are S3 classes (`S3method`). The object can then be manipulated at one's will using various extraction, summary or plotting functions. Whenever possible, we develop a graphical user interface of the various functions of a coherent whole, to facilitate the use of this package.

Function	Description
<code>dkde</code>	Derivatives of kernel density estimator, as defined in Equation 3.
<code>h.amise</code>	AMISE for optimal bandwidth selectors (Equation 10).
<code>h.mlcv</code>	Maximum-likelihood cross-validation bandwidth selection (Equation 12).
<code>h.ucv</code>	Unbiased cross-validation bandwidth selection (Equation 14).
<code>h.bcv</code>	Biased cross-validation bandwidth selection (Equations 15 and 16) .
<code>h.ccv</code>	Complete cross-validation bandwidth selection (Equation 17).
<code>h.mcv</code>	Modified cross-validation bandwidth selection (Equation 18).
<code>h.tcv</code>	Trimmed cross-validation bandwidth selection (Equation 19).

Table 10: Summary of contents of the package.

References

- Alekseev, V. G. (1972). Estimation of a probability density function and its derivatives. *Mathematical notes of the Academy of Sciences of the USSR*, **12**(5), 808–811.
- Alexandre, B. T. (2009). *Introduction to Nonparametric Estimation*. Springer-Verlag, New York.
- Bhattacharya, P. K. (1967). Estimation of a probability density function and Its derivatives. *Sankhya: The Indian Journal of Statistics, Series A*, **29**, 373–382.
- Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of kernel density estimates. *Biometrika*, **71**, 353–360.
- Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford.
- Bowman, A. W. and Azzalini, A. (2013). R package `sm`: nonparametric smoothing methods (version 2.2-5). <http://www.stats.gla.ac.uk/~adrian/sm>, http://azzalini.stat.unipd.it/Book_sm
- Chiu, S.T. (1991a). Some stabilized bandwidth selectors for nonparametric regression. *Ann. Stat.* **19**, 1528–1546.
- Chiu, S.T. (1991b). Bandwidth selection for kernel density estimation. *Ann. Stat.* **19**, 1883–1905.
- Chiu, S.T. (1992). An automatic bandwidth selector for kernel density estimation. *Biometrika*, **79**, 771–782.
- Duong, T. (2007). `ks`: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software*. **21**(7).
- Duong, T. and Hazelton, M.L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, **32**, 485–506.
- Duong, T. and Hazelton, M.L. (2003). Plug-in bandwidth selectors for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, **15**, 17–30.
- Duong, T. and Matt, W. (2013). `feature`: Feature significance for multivariate kernel density estimation. R package version 1.2.9. <http://CRAN.R-project.org/package=feature>
- Duin, R. P. W. (1976). On the choice of smoothing parameters of Parzen estimators of probability density functions. *IEEE Transactions on Computers*, **C-25**, 1175–1179.
- Feluch, W. and Koronacki, J. (1992). A note on modified cross-validation in density estimation. *Computational Statistics and Data Analysis*, **13**, 143–151.

- Guidoum, A. C. (2013). **kedd**: Kernel estimator and bandwidth selection for density and its derivatives. R package version 1.0.0. <http://CRAN.R-project.org/package=kedd>
- Habbema, J. D. F., Hermans, J., and Van den Broek, K. (1974). A stepwise discrimination analysis program using density estimation. *Compstat 1974: Proceedings in Computational Statistics*. Physica Verlag, Vienna.
- Heidenreich, N. B., Schindler, A. and Sperlich, S. (2013). Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *Advances in Statistical Analysis*.
- Jeffrey, S. S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag, New York.
- Jones, M. C. and Kappenman, R. F. (1991). On a class of kernel density estimate bandwidth selectors. *Scandinavian Journal of Statistics*, **19**, 337–349.
- Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, **91**, 401–407.
- Olver, F. W., Lozier, D. W., Boisvert, R. F. and Clark, C. W. (2010). *NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, USA.
- Peter, H. and Marron, J.S. (1987). Estimation of integrated squared density derivatives. *Statistics and Probability Letters*, **6**, 109–115.
- Peter, H. and Marron, J.S. (1991). Local minima in cross-validation functions. *Journal of the Royal Statistical Society, Series B*, **53**, 245–252.
- R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. <http://www.R-project.org/>
- Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65–78.
- Schuster, E. F. (1969). Estimation of a probability density function and its derivatives. *The Annals of Mathematical Statistics*, **40**(4), 1187–1195.
- Scott, D. W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization*. New York: Wiley.
- Scott, D.W. and George, R. T. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, **82**, 1131–1146.
- Sheather, S. J. (2004). Density estimation. *Statistical Science*, **19**, 588–597.
- Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683–690.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC. London.
- Singh, R. S. (1990). Applications of estimators of a density and its derivatives **39**(3), 357–363.
- Stoker, T. M. (1993). Smoothing bias in density derivative estimation. *Journal of the American Statistical Association*, **88**, 855–863.
- Stute, W. (1992). Modified cross validation in density estimation. *Journal of Statistical Planning and Inference*, **30**, 293–305.
- Tristen, H. and Jeffrey, S. R. (2008). Nonparametric Econometrics: The **np** Package. *Journal of Statistical Software*. **27** (5).

- Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.
- Wand, M.P. and Ripley, B. D. (2013). **KernSmooth**: Functions for kernel smoothing for Wand and Jones (1995). R package version 2.23-10. <http://CRAN.R-project.org/package=KernSmooth>
- Wolfgang, H. (1991). *Smoothing Techniques, With Implementation in S*. Springer-Verlag, New York.
- Wolfgang, H., Marron, J. S. and Wand, M. P. (1990). Bandwidth choice for density derivatives. *Journal of the Royal Statistical Society, Series B*, 223–232.
- Wolfgang, H., Marlene, M., Stefan, S. and Axel, W. (2004). *Nonparametric and Semiparametric Models*. Springer-Verlag, Berlin Heidelberg.
- Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.