Statistics

On bandwidth parameter choices for discrete nonparametric kernel estimator

Choix de fenêtres de lissage pour un estimateur non paramétrique à noyau discret

Tristan Senga Kissé

INRA, UMR 1069, Sol Agro et Hydrosystème SpatioLisation, 35000 Rennes, France

ABSTRACT

This note concentrates on the nonparametric estimation of a probability mass function (p.m.f.) using discrete associated kernels. An expression of the optimal bandwidth minimizing the asymptotic part of the global squared error is given. Some asymptotic expressions of bias and variance of the cross-validation criterion are also presented. At last, the two bandwidth selection procedures are illustrated through some simulations and an application on a real count data set.

© 2016 Académie des sciences. Published by Elsevier Masson SAS. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

RÉSUMÉ

Cette note se focalise sur l’estimation non paramétrique à noyau associé discret d’une fonction de masse de probabilité. Une expression de la fenêtre optimale minimisant la partie asymptotique de l’erreur quadratique globale est donnée. Des expressions asymptotiques pour le biais et la variance d’un critère de sélection par validation croisée sont également présentées. Enfin, les deux méthodes de choix de fenêtre sont illustrées par des simulations et une application sur des données réelles.

© 2016 Académie des sciences. Published by Elsevier Masson SAS. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Let \((X_i)_{i=1,2,\cdots,n}\) be a sample of i.i.d. discrete random variables (r.v.) having a p.m.f. \(f(x) = \Pr(X_i = x) > 0\) on support \(S\) (e.g., non-negative integers set \(\mathbb{N}\)). The discrete kernel estimator \(\hat{f}_n\) of \(f\) can be expressed as

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

\(1\)

E-mail address: tristan.sengakiesse@rennes.inra.fr.

http://dx.doi.org/10.1016/j.crma.2016.02.012

1631-073X/© 2016 Académie des sciences. Published by Elsevier Masson SAS. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
where $h = h(n) > 0$ is an arbitrary sequence of smoothing parameters that fulfills $\lim_{n \to \infty} h(n) = 0$ and $K_{s,h}(\cdot)$ is a p.m.f., called discrete associated kernel, with r.v. $K_{s,h}$ on support $S_x$ (containing target $x$ and not depending on $h$) satisfying

$$H_1 : \lim_{h \to 0} E(K_{s,h}(x)) = x, \quad \text{and} \quad H_2 : \lim_{h \to 0} \text{Var}(K_{s,h}(x)) = 0;$$

without loss of generality, we assume $S_x \subseteq S$ in all this work. The nonparametric kernel estimation has already largely discussed in the literature for the probability density function (p.d.f.), in particular for choosing smoothing parameters; see Bowman [2] or Marron [6]. In contrast, the discrete kernel estimation for p.m.f. has received more less attention than that for p.d.f. In fact, until now, it commonly results from the discretization of the continuous case by considering the ordinal variables as being continuous; see Atchison and Aitken [1], Titterington [8] or Simonoff and Tutz [7]. For a specific investigation of count data, Kokonendji et al. [4] and Kokonendji and Senga Kiessé [3] have developed the nonparametric kernel estimation using some discrete associated kernels. This opens way to a specific treatment of the bandwidth selections of discrete kernel estimators by adapting classical methods, which are well known for the continuous case. This note pursues the first works realized on this subject, with a first attempt to provide an expression of the optimal bandwidth and by investigating the asymptotic properties of the cross-validation criterion for $f_n$ in (1).

2. Discrete associated kernels

Let us first introduce the following hypotheses on modal probability of $K_{s,h}$:

$$H_1' : K_{s,h}(x) = 1 - hA(K_{s,h}) + O(h^2),$$

with $\sum_{y \in S_x \setminus \{x\}} K_{s,h}(y) = hA(K_{s,h}) + O(h^2) \to 0$ as $h \to 0$, where $A(K_{s,h})$ is bounded away from 0 for $h \to 0$. It results in the following proposition.

**Proposition 2.1.** Consider a fixed point $x \in S$ and the bandwidth $h > 0$. Under assumption $H_1'$, the expectation and variance of the discrete associated kernel $K_{s,h}$ fulfill $H_1$ and $H_2$.

**Proof.** For the assumption $H_1$, the expectation of $K_{s,h}$ comes directly from

$$\mathbb{E}(K_{s,h}) = xK_{s,h}(x) + \sum_{y \in S_x \setminus \{x\}} yK_{s,h}(y) = x + B(x; h),$$

with $B(x; h) = -xhA(K_{s,h}) + \sum_{y \in S_x \setminus \{x\}} yK_{s,h}(y) + O(h^2) \to 0$ for $h \to 0$. For the assumption $H_2$, the variance of $K_{s,h}$ can be successively expressed as

$$\text{Var}(K_{s,h}) = \sum_{y \in S_x} y^2K_{s,h}(y) - \left( \sum_{y \in S_x} yK_{s,h}(y) \right)^2 = x^2[K_{s,h}(x) - 1] + D(x; h),$$

with $D(x; h) = \sum_{y \in S_x \setminus \{x\}} y^2K_{s,h}(y) + x^2 - \left[ x + \sum_{y \in S_x \setminus \{x\}} yK_{s,h}(y) \right]^2 \to 0$ when $h \to 0$ under the hypotheses $H_1'$. Indeed, when $h \to 0$, we have both $K_{s,h}(y) \to 0$ for $y \neq x$ and $K_{s,h}(x) \to 1$ for $y = x$. \[\square\]

Then, the following assumption can be formulated on the variance of $K_{s,h}$:

$$H_2' : \text{Var}(K_{s,h}) = hV(K_{s,h}) + O(h^2),$$

leading to some hypotheses $H_1'$ and $H_2'$ less general than $H_1$ and $H_2$; $A(K_{s,h})$ and $V(K_{s,h})$ do not obligatory depend on $x$ and $h$, as in the next example.

**Example.** (See Kokonendji and Zocchi [5].) Let $a_1, a_2$ be fixed integers and $h_1, h_2$ be smoothing parameters. For any fixed $x \in S = \mathbb{Z}$, consider the r.v. $K_{a_1,a_2;x,h_1,h_2}$ of discrete generalized triangular associated kernels defined on supports $S_{a_1,x} = \{x-1, x-2, \ldots, x-a_1\}$ and $S_{a_2,x} = \{x+1, x+2, \ldots, x+a_2\}$ and whose p.m.f. is

$$K_{a_1,a_2;x,h_1,h_2}(y) = \frac{1}{P} \left\{ \begin{array}{c} 1 - \left( \frac{x-y}{a_1+1} \right)^{h_1} 1_{S_{a_1,x}}(y) + 1 - \left( \frac{y-x}{a_2+1} \right)^{h_2} 1_{S_{a_2,x}}(y) \end{array} \right\},$$

where $P \equiv (a_1 + a_2 + 1) - (a_1 + 1)^{-h_1} \sum_{k=1}^{a_1} k^{h_1} - (a_2 + 1)^{-h_2} \sum_{k=1}^{a_2} k^{h_2}$. $P(a_1, a_2, h_1, h_2)$ is the normalizing constant. For $h$ sufficiently small, one has

$$K_{a_1,a_2;x,h_1}(x) = 1 - \sum_{i=1}^{a_1} \left\{ h_1A(a_i) + O(h_1^2) \right\} \quad \text{and} \quad \text{Var}(K_{a_1,a_2;x,h_1}) = \sum_{i=1}^{a_1} \left\{ h_1V(a_i) + O(h_1^2) \right\},$$

with $A(a_i) = a_i \log(a_i + 1) - \sum_{k=1}^{a_i} \log(k)$ and $V(a_i) = a_i(2a_i^2 + 3a_i + 1) \log(a_i + 1)/6 - \sum_{k=1}^{a_i} k^2 \log(k)$. Hence, the assumptions $H_1'$ and $H_2'$ are fulfilled.
3. Nonparametric kernel estimator

This section presents an optimal $h$-value minimizing an approximate of the mean integrated squared error (MISE) of $\tilde{f}_n$ in (1), in comparison with the $h$-value minimizing cross-validation function (Kokonendji and Senga Kiessé [3]).

3.1. Minimization of MISE

The $\tilde{f}_n$’s variance and bias have been established by Kokonendji and Senga Kiessé [3] under $H1$ and $H2$; by taking into account $H1’$ and $H2’$, the global error is

$$MISE(h) = \sum_{x \in S} Var(\tilde{f}_n(x)) + \sum_{x \in S} Bias^2(\tilde{f}_n(x)) = AMISE(h) + o \left( \frac{1}{n} + h^2 \right) + O(h^2),$$

where

$$AMISE(h) = \frac{1}{n} \sum_{x \in S} f(x) \left[ (1 - h A(K_{x,h})) f(x) - f(x) \right] + \frac{1}{4} h^2 \sum_{x \in S} \{V(K_{x,h}) f^{(2)}(x)\}^2,$$

is the approximated MISE for $h$ sufficiently small, with $f^{(2)}$ a finite difference of second order. Hence, an approximate value $\hat{h}_{opt} = \arg\min_{h>0} AMISE(h)$ of the true optimal bandwidth $h_{opt} = \arg\min_{h>0} MISE(h)$ can be given by

$$\hat{h}_{opt}(n, f) = \frac{\sum_{x \in S} f(x) A(K_{x,h})}{\sum_{x \in S} f(x)(A(K_{x,h}))^2 + (n/4) \sum_{x \in S} [V(K_{x,h}) f^{(2)}(x)]^2},$$

which tends to 0 as $n \to \infty$ with $0 < \sum_{x \in S} [V(K_{x,h}) f^{(2)}(x)]^2 < \infty$. It results in an asymptotic relationship such that $\hat{h}_{opt} \sim k_0 n^{-1}$ with $k_0 = 4 \frac{\sum_{x \in S} f(x) A(K_{x,h})}{\sum_{x \in S} [V(K_{x,h}) f^{(2)}(x)]^2}$. Thus, the discrete generalized triangular associated kernel presented as an example has optimal bandwidth

$$\hat{h}_{opt}(n, a_i, f) = \frac{A(a_i)}{2(A(a_i))^2 + (n/2) [V(a_i)]^2 \sum_{x \in S} [f^{(2)}(x)]^2}, \quad i = 1, 2. \quad (2)$$

3.2. Minimization of cross-validation function

We are interested in a bandwidth $h_{cv}$ minimizing a cross-validation score function $CV$ with respect to $h$ such that $h_{cv} = \arg\min_{h>0} CV(h)$ with

$$CV(h) = \sum_{x \in S} \left\{ \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(X_i) \right\}^2 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{x,h}(X_i) \cdot \quad (3)$$

The CV’s mean and variance for fixed $h$ are provided in the following theorem.

**Theorem 3.1.** Consider a fixed point $x \in S$ and the bandwidth $h = h(n) > 0$ such that $\lim h = 0$. Assume that $CV$ is the cross-validation function for the nonparametric estimator in (1) of p.m.f. $f$. Then, we have

$$\mathbb{E}[CV(h)] = AMISE(h) - \sum_{x \in S} f^2(x) + O \left( \frac{1}{n} + h^2 \right)$$

and

$$\text{Var}(CV(h)) = \frac{1}{n} \sum_{x \in S} \left\{ \sum_{x \in S} K_{x,h}^2(x) - 2K_{x,h}(x) \right\}^2 f^3(x) + \frac{1}{n} \left\{ \sum_{x \in S} f^2(x) \right\}^2 + O \left( \frac{h^2}{n} + \frac{1}{n^2} \right),$$

where $K_{x,h}$ is a discrete kernel satisfying the assumption $H1’$.

**Proof.** Let us present the cross-validation score function in (3) as

$$CV(h) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{x \in S} K_{x,h}^2(X_i) + \frac{2}{n^2} \sum_{j < i} H_{ij},$$

with $H_{ij} = \sum_{x \in S} K_{x,h}(X_i) K_{x,h}(X_j) - 2K_{x,h}(X_j)$. One has $\mathbb{E}[CV(h)] = (1/n) \sum_{x \in S} \mathbb{E}[K_{x,h}^2(X_1)] + (1 - 1/n)\mathbb{E}(H_{ij})$. For the first term of $\mathbb{E}[CV(h)]$, we get:
\[
\sum_{x \in S} \mathbb{E}[K^2_{x,h}(X_1)] = \sum_{x \in S} K^2_{x,h}(x) f(x) + \sum_{x \in S} \sum_{y \in S \setminus \{x\}} K^2_{x,h}(y) f(y);
\]

where the second term in the previous equation is finite. Now let us express

\[
\mathbb{E}[K_{x_1,h}(X_2)] = \sum_{x \in S} \mathbb{E}[K_{x_1,h}(X_2)] f(x), \quad \mathbb{E}[K_{x_1,h}(X_1)] = \sum_{y \in S} K_{x_1,h}(y) f(y) = \mathbb{E}(f(K_{x_1,h})),
\]

and \( \mathbb{E}(f(K_{x_1,h})) = f(x) + (1/2) f^{(2)}(x) \text{Var}(K_{x_1,h}) + o(h) \) obtained by using a discrete Taylor expansion around \( x \). For the second term of \( \mathbb{E}(\text{CV}(h)) \), it ensues

\[
\mathbb{E}(H_{ij}) = \frac{h^2}{4} \sum_{x \in S} \left[ V(K_{x_1,h}) f^{(2)}(x) \right]^2 - \sum_{x \in S} f^2(x) + o(h^2) + O(h^2),
\]

where \( o(h^2) \) is asymptotically dominated by \( O(h^2) \). Hence, it results \( \mathbb{E}(\text{CV}(h)) \).

For \( \text{CV} \)'s variance, we begin by calculating the variance of the first term as

\[
\frac{1}{n^2} \text{Var} \left( \sum_{x \in S} K^2_{x,h}(X_1) \right) = \frac{1}{n^3} \left[ \left( \sum_{x \in S} K^2_{x,h}(x) \right)^2 f(x) - \left( \sum_{x \in S} K^2_{x,h}(x) f(x) \right)^2 \right] + Q_n(x; h),
\]

with

\[
n^3 Q_n(x; h) = \sum_{y \in S \setminus \{x\}} \left[ \{ \sum_{x \in S} K^2_{x,h}(y) \} f(y) + \left( \sum_{x \in S} K^2_{x,h}(x) f(y) \right)^2 - \left( \sum_{y \in S \setminus \{x\}} \sum_{x \in S} K^2_{x,h}(y) f(y) \right)^2 \right],
\]

where \( Q_n(x; h) = o(1/n^3) \); in fact, we can assume \( \text{Var}[\sum_{x \in S} \left( n^2 \sum_{i=1}^n K^2_{x,h}(X_i) \right)] = O(n^{-3}) \). Considering the variance of the second term of \( \text{CV} \), we have

\[
\frac{1}{n^2} \text{Var} \left( \sum_{j<i} H_{ij} \right) = \frac{1}{n^2} \text{Var}(H_{ij}) + \frac{1}{n} \left( 1 - \frac{3}{n} \right) \text{Cov}(H_{ij}, H_{ik}) + O \left( \frac{1}{n^3} \right).
\]

Without giving all details, we get \( \mathbb{E}(H_{ij}^2) = \sum_{x \in S} \left( \sum_{x \in S} K^2_{x,h}(x) - 2K_{x,h}(x) \right)^2 f^2(x) + o(h^2) \). Then, by using expression of \( \mathbb{E}(H_{ij}) \) calculated previously, we have

\[
\text{Var}(H_{ij}) = \sum_{x \in S} \left[ \left( \sum_{x \in S} K^2_{x,h}(x) - 2K_{x,h}(x) \right)^2 f^2(x) - \left( \sum_{x \in S} f^2(x) \right)^2 \right] + \frac{h^2}{2} \sum_{x \in S} \left[ V(K_{x_1,h}) f^{(2)}(x) \right]^2 \sum_{x \in S} f^2(x) + o(h^2).
\]

In addition, it can be shown that

\[
\mathbb{E}(H_{ij}H_{ik}) = \sum_{x \in S} \left[ \left( \sum_{x \in S} K^2_{x,h}(x) - 2K_{x,h}(x) \right)^2 f^3(x) + o(h^2) \right];
\]

then, by considering \( \text{Cov}(H_{ij}, H_{ik}) = \mathbb{E}(H_{ij}H_{ik}) - \mathbb{E}(H_{ij})\mathbb{E}(H_{ik}) \), we have

\[
\text{Cov}(H_{ij}, H_{ik}) = \sum_{x \in S} \left[ \left( \sum_{x \in S} K^2_{x,h}(x) - 2K_{x,h}(x) \right)^2 f^3(x) - \left( \sum_{x \in S} f^2(x) \right)^2 \right] + \frac{h^2}{2} \sum_{x \in S} \left[ V(K_{x_1,h}) f^{(2)}(x) \right]^2 \sum_{x \in S} f^2(x) + o(h^3) + O(h^2),
\]

with \( \text{Cov}(H_{ij}, H_{ik}) = 0 \). Hence the desired result on \( \text{Var}(\text{CV}(h)) \).

4. Illustrations

This section illustrates the bandwidth choices; discrete symmetric and generalized triangular kernels are applied for simulated and real data, respectively.
Table 1
Averages $ISE, \hat{h}_{opt}$ and $h_{cv}$ for nonparametric estimator using triangular kernels.

<table>
<thead>
<tr>
<th>Sample size $n$</th>
<th>$h_{opt}$</th>
<th>$ISE(h_{opt})$</th>
<th>$\hat{h}_{opt}$</th>
<th>$ISE(\hat{h}_{opt})$</th>
<th>$h_{cv}$</th>
<th>$TSE(h_{cv})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a = 2$</td>
<td>25</td>
<td>0.49</td>
<td>0.0028</td>
<td>0.23</td>
<td>0.0116</td>
<td>0.73</td>
</tr>
<tr>
<td>100</td>
<td>0.16</td>
<td>0.0011</td>
<td>0.12</td>
<td>0.0043</td>
<td>0.23</td>
<td>0.0053</td>
</tr>
<tr>
<td>$a = 3$</td>
<td>25</td>
<td>0.22</td>
<td>0.0034</td>
<td>0.10</td>
<td>0.0140</td>
<td>0.43</td>
</tr>
<tr>
<td>100</td>
<td>0.06</td>
<td>0.0009</td>
<td>0.04</td>
<td>0.0057</td>
<td>0.07</td>
<td>0.0062</td>
</tr>
<tr>
<td>$a = 4$</td>
<td>25</td>
<td>0.12</td>
<td>0.0033</td>
<td>0.05</td>
<td>0.0196</td>
<td>0.23</td>
</tr>
<tr>
<td>100</td>
<td>0.03</td>
<td>0.0006</td>
<td>0.02</td>
<td>0.0065</td>
<td>0.04</td>
<td>0.0068</td>
</tr>
</tbody>
</table>

Fig. 1. Estimations of the number of goals per match using a discrete generalized triangular kernel.

4.1. Simulations

Consider simulated data of sample size $n$ from a Poisson distribution $f$ with mean $\mu = 2$. The estimator $\hat{f}_n$ using symmetric triangular kernels is applied with $a_1 = a_2 = a \in \{2, 3, 4\}$, as an example. The performance of $\hat{f}_n$ is evaluated via the integrated squared error $ISE(h) = \sum_{x \in \mathbb{N}} (\hat{f}_n(x) - f(x))^2$ determined using $h_{opt}$, $h_{cv}$ and true $h_{opt}$ computed numerically since $f$ is known. In Table 1, the averages $TSE$, $\hat{h}_{opt}$ and $h_{cv}$ are calculated based on 100 replications of the simulated data. The value $\hat{h}_{opt}$ is competitive since globally we have $TSE(\hat{h}_{opt}) \leq TSE(h_{cv})$ for chosen values of $a$; note that $h_{opt}$ is underestimated by $\hat{h}_{opt}$ and overestimated by $h_{cv}$. At last, by estimating $f$ using empirical frequency of data, $TSE$ is equal to 0.0308 for $n = 25$ and 0.0088 for $n = 100$.

4.2. Application

The estimator $\hat{f}_n$ is applied using a generalized triangular kernel with $(a_1, a_2) = (1, 2)$ on count data, describing a number of goals per match from the French football championship (Kokonendji et al. [4]). By a cross-validation procedure, we have $h_{1cv} = 0.170$ and $h_{2cv} = 0.085$, while the expression (2) results in $\hat{h}_{opt,01} = 0.156$ and $\hat{h}_{opt,02} = 0.031$, obtained by replacing the unknown p.m.f. $f$ in (2) with its empirical frequency estimate. The quality of the estimate provided using $(h_{1cv}, h_{2cv})$ is close to that of the one calculated using $(\hat{h}_{opt,01}, \hat{h}_{opt,02})$, since $ISE$ is, respectively, equal to $3.0081 \times 10^{-4}$ and $3.0082 \times 10^{-4}$ (Fig. 1).

References