A MINIMUM-ENTROPY ESTIMATOR FOR REGRESSION PROBLEMS WITH UNKNOWN DISTRIBUTION OF OBSERVATION ERRORS

L. PRONZATO AND E. THIERRY

ABSTRACT. We consider a nonlinear regression model, with independent observation errors identically distributed with an unknown probability density function $f(\cdot)$. Instead of minimising the empirical version of the entropy of $f(\cdot)$ based on the residuals, which corresponds to maximum likelihood estimation and requires the knowledge of $f(\cdot)$, we minimise the entropy of a (symmetrized) kernel estimate $\hat{f}_{n,h}$ of $f(\cdot)$, constructed from the residuals. Two examples are presented to illustrate the good finite-sample behaviour of this estimator (accuracy, robustness). Some (preliminary) consistency results are given.

1. INTRODUCTION

Consider a nonlinear regression model, with observations given by

$$y_i = \eta(\bar{\boldsymbol{\theta}}, \xi_i) + \epsilon_i \,, \; i = 1, \ldots, n \,, \tag{1.1}$$

where $\bar{\boldsymbol{\theta}} \in \Theta \subset I\!R^p$ is the unknown true value of the model parameters, with Θ a compact set, $\eta(\boldsymbol{\theta}, \xi)$ is the model response for parameters $\boldsymbol{\theta}$ and experimental conditions ξ, $\xi \in \mathcal{X} \subset I\!R^q$, and $\{\epsilon_i\}$ is a sequence of independently identically distributed (i.i.d.) errors, with probability density function (p.d.f.) $f(\cdot)$. We assume that $f(\cdot)$ is symmetric ($f(x) = f(-x)$), two times continuously differentiable, with derivatives $f'(\cdot)$ and $f''(\cdot)$, and that the Fisher information $\mathcal{I}(f) = \int_{-\infty}^{\infty} [f'(x)]^2 / f(x)\, dx$ exists. The function $\eta(\boldsymbol{\theta}, \xi)$ is assumed to be two times continuously differentiable w.r.t. $\boldsymbol{\theta}$. We define

$$e_i(\boldsymbol{\theta}) = y_i - \eta(\boldsymbol{\theta}, \xi_i)$$

and denote $\mathbf{e}_1^n(\boldsymbol{\theta}) = [e_1(\boldsymbol{\theta}), \ldots, e_n(\boldsymbol{\theta})]$. The vectors (ϵ_i, ξ_i) are assumed to be stochastically independent, that is, the design is *not sequential* (one can refer *e.g.* to [8, 16] for the difficulties of making statistical inference when the experiment is designed sequentially). We also assume that ϵ_i and ξ_i are mutually independent, and denote by μ the design measure, with $\int_{\mathcal{X}} \mu(d\xi) = 1$, see, *e.g.* [7]. We assume, moreover, that μ is a discrete measure, with $m < \infty$ support points $\xi^1, \ldots, \xi^m$ receiving weights $w_1, \ldots, w_m$. Since a random permutation of the observations does not modify the estimation of $\boldsymbol{\theta}$, the ξ_i's can be considered as forming an i.i.d. sequence, independent of $\{\epsilon_i\}$, with probability measure μ. We shall denote $\mathbf{M}(\bar{\boldsymbol{\theta}}, \mu)$ the Fisher information matrix

$$\mathbf{M}(\bar{\boldsymbol{\theta}}, \mu) = \mathcal{I}(f) \left(\int_{\mathcal{X}} \frac{\partial \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta}}_{|\bar{\boldsymbol{\theta}}} \frac{\partial \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta}^\top}_{|\bar{\boldsymbol{\theta}}} \mu(d\xi) \right) , \tag{1.2}$$

and assume that μ is not degenerate ($\mathbf{M}(\bar{\boldsymbol{\theta}}, \mu)$ is positive-definite).

Any reasonable approach for estimating $\boldsymbol{\theta}$ from the observations $y_1, \ldots, y_n$ aims at minimising some measure of the dispersion of the errors $\mathbf{e}_1^n(\boldsymbol{\theta})$. For instance, the Maximum Likelihood (ML) estimator of $\boldsymbol{\theta}$ minimises $-\int \log f_{\boldsymbol{\theta}}(y, \xi) dG_n(y, \xi)$, with G_n the empirical distribution of the observations y_i and experimental conditions ξ_i and $f_{\boldsymbol{\theta}}(y, \xi) = f[y - \eta(\boldsymbol{\theta}, \xi)]$; that is, $\hat{\theta}_{ML}$ simply minimises the sample version of the entropy of $f(\cdot)$, evaluated at $\mathbf{e}_1^n(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\min_{\boldsymbol{\theta} \in \Theta} - \sum_{i=1}^{n} \log f[e_i(\boldsymbol{\theta})] \,. \tag{1.3}$$

Date: November 6, 2000.
Key words and phrases. Parameter estimation, entropy, regression, maximum likelihood.

Under standard assumptions, see, *e.g.*, [9], $\hat{\boldsymbol{\theta}}_{ML}$ possesses the following asymptotic properties

$$\begin{aligned} &\text{consistency: } \hat{\boldsymbol{\theta}}_{ML} \xrightarrow{a.s.} \bar{\boldsymbol{\theta}}\,; \\ &\text{asymptotic normality: } \sqrt{n}(\hat{\boldsymbol{\theta}}_{ML}-\bar{\boldsymbol{\theta}}) \xrightarrow{d} z \sim \mathcal{N}(\mathbf{0}, \mathbf{M}^{-1}(\bar{\boldsymbol{\theta}},\mu))\,; \\ &\text{asymptotic efficiency: } \mathbf{M}(\bar{\boldsymbol{\theta}},\mu)^{-1} \text{ is the Cramer-Rao lower bound}\,. \end{aligned} \tag{1.4}$$

When $f(\cdot)$ is unknown, using a wrong distribution in the calculation of $\hat{\boldsymbol{\theta}}_{ML}$ makes the approach suboptimal. Robustness w.r.t. uncertainty on $f(\cdot)$ can be taken into account for instance by using the minimax approach of [14, 10], which corresponds to ML-estimation for the p.d.f. minimising $\mathcal{I}(f)$ in a given class (we shall come back to the robustness issue in Section 3, where robustness with respect to outliers is considered through an example). A quite different approach is considered in this paper: we estimate $\boldsymbol{\theta}$ by minimising the entropy of an estimate of $f(\cdot)$ based on the empirical distribution of the errors $\mathbf{e}_1^n(\boldsymbol{\theta})$.

Let $K(\cdot)$ denote a kernel weighting function (a Borel function) such that $\sup_{-\infty<y<\infty}|K(y)|<\infty$, $\lim_{y\to\infty}|yK(y)|=0$, $\int_{-\infty}^{\infty}|K(y)|dy<\infty$, $\int_{-\infty}^{\infty}K(y)dy=1$. We assume that $K(\cdot)$ is differentiable, with $K'(\cdot)$ its derivative, symmetric ($K(y)=K(-y)$) and positive. For any p.d.f. $f(\cdot)$, let $\hat{f}_{n,h}(\cdot)$ denote its kernel estimate based on $X_1,\ldots,X_n$, that is,

$$\hat{f}_{n,h}(x) = \hat{f}_{n,h}(x|X_1,\ldots,X_n) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\,.$$

The bandwidth h will be written h_n when it is taken as a function of the sample size n. Much attention has been paid to conditions under which $\hat{f}_{n,h_n}(\cdot)$ converges to $f(\cdot)$, in various senses, when $X_1,\ldots,X_n$ is i.i.d. with the p.d.f. $f(\cdot)$. For instance, [13] shows that $\hat{f}_{n,h_n}(x) \xrightarrow{P} f(x)$ at all points of continuity of $f(\cdot)$ when $h_n\to 0$ and $nh_n\to\infty$ as $n\to\infty$, and in that case the sequence $\{\hat{f}_{n,h_n}(x)\}$ is also asymptotically normal. It is shown in the same paper that when the Fourier transform of $K(y)$ is absolutely integrable and $f(x)$ is uniformly continuous, then $\hat{f}_{n,h_n}(x)$ is uniformly consistent in probability, that is, $\sup_{-\infty<x<\infty}|\hat{f}_{n,h_n}(x)-f(x)| \xrightarrow{P} 0$, if $h_n\to 0$ and $nh_n^2\to\infty$ when $n\to\infty$. Strong uniform consistency of $\hat{f}_{n,h_n}(x)$ ($\sup_{-\infty<x<\infty}|\hat{f}_{n,h_n}(x)-f(x)| \xrightarrow{a.s.} 0$) is proved in [2] under a different set of assumptions. Devroye and Györfi [3] show that $\int_{-\infty}^{\infty}|\hat{f}_{n,h_n}(x)-f(x)|dx \xrightarrow{a.s.} 0$ if and only if $h_n\to 0$ and $nh_n/\log n\to\infty$ as $n\to\infty$. Under reasonable assumptions on $f(\cdot)$, $\eta(\boldsymbol{\theta},\xi)$, $K(\cdot)$ and h_n, $\hat{f}_{n,h_n}[x|\mathbf{e}_1^n(\boldsymbol{\theta})]$ thus converges to $g(x)$ given by

$$g(e) = g(e,\boldsymbol{\theta},\bar{\boldsymbol{\theta}},\mu) = \int_{\mathcal{X}} f[e-\eta(\bar{\boldsymbol{\theta}},\xi)+\eta(\boldsymbol{\theta},\xi)]\,\mu(d\xi)\,. \tag{1.5}$$

The fact that the entropy of a p.d.f. $f(\cdot)$,

$$\mathrm{ent}(f) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx\,,$$

is invariant by translation makes $J_e(\boldsymbol{\theta}) = \mathrm{ent}\{\hat{f}_{n,h}[\,\cdot\,|\mathbf{e}_1^n(\boldsymbol{\theta})]\}$ not suitable as an estimation criterion, see also Remark 4.1 below. For that reason, we shall consider instead the criterion

$$J_e^s(\boldsymbol{\theta}) = \mathrm{ent}\{\hat{f}_{n,h}[\,\cdot\,|\mathbf{e}_1^n(\boldsymbol{\theta}),-\mathbf{e}_1^n(\boldsymbol{\theta})]\}\,, \tag{1.6}$$

the minimisation of which forces, in some sense, the errors to be close to zero. Again, under reasonable assumptions, $\hat{f}_{n,h_n}[\,\cdot\,|\mathbf{e}_1^n(\boldsymbol{\theta}),-\mathbf{e}_1^n(\boldsymbol{\theta})]$ converges to

$$g^s(e) = g^s(e,\boldsymbol{\theta},\bar{\boldsymbol{\theta}},\mu) = \frac{1}{2}\int_{\mathcal{X}}\{f[e-\eta(\bar{\boldsymbol{\theta}},\xi)+\eta(\boldsymbol{\theta},\xi)]+f[e+\eta(\bar{\boldsymbol{\theta}},\xi)-\eta(\boldsymbol{\theta},\xi)]\}\,\mu(d\xi)\,. \tag{1.7}$$

We shall denote by $\hat{\boldsymbol{\theta}}_e$ the *minimum-entropy estimator*

$$\hat{\boldsymbol{\theta}}_e = \arg\min_{\boldsymbol{\theta}\in\Theta} J_e^s(\boldsymbol{\theta})\,. \tag{1.8}$$

Section 2 briefly describes the Least Squares (LS) cross validation method for choosing the bandwidth h. The finite sample behaviour of the estimator $\hat{\boldsymbol{\theta}}_e$ is considered in Section 3 through examples. Notice that when $n = p$ and $\mathbf{y} = (y_1, \dots, y_n)^\top$ belongs to the expectation surface $\mathcal{S}$, defined by $\mathcal{S} = \{\boldsymbol{\eta}(\boldsymbol{\theta}, \Xi) = [\eta(\boldsymbol{\theta}, \xi_1), \dots, \eta(\boldsymbol{\theta}, \xi_n)]^\top \ , \ \boldsymbol{\theta} \in \Theta\}$, then, for any $h > 0$ $J_e^s(\boldsymbol{\theta})$ is minimum when each error $e_i(\boldsymbol{\theta})$ is equal to zero, so that $\hat{\boldsymbol{\theta}}_e$ coincides with the Least Squares (LS) estimator. Section 4 gives some properties of $\text{ent}(g^s)$ and its derivatives w.r.t. $\boldsymbol{\theta}$. The asymptotic behaviour of $\hat{\boldsymbol{\theta}}_e^n$, which differs from $\hat{\boldsymbol{\theta}}_e$ by a truncation of the integral in the entropy criterion (1.6), is considered in Section 5, where $\hat{\boldsymbol{\theta}}_e^n$ is shown to be strongly consistent. Finally, Section 6 draws some conclusions.

2. Bandwith selection by LS cross-validation

Consider the integrated squared error loss function

$$I(h) = \int_{-\infty}^{\infty} [\hat{f}_{n,h}(x) - f(x)]^2 dx = \int_{-\infty}^{\infty} \hat{f}_{n,h}(x)^2 dx - 2\int_{-\infty}^{\infty} \hat{f}_{n,h}(x) f(x) dx + \int_{-\infty}^{\infty} f(x)^2 dx \,.$$

The last terms does not depend on h. The first one can easily be evaluated numerically, and may even have a close-form expression when $K(\cdot)$ has a simple form. This is the case for instance for the Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \,, \tag{2.1}$$

which gives

$$\int_{-\infty}^{\infty} \hat{f}_{n,h}(x)^2 dx = \frac{1}{2n^2 h\sqrt{\pi}} \sum_{i,j=1}^{n} \exp\left[-\frac{(X_i - X_j)^2}{4h^2}\right] .$$

Evaluating the second term by leave-one-out cross validation, that is

$$\int_{-\infty}^{\infty} \hat{f}_{n,h}(x) f(x) dx \simeq \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} K\left(\frac{X_i - X_j}{h}\right) ,$$

one gets the criterion

$$C(h) = \int_{-\infty}^{\infty} \hat{f}_{n,h}(x)^2 dx - \frac{2}{n-1} \sum_{i=1}^{n} \left[\hat{f}_{n,h}(X_i) - \frac{K(0)}{nh}\right] , \tag{2.2}$$

to be minimised w.r.t. h. One can refer to [11] for a comparison with other methods for bandwidth selection. Note that when h is selected by LS cross validation, each evaluation of $J_e^s(\boldsymbol{\theta})$ during the determination of $\hat{\boldsymbol{\theta}}_e$, see (1.8), requires the minimisation of a criterion of the form (2.2), so that $h = h(\boldsymbol{\theta})$ in (1.6).

3. Examples

3.1. **Example 1.** We take $\eta(\theta, \xi) = \exp(-\theta\xi)$, $\bar{\theta} = 1$, in the model (1.1), with 20 observations at $\xi = 0.5, 0.55, \dots, 1.45$, and use the Gaussian kernels (2.1).

We consider first the case where the errors ϵ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$, with $\sigma = 0.1$, and estimate h by LS cross-validation. Figure 1 gives the empirical distributions of $\hat{\theta}_e$ given by (1.8) and of the ML estimator $\hat{\theta}_{ML}$, (which coincides here with the LS estimator $\hat{\theta}_{LS}$) obtained for 1000 repetitions of the experiment, together with the density of the normal distribution $\mathcal{N}(\bar{\theta}, \mathbf{M}(\bar{\theta}, \mu)^{-1}/20)$. Note the good agreement between the empirical and asymptotic densities of $\hat{\theta}_{ML}$ and the larger dispersion for $\hat{\theta}_e$.

Figure 2 presents the densities of $\hat{\theta}_e$ and $\mathcal{N}(\bar{\theta}, \mathbf{M}(\bar{\theta}, \mu)^{-1}/20)$ when h is fixed to 0.1. The accuracy of the estimation by $\hat{\theta}_e$ is now similar to that of ML estimation. Figure 3 gives the empirical standard deviation of $\hat{\theta}_e$ as a function of h, for h between 0.01 and 0.2. Oversmoothing appears to have little influence on the accuracy of estimation by $\hat{\theta}_e$, whereas undersmoothing produces a larger dispersion of the estimates. The poor results obtained for $\hat{\theta}_e$ in Figure 1 are thus due to undersmoothing of the estimated density by LS cross-validation, a phenomenon noticed in [11].

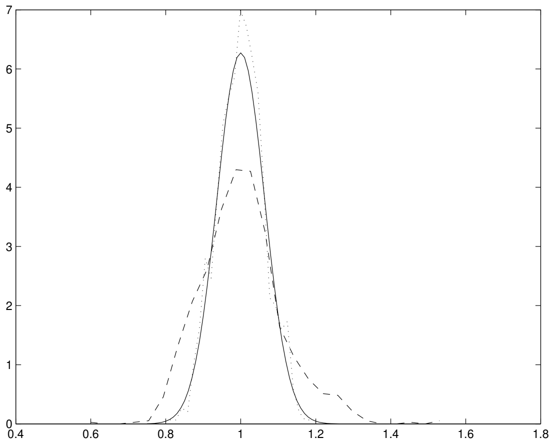

FIGURE 1. Empirical densities of $\hat{\theta}_e$ with h selected by LS cross validation (- - -) and $\hat{\theta}_{ML}$ $(\cdots)$ obtained for 1000 repetitions of the experiment (with normal errors), and asymptotic density (1.4) (—)

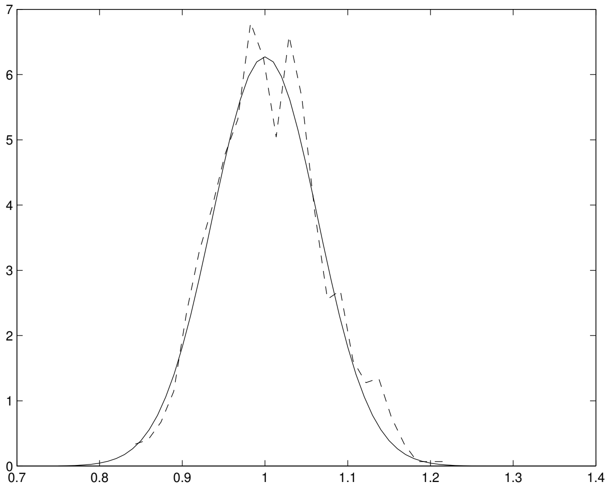

FIGURE 2. Empirical density of $\hat{\theta}_e$ with $h = 0.1$ (- - -) obtained for 1000 repetitions of the experiment (with normal errors), and asymptotic density (1.4) (—)

Consider now the case where the errors ϵ_i are independently uniformly distributed in $[-0.4, 0.4]$, with h fixed at 0.1 for the determination of $\hat{\theta}_e$. Figure 4 presents the empirical densities of $\hat{\theta}_e$ and $\hat{\theta}_{LS}$, the LS-estimator, which corresponds to the ML estimator under the (wrong) hypothesis that the errors are normally distributed. Estimation by entropy minimisation is clearly more accurate than by LS estimation.

When the errors ϵ_i are i.i.d. with the Laplace density $f(x) = [1/(\sigma\sqrt{2})]\exp(-\sqrt{2}|x|/\sigma)$, $\sigma = 0.15$, the asymptotic variance of $\hat{\theta}_{ML}$, given by (1.4), is $4.552\ 10^{-3}$. The empirical variance of $\hat{\theta}_e$ with $h = 0.075$ (respectively of $\hat{\theta}_{LS}$) based on 1000 repetitions of the experiment is approximately $7.8\ 10^{-3}$ (respectively $8.4\ 10^{-3}$). Although Gaussian kernels are not very appropriate here for estimating $f(\cdot)$, $\hat{\theta}_e$ gives a more accurate estimation than $\hat{\theta}_{LS}$. A consequence of using such inappropriate kernels is that simulations with larger values of h give now larger dispersions for $\hat{\theta}_e$, contrary to what happens in Figure 3.

Consider finally the same situation as for Figure 1 (normal errors, h selected by LS cross validation), but with the last observation, at $\xi_{20} = 1.45$, replaced by an outlier: $y_{20} = 5$. Figure 5 presents the true model response $\eta(\bar{\theta}, \xi)$ as a function of ξ (—), together with the responses obtained respectively for $\hat{\theta}_e$

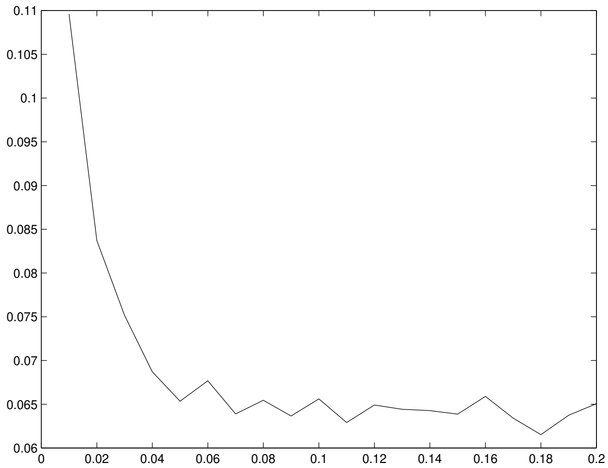

FIGURE 3. Empirical standard deviation of $\hat{\theta}_e$ (- - -), obtained for 1000 repetitions of the experiment (with normal errors), as a function of h

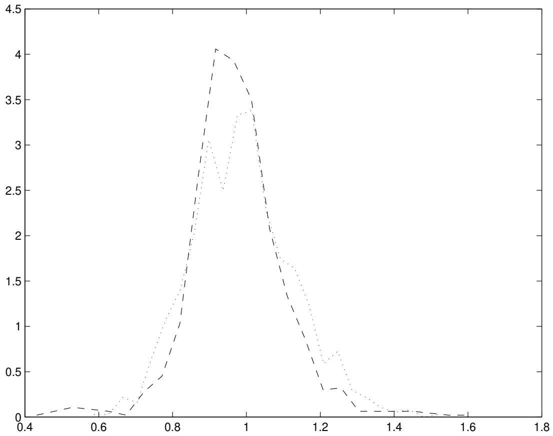

FIGURE 4. Empirical densities of $\hat{\theta}_e$ with $h = 0.1$ (- - -) and $\hat{\theta}_{LS}$ ($\cdots$) obtained for 1000 repetitions of the experiment (with uniform errors)

(- - -) and $\hat{\theta}_{ML}$ ($\cdots$). The fit is very good for $\hat{\theta}_e$, whereas $\hat{\theta}_{LS}$ is strongly influenced by the presence of the outlier. The reason for the robustness of $\hat{\theta}_e$ w.r.t. the presence of outlying points can be easily understood from Figure 6, which presents the estimated density $\hat{f}_{n,h}[\,\cdot\,|\mathbf{e}_1^n(\hat{\theta}_e), -\mathbf{e}_1^n(\hat{\theta}_e)]$. The two isolated peaks correspond to the errors $-e_{20}(\hat{\theta}_e)$ and $e_{20}(\hat{\theta}_e)$. The entropy of the estimated density is the same whatever the location of such peaks provided they are isolated. This property is similar to that of so-called redescending M-estimators, see [10]. However, when h is selected by a data-driven procedure, like LS cross validation for instance, the threshold used by $\hat{\boldsymbol{\theta}}_e$ for considering an observation as an outlier has the advantage of being data-dependent and not fixed *a priori*.

3.2. **Example 2.** Consider a parameter estimation problem, in a signal-processing context where interferences are present. A known binary discrete time signal $u(t) \in \{-1, 1\}$ passes through a Finite Impulse Response (FIR) filter A of length 5, which we formally write as $A(q) = \sum_{i=0}^{4} a_i q^{-i}$, with q^{-1} the delay operator. The observed output $y_t = y(t)$ is corrupted by a random i.i.d. sequence $e(t)$ and by interferences $z(t)$ given by the action of an unknown FIR filter $B(q)$ on an unknown binary signal

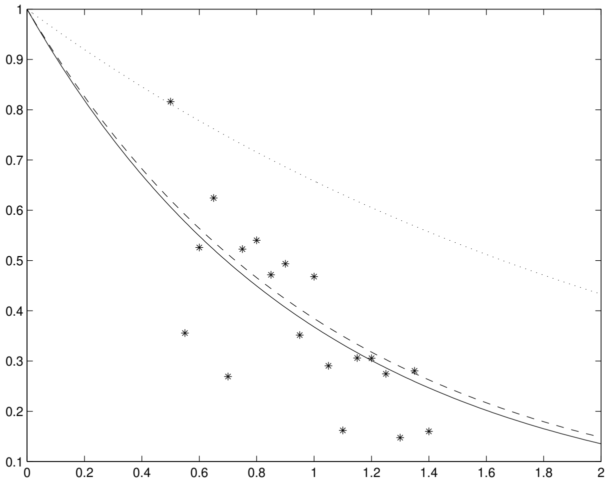

FIGURE 5. $\eta(\bar{\theta},\xi)$ (—), $\eta(\hat{\theta}_e,\xi)$ (- - -) and $\eta(\hat{\theta}_{ML},\xi)$ ($\cdots$) as functions of ξ; the observations $y_1 \ldots y_{19}$ are represented by stars, $y_{20} = 5$

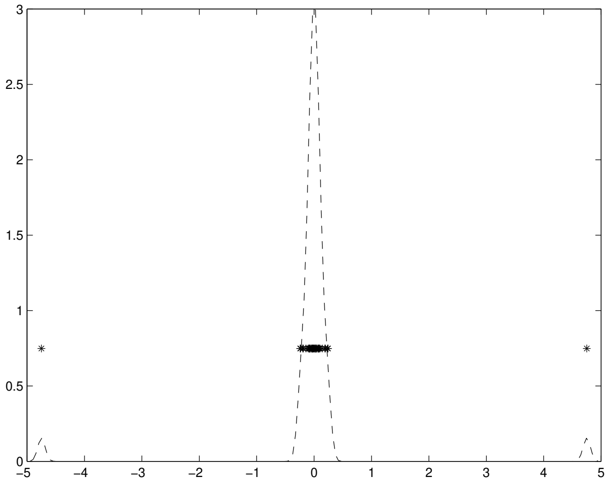

FIGURE 6. Estimated density $\hat{f}_{n,h}[\,\cdot\,|\mathbf{e}_1^n(\hat{\theta}_e), -\mathbf{e}_1^n(\hat{\theta}_e)]$ (- - -); the locations of the errors $\mathbf{e}_1^n(\hat{\theta}_e), -\mathbf{e}_1^n(\hat{\theta}_e)$ are indicated by stars

$v(t) \in \{-1,1\}$. The problem is to estimate the parameters $\boldsymbol{\theta} = (a_0,\ldots,a_4)$ from the inputs $u(t)$ and observations $y_t = A(q)u(t)+\epsilon(t)$, with $\epsilon(t) = e(t)+z(t)$. We take $e(t) \sim \mathcal{N}(0,\sigma^2)$, with $\sigma = 0.1$, and $B(q) = 1-0.5q^{-1}+0.2q^{-2}$. Figure 7 gives an histogram of the errors $\epsilon(t)$. Although they are not stochastically independent, we use the estimator $\hat{\boldsymbol{\theta}}_e$ given by (1.8). Figure 8 gives the empirical densities of $\hat{\boldsymbol{\theta}}_e$ (full line) and $\hat{\boldsymbol{\theta}}_{LS}$ (dashed line) when the true value of $\boldsymbol{\theta}$ is $\bar{\boldsymbol{\theta}} = (1,-0.5,0.2,-0.3,0.1)$, obtained from 500 input-output sequences of length 100. The empirical Mean-Squared Errors (MSE) for the components of $\hat{\boldsymbol{\theta}}_{LS}$ and $\hat{\boldsymbol{\theta}}_e$ are respectively $0.0127, 0.0135, 0.0138, 0.0131, 0.0139$ and $0.0027, 0.0020, 0.0031, 0.0029, 0.0030$. The MSE is thus reduced by a factor between 4 and 7 when $\hat{\boldsymbol{\theta}}_e$ is used instead of $\hat{\boldsymbol{\theta}}_{LS}$.

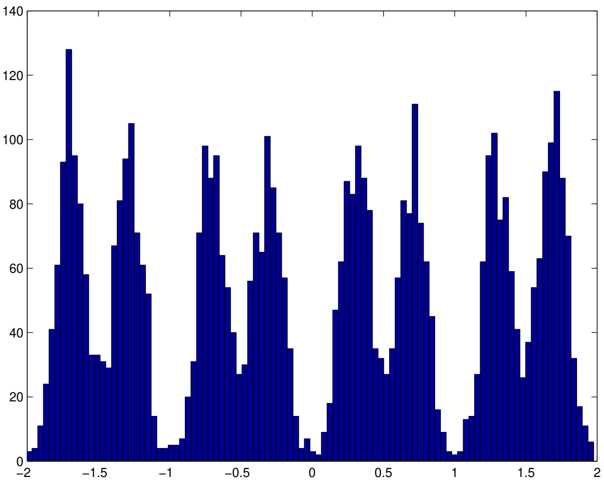

FIGURE 7. Histogram of the errors $\epsilon(t)$ in Example 2

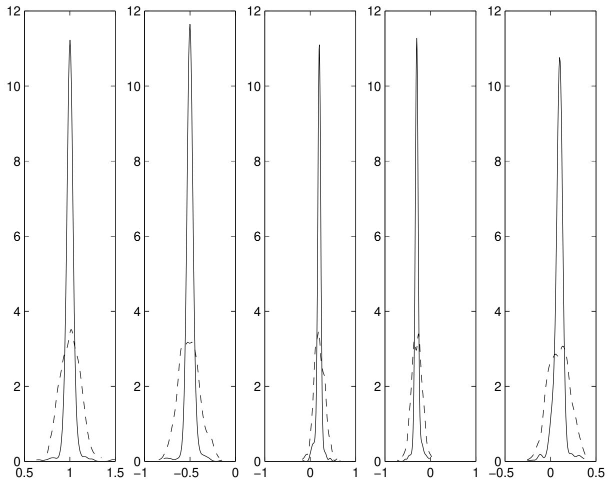

FIGURE 8. Empirical densities of $\hat{\theta}_e$ with $h = 0.1$ (—) and $\hat{\theta}_{LS}$ (- - -) obtained for 500 input-output sequences of length 100 in Example 2

4. Some properties of the entropy of g^s

The asymptotic normality of $\mathrm{ent}[\hat{f}_{n,h_n}(x|X_1,\ldots,X_n)]$ is proved in [6] for the double exponential kernel when $f(\cdot)$ has a finite moment of order larger than 2, $\mathcal{I}(f) < \infty$ and $\int_{-\infty}^{\infty}[f''(x)]^2/f(x)\,dx < \infty$; that is,

$$\sqrt{n}\{\mathrm{ent}[\hat{f}_{n,h_n}(x|X_1,\ldots,X_n)] - \mathrm{ent}(f)\} \xrightarrow{d} z \sim \mathcal{N}(0, \mathrm{Var}[\log f])\,.$$

Other consistency results for $\mathrm{ent}[\hat{f}_{n,h_n}(x|X_1,\ldots,X_n)]$ can be found, *e.g.*, in [12]. Under suitable assumptions, the criterion $J_e^s(\boldsymbol{\theta})$ (1.6) thus converges to $\mathrm{ent}(g^s)$, with $g^s(e)$ given by (1.7), hence the interest of studying the properties of $\mathrm{ent}(g^s)$.

First, we compute the derivatives of $g^s(e) = g^s(e, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}, \mu)$ w.r.t. $\boldsymbol{\theta}$. One has

$$\begin{aligned}
\frac{\partial g^s(e)}{\partial \boldsymbol{\theta}} &= \frac{1}{2} \int_{\mathcal{X}} \{f'[e - \eta(\bar{\boldsymbol{\theta}}, \xi) + \eta(\boldsymbol{\theta}, \xi)] - f'[e + \eta(\bar{\boldsymbol{\theta}}, \xi) - \eta(\boldsymbol{\theta}, \xi)]\} \frac{\partial \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta}} \, \mu(d\xi) \,, \\
\frac{\partial^2 g^s(e)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} &= \frac{1}{2} \int_{\mathcal{X}} \{f'[e - \eta(\bar{\boldsymbol{\theta}}, \xi) + \eta(\boldsymbol{\theta}, \xi)] - f'[e + \eta(\bar{\boldsymbol{\theta}}, \xi) - \eta(\boldsymbol{\theta}, \xi)]\} \frac{\partial^2 \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \, \mu(d\xi) \\
&\quad + \frac{1}{2} \int_{\mathcal{X}} \{f''[e - \eta(\bar{\boldsymbol{\theta}}, \xi) + \eta(\boldsymbol{\theta}, \xi)] + f''[e + \eta(\bar{\boldsymbol{\theta}}, \xi) - \eta(\boldsymbol{\theta}, \xi)]\} \frac{\partial \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta}} \frac{\partial \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta}^\top} \, \mu(d\xi) \,.
\end{aligned}$$

Therefore,

$$\frac{\partial g^s(e)}{\partial \boldsymbol{\theta}}_{|\bar{\boldsymbol{\theta}}} = \mathbf{0} \,, \quad \text{and} \quad \frac{\partial^2 g^s(e)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}_{|\bar{\boldsymbol{\theta}}} = f''(e) \int_{\mathcal{X}} \frac{\partial \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta}}_{|\bar{\boldsymbol{\theta}}} \frac{\partial \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta}^\top}_{|\bar{\boldsymbol{\theta}}} \mu(d\xi) \,.$$

Now,

$$\frac{\partial \mathrm{ent}(g^s)}{\partial \boldsymbol{\theta}} = - \int_{-\infty}^{\infty} [1 + \log g^s(e, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}, \mu)] \frac{\partial g^s(e, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}, \mu)}{\partial \boldsymbol{\theta}} \, de$$

so that $\partial \mathrm{ent}(g^s)/\partial \boldsymbol{\theta}_{|\bar{\boldsymbol{\theta}}} = \mathbf{0}$, and

$$\begin{aligned}
\frac{\partial^2 \mathrm{ent}(g^s)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} &= - \int_{-\infty}^{\infty} \frac{1}{g^s(e, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}, \mu)} \frac{\partial g^s(e, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}, \mu)}{\partial \boldsymbol{\theta}} \frac{\partial g^s(e, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}, \mu)}{\partial \boldsymbol{\theta}^\top} \, de \\
&\quad - \int_{-\infty}^{\infty} [1 + \log g^s(e, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}, \mu)] \frac{\partial^2 g^s(e, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}, \mu)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \, de \,.
\end{aligned}$$

Noticing that $(f \log f)'' = (1 + \log f) f'' + (f')^2/f$, one gets $\partial^2 \mathrm{ent}(g^s)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top_{|\bar{\boldsymbol{\theta}}} = \mathbf{M}(\bar{\boldsymbol{\theta}}, \mu)$, see (1.2). The criterion $\mathrm{ent}(g^s)$ is thus locally concave at $\bar{\boldsymbol{\theta}}$, with a stationary solution (zero derivative) at $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$, and $g^s(\cdot, \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}, \mu) = f(\cdot)$, which is consistent with the property that convolution increases entropy, see [5].

Remark 4.1. Consider the criterion $\mathrm{ent}(g)$, with $g(e)$ given by (1.5), and assume that some parameter, say θ_1, appears in an additive term, independent of ξ, in $\eta(\boldsymbol{\theta}, \xi)$; that is, $\eta(\boldsymbol{\theta}, \xi) = \eta_1(\theta_1) + \eta_2(\theta_2, \ldots, \theta_p, \xi)$. Then, $\partial g(e)/\partial \theta_1 = \partial \eta_1(\theta_1)/\partial \theta_1 \, dg(e)/de$, which gives $\partial \mathrm{ent}(g)/\partial \theta_1 = 0$. Therefore, θ_1 cannot be estimated by minimising $\mathrm{ent}(g)$. As mentioned in the introduction, this is due to the invariance of the entropy of a p.d.f. by translation. Calculations similar to those above give

$$\frac{\partial^2 \mathrm{ent}(g)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}_{|\bar{\boldsymbol{\theta}}} = \mathcal{I}(f) \left(\int_{\mathcal{X}} \frac{\partial \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta}}_{|\bar{\boldsymbol{\theta}}} \frac{\partial \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta}^\top}_{|\bar{\boldsymbol{\theta}}} \mu(d\xi) - \int_{\mathcal{X}} \frac{\partial \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta}}_{|\bar{\boldsymbol{\theta}}} \mu(d\xi) \int_{\mathcal{X}} \frac{\partial \eta(\boldsymbol{\theta}, \xi)}{\partial \boldsymbol{\theta}^\top}_{|\bar{\boldsymbol{\theta}}} \mu(d\xi) \right) \,,$$

so that $\partial^2 \mathrm{ent}(g^s)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top_{|\bar{\boldsymbol{\theta}}} - \partial^2 \mathrm{ent}(g)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top_{|\bar{\boldsymbol{\theta}}}$ is nonnegative-definite, which gives another reason for preferring $\mathrm{ent}(g^s)$ to $\mathrm{ent}(g)$.

5. Asymptotic behaviour of a truncated estimator

We shall denote $\hat{f}_{n,h}(x) = \hat{f}_{n,h}[x|\mathbf{e}_1^n(\boldsymbol{\theta}), -\mathbf{e}_1^n(\boldsymbol{\theta})]$. Although $J_e^s(\boldsymbol{\theta})$ depends on n, see (1.6), and $\mathrm{ent}(g^s)$ depends on $\boldsymbol{\theta}$, see (1.7), these dependencies will be omitted in the notations. Besides the assumptions of Section 1, we shall assume that the first three derivatives of $f(\cdot)$ are bounded, that $\int_{-\infty}^{\infty} |y| K(y) dy < \infty$ and that the s-th derivative $K^{(s)}(\cdot)$ of $K(\cdot)$ is a continuous function of bounded variation for $s = 1, 2$, denoted by v_s.

We use an approach similar to [4], and consider the criterion

$$J_e^s(\boldsymbol{\theta}, A_n, h_n) = \mathrm{ent}_{A_n}(\hat{f}_{n,h_n}) \,, \tag{5.1}$$

where

$$\mathrm{ent}_A(f) = - \int_{-A}^{A} f(x) \log f(x) dx \,.$$

We denote the associated estimator by $\hat{\boldsymbol{\theta}}_e^n$, $\hat{\boldsymbol{\theta}}_e^n = \arg\min_{\boldsymbol{\theta} \in \Theta} J_e^s(\boldsymbol{\theta}, A_n, h_n)$.

5.1. **Almost sure behaviour of the derivatives of $J_e^s(\boldsymbol{\theta}, A_n, h_n)$.** We shall make use of the following extension of Schuster Lemma [15].

Lemma 5.1. *Consider*

$$\hat{g}^{(r)}_{n,h_n}(x) = \frac{1}{nh_n^{r+1}} \sum_{i=1}^{n} K^{(r)} \left[\frac{x - X_i - r(Z_i)}{h_n}\right] R(Z_i)\,,$$

where $\{X_i\}$ is an i.i.d. sequence, with p.d.f. $f(\cdot)$ such that its first $r+1$ derivatives are bounded; $\{Z_i\}$ is an i.i.d. sequence, with a discrete probability measure μ having m support points Z^j receiving weights w_j, $j = 1, \ldots, m$; $r(\cdot)$ and $R(\cdot)$ are two Borel functions. Assume that $\int_{-\infty}^{\infty} |y| K(y) dy < \infty$ and that the s-th derivative $K^{(s)}(\cdot)$ of $K(\cdot)$ is a continuous function of bounded variation v_s, for $s = 1, \ldots, r$. Then, for $\{\gamma_n\}$ a sequence of positive numbers such that $h_n = o(\gamma_n)$, there exist positive constants C_1 and C_2 such that, for n large enough,

$$\mathrm{Prob}\{\sup_x |\hat{g}^{(r)}_{n,h_n}(x) - g^{(r)}(x)| > \gamma_n\} \leq C_1 \exp(-C_2 n \gamma_n^2 h_n^{2r+2})\,,$$

where $g^{(r)}(x) = \int f^{(r)}[x - r(z)] R(z)\, \mu(dz) = \sum_{j=1}^{m} w_j f^{(r)}[x - r(Z^j)] R(Z^j)$.

Proof. One can write $\hat{g}^{(r)}_{n,h_n}(x) = \sum_{j=1}^{m} (n_j/n) \hat{f}^{(r)}_{n_j,h_n}[x - r(Z^j)] R(Z^j)$, with

$$\hat{f}^{(r)}_{n_j,h_n}(x) = \frac{1}{n_j h_n^{r+1}} \sum_{k=1}^{n_j} K^{(r)} \left[\frac{x - X_{j_k}}{h_n}\right]$$

and $\sum_{j=1}^{m} n_j = n$. Therefore,

$$\begin{aligned}
|\hat{g}^{(r)}_{n,h_n}(x) - g^{(r)}(x)| \quad \leq \quad & \sum_{j=1}^{m} \frac{n_j}{n} |R(Z^j)|\, |\hat{f}^{(r)}_{n_j,h_n}[x - r(Z^j)] - E_X \hat{f}^{(r)}_{n_j,h_n}[x - r(Z^j)]| \\
& + \sum_{j=1}^{m} \frac{n_j}{n} |R(Z^j)|\, |E_X \hat{f}^{(r)}_{n_j,h_n}[x - r(Z^j)] - f^{(r)}[x - r(Z^j)]| \\
& + \sum_{j=1}^{m} |R(Z^j)|\, f^{(r)}[x - r(Z^j)] \left|\frac{n_j}{n} - w_j\right|\,,
\end{aligned}$$

where E_X denotes the expectation w.r.t. the X_i's. Lemma 2.3 of [15] gives

$$\sup_x |E_X \hat{f}^{(r)}_{n_j,h_n}(x) - f^{(r)}(x)| \leq C_0 h_n$$

for a constant C_0 not depending on h_n. Lemma 2.2 in the same paper gives

$$\sup_x |\hat{f}^{(r)}_{n_j,h_n}(x) - E_X \hat{f}^{(r)}_{n_j,h_n}(x)| \leq \frac{v_r}{h_n^{r+1}} \sup_x |F_{n_j}(x) - F(x)|\,,$$

with F_{n_j} the empirical distribution function based on $X_{j_1}, \ldots, X_{j_{n_j}}$. Since $h_n = o(\gamma_n)$, one has for n large enough

$$\begin{aligned}
\mathrm{Prob}\{\sup_x |\hat{g}^{(r)}_{n,h_n}(x) - g^{(r)}(x)| > \gamma_n\} \quad \leq \quad & \mathrm{Prob}\left\{R_* \frac{v_r}{h_n^{r+1}} \sum_{j=1}^{m} \frac{n_j}{n} \sup_x |F_{n_j}(x) - F(x)| > \frac{\gamma_n}{3}\right\} \\
& + \mathrm{Prob}\left\{R_* B \sum_{j=1}^{m} \left|\frac{n_j}{n} - w_j\right| > \frac{\gamma_n}{3}\right\}\,,
\end{aligned}$$

with B the bound on $f^{(r)}(\cdot)$ and $R_* = \max_{j=1,\ldots,m} |R(Z^j)|$. Denote the probabilities of the right-hand side by P_1 and P_2. One has $P_2 \leq \mathrm{Prob}\{m R_* B \max_{j=1,\ldots,m} |n_j/n - w_j| > \gamma_n/3\}$, and thus, from Lemma 2.1 of [15],

$$P_2 \leq C \exp[-2n\gamma_n^2/(9m^2 B^2 R_*^2)]\,,$$

with C a universal constant. P_1 can be written as $P_1 = \mathrm{Prob}\{\alpha_n \sum_{j=1}^m (n_j/n) \sup_x |F_{n_j}(x) - F(x)| > \gamma_n\}$, with $\alpha_n = 3R_* v_r / h_n^{r+1}$, and thus $P_1 \le \sum_{j=1}^m P_{1,j}$, with $P_{1,j} = \mathrm{Prob}\{\alpha_n \sup_x |F_{n_j}(x) - F(x)| > \gamma_n\}$. For n_j given, Lemma 2.1 of [15] gives $\mathrm{Prob}\{\alpha_n \sup_x |F_{n_j}(x) - F(x)| > \gamma_n\} \le C\exp[-2n_j(\gamma_n/\alpha_n)^2]$, and therefore, from the binomial distribution of n_j, $P_{1,j} \le C[1-(1-\beta_n)w_j]^n$, with $\beta_n = \exp[-2(\gamma_n/\alpha_n)^2]$. Letting h_n tend to zero, one gets $P_{1,j} \le C\exp[-n\gamma_n^2 h_n^{2r+2} w_j/(9R_*^2 v_r^2)]$ and thus, for n large enough,

$$P_1 \le mC\exp\left[-\frac{n\gamma_n^2 h_n^{2r+2} w_*}{9R_*^2 v_r^2}\right]$$

with $w_* = \min_{j=1,\ldots,m} w_j$. Together with the inequality obtained for P_2, this completes the proof. $\square$

The lemma above yields the following property.

Corollary 5.2. *For $\{\gamma_n\}$ a sequence of positive numbers such that $h_n = o(\gamma_n)$, there exist C_1 and C_2 positive such that, $\forall \boldsymbol{\theta} \in \Theta$, $\forall u, v = 1, \ldots, p$,*

$$\begin{aligned}
\mathrm{Prob}\{\sup_x |\hat{f}_{n,h_n}(x) - g^s(x)| > \gamma_n\} &\le C_1 \exp(-C_2 n\gamma_n^2 h_n^2)\,, \\
\mathrm{Prob}\left\{\sup_x \left|\frac{\partial \hat{f}_{n,h_n}(x)}{\partial\theta_u} - \frac{\partial g^s(x)}{\partial\theta_u}\right| > \gamma_n\right\} &\le C_1 \exp(-C_2 n\gamma_n^2 h_n^4)\,, \\
\mathrm{Prob}\left\{\sup_x \left|\frac{\partial^2 \hat{f}_{n,h_n}(x)}{\partial\theta_u\partial\theta_v} - \frac{\partial^2 g^s(x)}{\partial\theta_u\partial\theta_v}\right| > \gamma_n\right\} &\le C_1 \exp(-C_2 n\gamma_n^2 h_n^6)\,.
\end{aligned}$$

Proof. For $\boldsymbol{\theta}$ given, this is a direct consequence of Lemma 5.1. Indeed, we have

$$\begin{aligned}
\hat{f}_{n,h_n}(x) &= \frac{1}{2nh_n}\sum_{i=1}^n \left\{K\left[\frac{x-y_i+\eta(\boldsymbol{\theta},\xi_i)}{h_n}\right] + K\left[\frac{x+y_i-\eta(\boldsymbol{\theta},\xi_i)}{h_n}\right]\right\}\,, \\
\frac{\partial \hat{f}_{n,h_n}(x)}{\partial\theta_u} &= \frac{1}{2nh_n^2}\sum_{i=1}^n \left\{K^{(1)}\left[\frac{x-y_i+\eta(\boldsymbol{\theta},\xi_i)}{h_n}\right] - K^{(1)}\left[\frac{x+y_i-\eta(\boldsymbol{\theta},\xi_i)}{h_n}\right]\right\}\frac{\partial\eta(\boldsymbol{\theta},\xi_i)}{\partial\theta_u}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 \hat{f}_{n,h_n}(x)}{\partial\theta_u\partial\theta_v} &= \frac{1}{2nh_n^3}\sum_{i=1}^n \left\{K^{(2)}\left[\frac{x-y_i+\eta(\boldsymbol{\theta},\xi_i)}{h_n}\right] + K^{(2)}\left[\frac{x+y_i-\eta(\boldsymbol{\theta},\xi_i)}{h_n}\right]\right\}\frac{\partial\eta(\boldsymbol{\theta},\xi_i)}{\partial\theta_u}\frac{\partial\eta(\boldsymbol{\theta},\xi_i)}{\partial\theta_v} \\
&\quad + \frac{1}{2nh_n^2}\sum_{i=1}^n \left\{K^{(1)}\left[\frac{x-y_i+\eta(\boldsymbol{\theta},\xi_i)}{h_n}\right] - K^{(1)}\left[\frac{x+y_i-\eta(\boldsymbol{\theta},\xi_i)}{h_n}\right]\right\}\frac{\partial^2\eta(\boldsymbol{\theta},\xi_i)}{\partial\theta_u\partial\theta_v}\,.
\end{aligned}$$

The bounds obtained can be made independent of $\boldsymbol{\theta}$ by using the compactness of Θ and the continuity of $\eta(\boldsymbol{\theta},\xi)$ and its first two derivatives with respect to $\boldsymbol{\theta}$. $\square$

Let $H(\cdot)$ denote a monotonic strictly increasing function such that $\forall U$, $\sup_{|z|<U} 1/g^s(z) \le H(U)$, see [4], and $H^{-1}(\cdot)$ be the inverse function of $H(\cdot)$. We can prove the following result on the almost sure behaviour of the first two derivatives of $J_e^s(\boldsymbol{\theta}, A_n, h_n)$.

Lemma 5.3. *For any $\boldsymbol{\theta} \in \Theta$, $\partial J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta} \xrightarrow{a.s.} \partial\mathrm{ent}(g^s)/\partial\boldsymbol{\theta}$ and $\partial^2 J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top \xrightarrow{a.s.} \partial^2\mathrm{ent}(g^s)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ as $n \to \infty$ when $h_n = n^{-1/7}$, $A_n = H^{-1}(n^{1/48})$.*

Proof. We start with $\partial J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta}$. Direct calculations give

$$\frac{\partial J_e^s(\boldsymbol{\theta}, A_n, h_n)}{\partial\theta_l} = -\int_{-A_n}^{A_n} \frac{\partial \hat{f}_{n,h_n}(x)}{\partial\theta_l}[1 + \log \hat{f}_{n,h_n}(x)]\,dx\,, \tag{5.2}$$

and therefore

$$\begin{aligned}\left|\frac{\partial J_e^s(\boldsymbol{\theta},A_n,h_n)}{\partial\theta_l}-\frac{\partial \mathrm{ent}(g^s)}{\partial\theta_l}\right| &< 2\int_{A_n}^{\infty}\left|\frac{\partial g^s(x)}{\partial\theta_l}[1+\log g^s(x)]\right|dx\\ &+\int_{-A_n}^{A_n}\left|\frac{\partial g^s(x)}{\partial\theta_l}[1+\log g^s(x)]-\frac{\partial \hat f_{n,h_n}(x)}{\partial\theta_l}[1+\log \hat f_{n,h_n}(x)]\right|dx\,.\end{aligned}$$

The first term on the right-hand side tends to zero when A_n tends to infinity. Denote the second one by B_n. One has

$$\begin{aligned}B_n &< \int_{-A_n}^{A_n}\left|\frac{\partial g^s(x)}{\partial\theta_l}-\frac{\partial \hat f_{n,h_n}(x)}{\partial\theta_l}\right|dx+\int_{-A_n}^{A_n}\left|\frac{\partial g^s(x)}{\partial\theta_l}\log\frac{g^s(x)}{\hat f_{n,h_n}(x)}\right|dx\\ &+\int_{-A_n}^{A_n}\left|\left(\frac{\partial g^s(x)}{\partial\theta_l}-\frac{\partial \hat f_{n,h_n}(x)}{\partial\theta_l}\right)\log \hat f_{n,h_n}(x)\right|dx\,,\end{aligned}$$

and, since $|\log u|\le|1/u-1|+|u-1|$ and $|\log u|<u+1/u$,

$$\begin{aligned}B_n &< \int_{-A_n}^{A_n}\left|\frac{\partial g^s(x)}{\partial\theta_l}-\frac{\partial \hat f_{n,h_n}(x)}{\partial\theta_l}\right|dx+\int_{-A_n}^{A_n}\left|\frac{\partial g^s(x)}{\partial\theta_l}\right|\left[\left|\frac{g^s(x)-\hat f_{n,h_n}(x)}{\hat f_{n,h_n}(x)}\right|+\left|\frac{g^s(x)-\hat f_{n,h_n}(x)}{g^s(x)}\right|\right]dx\\ &+\int_{-A_n}^{A_n}\left|\frac{\partial g^s(x)}{\partial\theta_l}-\frac{\partial \hat f_{n,h_n}(x)}{\partial\theta_l}\right|\left(\hat f_{n,h_n}(x)+\frac{g^s(x)}{g^s(x)\hat f_{n,h_n}(x)}\right)dx\,.\end{aligned}$$

Since $f'(\cdot)$ is bounded, the same is true for $\partial g^s(x)/\partial\theta_l$, and $|\partial g^s(x)/\partial\theta_l|<G_1$ for some constant $G_1>0$. Consider the two events

$$\begin{aligned}E_1 &: \sup_z|\hat f_{n,h_n}(z)-g^s(z)|<\gamma_n\,,\\ E_2 &: \sup_z|\partial \hat f_{n,h_n}(z)/\partial\theta_l-\partial g^s(z)/\partial\theta_l|<\gamma_n\,.\end{aligned}\tag{5.3}$$

When they are realized, and provided that $H(A_n)\gamma_n\to 0$ as $n\to\infty$, we have for n large enough

$$B_n<\gamma_n H(A_n)+3G_1\gamma_n H^2(A_n)+\gamma_n[1+2H^2(A_n)]\,.$$

Take $\gamma>0$ and $\gamma_n=\gamma/[2(2+3G_1)H^2(A_n)]$. Choose $h_n=n^{-\alpha}$, $A_n=H^{-1}(n^\beta)$, with α, β, such that $\alpha>0$, $\beta>0$, $\alpha>2\beta$, $1-4\alpha-4\beta=\delta>0$ (for instance, $\alpha=1/7$, $\beta=1/48$ give $\delta=29/84$). From Corollary 5.2, it gives for n large enough

$$\begin{aligned}\mathrm{Prob}\{|\partial J_e^s(\boldsymbol{\theta},A_n,h_n)/\partial\theta_l-\partial\mathrm{ent}(g^s)/\partial\theta_l|>\gamma\} &\le \mathrm{Prob}\{\overline{E_1}\}+\mathrm{Prob}\{\overline{E_2}\}\\ &\le C_1\exp(-C_2 n\gamma_n^2h_n^2)+C_1\exp(-C_2n\gamma_n^2h_n^4)\\ &< 2C_1\exp(-C_2n\gamma_n^2h_n^4)\,,\end{aligned}$$

since $h_n\to 0$ as $n\to\infty$. Therefore, for any $\gamma>0$, for any $\boldsymbol{\theta}\in\Theta$ and any $l=1,\ldots,p$,

$$\mathrm{Prob}\{|\partial J_e^s(\boldsymbol{\theta},A_n,h_n)/\partial\theta_l-\partial\mathrm{ent}(g^s)/\partial\theta_l|>\gamma\}<2C_1\exp[-C_2\gamma^2/[2(2+3G_1)]^2n^\delta]\,.$$

This implies $\sum_{n=1}^\infty\mathrm{Prob}\{|\partial J_e^s(\boldsymbol{\theta},A_n,h_n)/\partial\theta_l-\partial\mathrm{ent}(g^s)/\partial\theta_l|>\gamma\}<\infty$, and by Borel-Cantelli arguments $\partial J_e^s(\boldsymbol{\theta},A_n,h_n)/\partial\boldsymbol{\theta}\xrightarrow{a.s.}\partial\mathrm{ent}(g^s)/\partial\boldsymbol{\theta}$ as $n\to\infty$.

The case of $\partial^2J_e^s(\boldsymbol{\theta},A_n,h_n)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ can be treated similarly. One gets

$$\frac{\partial^2J_e^s(\boldsymbol{\theta},A_n,h_n)}{\partial\theta_l\partial\theta_m}=-\int_{-A_n}^{A_n}\frac{\partial^2\hat f_{n,h_n}(x)}{\partial\theta_l\partial\theta_m}[1+\log\hat f_{n,h_n}(x)]\,dx-\int_{-A_n}^{A_n}\frac{\partial\hat f_{n,h_n}(x)}{\partial\theta_l}\frac{\partial\hat f_{n,h_n}(x)}{\partial\theta_m}\frac{1}{\hat f_{n,h_n}(x)}\,dx\,,$$

and therefore

$$\begin{aligned}\left|\frac{\partial^2 J_e^s(\boldsymbol{\theta},A_n,h_n)}{\partial\theta_l\partial\theta_m}-\frac{\partial^2\mathrm{ent}(g^s)}{\partial\theta_l\partial\theta_m}\right| &< 2\int_{A_n}^{\infty}\left|\frac{\partial^2 g^s(x)}{\partial\theta_l\partial\theta_m}[1+\log g^s(x)]\right|dx+2\int_{A_n}^{\infty}\left|\frac{\partial g^s(x)}{\partial\theta_l}\right|\left|\frac{\partial g^s(x)}{\partial\theta_m}\right|\frac{1}{g^s(x)}\,dx\\ &+\int_{-A_n}^{A_n}\left|\frac{\partial\hat f_{n,h_n}(x)}{\partial\theta_l}\frac{\partial\hat f_{n,h_n}(x)}{\partial\theta_m}\frac{1}{\hat f_{n,h_n}(x)}-\frac{\partial g^s(x)}{\partial\theta_l}\frac{\partial g^s(x)}{\partial\theta_m}\frac{1}{g^s(x)}\right|dx\\ &+\int_{-A_n}^{A_n}\left|\frac{\partial^2 g^s(x)}{\partial\theta_l\partial\theta_m}[1+\log g^s(x)]-\frac{\partial^2\hat f_{n,h_n}(x)}{\partial\theta_l\partial\theta_m}[1+\log\hat f_{n,h_n}(x)]\right|dx\,.\end{aligned}$$

The first two terms on the right-hand side tend to zero when A_n tends to infinity. Denote the third (resp. fourth) one by C_n (resp. D_n). One has

$$\begin{aligned}C_n &< \int_{-A_n}^{A_n}\left|\frac{\partial\hat f_{n,h_n}(x)}{\partial\theta_l}\left[\frac{\partial\hat f_{n,h_n}(x)}{\partial\theta_m}-\frac{\partial g^s(x)}{\partial\theta_m}\right]\right|\frac{g^s(x)}{g^s(x)\hat f_{n,h_n}(x)}\,dx\\ &+\int_{-A_n}^{A_n}\left|\frac{\partial g^s(x)}{\partial\theta_m}\left[\frac{\partial\hat f_{n,h_n}(x)}{\partial\theta_l}-\frac{\partial g^s(x)}{\partial\theta_l}\right]\right|\frac{g^s(x)}{g^s(x)\hat f_{n,h_n}(x)}\,dx\\ &+\int_{-A_n}^{A_n}\left|\frac{\partial g^s(x)}{\partial\theta_l}\frac{\partial g^s(x)}{\partial\theta_m}\frac{[g^s(x)-\hat f_{n,h_n}(x)]}{g^s(x)\hat f_{n,h_n}(x)}\right|dx\,,\end{aligned}\tag{5.4}$$

and

$$\begin{aligned}D_n &< \int_{-A_n}^{A_n}\left|\frac{\partial^2 g^s(x)}{\partial\theta_l\partial\theta_m}-\frac{\partial^2\hat f_{n,h_n}(x)}{\partial\theta_l\partial\theta_m}\right|dx+\int_{-A_n}^{A_n}\left|\frac{\partial^2 g^s(x)}{\partial\theta_l\partial\theta_m}\log\frac{g^s(x)}{\hat f_{n,h_n}(x)}\right|dx\\ &+\int_{-A_n}^{A_n}\left|\left(\frac{\partial^2 g^s(x)}{\partial\theta_l\partial\theta_m}-\frac{\partial^2\hat f_{n,h_n}(x)}{\partial\theta_l\partial\theta_m}\right)\log\hat f_{n,h_n}(x)\right|dx\,,\end{aligned}$$

and, since $|\log u|\le|1/u-1|+|u-1|$ and $|\log u|<u+1/u$,

$$\begin{aligned}D_n &< \int_{-A_n}^{A_n}\left|\frac{\partial^2 g^s(x)}{\partial\theta_l\partial\theta_m}-\frac{\partial^2\hat f_{n,h_n}(x)}{\partial\theta_l\partial\theta_m}\right|dx+\int_{-A_n}^{A_n}\left|\frac{\partial^2 g^s(x)}{\partial\theta_l\partial\theta_m}\right|\left[\left|\frac{g^s(x)-\hat f_{n,h_n}(x)}{\hat f_{n,h_n}(x)}\right|+\left|\frac{g^s(x)-\hat f_{n,h_n}(x)}{g^s(x)}\right|\right]dx\\ &+\int_{-A_n}^{A_n}\left|\frac{\partial^2 g^s(x)}{\partial\theta_l\partial\theta_m}-\frac{\partial^2\hat f_{n,h_n}(x)}{\partial\theta_l\partial\theta_m}\right|\left(\hat f_{n,h_n}(x)+\frac{g^s(x)}{g^s(x)\hat f_{n,h_n}(x)}\right)dx\,.\end{aligned}$$

Since $f'(\cdot)$ and $f''(\cdot)$ are bounded, $|\partial g^s(x)/\partial\theta_{l,m}|<G_1$ and $|\partial^2 g^s(x)/\partial\theta_l\partial\theta_m|<G_2$ for some positive constants G_1, G_2. Consider the event E_1 given by (5.3) together with

$$\begin{aligned}E_2 &: \sup_z\max_{i=l,m}|\partial\hat f_{n,h_n}(z)/\partial\theta_i-\partial g^s(z)/\partial\theta_i|<\gamma_n\,,\\ E_3 &: \sup_z|\partial^2\hat f_{n,h_n}(z)/\partial\theta_l\partial\theta_m-\partial^2 g^s(z)/\partial\theta_l\partial\theta_m|<\gamma_n\,.\end{aligned}$$

When they are realized, and provided that $H(A_n)\gamma_n\to 0$ as $n\to\infty$, we have for n large enough

$$C_n<4G_1\gamma_nH^2(A_n)+2G_1^2\gamma_nH^3(A_n)\,,\quad D_n<\gamma_nH(A_n)+3G_2\gamma_nH^2(A_n)+\gamma_n[1+2H^2(A_n)]\,.$$

Take again $\gamma>0$ and $\gamma_n=\gamma/[4G_1^2H^3(A_n)]$. Choose $h_n=n^{-\alpha}$, $A_n=H^{-1}(n^\beta)$, with α, β, such that $\alpha>0$, $\beta>0$, $\alpha>3\beta$, $1-6\alpha-6\beta=\delta>0$ (for instance, $\alpha=1/7$, $\beta=1/48$ give $\delta=1/56$). From Corollary 5.2, it gives for n large enough

$$\begin{aligned}\mathrm{Prob}\{|\partial^2 J_e^s(\boldsymbol{\theta},A_n,h_n)/\partial\theta_l\partial\theta_m &- \partial^2\mathrm{ent}(g^s)/\partial\theta_l\partial\theta_m|>\gamma\}\le\mathrm{Prob}\{\overline{E_1}\}+\mathrm{Prob}\{\overline{E_2}\}+\mathrm{Prob}\{\overline{E_3}\}\\ &\le C_1\exp(-C_2n\gamma_n^2h_n^2)+C_1\exp(-C_2n\gamma_n^2h_n^4)+C_1\exp(-C_2n\gamma_n^2h_n^6)\\ &< 2C_1\exp(-C_2n\gamma_n^2h_n^6)\,,\end{aligned}$$

since $h_n \to 0$ as $n \to \infty$. Therefore, for any $\gamma > 0$ and any $l, m = 1, \ldots, p$,

$$\mathrm{Prob}\{|\partial^2 J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\theta_l\partial\theta_m - \partial^2 \mathrm{ent}(g^s)/\partial\theta_l\partial\theta_m| > \gamma\} < 2C_1 \exp[-C_2\gamma^2/(16G_1^4)n^\delta]$$

and $\partial^2 J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top \xrightarrow{a.s.} \partial^2 \mathrm{ent}(g^s)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ as $n \to \infty$ by Borel-Cantelli arguments. □

5.2. **Strong consistency of $\hat{\boldsymbol{\theta}}_e^n$.** One can easily show that Lemma 5.3 implies strong consistency of $\hat{\boldsymbol{\theta}}_e^n$ for a suitable choice of h_n and A_n.

Theorem 5.4. *When $h_n = n^{-1/7}$, $A_n = H^{-1}(n^{1/48})$, there exists a sequence $\{\hat{\boldsymbol{\theta}}_e^n\}$ of stationary solutions for $J_e^s(\boldsymbol{\theta}, A_n, h_n)$ (that is, satisfying $\partial J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta}_{|\hat{\boldsymbol{\theta}}_e^n} = \mathbf{0}$), such that $\hat{\boldsymbol{\theta}}_e^n \xrightarrow{a.s.} \bar{\boldsymbol{\theta}}$ as $n \to \infty$. Moreover, $\hat{\boldsymbol{\theta}}_e^n$ corresponds to a (local) minimum of $J_e^s(\boldsymbol{\theta}, A_n, h_n)$ for n larger than some n_0.*

Proof. Let $\mathcal{S}(\mathbf{c}, \rho)$ and $\mathcal{B}(\mathbf{c}, \rho)$ respectively denote the sphere and the ball of centre $\mathbf{c}$ and radius ρ. We showed in Section 4 that $\mathrm{ent}(g^s)$ is locally convex at $\bar{\boldsymbol{\theta}}$. Take ρ small enough for $\mathrm{ent}(g^s)$ to be convex in $\mathcal{B}(\bar{\boldsymbol{\theta}}, \rho)$. It implies that $\exists\alpha > 0$ such that $\forall\boldsymbol{\theta} \in \mathcal{S}(\bar{\boldsymbol{\theta}}, \rho)$, $(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^\top \partial\mathrm{ent}(g^s)/\partial\boldsymbol{\theta} > \alpha\rho$. Lemma 5.3 implies $\forall\beta > 0$, $\exists N(\boldsymbol{\theta})$ such that $\mathrm{Prob}\{\forall n > N(\boldsymbol{\theta}), \|\partial J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta} - \partial\mathrm{ent}(g^s)/\partial\boldsymbol{\theta}\| < \alpha/2\} > 1 - \beta$, and therefore, $\mathrm{Prob}\{\forall n > N(\boldsymbol{\theta}), (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^\top \partial J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta} > \alpha\rho/2\} > 1 - \beta$. From the continuity of $\partial J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta}$ w.r.t. $\boldsymbol{\theta}$, by taking the supremum of $N(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \mathcal{S}(\bar{\boldsymbol{\theta}}, \rho)$, we obtain that $\exists\hat{\boldsymbol{\theta}}_e^n$ in $\mathcal{B}(\bar{\boldsymbol{\theta}}, \rho)$ such that $\partial J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta}_{|\hat{\boldsymbol{\theta}}_e^n} = \mathbf{0}$. This property being true for any ρ small enough, $\hat{\boldsymbol{\theta}}_e^n \xrightarrow{a.s.} \bar{\boldsymbol{\theta}}$ as $n \to \infty$.

Using the same approach as in [9] one can show that the stationary solution $\hat{\boldsymbol{\theta}}_e^n$ corresponds to a (local) minimum of $J_e^s(\boldsymbol{\theta}, A_n, h_n)$ for n larger than some n_0. □

Remark 5.5. To prove that $\hat{\boldsymbol{\theta}}_e^n$ corresponds to a (local) minimum of $J_e^s(\boldsymbol{\theta}, A_n, h_n)$ for n larger than some n_0 in Theorem 5.4, one only needs that the property $\partial^2 J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top \xrightarrow{a.s.} \partial^2\mathrm{ent}(g^s)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ holds at $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$. Since $\partial g^s(x)/\partial\boldsymbol{\theta} = \mathbf{0}$ at $\bar{\boldsymbol{\theta}}$, one gets a better bound for the term C_n, see (5.4), in the proof of Lemma 5.3, namely $C_n < 2\gamma_n^2 H^2(A_n)$. One can easily check that taking $h_n = n^{-\alpha}$, $A_n = H^{-1}(n^\beta)$, with α, β, such that $\alpha > 0$, $\beta > 0$, $\alpha > 2\beta$, $1 - 6\alpha - 4\beta = \delta > 0$ is enough. For instance, $\alpha = 1/7$ and $\beta = 1/32$ give $\delta = 1/56$.

Remark 5.6. The fact that $\partial J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta}$ cannot be written as a sum of i.i.d. variables, see (5.2), makes the proof of asymptotic normality of $\hat{\boldsymbol{\theta}}_e^n$ difficult. However, we conjecture that for a suitable choice of A_n and h_n, $\sqrt{nh_n^\gamma}(\hat{\boldsymbol{\theta}}_e^n - \bar{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{M}^{-1}(\bar{\boldsymbol{\theta}}, \mu))$ for some $\gamma > 0$.

Remark 5.7. A different approach for proving the strong consistency of $\hat{\boldsymbol{\theta}}_e^n$ would be to show the almost sure equicontinuity of $J_e^s(\boldsymbol{\theta}, A_n, h_n)$, see [1]. Although we did not pursue this line, it seems more capable than the approach we used here of an extension to the proof of strong consistency of $\hat{\boldsymbol{\theta}}_e$ given by (1.8).

6. Conclusions and further developments

Minimising the entropy of a (symmetrized) kernel estimate of the distribution of errors, constructed from the residuals, seems to be a promising alternative to maximum-likelihood estimation for the case where this distribution is unknown. An example illustrated the accuracy of the estimation: provided that the bandwidth of the kernel estimate is large enough, the precision is similar to that obtained for least-squares estimation in the case of normal errors, but is better for errors having a uniform or Laplace density. The approach is also robust w.r.t. to the presence of outliers. The ultimate goal would be to obtain an estimator with asymptotic properties similar to those of the maximum-likelihood estimator, even though the distribution of errors is unknown (particular interest is for asymptotic efficiency, which could be called *"blind asymptotic efficiency"* in this case). Strong consistency of the estimator corresponding to a truncated estimate of the entropy (the integral is evaluated on an interval $[-A_n, A_n]$) has been proved. However, easy calculations show that this estimator is not asymptotically efficient (its covariance matrix decreases as slowly as $[nh_n^\gamma]^{-1}$, $\gamma > 0$, with h_n, the bandwidth of the kernel estimator, tending to zero as the sample size $n \to \infty$. Further work is required to check if blind asymptotic efficiency can be obtained for $\hat{\boldsymbol{\theta}}_e$ given by (1.8), that is, when the entropy is evaluated by integrating on $(-\infty, \infty)$. Finally, one may notice that the results obtained in this preliminary work are not restricted to regression problems, and could be of interest in other circumstances where maximum likelihood estimation cannot be used because the distribution of errors is unknown.

References

[1] G. Benitz. Asymptotic results for maximum likelihood estimation with an array of sensors. *IEEE Transactions on Information Theory*, 39(4):1374–1385, 1993.

[2] M. Bertrand-Retali. Convergence uniforme d'un estimateur de la densité par la méthode du noyau. *Rev. Roum. Math. Pures et Appl.*, 23(3):361–385, 1978.

[3] L. Devroye and L. Györfi. *Nonparametric density dstimation: The L_1 view.* Wiley, New York, 1985.

[4] Y. Dmitriev and F. Tarasenko. On the estimation of functionals of the probability density and its derivatives. *Theory of Probability and its Applications*, 18(3):628–633, 1973.

[5] D. Donoho. On minimum entropy deconvolution. In D. Findley, editor, *Applied Time Series Analysis II*, pages 565–608. Academic Press, New York, 1981.

[6] P. Eggermont and V. LaRiccia. Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE Transactions on Information Theory*, 45(4):1321–1326, 1999.

[7] V. Fedorov. *Theory of Optimal Experiments.* Academic Press, New York, 1972.

[8] I. Ford, D. Titterington, and C. Wu. Inference and sequential design. *Biometrika*, 72(3):545–551, 1985.

[9] C. Fourgeaud and A. Fuchs. *Statistique.* Dunod, Paris, 1967.

[10] P. Huber. *Robust Statistics.* John Wiley, New York, 1981.

[11] C. Loader. Bandwidth selection: classical or plug-in? *Annals of Statistics*, 27(2):415–438, 1999.

[12] A. Mokkadem. Estimation of the entropy and information of absolutely continuous random variables. *IEEE Transactions on Information Theory*, 35(1):193–196, 1989.

[13] E. Parzen. On estimation of a probability density function and mode. *Annals of Math. Stat.*, 35:1065–1076, 1962.

[14] B. Polyak and Y. Tsypkin. Robust identification. *Automatica*, 16:53–63, 1980.

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

[16] C. Wu. Asymptotic inference from sequential design in a nonlinear situation. *Biometrika*, 72(3):553–558, 1985.

Laboratoire I3S, CNRS/Université de Nice-Sophia Antipolis, bât. Euclide, Les Algorithmes, 2000 route des Lucioles, BP 121, 06903 Sophia-Antipolis Cedex, France, Tel: 33 (0)4 92 94 27 08; Fax: 33 (0)4 92 94 28 98

E-mail address: {pronzato,et}@i3s.unice.fr