# Estimation semi-paramétrique par minimisation de l'entropie des résidus, application en traitement d'images

Eric Wolsztynski, Eric Thierry, Luc Pronzato

Laboratoire I3S
Les Algorithmes, 2000, route des lucioles - bât. Euclide B BP.121, 06903 Sophia Antipolis - Cedex, France
{wolsztyn, et, pronzato}@i3s.unice.fr

**Résumé** − Nous considérons un problème d'estimation semi-paramétrique en régression non linéaire, où le paramètre de nuisance (de dimension infinie) est la densité $f$ du bruit additif, dont on suppose uniquement qu'elle est symétrique en 0. Nous proposons ici une extension au cas multivariable de l'estimateur présenté dans [7] qui minimise l'entropie de l'échantillon symétrisé des résidus. Des résultats en traitement d'images illustrent les bonnes propriétés de cette méthode d'estimation.

**Abstract** − We consider a semiparametric estimation problem in nonlinear regression, for which the infinite-dimensional nuisance parameter is the density $f$ of the additive noise. We only suppose $f$ to be symmetric about 0. We propose a multivariate extension of the estimator presented in [7] that minimizes the entropy of the symmetrized residuals. Some results in image processing illustrate the properties of the minimum entropy estimation method.

## 1 Introduction

Ce travail reprend l'estimateur par Minimum d'Entropie (ME) présenté dans [7] dans le contexte de l'estimation semiparamétrique pour des modèles de régression non linéaire. Nous considérons $n$ observations données pour les points de mesure (déterministes ou aléatoires) $X_i \in \mathcal{X} \subset \mathbb{R}^q$ par

$$Y_i = \eta(\bar{\theta}, X_i) + \varepsilon_i\,, \quad i = 1, \ldots, n\,, \qquad (1)$$

où $(\varepsilon_i)$ est une suite de variables aléatoires i.i.d. de densité inconnue $f$, $\eta(\theta, x)$ est une fonction connue, bornée sur $\Theta \times \mathcal{X}$ et deux fois continûment différentiable en $\theta$ pour tout $x \in \mathcal{X}$, et $\bar{\theta} \in \text{int}(\Theta)$ est la vraie valeur (inconnue) du vecteur de paramètres $\theta \in \Theta = \overline{\text{int}(\Theta)} \subset \mathbb{R}^p$ que l'on cherche à estimer. $f$ est supposée symétrique en 0, régulière et de support infini. L'approche que nous proposons consiste à minimiser une estimée de l'entropie des résidus (fonctions de $\theta$)

$$e_i(\theta) = Y_i - \eta(\theta, X_i) = \varepsilon_i + \eta(\bar{\theta}, X_i) - \eta(\theta, X_i) \quad (2)$$

dont la densité à $X_i$ fixé est donnée par

$$f_{e,\,X_i}(u) = f(u - \eta(\bar{\theta}, X_i) + \eta(\theta, X_i)).$$

On peut montrer [8] grâce à des résultats classiques de théorie de l'information que l'entropie de Shannon $H(f_e^s) = -\int f_e^s(e) \log f_e^s(e) de$ de la loi marginale des résidus (symétrisés), donnée par

$$f_e^s(u) = \int_{\mathcal{X}} \frac{1}{2} \left[ f_{e,\,X_i}(u) + f_{e,\,X_i}(-u) \right] \mu(dx), \quad (3)$$

est minimale en $\theta = \bar{\theta}$, où elle coïncide avec l'entropie de $f$. Quand la densité $f$ est inconnue, $f_{e,\,X}$ et $f_e^s$ sont également inconnues, et le critère $H(f_e^s)$ ne peut pas être utilisé. Le critère d'estimation de $\theta$ sera alors une estimée de l'entropie des résidus (symétrisés).

Dans ce qui suit nous supposons que l'information de Fisher de $f$ est finie et que la matrice d'information de Fisher, pour une mesure $\mu$ sur $x$, est de rang plein pour tout $\theta$ dans un voisinage de $\bar{\theta}$.

Nous supposons vérifiée la condition d'identifiabilité $\int_{\mathcal{X}} [\eta(\theta,x) - \eta(\bar{\theta},x)]^2 \mu(dx) = 0 \;\Rightarrow\; \theta = \bar{\theta}$.

## 2 Un premier estimateur

Un premier estimateur, proposé dans [7], minimise l'entropie empirique d'une estimée de la densité des résidus. On utilise dans la méthode l'échantillon symétrisé des résidus car l'entropie est invariante par translation, ce qui ne permet pas l'estimation de composantes constantes dans un modèle de régression. Les $2n$ résidus $\pm e_i(\theta)$ permettent ainsi d'obtenir une estimée symétrique de leur densité $f_e^s$ en utilisant des techniques de lissage. On peut considérer en particulier l'estimateur à noyaux

$$\hat{f}_n^{\theta}(u) = \frac{1}{2\,n} \sum_{i=1}^{n} \left[ K_{h_n}(u - e_i(\theta)) + K_{h_n}(u + e_i(\theta)) \right]. \tag{4}$$

La fonction noyau $K_{h_n}(.) = 1/h_n K(./h_n)$ utilisée est symétrique en 0 (on considère des noyaux réguliers usuels, par exemple la loi normale centrée réduite). La dépendance en $\theta$ du critère est donc exprimée dans la construction de $\hat{f}_n^{\theta}$. L'estimateur par substitution de l'entropie de $f_e^s$ est ensuite obtenu par

$$\hat{H}_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_n^{\theta}(e_i(\theta))\, U_n[e_i(\theta)], \tag{5}$$

où $U_n$ est une troncature lisse des grandes valeurs des résidus (voir [7]). $\hat{H}_n(\theta)$ est deux fois continûment différentiable en $\theta \in \mathrm{int}(\Theta)$. Des exemples pour des résidus univariés sont donnés dans [7] et [8]. La convergence en probabilité de l'estimateur par ME $\hat{\theta}_{ME}^n = \arg\min_\theta \hat{H}_n(\theta)$ pour le problème de position (sous certaines conditions sur les queues de distribution) est donnée dans [9], ainsi que la convergence en probabilité de $\nabla_\theta^2 \hat{H}_n(\hat{\theta}_{ME}^n)$ vers $\nabla_\theta^2 H(\bar{\theta}) = i(f)$, l'information de Fisher pour la position, $i(f) = \int (f'/f)^2 f$.

L'adaptativité (au sens de Bickel) de $\hat{\theta}_{ME}^n$, c'est-à-dire la propriété que l'estimateur demeure *asymptotiquement efficace*, au sens du Maximum de Vraisemblance (MV) dans le cadre paramétrique, quand $n \to \infty$ et malgré le manque de connaissance sur le paramètre de nuisance $f$ de dimension infinie, reste une question ouverte (une définition de l'adaptativité est donnée dans [2], [4]). Nous pouvons montrer [8], toujours pour le modèle de position, qu'en partitionnant l'échantillon des données, la procédure d'estimation par minimum d'entropie des résidus coïncide avec l'approche en deux étapes de Stone-Bickel. Dans le cas de l'optimisation directe (sans partition des données), la normalité asymptotique de $\nabla_\theta^2 \hat{H}_n(\hat{\theta}_{ME}^n)$ reste cependant à prouver.

## 3 Cas de données multivariées

Lorsque la dimension $d$ des observations dépasse 2 ou 3, les estimateurs à noyaux deviennent peu performants pour des échantillons de taille raisonnable. Cette chute des performances est due à la difficulté de la sélection du paramètre de lissage $h$ (ou de la matrice $H$ selon l'approche choisie), qui doit être rapidement trop grand pour continuer à être performant [6].

Nous envisageons une approche qui utilise l'estimateur par $k^e$ plus proche voisin (kPPV) de l'entropie d'un échantillon, présenté pour le cas multivariable dans [3], où une preuve de la convergence en probabilité est donnée pour des conditions faibles sur $f$. L'entropie est estimée à partir de l'information apportée directement par la répartition des points de l'échantillon, ce qui permet de se passer de l'estimation de $f_e^s$ (bien que cet estimateur de l'entropie puisse s'interpréter comme un estimateur par substitution à noyaux uniformes). Considérons les résidus symétrisés $\pm e_j(\theta)$, $j = 1, \ldots, n$. Soit $\rho_{i,\,k}(\theta)$ la distance euclidienne entre l'un de ces points $z_i(\theta), i = 1, \ldots, 2n$, et son $k^e$ plus proche voisin. L'estimateur de $\theta$ par ME que nous suggérons minimise alors

$$H_{k,\,n}(\theta) = \log \bar{\rho}_k(\theta)^d + \log\left[c_1(d)(2n-1)\right] - \psi(k), \tag{6}$$

où $\bar{\rho}_k(\theta) = \left(\Pi_{i=1}^{2n} \rho_{i,\,k}(\theta)\right)^{1/(2n)}$ est la moyenne géométrique des $k$-distances à tous les points $z_i(\theta)$, $\psi(k) = \Gamma'(k)/\Gamma(k)$ est la fonction digamma, et $c_1(d) = 2\pi^{d/2}/(d\Gamma(d/2))$ est le volume de la boule unité de $\mathbb{R}^d$. Notons que seul le premier terme de (6) intervient dans l'estimation de $\theta$.

Le choix du paramètre $k$ ne semble pas aussi critique que celui du paramètre de lissage $h$ dans la méthode des noyaux. La sélection de $k$ n'implique pas d'étape d'optimisation, au contraire des approches par noyaux adaptatives (aux données) optimales classiques. Dans le cadre de l'estimation de $\theta$, il est néanmoins nécessaire de choisir $k > p$,

de manière à éviter les singularités possibles.

# 4 Simulations

Nous présentons ici des résultats de simulations obtenus pour des problèmes de traitement d'image, où l'on effectue une recherche exhaustive pour $\theta$ sur une grille finie. L'entropie est un critère naturel dans ce contexte, cette quantité décrivant selon la théorie du codage la longueur minimale de description nécessaire pour les donnés considérées (image des différences à transmettre). Minimiser l'entropie des erreurs entre deux signaux ou deux images revient donc à sélectionner les valeurs des paramètres pour lesquelles le taux de compression maximal est obtenu.

Nous considérons dans chaque exemple deux copies bruitées d'une même image. La deuxième image a cependant subi un changement homogène d'intensité lumineuse, ce qui correspond à une translation des valeurs des niveaux de couleur de chaque pixel. Dans ce contexte, il est donc préférable de ne pas symétriser l'échantillon des résidus, de manière à bénéficier de la propriété d'invariance par translation de l'entropie.

Les observations correspondent à un bloc de pixels de taille fixée pris dans la première image, (*cf.* Figs 1,a puis 1,c). Considérons le problème de retrouver les coordonnées du bloc qui lui correspond dans la deuxième image (la deuxième copie peut donc être "décalée" par rapport à la première ; nous choisissons ici de ne pas lui faire subir de décalage, ce qui ne change pas le problème). Nous appellerons ce bloc le bloc optimal ; ses coordonnées forment le vecteur de paramètres $\theta$, de dimension 2. La dimension des observations est donnée par le nombre de canaux de couleurs de l'image. Nous considérons ici soit des images en noir et blanc (données de dimension 1, Figs 1,a et 1,b, images 176×144), soit des images en couleurs (données de dimension 3, Figs 1,c et 1,d, images 352×288). Nous considérons de plus la présence de pixels "aberrants" dans l'image de référence et/ou dans l'image de travail (la deuxième copie).

Dans le premier exemple, les données univariées sont contaminées par un bruit gaussien de variance 10. La deuxième copie a subi une variation d'intensité lumineuse de 10 unités. Les observations $Y$

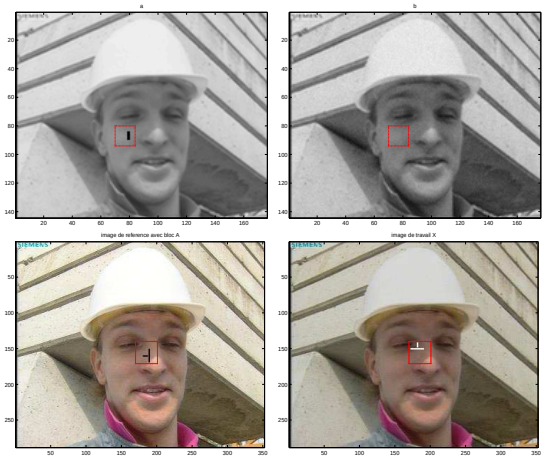

FIG. 1 – *images a, b, c, d avec blocs optimaux*

sont un bloc $A$ de taille $15 \times 15$, dont le coin supérieur gauche a pour coordonnées $\bar{\theta} = (80, 70)^t$ dans l'image de référence. $A$ comporte un ensemble de taille (2×6) de valeurs aberrantes (pixels noirs). Les valeurs des paramètres des différents estimateurs par minimum d'entropie sont $k = 5$ et $h_n = 2.345\hat{\sigma}\,(2n)^{-1/5}$ où $\hat{\sigma}$ est l'écart-type de l'échantillon (paramètre de lissage optimal, au sens de l'erreur quadratique intégrée moyenne, pour des noyaux gaussiens, voir [1]). Nous comparons dans cet exemple les approches avec puis sans symétrisation des résidus. La Table 1 contient les moyennes des estimées obtenues respectivement par ME avec kPPV et ME par substitution (MEs), Hellinger (DHM), moindres carrés (MC), et M-estimateur de Huber (M-est), pour 100 répétitions de la même expérience, pour des résidus symétrisés (S) puis non-symétrisés (NS).

TAB. 1 – *moyennes des estimées pour 100 répétitions avec une image N&B, blocs 15×15, bruit gaussien $\mathcal{N}(0, 10)$, résidus symétrisés (S) puis non-symétrisés (NS). $\bar{\theta} = (80, 70)^t$.*

| | kPPV | MEs | DHM | M-est | MC |
|---|---|---|---|---|---|
| S | 81.11 | **80.00** | 74.64 | 82.15 | 86.29 |
| | 64.98 | **69.99** | 84.48 | 64.45 | 65.87 |
| NS | 80.04 | **80.01** | 74.09 | 81.89 | 86.34 |
| | 71.47 | **70.09** | 86.00 | 64.43 | 65.84 |

Dans le deuxième exemple, les données de di-

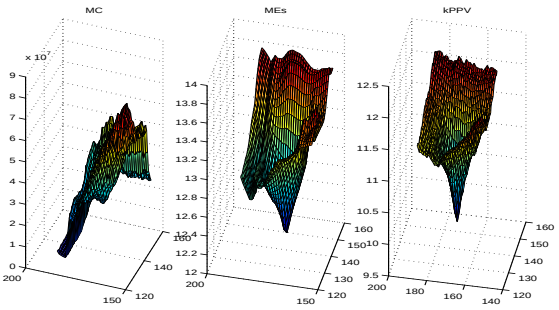


FIG. 2 – *critères vs $\theta$ pour une image couleur et des blocs 32×32 ; $\bar{\theta} = (140, 170)^t$.*

mension 3 sont contaminées par un bruit gaussien de variance 10, et les blocs optimaux dans les deux images comportent des pixels aberrants (de positions et de couleurs différentes). Les observations $Y$ sont un bloc $A$ de taille $32 \times 32$, dont le coin supérieur gauche a pour coordonnées $\bar{\theta} = (140, 170)^t$ dans l'image de référence. La deuxième image a subi une variation d'intensité lumineuse de 40 unités.

La Figure 2 montre un tracé des critères en fonction des coordonnées $\theta = (\theta_1, \theta_2)^t$. Elle illustre de gauche à droite le comportement des moindres carrés (MC), du critère de l'entropie par substitution MEs utilisant des produits d'estimateurs à noyaux univariés avec $h_n^j = \sigma_j (2n)^{-1/(d+4)}$ pour chaque composante $j$ des données [5] (on utilise ici la valeur exacte de l'écart-type du bruit, et non l'estimée de celui des résidus), et le comportement robuste du critère utilisant les $k^e$ plus proches voisins (kPPV), pour lequel le minimum est facilement identifiable.

Les premiers résultats que nous avons obtenus suggèrent que l'estimation semiparamétrique par minimum d'entropie est une approche très robuste et qui semble efficace pour des échantillons de taille raisonnable. L'approche par plus proches voisins permet d'appliquer le critère d'entropie minimale à des données de dimension supérieure à 1 tout en conservant des performances intéressantes. L'estimateur est en particulier peu sensible aux changements d'intensité entre deux images.

# Références


[1] A. Berlinet and L. Devroye. *A comparison of kernel density estimates.* Publications de l'Institut de Statistique de l'Université de Paris, 38(3) :3–59, 1994.

[2] P.J. Bickel. *On adaptive estimation.* Annals of Statistics, 10 :647–671, 1982.

[3] M.N. Goria, N.N. Leonenko, V.V. Mergel, and P.L. Novi Inverardi. *A new class of random vector entropy estimators and its applications in testing statistical hypotheses.* Journal of Nonparametric Statistics, 2005.

[4] C. Manski. *Adaptive estimation of nonlinear regression models.* Econometric Reviews, 3(2) :145–194, 1984.

[5] D.W. Scott. *Multivariate Density Estimation : Theory, Practice, and Visualization.* Wiley, 1992.

[6] B.A. Türlach. *Fast implementation of density-weighted average derivative estimation.* Computationally Intensive Statistical Methods, 26 :28–33, 1994.

[7] E. Wolsztynski, E. Thierry, and L. Pronzato. *Estimation semiparamétrique adaptative par minimum d'entropie.* In Proc. 36e Journées Françaises de Stat, Montpellier, 2004.

[8] E. Wolsztynski, E. Thierry, and L. Pronzato. *Minimum-entropy estimation in semi-parametric models.* Signal Processing, 2005.

[9] E. Wolsztynski, E. Thierry, and L. Pronzato. *Consistency of a minimum-entropy estimator of location.* Internal Report No I3S/RR-2004-38-FR, 30 pages, www.i3s.unice.fr/˜mh/RR/rapports.html, 2004.