Minimum Entropy Estimation in Semi-Parametric Models: a candidate for adaptive estimation?

Luc Pronzato, Éric Thierry and Éric Wolsztynski

Laboratoire I3S, Université de Nice-Sophia Antipolis/CNRS
{pronzato,et,wolsztyn}@i3s.unice.fr

Summary. In regression problems with errors having an unknown density f, least squares or robust M-estimation is the usual alternative to maximum likelihood, with the loss of asymptotic efficiency as a consequence. The search for efficiency in the absence of knowledge of f (adaptive estimation) has motivated a large amount of work, see in particular (Stein, 1956; Stone, 1975; Bickel, 1982) and the review paper (Manski, 1984). The present paper continues the work initiated in (Pronzato and Thierry, 2001a,b). The estimator is obtained by minimizing (an estimate of) the entropy of the symmetrized residuals. Connections and differences with previous work are indicated. The focus is mainly on the location model but we show how the results can be extended to nonlinear regression problems, in particular when the design consists of replications of a fixed design. Numerical results illustrate that asymptotic efficiency is not necessarily in conflict with robustness.

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

1 Introduction

Consider a 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 $\theta \in \Theta \subset \mathbb{R}^p$, (ε_i) forms a sequence of i.i.d. random variables with p.d.f. f and $\eta(\theta, x)$ is a known function of θ and x, the design variable. Most of the paper is devoted to the simplest case, the location model, where $\eta(\theta, x) = \theta$, but extension to nonlinear regression is considered in Section 4.

If f is known, Maximum Likelihood (ML) estimation can be used and, under standard assumptions, is *asymptotically efficient*: $\sqrt{n}(\hat{\theta}^n_{ML} - \bar{\theta}) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \mathbf{M}_F^{-1})$,

with $\mathbf{M}_F$ the Fisher information matrix. Suppose now that the density f is only known to be symmetric about zero (further regularity assumptions about f will be used in Section 3). The model 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. This corresponds to the approach taken by Stein in his seminal paper (1956). The absence of knowledge of f usually induces a loss of efficiency. An estimator that remains asymptotically efficient in these conditions is called *adaptive* (see Bickel (1982) for a precise definition and Begun et al. (1983) for a necessary condition for adaptive estimation). 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 from residuals obtained with a preliminary $\sqrt{n}$-consistent estimator. It is this second approach which has been further developed by Bickel (1982), see also Manski (1984). Here we shall follow the approach taken in (Pronzato and Thierry, 2001a,b) (PT-2001a,b in what follows) and minimise the entropy of a kernel estimate of f based on the $2n$ symmetrized residuals $e_i(\theta), -e_i(\theta)$, $i = 1, \ldots, n$, with $e_i(\theta) = Y_i - \eta(\theta, x_i)$ the ith residual in the regression model (1). The motivation is given in Section 2, where the connections and differences with the Stone-Bickel approach are detailed. The simplest case of a location problem is considered in Section 3 where adaptivity of a particular minimum-entropy estimator is proved. Extension to nonlinear regression is considered in Section 4, with an example illustrating the robustness properties of the estimator.

2 Minimizing entropy

The ML estimator $\hat{\theta}_{ML}^n$ (for f known) minimizes

$$\bar{H}_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log f[e_i(\theta)] \tag{2}$$

with respect to θ. The initial motivation in (PT-2001a,b) comes from the simple observations that (i) $\bar{H}_n(\bar{\theta}) = -(1/n)\sum_{i=1}^{n} \log f(\varepsilon_i)$ is an empirical version of $H(f) = -\int \log[f(x)]f(x)dx$, (ii) the entropy of a distribution is a measure of its dispersion, (iii) the entropy of the true distribution of the symmetrized residuals is minimum at $\theta = \bar{\theta}$ (symmetrized residuals are used because the entropy is shift-invariant). A possible construction when f is unknown is as follows: construct a kernel estimate $\hat{f}_n^\theta$ from the symmetrized residuals (which ensures that $\hat{f}_n^\theta$ is symmetric); compute its entropy, which forms the estimation criterion $\hat{H}_n(\theta)$ to be minimized. In (PT-2001a,b), $\hat{H}_n(\theta)$ is given by

$$\hat{H}_n(\theta) = -\int_{-A_n}^{A_n} \log[\hat{f}_n^\theta(x)]\hat{f}_n^\theta(x)dx \tag{3}$$

with (A_n) a suitably increasing sequence of positive numbers (to be chosen in accordance with the decrease of the bandwidth h_n of the kernel estimate $\hat{f}_n^\theta$). Other estimates of the entropy of $\hat{f}_n^\theta$ will be used in what follows.

In the location model, this construction can be justified following an approach similar to Beran (1978). The distribution of the observations Y_i has the density $g(y) = f(y - \bar{\theta})$. Define $\hat{\beta} = (\hat{\theta}, \hat{f})$ in the semi-parametric model, where $\hat{\theta}$ is a postulated value for $\bar{\theta}$ and $\hat{f}$ a postulated symmetric p.d.f., and let $g_{\hat{\beta}}(y)$ be the associated density for the observations, $g_{\hat{\beta}}(y) = \hat{f}(y - \hat{\theta})$. Assume that a kernel estimate $\hat{g}_n$ of g has been formed, based on the observations $Y_1, \ldots, Y_n$. The estimator in (Beran, 1978) is based on the Hellinger distance between $\hat{g}_n$ and $g_{\hat{\beta}}$. Here we use the Kullback-Leibler divergence $L(\hat{g}_n, g_{\hat{\beta}}) = \int \log[\hat{g}_n(y)/g_{\hat{\beta}}(y)]\hat{g}_n(y)dy$. Straightforward calculation shows that the symmetric $\hat{f}$ and parameter $\hat{\theta}$ that minimize $L(\hat{g}_n, g_{\hat{\beta}})$ respectively correspond to $\hat{f}_n = \hat{f}_n^{\hat{\theta}^n}$ with $\hat{f}_n^\theta(u) = [\hat{g}_n(u + \theta) + \hat{g}_n(-u + \theta)]/2$ and $\hat{\theta}^n = \arg\min_\theta H(\hat{f}_n^\theta)$. One may then notice that $\hat{f}_n^\theta$ is a kernel estimate based on the symmetrized residuals $Y_i - \theta, -Y_i + \theta$, and $\hat{\theta}^n$ minimizes its entropy.

This presents some similarities with the Stone-Bickel approach (1975; 1982), which relies on an approximation of the score function, that is, the derivative of $\bar{H}_n(\theta)$, given by (2), with respect to θ. The construction is in two stages: first, an estimator $\hat{\theta}_1^n$ asymptotically locally sufficient (in the sense of Le Cam) is constructed; second, an approximated score function is used to perform a Newton-Raphson step from $\hat{\theta}_1^n$. Stone (1975) shows that the construction is adaptive for the location model with symmetric errors; Bickel (1982) and then Manski (1984) extend the result to other models, including non linear regression, see also Andrews (1989). Although the construction of $\hat{H}_n(\theta)$ may rely on a similar kernel estimate, there are some key differences between the Stone-Bickel approach and the minimum-entropy method presented here. First, the estimation criterion $\hat{H}_n$, see (3,4), is minimized through a *series of* minimization steps, using an optimization algorithm. Second, all the data are treated similarly, whereas, for technical reasons, the developments in (Bickel, 1982; Manski, 1984) rely on sample splitting: m observations are used to construct a preliminary parameter estimate $\hat{\theta}^m$ to form residuals and a score function estimate; the $n - m$ remaining observations are used for the Newton-Raphson step from $\hat{\theta}_1^n$, with the requirement that $m \to \infty$ and $m/n \to 0$ as $n \to \infty$. One may notice that the numerical results presented in (Manski, 1984) show that this sample splitting degrades the performance of the estimator. The estimator proposed by Andrews (1989) does not rely on sample splitting and is defined to minimize an estimation criterion rather than to be a one-step estimator. However, the criterion (likelihood) is also constructed from a $\sqrt{n}$-consistent preliminary estimate $\hat{\theta}_1^n$, used to form a kernel estimate of f and hence of the likelihood function. On the contrary, $\hat{H}_n$ does not depend on any preliminary estimate.

One motivation for minimizing the entropy of the distribution of the residuals is that it allows a lot of flexibility: many methods are available to estimate the entropy $\hat{H}_n(\theta)$, and kernel estimation 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. Studying the application of these methods to semi-parametric estimation by entropy minimisation is quite challenging.

Define $\hat{\theta}^n = \arg\min_{\theta\in\Theta} \hat{H}_n(\theta)$, with Θ a compact subset of $\mathbb{R}^p$ and $\hat{H}_n(\theta)$ some estimate of the entropy of the distribution of the symmetrized residuals in the model (1). We assume that $\bar{\theta} \in \mathrm{int}(\Theta)$ and that $\hat{H}_n(\theta)$ is two times continuously differentiable with respect to $\theta \in \Theta$. 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 $\xrightarrow{\mathrm{d}}$; $\nabla F(\theta)$ and $\nabla^2 F(\theta)$ denote the first and second order derivatives of the function F with respect to θ. The adaptivity of $\hat{\theta}^n$ can be proved by following the steps below (leaving aside some usual measurability conditions, see, e.g., Lemmas 1,2 and 3 of Jennrich (1969)):

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

B) show that $\nabla^2 \hat{H}_n(\theta) \overset{\theta,\mathrm{p}}{\rightsquigarrow} \nabla^2 H(\theta)$, 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 $n\to\infty$.

A proves that $\hat{\theta}^n \xrightarrow{\mathrm{p}} \bar{\theta}$. A and B imply that $\nabla^2 \hat{H}_n(\hat{\theta}^n) \xrightarrow{\mathrm{p}} \mathbf{M}_2 = \nabla^2 H(\bar{\theta}) \succ 0$. Consider the following Taylor development of $\nabla \hat{H}_n(\theta)$ at $\theta = \hat{\theta}^n$, similar to that used in Jennrich (1969) for LS estimation: $\nabla \hat{H}_n(\hat{\theta}^n) = \mathbf{0} = \nabla \hat{H}_n(\bar{\theta}) + (\hat{\theta}^n - \bar{\theta})^\top \nabla^2 H[\alpha_n \hat{\theta}^n + (1-\alpha_n)\bar{\theta}]$, for some $\alpha_n \in [0,1]$. C then implies $\sqrt{n}(\hat{\theta}^n - \bar{\theta}) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \mathbf{M}_2^{-1}\mathbf{M}_1\mathbf{M}_2^{-1})$ and adaptivity is proved provided that $\mathbf{M}_2^{-1}\mathbf{M}_1\mathbf{M}_2^{-1} = \mathbf{M}_F^{-1}$, the inverse of the Fisher information matrix for the model (1). Step C allows some freedom in the choice of the function $\bar{H}_n(\theta)$. However, a natural candidate is (2), for which asymptotic normality of $\sqrt{n}\nabla \bar{H}_n(\bar{\theta})$ holds under standard assumptions, with $\mathbf{M}_1 = \mathbf{M}_F$.

A most important remark at this point is that (uniform) $\sqrt{n}$-consistency of the entropy estimate $\hat{H}_n(\theta)$ is not a prerequisite for $\sqrt{n}$-consistency of $\hat{\theta}^n$ (we only need $\sqrt{n}$-consistency of $\Delta_n(\bar{\theta})$). It is important because $\sqrt{n}$-consistency does not seem to be a widespread property among entropy estimation methods, see Beirlant et al. (1997).

3 Adaptive estimation in the location model

The residuals are given by $e_i(\theta) = Y_i - \theta$. Consider the kernel estimates

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

where the kernel K is a p.d.f. symmetric about 0 that satisfies $\int |u|K(u)du < \infty$, K and its first two derivatives being continuous and of bounded variation, see Schuster (1969) (the density of the standard normal satisfies these conditions). We assume that f has unbounded support, f and its derivatives $f^{(s)}$ are bounded for $s = 1, 2, 3$, $H(f) < \infty$ and f has a finite Fisher information for location, $i(f) = \int [f'(x)]^2/f(x)dx < \infty$. Consider the entropy estimate given by

$$\hat{H}_n = -\frac{1}{n} \sum_{i=1}^{n} \log\{f_{n,i}^{\theta}[e_i(\theta)]\} U_n[e_i(\theta)] \tag{4}$$

where $f_{n,i}^{\theta}(u) = (1/2)[k_{n,i}^{\theta}(u) + k_{n,i}^{\theta}(-u)]$ and $U_n(x) = u(|x|/A_n - 1)$, with $u(z) = 1$ for $z \leq 0$, 0 for $z \geq 1$ and $u(z)$ varying smoothly in between, $u'(0) = u'(1) = 0$, and $\max_z |u'(z)| = d_1 < \infty$, $\max_z |u''(z)| = d_2 < \infty$. $\hat{H}_n(\theta)$ is then continuous (and two times continuously differentiable) in θ for any n. As in (Dmitriev and Tarasenko, 1973), we assume that exists a (strictly increasing) function $B(x)$ such that for all x, $B(x) \geq \sup_{|y|\leq x} 1/f(y)$. Define $B_n = B(2A_n + L)$. Using (Dmitriev and Tarasenko, 1973, Theorem 4) and (Newey, 1991, Corollary 3.1) one can then show that $\hat{H}_n(\theta) \overset{\theta,\mathrm{p}}{\rightsquigarrow} H(\theta)$ as $n \to \infty$, provided that A_n (and thus B_n) increases slowly enough, and the bandwidth h_n of the kernel estimator decreases slowly enough ($B_n = n^{\alpha}$, $h_n = 1/[n^{\alpha} \log n]$ with $\alpha < 1/3$ is suitable). Here, $H(\theta)$ is the entropy of the true distribution of the symmetrized residuals for θ, $H(\theta) = -\int \log[\pi^{\theta}(x)]\pi^{\theta}(x)dx$ with $\pi^{\theta}(x) = (1/2)[f(x + \theta - \bar{\theta}) + f(x - \theta + \bar{\theta})]$, which is minimum at $\theta = \bar{\theta}$. This proves point A of Section 2, and thus the consistency of $\hat{\theta}^n$ that minimizes $\hat{H}_n$; the details will be presented elsewhere. Similarly, with slightly stronger conditions on f one can prove point B, that is, $\nabla^2 \hat{H}_n(\theta) \overset{\theta,\mathrm{p}}{\rightsquigarrow} \nabla^2 H(\theta)$, with $\nabla^2 H(\bar{\theta}) = i(f)$, e.g. when $B_n = n^{\alpha}$, $h_n = 1/[n^{\alpha} \log n]$, $\alpha < 1/7$. The adaptivity of $\hat{\theta}^n$, i.e. step C, would then follow from

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

The conditions required on the functions f, K and u for this to hold are currently under investigation. One may notice that a difficulty which is not present in the Stone-Bickel approach is due to the fact that $(k_{n,i}^{\bar{\theta}})'(-x) \neq -(k_{n,i}^{\bar{\theta}})'(x)$.

Example 1. In this example and Example 2 we take $A_n = \infty$ in (3) and $U_n(x) \equiv 1$ in (4); the bandwidth h of the kernel estimator is chosen by the double kernel method (Berlinet and Devroye, 1994) and based on residuals obtained from a robust M-estimator. Table 1 gives the empirical value $\hat{C}_n$

of $\mathtt{E}\{\hat{v}_n \hat{v}_n{}^\top\}$, with $\hat{v}_n = \sqrt{n}(\hat{\theta}^n - \bar{\theta})$, obtained from 100 repetitions of the estimation procedure, for different choices of the estimator $\hat{\theta}$ and p.d.f. f. The number of observations is $n = 100$.

Table 1. Value of $\hat{C}_n$ in the location model for the LS, Minimum Hellinger Distance (Beran, 1978) and Minimum Entropy estimators ((3) for ME_1 and (4) for ME_2). The errors are standard normal, bi-exponential ($f(x) = (1/\sqrt{2})\exp(-\sqrt{2}|x|)$), and Student's t_ν with $\nu = 3, 5$ and 10 degrees of freedom

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

4 Adaptive estimation in nonlinear regression

We return to the nonlinear regression model (1), and assume that the design consists of replications at fixed points: the design measure ξ has a finite number of support points $X^1, \ldots, X^m$, X^j receiving the weight ξ^j. Therefore, for a total of n observations, $n_j = n\xi^j$ are made at $X = X^j$.

Consider first a 2-stage estimation method, with m minimum entropy estimations at the first step and one (weighted) LS estimation at the second. For each X^j, $j = 1, \ldots, m$, 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 an adaptive estimator: we have $\sqrt{n_j}[\hat{\eta}^j - \eta(\bar{\theta}, X^j)] \xrightarrow{\mathrm{d}} \mathcal{N}(0, i^{-1}(f))$ as $n_j \to \infty$. Having solved m such location problems, we form a LS estimation problem by considering the estimated responses $\hat{\eta}^j$ as pseudo-observations, and minimize $J_n^{WLS}(\theta) = \sum_{j=1}^m \xi^j [\eta(\theta, X^j) - \bar{\eta}^j]^2$. Assume that $\bar{\theta} \in \mathrm{int}(\Theta)$, with Θ a compact subset of $\mathbb{R}^p$, that $\eta(\theta, X^j)$ is two times continuously differentiable with respect to $\theta \in \Theta$ for any j, that the Fisher information matrix $\mathbf{M}_F(\bar{\theta}) = i(f) \sum_{j=1}^m \xi^j \nabla\eta(\bar{\theta}, X^j)[\nabla\eta(\bar{\theta}, X^j)]^\top$ is definite positive and that the identifiabilty condition $\{[\eta(\theta, X^j) = \eta(\bar{\theta}, X^j), j = 1, \ldots, m] \Rightarrow \theta = \bar{\theta}\}$ is satisfied. One can then show that the 2-stage LS estimator $\hat{\theta}_{LS}^n$ that minimizes $J_n^{WLS}(\theta)$ satisfies $\sqrt{n}(\hat{\theta}^n - \bar{\theta}) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \mathbf{M}_F^{-1}(\bar{\theta}))$.

Although $\hat{\theta}_{LS}^n$ is asymptotically efficient, one can expect a one-stage adaptive estimator to exhibit a better finite sample behavior. Using a justification similar to that given in Section 2 for the location model, we suggest the following procedure: (i) form kernel estimates $\hat{f}^{j,\theta}$ of the distribution of (symmetrized) residuals *for each design point* X^j *separately* and compute their respective entropies $H(\hat{f}^{j,\theta})$, (ii) compute

$$\hat{\theta}^n = \arg\min_{\theta\in\Theta} \mathtt{E}_\xi\{\hat{H}_n(\theta, X)\} \text{ with } \hat{H}_n(\theta, X^j) = H(\hat{f}^{j,\theta}), j = 1, \ldots, m. \quad (5)$$

The adaptivity of the method would then follow from adaptivity in the location model.

This approach does not extend to more general designs (whereas, for instance, randomized designs are considered in (Bickel, 1982) and in (Manski, 1984) the empirical distribution of design points converges a.s. at rate $\sqrt{n}$ to some distribution with non singular variance; see also the conditions imposed on the design in (Jennrich, 1969)). The difficulty comes from the fact that estimating the entropy of the symmetrized residuals at each design point X is not possible without replications of observations. One possibility, which is used in (PT-2001a,b), is to mix all (symmetrized) residuals together and estimate the entropy $\hat{H}_n(\theta)$ of their distribution by (3) or (4). Replace ξ by ξ_n in (5), where ξ_n is empirical measure of the design points X_i. Let U be a random variable with distribution conditional on X given by $\hat{f}^{j,\theta}$ of (5). Then, $\mathtt{E}_{\xi_n}\{\hat{H}_n(\theta, X)\} = \mathcal{H}(U|X)$ the conditional entropy of U given X. From a standard result in information theory, $\mathcal{H}(U|X) \leq \mathcal{H}(U) = \hat{H}_n(\theta)$. The entropy $\hat{H}_n(\theta)$ obtained by mixing up all residuals, which can be constructed for any design, is thus an upper bound on the criterion $\mathtt{E}_{\xi_n}\{\hat{H}_n(\theta, X)\}$ minimized in (5). Studying the adaptivity of the corresponding estimator will form the subject of further work. Some preliminary numerical results are presented below.

Example 2. Consider the regression model $\eta(\theta, x) = \theta_1 \exp(-\theta_2 x)$, with $\bar{\theta} = (100, 2)^\top$, the design measure is supported at $X^j = 1+(j-1)/9$, $j = 1, \ldots, 10$, with $\xi^j = 1/10$ for all j. The results are in Table 2, with the same notations as in Example 1. We use 100 repetitions with 100 observations for each. ME uses (3), both MHD from (Beran, 1978) and ME mix all residuals together. In the last three lines of the table we introduce q outliers ε_k ($\mathcal{N}(2, 10)$, randomly allocated among the X^j's) in addition to the n observations ($\hat{C}_n$ is still computed for $n = 100$, f is the bi-exponential).

References

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

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

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

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

Table 2. Values (T_n, D_n) of the trace ($\times 10^{-3}$) and determinant ($\times 10^{-2}$) of $\hat{C}_n$ in Example 2 for different f and estimators (with q outliers for the last three lines)

f	$\mathcal{N}(0,1)$	exp	t_3	t_5	t_{10}
optimum ($\hat{C}_n = \mathbf{M}_F^{-1}$)	(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)
q (nb. outliers, f = exp)	20	60	80		
LS	(35.0, 4.2)	(128.4, 7.3)	(20.3, 12.4)		
MHD	(5.3, 1.4)	(17.9, 3.2)	(2.9, 2.5)		
ME	(4.9, 0.8)	(17.9, 2.2)	(2.9, 1.6)		

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

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

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

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*, 18(3):628–633, 1973.

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

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

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

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

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

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

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

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