Nonparametric estimation for probability mass function with Disake

an R package for discrete associated kernel estimators

W. E. Wansouwé* — C. C. Kokonendji** — D. T. Kolyang*

* Department of Computer Science
The University of Maroua
P.O. box 55 MAROUA
CAMEROON
ericwansouwe@gmail.com, dtaiwe@yahoo.fr

** Laboratoire de Mathématiques de Besançon-UMR 6623 CNRS-UFC
Université de Franche-Comté
16 route de Gray, 25030 BESANÇON Cedex
FRANCE
celestin.kokonendji@univ-fcomte.fr

RÉSUMÉ. La méthode des noyaux est l'une des techniques d'estimation les plus répandues en statistique non paramétrique. Nous introduisons un module en R, **Disake**, pour l'estimation d'une distribution de probabilité par noyaux associés discrets. Dans l'estimation par noyau, deux choix importants sont à faire : le choix du noyau et celui de la fenêtre de lissage. Le module **Disake** implémente principalement les noyaux associés discrets ainsi que la validation croisée et l'approche bayésienne locale pour la sélection du paramètre de lissage. Des applications sur des données simulées et réelles montrent que le noyau binomial est approprié pour les échantillons de petite ou moyenne taille et, l'estimateur naïf ou le noyau triangulaire discret, indiqué pour les échantillons de grande taille.

ABSTRACT. Kernel smoothing is one of the most widely used nonparametric data smoothing techniques. We introduce a new R package, **Disake**, for computing discrete associated kernel estimators for probability mass function. When working with a kernel estimator, two choices must be made: the kernel function and the smoothing parameter. The **Disake** package focuses on discrete associated kernels and also on cross-validation and local Bayesian techniques to select the appropriate bandwidth. Applications on simulated data and real data show that the binomial kernel is appropriate for small or moderate count data while the empirical estimator or the discrete triangular kernel is indicated for large samples.

MOTS-CLÉS : Module R, noyau associé discret standard, validation croisée.

KEYWORDS : R package, standard discrete associated kernel, cross-validation.

1. Introduction

Nonparametric density estimation is one of the most researched and still active areas in statistical theory, and the techniques and the theory are highly sophisticated. A lot of developments in statistics have taken place around the themes, methods, and mathematics of density estimation. Density estimation has experienced a wide explosion of interest over the last 20 years. The problem of density estimation can be used to obtain information on symmetry or multimodality of the sample law. It may also be used in the problem of estimation of the failure rate function and the estimation in the regression model. The well-known and popular technique of nonparametric density estimation is the kernel density estimator ; see [19] and also [17].

Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with an unknown probability density function (p.d.f.) f on $\mathbb{R}$. A continuous kernel estimator $\widehat{f}_n$ of f can be defined in the two following ways :

$$\widehat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \quad (1)$$

$$= \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i), x \in \mathbb{R}, \quad (2)$$

where $K(.)$ is the continuous kernel function which is typically a bona-fide p.d.f. with zero mean and unit variance, $h = h(n) > 0$ is an arbitrary sequence of smoothing parameters (or bandwidths) that fulfills $\lim_{n\to\infty} h(n) = 0$, and $K_{x,h}(.)$ will be the "continuous associated kernel" with the target x and the bandwidth h. Following the well-known expression (1) for unbounded supports of f, $K(.)$ is classically symmetric and, therefore, the associated kernel is written as :

$$K_{x,h}(.) = \frac{1}{h}K\left(\frac{x - .}{h}\right). \quad (3)$$

But the way from (2) to (1) is not always possible like for asymmetric associated kernels with respect to the target x. In the expressions (1) and (2), the bandwidth plays the role of a dispersion parameter around the target ; this can be easily illustrated through the symmetric Gaussian associated kernel $N_{x,h}$ with mean x (the target) and standard deviation h (the bandwidth) where $K = N_{0,1}$; see [22]. One can refer to [19] and [17] where the expression (1) has been mentioned for the first time. For recent references, one can see also [27]. The works usually cited [6], [21] and [24] concern some generalities on (supposed) continuous data. For functional data, one can refer to [7]. The contributions [25] and [26] are concerned with ordered categorical and discrete data *always* using the continuous kernels. The second expression (2), that we will use in this paper, is for adapting a "type of continuous kernel" generally asymmetric (such beta and gamma) to

the support of f ; see [4, 5]. For inverse Gaussian and reciprocal inverse Gaussian kernels, see [13] and [20] respectively. The case of a bounded support (from two or one end) of f to estimate induces a choice of type of asymmetric kernel, while the symmetric continuous kernel K does not have any important proper effects and can be used indifferently for smoothing functions on unbounded supports.

In order to estimate a probability mass function (p.m.f.) on $\mathbb{T}$ (*e.g.* $\mathbb{N}$ + $p\mathbb{N}$ for $p \geq 0$, $\mathbb{Z}^d$, $\{0, 1, \ldots, N\}^d$, for $d \in \mathbb{N}\backslash\{0\}$) using a discrete kernel method, the empirical or naive estimator is often used because of its good asymptotic properties. However, this Dirac type kernel estimator is not appropriate with small samples sizes. Furthermore, its great default is that it does not take into account observations around the target because its bandwidth is null or does not exist. Except for the naive estimator, the authors [2] have been the pioneers of discrete kernel estimators in the sense of (2). But the discrete kernel used has a unique shape and is appropriate for categorical data and finite discrete distributions. Thus, the case of discrete kernels for count data is not investigated in the sense of [10], except a first attempt of [16]. That attempt is only experimental and applied on univariate count data (*i.e.* $\mathbb{T} = \mathbb{N}$). A necessity of a discrete smoothing using discrete kernels out of the Dirac kernel is illustrated in Figures 1 and 5 ; for example, a binomial discrete kernel estimator is more interesting than the empirical estimator for a small sample size. A discrete associated kernel which asymptotically tends to Dirac type kernel has been recently defined and built. It results in many applications of the discrete associated kernel method in literature such that nonparametric estimations of discrete weighted function [9] and regression count function [12].

When working with a kernel estimator of the density function two choices must be made : the kernel function K and the smoothing parameter or bandwidth h. In practice, the selection of K can be easily adapted to the support of the density to be estimated ; but the choice of an efficient method for the calculation 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 undersmoothed estimator, with high variability. On the contrary, if the value of h is large, the resulting estimator will be oversmoothed.

Some methods have been investigated for selecting bandwidth parameter but the commonly used is the least squares cross-validation. A Bayesian approach has been also recently introduced [29] in the case of binomial kernel. Despite the great number of packages implemented for nonparametric estimation in continuous cases, to the best of our knowledge, the R packages to estimate p.m.f. have been far less investigated.

For the above reasons, we have implemented the package **Disake** for which a new version is available since january 2015 [28]. This package completes the papers cited previously which practically show the usefulness of the discrete associated kernel approach. The aim of this paper is to describe the package, and also to summarize and conveniently present discrete associated kernels recently introduced for nonparametric estimation. The rest of this paper is as follows. Section 2 presents the definition of a discrete associated

kernel estimator ; Section 3 details the bandwidth and the kernel selection procedures ; Section 4 explains the implemented functions of the package and shows illustrations with simulated and real data. This is followed by a comparison among discrete associated kernels estimators ; Section 5 is devoted to conclusion.

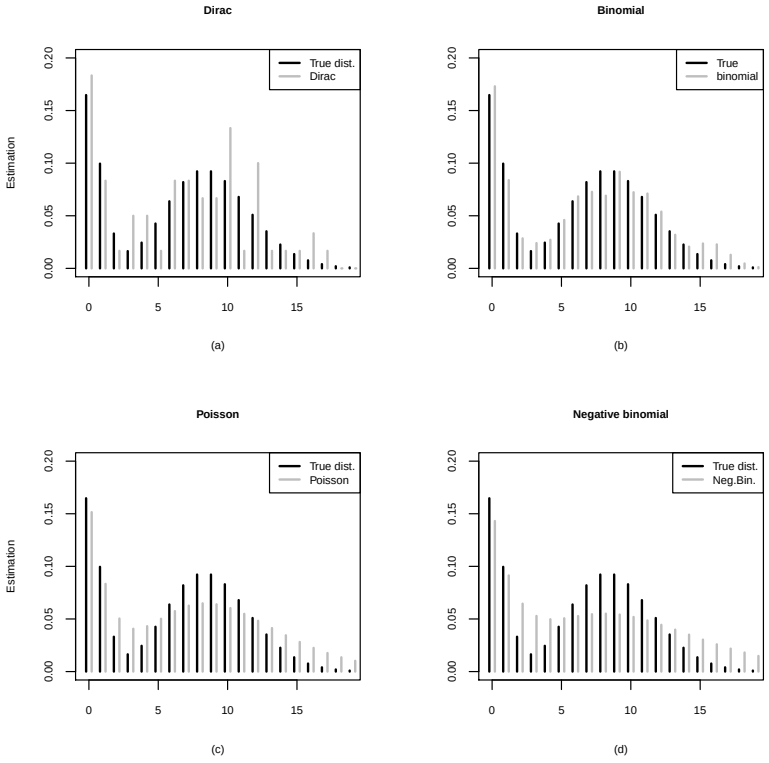

Figure 1 – Discrete smoothing of simulated data of $f = 0.3\mathcal{P}(0.6) + 0.7\mathcal{P}(9)$ with $n = 60$ using the standard discrete associated kernels : Dirac (a), binomial (b), Poisson (c) and negative binomial (d). Except for Dirac, bandwidth parameter is obtained by cross-validation.

2. Discrete associated kernels

In order to simplify, we assume that the support $\mathbb{T}$ of the p.m.f. f, to be estimated, is the count set $\mathbb{N} := \{0, 1, \ldots\}$. We then consider, for $\mathbb{T} = \mathbb{N}$, the topology inherited from the standard one of the real line number $\mathbb{R}$.

2.1. Definitions and properties

The discrete associated kernel introduced in (2) is defined as follows :

Definition 2.1. *Let $\mathbb{T}$ be the discrete support of the p.m.f. f, to be estimated, x a fixed target in $\mathbb{T}$ and $h > 0$ a bandwidth. A p.m.f. $K_{x,h}(.)$ on support $\mathbb{S}_x$ (not depending on h) is said to be an associated kernel, if it satisfies the following conditions :*

$$x \in \mathbb{S}_x, \tag{4}$$

$$\lim_{h \to 0} \mathbb{E}(\mathcal{K}_{x,h}) = x, \tag{5}$$

$$\lim_{h \to 0} \mathbb{V}ar(\mathcal{K}_{x,h}) = 0, \tag{6}$$

where $\mathcal{K}_{x,h}$ is the discrete random variable whose p.m.f. is $K_{x,h}(.)$.

In order to construct a discrete associated kernel $K_{x,h}(.)$ from a parametric discrete probability distribution K_θ, $\theta \in \Theta \subset \mathbb{R}^d$ on the support $\mathbb{S}_\theta$ such that $\mathbb{S}_\theta \cap \mathbb{T} \neq \emptyset$, we need to establish a correspondence between $(x, h) \in \mathbb{T} \times (0, \infty)$ and $\theta \in \Theta$ [11]. In what follows, we will call $K \equiv K_\theta$ the *type of discrete kernel* to make a difference from the classical notion of continuous kernel (1). In this context, the choice of the discrete associated kernel becomes important as well as that of the bandwidth. Moreover, we distinguish the discrete associated kernels said sometimes of "second order" of those said of "first order" which verify the two first conditions (4) and (5).

Discrete associated kernel estimator

Let us give the first properties of the estimator (2) with a discrete associated kernel.

Proposition 2.2. *(Proposition 1 of [11]). Let $X_1, X_2, \ldots, X_n$ be an n random sample i.i.d. from the unknown p.m.f. f on $\mathbb{T}$. Let $\widehat{f}_n = \widehat{f}_{n,h,K}$ be an estimator (2) of f with a discrete associated kernel. Then, for all $x \in \mathbb{T}$ and $h > 0$, we have*

$$\mathbb{E}\{\widehat{f}(x)\} = \mathbb{E}\{f(\mathcal{K}_{x,h})\},$$

where $\mathcal{K}_{x,h}$ is the random variable associated to the p.m.f. $K_{x,h}$ on $\mathbb{S}_x$. Furthermore, we have $\widehat{f}_n(x) \in [0, 1]$ for all $x \in \mathbb{T}$ and

$$\sum_{x \in \mathbb{T}} \widehat{f}_n(x) = C,$$

where $C = C(n; h, K)$ *is a positive and finite constant if* $\sum_{x\in\mathbb{T}} K_{x,h}(y) < \infty$ *for all* $y \in \mathbb{T}$.

Notice that $C = 1$ for the estimators (2) of Dirac discrete uniform kernel. In general we have $C \neq 1$ for the estimators (2), as with discrete triangular associated kernels and standard discrete kernels, but is always near 1. In practice, we calculate the constant C depending on observations before normalizing $\widehat{f}_n$ to be a p.m.f. Without loss of generality, from now we assume $C = 1$.

Pointwise consistency and asymptotic normality

In the following section, we point out the results for the asymptotic behavior of the mean squared error (MSE) of $\widehat{f}_n(x)$ and the global consistency of $\widehat{f}_n(x)$ in the sense of the mean integrated squared error (MISE). The first consistency result concerns the MSE of $\widehat{f}_n(x)$.

Theorem 1. *(Theorem 1 of [11]). Under assumptions* (4) - (6)*, for any fixed* $x \in \mathbb{T}$*, one has*

$$\lim_{n\to 0} \mathbb{E}\{\widehat{f}_n(x) - f(x)\}^2 = 0.$$

Theorem 2. *(Theorem 2.4 of [1]). Under assumptions* (4) - (6)*, for any fixed* $x \in \mathbb{T}$*, we have*

$$\widehat{f}_n(x) \xrightarrow{a.s.} f(x) \qquad as \quad n \to \infty,$$

where $\xrightarrow{a.s.}$ *denotes almost sure convergence.*

Global consistency

For the global consistency, the criterion to use is the MISE of $\widehat{f}_n = \widehat{f}_{n,h,K}$ defined as

$$MISE(n, h, K, f) = \sum_{x\in\mathbb{T}} Var(\widehat{f}_n(x)) + \sum_{x\in\mathbb{T}} bias^2(\widehat{f}_n(x)).$$

Theorem 3. *(Theorem 4 of [11]). Let* f *be a p.m.f. on* $\mathbb{T}$ *with* $\lim_{x\to\infty} f(x) = 0$*. Then the estimator* (2) $\widehat{f}_n = \widehat{f}_{n,h,K}$ *of* f *with any discrete associate kernel is such that, for* $n \to \infty$ *and* $h = h(n) \to 0$*, we have the behavior*

$$MISE(n, h, K, f) = \frac{1}{n}\sum_{x\in\mathbb{T}} f(x)[\{Pr(\mathcal{K}_{x,h} = x)\}^2 - f(x)]$$
$$+ \sum_{x\in\mathbb{T}} [f\{\mathbb{E}(\mathcal{K}_{x,h})\} - f(x) + \frac{1}{2}Var(\mathcal{K}_{x,h})f^{(2)}(x))]^2 + o(\frac{1}{n} + h^2),$$

where $f^{(2)}$ is the finite difference of second order given by

$$f^{(2)}(x) = \begin{cases} \{f(x+2) - 2f(x) + f(x-2)\}/4 & \text{if} \quad x \in \mathbb{N} \setminus \{0,1\} \\ \{f(3) - 3f(1) + 2f(0)\}/4 & \text{if} \quad x = 1 \\ \{f(2) - 2f(1) + f(0)\}/2 & \text{if} \quad x = 0. \end{cases}$$

Application : In the very particular case of the Dirac type kernel estimator and unbiased $\widehat{f}_{n,0,D}$, the MISE is equal to the integrated variance

$$MISE(n,0,D,f) = \frac{1}{n}\sum_{x\in\mathbb{T}} f(x)\{1-f(x)\} = \frac{1}{n}\left\{1 - \sum_{x\in\mathbb{T}} f^2(x)\right\}.$$

This exact result is used as reference in comparaison to the MISE of the others discrete associated kernel estimators, because $0 \leq \sum_{x\in\mathbb{T}} f^2(x) < 1$ and therefore we have the global consistency of the naive estimator as $MISE(n,0,D,f) \to 0$ when $n \to 0$.

2.2. Examples of discrete associated kernels

In this section, we examine the case of the so-called *standard discrete kernels*, represented in Figure 3, which are discrete associated kernels of the first order (i.e. not verifying the condition (6) in the Definition 2.1.) and two non standard discrete associated kernels. The three first kernels are built from usual discrete probability distributions of binomial, Poisson and negative binomial (see [8]). They are also useful for smoothing a p.m.f f on $\mathbb{T} = \mathbb{N}$ or distributions of count data with small samples. For all $x \in \mathbb{N}$ and $h > 0$, the discrete random variable $\mathcal{K}_{x,h}$ of standard discrete kernels satisfies, among others, the condition

$$\lim_{h\to 0} \mathrm{Var}(\mathcal{K}_{x,h}) \in \mathcal{V}(0), \tag{7}$$

where $\mathcal{V}(0)$ is a neighborhood of 0 which does not depend on x.

Binomial kernel : If we consider a binomial distribution $\mathcal{B}(N,p)$ with $N \in \mathbb{N}\backslash\{0\}$ and $p \in (0,1]$, we associate the random variable $\mathcal{B}_{x,h}$ corresponding to the binomial kernel $B_{x,h}$ following the distribution $\mathcal{B}\{x+1,(x+h)/(x+1)\}$ on $\mathbb{S}_x = \{0,1,\ldots,x+1\}$ for any $x \in \mathbb{N}$ and $h \in (0,1]$:

$$B_{x,h}(y) = \frac{(x+1)!}{y!(x+1-y)!}\left(\frac{x+h}{x+1}\right)^y \left(\frac{1-h}{x+1}\right)^{x+1-y} \mathbb{1}_{\mathbb{S}x}(y),$$

where $\mathbb{1}_{\mathbb{A}}$ denotes the indicator function of any given event A. It is an underdispersed discrete kernel (i.e. $\mathrm{Var}(\mathcal{B}_{x,h}) = (x+h)(1-h)/(x+1)$ smaller than $\mathbb{E}(\mathcal{B}_{x,h}) = x+h$) having its mode around $x+h$. The binomial kernel satisfies the three assumptions (4), (5) and (7) with $\nu(0) = [0,1)$. The bias and the variance of the corresponding estimator (2), for any $x \in \mathbb{N}$, are written as :

$$\mathrm{bias}\{\widehat{f}_n(x)\} = f(x)\{B_{x,h}(x) - 1\} + \sum_{y\in\mathbb{S}_x\setminus\{x\}} f(y)B_{x,h}(y)$$

and

$$n\,\mathrm{Var}(\widehat{f}_n(x)) = f(x)B^2_{x,h}(x) + \sum_{y\in\mathbb{S}_x\setminus\{x\}} f(y)B^2_{x,h}(y) - \left[f(x) + \sum_{y\in\mathbb{S}_x}\{f(y)-f(x)\}B_{x,h}(y)\right]^2.$$

The MISE of this estimator is not consistent but can be more smaller than those of estimators with discrete associated kernels and Dirac type kernel for some sample sizes not so large as shown in Figure 2.

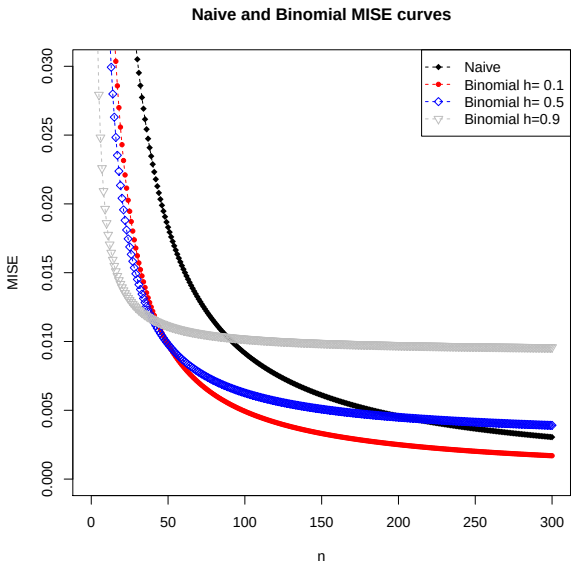

Figure 2 – Naive and binomial MISE for $f = 0.3\mathcal{P}(0.6) + 0.7\mathcal{P}(9)$ and different values of h.

Poisson kernel : Consider a Poisson distribution $P(\lambda)$ with $\lambda > 0$. For any x fixed in $\mathbb{T} = \mathbb{N}$ and $h > 0$, the corresponding random variable $\mathcal{P}_{x,h}$ associated to the Poisson kernel $P_{x,h}$ follows the distribution $P(x+h)$ with support $\mathbb{S}_x = \mathbb{N}$ and p.m.f.

$$P_{x,h}(y) = \frac{(x+h)e^{-(x+h)}}{y!}\mathbb{1}_{\mathbb{S}x}(y).$$

Note that the discrete kernel proposed by [16] inverts x and y in the expression $P_{x,h}(y)$ above and does not allow any mathematical study of properties. Our Poisson kernel $P_{x,h}$ is equidispersed (i.e. $\mathbb{E}(\mathcal{P}_{x,h}) = \mathrm{Var}(\mathcal{P}_{x,h}) = x+h$) and has its mode between $x+h-1$ and $x+h$. Then, the discrete kernel $P_{x,h}$ fulfills assumptions (4) and (5) except (7). The corresponding estimator $\widehat{f}_n(x)$ has the pointwise bias

$$\mathrm{bias}\,\{\widehat{f}_n(x)\} = f(x)\{P_{x,h}(x) - 1\} + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)P_{x,h}(y)$$

which does not tend to 0 when $h \to 0$. Its pointwise variance can be written as

$$n\mathrm{Var}(\widehat{f}_n(x)) = f(x)P^2_{x,h}(x) + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)P^2_{x,h}(y) - \left[f(x) + \sum_{y\in\mathbb{N}} \{f(y) - f(x)\}P_{x,h}(y)\right]^2.$$

This $\widehat{f}_n$ is not consistent in the sense of small MISE but can be more interesting than the naive estimator, for small or moderate sample sizes.

Negative binomial kernel : In the case of the negative binomial distribution $\mathcal{BN}(\lambda, p)$ with $\lambda > 0$ and $p > 0$, we define the negative binomial kernel $\mathcal{BN}_{x,h}$ with the random variable $BN_{x,h}$ following the distribution $BN\{x+1, (x+1)/(2x+1+h)\}$ on $S_x = \mathbb{N}$ for any $x \in \mathbb{N}$ and $h > 0$:

$$BN_{x,h}(y) = \frac{(x+y)!}{x!y!}\left(\frac{x+h}{2x+1+h}\right)^y \left(\frac{x+1}{2x+1+h}\right)^{x+1} \mathbb{1}_{\mathbb{S}x}(y).$$

It is an overdispersed discrete kernel (i.e. $Var(\mathcal{BN}_{x,h}) = (x+h)\{1+(x+h)/(x+1)\}$ greater than $\mathbb{E}(\mathcal{BN}_{x,h}) = x+h$ having its mode around $x+h$. This discrete kernel satisfies assumptions (4),(5) but not (7). For an estimator (2) with a negative binomial kernel, the pointwise bias is given as :

$$\mathrm{bias}\,\{\widehat{f}_n(x)\} = f(x)\{BN_{x,h}(x) - 1\} + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)BN_{x,h}(y)$$

and the pointwise variance can be written as

$$n\mathrm{Var}\{\widehat{f}_n(x)\} = f(x)BN^2_{x,h}(x) + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)BN^2_{x,h}(y) - \left[f(x) + \sum_{y\in\mathbb{N}} \{f(y) - f(x)BN_{x,h}(y)\}\right]^2.$$

Similarly to the previous cases, this estimator $\widehat{f}_n$ is not consistent in the sense of small MISE ; but, it can be more interesting than the naive estimator for some small samples

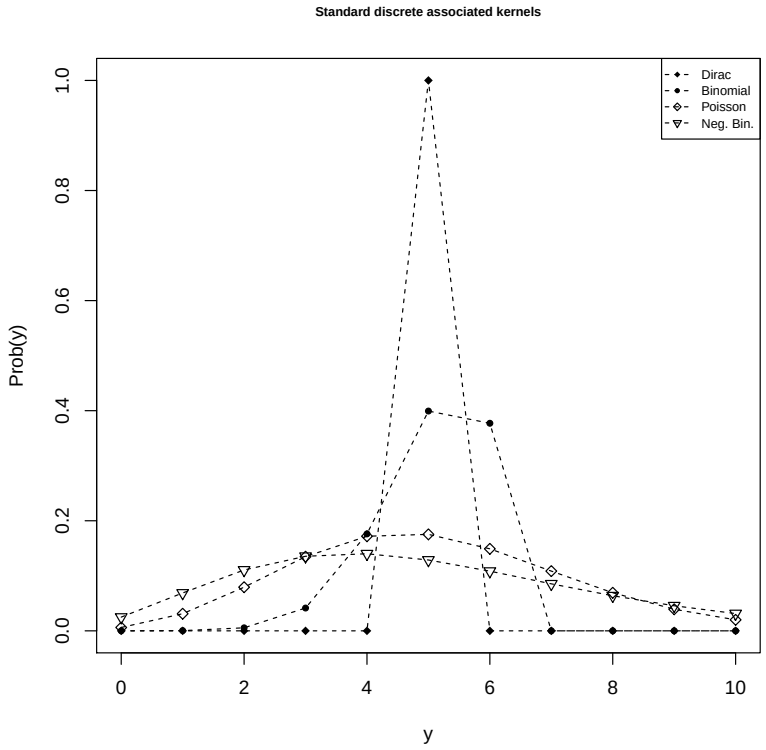

Figure 3 – Shapes of standard discrete associated kernels of type Dirac $\mathcal{D}(x)$, Poisson $\mathcal{P}(x+h)$, binomial $\mathcal{B}\{x+1,(x+h)/(x+1)\}$ and negative binomial $\mathcal{BN}\{x+1,(x+1)/(2x+1+h)\}$ for $x=5$ and $h=0.1$.

sizes.

Discrete triangular kernel : The following class of symmetric discrete kernels has been proposed in [10]. It generalizes the classical triangular kernel and might be constructed as follows. First, the support $\mathbb{T}$ 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 any x fixed in $\mathbb{T}$, consider the random variable $\mathcal{T}_{a;x,h}$ defined on $\mathbb{S}_x=\{x,x\pm1,\ldots,x\pm a\}$ and whose p.m.f. is given by

$$K_{x,h}(y)=\frac{(a+1)^h-|y-x|^h}{P(a,h)}\mathbb{1}_{\mathbb{S}_x}(y),$$

where $P(a,h) = (2a+1)(a+1)^h - 2\sum_{k=0}^{a} k^h$ is the normalizing constant. Since $K_{x,h}$ is symmetric around x, assumptions (4) and (5) are satisfed. A package for the asymmetric version of discrete triangular kernel is also available [23]. As for the variance term (6),

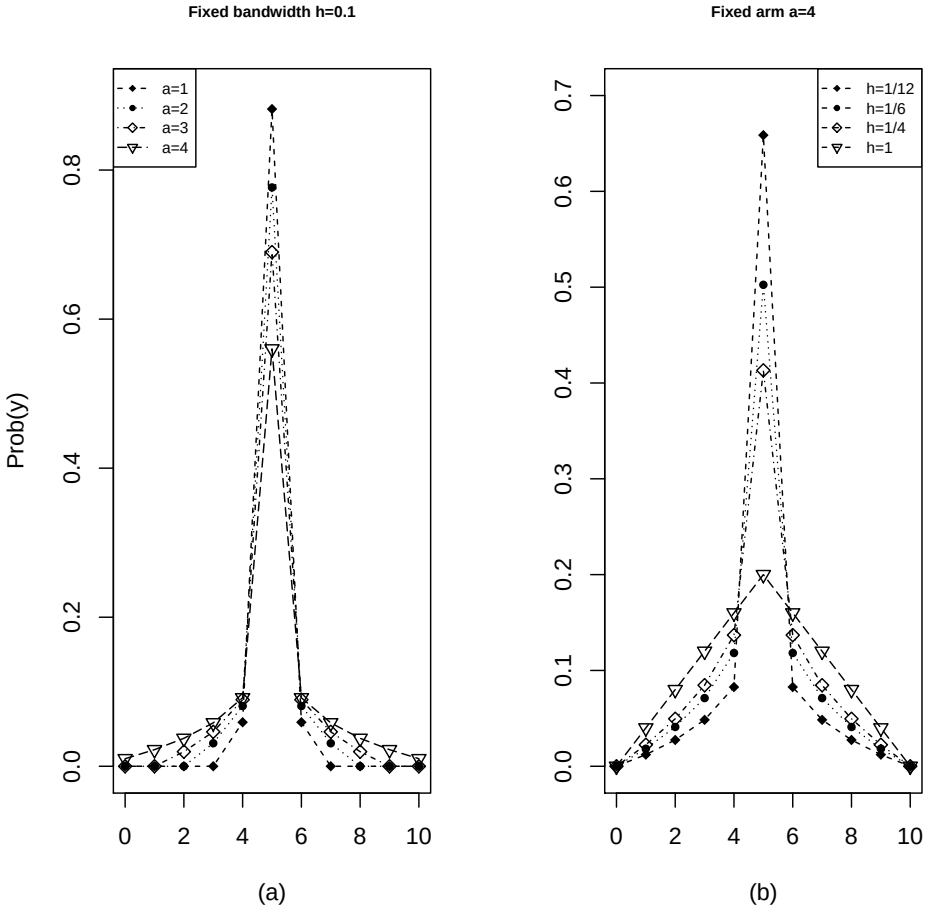

Figure 4 – Shapes of discrete triangular kernel with same target $x = 5$: (a) Different values of the arm a for fixed bandwidth $h = 0.1$; (b) Different values of the bandwidth h for fixed arm $a = 4$.

note that, for $a \in \mathbb{N}$ fixed, one has

$$V(a,h) = \frac{1}{P(a,h)} \left\{ \frac{a(2a+1)(a+1)^{h+1}}{3} - 2\sum_{k=0}^{a} k^{h+2} \right\}$$

$$\simeq \left\{ \frac{a(2a^2+3a+1)}{3} \log(a+1) - 2\sum_{k=1}^{a} k^2 \log(k) \right\} h + O(h^2),$$

which does not depend on $x = \mathbb{E}(\mathcal{T}_{a;x,h})$ and tends to 0 when $h \to 0$. The last approximation holds for h sufficiently small. The bias and the variance of the corresponding estimator (2) are written as :

$$\text{bias}\,\{\widehat{f}_n(x)\} = f(x) \left\{ \frac{(a+1)^h}{P(a,h)} - 1 \right\} + \sum_{y \in \mathbb{S}_x \setminus \{x\}} f(y) \Pr(\mathcal{T}_{a;x,h} = y)$$

which tends to 0 when $h \to 0$, and

$$\begin{aligned} n\text{Var}(\widehat{f}_n(x)) \quad &= \left[f(x) \left\{ \frac{(a+1)^h}{P(a,h)} \right\}^2 + \sum_{y \in \mathbb{S}_x \setminus \{x\}} f(y) \{\Pr(\mathcal{T}_{a;x,h} = y)\} \right] \\ &\quad - \left[f(x) + \sum_{y \in \mathbb{S}_x} \{f(y) - f(x)\} \Pr(\mathcal{T}_{a;x,h} = y) \right]^2 \end{aligned}$$

which tends to the variance $n^{-1} f(x)\{1 - f(x)\}$ of the naive estimator when $h \to 0$. The estimator (2) $\widehat{f}_n$ with discrete triangular kernels is consistent in the sense of MISE. Figure 4 shows the effect of the arm a and the bandwidth h of the discrete triangular distribution. The MISE curves can also be seen in Figure 5 for $a = 1$. This small value of the arm a can be considered as the worst case ; the discrete triangular in this case performs like the Dirac kernel as very few points around the target are involved in computing the estimation at the given point x.

Dirac discrete uniform kernel : A discrete kernel estimator for categorical or finite data has been introduced in [2]. Its asymmetric discrete associated kernel that we here label DiracDU (Dirac discrete uniform) has been deduced in [11] as follows. For fixed $c \in \{2, 3, \ldots\}$ the number of categories, $\mathbb{S}_c = \{0, 1, \ldots, c-1\}$ and the discrete kernel in (2) might be

$$K_{x,h}(y) = (1-h)\mathbb{1}_x(y) + \frac{h}{c-1}\mathbb{1}_{\mathbb{S}_c \setminus \{x\}}(y),$$

where h belongs to $(0, 1]$ and $x \in \mathbb{S}_c$. In addition, the target x can be considered as the reference point of X and the smoothing parameter h is such that $1 - h$ is the success

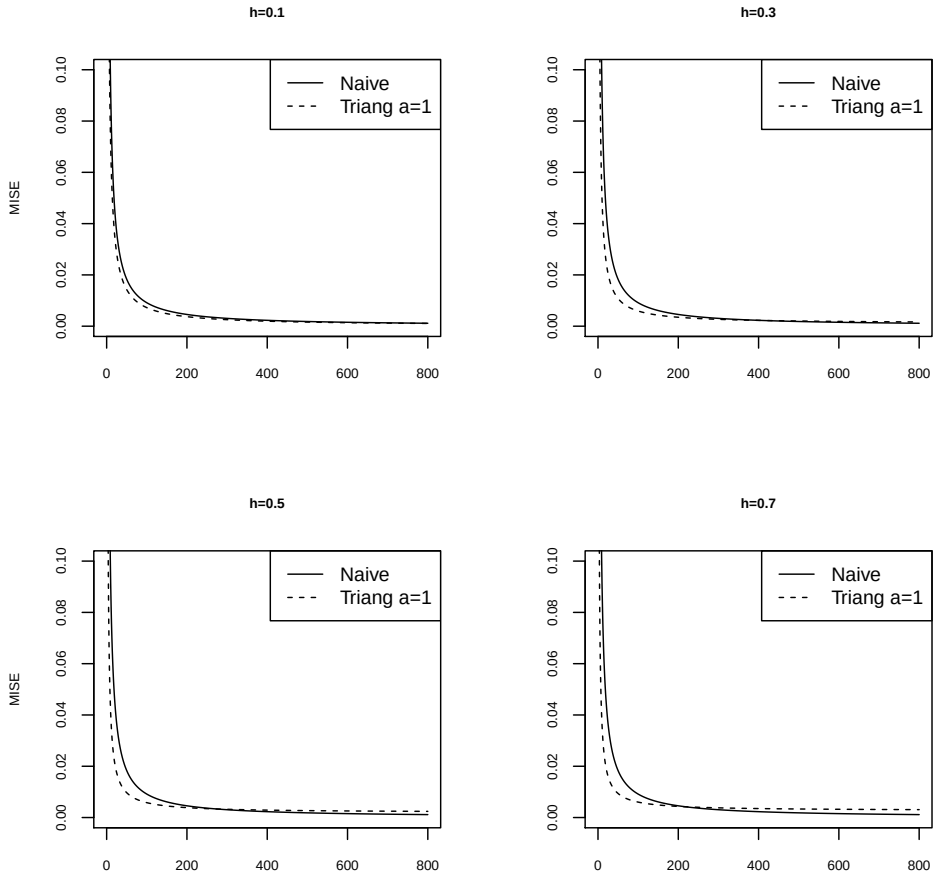

Figure 5 – Mise curves of the discrete triangular kernel with different values of h for a=1.

probability of the reference point. Finally, if the bandwidth h goes to 0, then, the random variable $\mathcal{A}_{c;x,h}$ associated to $K_{x,h}$ will satisfy (4), (5), (6). Its mean and variance are such that

$$\mathbb{E}(\mathcal{A}_{c;x,h}) = x + h\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right),$$

$$Var(\mathcal{A}_{c;x,h}) = -\left\{\frac{c^2(-2x+c-1)^2}{4(c-1)^2}\right\}h^2 + \left\{\frac{c(6x^2+2c^2-3c+1-6xc+6x)}{6(c-1)}\right\}h.$$

The bias and the variance of the corresponding estimator (2) are written as :

$$\text{bias}\,\{\widehat{f}_n(x)\} = \frac{-hc}{c-1}f(x) + \frac{h}{c-1}\sum_{i=0}^{c-1} f(i),$$

which also tends to 0 when $h \to 0$, and

$$\begin{aligned} n\text{Var}(\widehat{f}_n(x)) &= \left[f(x)(1-h)^2 + \frac{h^2}{(c-1)^2}\left\{\sum_{i=0}^{c-1} f(i) - f(x)\right\}\right] \\ &\quad - \left[f(x)(1-h) + \frac{h}{(c-1)}\left\{\sum_{i=0}^{c-1} f(i) - f(x)\right\}\right]^2 \end{aligned}$$

which also tends to the variance $n^{-1}f(x)\{1-f(x)\}$ of the naive estimator when $h \to 0$.

3. Kernel choice and bandwidth selection

In this section, we will consider the kernel choice and the bandwidth selection problems which occur generally in nonparametric estimation.

3.1. Kernel choice

The ideal discrete kernel satisfies

$$K_{id} = \arg\min_{\text{K}} \text{MISE}(n,h,K,f) = h_{id}(n,K,f).$$

Since the discrete (associated) kernel $K_{x,h}$ depends on the support $\mathbb{T}$ of f and also on each target $x \in \mathbb{T}$, we have to restrict us to a specific class of discrete kernels for realizing the optimization.

Thus, without loss of generality, we consider two random variables $\mathcal{K}_{x,h}^{[1]}$ and $\mathcal{K}_{x,h}^{[2]}$ connecting to discrete associated kernels (of first or second order) $\mathcal{K}_{x,h}^{[1]}$ and $\mathcal{K}_{x,h}^{[2]}$ on comparable supports $\mathbb{S}_x^{[1]}$ and $\mathbb{S}_x^{[2]}$ respectively. Up to

$$\mathbb{E}(\mathcal{K}_{x,h}^{[1]}) = \mathbb{E}(\mathcal{K}_{x,h}^{[2]}), \qquad \forall x \in \mathbb{T} \text{ and } h > 0,$$

the discrete kernel $K^{[1]}$ is said to be better than the discrete kernel $K^{[2]}$ if and only if

$$\text{Var}(\mathcal{K}_{x,h}^{[1]}) \leq \text{Var}(\mathcal{K}_{x,h}^{[2]}).$$

After all, for $h > 0$ and a p.m.f.f, the choice of a discrete kernel depends on the sample size n. For n large, the Dirac type kernel or a discrete associated kernel like discrete triangular kernel will be sufficient to get a good discrete estimating. Concerning small sample sizes for which the Dirac type kernel is not appropriate, the use of a discrete kernel of first order or a discrete associated kernel is more interesting.

3.2. Bandwidth selection methods

Similarly, the bandwidth selection is generally realized in the sense of MISE by approaching the ideal value of the bandwidth defined as

$$h_{id} = \arg\min_{\mathrm{h}>0} \mathrm{MISE}(n, h, K, f) = h_{id}(n, K, f).$$

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, [3], [15] and references therein. We simply propose two procedures for the bandwidth selection without making here a study on their consistencies : cross-validation and Bayesian procedures.

Cross-validation technique

For a given discrete kernel $K_{x,h}$ with $x \in \mathbb{T}$ and $h > 0$, we can prove that the optimal bandwidth h_{cv} of h is obtained by cross-validation as

$$h_{cv} = \arg\min_{\mathrm{h}>0} CV(h),$$

where

$$\begin{aligned} CV(h) &= \sum_{x\in\mathbb{T}} \{\widehat{f}_n(x)\}^2 - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{n,-i}(X_i) \\ &= \sum_{x\in\mathbb{T}} \left\{\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\right\}^2 - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} K_{X_i,h}(X_j) \end{aligned}$$

with $\widehat{f}_{n,-i}(y) = (n-1)^{-1}\sum_{j\neq i} K_{y,h}(X_j)$ being computed as $\widehat{f}_n(y)$ by excluding the observation X_i. This method is applied to all the estimators (2) with discrete kernels cited in this paper, independently on the support $\mathbb{T}$ of f to be estimated.

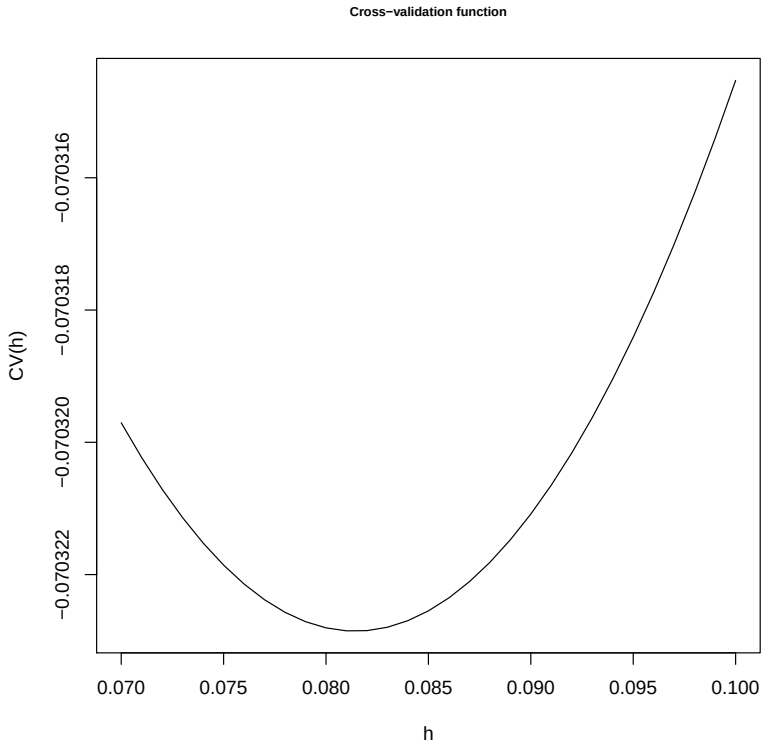

Figure 6 – Cross-validation function curve.

Bayesian approach

The Bayesian approach for bandwidth selection when using binomial kernel has been introduced in [29] where the authors investigated three different procedures : global, local and adaptive. 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 $\pi(.)$. Under squared error loss function, the Bayesian bandwidth selector is the posterior mean [29].

First, as we have mentioned above, $f(x)$ can be approximated by

$$f(x|h) = f_h(x) = \sum_{y \in \mathbb{T}} f(y) B_{x,h}(y) = \mathbb{E}\{B_{x,h}(\Upsilon)\},$$

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

Indeed, let $\pi(h)$ denotes the beta prior density of h. By the Bayes theorem, the posterior of h at the point of estimation x takes the form

$$\pi(h|x) = \frac{f_h(x)\pi(x)}{\int f_h(x)\pi(h)dh}.$$

Since f_h is unknown, we use $\widehat{f_h}$ as natural estimator of f_h, and hence we can estimate the posterior by

$$\pi(h|x, X_1, X_2, \ldots, X_n) = \frac{\widehat{f_h}(x)\pi(x)}{\int \widehat{f_h}(x)\pi(h)dh}.$$

Under the squared error loss, the Bayes estimator of the smoothing parameter $h(x)$ is the posterior mean and is given by

$$\widehat{h}_n(x) = \int h\widehat{\pi}(h|x, X_1, X_2, \ldots, X_n)dh.$$

4. The package

In this section, we will illustrate the package **Disake** through the functions that we implemented.

4.1. Implemented functions

The R package **Disake** contains eight functions. The functions dak, $pmfe$, $bwcv$ are implemented in the first version of the package while kf, $kpmfe$, $CVbw$, $Baysbw$, $sumry$ are in the second version. In Table 1 we can find a description of some functions.

Tableau 1 – Summary of contents of the package

Function	Description
dak	Discrete associate kernel function for standard kernels
$pmfe$	Probability mass function estimator when using standard kernels
$CVbw$	Cross-validation function
$Baysbw$	Bayesian procedure to select the bandwidth when using binomial kernel
$sumry$	Summary function of the results of computations.

The function dak corresponds to the discrete associated kernel function. Three options are allowed for the kernel function : "bino" for binomial, "pois" for Poisson and "nebi"

for the negative binomial kernel. As for dak, kf computes the discrete associated kernel function for DiracDU ("dirdu"), discrete triangular ("triang") and binomial ("bino") kernels. The following code helps to plot graphics as shown in Figure 3.

```
R > x<-4
R > h<-0.1
R > y<-0:10
R > k_d<-dak(x,y,h,"bino")
R > plot(y,k_d,pch=18,xlab="y",ylab="Prob(y)")
R > lines(y,k_d, pch=18,lty=2)
```

The functions $pmfe$ and $kpmfe$ compute the p.m.f. estimator. As for dak and kf functions, three options are also available in each function for the kernel. The function $bwcv$ as well as the $CVbw$ computes the bandwidth using the cross-validation method. The same possibilities occur here for the kernel : "pois", "nebi", "bino", "triang" and "dirdu".

When using binomial kernel, two functions are available to select the appropriate bandwidth : $CVbw$ and $Baysbw$. The last function $Baysbw$ computes the bandwidth through the local Bayesian procedure. In the binomial case, one needs to precise the type of bandwidth selection : "CV" for cross-validation technique and "bays" for Bayesian procedure. The function $sumry$ gives a summary of the results of computation given a sample and a kernel. It returns the value of the bandwidth parameter h_n, the normalizing constant C_n, the values of the empirical distribution f_0 at each point of the support $\mathbb{S}_x$ and the values of estimated p.m.f. f_n after normalization.

Notice that one can also use his own bandwidth in the $sumry$ function.

4.2. Illustrations

In this section, we illustrate some previous results. We first use simulated data from a mixture of two Poisson distributions with respective means $\mu_1 = 0.6$ and $\mu_2 = 9$; we will also use real data from Algerian football Championship and national Cup [29].

Simulated data

In the following example, we illustrate some functions of the package using a sample of 60 simulated data from a mixture of two Poisson distributions with respective means $\mu_1 = 0.6$ and $\mu_2 = 9$ with proportions 0.3 and 0.7 respectively :

$$f(x) = 0.3\frac{e^{-0.6}0.6^x}{x!} + 0.7\frac{e^{-9}9^x}{x!}, \quad x \in \mathbb{N}.$$

This probability mass function f defined on $\mathbb{N}$ has a bimodality with the maximum at $x = 0$ ($f(0) = 0.164$), a local minimum at $x = 3$ ($f(3) = 0.016$), a local maximum at $x = 8$ ($f(8) = 0.092$), and a tail from $x = 24$.

The plot shown in Figure 1 reveals that the empirical estimator is not appropriate for small or moderate data because it does not smooth and estimate well. Contrary to the naive estimator, the Poisson and negative binomial kernels are smoothing well, but do not give a good approximation of the function at each point of the support ; while estimating the value of f at x, these kernels take into account many points and even where there is no observation in the sample. The binomial kernel is the one which not only smooths well but also gives a good estimation as shown in Figure 1.

All computations were done by using the R software [18]. The intersection points as shown in Figure 2 between the MISE curves of the discrete binomial kernel estimator and the MISE of the naive estimator point out the superior limit of n (depending on the distribution function and the bandwidth) for which this estimator is more efficient than the naive. Beyond this limit, the naive estimator is better and its MISE tends to 0. Among estimators (2) with standard discrete kernels, the binomial one is more interesting than the other kernels. Its MISE curve does not converge but it is the appropriate kernel for small samples. Figures 5 and 2 show that for large sample, we need to use the naive estimator or a discrete triangular kernel. One can see in Figure 1 that for $(x > 23)$, estimators using Poisson and negative binomial continue to give an estimation of the distribution at these points even if there is no observation.

Since, there is no procedure in R to simulate data from a mixture of Poisson, we wrote a simple function to perform it. The function $rsimpois(n, \lambda_1, \lambda_2, p, 1-p)$ creates data of size n from a mixture of two Poisson distributions with means λ_1, λ_2 and proportions p, $1-p$ respectively.

The following code computes the cross-validation bandwidth, using the binomial associated kernel for a sample of 60 simulated data from a mixture of two Poisson distributions with respective means $\mu_1 = 0.6$ and $\mu_2 = 9$.

```
R > Vec <- rsimpois(60,0.6,9,0.3,0.7)
R > bws  <- seq(0,1,by=0.001)
R > CV <- bwcv(Vec,bws,"bino")
R > CV$CV_bw
 [1] 0.081
```

and one can plot the cross-validation function shown in Figure 6 using the following code :

```
R > plot(bws,CV$CV,type="l")
```

Real data

Some nonparametric estimation f of the distribution of count data from the Algerian football championship have been realized (Table 2) in [29] .

The authors did not compare the three standard discrete kernels to smooth this distribution but investigated another approach to select the bandwidth for binomial kernel only. We will use the three standard discrete kernel estimators and then compare them. Figure 7 shows the discrete smoothing of that distribution of real data. Once more, it points out

that the binomial kernel with bandwidth obtained by local Bayesian procedure is better the other kernels.

Tableau 2 – Number of goals per player in the Algerian football competitions with sample size $n = 69$ (season 2009 - 2010) [29].

Goals	0	1	2	3	4	5	6	7	8	9	10	11
Players	1	2	3	5	5	4	4	3	4	5	4	4
Goals	12	13	14	15	16	17	18	19	20	21	22	23
Players	3	2	2	2	2	3	3	3	2	1	1	1

For this specific data, we also realized that the cross validation does not give a good result (the curve is not very sympathetic). It has the minimum very closed to 1. Thus, it seems that the estimated value of the distribution at each point x is the empirical value of the distribution at $x + 1$. Instead of using cross-validation to select the bandwidth, [29] introduced a local Bayesian approach where the bandwidth h is treated as a random variable. It can be seen in Figure 7 that this method is better than the cross-validation technique. Smoothing through this discrete triangular kernel does respect the multimodality of the distribution ; but the estimating of the function at each point of the support is not good. Smoothing these data through Poisson and negative binomial kernels does not respect the muldimodality of the distribution. Thus, we will not get all the information about the distribution if we fit the data with one of these last three kernels estimators.

5. Conclusion

The **Disake** package completes the study of discrete kernels estimators introduced for nonparametric estimation of probability mass function. Figures 1 and 7 show the importance of the kernel choice as well as the bandwidth selection. In fact, there is a kernel effect : for small samples, the empirical estimator is not appropriate to estimate the probability mass function. We, then, need a discrete associated kernel like binomial, Poisson or negative binomial. But theoretical studies and simulations show that binomial kernel is the only one which smooths and estimates well. Poisson and negative binomial kernels can smooth well but do not estimate as well as the binomial kernel. In practice, the binomial kernel is the appropriate kernel for small size count data. For large samples size, the naive kernel or a discrete triangular is more indicated [11].

The new version of **Disake** package [28], includes discrete triangular and DiracDU kernels. It also contains functions to handle the bandwidth selection problems through cross-validation and local Bayesian procedures. We plan also to extend **Disake** to regres-

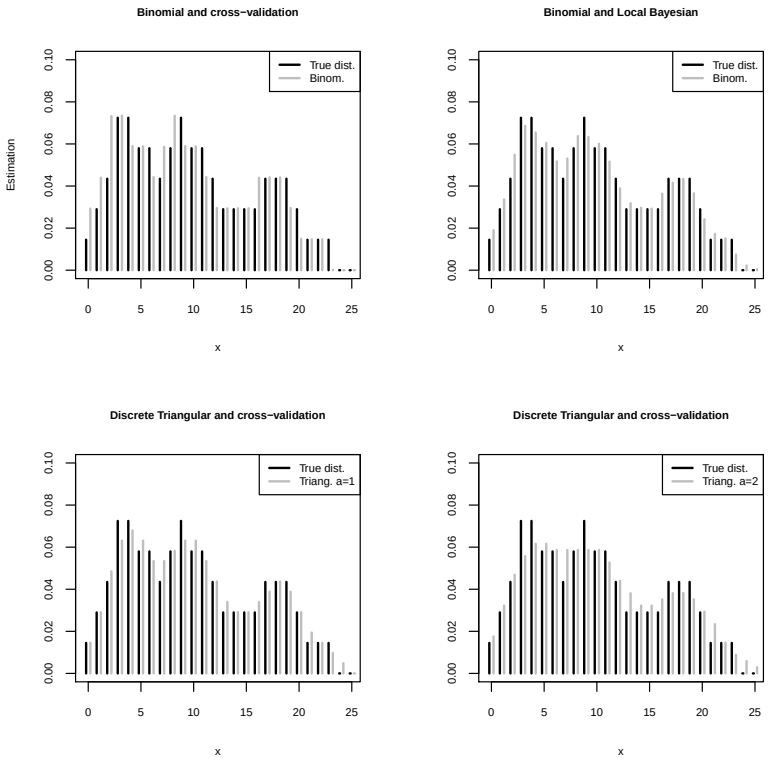

Figure 7 – Smoothing of Number of goals per player in the Algerian football competitions with sample size $n = 69$ (season 2009 - 2010) of Table 2 using standard discrete associated kernels.

sion estimation. Thus, we will have a complete overview of discrete associated kernel estimators in the case of univariate data. The case of multivariate data needs also to be taken in consideration. We think that **Disake** package can be of interest to nonparametric practitioners of different scientific fields.

Acknowledgments

We sincerely thank 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 The University of Maroua.

6. Bibliographie

[1] ABDOUS, B. & KOKONENDJI, C.C. « Consistency and asymptotic normality for discrete associated-kernel estimator », *African Diaspora Journal of Mathematics*, vol. 8, n° 2, pp. 63 - 70, 2009.

[2] AITCHISON, J. & AITKEN, C.G.G. « Multivariate binary discrimination by the kernel method », *Biometrika*, vol. 63, n° 3, pp. 413 - 420, 2000.

[3] BOWMAN, A. « An alternative method of cross-validation for the smoothing of density estimates », *Biometrika*, vol. 71, n° 2, pp. 353 - 360, 1984.

[4] CHEN, S.X. « Beta kernels estimators for density functions », *Computational Statistics & Data Analysis*, vol. 31, n° 3, pp. 131 - 145, 1999.

[5] CHEN, S.X. « Gamma kernel estimators for density functions », *Annals of the Institute of Statistical Mathematics*, vol. 52, n° 3, pp. 471 - 480, 2000.

[6] DEVROYE, L. « A Course in Density Estimation », *Birkhäuser, Boston.*, 1987.

[7] FERRATY, F. & VIEU, P. « Nonparametric Functional Data Analysis : Theory and Practice », *Springer, Berlin.*, 2006.

[8] JOHNSON, N.L., KEMP, A.W. & KOTZ, S. « Univariate Discrete Distributions », *3rd ed., John Wiley & Sons, Hoboken, NewJersey*, 2005.

[9] KOKONENDJI, C.C., MIZÈRE, D. & BALAKRISHNAN, N. « Connections of the Poisson weight function to overdispersion and underdispersion », *Journal of Statistical Planning and Inference*, vol. 138, n° 5, pp. 1287 - 1296, 2008.

[10] KOKONENDJI, C.C., SENGA KIESSÉ, T. & ZOCCHI, S.S. « Discrete triangular distributions and non-parametric estimation for probability mass function », *Journal of Nonparametric Statistics*, vol. 19, n° 6-8, pp. 241 - 254, 2007.

[11] KOKONENDJI, C.C. & SENGA KIESSÉ, T. « Discrete associated kernel method and extensions », *Statistical Methodology*, vol. 8, n° 6, pp. 497 - 516, 2011.

[12] KOKONENDJI, C.C., SENGA KIESSÉ, T.& DEMÉTRIO, C.G.B. « Appropriate kernel regression on a count explanatory variable and applications », *Advances and Applications in Statistics*, vol. 12, n° 1, pp. 99 - 126, 2009.

[13] LIBENGUÉ, F.G. « Méthode Non-paramétrique par Noyaux Associés Mixtes et Applications », *Unpublished Ph.D. Thesis*(in French), University of Franche-Comté Besançon, France University of Ouagadougou, Burkina Faso, 2013.

[14] MALEC, P. & SCHIENLE, M. « Nonparametric kernel density estimation near the boundary », *Computational Statistics and Data Analysis* vol. 72, pp. 57 - 76, 2014.

[15] MARRON, J.S. « A comparison of cross-validation techniques in density estimation », *The Annals of Statistics*, vol. 15, n° 1, pp. 152 - 162, 1987.

[16] MARSH, L.C. & MUKHOPADHYAY, K. « Discrete Poisson kernel density estimation with an application to wildcat coal strikes », *Applied Economics Letters*, vol. 6, n° 6, pp. 393 - 396, 1999.

[17] PARZEN, E. « On estimation of a probability density function and mode », *Annals of Mathematical Statistics*, vol. 33, n° 3, pp. 1065 - 1076, 1962.

[18] R DEVELOPMENT CORE TEAM « R : A Language and Environment for Statistical Computing, R Foundation for Statistical Computing », *Vienna, Austria, ISBN 3-900051-07-0, URL http ://www.R-project.org*, 2008.

[19] ROSENBLATT, M. « Remarks on some nonparametric estimates of a density function », *Annals of Mathematical Statistics*, vol. 27, n° 13, pp. 832 - 837, 1956.

[20] SCAILLET, O. « Density estimation using inverse and reciprocal inverse Gaussian kernels », *Journal of Nonparametric Statistics*, vol. 16, n° 1-2, pp. 217 - 226, 2004.

[21] SCOTT, D.W. « Multivariate Density Estimation : Theory, Practice, and Visualization », *Wiley, New York.*, 1992.

[22] SENGA KIESSÉ, T. « Nonparametric Approach by Discrete Associated-Kernel for Count Data », *Ph.D. manuscript*, University of Pau, URL http ://tel.archives-ouvertes.fr/tel-00372180/fr/ (in French), 2008.

[23] SENGA KIESSÉ, T., LIBENGUÉ, F.G., ZOCCHI, S.S. & KOKONENDJI, C.C. « The R package for general discrete triangular distributions. ». R package Version 1.0, URL `http://www.CRAN.R-project.org/package=TRIANGG` , 2010.

[24] SILVERMAN, B.W. « Density Estimation for Statistics and Data Analysis », *Chapman & Hall, London.*, 1986.

[25] SIMONOFF, J.S. « Smoothing Methods in Statistics », *Springer, New York*, 1996.

[26] SIMONOFF, J.S. & TUTZ, G. « Smoothing methods for discrete data », *In : M.G. Schimek (Ed.), Smoothing and Regression : Approaches, Computation, and Application. Wiley, New York*, pp. 193 - 228, 2000.

[27] TSYBAKOV, A.B. « Introduction à l'estimation non-paramétrique », *Springer, Paris*, 2004.

[28] WANSOUWÉ, W.E., KOKONENDJI,C.C. KOLYANG, D.T. « Disake : an R package for discrete associated kernel estimators ». R package Version 1.5, URL `http://www.CRAN.R-project.org/package=Disake` , 2015.

[29] ZOUGAB, N., ADJABI, S. & KOKONENDJI, C.C. « Binomial kernel and Bayes local bandwidth in discrete functions estimation », *Journal of Nonparametric Statistics*, vol. 24, n° 3, pp. 783 - 795, 2012.