Nouveaux résultats sur l'estimation adaptative de la densité dans un modèle semi-paramétrique

Jean-Baptiste Aubin, Samuela Leoni-Aubin

NOUVEAUX RESULTATS SUR L'ESTIMATION ADAPTATIVE DE LA DENSITE DANS UN MODELE SEMIPARAMETRIQUE

Jean-Baptiste AUBIN & Samuela LEONI-AUBIN

jean-baptiste.aubin@utc.fr, U.T.C., Rue P. de Roberval - BP 20529, 60205 Compiègne.
samuela.leoni@insa-lyon.fr, I.N.S.A. de Lyon, 20, Rue A. Einstein, 69621 Villeurbanne.

ABSTRACT

A $m-$sample semiparametric model in which the ratio of $m-1$ probability density functions with respect to the mth is of a known parametric form without reference to any parametric model is considered. This model arises naturally from retrospective studies and multinomial logistic regression model. An adaptative projection density estimator is constructed by smoothing the increments of the maximum semiparametric empirical likelihood estimator of the underlying distribution function, using the combined data from all the samples. Some asymptotic results on the proposed projection density estimator are established.

RESUME

Nous disposons de m échantillons pour lesquels les rapports des $m-1$ premières densités de probabilité par rapport à la $m-$ième sont de forme paramétrique connue. Ce modèle semiparamétrique apparaît naturellement notamment dans le modèle de régression logistique multinomiale. Nous introduisons de nouveaux estimateurs par projection adaptatifs semiparamétriques des m densités à partir des observations combinées de tous les échantillons. Nous établissons des vitesses de convergence suroptimales pour l'erreur quadratique intégrée dans un sous-ensemble dense dans l'ensemble des densités possibles.

MOTS CLES: Modèles Semi et Non Paramétriques, Statistique Mathématique

1. INTRODUCTION

Nous considérons m échantillons aléatoires indépendants $X_{i1}, \ldots, X_{in_i}$, $i = 1, \ldots, m$ de densités inconnues de probabilité respectives $g_i(x) = \mathrm{d}G_i(x)$, $i = 1, \ldots, m$ à estimer.
Soit le modèle semiparamétrique à rapport de densités suivant:

$$g_i(x) = w(x, \theta_i) g_m(x), \qquad i = 1, \ldots, m-1, \tag{1}$$

où $w(x, \theta_i) := \exp\{\theta_{1,i} + s(x)\theta_{2,i}\}$ est une fonction connue, positive et bornée et $\theta_i := \left[\theta_{1,i}, \theta_{2,i}^t\right]^t$, $i = 1, \ldots, m-1$ est un vecteur de paramètres de dimension finie d.

De récents travaux (Cheng et Chu (2004), Fokianos (2004), Aubin et Leoni-Aubin (2008-2) et Qin et Zhang (2004)) ont consideré le problème de l'estimation de la $m$ième densité dans un modèle à rapport de densités pour m échantillons (Fokianos (2004) et Aubin et Leoni-Aubin (2008-2)) ou à deux échantillons (Cheng et Chu (2004) et Qin et Zhang (2004)).
La première étape consiste en l'estimation des paramètres de dimension finie en maximisant une fonction de vraisemblance semiparamétrique, la seconde, en l'obtention d'un estimateur de la fonction de répartition inconnue G_m par maximum de vraisemblance semiparamétrique en posant des poids sur toutes les observations. Un lissage (utilisant un noyau pour Cheng et Chu (2004), Fokianos (2004) et Qin et Zhang (2004) ou une méthode de projection pour Aubin et Leoni-Aubin (2008-2)) mène alors à une nouvelle estimation de la densité.
Plus spécifiquement, Fokianos (2004) a montré que dans le modèle à rapport de densités (1), les observations combinées mènent à des estimateurs à noyau des distributions inconnues asymptotiquement plus efficaces, dans le sens où ces estimateurs, s'ils ont un biais identique à celui obtenu par les estimateurs classiques à noyau, ont une variance plus faible.
Des résultats analogues ont été démontrés par Aubin et Leoni-Aubin (2008-2) dans le cas d'un lissage par projection. De plus, une amélioration par rapport aux estimateurs semiparamétriques à noyau a été constatée dans le cas où la base de projection est convenablement choisie, c'est-à-dire dans le cas où les coefficients de Fourier décroîssent assez vite.
Le but de cette contribution est d'estimer des densités inconnues en deux étapes, en utilisant les données combinées de tous les échantillons, donc en tirant partie de l'information contenue dans tous les échantillons. Premièrement, en appliquant la méthode du maximum de vraisemblance empirique au modèle (1), deuxièmement, en procédant à une estimation par projection adaptative de la densité recherchée. La méthode adaptative utilisée est décrite dans Bosq (2005). Elle permet d'atteindre dans un ensemble dense dans celui des densités "possibles" des vitesses suroptimales pour l'erreur quadratique intégrée, et ainsi de faire mieux que les estimateurs présentés précédemment.
Dans la seconde partie nous rappelons la méthode d'estimation des paramètres de dimension finie dans le modèle (1) basée sur la vraisemblance empirique. La partie 2 définit l'estimateur par projection adaptative de la densité inconnue et la partie 3 présente quelques propriétés asymptotiques de l'estimateur considéré. En particulier, nous démontrons que, lorsque la base de projection est choisie de telle sorte que les coefficients de Fourier s'annulent à partir d'un certain rang, l'estimateur proposé atteint des vitesses de convergence pour l'erreur quadratique intégrée suroptimales, faisant ainsi mieux que les estimateurs semiparamétriques présentés jusqu'alors.

1. ESTIMATEUR DE LA DENSITÉ SEMIPARAMÉTRIQUE PAR PROJECTION ADAPTATIVE

Soient m échantillons dont les m densités correspondantes satisfont l'équation (1), soit de plus $n := \sum_{i=1}^{m} n_i$ le nombre total d'observations . On estime d'abord le paramètre θ comme "argsup" de la log-vraisemblance empirique profile basée sur toutes les observations des m échantillons $\{X_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, m\}$

$$l(\theta) = -\sum_{i=1}^{m}\sum_{j=1}^{n_i} \log\left[1 + \sum_{k=1}^{m-1} \frac{n_k}{n}\{w(X_{ij}, \theta_k) - 1\}\right] + \sum_{i=1}^{m-1}\sum_{j=1}^{n_i} \log[w(X_{ij}, \theta_i)].$$

On montre que sous certaines conditions (voir Aubin et Leoni-Aubin (2008)) $\widehat{\theta} \to \theta_0$ *p.s.* (par rapport à G_m)

Cheng et Chu (2004), Fokianos (2004) et Qin et Zhang (2004) proposent des estimateurs semiparamétriques en modifiant les estimateurs classiques à noyau (essentiellement en lissant les sauts de $\widehat{G}_i$, $i = 1, \ldots, m$, où $\widehat{G}_i$ est l'estimateur du maximum de vraisemblance de G_i).
Aubin et Leoni-Aubin (2008-2) introduisent des estimateurs semiparamétriques par projection. La méthode de projection consiste en la projection de la densité à estimer sur un espace de dimension finie (par exemple celui généré par les premières composantes d'une base de l'espace des densités possibles). On estime ensuite cette projection par une méthode des moments.
Nous supposons que pour $l = 1, \ldots, m$, le $l-$ième échantillon admet la densité g_l par rapport à μ telle que $g_l \in L^2(\mu)$, où μ est une mesure finie de mesure totale σ_μ. L'espace de Hilbert $(L^2(\mu), \|.\|)$ est supposé séparable, de dimension infinie et muni d'une base $e_1, e_2, \ldots$ orthonormale.
L'estimateur par projection classique pour $g_m = \sum_{j=1}^{\infty} a_j e_j$ est $\overline{g}_{m_{n_m}} = \sum_{j=1}^{k_{n_m}} \overline{a}_{j,n_m} e_j$, où (k_{n_m}) est une suite d'indices de troncature telle que $k_{n_m} \leq n_m$, $(n_m/k_{n_m}) \uparrow \infty$ et $(k_{n_m}) \uparrow \infty$ lorsque $n_m \uparrow \infty$. $\overline{a}_{j,n_m} = \frac{1}{n_m}\sum_{i=1}^{n_m} e_j(X_{mi})$ est un estimateur sans biais du $j-$ième coefficient de Fourier a_j (de la densité g_m). Nous supposons de plus que la base $(e_j)_{j\in\mathbb{N}^*}$ est uniformément bornée (telle que $\exists M < \infty : \sup_j \|e_j\|_\infty < M$).
Comme dans Aubin et Leoni-Aubin (2008-2), il est possible de tirer profit de l'information de tous les échantillons (au lieu de ne considérer que le dernier) pour estimer plus efficacement g_m. Nous proposons d'utiliser un indice de troncature aléatoire dans l'estimateur adaptatif semiparamétrique par projection suivant:

$$\widehat{g}_{m_n} = \sum_{j=1}^{\widehat{k}_n} \widehat{a}_{j,n} e_j, \qquad \text{avec} \qquad \widehat{a}_{j,n} := \sum_{i=1}^{m}\sum_{j=1}^{n_i} \widehat{p}_{ij} e_j(X_{ij}), \tag{2}$$

où $\widehat{k}_n := \max\{j : 1 \leq j \leq k_n : |\widehat{a}_{j,n}| \geq \gamma_n\}$, (k_n) étant une suite d'entiers croissant vers l'infini moins vite que n et $\widehat{p_{ij}}$ l'estimateur de vraisemblance empirique de $p_{ij} := dG_m(X_{ij})$. Par ailleurs, la suite de seuils (γ_n) est donnée par $\gamma_n = \min\left(c\sqrt{\frac{\log n}{n}}, 1\right)$, $c > 0$. Notons que le choix de c est laissé à l'utilisateur et qu'un tel choix influe sur $\widehat{k}_n$.
Par ailleurs, afin d'assurer l'existence de $\widehat{k}_n$, nous prenons la première composante de la base de projection e_1 constante et égale à $\frac{1}{\sqrt{\sigma_\mu}}$.
Notons que l'équation (2) définit un estimateur de la densité semiparamétrique puisqu'il dépend à la fois de la fonction de répartition inconnue et des paramètres du modèle (1).
Par la suite, nous nous intéressons au comportement asymptotique de l'indice de troncature. Il apparaît que celui-ci se fixe presque sûrement sur le dernier terme non nul du développement de g_m sur la base orthonormale $(e_j)_{j\in\mathbb{N}^*}$ de l'espace des densités "possibles" considérée. De façon analogue, il tend presque sûrement vers l'infini si le nombre de termes dans le développement est infini. Un encadrement plus précis de l'indice de troncature est fourni par la suite. Nous distinguons par la suite les fonctions admettant un développement fini par rapport à la base $(e_j)_{j\in\mathbb{N}^*}$ (définissant le sous-espace $\mathcal{G}_0$) des autres (définissant le sous-espace $\mathcal{G}_1$).
Des propriétés suroptimales de l'estimateur sont ensuite données. Celles-ci concernent l'espérance quadratique intégrée pour $\mathcal{G}_0$ dense dans l'espace $L^2(\mu)$ des densités "possibles".
Enfin, nous précisons le comportement de l'estimateur sur le complémentaire de $\mathcal{G}_0$ dans $L^2(\mu)$, $\mathcal{G}_1$.

Dans la suite, nous distinguons donc:

$$\begin{aligned}
\mathcal{G}_0(K) &:= \left\{g_m \in L^2(\mu) : a_K \neq 0, a_j = 0, j > K\right\}, \\
\mathcal{G}_0 &:= \bigcup_{K=1}^{\infty} \mathcal{G}_0(K) \\
\mathcal{G}_1 &:= L^2(\mu) - \mathcal{G}_0.
\end{aligned}$$

Sur $\mathcal{G}_0$, la vitesse pour l'espérance quadratique intégrée est meilleure que celle obtenue par la méthode d'estimation à noyau, à la fois dans les cas classique (nonparamétrique) et semiparamétrique (voir Cheng et Chu (2004), Fokianos (2004) et Qin et Zhang (2004)).
Nous déduisons enfin les estimateurs pour les autres densités g_l, $l = 1, \ldots, m-1$ comme suit:

$$\widehat{g}_{l_n} = \sum_{j=1}^{\widehat{k}_n} \widehat{a}_{j,n} e_j, \qquad \text{avec} \qquad \widehat{a}_{j,n} := \sum_{i=1}^{n} \widehat{p}_i w(T_i, \widehat{\theta}_l) e_j(T_i), \qquad l = 1, \ldots, m-1.$$

Ces estimateurs jouissent des mêmes propriétés asymptotiques que (2).

Nous démontrons que, pour $g_m \in \mathcal{G}_0(K)$, $\widehat{k}_n$ est un estimateur de K et que de façon analogue, sur $\mathcal{G}_1$, $\widehat{k}_n$ tend vers l'infini.

1 Comportement asymptotique de $\widehat{g}_{m_n}$

Dans cette partie, nous considérons l'erreur quadratique intégrée asymptotique de l'estimateur de la densité semiparamétrique $\widehat{g}_{m_n}$ (défini en (2)) comme mesure de sa qualité.
Pour étudier les propriétés statistiques de $\widehat{g}_{m_n}$, nous serons amenés à considérer

$$\tilde{g}_{m_n} = \sum_{j=1}^{\widehat{k}_n} \tilde{a}_{j,n} e_j,$$

avec $\tilde{a}_{j,n}$ défini plus haut.

Théorème 1.1 *Si $\widehat{\theta} \to \theta_0$ p.s. et si $g_m \in \mathcal{G}_0(K)$, alors, pour n assez grand,*

$$\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathcal{O}\left(\frac{1}{n}\right).$$

Notons que la vitesse optimale "classique"dans le cas semiparamétrique (voir Aubin et Leoni-Aubin (2008-2)) est en $\mathcal{O}\left(\frac{k_n}{n}\right) + \sum_{j>k_n} a_j^2$.

Nous nous intéressons à présent au comportement asymptotique de l'estimateur lorsque g_m n'admet pas de développement fini par rapport à la base de projection. Voici un encadrement asymptotique de $\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2$ dans ce cas.

Théorème 1.2 *Si $\widehat{\theta} \to \theta_0$ p.s. et si $g_m \in \mathcal{G}_1$, alors, pour n et c assez grands, on a presque sûrement*

$$\mathcal{O}\left(\frac{q((1+\varepsilon)\gamma_n)}{n}\right) + \sum_{j>q((1-\varepsilon')\gamma_n)\wedge k_n} a_j^2 \leq \mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 \leq \mathcal{O}\left(\frac{k_n}{n}\right) + \sum_{j>q((1+\varepsilon)\gamma_n)} a_j^2.$$

2 Conclusions et Perspectives

Nous avons défini une nouvelle classe d'estimateurs semiparamétriques par projection de la densité. En plus de d'élargir les conditions usuelles de travail sur des problèmes à m échantillons, cette méthode a l'avantage de conduire, pour des bases de projection bien

choisies par l'utilisateur, à des estimateurs pouvant atteindre des vitesses suroptimales (paramétriques) pour l'erreur quadratique intégrée asymptotique.

En outre, les logiciels statistiques usuellement utilisés sont bien adaptés pour calculer ces estimateurs. A ce propos, nous développons un "package" relatif à cette méthode pour le logiciel libre R.

Enfin, le choix aléatoire de l'indice de troncature $\widehat{k}_n$ s'appuie sur des choix préalables de constantes influant grandement sur la valeur de $\widehat{k}_n$ à distance finie. Il serait oppportun de procéder à de nombreuses simulations numériques afin de préciser le comportement de notre estimateur, d'un côté par rapport à différents choix de ces constantes, et d'un autre par rapport à ses concurrents naturels, les estimateurs semiparamétriques à noyau ou à projection non adaptatifs.

References

[1] Aubin, J.B. et Leoni-Aubin, S. (2008) Estimation adaptative de la densité par projection dans un modèle semiparamétrique à rapport de densités. *Annales de L'ISUP* **52** n. spécial.

[2] Aubin, J.B. et Leoni-Aubin, S. (2008-2) Projection density estimation under a m-sample semiparametric model. *Computational Statistics and Data Analysis* **52**(5), 2451–2468.

[3] Bosq, D. (2005) *Inférence et prévision en grandes dimensions*. Economica, Paris.

[4] Cheng, K.F. and Chu, C.K. (2004) Semiparametric density estimation under a two-sample density ratio model. *Bernoulli* **10**(4), 583–604.

[5] Fokianos, K. (2004) Merging information for semiparametric density estimation. *J. R. Statist. Soc. B* **66**(4), 941–958.

[6] Qin, J. et Zhang, B. (2005) Density estimation under a two-sample semiparametric model. *Nonparametric Statistics* **17** (6), 665–683.