ENTROPY MINIMIZATION FOR PARAMETER ESTIMATION PROBLEMS WITH UNKNOWN DISTRIBUTION OF THE OUTPUT NOISE

L. Pronzato, É. Thierry

Laboratoire I3S, CNRS/UNSA
Bât Euclide, Les algorithmes
2000 route des Lucioles, BP 121
06903 Sophia Antipolis Cedex, France

ABSTRACT

We consider the situation where the parameters $\boldsymbol{\theta}$ of a linear regression model have to be estimated from observations corrupted by an additive noise with unknown distribution f. Since maximum likelihood estimation cannot be used, we estimate $\boldsymbol{\theta}$ by minimizing the entropy of a kernel estimate of f, constructed from the residuals. An example of parameter estimation in the presence of interference with random binary signal is presented.

1. INTRODUCTION

Consider a linear regression model, with observations given by

$$y_i = \mathbf{r}^\top(\xi_i)\bar{\boldsymbol{\theta}} + \epsilon_i\,,\ i = 1, \ldots, n\,, \tag{1}$$

where $\bar{\boldsymbol{\theta}} \in \Theta \subset I\!R^p$ is the unknown true value of the model parameters, with Θ a compact set, $\mathbf{r}^\top(\xi_i)$ is the regressor for experimental conditions ξ, $\xi \in \mathcal{X} \subset I\!R^q$, and (ϵ_i) is a sequence of independently identically distributed (i.i.d.) errors, with probability density function (p.d.f.) $f(\cdot)$. The experimental conditions ξ_i may correspond for instance to passed values of an input sequence applied to a dynamical system, $\xi_i = (u_i, u_{i-1}, \ldots, u_{i-q})$, see Section 4.

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. We define

$$e_i(\boldsymbol{\theta}) = y_i - \mathbf{r}^\top(\xi_i)\boldsymbol{\theta}$$

and denote $\mathbf{e}_1^n(\boldsymbol{\theta}) = [e_1(\boldsymbol{\theta}), \ldots, e_n(\boldsymbol{\theta})]$. The variables ξ_i and ϵ_i are mutually independent, that is the design is *not sequential*: ξ_i does not depend on passed observations y_{i-1}, $y_{i-2}\ldots$ (this also covers the case where (ξ_i) is a deterministic sequence). Note, however, that the ξ_i's may be correlated, as it is the case for the dynamical example mentioned above. For the sake of simplicity, we assume that they can only take a finite number of values, that is, $\mathcal{X}$ is a finite set $\mathcal{X} = \{\xi^1, \ldots, \xi^m\}$. We also assume that the sequence (ξ_i) is ergodic: when n tends to infinity, each element ξ^j of $\mathcal{X}$ is used a fraction w_j of times. We denote by μ the discrete measure that attaches weights w_j to the support points ξ^j, $j = 1, \ldots, m$. Since a random permutation of the observations does not modify the estimation of $\boldsymbol{\theta}$, the ξ_i's can then be considered as forming an i.i.d. sequence, independent of (ϵ_i), with probability measure μ. We denote by $\mathbf{M}(\mu)$ the average Fisher information matrix *per* sample

$$\mathbf{M}(\mu) = \mathcal{I}(f)\,\left(\int_{\mathcal{X}} \mathbf{r}(\xi)\mathbf{r}^\top(\xi)\,\mu(d\xi)\right)\,, \tag{2}$$

and assume that $\mathcal{X}$ and μ are such that $\mathbf{M}(\mu)$ is positive-definite. When ξ corresponds to the input of a linear stationary model, the analytic expression of the matrix $\mathbf{M}(\mu)$ can easily be derived, see [2, 3]. For instance, consider the case where the model is a Finite Impulse Response (FIR) filter, that is, $\mathbf{r}(\xi_i) = (u_i, u_{i-1}, \ldots, u_{i-q})^\top$, (u_i) is stationary with power spectral density $p_u(\omega)$, and the sampling period T is normalized to 1. The matrix $\mathbf{M}(\mu)$ can then be written as

$$\mathbf{M}(\mu) = \mathbf{M}(p_u) = \frac{\mathcal{I}(f)}{2\pi}\int_{-\pi}^{\pi} \mathbf{b}(j\omega)\mathbf{b}^T(-j\omega)p_u(\omega)\,d\omega\,,$$

with $\mathbf{b}(j\omega) = [1\, e^{-j\omega}\, e^{-2j\omega}\, \ldots\, e^{-qj\omega}]^\top$ which gives, assuming that the signal u_i is real,

$$\mathbf{M}(p_u) = \frac{\mathcal{I}(f)}{\pi}\int_0^{\pi} \bar{\mathbf{M}}(\omega)p_u(\omega)\,d\omega\,, \tag{3}$$

with

$$\bar{\mathbf{M}}(\omega) = \begin{pmatrix} 1 & \cos(\omega) & \cdots & \cos(q\omega) \\ \cos(\omega) & 1 & \cdots & \cos[(q-1)\omega] \\ \vdots & \vdots & \vdots & \vdots \\ \cos(q\omega) & \cos[(q-1)\omega] & \cdots & 1 \end{pmatrix}.$$

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 $f_{\boldsymbol{\theta}}(y,\xi) = f[y - \mathbf{r}^\top(\xi)\boldsymbol{\theta}]$; that is, $\hat{\boldsymbol{\theta}}_{ML}$ simply minimises the sample version of the (Shannon) entropy of $f(\cdot)$, evaluated at $\mathbf{e}_1^n(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\min_{\boldsymbol{\theta}\in\Theta} -\sum_{i=1}^{n} \log f[e_i(\boldsymbol{\theta})]\,. \tag{4}$$

Under standard assumptions, see, *e.g.*, [1], $\hat{\boldsymbol{\theta}}_{ML}$ possesses the asymptotic properties of consistency: $\hat{\boldsymbol{\theta}}_{ML} \xrightarrow{a.s.} \bar{\boldsymbol{\theta}}$; asymptotic normality:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \bar{\boldsymbol{\theta}}) \xrightarrow{d} z \sim \mathcal{N}(\mathbf{0}, \mathbf{M}^{-1}(\mu))\,;$$

and asymptotic efficiency: $\mathbf{M}(\mu)^{-1}$ is the Cramer-Rao lower bound.

When $f(\cdot)$ is unknown, using a wrong distribution in the calculation of $\hat{\boldsymbol{\theta}}_{ML}$ makes the approach suboptimal. For instance, the Least Squares (LS) estimator satisfies $\sqrt{n}(\hat{\boldsymbol{\theta}}_{LS} - \bar{\boldsymbol{\theta}}) \xrightarrow{d} z \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathcal{I}(f)\mathbf{M}^{-1}(\mu))$, with σ^2 the variance of ϵ_i, and, for σ^2 fixed, the minimum value of $\sigma^2\mathcal{I}(f)$ is 1 and is obtained for the normal distribution: the LS estimator is thus suboptimal, in terms of precision of the estimation, for any distribution other than the normal. The approach we suggest tries to estimate $f(\cdot)$ and $\boldsymbol{\theta}$ simultaneously by minimizing the entropy of an estimate of $f(\cdot)$ based on the empirical distribution of the errors $\mathbf{e}_1^n(\boldsymbol{\theta})$, with the objective of approaching the ML estimator even when $f(\cdot)$ is unknown. One can refer to [4, 5] for a more detailed exposition.

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. $f(\cdot)$. Using the results in [6, 7, 8], one can show that the kernel estimate $\hat{f}_{n,h_n}[x|\mathbf{e}_1^n(\boldsymbol{\theta})]$ converges to $G(x)$ given by

$$G(x) = G(x,\boldsymbol{\theta},\bar{\boldsymbol{\theta}},\mu) = \int_{\mathcal{X}} f[x + \mathbf{r}^\top(\xi)(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})]\,\mu(d\xi) \tag{5}$$

under reasonable assumptions on $f(\cdot)$, $\eta(\boldsymbol{\theta},\xi)$, $K(\cdot)$ and h_n. Now, since the entropy of a p.d.f. $f(\cdot)$, which is given by

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

is invariant by translation, the minimization of the criterion $J_e(\boldsymbol{\theta}) = \text{ent}\{\hat{f}_{n,h}[\,\cdot\,|\mathbf{e}_1^n(\boldsymbol{\theta})]\}$ not suitable for estimating $\boldsymbol{\theta}$, and we shall consider instead

$$J_e^s(\boldsymbol{\theta}) = \text{ent}\{\hat{f}_{n,h}[\,\cdot\,|\mathbf{e}_1^n(\boldsymbol{\theta}), -\mathbf{e}_1^n(\boldsymbol{\theta})]\}\,, \tag{6}$$

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

$$\begin{aligned} G^s(x) \quad &= \quad G^s(x,\boldsymbol{\theta},\bar{\boldsymbol{\theta}},\mu) = \frac{1}{2}\int_{\mathcal{X}}\{f[x + \mathbf{r}^\top(\xi)(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})] \\ &\qquad + f[x - \mathbf{r}^\top(\xi)(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})]\}\,\mu(d\xi)\,. \end{aligned} \tag{7}$$

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

$$\hat{\boldsymbol{\theta}}_e = \arg\min_{\boldsymbol{\theta}\in\Theta} J_e^s(\boldsymbol{\theta})\,. \tag{8}$$

Section 2 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 (6), is considered in Section 3. An example of parameter estimation for a Finite Impulse Response (FIR) model in presence of interference with an unknown binary signal is presented in Section 4. Finally, Section 5 draws some conclusions.

2. SOME PROPERTIES OF $\text{ent}(G^S)$

Under suitable assumptions, see, *e.g.*, [9, 10], $J_e^s(\boldsymbol{\theta})$ given by (6) converges to $\text{ent}(G^s)$, with $G^s(e)$ given by (7), hence the interest of studying the properties of $\text{ent}(G^s)$.

Easy calculation gives

$$\frac{\partial G^s(e)}{\partial\boldsymbol{\theta}}_{|\bar{\boldsymbol{\theta}}} = \mathbf{0}\,,\ \frac{\partial^2 G^s(e)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}_{|\bar{\boldsymbol{\theta}}} = f''(e)\int_{\mathcal{X}}\mathbf{r}(\xi)\mathbf{r}^\top(\xi)\,\mu(d\xi)$$

and

$$\begin{aligned} \frac{\partial\text{ent}(G^s)}{\partial\boldsymbol{\theta}} \quad &= \quad -\int_{-\infty}^{\infty}[1 + \log G^s(e,\boldsymbol{\theta},\bar{\boldsymbol{\theta}},\mu)] \\ &\qquad \frac{\partial G^s(e,\boldsymbol{\theta},\bar{\boldsymbol{\theta}},\mu)}{\partial\boldsymbol{\theta}}\,de\,; \end{aligned}$$

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

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

Noticing that $(f\log f)'' = (1 + \log f)f'' + (f')^2/f$, one gets $\partial^2\text{ent}(G^s)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top_{|\bar{\boldsymbol{\theta}}} = \mathbf{M}(\bar{\boldsymbol{\theta}},\mu)$, see (2). The criterion

ent(G^s) is thus locally convex 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 [11].

3. STRONG CONSISTENCY OF A TRUNCATED ESTIMATOR

We use an approach similar to [12], and consider the criterion

$$J_e^s(\boldsymbol{\theta}, A_n, h_n) = \text{ent}_{A_n}(\hat{f}_{n,h_n}), \tag{9}$$

where

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

We denote the associated estimator by $\hat{\boldsymbol{\theta}}_e^n$,

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

Besides the assumptions made in the introduction, we 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 $s = 1, 2$. Let $H(\cdot)$ denote a monotonic strictly increasing function such that $\forall U$, $\sup_{|z|<U} 1/G^s(z) \leq H(U)$, see [12], 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 $J_e^s(\boldsymbol{\theta}, A_n, h_n)$, see [5].

Lemma 1 *For any $\boldsymbol{\theta} \in \Theta$,*

$$\begin{aligned} \partial J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta} &\xrightarrow{a.s.} \partial \text{ent}(G^s)/\partial\boldsymbol{\theta} \text{ and} \\ \partial^2 J_e^s(\boldsymbol{\theta}, A_n, h_n)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top &\xrightarrow{a.s.} \partial^2 \text{ent}(G^s)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top \end{aligned}$$

as $n \to \infty$ when $h_n = n^{-1/7}$, $A_n = H^{-1}(n^{1/48})$.

One can then show that Lemma 1 implies strong consistency of $\hat{\boldsymbol{\theta}}_e^n$ for a suitable choice of h_n and A_n, see [5].

Theorem 1 *When $h_n = n^{-1/7}$, $A_n = H^{-1}(n^{1/48})$, there exists a sequence $(\hat{\boldsymbol{\theta}}_e^n)$ 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.*

4. EXAMPLE

Consider a parameter estimation problem for a FIR model in presence of interferences. We observe $y_t = A(q)u_t + B(q)v_t + \zeta_t$, where $A(q) = \sum_{i=0}^{q} a_i q^{-i}$, with q^{-1} the delay operator, corresponds to a FIR filter with unknown parameters $\boldsymbol{\theta} = (a_0, \ldots, a_q)$, $B(q)$ is also an unknown FIR filter, ζ_t corresponds to an i.i.d. sequence of errors, u_t is a known input signal, v_t is an unknown interfering signal. With the same notation as in (1), we thus have $\epsilon_t = B(q)v_t + \zeta_t$ the output noise. Note that the sequence (ϵ_t) is correlated even if the v_t's are i.i.d., due to the action of the filter B. However, the results in [13] show that $\hat{f}_{n,h_n}[x|\mathbf{e}_1^n(\bar{\theta}), -\mathbf{e}_1^n(\bar{\theta})]$ still converges to $G^s(x)$ given by (7) for a suitably decreasing sequence (h_n) and the results below show that the estimator $\hat{\boldsymbol{\theta}}_e$ given by (8) still possesses attractive properties in presence of correlated errors.

In the simulations below, (v_t) is an independent binary sequence, $v_t = \pm 1$, $B(q) = 1 - 0.5q^{-1} + 0.2q^{-2}$, $q = 4$, $\bar{\boldsymbol{\theta}} = (1, -0.5, 0.2, -0.3, 0.1)$ and $\zeta_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. We take $u_t = \sum_{i=0}^{3} \cos[(2i+1)\pi t/8]$, which corresponds to a D-optimal input signal with unit average power for estimating $\boldsymbol{\theta}$, see [3]: it maximizes $\det \mathbf{M}(p_u)$ under the constraint $(1/\pi)\int_0^\pi p_u(\omega)d\omega = 1$. We set h to 0.1 in the computation of $\hat{\boldsymbol{\theta}}_e$ and the minimization of $J_e^s(\boldsymbol{\theta})$, see (6), is initialized at $\hat{\boldsymbol{\theta}}_{LS}$.

Figure 1 gives an histogram of the errors $\epsilon(t)$ and Figure 2 gives a typical realization of the density reconstructed from the residuals $e_i(\hat{\boldsymbol{\theta}}_e)$ (full line) and $e_i(\hat{\boldsymbol{\theta}}_{LS})$ (dashed line) with the bandwidth $h = 0.1$. The location of the residuals on the horizontal axis is indicated by stars ($\hat{\boldsymbol{\theta}}_e$) and crosses ($\hat{\boldsymbol{\theta}}_{LS}$). It is clear from these figures that a better reconstruction of the density of the errors is obtained when using $\hat{\boldsymbol{\theta}}_e$, so that $\hat{\boldsymbol{\theta}}_e$ maximizes a rather good approximation of the likelihood function. Repeated simulations indicate that the mean-squared errors for the components of $\hat{\boldsymbol{\theta}}_{LS}$ are more that two times larger than those of $\hat{\boldsymbol{\theta}}_e$.

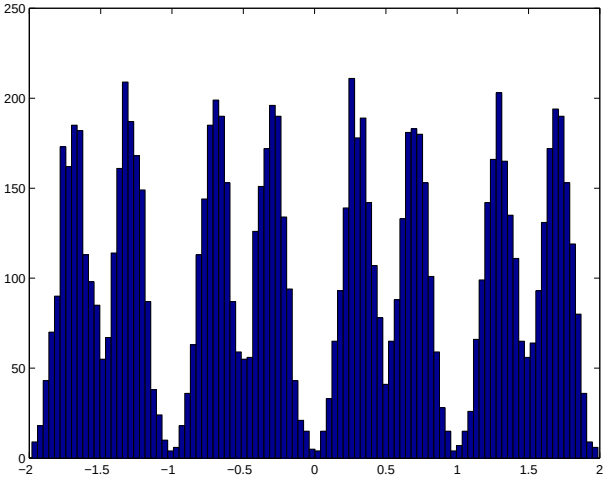

Fig. 1. Histogram of the errors $\epsilon(t)$

5. CONCLUSIONS

We suggest to minimise the entropy of a (symmetrized) kernel estimate of the distribution of output errors, constructed

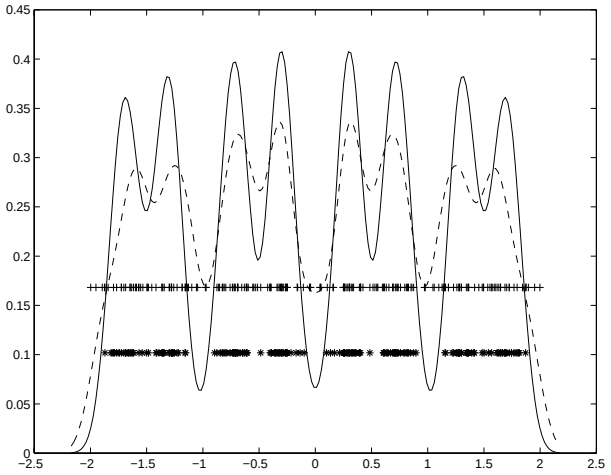

Fig. 2. Typical reconstruction of the density of the errors from the residuals: $\hat{\boldsymbol{\theta}}_e$ (full line and stars), $\hat{\boldsymbol{\theta}}_{LS}$ (dashed line and crosses)

from the residuals, as an alternative to LS estimation for the case where the distribution of these errors is unknown and maximum likelihood cannot be used. An example of estimation in presence of interferences with an unknown signal illustrates the attractive properties of the approach.

A far reaching target 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 (this concerns in particular asymptotic efficiency, which could be called *"blind asymptotic efficiency"* in this case). Such developments are currently under study.

6. REFERENCES

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

[2] G.C. Goodwin and R.L. Payne, *Dynamic System Identification: Experiment Design and Data Analysis*, Academic Press, New York, 1977.

[3] E. Walter and L. Pronzato, *Identification of Parametric Models from Experimental Data*, Springer, Heidelberg, 1997.

[4] L. Pronzato and E. Thierry, "A minimum-entropy estimator for regression problems with unknown distribution of observation errors," in *Proceedings of MaxEnt'2000*, Paris, July 2000, Kluwer.

[5] L. Pronzato and E. Thierry, "A minimum-entropy estimator for regression problems with unknown distribution of observation errors," Tech. Rep. 00–08, Laboratoire I3S, CNRS–Université de Nice-Sophia Antipolis, 06903 Sophia Antipolis, France, 2000, `http://www.i3s.unice.fr/~pronzato/` .

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

[7] M. Bertrand-Retali, "Convergence uniforme d'un estimateur de la densité par la méthode du noyau," *Rev. Roum. Math. Pures et Appl.*, vol. 23, no. 3, pp. 361–385, 1978.

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

[9] A. Mokkadem, "Estimation of the entropy and information of absolutely continuous random variables," *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 193–196, 1989.

[10] P.B. Eggermont and V.N. LaRiccia, "Best asymptotic normality of the kernel density entropy estimator for smooth densities," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1321–1326, 1999.

[11] D. Donoho, "On minimum entropy deconvolution," in *Applied Time Series Analysis II*, D.F. Findley, Ed., pp. 565–608. Academic Press, New York, 1981.

[12] Yu.G. Dmitriev and F.P. Tarasenko, "On the estimation of functionals of the probability density and its derivatives," *Theory of Probability and its Applications*, vol. 18, no. 3, pp. 628–633, 1973.

[13] J.V. Castellana and M.R. Leadbetter, "On smoothed probability density estimation for stationary processes," *Stochastic processes and their Applications*, vol. 21, pp. 179–193, 1986.