C. R. Acad. Sci. Paris, t. 332, Série I, p. 1–6, 2001
Statistique/*Statistics*
(Probabilités/*Probability Theory*)

Iterated Barron density estimators

Alain BERLINET, Gérard BIAU

Laboratoire de Probabilités et Statistique, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier cedex 5, France
Courriel : biau@ensam.inra.fr

(Reçu le 22 novembre 2000, accepté le 6 février 2001)

Abstract. Barron estimators are modified histograms depending on some reference probability density. They have been proved to have good consistency properties according to several information theoretic criteria. In this Note, Barron estimators are viewed as dynamical systems in a functional space with the reference density as initial state. Properties of stationary densities of these systems are investigated. It is then proved, under some assumptions, that the trajectories become almost surely stationary in a finite number of iterations.

Estimateurs de Barron itérés

Résumé. *Les estimateurs de Barron sont des histogrammes modifiés construits à partir d'une densité de probabilité dite de référence. Ces estimateurs sont connus pour posséder de bonnes propriétés de convergence au sens des critères de la théorie de l'information. Dans cette Note, nous envisageons les estimateurs de Barron comme des systèmes dynamiques fonctionnels pour lesquels la densité de référence joue le rôle de la condition initiale. Nous étudions les propriétés des densités stationnaires de ces systèmes et nous démontrons, sous certaines hypothèses, que les trajectoires deviennent presque sûrement stationnaires après un nombre fini d'itérations.*

Version française abrégée

Plaçons-nous sur l'espace mesurable $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ et désignons par f et f^* deux densités de probabilité définies par rapport à la mesure de Lebesgue λ. L'*entropie relative* (ou *information de Kullback-Leibler* ou I-*divergence*) de f par rapport à f^* est définie par

$$D(f, f^*) = \begin{cases} \int_{\mathbb{R}} f \ln(f/f^*)\,\mathrm{d}\lambda & \text{si } f \ll f^* \\ +\infty & \text{sinon.} \end{cases} \tag{1}$$

Dans de nombreux domaines d'application de la statistique (compression de données, réseaux de télécommunications, problèmes de classification, réseaux de neurones...), la convergence en entropie relative joue un rôle prépondérant. Un certain nombre d'auteurs se sont donc naturellement intéressés au comportement d'estimateurs de la densité au sens de ce critère sous des hypothèses minimales. Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires réelles indépendantes et de même

Note présentée par Paul DEHEUVELS.

loi à densité inconnue f. Barron, Györfi et van der Meulen [3] ont montré qu'il était possible de construire un estimateur f_n de f basé sur $X_1, \ldots, X_n$ convergeant presque sûrement en entropie relative et en entropie relative moyenne, à condition qu'il existe une densité de probabilité g (dite *de référence*) telle que $D(f, g) < \infty$. Soulignons au passage que cette dernière condition implique, d'après (1), l'absolue continuité de f par rapport à g. Cet estimateur f_n a été initialement introduit par Barron [2]. Il est défini comme histogramme modifié de la manière suivante :

- soit $\{m_n\}_{n \geq 2}$ une suite d'entiers tels que $1 < m_n \leq n$, soit $h_n = 1/m_n$ et soit $a_n = 1/(nh_n+1)$;
- soit g la densité de référence (connue) associée à la loi de probabilité ν ;
- soit $\{\mathcal{P}_n\}_{n \geq 2} = \{A_{n1}, A_{n2}, \ldots, A_{nm_n}\}_{n \geq 2}$ une suite de partitions de $\mathbb{R}$ telles que $\nu(A_{ni}) = h_n$, $i = 1, \ldots, m_n$;

Alors, en notant, pour $x \in \mathbb{R}$, $A_n(x)$ l'élément de $\mathcal{P}_n$ qui contient x et μ_n la mesure empirique construite à partir de l'échantillon $X_1, \ldots, X_n$, on a :

$$f_n(x) = \left[(1 - a_n)\frac{\mu_n\big(A_n(x)\big)}{h_n} + a_n\right] g(x) = \frac{n\mu_n\big(A_n(x)\big) + 1}{nh_n + 1}\, g(x).$$

La première des deux égalités précédentes montre que l'estimateur f_n (dit *estimateur de Barron*) est "voisin" de l'estimateur histogramme usuel $\hat{f}_n(x) = \big[\mu_n\big(A_n(x)\big)\, g(x)\big]/h_n$, relatif à la mesure ν. Pour une analyse des propriétés de l'estimateur de Barron f_n, on pourra se reporter à Berlinet, Györfi et van der Meulen [5], Berlinet, Vajda et van der Meulen [6] et à Györfi, Liese, Vajda et van der Meulen [9]. L'estimateur de Barron est en fait construit comme une déformation par morceaux de la densité de référence g. Cet estimateur peut être vu, à échantillon fixé, comme une fonctionnelle admettant en entrée la densité de référence g et fournissant en sortie une nouvelle densité de probabilité, disons Bg, estimateur de la densité inconnue f. Puisque Bg est une densité de probabilité, nous pouvons alors la considérer comme densité de référence de l'estimateur de Barron : cela conduit à un nouvel estimateur de f, noté $B^2 g$. L'estimateur $B^2 g$ peut alors devenir densité de référence à son tour et ainsi de suite... Ce processus itératif nous a naturellement amenés à considérer le système dynamique à espace d'état fonctionnel défini de la façon suivante :

$$\begin{array}{rccc} B: & \mathcal{D} & \to & \mathcal{D} \\ & g & \mapsto & Bg, \end{array}$$

où $\mathcal{D}$ représente l'ensemble des densités de probabilité sur $\mathbb{R}$. Deux questions se posent alors.

1) Le système dynamique B est-il ou non stationnaire ?
2) Quelles sont les propriétés statistiques des éventuels points fixes de B ?

Le but de la présente Note est de répondre à la question 1). La question 2) est actuellement à l'étude en liaison avec les résultats de Berlinet et Brunel [4] sur la validation croisée. Des résultats de simulation figurent dans Biau [7].

Nous commençons par une remarque élémentaire : une densité g est stationnaire pour B si et seulement si $\mu_n(A_{ni}) = h_n$ pour $i = 1, \ldots, m_n$. Ceci ne peut se produire que si m_n (le nombre de classes) est un diviseur de n (le nombre d'observations), ce que nous supposerons dans la suite. En pratique, on pourra éventuellement ne considérer que $m_n[n/m_n]$ données $\big([.]$ désigne la partie entière d'un nombre réel$\big)$. Ceci n'est pas dramatique dans la mesure où m_n est très petit par rapport à n. Plus précisément, la théorie asymptotique fait ressortir que m_n doit être de l'ordre de $n^{1/3}$ (Barron, Györfi et van der Meulen [3]). Sous cette condition de divisibilité, nous montrons que la trajectoire $\{B^p g\}_{p \geq 0}$ générée à partir d'une densité initiale g dont la fonction de répartition est inversible devient presque sûrement stationnaire après un nombre fini d'itérations. Nous introduisons pour ce faire les quantiles d'ordre ih_n $(i = 1, \ldots, m_n - 1)$ de la densité $B^p g$, soit $q_1^p, q_2^p, \ldots, q_{m_n-1}^p$, et nous posons, lorsque $m_n \geq 3$, les hypothèses suivantes $\big((X_{(1)}, \ldots, X_{(n)})$ désigne le vecteur des statistiques d'ordre des observations $X_1, \ldots, X_n\big)$:

HYPOTHÈSES (HLi) ($i = 2, \ldots, m_n - 1$). – *Pour p assez grand,* $q_i^p > X_{(\frac{in}{m_n}-1)}$.

Le théorème obtenu s'énonce alors ainsi :

THÉORÈME 1. – *Supposons que m_n divise n et que l'hypothèse* (H)=(HL2)+...+(HL$m_n - 1$) *soit vérifiée dès que $m_n \geq 3$. Alors chaque suite $\{q_i^p\}_{p\geq 0}$ ($i = 1, \ldots, m_n - 1$) devient presque sûrement stationnaire après un nombre fini d'itérations. Chaque suite atteint presque sûrement un point limite qui appartient à l'intervalle quantile empirique d'ordre ih_n, c'est-à-dire à l'intervalle* $[X_{(\frac{in}{m_n})}, X_{(\frac{in}{m_n}+1)}[$.

De ce dernier théorème, on déduit que la trajectoire $\{B^p g\}_{p\geq 0}$ devient presque sûrement stationnaire après un nombre fini d'itérations. L'étude des propriétés statistiques (vitesses de convergence lorsque n tend vers l'infini) de la densité limite ainsi que la détermination de conditions assurant l'hypothèse (H) font l'objet d'un travail en cours.

1. Introduction

Let f and f^* be two probability densities with respect to the Lebesgue measure λ on the Borel line $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$. The *information divergence* (also called I-*divergence* or *relative entropy* or *Kullback-Leibler information number*) $D(f, f^*)$ of f with respect to f^*, defined by (1), is widely used in many fields of statistics (data compression, telecommunication networks, classification, pattern recognition, neural networks...) This motivates the construction of nonparametric estimators consistent in information divergence. Suppose that we observe independent real random variables $X_1, \ldots, X_n$ with common unknown density f. Barron, Györfi and van der Meulen [3] showed that one can construct an estimator of the density f, say f_n, consistent in information divergence and expected information divergence, *i.e.*

$$\lim_{n\to\infty} \mathbf{E}D(f, f_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} D(f, f_n) = 0 \text{ a.s.}$$

This density estimator, first introduced in Barron [2], is defined as follows:

- denote by g a known probability density on $\mathbb{R}$ and by ν the associated distribution;
- define a sequence of integers $\{m_n\}_{n\geq 2}$ such that $1 < m_n \leq n$ and let $h_n = 1/m_n$;
- introduce a sequence of partitions $\{\mathcal{P}_n\}_{n\geq 2} = \{A_{n1}, \ldots, A_{nm_n}\}_{n\geq 2}$ such that the A_{ni}'s are intervals with $\nu(A_{ni}) = h_n$ for $i = 1, \ldots, m_n$;
- finally consider, for $a_n = 1/(nh_n + 1)$, the following density estimator f_n:

$$f_n(x) = \left[(1 - a_n)\,\frac{\mu_n\big(A_n(x)\big)}{h_n} + a_n\right] g(x) = \frac{n\mu_n\big(A_n(x)\big) + 1}{nh_n + 1}\, g(x), \tag{2}$$

where μ_n stands for the empirical measure associated with the sample $X_1, \ldots, X_n$ and $A_n(x)$ equals A_{ni} if $x \in A_{ni}$. The estimator defined in (2), usually called *Barron estimator*, is a mixture of a histogram-type density estimator and of the reference density g itself. Barron, Györfi and van der Meulen [3] proved that this estimator is consistent both in information divergence and expected information divergence, provided $D(f, g) < \infty$, $h_n \to 0$ and $nh_n \to \infty$ as n grows to infinity. It is worth noting that the condition $D(f, g) < \infty$ implies that f is absolutely continuous with respect to g. Observe also that the information divergence may be infinite with positive probability for the standard histogram. For further results on Barron estimators, we refer to Berlinet, Györfi and van der Meulen [5], Györfi, Liese, Vajda and van der Meulen [9] and Berlinet, Vajda and van der Meulen [6] who extend the I-divergence consistency properties of Barron estimators respectively to the χ^2-divergence and to more general ϕ-divergences.

In the present Note, Barron density estimators are viewed as dynamical systems in a functional space with initial state the reference probability density g. More precisely, we address the following question: what happens if the reference density in (2) is replaced in a second step by the Barron estimator itself? And, more generally, what happens if we iterate the process? In Section 2, some elementary properties of stationary densities of this dynamical system are studied. In Section 3 we show, under suitable conditions, that the trajectories of the system become almost surely stationary in a finite number of iterations.

2. Iteration of Barron estimators

The Barron estimation process can be seen as a system that, given an input probability density g, supplies an output probability density $f_n = Bg$ (the dependence of B upon n being understood). Thus, if we denote by $\mathcal{D}$ the set of probability densities on $\mathbb{R}$, we are led to consider the dynamical system $B : \mathcal{D} \to \mathcal{D}$ which associates to each g $(= B^0 g)$ in $\mathcal{D}$ a trajectory $\{B^p g\}_{p\geq 0}$ of densities. Following Lasota and Mackey [10], we call any element g of $\mathcal{D}$ satisfying the equation $Bg = g$ a *stationary* (*probability*) *density* of B. Then two questions arise.

1) Is the dynamical system B stationary?
2) What are the statistical properties of the possible stationary densities of B?

The aim of the present Note is to answer question 1). Question 2) is still under study, in connection with the results obtained by Berlinet and Brunel [4]. Simulation results are available in Biau [7].

We start with an elementary remark: a density g is stationary for B if and only if $\mu_n(A_{ni}) = h_n$ for $i = 1, \ldots, m_n$. This can only happen if m_n (the number of classes) is a divisor of n (the number of observations), which we will assume in the sequel. In practice, we may restrict ourselves to $m_n[n/m_n]$ data ([.] stands for the integer part of a real number). This does not harm to much since m_n is very small with respect to n. More precisely, the asymptotic theory reveals that m_n should be of order $n^{1/3}$ (Barron, Györfi and van der Meulen [3]). Also, a slight modification of the estimate could allow to remove this divisibility condition. Assuming that n is a multiple of m_n, it is easy to construct many stationary probability densities of B. The most "simple" are the following ones, called *Random Partition* or *Statistically Equivalent Blocks* (S.E.B.) *estimators* (Anderson [1], Bosq and Lecoutre [8], Tukey [11]):

DEFINITION 1. – Let $t_n = [n/m_n]$ and let $(X_{(1)}, \ldots, X_{(n)})$ be the order statistics vector of the sample $X_1, \ldots, X_n$. Construct a partition of $[X_{(1)}, X_{(n)}]$ into m_n intervals A_{ni} $(i = 1, \ldots, m_n)$ each of them containing exactly the same number of observations up to one (either t_n or $t_n + 1$). Then a *Random Partition* or *Statistically Equivalent Blocks* (S.E.B.) estimator b_n is defined as follows:

$$b_n(x) = \frac{t_n(x)}{n\lambda(A_n(x))} \text{ if } x \in [X_{(1)}, X_{(n)}],\ \ b_n(x) = 0 \text{ otherwise.}$$

PROPOSITION 1. – *Assume that m_n is a divisor of n. Then the S.E.B. estimator b_n is a stationary density of B.*

Proof. – Let us denote by $]-\infty, q_1],]q_1, q_2], \ldots,]q_{m_n-1}, +\infty[$ the classes defining the estimator b_n associated with the probability distribution ν_n. By construction:

$$\nu_n\Big(]-\infty, q_1]\Big) = \nu_n\Big(]q_1, q_2]\Big) = \ldots = \nu_n\Big(]q_{m_n-1}, +\infty[\Big) = 1/m_n = h_n.$$

It follows that $q_1, \ldots, q_{m_n-1}$ are the quantiles of order ih_n $(i = 1, \ldots, m_n-1)$ of b_n. To prove that b_n is stationary, it remains to show that $\mu_n(]-\infty, q_1]) = \mu_n(]q_1, q_2]) = \ldots = \mu_n(]q_{m_n-1}, +\infty[) = h_n$. This is true, by the very construction of the S.E.B. estimator. □

3. Asymptotic results

In this section, to avoid useless technicalities, we assume that m_n is a divisor of n and that the distribution function of the initial density g is invertible. We give results about the asymptotics of the trajectory $\{B^p g\}_{p\geq 0}$. To this end, we denote, for each p in $\mathbb{N}$:

– $q_1^p, q_2^p, \ldots, q_{m_n-1}^p$ the quantiles of order ih_n $(i = 1, \ldots, m_n - 1)$ of the density $B^p g$ associated with the partition $\{]-\infty, q_1^p],]q_1^p, q_2^p], \ldots,]q_{m_n-1}^p, +\infty[\} = \{A_{n1}^p, \ldots, A_{nm_n}^p\}$;

– $\alpha_1^p, \ldots, \alpha_{m_n}^p$ the *Barron coefficients* associated with the partition above, namely:

$$\alpha_i^p = \frac{n\mu_n(A_{ni}^p) + 1}{nh_n + 1}, \quad i = 1, \ldots, m_n.$$

As enlightened by equality (2), the study of the sequence $\{B^p g\}_{p\geq 0}$ reduces to the study of the two sequences $\left\{(q_1^p, \ldots, q_{m_n-1}^p)\right\}_{p\geq 0}$ and $\left\{(\alpha_1^p, \ldots, \alpha_{m_n}^p)\right\}_{p\geq 0}$. Whenever $m_n \geq 3$, we need the following hypotheses:

HYPOTHESES (HLi) $(i = 2, \ldots, m_n - 1)$. – *For large enough* p, $q_i^p > X_{(\frac{in}{m_n}-1)}$.

Assumption (HLi) means that, for large enough p, each value of the sequence $\{q_i^p\}_{p\geq 0}$ falling to the left of $X_{(\frac{in}{m_n})}$ is actually "close" to this observation.

THEOREM 1. – *Assume that* m_n *divides* n *and that the hypothesis* (H)=(HL2)+…+(HLm_n-1) *holds whenever* $m_n \geq 3$. *Then each sequence* $\{q_i^p\}_{p\geq 0}$ $(i = 1, \ldots, m_n - 1)$ *becomes almost surely stationary in a finite number of iterations. Each sequence reaches almost surely a limit point belonging to the empirical quantile interval of order* ih_n, *namely* $[X_{(\frac{in}{m_n})}, X_{(\frac{in}{m_n}+1)}[$.

Sketch of proof. – We claim that, for each quantile sequence $\{q_i^p\}_{p\geq 0}$, four dynamics have to be considered. We point out these dynamics with the abbreviations **R**(**i**) (**R** means "Right", **i** means the i-th quantile), **I**(**i**) (**I** means "Inside"), **L**(**i**) (**L** means "Left") and **O**(**i**) (**O** means "Oscillation").

– *First dynamics*: **R**(**i**). Description: for large enough p, the sequence $\{q_i^p\}_{p\geq 0}$ remains in the interval $[X_{(\frac{in}{m_n}+1)}, +\infty[$.

– *Second dynamics*: **I**(**i**). Description: there exists p_0 in $\mathbb{N}$ such that $q_i^{p_0}$ belongs to the interval $[X_{(\frac{in}{m_n})}, X_{(\frac{in}{m_n}+1)}[$. In that case, one shows that $q_i^p = q_i^{p_0}$ for all $p \geq p_0$: the quantile is stationary.

– *Third dynamics*: **L**(**i**). Description: for large enough p, the sequence $\{q_i^p\}_{p\geq 0}$ remains in the interval $]-\infty, X_{(\frac{in}{m_n})}[$.

– *Fourth dynamics*: **O**(**i**). Description: the sequence $\{q_i^p\}_{p\geq 0}$ takes infinitely many values in both intervals $]-\infty, X_{(\frac{in}{m_n})}[$ and $[X_{(\frac{in}{m_n}+1)}, +\infty[$.

A tedious analysis of the four dynamics above and their possible interactions, together with Helly's selection theorem, show that, under the assumption (H), the dynamics $\mathbf{I(1)}, \ldots, \mathbf{I(m_n-1)}$ is almost surely the only one possible for the sequences $\{q_i^p\}_{p\geq 0}$, $i = 1, \ldots, m_n - 1$. A detailed proof can be found in Biau [7]. □

The following corollary is easily deduced from Theorem 1:

COROLLARY 1. – *Under the same conditions as in Theorem* 1:
– *Each sequence* $\{\alpha_i^p\}_{p\geq 0}$ $(i = 1, \ldots, m_n)$ reaches almost surely the value 1 in a finite number of iterations.
– The sequence of probability densities $\{B^p g\}_{p\geq 0}$ becomes almost surely stationary in a finite number of iterations.

We developed an algorithm automatically generating the sequence of quantiles. The code was implemented with the mathematical macro-language *Matlab* and is available on request. Details and results of simulations are to be found in Biau [7]. Questions regarding statistical properties of the stationary density (rates of convergence as n tends to infinity) as well as conditions ensuring (H) are currently under study, in connection with recent optimality results obtained in [4].

References

[1] Anderson T.W., Some nonparametric multivariate procedures based on statistically equivalent blocks, in : Multivariate Analysis, Academic, New York, 1966, pp. 5–27.

[2] Barron A.R., The convergence in information of probability density estimators, in : Proceedings of the International Symposium of IEEE on Information Theory, Kobe, Japan, June 19–24, 1988.

[3] Barron A.R., Györfi L., van der Meulen E., Distribution estimation consistent in total variation and in two types of information divergence, IEEE Transactions on Information Theory 38 (1992) 1437–1454.

[4] Berlinet A., Brunel E., Choix optimal du nombre de classes pour l'estimateur de Barron de la densité, Comptes Rendus de l'Académie des Sciences de Paris 331 (2000) 713–716.

[5] Berlinet A., Györfi L., van der Meulen E., Asymptotic normality of relative entropy in multivariate density estimation, Publications de l'Institut de Statistique de l'Université de Paris 41 (1997) 3–27.

[6] Berlinet A., Vajda I., van der Meulen E., About the asymptotic accuracy of Barron density estimates, IEEE Transactions on Information Theory 44 (1998) 999–1009.

[7] Biau G., Méthodes itératives en estimation fonctionnelle et systèmes dynamiques, Thèse de doctorat, Université Montpellier II, 2000.

[8] Bosq D., Lecoutre J.P., Théorie de l'estimation fonctionnelle, Economica, Paris, 1987.

[9] Györfi L., Liese F., Vajda I., van der Meulen E., Distribution estimates consistent in χ^2-divergence, Statistics 32 (1998) 31–57.

[10] Lasota A., Mackey M.C., Chaos, fractals, and noise – Stochastic aspects of dynamics, Springer-Verlag, New York, 1994.

[11] Tukey J.W., Nonparametric estimation II. Statistically equivalent blocks and tolerance regions. The continuous case, The Annals of Mathematical Statistics 18 (1947) 529–539.