MINIMUM ENTROPY ESTIMATION IN SEMI PARAMETRIC MODELS

Éric Wolsztynski, Éric Thierry, Luc Pronzato

Laboratoire I3S
Université de Nice-Sophia Antipolis — CNRS
Bât. Euclide, Les Algorithmes
2000 route des Lucioles, BP 121
06903 Sophia Antipolis cedex, France

ABSTRACT

This paper is a continuation of the work initiated in [1, 2]: we estimate parameters in a regression model, linear or not, by minimizing (an estimate of) the entropy of the symmetrized residuals, obtained by a kernel estimation of their distribution. The objective is to obtain efficiency in the absence of knowledge of the density f of the observation errors, which is called adaptive estimation, see in particular [3, 4, 5] and the review paper [6]. Connections and differences with previous work are indicated. Numerical results illustrate that asymptotic efficiency is not necessarily in conflict with robustness.

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 $\theta \in \Theta \subset I\!R^p$, (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 the design variable x. Maximum Likelihood (ML) estimation can be used when f is known, 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. When the density f is only known to be symmetric about zero, the model can then be termed *semi-parametric*, with θ and f respectively its parametric and non-parametric parts: f can be considered as an infinite-dimensional nuisance parameter for the estimation of θ, see [3]. This nuisance parameter generally induces a loss of efficiency, and an estimator that remains asymptotically efficient in these conditions is called *adaptive*, see [5, 7]. Beran [8] and Stone [4] 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 estimation of f from residuals obtained with a preliminary $\sqrt{n}$-consistent estimator. This second approach has been further developed by Bickel [5], see also [6].

The method we suggest consists in minimizing the entropy of a kernel estimate of f based on symmetrized residuals, see [1, 2]. The motivation and the connections and differences with the Stone-Bickel approach will be presented in Section 2. We can already mention that an advantage of the minimum-entropy approach is its flexibility: different methods can be used to estimate the entropy of f, each one corresponding to a different method for estimating θ. Also, the quest for an adaptive estimator can be decomposed into a series of steps, which, in the case of minimum-entropy estimators, are similar to those one encounters for LS or ML estimation. The simplest case of a location problem is considered in Section 3 and the extension to nonlinear regression in Section 4. Section 5 illustrates the robustness properties of the estimator through an example.

2. MINIMUM ENTROPY ESTIMATION

Define $e_i(\theta) = Y_i - \eta(\theta, x_i)$, the ith residual in the regression model (1). When f is known, the ML estimator $\hat{\theta}^n_{ML}$ minimizes

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

with respect to θ. We may then observe 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 distribution of the symmetrized residuals is minimum at $\theta = \bar{\theta}$. This was the motivation in [1, 2] for minimizing an estimate of the entropy of the distribution of the residuals in the case where f is unknown (in fact, since the entropy is shift-invariant, we need to use the symmetrized residuals $[e_i(\theta), -e_i(\theta)]$). Our construction is as follows: we construct a kernel estimate $\hat{f}^\theta_n$ from the symmetrized residuals (which ensures that $\hat{f}^\theta_n$ is symmetric); we compute its

entropy, which forms the estimation criterion $\hat{H}_n(\theta)$ to be minimized. In [1, 2], $\hat{H}_n(\theta)$ was given by

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

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

Consider the special case of the location model. We can then follow the same lines as Beran in [9], although his approach is based on the Hellinger distance and ours relies on the Kullback-Leibler divergence. 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 an estimate $\hat{g}_n$ of g is known, e.g. a kernel estimate based on the observations $Y_1, \ldots, Y_n$. Straightforward calculation shows that the symmetric $\hat{f}$ and parameter $\hat{\theta}$ that minimize the Kullback-Leibler divergence

$$L(\hat{g}_n, g_{\hat{\beta}}) = \int \log[\hat{g}_n(y)/g_{\hat{\beta}}(y)]\hat{g}_n(y)dy$$

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 $\hat{\theta}^n = \arg\min_\theta H(\hat{f}_n^\theta)$. Therefore, *$\hat{\theta}^n$ minimizes the entropy of a kernel estimate $\hat{f}_n^\theta$ based on the symmetrized residuals $Y_i - \theta, -Y_i + \theta$.*

This approach clearly has some similarities with the Stone-Bickel approach [4, 5]. They estimate θ in two stages: first, an asymptotically locally sufficient (in the sense of Le Cam) estimator $\hat{\theta}_1^n$ is constructed; second, an approximated score function, the derivative of $\bar{H}_n(\theta)$ with respect to θ, with $\bar{H}_n(\theta)$ given by (2), is used to perform a Newton-Raphson step from $\hat{\theta}_1^n$. The adaptivity of the method for the location model with symmetric errors has been proved first by Stone [4], and Bickel [5] and Manski [6] have extended the result to other models, including non linear regression. Although the construction of $\hat{H}_n(\theta)$ may rely on a similar kernel estimate, we can mention some important differences between the Stone-Bickel approach and the minimum-entropy method presented here. First, the estimation criterion $\hat{H}_n$, (3) or (5), 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 [5, 6] 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 expect sample splitting to degrade the performance of the estimator, which is confirmed by the results presented in [6]. The estimator proposed by Andrews [10] has the advantage of not relying on sample splitting. Also, it is defined by the minimization of an estimation criterion and not by a single Newton-Raphson step. However, his approach is different from the one suggested here: his criterion (likelihood function) is 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 opposite, $\hat{H}_n$ does not depend on any preliminary estimate.

As already mentioned in introduction, an important 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, if we use kernel estimation this is only one possibility. One may refer to [11] for a survey which includes plug-in, sample spacing and nearest neighbor methods.

The asymptotic properties of the minimum entropy estimator $\hat{\theta}^n = \arg\min_{\theta\in\Theta} \hat{H}_n(\theta)$ can be uncovered by following the three steps indicated below. Here we assume that Θ is an open subset of $\mathbb{R}^p$ and $\hat{H}_n(\theta)$ some estimate of the entropy of the distribution of the symmetrized residuals in model (1). We also assume that $\bar{\theta} \in \Theta$, that Θ is locally convex at $\bar{\theta}$ and that $\hat{H}_n(\theta)$ is two times continuously differentiable with respect to $\theta \in \Theta$. Convergence in probability when $n \to \infty$ is denoted $\xrightarrow{\mathrm{p}}$ ($\overset{\theta,\mathrm{p}}{\rightsquigarrow}$ when the convergence is uniform with respect to θ), and convergence in distribution is 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.

Leaving aside some usual measurability conditions, see, e.g., Lemmas 1,2 and 3 of [12], the main steps are as follows:

(i) 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 $\theta \neq \bar{\theta}$;

(ii) 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 ($\succ 0$);

(iii) 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 $n \to \infty$.

(i) proves that $\hat{\theta}^n \xrightarrow{\mathrm{p}} \bar{\theta}$, (i) and (ii) 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 [12] 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 $\alpha_n \in [0, 1]$. (iii) 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 would result from $\mathbf{M}_2^{-1}\mathbf{M}_1\mathbf{M}_2^{-1} = \mathbf{M}_F^{-1}$, the inverse of the Fisher information matrix for the model (1). Step (iii) allows some freedom in the choice of the function $\bar{H}_n(\theta)$. For

instance, under standard assumptions asymptotic normality of $\sqrt{n}\nabla\bar{H}_n(\bar{\theta})$ holds for (2), with $\mathbf{M}_1 = \mathbf{M}_F$.

One may notice that (uniform) $\sqrt{n}$-consistency of the entropy estimate $\hat{H}_n(\theta)$ is not a prerequisite for $\sqrt{n}$-consistency of $\hat{\theta}^n$.

3. LOCATION MODEL

We shall use kernel estimates given by

$$k^{\theta}_{n,i}(u) = \frac{1}{(n-1)h_n} \sum_{j=1, j\neq i}^{n} K\left[\frac{u - e_j(\theta)}{h_n}\right], \quad (4)$$

with $e_i(\theta)$ the residuals, $e_i(\theta) = Y_i - \theta$, $i = 1, \ldots, n$. Here 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 [13] (these conditions are satisfied e.g. by the density of the standard normal). 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

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

where $f^{\theta}_{n,i}(u) = (1/2)[k^{\theta}_{n,i}(u) + k^{\theta}_{n,i}(-u)]$ and $U_n(u)$ is a function equal to zero for large values of u. We take $U_n(x) = u(|x|/A_n - 1)$, with $u(z) = 1$ for $z \leq 0$, 0 for $z \geq 1$ and $u(z)$ varying smoothly between 0 and 1, $u'(0) = u'(1) = 0$, and $\max_z |u'(z)| = d_1 < \infty$, $\max_z |u''(z)| = d_2 < \infty$. $\hat{H}_n(\theta)$ is then two times continuously differentiable in θ for any n. As in [14], we assume there exists a function $B(x)$ such that for all x, $B(x) \geq \sup_{|y|\leq x} 1/f(y)$ ($B(x)$ can be assumed to be strictly increasing without any loss of generality). Define $B_n = B(2A_n + L)$. Using [14], Theorem 4, and [15], 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 $\theta = \bar{\theta}$ (see also [2]). This proves point (i) of Section 2, and thus the consistency of $\hat{\theta}^n$ that minimizes $\hat{H}_n$. Similarly, with slightly stronger conditions on f one can prove (ii), that is, $\nabla^2\hat{H}_n(\theta) \overset{\theta;\mathrm{p}}{\rightsquigarrow} \nabla^2 H(\theta)$, with $\nabla^2 H(\bar{\theta}) = i(f)$, when $B_n = n^{\alpha}$, $h_n = 1/[n^{\alpha}\log n]$, $\alpha < 1/7$. The adaptivity of $\hat{\theta}^n$, i.e. step (iii), would then follow from

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

One may notice that for checking this condition a difficulty which is not present in the Stone-Bickel approach is due to the fact that $(k^{\bar{\theta}}_{n,i})'(-x) \neq -(k^{\bar{\theta}}_{n,i})'(x)$.

4. NONLINEAR REGRESSION

Consider now a nonlinear regression model (1), for which the experiment consists of repetitions at fixed points $X^1, \ldots, X^m$. If ξ^j denotes the weight of point X^j, $n_j = n\xi^j$ of the n observations are made at $X = X^j$.

Using a justification similar to that given in Section 2 for the location model, we arrive at 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

$$\hat{\theta}^n = \arg\min_{\theta\in\Theta} \mathtt{E}_{\xi}\{\hat{H}_n(\theta, X)\}$$
$$\text{with} \quad \hat{H}_n(\theta, X^j) = H(\hat{f}^{j,\theta}), j = 1, \ldots, m . \quad (6)$$

The adaptivity of this method would follow from adaptivity in the location model. However, this approach does not extend to more general experiments since estimating the entropy of the symmetrized residuals *at each* X is not possible without repetitions of observations. Hence the approach used in [1, 2] that consists in mixing all (symmetrized) residuals together and estimating the entropy $\hat{H}_n(\theta)$ of their distribution, see (3). Replace ξ by ξ_n in (6), where ξ_n is an empirical measure of the points X_i. Let U be a random variable with distribution conditional to X given by $\hat{f}^{j,\theta}$ of (6). Then, $\mathtt{E}_{\xi_n}\{\hat{H}_n(\theta, X)\} = \mathcal{H}(U|X)$ is the conditional entropy of U given X, and $\mathcal{H}(U|X) \leq \mathcal{H}(U) = \hat{H}_n(\theta)$. The entropy $\hat{H}_n(\theta)$ obtained by mixing up all residuals is thus an upper bound on the criterion that is minimized in (6). Studying the adaptivity of the corresponding estimator will form the subject of further work. Some preliminary numerical results are presented below.

5. EXAMPLE

Consider the regression model $\eta(\theta, x) = \theta_1 \exp(-\theta_2 x)$ with $\bar{\theta} = (100, 2)^{\top}$, the design points are $X^j = 1 + (j-1)/9$ $j = 1, \ldots, 10$, with $\xi^j = 1/10$ for all j. We take $U_n(x) \equiv 1$ in (5); the kernel bandwidth h is chosen by the double kernel method [16] and based on residuals obtained from a robust M-estimator. Table 1 gives the trace and determinant of the empirical covariance matrix $\hat{C}_n$ of $\sqrt{n}(\hat{\theta} - \bar{\theta})$, obtained from 100 repetitions of the estimation procedure, for different choices of the estimator $\hat{\theta}$ and distribution f, using $n = 100$ observations. The Minimum Entropy estimator ME is based on (5), using gaussian kernels with residuals symmetrized beforehand, that is, $k^{\theta}_{2n,i}(u)$ given by (4) with $2n$ residuals $e^s_{2n} = [e_n; -e_n]$. The Minimum Hellinger Distance estimator MHD [9] and ME mix all residuals.

Table 1. Values T_n, D_n of the trace and determinant of the empirical covariance matrix $\hat{C}_n$ with f the 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}
T_n	LS	7498	8304	6968	8299	7242
	MHD	9172	3381	6411	8001	7106
	ME	9267	3405	4620	7753	6786
	$Tr(M_F^{-1})$	6171	3086	3086	4937	5834
D_n	LS	85.6	204.7	128.4	166.8	145.1
	MHD	124.0	39.1	56.7	121.1	158.4
	ME	116.2	34.4	36.4	105.3	120.2
	$\det(M_F^{-1})$	84.3	21.1	21.1	53.9	75.3

Table 2. Trace and det of empirical MSE matrix with q outliers in addition to n observations, f the bi-exponential

	q	0	20	40	60	80
$Tr(MSE_n)$	LS	13607	35039	45150	128408	20347
	MHD	3844	5326	35550	17870	2864
	ME	3868	4771	5858	17866	2851
$\det(MSE_n)$	LS	356.7	418.0	953.6	729.2	1244.9
	MHD	47.3	143.9	9009.6	316.0	246.0
	ME	42.1	101.7	209.1	192.7	132.4

We now add q outliers ($\mathcal{N}(2, 10)$, randomly allocated among the ε_j's, $j = 1, \ldots, n+q$) to the n observations obtained for f the bi-exponential; MSE_n is computed for $n = 100$. Table 2 shows these outliers have little influence over the ME estimator. This result would deserve further studies, following those for the MHD estimator, see respectively [17, 18] and [9] for the parametric and semiparametric cases.

6. REFERENCES

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

[2] 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 2001.

[3] C. Stein, "Efficient nonparametric testing and estimation," in *Proc. 3rd Berkeley Symp. Math. Stat. Prob.*, Berkeley, 1956, vol. 1, pp. 187–196, University of California Press.

[4] C.J. Stone, "Adaptive maximum likelihood estimators of a location parameter," *Annals of Statistics*, vol. 3, no. 2, pp. 267–284, 1975.

[5] P.J. Bickel, "On adaptive estimation," *Annals of Statistics*, vol. 10, pp. 647–671, 1982.

[6] C.F. Manski, "Adaptive estimation of nonlinear regression models," *Econometric Reviews*, vol. 3, no. 2, pp. 145–194, 1984.

[7] 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*, vol. 11, no. 2, pp. 432–452, 1983.

[8] R. Beran, "Asymptotically efficient rank estimates in location models," *Annals of Statistics*, vol. 2, pp. 63–74, 1974.

[9] R. Beran, "An efficient and robust adaptive estimator of location," *Annals of Statistics*, vol. 6, no. 2, pp. 292–313, 1978.

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

[11] J. Beirlant, E.J. Dudewicz, L. Györfi, and E.C. van der Meulen, "Nonparametric entropy estimation; an overview," *Intern. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[12] R.L. Jennrich, "Asymptotic properties of nonlinear least squares estimation," *Annals of Math. Stat.*, vol. 40, pp. 633–643, 1969.

[13] E.F. Schuster, "Estimation of a probability density function and its derivatives," *Annals of Math. Stat.*, vol. 40, pp. 1187–1195, 1969.

[14] 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.

[15] W.K. Newey, "Uniform convergence in probability and stochastic equicontinuity," *Econometrica*, vol. 9, no. 4, pp. 1161–1167, 1991.

[16] A. Berlinet and L. Devroye, "A comparison of kernel density estimates," *Publications de l'institut de statistique de l'Universit de Paris*, vol. 38, pp. 3–59, 1994.

[17] R. Beran, "Minimum Hellinger distance estimates for parametric models," *Annals of Statistics*, vol. 5, no. 3, pp. 445–463, 1977.

[18] B.G. Lindsay, "Efficiency versus robustness: the case for minimum Hellinger distance and related methods," *Annals of Statistics*, vol. 22, no. 2, pp. 1081–1114, 1994.