Un algorithme rapide d'estimation des densités de probabilité basé sur le plug-in : Application à des données génétiques

M. TROUDI[1 2], S. SAOUDI[2] ET A. M. ALIMI[3]

[1]REGIM, Ecole Nationale des Ingénieurs de Sfax, Route de Gabès, Sfax, Tunisie

molkaghorbel@gmail.com, adel.alimi@enis.rnu.tn

[2]Département Système de Communication, ENSTB, 29279 Plouzané, France

samir.saoudi@enst-Bretagne.fr

Résumé - Nous proposons ici une version rapide de l'algorithme itératif plug-in permettant l'estimation itérative du pas optimal à partir du noyau optimal au sens de l'erreur quadratique moyenne intégrée (EQMI). A chaque itération, le principal paramètre intervenant dans le développement limité du EQMI est calculé analytiquement. Différentes distributions sont générées pour tester l'efficacité de l'algorithme proposé. Ces algorithmes sont ensuite appliqués à des données génétiques dans le but d'estimer la neutralité de populations berbères de Tunisie.

Abstract - Here, we propose a fast version of the iterative plug-in procedure for the optimal smoothing parameter of the Kernel probability density function (pdf) estimator. Such procedure considers in the same time the optimal bandwidth and the optimal Kernel in the mean of MISE criterion. For each time, we approximate analytically a factor J(f), which is linked to the second order derivative of the pdf. Different random variables with difficult distributions are generated in order to prove the efficiency of the proposed optimal estimator. So, these algorithms are applied to a Tunisian genetics data in order to give a better characterisation of neutrality population

1. Introduction

Par une vision probabiliste de la neutralité en génétique des populations, la transmission génétique aboutit à un modèle multinomial[2]. Les tests développés dans [3, 8] aboutissent au calcul d'un estimateur de la neutralité comme la statistique de Tajima ou celle de Fu. Pour évaluer leur significativité, des simulations de populations neutres de mêmes caractéristiques que les populations étudiées sont mises en œuvre en estimant les densités de probabilité (d.p)de la statistique D de Tajima ou F_s de Fu. La probabilité que la variable aléatoire générée (D ou F_s) soit inférieure à la valeur déduite de l'échantillon permet de conclure sur la neutralité de la population. La fiabilité des résultats obtenus est tributaire de la qualité de l'estimateur des (d.p). Pour plus de précision, nous sélectionnons une méthode d'estimation parmi les non paramétriques telles que l'histogramme, la méthode du noyau et ses variantes [6] [7] et celles basées sur les fonctions orthogonales [4].

Dans ce papier, nous focalisons notre étude sur les variantes de la méthode du noyau en approfondissant la question délicate de la recherche du pas optimal. L'état de l'art statistique permet de recenser de nombreuses méthodes développées dans ce sens. Les méthodes "Rule of thumb ", "cross-validation" et "plug-in" sont réputées par leur performance au sens de l'EQMI[1] [5].

Une version rapide de l'algorithme plug-in [5] d'estimation du pas optimal h_n basée sur une approximation itérative de l'intégrale de la dérivée seconde élevée au carré de la dp *J(f),* est proposée. Pour cela, une suite de $h_n^{(k)}$ est construite au fil des itérations, k étant le nombre d'itérations et n la taille de l'échantillon.

L'approximation de *J(f)* peut être obtenue en dérivant l'expression analytique de l'estimateur du noyau optimal. Cela permet de n'estimer la densité de probabilité qu'une seule fois alors que l'approximation numérique de *J(f)* implique son estimation à chaque itération.

Dans ce papier, nous rappelons les principaux théorèmes de convergence de la méthode du noyau au sens de l'EQMI puis nous décrivons l'algorithme présenté. Différentes distributions sont ensuite testées et une étude comparative entre l'algorithme itératif avec approche numérique de *J(f)* et celui avec approche analytique de *J(f)*, est menée. Enfin, son application à l'évaluation de la neutralité d'un échantillon extrait à partir d'une population berbère de Tunisie est mise en oeuvre.

2. Fondements théoriques

2.1. Rappels et notations

Soit l'estimateur à noyau développé par Parzen [6] :

$$\hat{f}_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right) \qquad (1)$$

où h_n est une suite de nombres positifs tendant vers zéro et K une densité de probabilité.

L'obtention de la convergence entre une densité de probabilité f et son estimée $\hat{f}$ est fonction du choix du noyau et du pas optimal h_n. La recherche de ces deux entités passe par la minimisation de l'erreur quadratique moyenne (EQM) ainsi que par la minimisation de l'erreur quadratique moyenne intégrée (EQMI). L'EQM tend vers 0 si nh_n tend vers l'infini lorsque n tend vers l'infini alors que l'EQMI exige une condition plus restrictive pour sa convergence : $n(h_n)^2$ doit tendre vers l'infini lorsque n tend vers l'infini.

2.2. Convergence du EQMI :

L'étude de la convergence en moyenne quadratique de $\hat{f}$ (EQM) montre que :

$$E\left[\left\| f_n - f \right\|_x^2\right] = A_n(x) + B_n(x) - C_n(x) \qquad (2)$$

avec

$$A_n(x) = \frac{1}{nh_n}\int_{-\infty}^{+\infty} K^2(u) f(x - uh_n)\, du$$

$$B_n(x) = \left[\int_{-\infty}^{+\infty} K(u)\{f(x - uh_n) - f(x)\} du\right]^2$$

$$C_n(x) = \frac{1}{n}\left[\int_{-\infty}^{+\infty} K(u) f(x - uh_n)\, du\right]^2$$

Suite à une étude asymptotique, la minimisation du EQMI pour une taille d'échantillon n fixée, est fonction des entités suivantes :

$$M(K) = \int_{-\infty}^{+\infty} K^2(u) \text{ et } J(f) = \int_{-\infty}^{+\infty} (f''(x))^2\, dx \qquad (3)$$

Avec f'' est la dérivée seconde de f.
La valeur de $\Delta(h_n)$ est minimale lorsque

$$h_n^* = n^{-\frac{1}{5}} . (J(f))^{-\frac{1}{5}} . (M(K))^{\frac{1}{5}} \qquad (4)$$

Cela donne une EQMI minimale de l'ordre de :

$$D^2(\hat{f}_n, f) = \frac{5}{4} n^{-\frac{4}{5}} (M(K))^{\frac{4}{5}} (J(f))^{\frac{1}{5}} \qquad (5)$$

3. Description de l'algorithme

3.1. Méthode itérative du noyau optimal

Nous rappelons dans ce paragraphe les étapes principales de l'algorithme itératif plug-in pour approcher le pas optimal dans la méthode du noyau avant de décrire une nouvelle approche analytique permettant de réduire la complexité de l'algorithme.

Etape 1: Détermination analytique de $M(K)$ (3).

Etape 2: Initialisation arbitraire de $J^{(0)}(f)$ afin de déterminer $h_n^{(0)}$, première valeur de h_n (4).

Etape 3: Estimation de la densité $f^{(0)}$ à partir $h_n^{(0)}$.

Etape 4: A la k^{ème} itération, déduction de $J^{(k)}(f)$ à partir de la densité $f^{(k-1)}$ (1). $h_n^{(k)}$ (4) est ensuite calculé et $f^{(k)}$ (1) ré-estimé. A chaque itération $J(f)$ est estimé numériquement permettant de déduire h_n et f.

Etape 5: Critère d'arrêt: $\left|h_n^{(k-1)} - h_n^{(k)}\right| < \varepsilon$.

3.2. Approximation analytique de $J(f)$ dans le cas du noyau optimal :

Une dérivation directe de l'expression analytique du noyau optimal permet de calculer analytiquement $J(f)$:

$$K(x) = \begin{cases} 0 \text{ if } |x| > \sqrt{5} \\ \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right) \text{ if } |x| \le \sqrt{5} \end{cases}$$

$$\hat{f}''(x) = \frac{1}{nh_n^3}\sum_{i=1}^{n} K''\left(\frac{x - x_i}{h_n}\right)$$

$$K''(x) = \begin{cases} 0 & si \quad |x| > \sqrt{5} \\ indéfini & si \quad |x| = \sqrt{5} \\ \frac{3\sqrt{5}}{50} & si \quad |x| < \sqrt{5} \end{cases}$$

Soit la fonction suivante $\beta(x)$ constante par intervalles et formant une partition sur la droite réelle :

$$\beta(x) = \left[\sum_{i}^{n} K''\left(\frac{x - X_i}{h_n}\right)\right]^2 = \left[\sum_{i \in A_n(x)} K''\left(\frac{x - X_i}{h_n}\right)\right]^2$$

avec $A_n(x)$ un sous ensemble d'entiers :

$$A_n(x) = \left\{0 \le i \le n; \frac{|x - X_i|}{h_N} \le \sqrt{5}\right\}$$

L'intégrale $J(f)$ est composée par la somme finie des dérivées secondes de la fonction noyau optimal. Le nombre de points non définis pour $\beta(x)$ étant fini, la contribution de ces points dans $J(f)$ est négligeable. Cela implique que :

$$J(f) = \int_{-\infty}^{+\infty}\left[\frac{1}{nh_n^2}\sum_{i=1}^{n} K''\left(\frac{x - x_i}{h_n}\right)\right]^2 dx$$

$$J(f) = \frac{9}{500}\frac{1}{n^2 h_n^6}\int_{-\infty}^{+\infty}\beta(x)dx$$

Les simulations montrent que la convergence est obtenue pour une puissance du paramètre h_n comprise entre 4 et 5. Cela se justifie par le fait que la dérivation de $J(f)$ s'apparente à la méthode d'approximation du noyau dont la variance a besoin d'être ajustée. Dans les simulations suivantes, nous considérons l'estimateur de $J(f)$ suivant :

$$\hat{J}(f) = \frac{9}{500}\frac{1}{n^2 h_n^{4.5}}\int_{-\infty}^{+\infty}\beta(x)dx \qquad (6)$$

3.3. Algorithme itératif du noyau optimal analytique

Les étapes de l'algorithme itératif du noyau optimal avec approche analytique de *J(f)* sont décrites ci-dessous:

Etape 1 : Détermination analytique de *M(K)* (3).

Etape 2 : Initialisation arbitraire de $J^{(0)}(f)$ afin de déterminer $h_n^{(0)}$, première valeur de h_n (4).

Etape 3 : À la k^{ème} itération, $J^{(k)}(f)$ est calculé directement à partir de l'échantillon X_i (7). $h_n^{(k)}$ (4) est ensuite calculé et $J(f^{(k)})$ (1) ré-estimé. A chaque itération *J(f)* est calculé analytiquement.

Etape 4 : Critère d'arrêt : $\left|h_n^{(k-1)} - h_n^{(k)}\right| < \varepsilon$.

L'approximation analytique de *J(f)* n'exige pas d'estimer la densité de probabilité à chaque itération ce qui permet de réduire le temps de calcul.

3.4. Simulations

L'algorithme proposé est testé sur différentes distributions. Les densités de probabilité sont estimées par les deux méthodes puis comparées avec la distribution théorique au sens de l'EQMI.

Nous présentons ici le cas d'une distribution mélange d'une loi normale et d'une loi uniforme ayant une densité de probabilité de la forme suivante :

$$f(x) = \pi_1 f_{\mu_1,\sigma_1}(x) + \pi_2 f_{a,b}(x) \quad (8)$$

avec μ_1=0.3, σ_1=0.2, a=-0.3, b=0.2.

Les probabilités a priori π_1 et π_2 sont respectivement de 0.75 et 0.25 et la taille de l'échantillon de 4000.

La figure 1 représente les deux différentes estimations de la densité de probabilité théorique. On peut observer que les estimations obtenues par les 2 méthodes sont comparables. Ce résultat est corroboré par les valeurs des EQMI présentées dans le tableau 1 : Il s'agit des mêmes valeurs dans les deux cas avec des variances très proches.

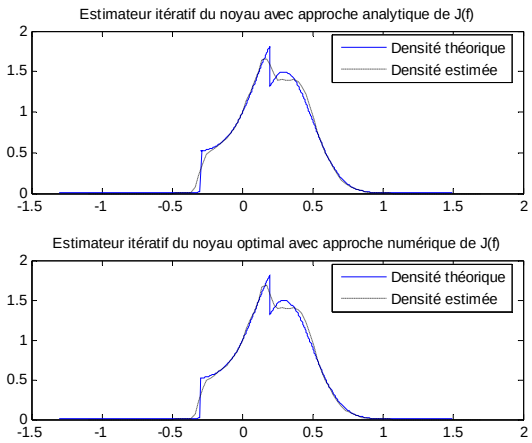

FIG. 1 : Comparaison de la densité théorique avec les deux densités estimées.

TAB. 1 : EQMI et variance de l'EQMI estimés à partir de 1000 simulations de la densité théorique en utilisant les deux estimateurs étudiés.

	EQMI	Variance
Algorithme itératif du noyau optimal (*J(f)* numérique)	0.0223	$2.6130.10^{-5}$
Algorithme itératif du noyau optimal (*J(f)* analytique)	0.0223	$2.6432.10^{-5}$

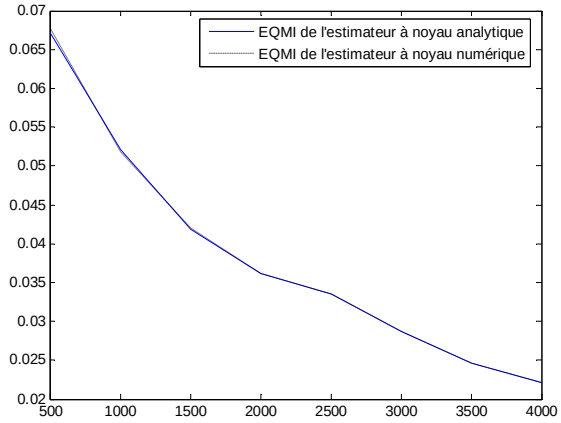

FIG. 2 : EQMI en fonction de la taille d'échantillon pour les deux estimateurs.

Dans la figure 2, l'EQMI est calculé en fonction de la taille d'échantillon (100 itérations). On peut observer que les deux courbes sont très proches et que les valeurs de l'EQMI sont faibles et tendent vers zéro lorsque la taille d'échantillon croît.

3.4. Etude de la complexité

Dans le cas de l'algorithme itératif du noyau optimal avec approche numérique de *J(f)*, *f* et *J(f)* sont estimés k fois, k étant le nombre d'itérations. L'estimation de *f* est O(2np), et l'estimation de *J(f)* est O(2p). La complexité de l'algorithme est de O(2knp). Pour l'algorithme itératif du noyau optimal avec approche analytique de *J(f)*, *f* n'est estimée qu'une seule fois. Le coût en temps de calcul pour cet algorithme est de O(2p(k+n)). k étant très faible comparativement à n, la complexité devient O(2pn).

4. Application à l'estimation de la neutralité de populations berbères de Tunisie

Dans ce paragraphe, nous proposons d'estimer des distributions de la statistique D de Tajima obtenues par simulation de populations neutres par les deux algorithmes présentés ci-dessus. La construction de populations neutres et le calcul de la statistique D sont réalisés par la simulation de généalogies de gènes en fonction de deux paramètres : θ défini comme étant égal à 4Nμ, N étant la taille de la population et μ, le nombre de mutations par génération et n la taille de l'échantillon. Une population est estimée neutre avec un risque de première espèce α égal

à 0,05, si la probabilité que la variable aléatoire D générée soit inférieure à la valeur de D observée dans l'échantillon (D_p), est supérieure à 0,02 ($P[D<D_p]>0.02$). Lorsque la probabilité calculée est proche de 0,02, il est difficile de conclure à la neutralité ou à la non neutralité des populations testées. La probabilité calculée étant une variable aléatoire, elle dépend directement des distributions de D générées. Il paraît alors plus logique d'évaluer la neutralité d'une population à partir d'une moyenne de probabilités calculées. Le gain en temps de calcul apporté dans ce contexte par l'approche analytique de *J(f)* paraît intéressant.

Pour illustrer notre propos, nous présentons une population litigieuse au sens de la neutralité dont les caractéristiques sont présentées dans la table 2.

Les moyennes des probabilités calculées pour les deux estimateurs sont présentées dans le tableau 3.

TAB. 2 : Caractéristiques des populations étudiées

Population	n	π	D_p
Sened	55	7.60471	-1.71764

Nous pouvons constater que les deux méthodes donnent des résultats très comparables avec un gain en temps de calcul pour la méthode itérative à noyau optimal avec approche de *J(f)* par calcul analytique.

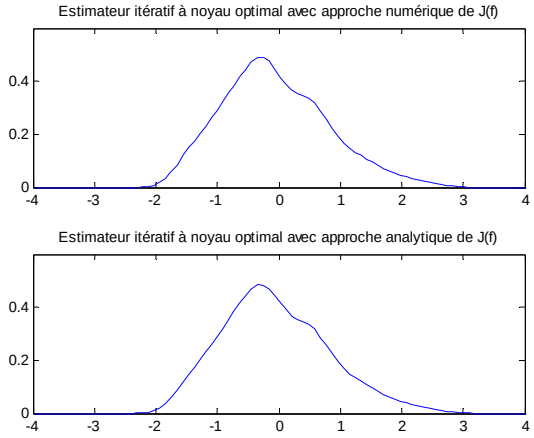

FIG. 3 : Estimation de la densité de probabilité de D pour la population de Sened avec les deux estimateurs étudiés

TAB. 3. Valeurs moyennes de $P[D<D_p]$ obtenues à partir des deux estimateurs itératifs à noyau étudiés.

	$P[D<D_p]$ (Valeur moyenne)	
Population	Méthode plug-in par approximation numérique de *J(f)*	Méthode plug-in par calcul analytique de *J(f)*
Sened	0.0213	0.0214

5. Conclusion et perspectives

Nous avons proposé un algorithme itératif et rapide pour l'estimation des densités de probabilité. Cet algorithme est basé sur l'étude du développement limité de l'erreur quadratique moyenne intégrée de l'estimateur à noyau.

L'entité notée *J(f)* qui correspond à l'intégrale de la dérivée seconde élevée au carré de la densité de probabilité à estimer, est approchée analytiquement à chaque itération dans le but d'estimer le pas optimal h_n.

Cet algorithme a été testé sur plusieurs simulations de densités multimodales généralement difficiles à estimer. L'estimateur proposé est comparable à l'estimateur itératif à noyau optimal basé sur une approche numérique de *J(f)*. Il permet, cependant de réduire la complexité de l'algorithme ce qui permet de diminuer les temps de calcul.

Cet estimateur peut être appliqué à plusieurs autres domaines. Nous pouvons citer, à titre d'exemple, l'estimation de la probabilité d'erreur des systèmes de communication numérique de type CDMA qui sera présenté dans nos futurs travaux. Nous projetons également d'appliquer cette idée à la méthode du noyau-difféomorphisme pour les densités de probabilité à support borné et de généraliser cet algorithme au cas multivarié.

Références

[1] A. W. Bowman, and A. Azzalini, *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, 1997.

[2] W. J. Ewens, *The sampling theory of selectively sampling alleles*. Theoretical population biology, Vol. 3, pp. 87 – 112, 1972.

[3] Y. X. Fu, *Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection*. Genetics, Vol. 147, pp. 915 – 925, 1997.

[4] P. Hall, *Comparison of two orthogonal series methods of estimating a density and its derivatives on interval*. J. Multivariate anal., Vol. 12, pp. 432 – 449, 1982.

[5] M.C. Jones, J.S. Marron, S.J. Sheather, *A brief survey of bandwidth selection for density estimation*, J. Amer. Stat. Assoc., Vol. 91, pp. 401 – 407, 1996.

[6] E. Parzen, *On estimation of a probability density function and mode*. Annals of mathematical statistics, Vol. 33, pp. 1065-1076, 1962.

[7] S. Saoudi, F. Ghorbel, A. Hillion, *Non parametric probability density function estimation on a bounded support: applications to shape classification and speech coding*. Applied Stochastic Models and Data Analysis, Vol. 10, pp. 215-231, 1994.

[8] F. Tajima, *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism,* Genetics, Vol. 123, pp. 585 – 595, 1989.