Estimation rapide non-paramétrique de la densité de la distribution d'entropie maximale pour les statistiques d'ordre

Richard Fischer [1] & Cristina Butucea [2] & Jean-François Delmas [3] & Anne Dutfoy [4]

[1] *EDF Recherche & Développement, Management des Risques Industriels, 1 Av. Général de Gaulle, 92141 Clamart; richard.fischer@edf.fr*

[2] *LAMA, Université Paris-Est - Marne-la-Vallée, 5 Bd. Descartes, 77454 Champs-sur-Marne; cristina.butucea@univ-mlv.fr*

[3] *CERMICS, Ecole Nationale des Ponts et Chaussées, 6 & 8 Av. Pascal, 77455 Champs-sur-Marne; delmas@cermics.enpc.fr*

[4] *EDF Recherche & Développement, Management des Risques Industriels, 1 Av. Général de Gaulle, 92141 Clamart; anne.dutfoy@edf.fr*

Résumé. L'objectif de cette communication est de présenter une méthode pour estimer, d'une façon non-paramétrique, la densité de la distribution d'entropie maximale des statistiques d'ordre ayant des marginales fixées. Ces densités, dont le support est inclus dans l'ensemble $S = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d; x_1 \leq x_2 \leq \ldots, \leq x_d\}$, admettent une forme produit. On souhaite estimer, à partir d'un échantillon i.i.d., une densité qui appartient à cette famille de lois. Étant donné la forme et le support spécial, nous proposons un modèle log-additif basé sur des séries de polynômes quasi-orthogonaux spécialement conçus pour ce problème. L'intérêt de cette méthode est qu'elle nous donne une véritable fonction de densité de statistiques d'ordre qu'on pourra utiliser lors des simulations de type Monte-Carlo. Nous démontrons que, si le logarithme de la densité appartient à une classe de type Sobolev anisotropique, on peut décomposer notre problème d'estimation d'une densité d-dimensionnelle à d problèmes univariés, on peut alors retrouver la vitesse de convergence univariée optimale dans le sens minimax pour une classe Sobolev de log-densités.

Mots-clés. statistiques d'ordre, estimation non-paramétrique de densité, modèle log-additif, classe Sobolev

Abstract. The objective of this communication is to present a nonparametric method for estimating the density of the maximum entropy distribution of order statistics with fixed marginals. In particular, these densites have a product form with support included in the set $S = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d; x_1 \leq x_2 \leq \ldots, \leq x_d\}$. Our aim is to estimate a density belonging to this family based on an i.i.d. sample with a fast convergence rate. We propose a log-additive model based on series of quasi-orthogonal polynomials specially designed to suit this particular structure. The practical importance of this method is that we obtain a real density function of order statistics which we can use in Monte Carlo simulations. We show that if the logarithm of the density belongs to an anisotropic Sobolev class, we can

decompose our d-dimensional problem into d one-dimensional ones, thus achieving the optimal minimax univariate convergence rate for nonparametric estimations of densities whose logarithm belongs to a Sobolev class.

Keywords. Order statistics, nonparametric density estimation, log-additive model, Sobolev class

1 Distributions d'entropie maximale des statistiques d'ordre

Le problème initial est de donner la distribution d'entropie maximale d'un vecteur de statistiques d'ordre de marginales fixées. C'est à dire le vecteur $X = (X_1, \ldots, X_d)$ vérifie:

$$X_1 \leq X_2 \leq \ldots \leq X_d \quad \text{p.s.}$$

et la fonction de répartition de chaque marginale est fixée et notée par F_i, $1 \leq i \leq d$. Parmi les distributions admissibles, on cherche celle qui maximise l'entropie de Shannon donnée par, pour une variable aléatoire X de dimension d:

$$H(X) = \begin{cases} -\int_{\mathbb{R}^d} f_X(x) \log\left(f_X(x)\right) \, dx, & \text{si } X \text{ a une densité } f_X, \\ -\infty & \text{sinon.} \end{cases}$$

A l'aide de la théorie des copules, et de l'optimisation sous contraintes de dimension infinie développée par Borwein et al. (1994), nous avons obtenu le résultat suivant: si les marginales F_i sont absolument continues de densité f_i, $1 \leq i \leq d$, et vérifient

C1 $H(X_i) < +\infty$, $1 \leq i \leq d$,

C2 $F_i > F_{i+1}$ sur l'ensemble $\{1 > F_i, F_{i+1} > 0\}$, $1 \leq i \leq d-1$,

C3 $\sum_{i=2}^{d} f_i(s) \left|\log(F_{i-1}(s) - F_i(s))\right| \, ds < +\infty$,

alors il existe une unique distribution de statistiques d'ordre ayant comme marginales F_i, $1 \leq i \leq d$, qui maximise l'entropie. Cette distribution admet une densité qui est donnée par:

$$f^*(x) = f_1(x_1) \prod_{i=2}^{d} \frac{f_i(x_i)}{F_{i-1}\left(x_i\right) - F_i(x_i)} \exp\left(- \int_{x_{i-1}}^{x_i} \frac{f_i(s)}{F_{i-1}(s) - F_i(s)} \, ds \right) \mathbf{1}_S(x),$$

avec $S = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d; x_1 \leq x_2 \leq \ldots, \leq x_d\}$. En particulier, cette densité peut s'exprimer comme:

$$f^*(x) = \exp\left(\sum_{i=1}^{d} \ell_i(x_i) - \mathrm{a}_0 \right) \mathbf{1}_S(x), \tag{1}$$

où ℓ_i sont des fonctions mesurables et a_0 une constante de normalisation. F^* est donc de forme produit sur le simplexe S.

2 Modèle log-linéaire

Dans la suite, on suppose que l'on possède un échantillon i.i.d. $(X^j, j \in \mathbb{N})$ issu d'une distribution dont la densité f est de la forme (1). On se restreint à étudier des distributions dont le support est $\triangle = [0,1]^d \cap S$. On cherche à estimer la densité par une méthode non-paramétrique qui tient compte de sa structure spéciale et entraine une convergence plus rapide que pour l'estimation des densités multidimensionnelles. Étant donné la forme produit, nous proposons un modèle additif pour le logarithme de la densité. Pour chaque $1 \leq i \leq d$, soit $(\phi_{i,k}, k \in \mathbb{N})$ une suite de polynômes univariés orthonormes par rapport à la fonction de poids q_i, qui est la marginale unidimensionnelle de la mesure de Lebesgue sur $\triangle$. Nous construisons une telle suite à partir des polynômes de Jacobi. Il faut noter que l'ensemble des polynômes $(\phi_{i,k}; 1 \leq i \leq d, k \in \mathbb{N})$ n'est pas orthonormale par rapport à la mesure de Lebesgue sur $\triangle$, d'où la notion quasi-orthogonale. On estime f par f_θ donnée par, :

$$f_\theta(x) = \exp\left(\sum_{i=1}^{d}\sum_{k=1}^{m_i} \theta_{i,k}\phi_{i,k}(x_i) - \psi(\theta)\right), \quad \text{pour } x = (x_1, \ldots, x_d) \in \triangle,$$

avec $\psi(\theta) = \log\left(\int_\triangle \exp\left(\sum_{i=1}^{d}\sum_{k=1}^{m_i} \theta_{i,k}\phi_{i,k}(x_i)\right) dx\right)$ la constante de normalisation. Ce modèle est une version multivariée du modèle présenté par Barron et Sheu (1991), mais différent par rapport à la généralisation multidimensionnelle donnée par Wu (2010). On estime les paramètres $(\theta_{i,k}; 1 \leq i \leq d, 1 \leq k \leq m_i)$, en maximisant la vraisemblance associée à l'échantillon $(X^j, 1 \leq j \leq n)$. L'estimateur $\hat{\theta}$ est obtenu en résolvant le système d'équations suivant:

$$\int_\triangle \phi_{i,k}(x_i) f_{\hat{\theta}}(x)\, dx = \frac{1}{n}\sum_{j=1}^{n} \phi_{i,k}(X_i^j),$$

pour tous $1 \leq i \leq d, 1 \leq k \leq m_i$. Nous remarquons que l'estimateur $f_{\hat{\theta}}$ est une véritable densité: elle est non-négative et s'intègre à un.

3 Résultats

On mesure la discrépance entre la vraie densité f et son estimateur $\hat{\theta}$ par la divergence de Kullback-Leibler $D\left(f \| f_{\hat{\theta}}\right)$ définie comme:

$$D\left(f \| f_{\hat{\theta}}\right) = \int_\triangle f \log\left(\frac{f}{f_{\hat{\theta}}}\right).$$

Supposons que les fonctions ℓ_i, $1 \leq i \leq d$ appartiennent à des classes de Sobolev $W_{r_i}^2$, $r_i \in \mathbb{N}$ et $r_i > d$. Nous démontrons que si on laisse $m_i = m_i(n)$, $1 \leq i \leq d$ tendre vers

l'infini avec n, tel que $(\sum_{i=1}^d m_i^{2d})(\sum_{i=1}^d m_i^{-2r_i})$ et $(\sum_{i=1}^d m_i^{2d})(\sum_{i=1}^d m_i/n)$ tendent vers 0, alors l'estimateur du maximum de vraisemblance existe avec grande probabilité et vérifie:

$$D\left(f\|f_{\hat{\theta}}\right) = O_p\left(\sum_{i=1}^{d}\left(m_i^{-2r_i} + \frac{m_i}{n}\right)\right).$$

En particulier, si on prend m_i de l'ordre $n^{1/(2r_i+1)}$, on obtient la vitesse $O_p(\sum_{i=1}^d n^{-2r_i/(2r_i+1)})$, qui est de plus de l'ordre $n^{-2\underline{r}/(2\underline{r}+1)}$, où $\underline{r} = \min(r_i)$. Ceci correspond à la vitesse optimale minimax pour l'estimation d'une densité univariée dont le logarithme appartient à l'espace de Sobolev $W_{\underline{r}}^2$ (cf. Barron et Yang (1999)). Comme les r_i varient dans chaque direction, l'estimateur est capable de s'adapter à des classes de densités de régularité anisotropiques. De même, la vitesse obtenue constitue un gain par rapport à la vitesse générale pour les classes de densités anisotropiques avec des paramètres de régularité r_i dans chaque direction, qui est de l'ordre de $n^{-2\tilde{r}/(2\tilde{r}+1)}$, avec $\tilde{r}$ défini via $1/\tilde{r} = \sum_{i=1}^d 1/r_i$ (cf. par exemple Birgé (1986)).

Bibliographie

[1] Barron, A. R. and Sheu, C.-H. (2000), Approximation of density functions by sequences of exponential families, *The Annals of Statistics*, 19(3), 1347–1369.

[2] Birgé, L. (1986), On estimating a density using Hellinger distance and some other strange facts, *Probability theory and related fields*, 71(2), 271–291.

[3] Borwein, J., Lewis, A. et Nussbaum, R. (1994), Entropy minimization, DAD problems, and doubly stochastic kernels, *Journal of Functional Analysis*, 123(2), 264–307.

[4] Wu, X. (2010), Exponential series estimator of multivariate densities, *Journal of Econometrics* 156(2), 354–366.

[5] Yang, Y. and Barron, A.R. (1999), Information-theoretic determination of minimax rates of convergence, *Annals of Statistics*, 27(5), 1564–1599.