Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 342 (2006) 693–696
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique

Polygone des fréquences pour des champs aléatoires ☆

Michel Carbon

Département de Statistique, ENSAI, rue Blaise-Pascal, 35172 Bruz, France

Reçu le 13 août 2005 ; accepté après révision le 6 février 2006

Présenté par Paul Deheuvels

Résumé

Dans cette Note, nous présentons le polygone des fréquences comme estimateur de la densité pour des champs aléatoires indexés sur un treillis. Nous déterminons la largeur de cellule optimale qui minimise asymptotiquement l'erreur moyenne quadratique intégrée (IMSE). On montre également que, sous des conditions très générales, le polygone des fréquences atteint la même vitesse de convergence pour l'IMSE que les estimateurs à noyaux. Il peut aussi atteindre la vitesse optimale de la convergence uniforme sous des conditions générales. En conséquence, du point de vue de la convergence uniforme ou de l'IMSE, le polygone des fréquences est un très bon estimateur de la densité. ***Pour citer cet article : M. Carbon, C. R. Acad. Sci. Paris, Ser. I 342 (2006).***

Abstract

Frequency polygons for random fields. The purpose of this Note is to investigate the frequency polygon as a density estimator for stationary random fields indexed by multidimensional lattice points space. Optimal bin widths which asymptotically minimize integrated errors (IMSE) are derived. Under mild regularity assumptions, frequency polygons achieve the same rate of convergence to zero of the IMSE as kernel estimators. They can also attain the rate of uniform convergence under general conditions. Frequency polygons thus appear to be very good density estimators with respect to both criteria of IMSE and uniform convergence. ***To cite this article: M. Carbon, C. R. Acad. Sci. Paris, Ser. I 342 (2006).***

1. Introduction

Le but de cette Note est d'étudier le polygone des fréquences comme estimateur de la densité pour des variables aléatoires présentant une interaction spatiale (exemple : taux de radioactivité dans une zone proche d'une centrale nucléaire, production de blé dans une région, etc.). Un approche non paramétrique de ce type devrait être conseillée avant tout choix ultérieur éventuel d'une famille paramétrique de densités. Le polygone des fréquences est construit en joignant les milieux des segments de l'estimateur histogramme. L'effort calculatoire est donc équivalent à celui de l'histogramme, mais ses performances sont nettement supérieures.

Considérons un champ aléatoire $\{X_{\mathbf{n}}\}$ strictement stationnaire indexé par $\mathbf{n}$ appartenant au treillis Z^N, $N \geqslant 1$, et défini sur un certain espace probabilisé $(\Omega, \mathcal{F}, P)$. Un tel point $\mathbf{n} = (n_1, \ldots, n_N)$ de Z^N sera appelé un site. Pour deux

☆ À la mémoire de Jean et Alain Gautier.

Adresse e-mail : carbon@ensai.fr (M. Carbon).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2006.02.019

ensembles de sites S et S', désignons par $\mathcal{B}(S) = \mathcal{B}(X_{\mathbf{n}}, \mathbf{n} \in S)$ et $\mathcal{B}(S') = \mathcal{B}(X_{\mathbf{n}}, \mathbf{n} \in S')$ les tribus engendrées par les variables aléatoires $\{X_{\mathbf{n}}\}$ où $\mathbf{n}$ décrit S et S' respectivement. On notera $\operatorname{dist}(S, S')$ la distance euclidienne entre S et S'. Nous supposerons que $\{X_{\mathbf{n}}\}$ est fortement mélangeant au sens où : il existe une fonction $\varphi(t) \downarrow 0$ quand $t \to \infty$, telle que, quels que soient $S, S' \subset Z^N$,

$$\begin{aligned}\alpha\big(\mathcal{B}(S), \mathcal{B}\big(S'\big)\big) &= \sup\big\{\big|P(AB) - P(A)P(B)\big|, A \in \mathcal{B}(S), B \in \mathcal{B}\big(S'\big)\big\} \\ &\leqslant h\big(\operatorname{Card}(S), \operatorname{Card}\big(S'\big)\big)\varphi\big(\operatorname{dist}\big(S, S'\big)\big),\end{aligned} \tag{1.1}$$

où $\operatorname{Card}(S)$ désigne le cardinal de S. Dans (1.1), h est une fonction positive symétrique, croissante en chaque variable. On supposera aussi que h vérifie l'une des deux conditions suivantes :

$$h(n, m) \leqslant \min\{m, n\} \tag{1.2}$$

ou

$$h(n, m) \leqslant C(n + m + 1)^{\tilde{k}} \tag{1.3}$$

avec $\tilde{k} \geqslant 1$ et $C > 0$.

Les conditions (1.2) and (1.3) sont les mêmes que celles de Neaderhouser [12] et Takahata [16] respectivement. Elles sont vérifiées par beaucoup de modèles spatiaux. Des exemples peuvent être trouvés dans Neaderhouser [12], Rosenblatt [13] et Guyon [7]. Pour des références sur les champs aléatoires, on pourra consulter, par exemple, Neaderhouser [12], Bolthausen [2], Guyon and Richardson [8], Guyon [7], Tran [17], Tran et Yakowitz [18], Carbon, Hallin et Tran [5], Carbon, Tran et Wu [6], Hallin, Lu et Tran [9–11] et Biau [1].

Supposons que $X_{\mathbf{n}}$ prenne des valeurs réelles et possède une densité f uniformément continue.

On écrira $\mathbf{n} \to \infty$ si $\min\{n_k\} \to \infty$ et $|n_j/n_k| < C$ pour $0 < C < \infty$, $1 \leqslant j, k \leqslant N$. Toutes les limites seront prises quand $\mathbf{n} \to \infty$.

On observe le champ aléatoire $\{X_{\mathbf{n}}\}$ sur une région rectangulaire $I_{\mathbf{n}}$ définie par :

$$I_{\mathbf{n}} = \big\{\mathbf{i}\colon\ \mathbf{i} \in Z^N,\ 1 \leqslant i_k \leqslant n_k,\ k = 1, \ldots, N\big\},$$

qu'on choisit d'étendre à la même vitesse dans toutes les directions.

Définissons $\hat{\mathbf{n}} = n_1 \ldots n_N$. On a ainsi collecté $\hat{\mathbf{n}}$ observations.

Dans le cas $N = 1$, pour le critère IMSE, les histogrammes atteignent la vitesse de convergence $n^{-2/3}$, et les estimateurs à noyaux la vitesse $n^{-4/5}$. La vitesse optimale de convergence uniforme des estimateurs non paramétriques de la densité dans le cas i.i.d. (voir Stone [15]) est $(n^{-1}\log n)^{1/3}$.

Sous des conditions de dépendance faible, on verra ici que la vitesse de convergence des polygones des fréquences est de l'ordre de $\hat{\mathbf{n}}^{-4/5}$ pour le critère IMSE. On établira aussi, sous certaines conditions de régularité, que la vitesse de convergence uniforme des polygones des fréquences est de l'ordre de $(\hat{\mathbf{n}}^{-1}\log(\hat{\mathbf{n}}))^{1/3}$.

Nous noterons x un point de R. La lettre C sera utilisée dans la suite pour remplacer toutes les constantes dont les valeurs ne sont pas importantes. La lettre D désignera un compact arbitraire de R.

On notera enfin : $\Psi_{\mathbf{n}} = \max(b; (\log\hat{\mathbf{n}}(\hat{\mathbf{n}}b)^{-1})^{1/2})$.

2. Préliminaires

Considérons une partition $\cdots < x_{-2} < x_{-1} < x_0 < x_1 < x_2 < \cdots$ de la droite réelle en intervalles $I_k = [(k-1)b, kb[$ de même largeur $b = b_{\mathbf{n}}$. Sans perte de généralité, on suppose que zéro est un des points de la partition. Considérons deux cellules adjacentes $I_0 = [-b, 0[$ et $I_1 = [0, b[$. Notons respectivement ν_0 and ν_1 le nombre d'observations situées dans ces deux cellules ($\nu_0 + \nu_1 = \hat{\mathbf{n}}$). Les valeurs de l'estimateur histogramme dans ces deux cellules sont respectivement : $f_0 = \nu_0\hat{\mathbf{n}}^{-1}b^{-1}$ et $f_1 = \nu_1\hat{\mathbf{n}}^{-1}b^{-1}$. Le polygone des fréquences est alors donné par :

$$f_{\hat{\mathbf{n}}}(x) = \left(\frac{1}{2} - \frac{x}{b}\right)f_0 + \left(\frac{1}{2} + \frac{x}{b}\right)f_1, \quad \text{pour } -\frac{b}{2} \leqslant x < \frac{b}{2}. \tag{2.1}$$

Nous supposons que b tend vers zéro quand $\mathbf{n} \to \infty$.

Notons, pour simplifier, $f_{\mathbf{j}|\mathbf{i}}$ la densité conditionnelle de $X_{\mathbf{j}}$ sachant $X_{\mathbf{i}}$. Nous aurons besoin, dans la suite, des hypothèses suivantes :

Hypothèse 1. Il existe une constante M_1 telle que, pour tout $\mathbf{i}$, $\mathbf{j}$, on ait :

$$\sup_{(x,y)\in R\times R} f_{\mathbf{j}|\mathbf{i}}(y|x) \leqslant M_1.$$

Hypothèse 2. f est deux fois continûment différentiable ; f'' est absolument continue par rapport à la mesure de Lebesgue sur R ; $f^{1/2} \in L^1$, $f' f^{-1/2} \in L^1$, $f^{(k)} \in L^2$ pour $k = 0, 1, 2, 3$.

3. Erreur moyenne quadratique intégrée et largeur de bande optimale

On définit la rugosité de la k-ième dérivée de f par :

$$R_k(f) = \int_{-\infty}^{+\infty} \left[f^{(k)}(x)\right]^2 \mathrm{d}x,$$

L'IMSE est la somme de deux termes : le carré du biais ponctuel intégré, et la variance ponctuelle intégrée.

Théorème 3.1. *Supposons que les Hypothèses* 1 *et* 2 *soient satisfaites, et que* $\varphi(k) = \mathrm{O}(k^{-\rho})$ *pour* $\rho > 2N + (3/2)$. *La valeur de la largeur de cellule minimisant asymptotiquement l'IMSE de l'estimateur polygone des fréquences vaut* :

$$b = b_{\mathbf{n}} = 2\left(\frac{15}{49R_2(f)}\right)^{1/5} \hat{\mathbf{n}}^{-1/5}$$

avec

$$IMSE = \frac{5}{12}\left(\frac{49R_2(f)}{15}\right)^{1/5} \hat{\mathbf{n}}^{-4/5} + \mathrm{O}\big(\hat{\mathbf{n}}^{-1}\big).$$

Ce théorème généralise les résultats de Scott [14], et de Carbon, Garel et Tran [4].

4. Convergence uniforme de l'estimateur polygone des fréquences

Hypothèse 2*. La densité f est C-lipschitzienne sur R.

Définissons :

$$\theta_1 = \frac{\rho + 3N}{\rho - 3N}, \qquad \theta_2 = \frac{N - \rho}{\rho - 3N}, \qquad \theta_3 = \frac{\rho + 3N}{\rho - (2\tilde{k} + 1)N}, \qquad \theta_4 = \frac{N - \rho}{\rho - (2\tilde{k} + 1)N}.$$

Théorème 4.1. *Supposons que* $\varphi(k) = \mathrm{O}(k^{-\rho})$ *pour* $\rho > 0$ *et que l'Hypothèse* 2^* *soit vérifiée.*

(i) *Si* (1.2) *est satisfaite et si* :

$$\hat{\mathbf{n}} b_{\mathbf{n}}^{\theta_1} (\log \hat{\mathbf{n}})^{\theta_2} \to \infty, \tag{4.1}$$

(ii) *ou si* (1.3) *est satisfaite et si* :

$$\hat{\mathbf{n}} b_{\mathbf{n}}^{\theta_3} (\log \hat{\mathbf{n}})^{\theta_4} \to \infty, \tag{4.2}$$

alors :

$$\sup_{x\in D} \left| f_{\hat{\mathbf{n}}}(x) - f(x) \right| = \mathrm{O}(\Psi_{\mathbf{n}}) \quad \textit{en probabilité.}$$

Exemple 4.1.

(i) Prenons $b = C\hat{\mathbf{n}}^{-1/5}$ où b est la largeur optimale trouvée dans la partie 3. Alors (4.1) est satisfaite si $\rho > 9N/2$, et (4.2) est satisfaite si $\rho > 2N + (5N\tilde{k})/2$.

(ii) Prenons $b = (\hat{\mathbf{n}}^{-1}\log\hat{\mathbf{n}})^{1/3}$. Alors $\Psi_{\mathbf{n}} = (\hat{\mathbf{n}}^{-1}\log\hat{\mathbf{n}})^{1/3}$, ce qui est la vitesse optimale dans le cas *i.i.d.* pour $N = 1$. Alors (4.1) est satisfaite si $\rho > 6N$, et (4.2) est satisfaite si $\rho > 3N + N\tilde{k}$.

5. Vitesse de convergence p.s. de $f_{\hat{\mathbf{n}}}$

Soit ε un nombre positif arbitrairement petit et soit $g(\mathbf{n}) = \prod_{i=1}^{N}(\log n_i)(\log\log n_i)^{1+\varepsilon}$, une fonction définie pour tous les $n_i > 2$ pour tout $i = 1, \ldots, N$.

Définissons :

$$\theta_1^* = \frac{\rho + 3N}{\rho - 5N}, \qquad \theta_2^* = \frac{N - \rho}{\rho - 5N}, \qquad \theta_3^* = \frac{\rho + 3N}{\rho - (2\tilde{k} + 3)N}, \qquad \theta_4 = \frac{N - \rho}{\rho - (2\tilde{k} + 3)N}.$$

Théorème 5.1. *Supposons $\varphi(k) = \mathrm{O}(k^{-\rho})$ avec $\rho > 0$ et que l'Hypothèse 2^* soit vérifiée.*

(i) *Si* (1.2) *est satisfaite et si* :

$$\hat{\mathbf{n}} b_{\mathbf{n}}^{\theta_1^*} (\log\hat{\mathbf{n}})^{\theta_2^*} \big(g(\mathbf{n})\big)^{-2N/(\rho - 5N)} \to \infty, \tag{5.1}$$

(ii) *ou si* (1.3) *est satisfaite et si* :

$$\hat{\mathbf{n}} b_{\mathbf{n}}^{\theta_3^*} (\log\hat{\mathbf{n}})^{\theta_4^*} \big(g(\mathbf{n})\big)^{-2N/(\rho - (2\tilde{k}+3)N} \to \infty, \tag{5.2}$$

alors :

$$\sup_{x\in D} \big| f_{\hat{\mathbf{n}}}(x) - f(x) \big| = \mathrm{O}(\Psi_{\mathbf{n}}) \quad a.s. \tag{5.3}$$

Pour une preuve des Théorèmes 4.1 et 5.1, on pourra consulter Carbon [3].

Références

[1] G. Biau, Spatial kernel density estimation, Math. Methods Statist. 12 (2003) 371–390.
[2] E. Bolthausen, On the central limit theorem for stationary random fields, Ann. Probab. 10 (1982) 1047–1050.
[3] M. Carbon, Frequency Polygons for Random Fields. Lucarne bleue, CREST, 2005-04, 2005.
[4] M. Carbon, B. Garel, L.T. Tran, Frequency polygon for weakly dependent processes, Statist. Probab. Lett. 33 (1997) 1–13.
[5] M. Carbon, M. Hallin, L.T. Tran, Kernel density estimation for random fields: the L_1 theory, J. Nonparametr. Statist. 6 (1996) 157–170.
[6] M. Carbon, L.T. Tran, B. Wu, Kernel density estimation for random fields (Density estimation for random fields), Statist. Probab. Lett. 36 (1997) 115–125.
[7] X. Guyon, Estimation d'un champ par pseudo-vraisemblance conditionnelle : Etude asymptotique et application au cas Markovien, in : Proc. 6th Franco-Belgian Meeting of Statisticians, 1987.
[8] X. Guyon, S. Richardson, Vitesse de convergence du théorème de la limite centrale pour des champs faiblement dépendants, Z. Wahrsch. Verw. Gebiete (1984) 297–314.
[9] M. Hallin, Z. Lu, L.T. Tran, Density estimation for spatial linear processes, Bernouilli 7 (2001) 657–688.
[10] M. Hallin, Z. Lu, L.T. Tran, Kernel density estimation for spatial linear processes: the L^1-theory, J. Multivariate Anal. 88 (2004) 61–75.
[11] M. Hallin, Z. Lu, L.T. Tran, Local linear spatial regression, Ann. Statist. 32 (6) (2004) 2469–2500.
[12] C.C. Neaderhouser, Convergence of block spins defined on random fields, J. Statist. Phys. 22 (1980) 673–684.
[13] M. Rosenblatt, Stationary Sequences and Random Fields, Birkhäuser, Boston, 1985.
[14] D.W. Scott, Frequency polygons, theory and applications, J. Amer. Statist. Assoc. 80 (1985) 348–354.
[15] C.J. Stone, Optimal uniform rate of convergence for non parametric estimators of a density function and its derivative, in: M.H. Revzi, J.S. Rustagi, D. Siegmund (Eds.), Recent Advances in Statistics: Papers in Honor of H. Chernoff, 1983, pp. 393–406.
[16] H. Takahata, On the rates in the central limit theorem for weakly dependent random fields, Z. Wahrsch. Verw. Gebiete 62 (1983) 477–480.
[17] L.T. Tran, Kernel density estimation on random fields, J. Multivariate Anal. 34 (1990) 37–53.
[18] L.T. Tran, S. Yakowitz, Nearest neighbor estimators for random fields, J. Multivariate Anal. 44 (1993) 23–46.