Ake: An R Package for Discrete and Continuous Associated Kernel Estimations

by Wanbitching E. Wansouwé, Sobom M. Somé and Célestin C. Kokonendji

Abstract  Kernel estimation is an important technique in exploratory data analysis. Its utility relies on its ease of interpretation, especially based on graphical means. The Ake package is introduced for univariate density or probability mass function estimation and also for continuous and discrete regression functions using associated kernel estimators. These associated kernels have been proposed due to their specific features of variables of interest. The package focuses on associated kernel methods appropriate for continuous (bounded, positive) or discrete (count, categorical) data often found in applied settings. Furthermore, optimal bandwidths are selected by cross-validation for any associated kernel and by Bayesian methods for the binomial kernel. Other Bayesian methods for selecting bandwidths with other associated kernels will complete this package in its future versions; particularly, a Bayesian adaptive method for gamma kernel estimation of density functions is developed. Some practical and theoretical aspects of the normalizing constant in both density and probability mass functions estimations are given.

Introduction

Kernel smoothing methods are popular tools for revealing the structure of data that could be missed by parametric methods. For real datasets, we often encounter continuous (bounded, positive) or discrete (count, categorical) data types. The classical kernels methods assume that the underlying distribution is unbounded continuous, which is frequently not the case; see, for example, Duong (2007) for multivariate kernel density estimation and discriminant analysis. A solution is provided for categorical data sets by Hayfield and Racine (2008). In fact, they used kernels well adapted for these categorical sets (Aitchison and Aitken, 1976). Throughout the present paper, the unidimensional support $T$ of the variable of interest can be $\{0, 1, \ldots, N\}$, $[a, b]$ or $[0, \infty)$ for a given integer $N$ and reals $a < b$.

The recently developed Ake package, implements associated kernels that seamlessly deal with continuous (bounded, positive) and discrete (categorical, count) data types often found in applied settings; see, for example, Libengué (2013) and Kokonendji and Senga Kiessé (2011). These associated kernels are used to smooth probability density functions (p.d.f.), probability mass functions (p.m.f.) or regression functions. The coming versions of this package will contain, among others, p.d.f. estimation of heavy tailed data (e.g., Ziane et al., 2015) and the estimation of other functionals. The bandwidth selection remains crucial in associated kernel estimations of p.d.f., p.m.f. or regression functions. Some methods have been investigated for selecting bandwidth parameters but the commonly used is the least squared cross-validation. A Bayesian approach has been also recently introduced by Zougab et al. (2012) in the case of a binomial kernel. This method can be extended to various associated kernels with other functionals. Despite the great number of packages implemented for nonparametric estimation in continuous cases with unbounded kernels, to the best of our knowledge, the R packages to estimate p.m.f. with categorical or count variables, p.d.f. with bounded or positive datasets, and regression functions have been far less investigated.

The rest of the paper is organized as follows. In Section Non-classical associated kernels, we briefly describe the definition of associated kernels and then illustrate examples in both continuous and discrete cases which are discussed. Then, the associated kernel estimator for p.d.f. or p.m.f. is presented and illustrated with some R codes in Section Density or probability mass function estimations. In particular, three bandwidth selection methods are available: cross-validation for any (continuous or discrete) associated kernel, the Bayesian local method for the binomial kernel and also a new theoretical Bayesian adaptive method for the gamma kernel. Also, some practical and theoretical aspects of the normalizing constant in both p.d.f. and p.m.f. estimations are given. Section Bandwidth selection for kernel regression involving associated kernels investigates the case of regression functions with two bandwidth selection techniques: cross-validation and also the Bayesian global method for the binomial kernel. Section Summary and final remarks concludes.
Non-classical associated kernels

Recall that the support $T$ of the p.m.f., p.d.f. or regression function, to be estimated, is any set $\{0,1,\ldots,N\}$, $[a,b]$ or $[0,\infty)$ for a given integer $N$ and reals $a \prec h$. The associated kernel in both continuous and discrete cases is defined as follows.

**Definition 34.2.1.** (Kokonendji and Senga Kiessé, 2011; Libengué, 2013) Let $T (\subseteq \mathbb{R})$ be the support of the p.m.f., p.d.f. or regression function, to be estimated, $x \in T$ a target and $h$ a bandwidth. A parametrized p.m.f. (respectively p.d.f.) $K_{x,h}(\cdot)$ of support $S_{x,h}(\subseteq \mathbb{R})$ is called “associated kernel” if the following conditions are satisfied:

\[
x \in S_{x,h},
\]
\[
E \left( Z_{x,h} \right) = x + a(x, h),
\]
\[
V \left( Z_{x,h} \right) = b(x, h),
\]

where $Z_{x,h}$ denotes the random variable with p.m.f. (respectively p.d.f.) $K_{x,h}$ and both $a(x, h)$ and $b(x, h)$ tend to 0 as $h$ goes to 0.

**Remark 34.2.2.** This definition has the following interesting interpretations:

(i) The function $K_{x,h}(\cdot)$ is not necessary symmetric and is intrinsically linked to $x$ and $h$.

(ii) The support $S_{x,h}$ is not necessary symmetric around $x$; it can depend or not on $x$ and $h$.

(iii) The condition (1) can be viewed as $\cup_{x \in T} S_{x,h} \supseteq T$ and it implies that the associated kernel takes into account the support $T$ of the density $f$, to be estimated.

(iv) If $\cup_{x \in T} S_{x,h}$ does not contain $T$ then this is the well-known problem of boundary bias.

(v) Both conditions (2) and (3) indicate that the associated kernel is more and more concentrated around $x$ as $h$ goes to 0. This highlights the peculiarity of the associated kernel which can change its shape according to the target position.

In order to construct an associated kernel $K_{x,h}(\cdot)$ from a parametric (discrete or continuous) probability distribution $K_{\theta}$, $\theta \in \Theta \subseteq \mathbb{R}^d$ on the support $S_{\theta}$ such that $S_{\theta} \cap T \neq \emptyset$, we need to establish a correspondence between $(x,h) \in T \times (0,\infty)$ and $\theta \in \Theta$; see Kokonendji and Senga Kiessé (2011).

In what follows, we will call $K \equiv K_{\theta}$ the type of kernel to make a difference from the classical notion of a continuous symmetric (e.g., Gaussian) kernel. In this context, the choice of the associated kernel becomes important as well as that of the bandwidth. Moreover, we distinguish the associated kernels said sometimes of “second order” of those said of “first order” which verify the two first conditions (1) and (2). The rest of this section is devoted to discuss examples of associated kernels in both discrete and continuous cases.

Discrete associated kernels

Among the discrete associated kernels found in literature, we here use the best in sense of Definition 34.2.1. Negative binomial and Poisson kernels are respectively overdispersed (i.e., $V(Z_{x,h}) > E(Z_{x,h})$) and equidispersed (i.e., $V(Z_{x,h}) = E(Z_{x,h})$) and thus are not recommended; see Kokonendji and Senga Kiessé (2011) for further details. The first associated kernel listed below, namely the binomial kernel, is the best of the first order or *standard* kernels which satisfies

\[
\lim_{h \to 0} V(Z_{x,h}) \in V(0),
\]

where $V(0)$ is a neighborhood of 0 which does not depend on $x$. The two other discrete associated kernels satisfy all conditions of Definition 34.2.1.

- The binomial (bin) kernel is defined on the support $S_x = \{0,1,\ldots,x+1\}$ with $x \in T := \mathbb{N} = \{0,1,\ldots\}$ and then $h \in (0,1]$:

\[
B_{x,h}(u) = \frac{(x+1)!}{u!(x+1-u)!} \left( \frac{x+h}{x+1} \right)^u \left( \frac{1-h}{x+1} \right)^{x+1-u} 1_{S_x}(u),
\]

where $1_A$ denotes the indicator function of any given event $A$. Note that $B_{x,h}$ is the p.m.f. of the binomial distribution $B(x+1; (x+h)/(x+1))$ with its number of trials $x+1$ and its success probability in each trial $(x+h)/(x+1)$. It is appropriate for count data with small or moderate sample sizes and, also, it satisfies (4) rather than (3); see Kokonendji and Senga Kiessé (2011) and also Zougab et al. (2012) for a bandwidth selection by Bayesian method.
• The following class of symmetric discrete triangular kernels has been proposed in Kokonendji et al. (2007). The support $T$ of the p.m.f. $f$ to be estimated, can be unbounded (e.g., $\mathbb{N}, \mathbb{Z}$) or finite (e.g., $\{0, 1, \ldots, N\}$). Then, suppose that $h$ is a given bandwidth parameter and $a$ is an arbitrary and fixed integer. For fixed arm $a \in \mathbb{N}$, the discrete triangular $(DTa)$ kernel is defined on $S_{T,a} = \{x, x+1, \ldots, x+a\}$ with $x \in T = \mathbb{N}$:

$$DT_{x,h,a}(u) = \frac{(a+1)^h - |u-x|^h}{P(a,h)} I_{S_{T,a}}(u),$$

where $P(a,h) = (2a+1)(a+1) - 2 \sum_{i=0}^{a} k^h$ is the normalizing constant. It is symmetric around the target $x$, satisfying Definition 34.2.1 and suitable for count variables; see Kokonendji and Zocchi (2010) for an asymmetric version. Note that $h \to 0$ gives the Dirac kernel.

Continuous associated kernels

One can find several continuous associated kernels in literature among the Birnbaum-Saunders of $\mathcal{LN}$, gamma of $\mathcal{GA}$, and reciprocal inverse Gaussian of $\mathcal{RIG}$. Continuous associated kernels

- $\mathcal{BE}$ kernel

The extended beta ($\mathcal{BE}$) kernel is defined on $S_{x,h,a,b} = [a,b] \subset T$ with $a < b < \infty$, $x \in T$ and $h > 0$ such that

$$BE_{x,h,a,b}(u) = \frac{(a-b)^{(x-a)/h} (b-u)^{(b-x)/h}}{(b-a)^{1+h} B (1 + (x-a)/h, 1 + (b-x)/h) I_{S_{x,h,a,b}}(u)}$$

where $B(r,s) = \int_0^1 t^{r-1}(1-t)^{s-1} dt$ is the usual beta function with $r > 0, s > 0$; see Libengué (2013). For $a = 0$ and $b = 1$, it corresponds to the beta kernel (Chen, 1999) which is the p.d.f. of the beta distribution with shape parameters $1 + x/h$ and $(1 - x)/h$. The extended beta kernel is appropriate for any compact support of observations.

- $\mathcal{GA}$ kernel

The gamma ($\mathcal{GA}$) kernel is given on $S_{x,h} = [0,\infty) \subset T$ with $x \in T$ and $h > 0$:

$$GA_{x,h}(u) = \frac{u_x}{\Gamma(1 + x/h)} h^{1 + x/h} \exp \left( - \frac{u_x}{h} \right) I_{[0,\infty)}(u),$$

where $\Gamma(v) = \int_0^\infty s^{v-1} \exp(-s) ds$ is the classical gamma function with $v > 0$; see Chen (2000). It is the p.d.f. of the gamma distribution $\mathcal{GA}(1 + x/h, h)$ with scale parameter $1 + x/h$ and shape parameter $h$. It is suitable for the non-negative real set $T = [0,\infty)$.

- $\mathcal{LN}$ kernel

The lognormal ($\mathcal{LN}$) kernel is defined on $S_{x,h} = [0,\infty) \subset T$ with $x \in T$ and $h > 0$ such that

$$LN_{x,h}(u) = \frac{1}{uh\sqrt{2\pi}} \exp \left( - \frac{1}{2} \left( \frac{1}{h} \log \left( \frac{u_x}{h} \right) - h \right)^2 \right) I_{S_{x,h}}(u);$$

see Libengué (2013) and also Igarashi and Kakizawa (2015). It is the p.d.f. of the classical lognormal distribution with mean $\log(x) + h^2$ and standard deviation $h$.

- $\mathcal{RIG}$ kernel

The reciprocal inverse Gaussian ($\mathcal{RIG}$) kernel is given on $S_{x,h} = (0,\infty) \subset T$ with $x \in T$ and $h > 0$:

$$RIG_{x,h}(u) = \frac{1}{\sqrt{2\pi hu}} \exp \left( - \frac{(x^2 + xh)^{1/2}}{2h} \left( \frac{u_x}{(x^2 + xh)^{1/2}} - 2 + \frac{(x^2 + xh)^{1/2}}{u_x} \right) \right) I_{S_{x,h}}(u);$$

where $P(a,h) = (2a+1)(a+1) - 2 \sum_{i=0}^{a} k^h$ is the normalizing constant. It is symmetric around the target $x$, satisfying Definition 34.2.1 and suitable for count variables; see Kokonendji and Zocchi (2010) for an asymmetric version. Note that $h \to 0$ gives the Dirac kernel.
Figure 1: Shapes of univariate (discrete and continuous) associated kernels: (a) DiracDU, discrete triangular \( a = 3 \) and binomial with same target \( x = 3 \) and bandwidth \( h = 0.13 \); (b) lognormal, inverse Gaussian, gamma, reciprocal inverse Gaussian, inverse gamma and Gaussian with same target \( x = 1.3 \) and \( h = 0.2 \).

see Scaillet (2004), Libengué (2013) and also Igarashi and Kakizawa (2015). It is the p.d.f. of the classical reciprocal inverse Gaussian distribution with mean \( 1/\sqrt{x^2 + \chi h} \) and standard deviation \( 1/h \).

Remark 34.2.3. The three continuous associated kernels inverse gamma, inverse Gaussian and Gaussian are not adapted for density estimation on supports \([0, \infty)\) and thus are not included in the Ake package; see Part (b) of Figure 1.

Indeed:

- The inverse gamma (IGA) kernel, defined on \( S_{x,h} = (0, \infty) = \mathbb{T} \) with \( x \in (0, 1/h) \) and \( h > 0 \) such that

\[
IGA_{x,h}(u) = \frac{h^{1-1/(x h)}}{1 - 1/(x h)} u^{-1/(x h)} \exp \left( - \frac{1}{h u} \right) 1_{(0,\infty)}(u)
\]

(Libengué, 2013), is graphically the worst since it does not well concentrate on the target \( x \). Note that it is the p.d.f. of the inverse gamma distribution with scale parameter \(-1 + 1/(x h)\) and scale parameter \(1/h\).

- Also, the inverse Gaussian (IG) kernel, defined on \( S_{x,h} = (0, \infty) = \mathbb{T} \) with \( x \in (0, 1/3h) \) and \( h > 0 \) by

\[
IG_{x,h}(u) = \frac{1}{\sqrt{2\pi hu}} \exp \left\{ - \frac{(1 - 3xh)^{1/2}}{2h} \left( \frac{u}{(1 - 3xh)^{1/2}} - 2 + \frac{(1 - 3xh)^{1/2}}{u} \right) \right\} 1_{S_{x,h}}(u)
\]

(Scailliet, 2004; Libengué, 2013), has the same graphical properties as the inverse gamma. Note that it is the p.d.f. of the inverse Gaussian distribution \( IG(1 + x/h, h) \) with scale parameter \( x/(1 - 3xh)^{1/2} \) and shape parameter \( 1/h \).

- From the well known Gaussian kernel

\[
K_{G,x,h}(u) = \frac{1}{h \sqrt{2\pi}} \exp \left\{ \frac{1}{2} \left( \frac{u - x}{h} \right)^2 \right\} 1_{\mathbb{R}}(u).
\]

It has the same shape at any target and thus is well adapted for continuous variables with unbounded supports but not for \([0, \infty)\) or compact set of \( \mathbb{R} \); see also Epanechnikov (1969) for another example of a continuous symmetric kernel.

Figure 1 shows some forms of the above-mentioned univariate associated kernels. The plots highlight the importance given to the target point and around it in discrete (a) and continuous (b) cases. Furthermore, for a fixed bandwidth \( h \), the Gaussian keeps its same shape along the support; however, they change according to the target for the other non-classical associated kernels. This
explains the inappropriateness of the Gaussian kernel for density or regression estimation in any bounded interval and of the DiracDU kernel for count regression estimation; see Part (b) of Figure 1. From Part (v) of Remark 34.2.2, the inverse gamma and inverse Gaussian are the worst since they do not well concentrate on the target \( x \); see Remark 34.2.3. These previous associated kernels can be applied to various functionals.

We have implemented in R the method `kern.fun` for both discrete and continuous associated kernels. Seven possibilities are allowed for the kernel function. We enumerate the arguments and results of the default `kern.fun.default` function in Table 1. The `kern.fun` is used as follows for the binomial kernel:

\[
\text{R> x <- 5; h <- 0.1; y <- 0:10; k_b <- kern.fun(x, y, h, "discrete", "bino")}
\]

\[\text{Density or probability mass function estimations}\]

The p.d.f. or p.m.f. estimation is an usual application of the associated kernels. Let \( X_1, \ldots, X_n \) be independent and identically distributed (i.i.d.) random variables with an unknown p.d.f. (respectively p.m.f.) \( f \) on \( T \). An associated kernel estimator \( \hat{f}_n \) of \( f \) is simply:

\[
\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(X_i), \quad x \in T.
\] (5)

Here, we point out some pointwise properties of the estimator (5) in both discrete and continuous cases.

**Proposition 34.3.1.** (Kokonendi and Senga Kiessé, 2011; Libengué, 2013) Let \( X_1, X_2, \ldots, X_n \) be an \( n \) random sample i.i.d. from the unknown p.m.f. (respectively p.d.f.) \( f \) on \( T \). Let \( \hat{f}_n = \hat{f}_{n,h,k} \) be an estimator (5) of \( f \) with an associated kernel. Then, for all \( x \in T \) and \( h > 0 \), we have

\[
\mathbb{E}\{\hat{f}(x)\} = \mathbb{E}\{f(Z_{x,h})\},
\]

where \( Z_{x,h} \) is the random variable associated to the p.m.f. (respectively p.d.f.) \( K_{x,h} \) on \( S_{x,h} \). Furthermore, for a p.m.f. (respectively p.d.f.), we have respectively \( \hat{f}_n(x) \in [0,1] \) (respectively \( \hat{f}_n(x) > 0 \)) for all \( x \in T \) and

\[
\int_{x \in T} \hat{f}_n(x) \nu(dx) = C_n,
\] (6)

where \( C_n = C(n; h, K) \) is a positive and finite constant if \( \int_{T} K_{x,h}(t) \nu(dx) < \infty \) for all \( t \in T \), and \( \nu \) is a count or Lebesgue measure on \( T \).

It is easy to see that \( C_n = 1 \) for the estimators (5) with DiracDU kernel or any classical (symmetric)
associated kernel. Indeed, for the DiracDU kernel estimation we have

\[
\sum_{x=0}^{c-1} f_n(x) = \sum_{x=0}^{c-1} \left\{ (1 - h) \mathbb{I}_{\{x\}}(X_1) + \frac{h}{c-1} \mathbb{I}_{(x+1)\{c\}}(X_1) \right\} \\
= (1 - h) + \frac{h}{c-1} (c - 1) \\
= 1.
\]

In general we have \( C_n \neq 1 \) for other discrete and also continuous associated kernels, but it is always close to 1. In practice, we compute \( C_n \) depending on observations before normalizing \( f_n \) to be a p.m.f. or a p.d.f. The following code helps to compute the normalizing constant, e.g., for gamma kernel

\[
\text{R> f}_{\text{C}_n} = \frac{\text{f}_n(x)}{\sup_{T} f_n(x)} v(dx), \quad x \in T.
\]

Another alternative is to use an adaptive normalization of (5) according to each target \( x \):

\[
\tilde{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{K_{x,h}(X_i)}{\sup_{T} K_{x,h}(X_i)} v(dx), \quad x \in T,
\]

but this approach, with similar results than (7), is not used here. The representations are done with the global normalization (7). In the package, we also compute the normalizing constant (6) for any data set.

In discrete cases, the integrated squared error (ISE) defined by

\[
\text{ISE}_0 = \sum_{x \in \mathbb{N}} \{f_n(x) - f_0(x)\}^2,
\]

is the criteria used to measure numerically the discrete smoothness of \( \tilde{f}_n \) from (7) with the empirical or naive p.m.f. \( f_0 \) such that \( \sum_{x \in \mathbb{N}} f_0(x) = 1 \); see, e.g., Kokonendi and Senga Kiessé (2011). Concerning the continuous variables, the histogram gives a graphical measure of comparison with \( f_n \); see, for example, Figure 2.

**Some theoretical aspects of the normalizing constant**

In this section, we present some theoretical aspects of the normalizing constant \( C_n \) of (6) and two examples in the continuous and discrete cases. We first recall the following result on pointwise properties of the estimator (5).

**Lemma 34.3.2.** (Kokonendi and Senga Kiessé, 2011; Libengué, 2013) Let \( x \in T \) be a target and \( h \equiv h_n \) a bandwidth. Assuming \( f \) is in the class \( \mathcal{C}^2(T) \) in the continuous case, then

\[
\text{Bias} \left\{ \hat{f}_n(x) \right\} = A(x,h) f'(x) + \frac{1}{2} \left( A^2(x,h) + B(x,h) \right) f''(x) + o(h^2).
\]

Similar expressions (8) hold in the discrete case, except that \( f' \) and \( f'' \) are finite differences of the first and second order respectively.

Furthermore, for the continuous case, if \( f \) is bounded on \( T \) then there exists \( r_2 = r_2(K_{x,h}) > 0 \) the largest
real number such that \( \|K_{\gamma,h}\|_2^2 := \int_{S_{\gamma,h}} K_{\gamma,h}^2(u)\,du \leq c_2(x)h^{-2}, 0 \leq c_2(x) \leq \infty \) and

\[
\mathbb{V}\left(\hat{f}_n(x)\right) = \frac{1}{n} f(x) \|K_{\gamma,h}\|_2^2 + o\left(\frac{1}{nh^2}\right). \tag{9}
\]

For discrete situations, the result (9) becomes

\[
\mathbb{V}\left(\hat{f}_n(x)\right) = \frac{1}{n} f(x)\{(P(Z_{\gamma,h} = x))^2 - f(x)\},
\]

where \( Z_{\gamma,h} \) denotes the discrete random variable with p.m.f. \( K_{\gamma,h}^1 \).

It is noticeable that the bias (8) is bigger than the one with symmetric kernels and thus can be reduced; see, e.g., Zhang (2010), Zhang and Karunamuni (2010) and Libengue (2013).

**Proposition 34.3.3.** Following notations in Lemma 34.3.2, the mean and variance of \( C_n \) of \((6)\) are respectively:

\[
\mathbb{E}\left(C_n\right) \simeq 1 + \frac{\sup(T)}{\inf(T)} \left\{ A(x,h)f'(x) + \frac{1}{2} \left[H^2(x,h) + B(x,h)\right] f''(x)\right\} \nu(dx), \tag{10}
\]

\[
\mathbb{V}\left(C_n\right) \simeq \begin{cases} 
\frac{1}{n} \int_{\inf(T)}^{\sup(T)} \left(f(x) \|K_{\gamma,h}\|_2^2\right) \,dx & \text{if } T \text{ is continuous,} \\
\frac{1}{n} \sum_{x=\frac{1}{n}} \left(f(x)\{(P(Z_{\gamma,h} = x))^2 - f(x)\}\right) & \text{if } T \text{ is discrete,}
\end{cases} \tag{11}
\]

where \( \inf(T) \) and \( \sup(T) \) are respectively the infimum and supremum of \( T \), the measure \( \nu \) is Lesbesgue or count on the support \( T \), and where "\( \simeq \)" stands for approximation.

**Proof.** From Lemma 34.3.2 and the Fubini theorem, we successively show (10) as follows:

\[
\mathbb{E}\left(C_n\right) = \mathbb{E}\left(\int_{\inf(T)}^{\sup(T)} \hat{f}_n(x)\nu(dx)\right) = \int_{\inf(T)}^{\sup(T)} \mathbb{E}\left(\hat{f}_n(x)\right)\nu(dx)
\]

\[
= \int_{\inf(T)}^{\sup(T)} \left[ \text{Bias} \left\{ \hat{f}_n(x) \right\} + f(x) \right] \nu(dx)
\]

\[
\simeq \int_{\inf(T)}^{\sup(T)} \left\{ A(x,h)f'(x) + \frac{1}{2} \left[H^2(x,h) + B(x,h)\right] f''(x)\right\} \nu(dx)
\]

\[
\simeq \int_{\inf(T)}^{\sup(T)} f(x)\nu(dx) + \int_{\inf(T)}^{\sup(T)} \left\{ A(x,h)f'(x) + \frac{1}{2} \left[H^2(x,h) + B(x,h)\right] f''(x)\right\} \nu(dx)
\]

\[
\simeq 1 + \int_{\inf(T)}^{\sup(T)} \left\{ A(x,h)f'(x) + \frac{1}{2} \left[H^2(x,h) + B(x,h)\right] f''(x)\right\} \nu(dx).
\]

The variance (11) is trivial from Lemma 34.3.2. \( \square \)

**Example 1.** Let \( f \) be an exponential density with parameter \( \gamma > 0 \). Thus, one has:

\[
f(x) = \gamma \exp(-\gamma x),
\]

\[
f'(x) = -\gamma^2 \exp(-\gamma x),
\]

\[
f''(x) = \gamma^3 \exp(-\gamma x).
\]

Consider the lognormal kernel with \( A(x,h) = x(\exp(3h^2/2) - 1), B(x,h) = x^2 \exp(3h^2)(\exp(h^2) - 1) \) and \( \|\mathcal{L}_{\gamma,h}\|_2^2 = 1/(2\pi h x^2/2) \); see Libengue (2013). Then, using the Taylor formula around \( h \), the expressions of \( \mathbb{E}\left(C_n\right) \) and \( \mathbb{V}\left(C_n\right) \) are:

\[
\mathbb{E}\left(C_n\right) \simeq 1 + \left[-1 + \exp\left(\frac{3h^2}{2}\right) + \frac{1}{2} \left(-1 + \exp\left(\frac{3h^2}{2}\right) + \exp\left(\frac{3h^2}{2} - 1 + h^2\right)\right)\right] \simeq 1 - \frac{h^2}{2}
\]

and

\[
\mathbb{V}\left(C_n\right) \simeq \gamma \left[2n 2\sqrt{\pi} \right]^{-1} \int_0^{\infty} z^{-1} \exp(-z)dz + \int_0^{\infty} z^{-1} \exp(-z)dz \simeq 16.2340 \text{ by computation with } R. \text{ Thus, the quantity } C_n \text{ cannot be equal to } 1.
\]

**Example 2.** Let \( f \) be a Poisson p.m.f. with parameter \( \lambda \) and thus \( f(x) = \lambda^x \exp(-\lambda)/x! \). The finite differences \( f^{(k)}(x) \) of order \( k \in \{1,2,\ldots\} \) at \( x \in \mathbb{N} \) are given by the recursive relation:

\[
f^{(k)}(x) = \{f^{(k-1)}(x)\}^{(1)} \text{ with } f^{(1)}(x) = \begin{cases} 
\{f(x + 1) - f(x - 1)\}/2, & \text{if } x \in \mathbb{N} \setminus \{0\}, \\
f(1) - f(0), & \text{if } x = 0,
\end{cases}
\]
and
\[ f^{(2)}(x) = \begin{cases} \frac{f(x+2) - 2f(x) + f(x-2)}{4}, & \text{if } x \in \mathbb{N} \setminus \{0,1\}, \\ \frac{f(3) - 3f(1) + f(0)}{4}, & \text{if } x = 1, \\ \frac{f(2) - 2f(1) + f(0)}{2}, & \text{if } x = 0. \end{cases} \]

Considering the binomial kernel with \( A(x,h) = x+h, B(x,h) = (x+h)(1-h)/(x+1) \), we successively obtain
\[
\mathbb{E}(C_n) \simeq 1 + h \exp(-\lambda) \left[ \frac{\lambda^3}{3!} + \frac{3\lambda^2}{2!} - 3\lambda - 2 \right]
+ \sum_{x=2}^{\infty} \frac{2(x+1)}{x} \exp(-\lambda) \left[ \frac{\lambda^{x+1}}{(x+1)!} - \frac{2\lambda^{x-1}}{(x-1)!} + \frac{x+3}{x+1} \left( \frac{\lambda^{x+2}}{(x+2)!} - \frac{2\lambda^{x-2}}{(x-2)!} \right) \right] + \frac{\lambda^3}{4} + 2\lambda^2 - \frac{17\lambda}{2} + 3
+ \sum_{x=2}^{\infty} \frac{x}{2} \left( \frac{\lambda^{x+1}}{(x+1)!} - \frac{\lambda^{x-1}}{(x-1)!} \right) + \frac{x^3 + x^2 + 1}{x+1} \left( \frac{\lambda^{x+2}}{(x+2)!} - \frac{2\lambda^{x-2}}{(x-2)!} \right)
\]
and
\[
\mathbb{V}(C_n) \simeq \frac{\exp(-\lambda)}{n} \sum_{x=1}^{\infty} \left[ \frac{\lambda^x}{x!} \left( (1+h) \frac{x+h}{x+1} \right)^2 - \frac{\lambda^x \exp(-\lambda)}{x!} \right].
\]

### Bandwidth selection

Now, we consider the bandwidth selection problems which are generally crucial in nonparametric estimation. Several methods already existing for continuous kernels can be adapted to the discrete case as the classical least-squares cross-validation method; see, for example, Bowman (1984), Marron (1987) and references therein. Here, we simply propose three procedures for the bandwidth selection: cross-validation, Bayesian local for binomial and adaptive for the gamma kernel. Also, a review of bayesian bandwidth selection methods is presented. Each time, the smoothing parameter selection is done with the non-normalized version \( \hat{f}_n \) of the estimator (5) before the global normalization \( \tilde{f}_n \) of (7).

### Cross-validation for any associated kernel

For a given associated kernel \( K_{x,h} \) with \( x \in \mathbb{T} \) and \( h > 0 \), the optimal bandwidth \( h_{cv} \) of \( h \) is obtained by cross-validation as \( h_{cv} = \arg \min_{h>0} CV(h) \) with
\[
CV(h) = \int_{x \in \mathbb{T}} \left( \hat{f}_n(x) \right)^2 \nu(dx) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{n,i}(X_i),
\]
where \( \hat{f}_{n,i}(X_i) = (n-1)^{-1} \sum_{j \neq i} K_{X_i,h}(X_j) \) is being computed as \( \hat{f}_n(X_i) \) by excluding the observation
\( X_i \) and \( \nu \) is the Lebesgue or count measure. This method is applied to all estimators (5) with associated kernels cited in this paper, independently on the support \( \mathbb{T} \) of \( f \) to be estimated.

Table 2 gives the arguments and results of the cross-validation function \( hcvd.fun \) defined for continuous data are below. The \( hcvd.fun \) is the corresponding function for discrete data. The \( hcvd.fun \) is performed with the Old Faithful geyser data described in Azzalini and Bowman (1990) and Härdle (2012). The dataset concerns waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. The following codes and Figure 2 give smoothing density estimation with various associated kernels of the waiting time variable.
Arguments | Description
---|---
Vec | The positive continuous data sample.
seq.bws | The sequence of bandwidths where to compute the cross-validation function.
ker | The associated kernel.
a0, a1 | The bounds of the support of extended beta kernel. Default values are respectively 0 and 1.
a | The arm of the discrete triangular kernel. Default value is 1.
c | The number of categories in DiracDU kernel. Default value is 2.

Results | Description
---|---
hcv | The optimal bandwidth obtained by cross-validation.
seq.h | The sequence of bandwidths used to compute hcv.
CV | The values of the cross-validation function.

Table 2: Summary of arguments and results of hcvc.fun.

![Density estimation of Old Faithful geyser data](image)

**Figure 2:** Smoothing density estimation of the Old Faithful geyser data (Azzalini and Bowman, 1990) by some continuous associated kernels with the support of observations $[43, 96] = \mathbb{T}$.
A review of Bayesian bandwidth selection

Bayesian inference grows out of the simple formula known as Bayes rule. Assume we have two random variables \(A\) and \(B\). A principle rule of probability theory known as the chain rule allows us to specify the joint probability of \(A\) and \(B\) taking on particular values \(a\) and \(b\), \(P(a,b)\), as the product of the marginal probability that \(A\) will take on value \(a\) given that \(B\) has taken on value \(b\), \(P(a|b)\), and the marginal probability that \(B\) takes on value \(b\), \(P(b)\). Which gives us:

\[
\text{Joint probability} = \text{Conditional Probability} \times \text{Marginal Probability}.
\]

Thus we have:

\[
P(a,b) = P(a|b)P(b).
\]

This expression (Bayes rule) indicates that we can compute the conditional probability of a variable \(A\) given the variable \(B\) from the conditional probability of \(B\) given \(A\). This introduces the notion of prior and posterior knowledge.

**Prior and posterior knowledge.** A prior probability is the probability available to us beforehand, and before making any additional observations. A posterior probability is the probability obtained from the prior probability after making additional observation to the prior knowledge available.

**Summarizing the Bayesian approach.** The Bayesian approach to parameter estimation works as follows:

1. Formulate our knowledge about a situation.
2. Gather data.
3. Obtain posterior knowledge that updates our beliefs.

How do we formulate our knowledge about a situation?

a. Define a distribution model which expresses qualitative aspects of our knowledge about the situation. This model will have some unknown parameters, which will be dealt with as random variables.

b. Specify a prior probability distribution which expresses our subjective beliefs and subjective uncertainty about the unknown parameters, before seeing the data.

After gathering the data, how do we obtain posterior knowledge?

c. Compute posterior probability distribution which estimates the unknown parameters using the rules of probability and given the observed data, presenting us with updated beliefs.

**Zougab et al. (2013)** proposed a Bayesian approach based upon a likelihood cross-validation approximation and a Markov chain Monte Carlo (MCMC) method for deriving the global optimal bandwidth using the famous binomial kernel. However, a global bandwidth does not generally provide a good estimator for complex p.m.f.'s, in particular for small and moderate sample sizes. Generally, the global discrete associated kernel estimator tends to simultaneously under- and oversmooth \(f(x)\).

In order to improve the global discrete associated kernel estimator, in particular for complex count data with small and moderate sample sizes, Zougab et al. (2012) and Zougab et al. (2013) adapted two versions of variable bandwidths for discrete associated kernel estimator and proposed Bayesian approaches for selecting these variable bandwidths. Note that these two versions are originally proposed for kernel density estimation (see, e.g., Sain and Scott 1996; Breiman et al. 1977; Abramson 1982; Brewer 2000; Zhang et al. 2006; Zougab et al. 2014a and Zhang et al. 2016).

Recently, Zougab et al. (2012) have considered the local discrete associated kernel estimator (balloon estimator in discrete case) and have derived the closed form of the variable bandwidth at each point \(x\) for which the p.m.f. is estimated by considering the binomial kernel estimator and locally treating the bandwidth as a random quantity with a beta prior distribution. This approach outperforms existing classical global methods, namely, MISE and CV in particular for small and moderate sample sizes. Zougab et al. (2013) have also proposed the adaptive discrete associated kernel estimator (sample-point estimator in discrete situation), which replaces \(b\) by \(h_i\) for each observation \(x_i\) with \(i = 1, \ldots, n\), and then employs the Bayesian approach for estimating the adaptive bandwidths \(h_i\). The authors have considered the binomial kernel and the beta prior for each variable bandwidth \(h_i\), and have shown that this approach performs better than the popular classical global selectors.
Bayesian estimation of localized bandwidth for the binomial kernel

An alternative to the cross-validation for bandwidth selection is by using Bayesian methods. These methods have been investigated with three different procedures: local, global and adaptive; see respectively Zougab et al. (2012, 2013, 2014b). In terms of integrated squared error and execution times, the local Bayesian outperforms the other Bayesian procedures. In the local Bayesian framework, the variable bandwidth is treated as parameter with prior \( \pi(\cdot) \). Under squared error loss function, the Bayesian bandwidth selector is the posterior mean; see Zougab et al. (2012).

First, as we have mentioned above, \( f(x|h) \) can be approximated by

\[
f(x|h) = f_\hat{h}(x) = \sum_{u \in X} f(u) B_{x,h}(u) = E\{B_{x,h}(X)\},
\]

where \( B_{x,h} \) is the binomial kernel and \( X \) is a random variable with p.m.f. \( f \). Now, considering \( h \) as a scale parameter for \( f_\hat{h}(x) \), the local approach consists of using \( f_\hat{h}(x) \) and constructing a Bayesian estimator for \( h(x) \).

Indeed, let \( \pi(h) \) denote the beta prior density of \( h \) with positive parameters \( \alpha \) and \( \beta \). By the Bayes theorem, the posterior of \( h \) at the point of estimation \( x \) takes the form

\[
\pi(h|x) = \frac{f_\hat{h}(x)\pi(h)}{\int f_\hat{h}(x)\pi(h)dh}.
\]

Since \( f_\hat{h} \) is unknown, we use \( f_\hat{h} \) as natural estimator of \( f_\hat{h} \), and hence we can estimate the posterior by

\[
\pi(h|x, X_1, X_2, \ldots, X_n) = \frac{\tilde{f}_\hat{h}(x)\pi(h)}{\int \tilde{f}_\hat{h}(x)\pi(h)dh}.
\]

Under the squared error loss, the Bayes estimator of the smoothing parameter \( h(x) \) is the posterior mean and is given by \( \tilde{h}_n(x) = \int h\tilde{r}(h|x, X_1, X_2, \ldots, X_n)dh \). Exact approximation is

\[
\tilde{h}_n(x) = \sum_{i=0}^{n} \sum_{k=0}^{x} \frac{x^k}{(x+1-X_i)!k!(X_i-k)!} B(X_i + \alpha - k + 1, x + \beta + 1 - X_i)
\]

\[
\sum_{i=0}^{n} \sum_{k=0}^{x} \frac{x^k}{(x+1-X_i)!k!(X_i-k)!} B(X_i + \alpha - k, x + \beta + 1 - X_i),
\]

where \( B(\cdot, \cdot) \) is the beta function; see Zougab et al. (2012) for more details.

Bayesian estimation of adaptive bandwidth for the gamma kernel

The bandwidth \( h \) in the gamma kernel density estimation can be allowed to be adaptive. This approach gives a variable bandwidth \( h_i \) for each observation \( X_i \) in place of the initial fixed bandwidth \( h \). Following Zougab et al. (2014a), we suggest using Bayesian methods to estimate such adaptive bandwidths or variable bandwidths \( h_i, i = 1, \ldots, n \). Thus, we treat \( h_i \) as a random variable with a prior distribution \( \pi(\cdot) \). The estimator (5) with gamma kernel of Section Continuous associated kernels and variable bandwidths are reformulated as

\[
\tilde{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} GA_{X,h_i}(X_i).
\]

The leave-one-out kernel estimator of \( f(X_i) \) deduced from (12) is

\[
\tilde{f}(X_i | \{X_{-i}\}, h_i) = \frac{1}{n-1} \sum_{j=1,j \neq i}^{n} GA_{X,h_i}(X_j),
\]

where \( \{X_{-i}\} \) denotes the set of observations excluding \( X_i \). The posterior distribution for each variable bandwidth \( h_i \) given \( X_i \) provided from the Bayesian rule is expressed as follow

\[
\pi(h_i | X_i) = \frac{\tilde{f}(X_i | \{X_{-i}\}, h_i)\pi(h_i)}{\int_0^\infty \tilde{f}(X_i | \{X_{-i}\}, h_i)\pi(h_i)dh_i}.
\]

We obtain the Bayesian estimator \( \tilde{h}_i \) of \( h_i \) by using the quadratic loss function

\[
\tilde{h}_i = E(h_i | X_i).
\]
In the following, we assume that each $h_i = h_i(n)$ has an inverse gamma prior distribution $IG_A(a, \beta)$ with the shape parameter $a > 0$ and scale parameter $\beta > 0$. The density of $IG_A(a, b)$ with $a, b > 0$ is defined as

$$
\Phi_{a, b}(z) = \frac{b^a}{\Gamma(a)} z^{-a-1} \exp(-b/z) \mathbb{1}_{(0, \infty)}(z). \quad (16)
$$

This allows us to obtain the closed form of the posterior density and the Bayesian estimator given by the following result.

**Theorem 34.3.4.** For fixed $i \in \{1, 2, \ldots, n\}$, consider each observation $X_i$ with its corresponding bandwidth $h_i$. Using the gamma kernel estimator (12) and the inverse gamma prior distribution $IG_A(a, \beta)$ given in (16) with $a > 1/2$ and $\beta > 0$ for each $h_i$, then:

(i) The posterior density (14) is the following weighted sum of inverse gamma

$$
\pi(h_i \mid X_i) = \frac{1}{D_{ij}} \sum_{j=1,j\neq i}^{n} \left\{ A_{ij} \Phi_{a+1/2, b_j}(h_i) \mathbb{1}_{(0, \infty)}(X_i) + C_j \Phi_{a+1, b_i}(h_i) \mathbb{1}_{(0)}(X_i) \right\}
$$

with $A_{ij} = [\Gamma(a + 1/2)]/(\beta^a X_j^{1/2} \sqrt{2\pi} B_{ij}^{a+1/2})$, $B_{ij} = X_i \log X_i - X_i \log X_j + X_j - X_i + \beta$, $C_j = \Gamma(a + 1)/[\beta^{-a}(X_j + \beta)^{a+1}]$ and $D_{ij} = \sum_{j=1,j\neq i}^{n} A_{ij} \mathbb{1}_{(0, \infty)}(X_i) + C_j \mathbb{1}_{(0)}(X_i)$.

(ii) The Bayesian estimator $\tilde{h}_i$ of $h_i$, given in (16), is

$$
\tilde{h}_i = \frac{1}{D_{ij}} \sum_{j=1,j\neq i}^{n} \left\{ A_{ij} B_{ij} \mathbb{1}_{(0, \infty)}(X_i) + \frac{(X_i + \beta) C_j}{a} \mathbb{1}_{(0)}(X_i) \right\}
$$

according to the previous notations of $A_{ij}$, $B_{ij}$, $C_j$ and $D_{ij}$.

**Proof.** (i) Let us represent $\pi(h_i \mid X_i)$ of (14) as the ratio of $N(h_i \mid X_i) := \int (X_i \mid \{X_{-i}, h_i\}) \pi(h_i)$ and $\int_0^\infty N(h_i \mid X_i) dh_i$. From (13) and (16) the numerator is, first, equal to

$$
N(h_i \mid X_i) = \frac{1}{n-1} \sum_{j=1,j\neq i}^{n} GA_{X_i,h_{i}}(X_j) \left( \frac{\Gamma(a)}{\beta^a (h_i)^{a+1}} \exp(-\beta/h_i) \right)
$$

$$
= \frac{\Gamma(a+1)}{(n-1)} \sum_{j=1,j\neq i}^{n} GA_{X_i,h_{i}}(X_j) \frac{\beta^a}{\beta^a (h_i)^{a+1}} \exp(-\beta/h_i).
$$

Following Chen (2000), we assume that for all $X_i \in (0, \infty)$ one has $1 + (X_i/h_i) \to \infty$ as $n \to \infty$. Using the Stirling formula $\Gamma(z+1) \simeq \sqrt{2\pi}z^{z+1/2} \exp(-z)$ as $z \to \infty$, the term of the sum in (17) can be successively calculated as

$$
\frac{GA_{X_i,h_{i}}(X_j)}{\beta^a (h_i)^{a+1}} \exp(-\beta/h_i) = \frac{X_j^{(X_i/h_i)}}{h_i \Gamma[1+(X_i/h_i)]} \frac{\exp(-X_j/h_i)}{\exp(-X_j + \beta - X_j \log X_j/h_i)}
$$

$$
= \frac{\beta^{-a} X_j^{(X_i/h_i)+a+1} \sqrt{2\pi} \exp(-X_j/h_i)(X_j/h_i)^{(X_i/h_i)+1/2}}{\beta^{-a} X_j^{1/2} \sqrt{2\pi} B_{ij}^{a+1/2}} \frac{\exp(-B_{ij}/h_i)}{h_i^{a+3/2} \Gamma(a + 1/2)}
$$

$$
= A_{ij} \Phi_{a+1, 2B_{ij}}(h_i), \quad (18)
$$

with $B_{ij} = X_i \log X_i - X_i \log X_j + X_j - X_i + \beta$, $A_{ij} = [X_j^{-1} \Gamma(a + 1/2)]/(\beta^{-a} X_j^{-1/2} \sqrt{2\pi} B_{ij}^{a+1/2})$ and $\Phi_{a+1/2, B_{ij}}(h_i)$ is given in (16).
Also, for \( X_i = 0 \), the term of the sum (17) can be expressed as follows

\[
\frac{G_{A_0,b_i}(x_i)}{\beta^{-a}h_i^{a+1}} \exp(-\beta/h_i) = \frac{\exp(-x_i/h_i)}{\beta^{-a}h_i^{a+2}} \exp(-\beta/h_i)
\]

\[
= \frac{\Gamma(a+1)}{\beta^{-a}(X_i + \beta)^{a+1}} \times \frac{(X_i + \beta)^{a+1} \exp[-(X_i + \beta)/h_i]}{h_i^{a+2}\Gamma(a+1)}
\]

\[
= C_j \Phi_{\alpha+1,X_i+\beta}(h_i),
\]

(19)

with \( C_j = \Gamma(a+1)/[\beta^{-a}(X_i + \beta)^{a+1}] \) and \( \Phi_{\alpha+1,X_i+\beta}(h_i) \) is given in (16). Combining (18) and (19), the expression of \( N(h_i \mid X_i) \) in (17) becomes

\[
N(h_i \mid X_i) = \frac{\Gamma(a)}{(n-1)} \sum_{j=1}^{n} \left\{ A_{ij} \Phi_{\alpha+1/2,\beta_j}(h_i) \mathbb{I}_{(0,\infty)}(X_i) + C_j \Phi_{\alpha+1+,\beta}(h_i) \mathbb{I}_{[0]}(X_i) \right\}.
\]

(20)

From (20), the denominator is successively computed as follows:

\[
\int_0^\infty N(h_i \mid X_i) \, dh_i = \frac{\Gamma(a)}{(n-1)} \sum_{j=1}^{n} \left\{ A_{ij} \int_0^\infty \Phi_{\alpha+1/2,\beta_j}(h_i) \mathbb{I}_{(0,\infty)}(X_i) \, dh_i + C_j \int_0^\infty \Phi_{\alpha+1+,\beta}(h_i) \mathbb{I}_{[0]}(X_i) \, dh_i \right\}
\]

\[
= \frac{\Gamma(a)}{(n-1)} \sum_{j=1}^{n} \left\{ A_{ij} \mathbb{I}_{(0,\infty)}(X_i) + C_j \mathbb{I}_{[0]}(X_i) \right\}
\]

\[
= \frac{\Gamma(a)}{(n-1)} D_{ij}
\]

(21)

with \( D_{ij} = \sum_{j=1}^{n} \left\{ A_{ij} \mathbb{I}_{(0,\infty)}(X_i) + C_j \mathbb{I}_{[0]}(X_i) \right\} \). Finally, the ratio of (20) and (21) leads to the result of Part (i).

(ii) Let us remember that the mean of the inverse gamma distribution \( IG(\alpha, \beta) \) is \( \beta/\alpha - 1 \). Thus, the expression of \( \pi(h_i \mid X_i) \) in (14) is given by

\[
\pi(h_i \mid X_i) = \frac{1}{D_{ij}} \sum_{j=1}^{n} \left\{ A_{ij} \Phi_{\alpha+1/2,\beta_j}(h_i) \mathbb{I}_{(0,\infty)}(X_i) + C_j \Phi_{\alpha+1+,\beta}(h_i) \mathbb{I}_{[0]}(X_i) \right\}
\]

and, therefore, \( \tilde{h}_i = \mathbb{E}(h_i \mid X_i) = \int_0^\infty h_i \pi(h_i \mid X_i) \, dh_i \) is finally

\[
\tilde{h}_i = \mathbb{E}(h_i \mid X_i) = \frac{1}{D_{ij}} \sum_{j=1}^{n} \left\{ A_{ij} \frac{B_{ij}}{\alpha - 1/2} \mathbb{I}_{(0,\infty)}(X_i) + \frac{(X_i + \beta)C_j}{\alpha} \mathbb{I}_{[0]}(X_i) \right\}
\]

This new method of selecting bandwidth by the Bayesian adaptive procedure will be implemented in a future version of the Ake package.

**Bandwidth selection for kernel regression involving associated kernels**

One of the most often encountered models in nonparametric statistics is the regression model. The function that provides the best prediction of a dependent variable \( y \) in terms of an independent variable \( x \) is the conditional expectation \( \mathbb{E}(y \mid x) = m(x) \). This is called regression function and its estimation from a sequence of \( n \) pairs \((x_i, y_i)\), \( i = 1, \ldots, n \) is a problem in statistics. We consider the case \((x, y) \in T \times R \). For simplicity, we take \( T = R \) if \( x \) is a continuous variable and \( T = N \) in the discrete case. The classical nonparametric regression model between two variables \( y \) and \( x \) is

\[
y_i = m(x_i) + \epsilon_i
\]

(22)

where \( y = (y_1, \ldots, y_n) \) is a response vector, \( x = (x_1, \ldots, x_n) \) is an explanatory vector, \( \epsilon = (\epsilon_1, \ldots, \epsilon_n) \) is the error following a Gaussian distribution with zero mean and finite variance \( \sigma^2 \), i.e., \( \epsilon_i \sim N(0, \sigma^2) \) and \( m : T \rightarrow R \) is the unknown regression function. Several methods have been proposed to estimate the regression function in the continuous case. We cite for example the histograms introduced by Tukey (1961) and studied by Geoffroy (1980) and Lecoutre (1990), the spline method which can be found in Reinsch (1967), Silverman (1985) and Wahba (1990) and also the regression using partition
proposed by Breiman et al. (1984).

As for the density or probability mass function, the estimate of the regression function by the kernel method is the most used because of its good asymptotic properties and interest in practice. Introduced initially for continuous density estimation by Rosenblatt (1956) and Parzen (1962), this method was adopted by Nadaraya (1964) and Watson (1964) for estimating the continuous regression function. It was also applied to smooth the discrete regression function \( m \) for \( x \in \mathbb{N} \). Some studies have been done to estimate the discrete regression function, using the Dirac-type kernel (naive estimator) or discrete kernels of Aitchison and Aitken (1976). However, the naive estimator is appropriate only when the sample size is large, and the discrete kernel of Aitchison and Aitken (1976) is only suitable for categorical data; see Hayfield and Racine (2008) and also Hayfield and Racine (2014). Kokonendji et al. (2009) adapted the Nadaraya (1964) and Watson (1964) kernel to the discrete unknown function \( m \), using the discrete associated kernels. In their work, using the integrated mean square error and the coefficient of determination \( R^2 \), they showed that the binomial or discrete triangular kernels are better compared to the optimal Epanechnikov kernel. In this section we present the theoretical foundations of the estimated regression function with continuous and discrete associated kernels.

Both in continuous and discrete cases, consider the relation between a response variable \( Y \) and an explanatory variable \( x \) given by

\[
Y = m(x) + \epsilon, \quad (23)
\]

where \( m \) is an unknown regression function from \( T \subseteq \mathbb{R} \) to \( \mathbb{R} \) and \( \epsilon \) the disturbance term with null mean and finite variance. Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be a sequence of i.i.d. random variables on \( T \times \mathbb{R}(\subseteq \mathbb{R}^2) \) with \( m(x) = E(Y|X=x) \) of (23). Using (continuous or discrete) associated kernels, the Nadaraya (1964) and Watson (1964) estimator \( \hat{m}_n \) of \( m \) is

\[
\hat{m}_n(x; h) = \frac{1}{n} \sum_{i=1}^{n} Y_i K_{x,h}(X_i) = \hat{m}_n(x), \quad \forall x \in T \subseteq \mathbb{R}, \quad (24)
\]

where \( h \equiv h_n \) is the smoothing parameter such that \( h_n \to 0 \) as \( n \to \infty \).

Besides the criterion of kernel support, we retain the root mean squared error (RMSE) and also the practical coefficient of determination given respectively by

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{m}_n(X_i))^2}
\]

and

\[
R^2 = \frac{\sum_{i=1}^{n} \{\hat{m}_n(X_i) - \bar{Y}\}^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2},
\]

with \( \bar{Y} = n^{-1}(Y_1 + \ldots + Y_n) \).

In discrete cases, the \( \text{reg.\ fun} \) function for (24) is used with the binomial kernel on milk data as follows. This dataset is about average daily fat (kg/day) yields from milk of a single cow for each of the first 35 weeks.

\[
R> \text{data("milk", package = "Ake")}
R> x <- milk$week
R> y <- milk$yield
R> h <- \text{reg.fun}(x, y, "discrete", "bino", 0.1)
R> h
Bandwidth h:0.1 Coef_det=0.9726

Number of points: 35; Kernel = Binomial

data
\[
\begin{array}{ll}
\text{Min.} & : 1.0 \quad \text{Min.} & : 0.0100 \\
1\text{st Qu.} & : 9.5 \quad 1\text{st Qu.} & : 0.2750 \\
\text{Median} & : 18.0 \quad \text{Median} & : 0.3680 \\
\text{Mean} & : 18.0 \quad \text{Mean} & : 0.3986 \\
3\text{rd Qu.} & : 26.5 \quad 3\text{rd Qu.} & : 0.6150 \\
\text{Max.} & : 35.0 \quad \text{Max.} & : 0.7200 \\
\text{eval.points} & : m_n \\
\text{Min.} & : 1.0 \quad \text{Min.} & : 0.0042 \\
1\text{st Qu.} & : 9.5 \quad 1\text{st Qu.} & : 0.27681
\end{array}
\]

The R Journal Vol. 8/2, December 2016 ISSN 2073-4859
Arguments Description

<table>
<thead>
<tr>
<th>Argumen</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vec</td>
<td>The explanatory data sample can be discrete or continuous.</td>
</tr>
<tr>
<td>y</td>
<td>The response variable.</td>
</tr>
<tr>
<td>ker</td>
<td>The associated kernel.</td>
</tr>
<tr>
<td>h</td>
<td>The sequence of bandwidths where to compute the optimal bandwidth.</td>
</tr>
<tr>
<td>a0, a1</td>
<td>The bounds of the support of extended beta kernel. Default values are respectively 0 and 1.</td>
</tr>
<tr>
<td>a</td>
<td>The arm of the discrete triangular kernel. Default value is 1.</td>
</tr>
<tr>
<td>c</td>
<td>The number of categories in DiracDU kernel. Default value is 2.</td>
</tr>
</tbody>
</table>

Results Description

<table>
<thead>
<tr>
<th>Result</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>kernel</td>
<td>The associated kernel.</td>
</tr>
<tr>
<td>hcv</td>
<td>The optimal bandwidth obtained by cross-validation.</td>
</tr>
<tr>
<td>CV</td>
<td>The values of the cross-validation function.</td>
</tr>
<tr>
<td>seqbws</td>
<td>The sequence of bandwidths used to compute hcv.</td>
</tr>
</tbody>
</table>

Table 3: Summary of arguments and results of hcvreg.fun.

<table>
<thead>
<tr>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.0</td>
<td>18.0</td>
<td>26.5</td>
<td>35.0</td>
</tr>
<tr>
<td>0.35065</td>
<td>0.39777</td>
<td>0.60942</td>
<td>0.70064</td>
</tr>
</tbody>
</table>

The above reg.fun is also used for continuous cases; see Figure 3 and Table 4 for the motorcycle impact data of Silverman (1985).

Bandwidth selection

We present two bandwidth selection methods for the regression: the well-known cross-validation for any associated kernel and the Bayesian global for the binomial kernel.

Cross-validation for any associated kernel

For a given associated kernel, the optimal bandwidth parameter is 
\[ \hat{h}_{cv} = \arg \min_{h > 0} LSCV(h) \]

where \( \tilde{m}_{-i}(X_i) \) is computed as \( \hat{m}_{-i} \) of (24) excluding \( X_i \); see, e.g., Kokonendji et al. (2009). The hcvreg.fun function to compute this optimal bandwidth is described in Table 3.

The following code helps to compute the bandwidth parameter by cross-validation on milk data. The associated kernel used is the discrete triangular kernel with arm \( a = 1 \).

```r
R> data("milk", package = "Ake")
R> x <- milk$week
R> y <- milk$yield
R> f <- hcvreg.fun(x, y, type_data = "discrete", ker = "triang", a = 1)
R> f$hcv
[1] 1.141073
```

When we consider the continuous associated kernel, one needs to set the type of data parameter to "continuous" in the hcvreg.fun function. Thus, the hcvreg.fun and reg.fun functions are used with gamma, lognormal, reciprocal inverse Gaussian and Gaussian kernel on the motor cycle impact data described in Silverman (1985). The observations consist of accelerometer reading taken through time in an experimentation on the efficiency of crash helmets. The results in Table 4 agree with the shapes of continuous associated kernels of Part (b) of Figure 1; see also Figure 3. In fact, since the lognormal kernel is well concentrated around the target \( x \), it gives the best \( R^2 \) which is 75.9%. The gamma and the reciprocal inverse Gaussian kernels give similar \( R^2 \) in the order 73%. Although the Gaussian kernel is well concentrated on the target, it gives the lower result of \( R^2 = 70.90\% \). This is mainly due to the symmetry of the kernel which cannot change its shapes according to the target.
Table 4: Some expected values of $R^2$ of nonparametric regressions of the motor cycle impact data (Silverman, 1985) by some continuous associated kernels.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>0.7320</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.7591</td>
</tr>
<tr>
<td>Rec. Inv. Gaussian</td>
<td>0.7328</td>
</tr>
<tr>
<td>Gaussian</td>
<td>0.7090</td>
</tr>
</tbody>
</table>

Figure 3: Nonparametric regressions of the motors cycle impact data (Silverman, 1985) by some continuous associated kernels.

Bayesian global for binomial kernel

Using Bayes theorem, the joint posterior distribution of $h$ given the observations is

$$
\pi(h|X_1, X_2, \ldots, X_n) \propto h^{n-1}(1-h)^{\beta-1} \left( \frac{1}{2} \sum_{i=1}^{n} \left( y_i - \hat{m}_{-i}(X_i) \right)^2 + b \right)^{-\frac{(a+2\beta)}{2}},
$$

where $\propto$ denotes proportional, the reals $a$ and $b$ are the parameters of the inverse gamma distribution $IG(a, b)$, and $\alpha$ and $\beta$ those of the beta distribution $Be(\alpha, \beta)$. The estimate $\hat{h}_{bay}$ of the smoothing parameter $h$ is given by Markov chain Monte Carlo (MCMC) techniques with Gibbs sampling:

$$
\hat{h}_{bay} = \frac{1}{N - N_0} \sum_{t=N_0+1}^{N} h(t),
$$

where $N_0$ is the burn-in period and $N$ the number of iterations; see Zougab et al. (2014b) for further details. It will be implemented in a future version of the Ake package.

Summary and final remarks

The Ake package offers easy tools for R users whose research involves kernel estimation of density functions and/or regression functions through associated kernels that are capable of handling all categorical, count and real positive datasets. Figure 1 shows the importance of the associated kernel choice as well as the bandwidth selection. In fact, symmetric (e.g., Gaussian) kernel estimators (respectively empirical estimators) are not suitable for bounded or positive continuous datasets (respectively discrete small samples). We then need an appropriate associated kernel. The binomial kernel is suitable for small size count data while the discrete triangular or the naive kernel are more indicated for large sample sizes. In continuous cases, the lognormal and gamma kernels give the best
estimation for positive data while the extended beta is suitable for any compact support.

This package includes various continuous and discrete associated kernels. It also contains functions to handle the bandwidth selection problems through cross-validation, local and global Bayesian procedures for binomial kernel and also the adaptive Bayesian procedure for the gamma kernel. In general, Bayesian choices of smoothing parameters will be better than their cross-validation counterparts. Future versions of the package will contain Bayesian methods with other associated kernels. Also, these associated kernels are useful for heavy tailed data p.d.f. estimation and can be added later in the package; see, e.g., Ziane et al. (2015). The case of multivariate data needs to be taken in consideration; see Kokonendji and Somé (2015) for p.d.f. estimation and Somé and Kokonendji (2016) for regression. We think that the Ake package can be of interest to nonparametric practitioners of different applied settings.

Acknowledgments

We sincerely thank the CRAN editors and two anonymous reviewers for their valuable comments. Part of this work was done while the first author was at “Laboratoire de Mathématiques de Besançon” as a Visiting Scientist, with the support of “Agence Universitaire de la Francophonie” (AUF).

Bibliography


Wanbitching E. Wansouwé
The University of Maroua
Higher Teachers' Training College
Department of Computer Science
P.O. Box 55 Maroua, Cameroon
ericwansouw@gmail.com

Sobom M. Somé
Université Bourgogne Franche-Comté
Laboratoire de Mathématiques de Besançon
16 route de Gray, 25030 Besançon cedex, France
sobom.some@univ-fcomte.fr

Célestin C. Kokonendji
Université Bourgogne Franche-Comté
Laboratoire de Mathématiques de Besançon
16 route de Gray, 25030 Besançon cedex, France
celestin.kokonendji@univ-fcomte.fr