LABORATOIRE

INFORMATIQUE, SIGNAUX ET SYSTÈMES
DE SOPHIA ANTIPOLIS
UMR 6070

MINIMUM-ENTROPY ESTIMATION IN SEMI-PARAMETRIC MODELS

Eric Wolsztynski, Eric Thierry, Luc Pronzato

Projet TOPMODEL

Rapport de recherche
ISRN I3S/RR–2004-16–FR

Juin 2004

LABORATOIRE I3S: Les Algorithmes / Euclide B – 2000 route des Lucioles – B.P. 121 –
06903 Sophia-Antipolis Cedex, France – Tél. (33) 492 942 701 – Télécopie : (33) 492 942 898
http://www.i3s.unice.fr/I3S/FR/

Minimum-Entropy Estimation in Semi-Parametric Models ⋆

Eric Wolsztynski, Eric Thierry and Luc Pronzato

Laboratoire I3S, Université de Nice-Sophia Antipolis/CNRS

Abstract

In regression problems where the density f of the errors is not known, maximum likelihood is unapplicable, and the use of alternative techniques like least squares or robust M-estimation generally implies inefficient estimation of the parameters. The search for adaptive estimators, that is, estimators that remain asymptotically efficient independently of the knowledge of f, has received a lot of attention, see in particular (Stein, 1956; Stone, 1975; Bickel, 1982) and the review paper (Manski, 1984). The paper considers a minimum-entropy parametric estimator that minimizes an estimate of the entropy of the distribution of the residuals. A first construction connects the method with the Stone-Bickel approach, where the estimation is decomposed into two steps. Then we consider a direct approach that does not involve any preliminary $\sqrt{n}$-consistent estimator. Some results are given that illustrate the good performance of minimum-entropy estimation for reasonable sample sizes when compared to standard methods, in particular concerning robustness in the presence of outliers.

Key words: Adaptive estimation, efficiency, entropy, parameter estimation, semi-parametric models, robustness, outliers

Résumé

Dans les problèmes de régression où la densité f des erreurs est inconnue, le maximum de vraisemblance ne peut pas être utilisé pour estimer les paramètres du modèle, et les techniques alternatives telles que moindres carrés ou M-estimation robuste impliquent en général la perte d'efficacité de l'estimation.

⋆ This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

Email address: {wolsztyn,et,pronzato}@i3s.unice.fr (Eric Wolsztynski, Eric Thierry and Luc Pronzato).

Beaucoup d'attention a été portée sur la recherche d'estimateurs adaptatifs, c'est-à-dire qui demeurent asymptotiquement efficaces indépendamment de la connaissance de f — voir en particulier (Stein, 1956; Stone, 1975; Bickel, 1982) et la revue des différentes approches existantes (Manski, 1984). Nous considérons ici un estimateur paramétrique qui minimise une estimée de l'entropie de la distribution des résidus. Une première construction permet d'établir le lien avec l'approche de référence de Stone et Bickel, où l'estimation est décomposée en deux étapes. Nous considérons ensuite une approche directe qui ne nécessite pas d'estimateur $\sqrt{n}$-consistant préliminaire. Quelques résultats illustrent dans les dernières parties les bonnes performances de l'estimation par minimum d'entropie pour des échantillons de taille raisonnable, en particulier concernant la robustesse de l'estimateur en présence de données aberrantes.

Mots-clés : Estimation adaptative, efficacité, entropie, estimation paramétrique, modèle semi-paramétrique, robustesse, données aberrantes

1 Problem statement

We consider a general nonlinear regression problem with observations

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

where $\bar{\theta}$ is the unknown value of the model parameters $\theta \in \Theta \subset I\!R^p$ and $\eta(\theta, x)$ is a known function of θ and the design variable $x \in \mathcal{X} \subset \mathbb{R}^d$. For F a function $\Theta \to I\!R$, $\nabla F(\theta)$ and $\nabla^2 F(\theta)$ will denote its first and second order derivatives with respect to θ, respectively a p-dimensional vector and a $p \times p$ symmetric matrix. For g a function $I\!R \to I\!R$, the first, second and third derivatives are simple denoted g', g'' and g'''. We shall assume the following throughout the paper. We suppose that $\bar{\theta} \in \mathrm{int}(\Theta)$, $\Theta = \overline{\mathrm{int}(\Theta)}$, and that $\eta(\theta, x)$ is bounded on $\Theta \times \mathcal{X}$ and two times continuously differentiable w.r.t. $\theta \in \mathrm{int}(\Theta)$ for any $x \in \mathcal{X}$, $\nabla\eta(\theta, x)$ and $\nabla^2\eta(\theta, x)$ being bounded on $\mathrm{int}(\Theta) \times \mathcal{X}$. The additive noise (ε_i) forms a sequence of i.i.d. random variables with p.d.f. f (with respect to the Lebesgue measure) that we suppose to be symmetric about zero. We further assume that f is two times continuously differentiable, with bounded derivatives $f'(.)$ and $f''(.)$, and that the Fisher information for location

$$\mathcal{I}(f) = \int_{-\infty}^{\infty} [f'(u)]^2 / f(u)\, du$$

exists. For a given measure μ on the design variable x, the Fisher information matrix $\mathbf{M}_F(\theta)$ associated with f and θ is given by

$$\mathbf{M}_F(\theta) = \mathcal{I}(f) \int_{\mathcal{X}} \nabla\eta(\theta, x)[\nabla\eta(\theta, x)]^\top \mu(dx)\,. \tag{2}$$

We suppose that $\mathbf{M}_F(\bar{\theta})$ has full rank and that the identifiability condition

$$\int_{\mathcal{X}} [\eta(\theta, x) - \eta(\bar{\theta}, x)]^2 \mu(dx) = 0 \Rightarrow \theta = \bar{\theta} \tag{3}$$

is satisfied.

In the classical situation where f is known, the Maximum Likelihood (ML) estimator $\hat{\theta}^n_{ML}$ is, under standard assumptions, *asymptotically efficient*, that is, asymptotically normal with minimum variance: $\sqrt{n}(\hat{\theta}^n_{ML} - \bar{\theta}) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \mathbf{M}_F^{-1}(\bar{\theta}))$. The difficulty here is that we need an estimator that does not require the knowledge of f, unlike ML estimation. Following Stein's formulation, see (Stein, 1956), the model (1) can then be termed *semi-parametric*, with θ and f respectively its parametric and non-parametric parts, and f can be considered as an infinite-dimensional nuisance parameter for the estimation of θ. In general, the presence of this nuisance parameter induces the loss of efficiency. An estimator that remains asymptotically efficient in these conditions is called *adaptive*. A precise definition is given in (Bickel, 1982) and Begun et al. (1983) give a necessary condition for adaptive estimation. The issue of adaptivity has motivated a large amount of work and major results have been derived by a series of authors. In particular, Beran (1974) and Stone (1975) proved that adaptive estimation in the location model was possible, using respectively adaptive rank estimates, and an approximation of the score function based on a kernel estimate of f constructed from residuals obtained with a preliminary $\sqrt{n}$-consistent estimator. It is this second approach that has been further developed by Bickel (1982), see also (Manski, 1984). Here we shall follow an approach that consists of minimizing the entropy of a kernel estimate constructed from the symmetrized residuals in the regression model (1). The approach is introduced and justified in Section 2. In Section 3 a two-step method is shown to coincide with the Stone-Bickel method, which is adaptive under suitable conditions. Direct (one step) entropy minimization is considered in section 4 for the location problem. An example illustrates the good performance of the approach for moderate sample sizes ($n = 100$ observations). Generalization to the more general case of nonlinear regression is considered in Section 5, with an example illustrating the robustness of the estimator with respect to outliers.

2 An introduction to minimum-entropy estimation

When f is known, the Maximum Likelihood estimator $\hat{\theta}^n_{ML}$ minimizes

$$\bar{H}_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log f[e_i(\theta)] \tag{4}$$

with respect to $\theta \in \Theta$. Since $\bar{H}_n(\bar{\theta}) = -(1/n)\sum_{i=1}^n \log f(\varepsilon_i)$ is an empirical version of the (Shannon) entropy

$$H(f) = -\int_{-\infty}^{\infty} \log[f(u)]f(u)du\,,$$

an intuitive idea is to base the estimation criterion on entropy. Indeed, the entropy of the residuals being a measure of their dispersion, its minimization forces them to gather. Consider the residuals $e_i(\theta)$ obtained from the observations in the regression model (1),

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

The density of $e_i(\theta)$, given X_i, is thus

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

Since entropy is invariant by translation, we shall consider the $2n$ symmetrized residuals [1] $e_i(\theta), -e_i(\theta)$, with corresponding density given X_i

$$f^s_{e,\,X_i}(u) = \frac{1}{2}\left[f(u - \eta(\bar{\theta}, X_i) + \eta(\theta, X_i)) + f(u + \eta(\bar{\theta}, X_i) - \eta(\theta, X_i))\right]. \tag{5}$$

We can show that the entropy of the marginal distribution of the symmetrized residuals,

$$f^s_e(u) = \int_{\mathcal{X}} f^s_{e,\,x}(u)\mu(dx) \tag{6}$$

is maximum for $\theta = \bar{\theta}$. Indeed,

$$H(f^s_e) = -\frac{1}{2}\int_{\mathcal{X}} \left[\int_{-\infty}^{\infty} f[u - \eta(\bar{\theta}, x) + \eta(\theta, x)] \log[f^s_e(u)]\, du\right] \mu(dx)$$

[1] Another possibility would be to use un-symmetrized residuals with the constraint that their median, or their mean, is zero.

$$
\begin{aligned}
&-\frac{1}{2}\int_{\mathcal{X}}\left[\int_{-\infty}^{\infty} f[u+\eta(\bar\theta,x)-\eta(\theta,x)]\log[f_e^s(u)]\,du\right]\mu(dx)\\
&\geq -\int_{\mathcal{X}}\left[\int_{-\infty}^{\infty} f(u)\log[f(u)]\,du\right]\mu(dx) = H(f)\,,
\end{aligned}
$$

see, *e.g.*, Ash (1965), Lemma 8.3.1 p. 238. Equality is obtained only if for μ-almost all x and almost all u (Lebesgue)

$$
f[u-\eta(\bar\theta,x)+\eta(\theta,x)] = f[u+\eta(\bar\theta,x)-\eta(\theta,x)] = f_e^s(u)\,.
$$

From the identifiability condition (3), this implies $\theta=\bar\theta$. The same is true for the conditional entropy of the symmetrized residuals

$$
\begin{aligned}
\mathtt{E}_\mu\{H(f_{e,X}^s)\} &= -\int_{\mathcal{X}}\left[\int_{-\infty}^{\infty} f_{e,x}^s(u)\log[f_{e,x}^s(u)]\,du\right]\mu(dx)\\
&\geq -\int_{\mathcal{X}}\left[\int_{-\infty}^{\infty} f(u)\log[f(u)]\,du\right]\mu(dx) = H(f)\,,
\end{aligned}
$$

with equality attained only if $H(f_{e,x}^s)=H(f)$ for μ-almost all x, which again implies $\theta=\bar\theta$. Notice that from a classical result in information theory,

$$
\mathtt{E}_\mu\{H(f_{e,X}^s)\} \leq H(f_e^s)\,.
$$

We can perform a local study of $H(f_e^s)$ around $\theta=\bar\theta$. Direct calculation gives

$$
\nabla f_e^s(u)_{|\theta=\bar\theta} = \mathbf{0}\,,\ \text{and}\ \nabla^2 f_e^s(u)_{|\theta=\bar\theta} = f''(u)\int_{\mathcal{X}}\nabla\eta(\bar\theta,x)\,[\nabla\eta(\bar\theta,x)]^\top\,\mu(dx).
$$

The entropy of f_e^s satisfies

$$
\begin{aligned}
\nabla H(f_e^s) &= -\int_{-\infty}^{\infty}[1+\log f_e^s(u)]\,\nabla f_e^s(u)\,du,\\
\nabla^2 H(f_e^s) &= -\int_{-\infty}^{\infty}\frac{1}{f_e^s(u)}\nabla f_e^s(u)[\nabla f_e^s(u)]^\top\,du\\
&\quad -\int_{-\infty}^{\infty}[1+\log f_e^s(u)]\,\nabla^2 f_e^s(u)\,du\,.
\end{aligned}
$$

Since $(f\log f)'' = (1+\log f)f'' + (f')^2/f$, we obtain

$$
\nabla H(f_e^s)_{|\theta=\bar\theta} = \mathbf{0}\,,\ \nabla^2 H(f_e^s)_{|\theta=\bar\theta} = \mathbf{M}_F(\bar\theta)\,,
$$

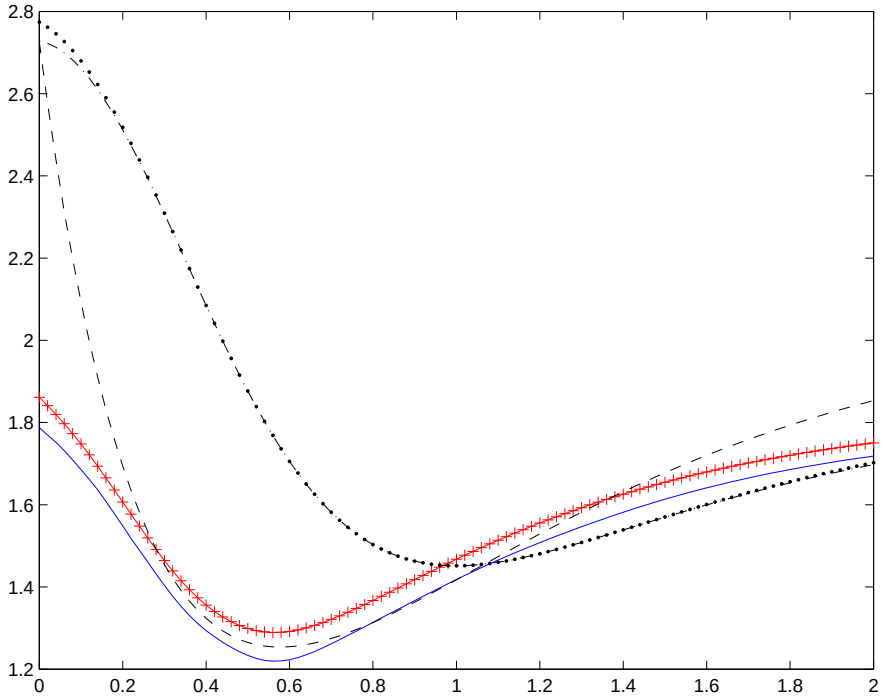

Fig. 1. Behaviors of different criteria plotted as functions of θ: $H(f_e^s)$ (dots) is almost undistinguishable from $\mathtt{E}_\mu\{H(f_{e,X}^s)\}$ (dash-dotted line), although always larger; the dashed line is for the ML criterion (4), the curve with crosses corresponds to the estimate (7) (with $A_n = \infty$), the full line to (13) (with $U_n \equiv 1$). The model is $\eta(\theta, x) = \exp(-\theta)$, with $n = 100$ observations, 10 at each $X^j = 1 + (j-1)/9$, $j = 1, \ldots, 10$, $\bar{\theta} = 1$, the ε_i have the Laplace density $f(u) = (1/\sqrt{2})\exp(-\sqrt{2}|u|)$.

that is, the entropy $H(f_e^s)$ is locally concave with zero derivative at $\theta = \bar{\theta}$. Similar results are obtained for the conditional entropy $\mathtt{E}_\mu\{H(f_{e,X}^s)\}$,

$$\forall x \in \mathcal{X}\,,\ \nabla f_{e,x}^s(u)_{|\theta=\bar{\theta}} = \mathbf{0}\,,\quad \nabla^2 f_{e,x}^s(u)_{|\theta=\bar{\theta}} = f''(u)\,\nabla\eta(\bar{\theta}, x)\,[\nabla\eta(\bar{\theta}, x)]^\top\,,$$

and

$$\nabla \mathtt{E}_\mu\{H(f_{e,X}^s)\}_{|\theta=\bar{\theta}} = \mathbf{0}\,,\ \nabla^2 \mathtt{E}_\mu\{H(f_{e,X}^s)\}_{|\theta=\bar{\theta}} = \mathbf{M}_F(\bar{\theta})\,.$$

Both entropies are presented in Figure 1 as functions of θ for a nonlinear one-parameter model.

The ML criterion $\bar{H}_n(\theta)$ given by (4) cannot be used since f is unknown. The situation is similar for $H(f_e^s)$ and $\mathtt{E}_\mu\{H(f_{e,X}^s)\}$ that use f and $\bar{\theta}$. To define a criterion approaching $H(f_e^s)$ we can simply plug a symmetric kernel estimate $\hat{f}_n^\theta$ of f_e^s in H, to obtain $\bar{H}_n(\theta) = H(\hat{f}_n^\theta)$. For technical reasons, we introduced a truncation of the integral in (Pronzato and Thierry, 2001a,b), and the criterion

$\hat{H}_n(\theta)$ to be minimized is then given by

$$\hat{H}_n(\theta) = -\int_{-A_n}^{A_n} \log[\hat{f}_n^\theta(u)]\hat{f}_n^\theta(u)du\,, \tag{7}$$

where (A_n) is a suitably (slowly) increasing sequence of positive numbers (to be chosen in accordance with the decrease of the bandwidth h_n of the kernel estimate $\hat{f}_n^\theta$, see Section 4). Such an estimate is plotted in Figure 1 (curve with crosses). Similarly, when a kernel estimate $f_{n_i}^{i,\theta}$ of the conditional distribution f_{e,X_i}^s can be constructed (design with replications, see Section 5.1), we can plug it in H to approach $\mathtt{E}_\mu\{H(f_{e,X}^s)\}$.

This construction can be formally justified following an approach similar to Beran (1978). We consider here the simple case of the location model, but the justification remains valid for a general nonlinear regression model when the design consists of replications and the entropy of f_{e,X_i}^s can be estimated for each i, see Section 5.1. In a location model, the distribution of the observations Y_i has the density $g(y) = f(y - \bar{\theta})$. Let us define $\hat{\beta} = (\hat{\theta}, \hat{f})$ as a couple of postulated values for θ and f in the semi-parametric model, with $\hat{f}$ symmetric. The associated density for the observations can then be expressed as $g_{\hat{\beta}}(y) = \hat{f}(y - \hat{\theta})$. Consider a kernel estimate $\hat{g}_n$ of g, formed from the observations $Y_1, \ldots, Y_n$,

$$\hat{g}_n(u) = \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{u - Y_i}{h_n}\right)$$

with K a symmetric kernel function, for instance the normal density. The estimator in (Beran, 1978) is obtained by minimizing the Hellinger distance [2] between $\hat{g}_n$ and $g_{\hat{\beta}}$. Consider instead the Kullback-Leibler divergence

$$\mathcal{L}(\hat{g}_n, g_{\hat{\beta}}) = \int_{-\infty}^{\infty} \log[\hat{g}_n(y)/g_{\hat{\beta}}(y)]\hat{g}_n(y)dy$$

as the ϕ-divergence to be minimized. We then need to maximize the term $\int_{-\infty}^{\infty} \log\left[\hat{f}(u - \hat{\theta})\right] \hat{g}_n(u)\, du$ with respect to $\hat{\beta}$. Write $\hat{f}(u) = [h(u)h(-u)]^{1/2}$ (h is not necessarily a p.d.f.), the maximum is obtained when h minimizes $\mathcal{L}\left([\hat{g}_n(u + \theta) + \hat{g}_n(-u + \theta)]/2, h(u)\right)$. Since $\mathcal{L}(p, q)$ is minimal when p and q are proportional, we need to select $h^\star(u) = [\hat{g}_n(u + \theta) + \hat{g}_n(-u + \theta)]/2$, and thus

$$\hat{f} = \hat{f}_n^\theta(u) = \frac{1}{2}\left[\hat{g}_n(u + \theta) + \hat{g}_n(-u + \theta)\right]$$

[2] One may refer, *e.g.*, to Beran (1977); Pak (1996) for parameter estimation based on the Hellinger distance in parametric models.

$$= \frac{1}{2nh_n} \sum_{i=1}^{n} K\left(\frac{u-(Y_i-\theta)}{h_n}\right) + K\left(\frac{u+(Y_i-\theta)}{h_n}\right). \tag{8}$$

The value of θ that minimizes $\mathcal{L}(\hat{g}_n, g_{\hat{\beta}})$ then corresponds to

$$\hat{\theta}^n = \arg\min_{\theta} H(\hat{f}_n^{\theta}), \tag{9}$$

with $\hat{f}_n^{\theta}$ given by (8), which is thus a kernel estimate based on the symmetrized residuals $Y_i - \theta, -Y_i + \theta$.

Other entropy estimates than (7) could be used to estimate θ (in particular, another type of plug-in estimate of the entropy of $\hat{f}_n^{\theta}$ will be considered in the next sections). In fact, one motivation for the entropy-minimization approach is that it allows a lot of flexibility: many methods are available to estimate the entropy $\hat{H}_n(\theta)$, and kernel estimation of the conditional distribution f^s_{e,X_i}, see (5), or of the marginal f^s_e, see (6), is only one possibility. One may refer to Beirlant et al. (1997) for a survey which includes plug-in, sample spacing and nearest neighbor methods. Different types of consistency results are obtained (weak, strong, L_2, $\sqrt{n}$, ...) depending on the method and the assumptions, in particular on f. The application of these methods to semi-parametric estimation via entropy minimization is quite challenging.

3 Adaptivity of a two-step method with data splitting

As in (Bickel, 1982) we consider randomized designs, for which the X_i's are i.i.d. with measure μ, independently of the measurement errors ε_i [3].

The method is termed *two-step* since it is based on a preliminary estimate $\hat{\theta}_1^n$ (which uses all data points $Y_1, \ldots, Y_n$). This estimate is assumed to be asymptotically locally sufficient (in the sense of Le Cam 1969). In our situation, the standard LS estimator $\hat{\theta}^n_{LS} = \arg\min_{\theta\in\Theta} \sum_{i=1}^n [Y_i - \eta(\theta, X_i)]^2$ can be used as preliminary estimate.

The method uses *data splitting* in the sense that the data set is split into two parts $(Y_1, \ldots, Y_m)$ and $(Y_{m+1}, \ldots, Y_n)$, with $m = m(n) \to \infty$ and $m/n \to 0$ as $n \to \infty$. The first data set $(Y_1, \ldots, Y_m)$ is used to construct an estimate $\hat{\theta}_1^m$ (similar to $\hat{\theta}_1^n$ but with m data points only) and to obtain residuals $e_i(\hat{\theta}_1^m)$, $i = 1, \ldots, m$. Then we construct a kernel estimate $\hat{f}_m$ from the $2m$ symmetrized

[3] More generally, one might consider the situation where the empirical distribution of the X_i's converges almost surely at rate $\sqrt{n}$ to some distribution function G on $\mathcal{X}$, see Manski (1984).

residuals $\pm e_i(\hat{\theta}_1^m)$, $i = 1, \ldots, m$,

$$\hat{f}_m(u) = \frac{1}{2m\, h_m} \sum_{i=1}^{m} \left[K\left(\frac{u - e_i(\hat{\theta}_1^m)}{h_m} \right) + K\left(\frac{u + e_i(\hat{\theta}_1^m)}{h_m} \right) \right] ,$$

with $K(.)$ a suitable kernel function (*e.g.* the normal density).

The estimator is obtained by performing one single modified Newton step for the minimization of an estimate $\hat{H}_n(\theta)$ of the entropy of the marginal f_e^s, see (6), starting at $\hat{\theta}_1^n$. The estimate $\hat{H}_n(\theta)$ that we consider is only valid for θ close to $\hat{\theta}_1^m$ but is such that the method coincides with that of Bickel (1982) and Manski (1984) [4] :

$$\hat{H}_n(\theta) = -\frac{1}{n-m} \sum_{i=m+1}^{n} \log \hat{f}_m[e_i(\theta)] \,. \tag{10}$$

A pure Newton step would give

$$\hat{\theta}^n = \hat{\theta}_1^n - [\nabla^2 \hat{H}_n(\hat{\theta}_1^n)]^{-1} \nabla \hat{H}_n(\hat{\theta}_1^n) \,,$$

with

$$\begin{aligned} \nabla \hat{H}_n(\theta) &= -\frac{1}{n-m} \sum_{i=m+1}^{n} \frac{\nabla \hat{f}_m[e_i(\theta)]}{\hat{f}_m[e_i(\theta)]} \\ &= \frac{1}{n-m} \sum_{i=m+1}^{n} \rho_m[e_i(\theta)]\, \nabla \eta(\theta, X_i) \,, \end{aligned} \tag{11}$$

where $\rho_m = (\hat{f}_m)'/\hat{f}_m$. Define the following truncated version of ρ_m,

$$\underline{\rho}_m(u) = \rho_m(u) U_m(u) \,,$$

where $U_m(u) = 0$ if $|u| > a_m$ of $\hat{f}_m(u) < b_m$ or $|(\hat{f}_m)'(u)| > c_m\, \hat{f}_m(u)$, and consider $\underline{\nabla} \hat{H}_n(\theta)$, given by (11) but with $\underline{\rho}_m$ substituted for ρ_m. The modified Newton step uses

$$\hat{\theta}^n = \hat{\theta}_1^n - \mathbf{M}_n^{-1}(\hat{\theta}_1^n) \underline{\nabla} \hat{H}_n(\hat{\theta}_1^n)$$

with $\mathbf{M}_n(\theta)$ an approximation of $\nabla^2 \hat{H}_n(\theta)$,

$$\mathbf{M}_n(\theta) = \frac{\mathcal{I}_n}{m-n} \sum_{i=m+1}^{n} \nabla \eta(\theta, X_i) [\nabla \eta(\theta, X_i)]^\top$$

[4] Their approach does not rely on entropy minimization, but constructs an approximation of the score function used in ML estimation. The estimate $\hat{H}_n$ below is thus more an approximation of $\bar{H}_n$, see (4), than of $H(f_e^s)$. See also the discussion in Section 6.

where

$$\mathcal{I}_n = \frac{1}{n-m} \sum_{i=m+1}^{n} \underline{\rho}_m^2[e_i(\hat{\theta}_1^m)] \,.$$

This construction of $\hat{\theta}^n$ coincides with the estimator of Bickel (1982) and Manski (1984) who show that it is adaptive when the tuning parameters h_m, a_m, b_m and c_m satisfy $h_m \to 0$, $a_m \to \infty$, $b_m \to 0$, $c_m \to \infty$, $m^{-1}a_m h_m^{-3} \to 0$ and $h_m c_m \to 0$ as $m \to \infty$.

In fact, as noticed in (Manski, 1984), data splitting is only used for technical reasons: the fact that the residuals $e_i(\theta)$, $i = m+1, \ldots, n$ and the kernel estimate $\hat{f}_m$ are based on independent samples facilitates the proof of adaptivity. Also, $\hat{f}_m$ in (10) does not depend explicitly on θ, unlike the estimator constructed in Section 4. However, Stone's estimator for location (1975) does not use data splitting, at the expense of a more delicate proof, see also Andrews (1989). Moreover, the numerical results presented in (Manski, 1984) show that data splitting degrades the performances of the estimator. For that reason, a more direct method is presented in the next section, although its adaptivity remains an open issue, see (Pronzato et al., 2004).

4 Direct entropy minimization

We consider the case of a location model $Y_i = \bar{\theta} + \varepsilon_i$, $i = 1, \ldots, n$, and refer to Section 5 for the extension to nonlinear regression. For any $\theta \in \Theta$ we form the n residuals $e_i(\theta) = Y_i - \theta$, $i = 1, \ldots, n$, and construct

$$\hat{f}_n^\theta(u) = \frac{1}{2}\left[k_n^\theta(u) + k_n^\theta(-u)\right] \tag{12}$$

to be used to compute $\hat{H}_n(\theta)$ given by (7), with

$$k_n^\theta(u) = \frac{1}{n\,h_n} \sum_{i=1}^{n} K\left(\frac{u - e_i(\theta)}{h_n}\right)$$

where $K(.)$ is symmetric about zero. $\hat{f}_n^\theta$ is then a kernel density estimate based on the $2n$ symmetrized residuals $\pm e_i(\theta)$. We assume some standard regularity assumptions for $K(.)$ (such that $\int_{-\infty}^{\infty} |u|K(u)du < \infty$, K is two times continuously differentiable with derivatives of bounded variation, see Schuster (1969)). A classical choice is the normal density $K(u) = 1/\sqrt{2\pi}\,\exp(-u^2/2)$. Alternatively, we also consider the following entropy estimator, to be compared to (10),

$$\hat{H}_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_{n,i}^\theta[e_i(\theta)]\, U_n[e_i(\theta)] \tag{13}$$

where $\hat{f}^{\theta}_{n,\,i}$ is similar to (12), but does not use $e_i(\theta)$, that is,

$$\hat{f}^{\theta}_{n,\,i}(u) = \frac{1}{2}\left[k^{\theta}_{n,\,i}(u) + k^{\theta}_{n,\,i}(-u)\right] \tag{14}$$

with

$$k^{\theta}_{n,\,i}(u) = \frac{1}{(n-1)\,h_n}\sum_{j=1,\,j\neq i}^{n} K\left(\frac{u - e_j(\theta)}{h_n}\right),\; i = 1,\ldots,n\,.$$

In (13), U_n defines a smooth truncation for large residuals, $U_n(z) = U(|z|/A_n - 1)$ with $U(z) = 1$ for $z \leq 0$, 0 for $z \geq 1$ and $U(z)$ varying smoothly between 0 and 1, with $U'(0) = U'(1) = 0$, $\max_z\,|U'(z)| = d_1 < \infty$, $\max_z\,|U''(z)| = d_2 < \infty$.

Define $\hat{\theta}^n = \arg\min_{\theta\in\Theta}\hat{H}_n(\theta)$, with $\hat{H}_n(\theta)$ given by (7) or (13). Notice that $\hat{H}_n(\theta)$ is two times continuously differentiable w.r.t. $\theta \in \mathrm{int}(\Theta)$. Convergence in probability when $n\to\infty$ will be denoted $\xrightarrow{\mathrm{p}}$ ($\overset{\theta,\mathrm{p}}{\rightsquigarrow}$ will be used when the convergence is uniform with respect to θ), and convergence in distribution will be denoted $\xrightarrow{\mathrm{d}}$. Under common measurability conditions (see, *e.g.*, Lemmas 2 and 3 of Jennrich (1969)) the standard, and rather general, approach for proving asymptotic normality (and hopefully asymptotic efficiency) of $\hat{\theta}^n$ minimizing some criterion $\hat{H}_n(\theta)$ can be decomposed into three steps:

A) show that $\hat{H}_n(\theta)\overset{\theta,\mathrm{p}}{\rightsquigarrow}H(\theta)$, $n\to\infty$, with $\hat{H}_n(\theta)$ continuous in θ for any n, and that $H(\bar{\theta}) < H(\theta)$ for any $\theta \neq \bar{\theta}$;

B) show that $\nabla^2\hat{H}_n(\theta)\overset{\theta,\mathrm{p}}{\rightsquigarrow}\nabla^2 H(\theta)$, $n\to\infty$, with $\nabla^2 H(\bar{\theta})$ positive definite ($\succ 0$);

C) decompose $\nabla\hat{H}_n(\bar{\theta})$ into $\nabla\bar{H}_n(\bar{\theta}) + \Delta_n(\bar{\theta})$, with $\sqrt{n}\nabla\bar{H}_n(\bar{\theta})\xrightarrow{\mathrm{d}}\mathcal{N}(\mathbf{0},\mathbf{M}_1)$ and $\sqrt{n}\Delta_n(\bar{\theta})\xrightarrow{\mathrm{p}}\mathbf{0}$ as $n\to\infty$.

The uniform convergence in (A) proves the weak consistency of $\hat{\theta}_n$ ($\hat{\theta}^n\xrightarrow{\mathrm{p}}\bar{\theta}$). (A) and (B) imply that $\nabla^2\hat{H}_n(\hat{\theta}^n)\xrightarrow{\mathrm{p}}\mathbf{M}_2 = \nabla^2 H(\bar{\theta}) \succ 0$ as $n\to\infty$. Finally, consider the following Taylor expansion of $\nabla\hat{H}_n(\theta)$ at $\theta = \hat{\theta}^n$,

$$\nabla\hat{H}_n(\hat{\theta}^n) = \mathbf{0} = \nabla\hat{H}_n(\bar{\theta}) + (\hat{\theta}^n - \bar{\theta})^{\top}\nabla^2 H[\alpha_n\hat{\theta}^n + (1-\alpha_n)\bar{\theta}]\,,$$

with $\alpha_n \in [0,1]$ (see Jennrich (1969) who uses a similar approach for LS estimation). (C) then implies asymptotic normality, that is, $\sqrt{n}(\hat{\theta}^n - \bar{\theta})\xrightarrow{\mathrm{d}}\mathcal{N}(\mathbf{0},\mathbf{M}_2^{-1}\,\mathbf{M}_1\,\mathbf{M}_2^{-1})$. The adaptivity of $\hat{\theta}^n$ would then directly follow from $\mathbf{M}_2^{-1}\mathbf{M}_1\mathbf{M}_2^{-1} = \mathbf{M}_F^{-1}(\bar{\theta})$, the inverse of the Fisher information matrix (2).

One may notice that step (C) allows some freedom in the choice of the function $\bar{H}_n(\theta)$, even though it would be natural to pick (4), for which the asymptotic normality $\sqrt{n}\nabla\bar{H}_n(\bar{\theta})\xrightarrow{\mathrm{d}}\mathcal{N}(\mathbf{0},\mathbf{M}_1)$ holds under standard assumptions, with $\mathbf{M}_1 = \mathbf{M}_F(\bar{\theta})$ (asymptotic properties of the ML estimator). Also, according

to the review (Beirlant et al., 1997), $\sqrt{n}$-consistency of $\hat{H}_n(\theta)$ is difficult to obtain, but notice that it is not a prerequisite for $\sqrt{n}$-consistency of $\hat{\theta}^n$ (we only need $\sqrt{n}\Delta_n(\bar{\theta})\xrightarrow{\mathrm{p}}0$, $n \to \infty$).

In the case of the location model, assuming that, additionally to the assumptions of previous sections, f has unbounded support, f and its derivatives f', f'' and f''' are bounded and that there exists a strictly increasing function B such that for all $u \in I\!R$, $B(u) \geq \sup_{|y|<u} 1/f(y)$, we can prove (A) for the entropy estimate (13) provided that $B_n = B(3A_n)$ (respectively h_n) increases (respectively decreases) slowly enough ($B_n = n^\alpha$, $h_n = 1/(n^\alpha \log n)$ with $\alpha < 1/3$ is suitable). The proof is based on (Dmitriev and Tarasenko, 1973, Theorem 4) and (Newey, 1991, Corollary 3.1). Similarly, with slightly stronger conditions on f we can prove (B), that is, $\nabla^2 \hat{H}_n(\theta) \overset{\theta;\mathrm{p}}{\rightsquigarrow} \nabla^2 H(\theta)$, with $\nabla^2 H(\bar{\theta}) = \mathcal{I}(f)$, the fisher information for location ($B_n = n^\alpha$, $h_n = 1/(n^\alpha \log n)$ with $\alpha < 1/7$ is suitable). A key step to prove adaptivity of $\hat{\theta}^n$ at step (C) would be to show that

$$-\frac{2}{\sqrt{n}} \sum_{i=1}^{n} \frac{(k_{n,i}^{\bar{\theta}})'(-\varepsilon_i)}{k_{n,i}^{\bar{\theta}}(\varepsilon_i) + k_{n,i}^{\bar{\theta}}(-\varepsilon_i)} U_n(\varepsilon_i) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{I}(f)) \,,$$

the term on the left-hand side being the major contribution to $\sqrt{n}\nabla \hat{H}_n(\bar{\theta})$ when $\hat{H}_n(\theta)$ is given by (13). The conditions required on the functions f, K and U for this to hold are currently under investigation [5].

We conclude this section by some simulation results that illustrate the good finite sample behavior of minimum-entropy estimation in the location model, see Table 1. The estimators used in the comparison are the ordinary Least Squares (LS) estimator, the Minimum Hellinger Distance (MHD) estimator (Beran, 1978) and the Minimum-Entropy estimators minimizing (7) for ME_1 and (13) for ME_2. We take $A_n = \infty$ in (7) and $U_n(x) \equiv 1$ in (13). After considering different smoothing techniques, using the broad study of Berlinet and Devroye (1994), we selected the bandwidth h_n of the kernel estimators (12) and (14) with the double kernel method, based on residuals obtained from a robust M-estimator. We compare the empirical value $\hat{C}_n$ of the covariance $\mathtt{E}\{\hat{v}_n \hat{v}_n{}^\top\}$, with $\hat{v}_n = \sqrt{n}(\hat{\theta}^n - \bar{\theta})$, for the different estimators, making 100 repetitions of the estimation procedure with $n = 100$ observations each. Also, we compare the methods for different noise distributions: the standard normal, the bi-exponential or Laplace density ($f(u) = (1/\sqrt{2})\exp(-\sqrt{2}|u|)$), and Student's t_ν distributions with $\nu = 3, 5$ and 10 degrees of freedom. The optimum asymptotic values of $\hat{C}_n$, $\mathbf{M}_F^{-1}(\bar{\theta})$, obtained for ML estimation, are also given in Table 1.

[5] One difficulty which is not present in two-step approach with data splitting of Section 3 is due to the fact that $(k_{n,i}^{\bar{\theta}})'(-u) \neq -(k_{n,i}^{\bar{\theta}})'(u)$. On the other hand, $\rho_m(-u) = -\rho_m(u)$ in (11).

Table 1
Values of $\hat{C}_n$ in the location model.

f	$\mathcal{N}(0,1)$	exp	t_3	t_5	t_{10}
$\mathbf{M}_F^{-1}$	1	0.5	0.5	0.8	0.9455
LS	1.09	0.94	1.13	0.96	1.03
MHD	1.12	0.72	0.50	0.86	1.0
ME_1	1.12	0.71	0.48	0.83	0.99
ME_2	1.19	0.74	0.57	0.84	0.98

5 Entropy minimization in nonlinear regression

5.1 Design with replications

Assume that the design consists of replications at fixed points $X^1, \ldots, X^m$, where X^j receives the weight μ_j. The design measure μ thus has a finite number of support points, and, for a total of n observations, $n_j = n\mu_j$ are made at $X = X^j$.

We consider first a two-stage method, with m minimum-entropy estimations at the first stage and one (weighted) LS estimation at the second. For each X^j, $j = 1, \ldots, m$, we solve a location problem. Let Y_{j_i} denote the observations made at $X = X^j$, $i = 1, \ldots, n_j$, and $\hat{\eta}^j$ denote the estimated response at $X = X^j$ obtained by a minimum-entropy estimator. If $\hat{\eta}^j$ is adaptive, $\sqrt{n_j}[\hat{\eta}^j - \eta(\bar{\theta}, X^j)] \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{I}^{-1}(f))$, $n_j \to \infty$. Having solved m location problems of this type, we form a LS estimation problem by considering the estimated responses $\hat{\eta}^j$ as pseudo-observations, and minimize $J_n(\theta) = \sum_{j=1}^m \mu_j[\eta(\theta, X^j) - \bar{\eta}^j]^2$. One can then show that the estimator $\hat{\theta}^n$ that minimizes $J_n(\theta)$ satisfies $\sqrt{n}(\hat{\theta}^n - \bar{\theta}) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \mathbf{M}_F^{-1}(\bar{\theta}))$. Adaptivity in nonlinear regression with replications thus directly follows from adaptivity in the location model.

One can expect a one-stage estimator to exhibit a better finite sample behavior than the two-stage procedure above. Using a justification similar to that given in Section 4 for the location model, we suggest the following method:

(i) form the kernel estimates $\hat{f}_{n_j}^{j,\theta}$ of the conditional distributions f_{e,X_j}^s of (symmetrized) residuals *for each design point X^j separately*, using (12) or (14) and estimate their respective entropies $\hat{H}_{n_j}(\theta, X^j)$, using (7) or (13);

(ii) compute the conditional entropy

$$\mathtt{E}_\mu\{\hat{H}_n(\theta, X)\} = \sum_{j=1}^{m} \mu_j \hat{H}_{n_j}(\theta, X^j)\,. \tag{15}$$

Again, the adaptivity of $\hat{\theta}^n$ that minimizes $\mathtt{E}_\mu\{\hat{H}_n(\theta, X)\}$ would follow from adaptivity in the location model. However, the conditional entropy (15) can only be estimated in the case of designs consisting of replications, and this approach does not extend to more general designs.

5.2 General situation: an upper bound on the conditional entropy

Suppose that in the situation of Section 5.1 we mix all (symmetrized) residuals together and estimate the entropy $\hat{H}_n(\theta)$ of their marginal distribution by (7) or (13). Replace μ by μ^n in (15), with μ^n the empirical measure of the design points X_i. Let U be a random variable with distribution conditional on $X = X^j$ given by $\hat{f}^{j,\theta}_{n_j}$. Then, $\mathtt{E}_{\mu^n}\{\hat{H}_n(\theta, X)\} = \mathcal{H}(U|X)$ the conditional entropy of U given X, and, as mentioned in Section 2, $\mathcal{H}(U|X) \leq \mathcal{H}(U) = \hat{H}_n(\theta)$, the entropy obtained by mixing up all residuals. The latter can be constructed for any design, and forms an upper bound on the criterion $\mathtt{E}_{\mu^n}\{\hat{H}_n(\theta, X)\}$ given by (15). Figure 2 presents $\hat{H}_n(\theta)$ and $\mathtt{E}_{\mu^n}\{\hat{H}_n(\theta, X)\}$ for the same nonlinear one-parameter model as in Figure 1 when the entropy estimation uses (13) (with $U_n \equiv 1$) ($\hat{H}_n(\theta)$ is also presented in Figure 1, where it can be seen that estimation by (7) or (13) produces similar results).

5.3 Example

We take $\eta(\theta, x) = \theta_1 \exp(-\theta_2 x)$, with $\bar{\theta} = (100, 2)^\top$, the design measure μ is supported at $X^j = 1 + (j-1)/9$, $j = 1, \ldots, 10$, with weights $\mu_j = 1/10$ for all j. We compare the performances of the ordinary Least Squares estimator (LS), the Minimum Hellinger Distance estimator (MHD) of Beran (1978) and the Minimum-Entropy estimator (ME) minimizing (7) with $A_n = \infty$ and h_n obtained by the double kernel method applied on residuals of a robust M-estimation. We make 100 replications of the experiment using $n = 100$ observations and compare the empirical covariance matrices $\hat{C}_n$ of $\hat{v}_n = \sqrt{n}(\hat{\theta}^n - \bar{\theta})$ for the different estimators and different distributions of the measurement errors ε_i (standard normal, bi-exponential and Student's t_ν distributions with $\nu = 3, 5$ and 10 degrees of freedom). Both MHD and ME mix all residuals together (kernel estimates of the marginal distributions are used). The trace and determinant of $\hat{C}_n$ are compared to $\mathrm{trace}(\mathbf{M}_F^{-1})$ and $\det(\mathbf{M}_F^{-1})$ obtained

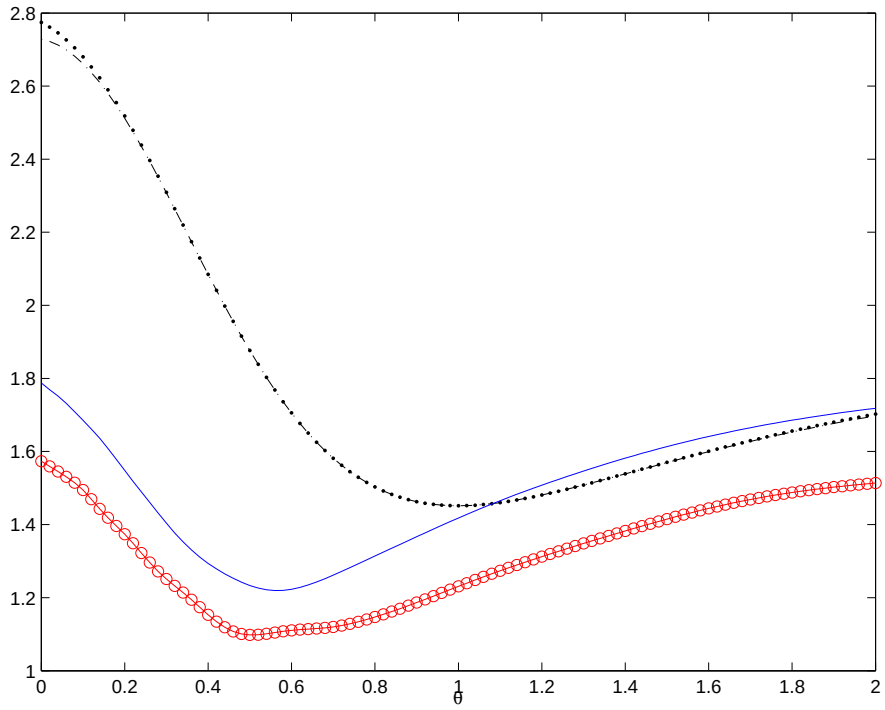

Fig. 2. Behaviors of different criteria plotted as functions of θ for the same example as in Figure 1: $H(f_e^s)$ (dots), $\mathtt{E}_\mu\{H(f_{e,X}^s)\}$ (dash-dotted line), $\hat{H}_n(\theta)$ (full line) and $\mathtt{E}_{\mu^n}\{\hat{H}_n(\theta, X)\}$ (circles).

Table 2
Values (T_n, D_n) of the trace ($\times 10^{-3}$) and determinant ($\times 10^{-2}$) of $\hat{C}_n$ for different f and estimators

f	$\mathcal{N}(0,1)$	exp	t_3	t_5	t_{10}
$\hat{C}_n = \mathbf{M}_F^{-1}$	(6.2, 0.8)	(3.1, 0.2)	(3.1, 0.2)	(4.9, 0.5)	(5.8, 0.75)
LS	(8.8, 1.2)	(13.6, 3.6)	(9.1, 2.0)	(9.7, 2.2)	(9.7, 2.1)
MHD	(9.1, 1.4)	(3.8, 0.5)	(6.4, 0.6)	(7.9, 1.4)	(7.1, 1.7)
ME	(9.2, 1.25)	(3.8, 0.4)	(4.9, 0.4)	(7.8, 1.2)	(6.8, 1.35)

asymptotically ($n \to \infty$) by ML estimation. The results are presented in Table 2.

Finally, in order to illustrate the robustness properties of the minimum-entropy estimator, we introduce q outliers, $q = 20, 40, 60, 80$, in addition to the $n = 100$ regular observations, when f corresponds to the Laplace distribution. They correspond to errors ε_i normally distributed $\mathcal{N}(10, 4)$ with $q/10$ observations at each of the X^j's.

Table 3 presents the results for trace$(\hat{C}_n)$ and det$(\hat{C}_n)$ obtained in this case and shows that minimum-entropy estimation (ME_1 for (7) and ME_2 for (13))

Table 3
Values (T_n, D_n) of the trace ($\times 10^{-3}$) and determinant ($\times 10^{-2}$) of $\hat{C}_n$ when q outliers are added to the $n = 100$ regular observations, f corresponds to the Laplace distribution.

q	20	40	60	80
LS	(84.25, 42.8)	(146.25, 67.0)	(184.45, 58.2)	(208.7, 58.5)
MHD	(4.25, 0.9)	(12.7, 2.25)	(23.4, 4.5)	(56.95, 19.8)
ME_1	(4.0, 0.5)	(9.8, 1.7)	(6.0, 0.9)	(6.7, 13.6)
ME_2	(3.95, 0.5)	(8.4, 1.7)	(5.5, 0.9)	(10.9, 39.7)

is robust with respect to the presence of outliers. Figure 3 illustrates the situation when $q = 40$ outliers have been introduced. On both sides of the figure the dashed line corresponds to the true density of the errors ε_i (Laplace). The full line corresponds to the estimated density $\hat{f}_n^{\hat{\theta}^n}$ of the residuals, with $\hat{\theta}^n$ the minimum-entropy estimator on the left-hand side and $\hat{\theta}^n$ the LS estimator on the right-hand side. The positions of the residuals along the horizontal axis are indicated by stars (the vertical position is arbitrary). Note that the minimum entropy estimator manages to separate the regular observations from the outliers, whereas all residuals remain mixed for the LS estimator. Once the residuals are parted into distinct clusters, with an estimated density $\hat{f}_n^{\hat{\theta}^n}$ showing three distinct modes (on the left and right for the outliers, near zero for the regular data), the entropy does not change when the left and right clusters are translated further away from zero. Hence, the minimum-entropy estimate is not modified when the distance of outliers to regular data points becomes very large (a property similar to so-called redescending M-estimators).

6 Perspectives and open questions

Although in Section 3 we pointed out some connection between minimum-entropy estimation and the approach of Bickel (1982) and Manski (1984), there exists a basic difference that should not be undervalued: they approximate the score-function used in ML estimation, or equivalently the ML criterion (4), whereas we approximate the entropy of the distribution of residuals. That the two become close when the parameters θ are in the neighborhood of $\bar{\theta}$ is clear from (4): for θ around $\bar{\theta}$, the residuals are close to the errors ε_i and their estimated distribution is close to f. However, the difference may be important further away from $\bar{\theta}$ (although Figure 1 indicates that (4), (7) and (13) remain very similar for a reasonable parameter range). Also, the dependence of the kernel estimates (12) or (14) in θ makes the derivation of the asymptotic properties of $\hat{\theta}^n$ minimizing (7) or (13) much more difficult than for the minimizer of (10). In particular, the adaptivity of $\hat{\theta}^n$ is still an open

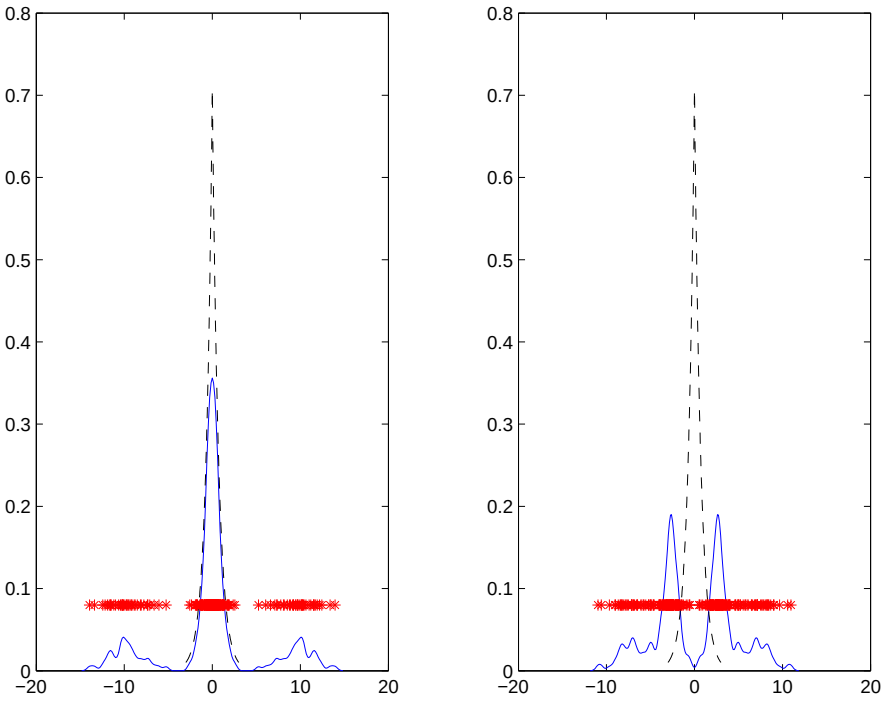

Fig. 3. 100 regular errors with Laplace distribution (dashed line), 40 outliers with $\varepsilon_i \sim \mathcal{N}(10, 4)$. Residuals and their estimated density: left for minimum entropy estimation, right for LS.

question.

The example of Section 5.3 shows that $\hat{\theta}^n$ is robust with respect to outliers, a property already investigated for the minimization of the Hellinger distance in (Beran, 1978). A similar study for $\hat{\theta}^n$ is still to be done. Displaying estimators asymptotically efficient for regular data and robust with respect to outliers would clearly be of practical importance in many signal and image processing applications.

The results presented here, and also most of those mentioned (Beran, 1978; Bickel, 1982; Manski, 1984, etc.), concern static systems only, in the sense that the errors ε_i are supposed to be independent (although the case of ε_i's and X_i's being interdependent is considered in (Manski, 1984)). The extension to α-mixing, or m-dependent sequences of errors would be important for signal processing applications (an example presented in (Pronzato and Thierry, 2001a,b) shows that minimum-entropy estimation still performs well in the presence of correlated errors — interference noise).

Finally, as mentioned in Section 2, several methods exist for entropy estimation, and each of them could be used to define a minimum-entropy estimator. Some do not require kernel smoothing of the empirical density of residuals, which could be considered as an advantage over the plug-in estimates used in this paper, see, *e.g.*, Vasicek (1976) for a sample-spacing method and

Kozachenko and Leonenko (1987) for an approach relying on nearest neighbors (in particular, the latter applies for samples in any dimension k, and could be used for minimum-entropy estimation in multiple regression where Y_i is then a k-dimensional vector). However, investigating the asymptotic properties of their associated minimum-entropy estimators seems a very difficult task. Another direction would be to consider recent developments in parametric estimation via divergence minimization. In a parametric context (which, for regression models, means that f is known), the asymptotically efficient estimator of Beran (1977), based on minimizing Hellinger distance, requires smoothing of the empirical distribution in order to compute its distance to a distribution with density. On the other hand, the approach used by Broniatowski (2003); Broniatowski and Keziou (2004) is based on a duality property that permits to estimate the divergences of interest without requiring smoothing. The application to the semi-parametric problem considered in the paper is an open and motivating issue.

References

Andrews, D., 1989. Asymptotics for semiparametric econometric models: III testing and examples. Cowles Foundation Discussion Paper No. 910, Yale University.

Ash, R., 1965. Information Theory. Wiley, New York, (Republished by Dover, New York, 1990).

Begun, J., Hall, W., Huang, W.-M., Wellner, J., 1983. Information and asymptotic efficiency in parametric-non parametric models. Annals of Statistics 11 (2), 432–452.

Beirlant, J., Dudewicz, E., Györfi, L., van der Meulen, E., 1997. Nonparametric entropy estimation; an overview. Intern. J. Math. Stat. Sci. 6 (1), 17–39.

Beran, R., 1974. Asymptotically efficient rank estimates in location models. Annals of Statistics 2, 63–74.

Beran, R., 1977. Minimum Hellinger distance estimates for parametric models. Annals of Statistics 5 (3), 445–463.

Beran, R., 1978. An efficient and robust adaptive estimator of location. Annals of Statistics 6 (2), 292–313.

Berlinet, A., Devroye, L., 1994. A comparison of kernel density estimates. Publications de l'institut de statistique de l'Université de Paris 38, 3–59.

Bickel, P., 1982. On adaptive estimation. Annals of Statistics 10, 647–671.

Broniatowski, M., 2003. Estimation through Kullback-Leibler divergence. Mathematical Methods of Statistics (to appear).

Broniatowski, M., Keziou, A., 2004. Parametric estimation and testing through divergences. Prepublication 2004-1, L.S.T.A., Université Paris 6.

Dmitriev, Y., Tarasenko, F., 1973. On the estimation of functionals of the probability density and its derivatives. Theory of Probability and its Appli-

cations 18 (3), 628–633.

Jennrich, R., 1969. Asymptotic properties of nonlinear least squares estimation. Annals of Math. Stat. 40, 633–643.

Kozachenko, L., Leonenko, N., 1987. On statistical estimation of entropy of random vector. Problems Infor. Transmiss. 23 (2), 95–101, (translated from Problemy Peredachi Informatsii, in Russian, vol. 23, No. 2, pp. 9-16, 1987).

Le Cam, L., 1969. Théorie Asymptotique de la Décision Statistique. Les Presses de l'Université de Montréal.

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

Newey, W., 1991. Uniform convergence in probability and stochastic equicontinuity. Econometrica 9 (4), 1161–1167.

Pak, R., 1996. Minimum Hellinger distance estimation in simple linear regression models; distribution and efficiency. Statistics & Probability Letters 26, 263–269.

Pronzato, L., Thierry, E., 2001a. Entropy minimization for parameter estimation problems with unknown distribution of the output noise. In: Proc. ICASSP'2001. Salt Lake City.

Pronzato, L., Thierry, E., 2001b. A minimum-entropy estimator for regression problems with unknown distribution of observation errors. In: Mohammad-Djafari, A. (Ed.), Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Proc. 20th Int. Workshop, Gif-sur-Yvette, France, July 2000. Am. Inst. of Physics, New York, pp. 169–180.

Pronzato, L., Thierry, E., Wolsztynski, E., 2004. Minimum entropy estimation in semi parametric models: a candidate for adaptive estimation? In: Di Bucchianico, A., Läuter, H., Wynn, H. (Eds.), mODa'7 – Advances in Model–Oriented Design and Analysis, Proceedings of the 7th Int. Workshop, Heeze (Netehrlands). Physica Verlag, Heidelberg, to appear.

Schuster, E., 1969. Estimation of a probability density function and its derivatives. Annals of Math. Stat. 40, 1187–1195.

Stein, C., 1956. Efficient nonparametric testing and estimation. In: Proc. 3rd Berkeley Symp. Math. Stat. Prob. Vol. 1. University of California Press, Berkeley, pp. 187–196.

Stone, C., 1975. Adaptive maximum likelihood estimators of a location parameter. Annals of Statistics 3 (2), 267–284.

Vasicek, O., 1976. A test for normality based on sample entropy. Journal of Royal Statistical Society B38 (1), 54–59.