Comportement asymptotique de l'estimateur à noyau de la densité, avec données discrétisées, pour des champs aléatoires dépendants et non-stationnaires

Michel Harel[1], Jean-François Lenain[2] and Joseph Ngatchou-Wandji[3,*]

[1] *Faculté des Sciences et Techniques, Limoges 87060 Limoges, Cedex, France*
[2] *IUFM du Limousin, Limoges Cedex 87036, France & IMT (UMR CNRS 5219, Université Paul Sabatier, Toulouse Cedex 31062, France*
[3,*] *EHESP de Rennes & Université de Lorraine, Institut Élie Cartan de Nancy, 54506 Vandoeuvre-lès-Nancy cedex, France.*

Résumé. Nous étudions le comportement asymptotique d'estimateurs à noyau de la densité pour des suites de données spatiales dépendantes discrétisées, localement non-stationnaire et *convergent* vers une suite stationnaire de données spatiales. Notre étude porte essentiellement sur le biais et la normalité asymptotique des estimateurs.

Mots-clés. Estimation à noyau, données spatiales, dépendance faible, non-stationnarité

Abstract. We investigate the asymptotic behavior of binned kernel density estimators for dependent and locally non-stationary random fields *converging* to stationary random fields. We focus on the study of the bias and the asymptotic normality of the estimators.

Keywords. Kernel estimator, spatial data, weak dependance, non-stationarity

1 Introduction

In many practical situations, one can be concerned with the statistical study of an unobserved stationary random fields $(X^*_{\mathbf{i}})_{\mathbf{i}\in\mathbb{Z}^N}, N \in \mathbb{N}$ at the place of which a sequence of random fields $(X_{\mathbf{i}})_{\mathbf{i}\in\mathbb{Z}^N}$ is observed and both series are linked by an equation of the form

$$X_{\mathbf{i}} = \vartheta(\mathbf{i}) + (1 + \zeta(\mathbf{i}))X^*_{\mathbf{i}},\ \mathbf{i} \in \mathbb{Z}^N, \tag{1.1}$$

where ϑ and ζ are some functions defined on $\mathbb{Z}^N$. Denoting by f^* the density function (with respect to Lebesgue measure) of the stationary distribution of $(X^*_{\mathbf{i}})_{\mathbf{i}\in\mathbb{Z}^N}$, statistical inferences of interest can be testing hypothesis on f^* or estimating this function. It is clear that such works can only be done trough the non-stationary $(X_{\mathbf{i}})_{\mathbf{i}\in\mathbb{Z}^N}$ defined in (1.1) and studied for instance in [6]. In the present paper, we show that under some conditions, the so-called binned kernel density estimator (BKDE) based on $(X_{\mathbf{i}})_{\mathbf{i}\in\mathbb{Z}^N}$ is consistent to f^*

and is asymptotically normal. We study the BKDE instead of the classical kernel density estimator (KDE) because of its lower coast computational advantage.

Let $n_1, \ldots, n_N$ be positive integers. Denote by $\mathbf{n}$ the N-dimensional vector $(n_1, ..., n_N)$ and by $I_{\mathbf{n}} = \prod_{k=1}^{N} [1, \ldots, n_k]$, a finite rectangular domain of the integer lattice points in N-dimensional Euclidian space $\mathbb{Z}^N$. Taking in (1.1) $\vartheta = \zeta = 0$, the Rosenblatt [7] kernel density estimator $\widehat{f}$ of f^* is defined by

$$\widehat{f}(x) = \frac{1}{\widehat{\mathbf{n}}h} \sum_{\mathbf{i} \in I_{\mathbf{n}}} K\left(\frac{x - X_{\mathbf{i}}}{h}\right), \ x \in \mathbb{R}, \tag{1.2}$$

where $\widehat{\mathbf{n}}$ is the finite product $n_1 \ldots n_N$, $h = h(\mathbf{n})$ is the smoothing parameter and K is a bounded integrable real-valued function defined on $\mathbb{R}$, called kernel.

It is well known that (1.2) has a high computation coast. The most popular way to reduce this coast is to prebin the data, an operation which leads to the BKDE studied for instance in Hall [2], Scott and Shearter [8], Hall and Wand [1] and Holmström [4]. The BKDE can be seen as aproximations of KDE, or as direct estimators of f^*. They have the general form

$$\widetilde{f}(x) = \frac{1}{\widehat{\mathbf{n}}h} \sum_{j \in \mathbf{Z}} K\left(\frac{x - a_j}{h}\right) \sum_{\mathbf{i} \in I_{\mathbf{n}}} T\left(\frac{X_{\mathbf{i}} - a_j}{\delta}\right), \ x \in \mathbb{R}, \tag{1.3}$$

where $\{a_j\} = \{a_0 + j\delta\}_{j \in \mathbb{Z}}$ is a given grid points with an arbitrary origin $a_0 \in \mathbb{R}$, T is a kernel with window width δ, and h and K are as above.

The aim is the study of the behavior of $\widetilde{f}$ for *local non-stationry* α-mixing random fields. That is, for α-mixing random fields $\{X_{\mathbf{i}}\}_{\mathbf{i} \in \mathbb{Z}^N}$ for which there exists a finite set of neighboring sites $I_{\mathbf{n}}^* \subset I_{\mathbf{n}}$ such that the sequence $\{X_{\mathbf{i}}\}_{\mathbf{i} \in I_{\mathbf{n}}^*}$ is possibly non-stationary and the sequence $\{X_{\mathbf{i}}\}_{\mathbf{i} \in I_{\mathbf{n}} - I_{\mathbf{n}}^*}$ is stationary with a stationary distribution different from at least that of one of the $X_{\mathbf{i}}$, $\mathbf{i} \in I_{\mathbf{n}}^*$. For example, for the series satisfying (1.1), we consider the cases where $\vartheta(\mathbf{i})$ and $\zeta(\mathbf{i})$ tend to zero as $\mathbf{i}$ tends to infinity.

Denote by $\lfloor x \rceil$ the integer closest to x and by $[x]$ the largest integer less than or equal to x. For $\delta > 0$ and an arbitrary $a_0 \in \mathbb{R}$, define the real-valued function a by $a(y) = \delta \lfloor (y - a_0)/\delta \rceil + a_0$, for rounding to the nearest value, or $a(y) = \delta \left[(y - a_0)/\delta\right] + a_0$, for rounding down. For a clear presentation, we restrict ourselves to the cases where $\widetilde{f}$ is defined with $T(y) = I(y \in [0, 1])$ or with $T(y) = I(y \in [-1/2, 1/2))$. More precisely, we consider

$$\widetilde{f}(x) \quad = \quad \frac{1}{nh} \sum_{\mathbf{i} \in I_{\mathbf{n}}} K\left(\frac{x - X_{\mathbf{i}}}{h} + \frac{\delta}{h} Z_{\mathbf{i}}\right), \ x \in \mathbb{R}, \tag{1.4}$$

where for $\mathbf{i} \in I_{\mathbf{n}}$, $Z_{\mathbf{i}} = \left(X_{\mathbf{i}} - a(X_{\mathbf{i}})\right)/\delta \in (0, 1)$ for rounding down or $Z_{\mathbf{i}} \in [-1/2, 1/2)$ for rounding to the nearest value.

In Section 2, we list the notations and the sequence of assumptions considered along the paper. In Section 3, we state our mains results.

2 General assumptions

For all $\mathbf{i} = (i_1, \ldots, i_N) \in I_{\mathbf{n}}$, $|\mathbf{i}| = \max_{k=1,\ldots,N} |i_k|$. We use the notation $\mathbf{n} \to \infty$ to mean that $\min_{k=1,\ldots,N} n_k \to \infty$ and $\max_{j,k=1,\ldots,N}(n_j/n_k) < C$ for some generic constant $C > 0$. The total variation norm of a real-value function ϖ is denoted by $||\varpi||_V$, and if $\int \varpi^p(x)dx < \infty$, its L_p-norm is defined by $||\varpi||_p = \left(\int \varpi^p(x)dx\right)^{1/p}$. For simplicity, we only treat the case where $I_{\mathbf{n}}^*$ contains only one single site $\mathbf{i_0}$ of $I_{\mathbf{n}}$. We make the following assumptions:

(A1) :

• For all $\mathbf{i} \in I_{\mathbf{n}}$, $X_{\mathbf{i}}$ has a cumulative distribution function $F_{\mathbf{i}}$ with density function $f_{\mathbf{i}}$, both continuous.
• There exists a strictly stationary random field $\{X_{\mathbf{i}}^*\}_{\mathbf{i} \in \mathbb{Z}^N}$ with continuous distribution and density functions F^* and f^* respectively.
• For $|\mathbf{j} - \mathbf{i}| > 0$, $(X_{\mathbf{i}}, X_{\mathbf{j}})$ has a continuous distribution function $F_{\mathbf{i},\mathbf{j}}$ with a continuous density function $f_{\mathbf{i},\mathbf{j}}$.
• For $|\mathbf{j} - \mathbf{i}| > 0$, $(X_{\mathbf{i}}^*, X_{\mathbf{j}}^*)$ has a continuous distribution function $F_{|\mathbf{i}-\mathbf{j}|}^*$ with a continuous density function $f_{|\mathbf{i}-\mathbf{j}|}^*$.
• For $|\mathbf{i} - \mathbf{i_0}| < |\mathbf{j} - \mathbf{i_0}|$ and $|\mathbf{i} - \mathbf{i}_0| \to \infty$, and for some non-increasing function η which satisfies $\sum_{\mathbf{i} \in I_{\mathbf{n}}} \eta\left(|\mathbf{i} - \mathbf{i}_0|\right) < \infty$, as $\mathbf{n} \to \infty$,

$$\left|F_{\mathbf{i},\mathbf{j}} - F_{|\mathbf{j}-\mathbf{i}|}^*\right|\Big|_{\mathsf{V}} = O\left(\eta\left(|\mathbf{i} - \mathbf{i}_0|\right)\right) \longrightarrow 0 \quad \text{and} \quad \|F_{\mathbf{i}} - F^*\|_{\mathsf{V}} = O\left(\eta\left(|\mathbf{i} - \mathbf{i}_0|\right)\right) \longrightarrow 0.$$

(A2) :

• The nonnegative function K is bounded, symmetric, absolutely continuous and piecewise differentiable with a bounded derivative, and is such that $\int K(x)dx = 1$, $\int x^2 K(x)dx$, $\int xK'(x)dx, \sup_{x \in \mathbb{R}} |K'(x)|$ and $\int |K'(x)|\, dx$ are finite.
• The sequences $h = h(\widehat{\mathbf{n}})$ and $\delta = \delta(\widehat{\mathbf{n}})$ are positive and are such that $h \longrightarrow 0$, $\delta \longrightarrow 0$, $\delta/h \longrightarrow 0$, $\widehat{\mathbf{n}}h \longrightarrow \infty$, as $\mathbf{n} \to \infty$.

(A3) :

• The sequences of random fields $\{X_{\mathbf{i}}\}_{\mathbf{i} \in \mathbb{Z}^N}$ and $\{X_{\mathbf{i}}^*\}_{\mathbf{i} \in \mathbb{Z}^N}$ are α-mixing with the same mixing rate. That is, for all $u, v > 0$,

$$\max_{U,V \subset \mathbb{Z}^N} \sup_{A \in \mathcal{U}_{U(u)}, B \in \mathcal{V}_{V(v)}} |P(A \cap B) - P(A)P(B)| = \alpha_{u,v}(m) \longrightarrow 0 \text{ as } m \to \infty,$$

where $\mathcal{U}_{U(u)}$ and $\mathcal{V}_{V(v)}$ are respectively the σ-algebras spanned by $(X_{\mathbf{i}}, \mathbf{i} \in U, \mathrm{Card}(U) \leq u)$ and $(X_{\mathbf{j}}, \mathbf{j} \in V, \mathrm{Card}(V) \leq v)$ with $\inf_{\mathbf{i} \in U(u), \mathbf{j} \in V(v)} |\mathbf{i} - \mathbf{j}| \geq m$, and $\alpha_{u,v}(m)$ is an increasing function of u and v, and a decreasing function of m. Here, we take $\alpha_{u,v}(n) = \zeta_{u,v}\, \alpha\,(n) \to 0$ as $n \to \infty$, for all u and $v > 0$, where $\zeta_{u,v}$ is an increasing function of u and v, and $\alpha\,(n)$ is a decreasing function of n.

• There exist $\nu > 0$ and $c \geq 3$, such that for all $u, v \in \mathbb{N}^*$, $u + v \leq c$, $u, v \geq 2$, $\sum_{r\geq 1}(r+1)^{N(c-u+1)-1}[\alpha(r)]^{1/\lambda} < \infty$ with $\lambda = (c+\nu)/\nu$, and for $h = h(\widehat{\mathbf{n}})$ there exists an increasing sequence $m = m(\widehat{\mathbf{n}})$ satisfying $hm^N \longrightarrow 0$, and $h^{-1/\lambda} \sum_{k=m}^{\infty} k^{N-1}\alpha(k)^{1/\lambda} \longrightarrow 0$ as $\mathbf{n} \to \infty$.

3 Main results

Our main results are on the study of the bias of $\widetilde{f}$ with respect to the stationary density f^*. The proofs of our results are based on the same techniques as those of [5] and [3].

Proposition 1 *Under the assumptions (A1)-(A2), the binned kernel estimator $\widetilde{f}$ is an asymptotically unbiased estimator of f^*. Moreover,*

$$E\left[\widetilde{f}(x)\right] - f^*(x) = E\left[\widehat{f}^*(x) - f^*(x)\right] + O\left(\frac{\delta}{h} + \frac{1}{\widehat{\mathbf{n}}h}\right) \tag{3.1}$$

where $E\left[\widehat{f}^*(x) - f^*(x)\right]$ stands for the bias of the Rosenblatt estimator of f^* based on $(X^*_{\mathbf{i}})_{\mathbf{i}\in\mathbb{Z}^N}$.

Proposition 2 *Under the assumptions (A1) and (A2)*

$$MISE\left(\widetilde{f}(x)\right) - MISE\left(\widehat{f}^*(x)\right) \;=\; O\left(\frac{1}{\widehat{\mathbf{n}}} + \frac{\delta}{h} + \delta\int_{\Delta} u du + \frac{1}{\widehat{\mathbf{n}}h}\right).$$

For $m = o(\widehat{\mathbf{n}}^{1/N})$, if

$$\frac{1}{\widehat{\mathbf{n}}h}\sum_{k=m}^{\infty} k^{N-1}\alpha(k)^{1/\lambda} \longrightarrow 0 \quad as \quad \mathbf{n} \to \infty,$$

then $\int E\left[\widetilde{f}(x) - f^*(x)\right]^2 dx$ tends to 0 as $\mathbf{n}$ tends to infinity.

Proposition 3 *Under the assumptions (A1)-(A3), for $m = o\left(\widehat{\mathbf{n}}^{1/N}\right)$ the mean square quadratic difference between $\widetilde{f}$ and $\widehat{f}$ is given by :*

$$E\left\{\left[\widetilde{f}(x) - \widehat{f}(x)\right]^2\right\} \;=\; O\left(\frac{\delta^2}{h^2}\right).$$

Theorem 1 *Under the assumptions (A1) and (A2), for $m = o(\widehat{\mathbf{n}}^{1/N})$, if*

$$\frac{1}{\widehat{\mathbf{n}}h}\; h^{-1/\lambda}\; \sum_{k=m}^{\infty} k^{N-1}\;\alpha\left(k\right)^{1/\lambda} \longrightarrow 0 \quad as \quad \mathbf{n} \to \infty,$$

then the binned kernel estimator $\widetilde{f}$ converges in mean square to f^*.

Theorem 1 is an immediate consequence of the following triangle inequality

$$\sqrt{E\left[\widetilde{f}(x)-f^{*}(x)\right]^{2}} \quad\leq\quad \sqrt{E\left[\widetilde{f}(x)-\widehat{f}(x)\right]^{2}}+\sqrt{E\left[\widehat{f}(x)-f^{*}(x)\right]^{2}}, \qquad (3.2)$$

Proposition 3 and the following lemma.

Lemma 1 *Under the assumptions (A1) and (A2), if there exists* $m=o\left(\widehat{\mathbf{n}}^{1/N}\right)$ *such that*

$$\frac{1}{\widehat{\mathbf{n}}h}h^{-1/\lambda}\sum_{k=m}^{\infty}k^{N-1}\ \alpha\left(k\right)^{1/\lambda}\longrightarrow 0 \quad as \quad \mathbf{n}\rightarrow\infty,$$

then the Rosenblatt estimator $\widehat{f}$ *converges to* f^{*} *in mean square. Moreover, if assumption (A3) is satisfied, then*

$$\begin{aligned} E\left[\widehat{f}(x)-f^{*}(x)\right]^{2} \quad &= \quad \frac{1}{\widehat{\mathbf{n}}h}\left\|K\right\|_{2}^{2}f^{*}\left(x\right)+o\left(\widehat{\mathbf{n}}^{-1}\right) \\ &\quad +O\left(\frac{m^{N}}{\widehat{\mathbf{n}}}+\frac{1}{\widehat{\mathbf{n}}h}\ h^{-1/\lambda}\sum_{k=m}^{\infty}k^{N-1}\alpha_{1,1}\left(k\right)^{1/\lambda}+\frac{1}{\widehat{\mathbf{n}}^{2}h^{2}}\right). \qquad (3.3)\end{aligned}$$

Proposition 3 shows that the first term in the right-hand side of (3.2) tends to zero, and Lemma 1 shows that the second term also tends to zero. This establishes Theorem 1.

Define

$$S_{\mathbf{n}}=\sqrt{\widehat{\mathbf{n}}h}\left[\widetilde{f}(x)-E\widetilde{f}(x)\right].$$

Theorem 2 *Assume that (A1)-(A3) hold. Let* $m=o\left(\ell\right)$ *and* $\ell=\widehat{\mathbf{n}}^{(1-\beta)/N}$, $\beta\in(0,1)$. *If*

$$\hat{\mathbf{n}}^{\beta}\zeta_{\ell^{N},\ell^{N}}\left[\alpha(m)+\sum_{i=1}^{\infty}i^{N-1}\alpha\big(i\big(m+\ell\big)\big)\right]\longrightarrow 0,\ \mathbf{n}\rightarrow\infty, \qquad (3.4)$$

then for $h=Cn^{\beta_{0}}, 0<\beta_{0}<\beta$, $S_{\mathbf{n}}$ *converges in distribution to a zero-mean Gaussian random variable with variance* $\sigma^{2}(x)=f^{*}(x)||K||_{2}^{2}$.

References

[1] Hall P. and Wand M.P. (1996). On the accuracy of binned kernel density estimators. *J. Multivar. Anal.* **56** 165-184.

[2] Hall P. (1982). The influence of rounding errors on some nonparametric estimators of a density and its derivatives. *SIAM J. Appl. Math.* **42** 390-399.

[3] Harel M., Lenain J.-F. and Ngatchou-Wandji J. (2013). Asymptotic normality of binned kernel density estimators for non-stationary dependent random variables *in Mathematical Statistics and Limit Theorems - Festschrift in Honour of Paul Deheuvels, ed. Hallin, M., Mason, D.M., Pfeifer, D. and Steinebach, J., Springer Proceedings in Mathematics & Statistics, Birkhäuser, Basel*, *in press.*

[4] Holström (2000). The accuracy and the computational complexity of a multivariate binned kernel density estimator. *J. Multivar. Anal.* **72** 264-309.

[5] Lenain J.F., Harel M. and Puri M.L. (2011). Asymptotic behavior of the kernel density estimator for dependent and nonstationary random variables with binned data. Festschrift in Honor of Professor P.K. Bhattacharya. Worldscientific. 396-420.

[6] Perrin O. and Senoussi R. (2000). Reducing non-stationary random fields to stationarity and isotropy using a space deformation. *Statist. Probab. Lett.* **48** 23-32.

[7] Rosenblatt M. (1971). Curve estimates. *Ann. Math. Stat.* **42** (6) 1815-1842.

[8] Sott D.W. and Sheather S.J. (1985). Kernel density estimation with binned data. *Comm. Statist. Theory Methods* **14** 1353-1359.