Discrete semiparametric regression models with associated kernel and applications

Tristan Senga Kiessé, Michaël Rivoire

Discrete semiparametric regression models with associated kernel and applications

T. Senga Kiessé[a]* and M. Rivoire[b,c]

[a] *Office National des Forêts, F-54840 Velaine-en-Haye, France;* [b] *INRA, Centre de Nancy, UMR 1092 Laboratoire d'Etude des Ressources Forêt-Bois, F-54280 Champenoux, France;* [c] *AgroParisTech, ENGREF, UMR 1092 Laboratoire d'Etude des Ressources Forêt-Bois, 14 rue Girardet, F-54000 Nancy, France*

This work is concerned with a semiparametric associated kernel estimator for count explanatory variables. The proposed semiparametric estimator is a multiplicative combination between a parametric model and a discrete nonparametric kernel estimator of Nadaraya–Watson type. In this semiparametric approach, the parametric model plays the role of the start function and the nonparametric kernel estimator is a correction factor of the parametric estimate. Some asymptotic properties of the discrete semiparametric kernel regression estimator are pointed out; in particular, we show its asymptotic normality and the order of the optimal bandwidth. The parametric part is illustrated by some nonlinear and generalised linear models; for the nonparametric estimator, we apply the discrete general triangular associated kernel providing bias reduction. The usefulness of the discrete semiparametric kernel regression estimator is shown on three practical examples in comparison with logistic, generalised linear and additive models.

Keywords: asymptotic normality; bandwidth optimal order; discrete regression; parametric regression model

AMS Subject Classification: 62G07; 62G08; 62F10

1. Introduction

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be the observations of the variables (X, Y) in $\mathcal{S} \times \mathbb{R}$ connected through the model

$$y_i = m(x_i) + e_i,$$

where $m : \mathcal{S} \mapsto \mathbb{R}$ is an unknown regression function to be estimated and e_i is assumed to be the residual from the real random variable ϵ_i with mean $\mathbb{E}(\epsilon_i) = 0$ and variance $\mathrm{Var}(\epsilon_i) = \sigma^2 < \infty$. To estimate the conditional mean function m, several approaches are available; one can consider the classical parametric regression models, some nonparametric techniques such as generalised additive models (GAMs, see Hastie and Tibshirani 1990) or local polynomials. Here, we are concerned with some multiplicative or additive procedures of nonparametric kernel and parametric

*Corresponding author. Email: tristan.senga-kiesse@onf.fr

regression models; for example, see Fan, Wu and Feng (2009) and Martins-Filho, Mishra and Ullah (2008) for some continuous models. Indeed, the nonparametric kernel estimator is known to be completely impartial to the special features of the function to be estimated; consequently, its combination with parametric models may lead to the improvement of the accuracy of the estimation. In this way, Abdous, Kokonendji and Senga Kiessé (2010) proposed a discrete semiparametric estimator, which is analogous to the continuous version presented by Glad (1998). The discrete semiparametric regression estimator uses the *discrete associated kernels* method introduced by Kokonendji, Senga Kiessé and Zocchi (2007) for estimating a function on a discrete support $\mathcal{S}$ as the non-negative integer set $\mathbb{N}$ (see Senga Kiessé 2009). Thus, we assume $\mathcal{S} = \mathbb{N}$ throughout this paper. In the discrete associated kernel methodology, the kernel $K_{x,h}(\cdot)$ is a probability mass function (p.m.f.) with support $\mathcal{S}_x$, which contains $x \in \mathbb{N}$ and does not depend on the smoothing parameter $h > 0$, such as

A1. $\lim_{h\to 0} \mathbb{E}(\mathcal{K}_{x,h}) = x$,
A2. $\lim_{h\to 0} \mathrm{Var}(\mathcal{K}_{x,h}) = 0$,

where $\mathcal{K}_{x,h}$ is the discrete random variable of p.m.f. $K_{x,h}(\cdot)$. In addition, the finite differences $g^{(k)}$, $k \in \mathbb{N} \setminus \{0\}$, of any count function $g : \mathbb{N} \to \mathbb{R}$ are used instead of the usual differentiation on $\mathbb{R}$ such as

$$g^{(k)}(x) = \{g^{(k-1)}(x)\}^{(1)} \text{ and } g^{(1)}(x) = \begin{cases} \{g(x+1) - g(x-1)\}/2 & \text{if } x \in \mathbb{N} \setminus \{0\} \\ g(1) - g(0) & \text{if } x = 0, \end{cases} \tag{1}$$

from which the finite difference of second order may be derived as

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

Within the semiparametric context, let us consider m as a discrete weighted parametric regression function given by

$$m(x) = l(x; \Theta) \times \omega(x), \quad x \in \mathbb{N}, \tag{3}$$

where $l(x; \Theta)$ is a nonrandom function relative to the parameter Θ and $x \mapsto \omega(x)$ is a positive nonparametric weight function. The discrete semiparametric associated kernel regression estimator results from a parametric estimation $\hat{l}(x) \equiv l(x; \hat{\Theta})$ of l multiplied by a nonparametric Nadaraya–Watson estimation $\hat{\omega}_n$ of ω as follows:

$$\hat{\omega}_n(x) = \sum_{i=1}^{n} \frac{Y_i K_{x,h}(X_i)}{\sum_{j=1}^{n} K_{x,h}(X_j)} \times \frac{1}{\hat{l}(X_i)}, \quad x \in \mathbb{N},$$

where $h = h(n) > 0$ is an arbitrary sequence of smoothing parameters that fulfils $\lim_{n\to\infty} h(n) = 0$ and $K_{x,h}(\cdot)$ is a suitably chosen discrete associated kernel function. Then, the discrete semiparametric estimator of m in Equation (3) is given by

$$\hat{m}_n(x) = \hat{l}(x) \times \hat{\omega}_n(x). \tag{4}$$

Concerning the parametric model, the smoothness of the function $l(x, t)$ with respect to t is required, and the estimator $\hat{\Theta}$ of Θ is obtained, for example, by the generalised least-squared

method. In the situation where the parametric function $l(x;\Theta)$ is mis-specified, the estimator $\hat{\Theta}$ of Θ converges in probability to a certain value Θ_0 such that $l(x;\Theta_0) \equiv l_0(x)$ is the best approximant to $m(x)$ with respect to the Kullback–Leibler distance of $l(x;\Theta)$ from the true function $m(x)$ as

$$\sum_{x\in\mathbb{N}} m(x)\log\frac{m(x)}{l(x;\Theta)} =: d\{m(\cdot), l(\cdot;\Theta)\};$$

see Abdous et al. (2010) and references therein for more details.

In this work, we establish the asymptotic normality of the discrete semiparametric kernel estimator. We use some parametric (logistic and generalised linear) models as start functions and a discrete associated kernel that provides bias reduction. The usefulness of the constructed discrete semiparametric regression model is illustrated on three practical data sets of agriculture, economy and agronomy in comparison with classical parametric regression models.

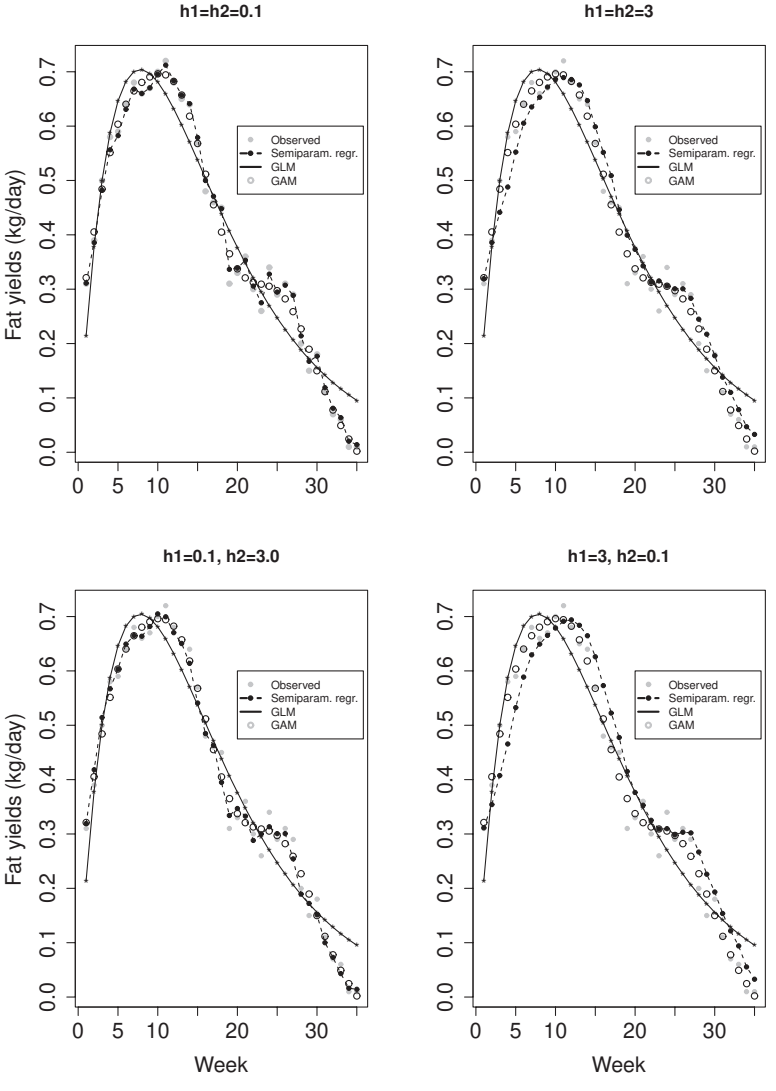

Figure 1. GLM (black curve), GAM (circles) and semiparametric regression using discrete general triangular associated kernels with $(a_1, a_2) = (3, 1)$ (black points) on average daily fat data.

The first example concerns the study of average daily fat (kg/day) yields from the milk of a single cow for each of the 35 first weeks denoted x_i (Kokonendji, Senga Kiessé and Demétrio 2009b). The quantity of fat in the milk increases during the first 14 weeks and decreases thereafter. The fitted curve comes from a generalised linear model (GLM): it is a normal model with a logarithmic link (McCullagh and Nelder 1989). This model does not fit well to data. In particular, it does not detect the plateau associated with observations $x = 19, 20, \ldots, 27$ (Figure 1). We will compare these results with those obtained by using our semiparametric model and GAM.

The second example given in Table 1 is a sales data set with multiple y_i at a given x_i (Kokonendji et al. 2009b). We analyse the amount of daily sales of a new product during the first 24 days. The 151 observations (x_i, y_i), $i = 1, \ldots, 24$, represent the day x_i and the corresponding mean of sales numbers $y_i \in \{y_{Ai}, y_{Bi}, \ldots, y_{Hi}\}$. The number of sale centres for each state (A, B, ..., H) is not available except for the state H, where this number is equal to one. We apply the GLM and GAM in comparison with the semiparametric model for fitting the sales data (Figure 2).

The third example deals with volume data from a forest beech tree (Table 2) provided by the French national research agency project 'EMERGE' (Compatible volume/biomass and nutrient content equations for fuelwood and forest resource; tools for sustainable and clear management); (Rivoire et al. 2010). On the stem of this tree, from the base (*ca.* 53 cm in diameter) to the tip (0 cm), 15 measures have been taken with a diameter tape. Cumulative stem volumes denoted y have been calculated to any possible diameter $x \in \{0, 1, \ldots, 53\}$ (cm) based on cone frustum volumes. More exactly, at the base of the tree, where the diameter is close to 53 cm, the cumulative volume is 0, whereas at the tip of the tree, the diameter is close to zero and the cumulative volume is the total stem volume. We apply the GAM, semiparametric model and parametric logistic one, since the tree data distribution has a sigmoïdal form (Figure 3).

The given examples indicate that the use of continuous semiparametric model of Glad (1998) may provide fitted values at continuous point as 0.3 or 1.2, while the predictor is an ordinal variable.

Table 1. Sales data.

x_i	y_{Ai}	y_{Bi}	y_{Ci}	y_{Di}	y_{Ei}	y_{Fi}	y_{Gi}	y_{Hi}
1	16.4	15.9	*5.0	16.1	16.2	15.4	16.2	16
2	21.4	18.9	22.2	19.7	17.2	20.7	19.9	23
3	22.0	19.7	24.2	20.5	18.1	22.4	21.1	20
4	20.4	19.2	23.0	19.8	17.6	21.7	20.5	23
5	18.2	18.1	20.4	18.4	16.8	19.8	∗8.8	24
6	16.1	16.6	17.7	16.6	15.7	17.6	∗6.5	12
7	14.2	15.0	15.2	14.8	14.5	15.3	∗4.0	13
8	12.6	13.5	13.0	13.0	13.3	13.3	∗1.8	9
9	11.1	12.0	11.1	11.5	12.1	11.5	9.9	9
10	9.9	10.6	9.5	10.0	10.9	9.9	8.5	8
11	8.7	9.4	8.2	8.8	9.8	8.6	7.5	10
12	7.8	8.3	7.1	7.7	8.7	7.4	6.8	8
13	6.9	7.3	6.2	6.7	7.8	6.5	6.3	7
14	–	6.4	5.4	5.9	6.9	5.6	5.9	2
15	5.4	5.6	4.7	–	6.1	4.9	5.6	*12
16	4.8	–	4.1	4.6	5.4	–	–	3
17	4.3	–	3.6	–	–	3.8	–	5
18	–	3.8	–	–	–	3.3	–	4
19	–	–	–	3.2	3.7	–	4.3	2
20	–	2.9	–	2.8	3.2	–	–	2
21	–	–	–	–	2.9	2.3	3.5	5
22	–	–	–	–	2.5	–	3.2	5
23	–	–	2.1	1.8	2.2	–	–	2
24	–	–	1.9	–	–	–	2.5	–

Notes: – denotes a missing observation and * can be considered as a strange value.

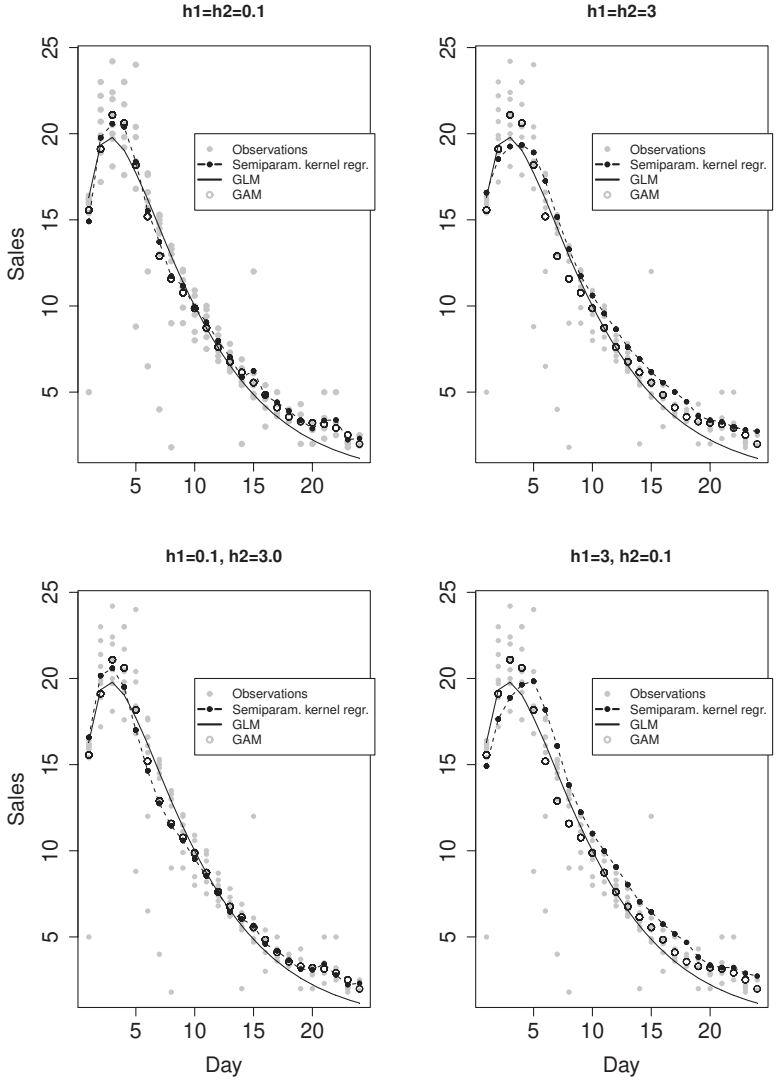

Figure 2. GLM (black curve), GAM (circles) and semiparametric regression using discrete general triangular associated kernels with $(a_1, a_2) = (3, 1)$ (black points) on sales data.

Table 2. Data of a beech tree.

x	0	1	2	3	4	5	6	7	8	9
y	1.51625	1.51621	1.51609	1.51575	1.51510	1.51384	1.51126	1.50838	1.50535	1.50195
x	10	11	12	13	14	15	16	17	18	19
y	1.49773	1.49436	1.48930	1.48450	1.47750	1.47750	1.47102	1.46363	1.45516	1.44807
x	20	21	22	23	24	25	26	27	28	29
y	1.43756	1.42577	1.39242	1.34545	1.33754	1.33329	1.32418	1.31423	1.30341	1.29166
x	30	31	32	33	34	35	36	37	38	39
y	1.27896	1.27225	1.25805	1.24281	1.22648	1.15572	1.04385	0.92528	0.80005	0.67887
x	40	41	42	43	44	45	46	47	48	49
y	0.53924	0.36795	0.27868	0.21124	0.19722	0.16811	0.13754	0.10548	0.07190	0.03674
x	50	51	52	53						
y	0	0	0	0						

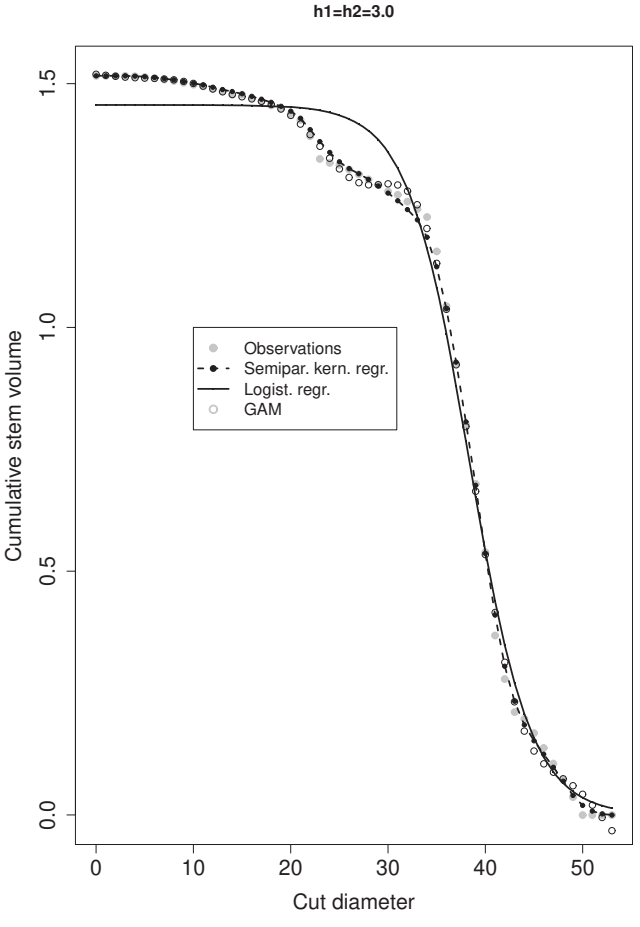

Figure 3. Logistic regression (black curve), GAM (circles) and semiparametric regression using discrete general triangular associated kernels with $(a_1, a_2) = (3, 1)$ (black points) on tree data.

This motivates the recommendation of a discrete model that focuses on ordinal covariates and has the same nature. Hence, the nonparametric correction in all the three examples is available only for discrete predictors even if the parametric models indeed treat predictors as continuous variables. Through these three applications, we point out that the discrete semiparametric associated kernel approach may produce better explanations of real data with both satisfying amounts of smoothing and goodness of fit.

The remainder of this paper is organised as follows. Section 2 is concerned with the bias, variance and asymptotic normality of the discrete semiparametric kernel regression estimator. Section 3 presents the result of the three applications. The optimal order of the bandwidth is shown under some assumptions for the discrete associated kernels used. Finally, Section 4 presents the concluding remarks.

2. Asymptotic properties

This section is concerned with the usual asymptotic results for the discrete semiparametric associated kernel regression estimator $\hat{m}_n$ in Equation (4). In particular, we demonstrate its

asymptotic normality; one can refer to Martins-Filho et al.(2008) for the asymptotic normality of the semiparametric estimator proposed by Glad (1998).

We state the bias and variance of $\hat{m}_n$ shown in Abdous et al.(2010). For $x \in \mathbb{N}$, let $l_0(x)$ be a fixed parametric start in Equation (3). Under assumptions A1 and A2, the discrete semiparametric estimator $\hat{m}_n$ in Equation (4) admits the following bias and variance:

$$\mathbb{E}\{\hat{m}_n(x)\} - m(x) = \left\{ l_0(x)\omega^{(2)}(x) + 2l_0(x)\omega^{(1)}(x) \left(\frac{f^{(1)}}{f} \right)(x) \right\} \frac{\mathrm{Var}(\mathcal{K}_{x,h})}{2} + O\left(\frac{1}{n}\right) + o(h), \tag{5}$$

$$\mathrm{Var}\{\hat{m}_n(x)\} = \frac{\sigma^2}{nf(x)} \{\Pr(\mathcal{K}_{x,h} = x)\}^2 + o\left(\frac{1}{n}\right), \tag{6}$$

where $f > 0$ is the p.m.f. of the regressor X and $f^{(1)}$, $m^{(1)}$ and $m^{(2)}$ are the finite differences of f as given in Equations (1) and (2). Hence, the consistency of the discrete semiparametric estimator $\hat{m}_n$ in Equation (4) is obtained through the asymptotic behaviour of its mean-squared error (MSE) as

$$\mathrm{MSE}(x) = \mathrm{Bias}^2\{\hat{m}_n(x)\} + \mathrm{Var}\{\hat{m}_n(x)\} \longrightarrow 0, \quad x \in \mathbb{N}.$$

Indeed, under assumptions A1 and A2, the asymptotic expansions of the bias in Equation (5) and variance in Equation (6) are such that

$$\mathrm{Bias}\{\hat{m}_n(x)\} = O\left(\frac{1}{\sqrt{n}}\right) + o(h) \longrightarrow 0 \quad \text{and} \quad \mathrm{Var}\{\hat{m}_n(x)\} = O\left(\frac{1}{n}\right) \longrightarrow 0,$$

as $h = h(n) \to 0$ and $n \to \infty$, since we assume $\mathrm{Var}(\mathcal{K}_{x,h}) = O(n^{-1/2})$. This assumption will be developed at the end of this section.

For the asymptotic normality, we need to recall the Lyapounov central limit theorem for triangular arrays (Wesolowski 1994).

THEOREM 2.1 *(Lyapounov) Assume that* $\{X_{n,j},\ j = 1, \ldots, k_n\}$ *are zero-mean independent random variables*, $n = 1, 2, \ldots$. *If*

$$\lim_{n\to\infty} \sum_{j=1}^{k_n} \mathbb{E}(X_{n,j}^2) = \Sigma^2 > 0,$$

$$\lim_{n\to\infty} \sum_{j=1}^{k_n} \mathbb{E}(|X_{n,j}|^3) = 0,$$

then $S_n = X_{n,1} + \cdots + X_{n,k_n}$ *converges in distribution to the normal law with the mean zero and the variance* Σ^2

$$S_n \xrightarrow{d} \mathcal{N}(0, \Sigma^2) \quad \text{as} \quad n \longrightarrow \infty.$$

The notation '$\xrightarrow{d}$' stands for convergence in distribution. Now, we are able to formulate the following theorem.

THEOREM 2.2 *For any fixed $x \in \mathbb{N}$, under assumptions A1 and A2, the semiparametric estimator $\hat{m}_n(x)$ converges in distribution to the normal law as follows*:

$$\sqrt{n}\{\hat{m}_n(x) - m(x)\} \xrightarrow{d} \mathcal{N}(\mu, \Lambda^2) \quad as\ n \longrightarrow \infty,$$

with the mean $\mu = \sqrt{n}\mathrm{Var}(\mathcal{K}_{x,h})\{\omega^{(2)}(x)l_0(x) + 2l_0(x)\omega^{(1)}(x)(f^{(1)}/f)(x)\}/2$ *and the variance* $\Lambda^2 = \sigma^2\{\Pr(\mathcal{K}_{x,h} = x)\}^2/f(x)$.

Proof For $x \in \mathbb{N}$ and $h > 0$, let us consider the semiparametric estimator $\hat{m}_n$ in Equation (4) and the sequence $\tilde{f}_n(x) = (1/n)\sum_{j=1}^{n} K_{x,h}(X_j)$. Using the discrete Taylor expansion of $\hat{l}(x)/\hat{l}(X_i)$ around $l_0(x)/l_0(X_i)$, we have

$$\begin{aligned}\hat{m}_n(x) \times \tilde{f}_n(x) = &\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)Y_i\frac{l_0(x)}{l_0(X_i)} + \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\frac{Y_i}{l_0(X_i)}\{\hat{l}(x) - l_0(x)\}\\ &- \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)Y_i\frac{l_0(x)}{l_0^2(X_i)}\{\hat{l}(X_i) - l_0(X_i)\}\{1 + o_p(1)\} \text{ a.s.}\end{aligned}$$

By using the equation $Y_i = l_0(X_i)\omega(X_i) + \epsilon_i$, where X_i and ϵ_i are independent variables, the terms

$$\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\frac{\epsilon_i}{l_0(X_i)}\{\hat{l}(x) - l_0(x)\} \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\epsilon_i\frac{l_0(x)}{l_0^2(X_i)}\{\hat{l}(X_i) - l_0(X_i)\}$$

are of order $o_p(h^2)$, and it ensues the following equalities:

$$\begin{aligned}\{\hat{m}_n(x) - m(x)\} \times \tilde{f}_n(x) = &\frac{l_0(x)}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\left\{\omega(X_i) + \frac{\epsilon_i}{l_0(X_i)}\right\} - l_0(x)\omega(x)\tilde{f}_n(x)\\ &+ \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\omega(X_i)\{\hat{l}(x) - l_0(x)\}\\ &- \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\omega(X_i)\frac{l_0(x)}{l_0(X_i)}\{\hat{l}(X_i) - l_0(X_i)\} + o_p(h^2)\\ = &A_n(x;h) + B_n(x;h) - C_n(x;h) + o_p(h^2). \qquad (7)\end{aligned}$$

For calculating the expectation of Equation (7), we begin with the first term A_n. Under assumptions A1 and A2 and using the discrete Taylor expansion such that

$$\begin{aligned}\mathbb{E}\{f(\mathcal{K}_{x,h})\} &= \mathbb{E}[f\{\mathbb{E}(\mathcal{K}_{x,h})\} + \{\mathcal{K}_{x,h} - \mathbb{E}(\mathcal{K}_{x,h})\}f^{(1)}\{\mathbb{E}(\mathcal{K}_{x,h})\} + o\{\mathcal{K}_{x,h} - \mathbb{E}(\mathcal{K}_{x,h})\}^2]\\ &= f\{\mathbb{E}(\mathcal{K}_{x,h})\} + o(h^2)\\ &= f(x) + \frac{1}{2}f^{(2)}(x)\mathrm{Var}(\mathcal{K}_{x,h}) + o(h^2),\end{aligned}$$

we have successively

$$\mathbb{E}\{A_n(x;h)\} = \frac{l_0(x)}{n}\mathbb{E}\left[\sum_{i=1}^{n} K_{x,h}(X_i)\{\omega(X_i)-\omega(x)\}\right] + \frac{l_0(x)}{n}\mathbb{E}\left\{\sum_{i=1}^{n} K_{x,h}(X_i)\frac{\epsilon_i}{l_0(X_i)}\right\} \quad (8)$$

$$= l_0(x)\left[\sum_{y\in\mathcal{S}_x}\omega(y)f(y)\Pr(\mathcal{K}_{x,h}=y) - \omega(x)\sum_{y\in\mathcal{S}_x} f(y)\Pr(\mathcal{K}_{x,h}=y)\right]$$

$$= l_0(x)[(\omega f)\{\mathbb{E}(\mathcal{K}_{x,h})\} - \omega(x)f\{\mathbb{E}(\mathcal{K}_{x,h})\}] + o(h^2)$$

$$= \frac{l_0(x)}{2}\{\omega^{(2)}(x)f(x) + 2\omega^{(1)}(x)f^{(1)}(x)\}\mathrm{Var}(\mathcal{K}_{x,h}) + o(h^2).$$

The expectations of the second and third terms B_n and C_n in Equation (7) are given by

$$\begin{aligned}
\mathbb{E}\{B_n(x;h)\} &= \mathbb{E}\{K_{x,h}(X_1)\omega(X_1)\}\mathbb{E}_{X_i}\{\hat{l}(x) - l_0(x)\} \\
&= \sum_{y\in\mathcal{S}_x}\omega(y)f(y)\Pr(\mathcal{K}_{x,h}=y)\mathbb{E}_{X_i}\{\hat{l}(x)-l_0(x)\} \\
&= \omega(x)f(x)\mathbb{E}_{X_i}\{\hat{l}(x)-l_0(x)\} + o(h^2), \\
\mathbb{E}\{C_n(x;h)\} &= l_0(x)\mathbb{E}\left\{K_{x,h}(X_1)\frac{\omega(X_1)}{l_0(X_1)}\right\}\mathbb{E}_{X_i}\{\widehat{l}(X_i) - l_0(X_i)\} \\
&= \omega(x)f(x)\mathbb{E}_{X_i}\{\widehat{l}(X_i) - l_0(X_i)\} + o(h^2).
\end{aligned}$$

It results in $\mathbb{E}[\{\hat{m}_n(x)-m(x)\}\times\tilde{f}_n(x)] = \mathbb{E}\{A_n(x;h)\} + o(h^2)$. Then, for the variance of Equation (7), we have

$$\mathrm{Var}\{A_n(x;h) + B_n(x;h) - C_n(x;h)\} = \frac{1}{n}\sigma^2 f(x)\{\Pr(\mathcal{K}_{x,h}=x)\}^2 + o\left(\frac{1}{n}\right),$$

with $\sigma^2 = \mathrm{Var}(\epsilon_i) < \infty$. This result is essentially due to the second term in Equation (8) given by

$$A_{1n}(x;h) = n^{-1}l_0(x)\sum_{i=1}^{n}\epsilon_i l_0^{-1}(X_i)K_{x,h}(X_i),$$

which is a sum of i.i.d. random variables; thus, we have $\mathbb{E}\{A_{1n}(x;h)\} = 0$ and, under assumptions A1 and A2,

$$\begin{aligned}
\mathrm{Var}\{A_{1n}(x;h)\} &= \frac{l_0^2(x)\mathbb{E}^2(\epsilon_1)}{n}\sum_{y\in\mathcal{S}_x} f(y)l_0^{-2}(y)\{\Pr(\mathcal{K}_{x,h}=y)\}^2 \\
&= \frac{f(x)\sigma^2}{n}\{\Pr(\mathcal{K}_{x,h}=x)\}^2 - \frac{1}{n}f^2(x)l_0^{-2}(x) + Q_n(x;h),
\end{aligned}$$

where

$$Q_n(x;h) = \frac{l_0^2(x)\sigma^2}{n}\sum_{y\in\mathcal{S}_x\setminus\{x\}} l_0^{-2}(y)f(y)\{\Pr(\mathcal{K}_{x,h}=y)\}^2 + \frac{1}{n}f^2(x)l_0^{-2}(x)$$

tends to 0 as $n \to \infty$ and $h = h(n) \to 0$. Indeed, let $y \in \mathcal{S}_x \setminus \{x\}$, we can find a constant $\eta = \eta(y) > 0$ such that

$$\begin{aligned}\Pr(\mathcal{K}_{x,h} = y) &\leq \Pr(|\mathcal{K}_{x,h} - x| > \eta) \\ &\leq \frac{1}{\eta^2}\mathbb{E}\{(\mathcal{K}_{x,h} - x)^2\} = \frac{1}{\eta^2}[\mathrm{Var}(\mathcal{K}_{x,h}) + \{\mathbb{E}(\mathcal{K}_{x,h}) - x\}^2] \longrightarrow 0 \text{ as } h \longrightarrow 0,\end{aligned}$$

and for $y = x$, we deduce the asymptotic modal probability $\Pr(\mathcal{K}_{x,h} = x) \to 1$ when $h \to 0$. The other terms in the variance of Equation (7) provide the order $o(h^2)$; we omit to detail here all these calculations. Rather, by applying the Lyapounov central limit theorem on A_{1n}, we have $\sqrt{n}A_{1n}(x; h) \xrightarrow{d} \mathcal{N}(0, f(x)\sigma^2\{\Pr(\mathcal{K}_{x,h} = x)\}^2)$.

Finally, by considering the convergence of $\tilde{f}_n$ to f states by Abdous and Kokonendji (2009), it results in

$$\begin{aligned}\sqrt{n}\{\hat{m}_n(x) - m(x)\}\tilde{f}_n(x) &= \sqrt{n}\{\hat{m}_n(x) - m(x)\}f(x) + o_p(1) \\ &= \mu f(x) + \sqrt{n}A_{1n}(x; h) + o_p(1), \\ \mathbb{E}\{\hat{m}_n(x) - m(x)\} &= \left\{\omega^{(2)}(x)l_0(x) + 2l_0(x)\omega^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right\}\frac{\mathrm{Var}(\mathcal{K}_{x,h})}{2} + o(h^2) \\ &= \frac{\mu}{\sqrt{n}} + o(h^2)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}\{\hat{m}_n(x) - m(x)\} &= \frac{1}{nf(x)}\sigma^2\{\Pr(\mathcal{K}_{x,h} = x)\}^2 + o\left(\frac{1}{n} + h^2\right) \\ &= \frac{1}{n}\Lambda^2 + o\left(\frac{1}{n} + h^2\right),\end{aligned}$$

with $\mu \equiv \mu(x; h, n)$ and $\Lambda^2 \equiv \Lambda^2(x; h)$. Hence, the desired result is obtained. ■

Remark 1 As a result, our estimator achieves $O(n^{-1/2})$ convergence rate; in addition, one can assume $\mathrm{Var}(\mathcal{K}_{x,h}) = O(n^{-1/2})$ and replace the assumption A2 with this. A more thorough treatment of the optimal order of the bandwidth h assuming $\mathrm{Var}(\mathcal{K}_{x,h}) = O(n^{-1/2})$ will be presented in Section 3.2 for the discrete associated kernels applied in this work.

3. Applications

This section presents the illustrations on data of average daily fat, sales data and cumulative stem volume. The data are fitted by the logistic model and GLM with parameter $\Theta = (\theta_1, \theta_2, \theta_3)$ in comparison to the GAM and semiparametric model using general discrete triangular associated kernels. The measure of error used is the root mean square error (RMSE) defined as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n}},$$

where $\hat{y}_j$ is the adjustment of the jth observation y_j and n is the number of observations.

In the following, we first present the parametric models (logistic and GLM) used as start functions for the discrete semiparametric model.

3.1. *Parametric models*

The GLM represents a normal model for the response variable Y_i with a logarithmic link. It has a linear predictor based on a combination of explanatory variables, such as

$$y_i = \theta_1 + \theta_2 x_i + \theta_2 \log x_i + e_i, \quad x_i \in \mathbb{N}.$$

The nonlinear model corresponds to a logistic one for the situation of population growth towards a limited value. It is given by

$$y_i = \frac{\theta_1^L}{1 + \exp\left\{-\left((x_i - \theta_2^L)/\theta_3^L\right)\right\}} + e_i, \quad x_i \in \mathbb{N}.$$

The fixed effect parameter θ_1^L is the asymptote towards which the population grows. The parameter θ_2^L is the midpoint and corresponds to the time at which $y_i = \theta_1^L/2$. The parameter θ_3^L is the scale and represents the distance on the time axis between the midpoint and the point where the response is $\theta_1^L/(1 + e^{-1})$.

Then, let us present an example of the discrete associated kernel constructed from a new discrete probability distribution introduced by Kokonendji and Zocchi (2010). It is a generalisation of the symmetric discrete triangular distributions (Kokonendji et al. 2007). We show the optimal order of the bandwidth parameter h such as $\mathrm{Var}(\mathcal{K}_{x,h}) = O(n^{-1/2})$ for these discrete associated kernels (Remark 1).

3.2. *Discrete associated kernel*

Let a_1 and a_2 be the fixed integers and h_1 and h_2 be the smoothing parameters. For any fixed $x \in \mathbb{Z}$, consider the random variable $\mathcal{DT}_{x;a_1,a_2,h_1,h_2}$ defined on supports $\mathcal{S}_{a_1,x} = \{x - 1, x - 2, \ldots, x - a_1\}$ and $\mathcal{S}_{x,a_2} = \{x, x + 1, \ldots, x + a_2\}$ and whose p.m.f. is

$$K_{x;a_1,a_2,h_1,h_2}(y) = \frac{1}{U(a_1, a_2, h_1, h_2)} \times \left\{ \left[1 - \left(\frac{x - y}{a_1 + 1} \right)^{h_1} \right] \mathbf{1}_{\mathcal{S}_{a_1,x}}(y) + \left[1 - \left(\frac{y - x}{a_2 + 1} \right)^{h_2} \right] \mathbf{1}_{\mathcal{S}_{x,a_2}}(y) \right\},$$

where

$$U(a_1, a_2, h_1, h_2) = (a_1 + a_2 + 1) - (a_1 + 1)^{-h_1} \sum_{k=1}^{a_1} k^{h_1} - (a_2 + 1)^{-h_2} \sum_{k=1}^{a_2} k^{h_2} \equiv U$$

is the normalising constant. Then, the mean is given by $\mathbb{E}(\mathcal{DT}_{x;a_1,a_2,h_1,h_2}) = x + V(a_1, a_2, h_1, h_2) \equiv x + V$ with

$$V = \frac{1}{U} \left\{ \frac{a_2(a_2 + 1) - a_1(a_1 + 1)}{2} + \sum_{k=1}^{a_1} k \left(\frac{k}{a_1 + 1} \right)^{h_1} - \sum_{k=1}^{a_2} k \left(\frac{k}{a_2 + 1} \right)^{h_2} \right\},$$

and the variance is $\mathrm{Var}(\mathcal{DT}_{x;a_1,a_2,h_1,h_2}) = W(a_1,a_2,h_1,h_2) - [V(a_1,a_2,h_1,h_2)]^2 \equiv W - [V]^2$ with

$$W = \frac{1}{U} \times \left\{ \frac{a_2(a_2+1)(2a_2+1) + a_1(a_1+1)(2a_1+1)}{6} - \sum_{k=1}^{a_1} k^2 \left(\frac{k}{a_1+1} \right)^{h_1} - \sum_{k=1}^{a_2} k^2 \left(\frac{k}{a_2+1} \right)^{h_2} \right\}.$$

Note that an R package for general discrete triangular distributions is available (Senga Kiessé, Libengué, Zocchi and Kokonendji 2010).

First, for showing the optimal order of the bandwidth assuming $\mathrm{Var}(\mathcal{DT}_{x;a,h}) = O(n^{-1/2})$, we consider the symmetric discrete triangular associated kernels $K_{x;a,h}$ with one arm $a = a_1 = a_2$ and one smoothing parameter $h = h_1 = h_2$. For h that is sufficiently small and $a \in \mathbb{N}$ fixed, we have the following expansion:

$$\begin{aligned}
\mathrm{Var}(\mathcal{DT}_{x;a,h}) &= \frac{1}{(a+1)^h U(a,h)} \left\{ \frac{a(2a+1)(a+1)^{h+1}}{3} - 2\sum_{k=1}^{a} k^{h+2} \right\} \\
&\simeq \left[\frac{a(2a+1)(a+1)}{3} \{1 + h\log(a+1)\} - 2\sum_{k=1}^{a} k^2 \{1 + h\log(k)\} \right] \\
&\quad \times \left[1 + h \left\{ (2a+1)\log(a+1) - 2\sum_{k=1}^{a} \log(k) \right\} \right]^{-1} \\
&= \left\{ \frac{a(2a^2+3a+1)}{3} \log(a+1) - 2\sum_{k=1}^{a} k^2 \log(k) \right\} h + O(h^2) \\
&= 2h\mathrm{Var}^*(\mathcal{DT}_{x;a,h}) + O(h^2).
\end{aligned}$$

It results in the following expression for the leading term of order $O(n^{-1/2})$ of the bias term in Equation (5) given by

$$\mathrm{Bias}^*\{\hat{m}_n(x)\} = h \left\{ l_0(x)\omega^{(2)}(x) + 2l_0(x)\omega^{(1)}(x) \left(\frac{f^{(1)}}{f} \right)(x) \right\} \mathrm{Var}^*(\mathcal{DT}_{x;a,h}).$$

Hence, the bandwidth h is of optimal order $O(n^{-1/2})$. This result can be generalised to the bandwidths h_i, $i = 1, 2$, for the discrete triangular associated kernels $K_{x;a_1,a_2,h_1,h_2}$, since

$$\begin{aligned}
\mathrm{Var}(\mathcal{DT}_{x;a_1,a_2,h_1,h_2}) &= \sum_{i=1}^{2} h_i \left\{ \frac{a_i(2a_i^2+3a_i+1)}{6} \log(a_i+1) - \sum_{k=1}^{a_i} k^2 \log(k) \right\} + O(h_1^2 + h_2^2) \\
&= \sum_{i=1}^{2} h_i \mathrm{Var}^*(\mathcal{DT}_{x;a_i,h_i}) + O(h_1^2 + h_2^2).
\end{aligned}$$

Then, for the bandwidth selection, one can directly use the optimal values of (h_1, h_2), which minimise the integrated squared error (ISE) such as $\mathrm{ISE}(h_1,h_2) = \sum_{x\in\mathbb{N}}\{\hat{m}_n(x) - m_0(x)\}^2$, where m_0 is the observed value. Another method should be to minimise the integrated MSE of the proposed discrete semiparametric regression estimator; thus, the bandwidth selection might be realised by a cross-validation score function. One can refer to Chiu (1991) for kernel density

estimation and to Kokonendji, Senga Kiessé and Balakrishnan (2009a) for semiparametric kernel estimation of p.m.f. Here, we do not investigate these different approaches and just propose some small and large values of h_1 and h_2 equal to 0.1 and 3.0 to point out the influence of both bandwidth parameters h_1 and h_2 on goodness of fit, degree of smoothing and boundary bias. To reduce this bias, we fix one bandwidth parameter and vary the other; in this way, we have an influence on both smoothing and fitting. Another possibility is to transform the arms for reducing bias as proposed by Kokonendji and Zocchi (2010). However, we do not exclude another discrete kernel as binomial or Poisson proposed in density estimation because of their advantage for small or moderate sample sizes (Senga Kiessé 2009), even if this advantage does not still hold for the regression.

At last, for both arms a_1 and a_2, in practice, they are small and equal to 1, 2 or 3 (Kokonendji and Zocchi 2010). Therefore, in what follows, we consider the general discrete triangular distributions with $a_1 = 3$ and $a_2 = 1$. We recommend these discrete distributions for using an associated kernel for our proposed estimator because of the advantages provided by the two smoothing parameters (h_1, h_2).

3.3. *Average daily fat*

Figure 1 indicates the difference between discrete semiparametric general triangular kernel model $(a_1, a_2) = (3, 1)$ and GAM, on the one hand, and GLM, on the other hand. Indeed, both first ones detect the plateau associated with the observations $x = 19, 20, \ldots, 27$, while the third does not detect it. The results given in Table 3 show that a better discrete semiparametric adjustment is obtained using bandwidth parameters $h_1 = h_2 = 0.1$ (giving the smallest RMSE); however, there is a lack of smoothing. The value of the RMSE increases and the degree of smoothing is improved when the values of h_1 and h_2 increase to 3.0. Then, we fix one of the bandwidth parameter and change the other. For $h_1 = 0.1$ fixed and h_2 varying to 3.0, the error RMSE increases, but we keep a good estimation at the right boundary $x = 35$ with a satisfying amount of smoothing, which is improved in comparison with the case $h_1 = h_2 = 0.1$. For $h_2 = 0.1$ fixed and h_1 varying to 3.0, we obtain a similar result by keeping a good adjustment on the left boundary $x = 1$.

3.4. *Sales data*

Table 4 and Figure 2 present the results corresponding to sales data. Similar to the previous example, the h-values $h_1 = h_2 = 0.1$ for the discrete semiparametric general triangular model

Table 3. RMSE (in %) calculated from the GLM, GAM and discrete semiparametric model with general triangular associated kernels on average daily fat data.

		Semiparametric regression with general discrete triangular kernel $a_1 = 3, a_2 = 1$			
GLM	GAM	$h_1 = h_2 = 0.1$	$h_1 = h_2 = 3.0$	$h_1 = 0.1$ and $h_2 = 3.0$	$h_1 = 3.0$ and $h_2 = 0.1$
5.129	2.438	1.075	3.886	2.032	4.998

Table 4. RMSE calculated from the logistic model, GAM and discrete semiparametric model with general triangular associated kernels on sales data.

		Semiparametric regression with general discrete triangular kernel $a_1 = 3, a_2 = 1$			
GLM	GAM	$h_1 = h_2 = 0.1$	$h_1 = h_2 = 3.0$	$h_1 = 0.1$ and $h_2 = 3.0$	$h_1 = 3.0$ and $h_2 = 0.1$
2.427	2.245	2.218	2.569	2.289	2.778

with $(a_1, a_2) = (3, 1)$ give the smallest RMSE but not the most satisfying amount of smoothing. In comparison to the parametric model, both satisfying degree of smoothing and fitting are obtained with $h_1 = h_2 = 3.0$. Furthermore, in general, the logistic model seems to underestimate the y-values contrary to the semiparametric associated kernel model and GAM.

In the two previous applications, the discrete semiparametric estimator using triangular associated kernel with $(a_1, a_2) = (3, 1)$ and $(h_1, h_2) = (0.1, 3.0)$ and GAM are closed in terms of goodness of fit and smoothing. Concerning the semiparametric model with discrete general triangular kernel $(a_1, a_2) = (3, 1)$, use of smoothing parameters $h_1 = h_2 = 3.0$ provides the most interesting results, considering the researched compromise between some good smoothing and fitting. Thus, some relative big bandwidths are recommended, considering the lack of smoothing, in spite of the fact that the optimal bandwidth is of order $O(n^{-1/2})$; however, a bandwidth of this optimal order would be recommended, considering the goodness of fit. For the last example, we directly apply the semiparametric model using these values of parameters with a logistic model as the start function.

3.5. *Tree data*

Here, the performance of the discrete semiparametric logistic kernel regression model $\hat{m}_n$ is illustrated on a tree data set (Table 2) having a distribution with sigmoïdal form in comparison with the purely logistic model and GAM. In Figure 3, the fitted curve for the logistic model does not succeed well in describing the variations of the distribution; it results in an error RMSE $= 5.378\%$. The semiparametric logistic model indicates that the use of discrete general triangular kernel with bandwidth parameters $h_1 = h_2 = 3.0$ and arms $(a_1, a_2) = (3, 1)$ provides some better amount of smoothing and adjustment on data. Thus, the capacity of the semiparametric estimator to detect the variations of the distribution to be estimated is clearly shown through the role of the nonparametric correction factor $\omega(x)$, $x = 0, 1, \ldots, 55$, (RMSE = 1.341%) at the opposite of the parametric model. In addition, the semiparametric associated kernel model gives good estimations at the left and right boundary points. The used semiparametric model is similar to GAM (RMSE = 1.757%) in terms of performance and thus the corresponding fitted points are closed to the observations in Figure 3; however, GAM does not adjust well at the right boundary.

4. Concluding remarks

This paper has investigated discrete semiparametric kernel estimators with logistic and normal models as the known start functions. The general discrete triangular associated kernel with left and right bandwidth parameters was used, which provided control on both the goodness of fit and degree of smoothing. Thus, the constructed semiparametric estimators outperformed the parametric models in the three examples given. They allowed to obtain some satisfying adjustments, amount of smoothing and reduction of boundary bias, which are often required. Several parametric models may be used as start functions in the semiparametric procedure such as some nonlinear Gompertz or log-linear models for count data. Similarly, other discrete associated kernels may be useful as binomial. Finally, the introduction of count explanatory variables and an optimal choice of bandwidth parameters may be studied for the discrete semiparametric associated kernel estimator.

Acknowledgement

We sincerely thank the anonymous referees and the Associate Editor for their valuable comments.

References

Abdous, B., and Kokonendji, C.C. (2009), 'Consistency and Asymptotic Normality for Discrete Associated-Kernel Estimator', *African Diaspora Journal of Mathematics*, 8, 63–70.

Abdous, B., Kokonendji, C.C., and Senga Kiessé, T. (2010), 'On Semiparametric Regression for Count Explanatory Variables', *Journal of Statistical Planning and Inference* (in revision).

Chiu, S.-T. (1991), 'The Effect of Discretization Error on Bandwidth Selection for Kernel Density Estimation', *Biometrika*, 78, 436–441.

Fan, J., Wu, Y., and Feng, Y. (2009), 'Local Quasi-Likelihood with a Parametric Guide', *The Annals of Statistics*, 37, 4153–4183.

Glad, I.K. (1998), 'A Note on Unconditional Properties of a Parametrically Guided Nadaraya–Watson Estimator', *Statistics and Probability Letters*, 37, 101–108.

Hastie, T.J., and Tibshirani, R.J. (1990), *Generalized Additive Models* (4th ed.), London: Chapman and Hall.

Kokonendji, C.C., and Zocchi, S.S. (2010), 'Extensions of Discrete Triangular Distributions and Boundary Bias in Kernel Estimation for Discrete Functions', *Statistics and Probability Letters*, 80, 1655–1662.

Kokonendji, C.C., Senga Kiessé, T., and Zocchi, S.S. (2007), 'Discrete Triangular Distributions and Non-Parametric Estimation for Probability Mass Function', *Journal of Nonparametric Statistics*, 19, 241–254.

Kokonendji, C.C., Senga Kiessé, T., and Balakrishnan, N. (2009a), 'Semiparametric Estimation for Count Data Through Weighted Distributions', *Journal of Statistical Planning and Inference*, 139, 3625–3638.

Kokonendji, C.C., Senga Kiessé, T., and Demétrio, C.G.B. (2009b), 'Appropriate Kernel Regression on a Count Explanatory Variable and Applications', *Advances and Applications in Statistics*, 12, 99–125.

Martins-Filho, C., Mishra, S., and Ullah, A. (2008), 'A Class of Improved Parametrically Guided Nonparametric Regression Estimators', *Econometric Reviews*, 27, 542–573.

McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Rivoire, M., Deleuze, C., Longuetaud, F., Saint-André, L., Morneau, F., Vallet, P., Bouvet, A., and Gauthier, A. (2010), 'Exploring the Variability of Biomass Distribution in Individual Forest Trees', *The International Forestry Review*, 12, 343.

Senga Kiessé, T. (2009), 'Nonparametric Approach by Discrete Associated-Kernel for Count Data', Ph.D. in Statistics, University of Pau, http://tel.archives-ouvertes.fr/tel-00372180/fr/.

Senga Kiessé, T., Libengué, F.G., Zocchi, S.S., and Kokonendji, C.C. (2010), 'The R Package for General Discrete Triangular Distributions', http://cran.r-project.org/web/packages/TRIANGG/index.html.

Wesolowski, J. (1994), 'The Lyapounov Central Limit Theorem for Factorizable Arrays', *Proceedings of the American Mathematical Society*, 122, 565–574.