APPLICATIONES MATHEMATICAE
22,4 (1995), pp. 515–529

W. NIEMIRO (Warszawa)

ESTIMATION OF NUISANCE PARAMETERS FOR INFERENCE BASED ON LEAST ABSOLUTE DEVIATIONS

Abstract. Statistical inference procedures based on least absolute deviations involve estimates of a matrix which plays the role of a multivariate nuisance parameter. To estimate this matrix, we use kernel smoothing. We show consistency and obtain bounds on the rate of convergence.

1. Introduction. Statistical inference procedures considered in this paper are related to M-functionals and M-estimators of *least absolute deviations* (LAD) type. Let (Y, X) be a random vector in $\mathbb{R} \times \mathbb{R}^d$. Fix $\alpha \in [0, 1]$ and define functions $\varrho, \psi : \mathbb{R} \to \mathbb{R}$ by

$$\varrho(s) = \tfrac{1}{2}|s| + \big(\alpha - \tfrac{1}{2}\big)s, \qquad \psi(s) = \tfrac{1}{2}\mathrm{sign}(s) + \alpha - \tfrac{1}{2}. \tag{1}$$

Of course, ψ is the derivative (more precisely, a subderivative) of ϱ. Let $Q : \mathbb{R}^d \to \mathbb{R}$ and $G : \mathbb{R}^d \to \mathbb{R}^d$ be given by

$$Q(t) = \mathbf{E}\varrho(Y - t^T X), \qquad G(t) = \mathbf{E}\psi(Y - t^T X)X. \tag{2}$$

Under mild assumptions, these functions are well-defined, $G = \nabla Q$ and there exists a unique $t_0 \in \mathbb{R}^d$ such that $G(t_0) = 0$ and $Q(t_0) = \min_t Q(t)$. Thus, t_0 is the value of an M-functional on the joint distribution of (Y, X). If $(Y_1, X_1), \ldots, (Y_n, X_n)$ is an i.i.d. sample from this distribution, put

$$Q_n(t) = \frac{1}{n}\sum_{i=1}^{n} \varrho(Y_i - t^T X_i), \qquad G_n(t) = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i - t^T X_i)X_i. \tag{3}$$

A random point t_n such that $Q_n(t_n) = \min_t Q_n(t)$ can be considered as

1991 *Mathematics Subject Classification*: Primary 62G07; Secondary 62F05, 62F12.

Key words and phrases: least absolute deviations, kernel density/regression estimation.

The research supported by KBN grant no. 80492/91/01.

an M-estimate. Note, in passing, that $G_n(t_n) \neq 0$ in general, because Q_n is not differentiable at t_n.

M-estimators corresponding to the ϱ-function given by (1) appear in regression analysis and discriminant analysis. For linear models, they are estimators of *regression quantiles*, introduced by Koenker and Basset (1978). For $\alpha = 1/2$ we obtain just LAD estimator. Two-class discrimination can be treated in much the same way as regression. For instance, we can assign by convention $Y_i = 1$ to observations X_i from the first class and $Y_i = -1$ to X_i from the second class. If we adopt a slightly different convention and set $\alpha = 1$ in (1), we obtain the function Q_n known in discriminant analysis as *perceptron criterion* (Hand 1981, Niemiro 1987, 1989; see also Section 3.2 for details). Asymptotic properties of LAD-type M-estimators and inference procedures related to them were investigated by many authors. Let us mention the monograph of Bloomfield and Steiger (1983), Rao (1988), Pollard (1991) and Niemiro (1992, 1993). Jurečková in a series of papers considered M-statistics in linear models with various choices of ϱ- (or ψ)-function, including LAD as a particular case (e.g. 1989, Jurečková and Sen, 1987, 1989). The best references for recent advances in this field are the proceedings of the two "L_1-Norm" conferences in Neuchâtel, edited by Dodge (1987, 1992).

Niemiro (1993) developed statistical inference procedures based on M-estimators with a general convex ϱ-function. The object of inference is the M-functional t_0. Suppose we are to build an approximate confidence region for t_0 or to test a hypothesis of the form $Ht_0 = c$, where H is a $p \times d$ matrix and $c \in \mathbb{R}^p$. The following asymptotic results help here. Under some regularity conditions,

$$n^{1/2}(t_n - t_0) \to_d N(0, D^{-1} V D^{-1}), \tag{4}$$

where

$$D = \nabla^2 Q(t_0) = \nabla G(t_0), \qquad V = \mathbf{E}\psi^2(Y - t_0 X) X X^T. \tag{5}$$

Therefore

$$n(t_n - t_0)^T D V^{-1} D (t_n - t_0) \to_d \chi^2(d). \tag{6}$$

Quite similarly, under the null hypothesis $Ht_0 = c$,

$$n(Ht_n - c)^T (H D^{-1} V D^{-1} H^T)^{-1} (Ht_n - c) \to_d \chi^2(p), \tag{7}$$

if H is of full rank p. If $\dot{t}_n$ is the constrained M-estimate, that is, $H\dot{t}_n = c$ and $Q_n(\dot{t}_n) = \min_{Ht=c} Q_n(t)$, then

$$n G_n(\dot{t}_n)^T D^{-1} H^T (H D^{-1} V D^{-1} H^T)^{-1} H D^{-1} G_n(\dot{t}_n) \to_d \chi^2(p). \tag{8}$$

These results are simple consequences of asymptotic representations of Bahadur–Ghosh type for the underlying M-estimators (Niemiro 1992). Of course, the matrices D and V are usually unknown. We need estimates of these matrices to construct an approximate confidence ellipsoid or to construct tests of approximately prescribed size. If $\widehat{D}_n$ and $\widehat{V}_n$ are consistent estimates, they can be substituted into (6)–(8) to obtain *statistics* with asymptotic χ^2 distribution. The obvious estimator $\widehat{V}_n = n^{-1}\sum\psi(Y_i - t_n^T X_i)X_i X_i^T$ is consistent. Estimation of D is much harder, due to the fact that ϱ is not differentiable. The function Q_n is piecewise linear and we cannot use its second derivative to estimate the second derivative of Q. Some smoothing technique resembling non-parametric density or regression estimation is necessary. The objective of this paper is to propose an estimator of D and to examine its properties.

Note that the classical assumptions imposed on regression models simplify the asymptotic theory, because they force matrices D and V to be proportional. Suppose

$$Y = t_0^T X + U, \tag{9}$$

where U is independent of X and $\mathbf{P}(U \le 0) = \alpha$. Then it is not hard to see that $V = \mathbf{E}XX^T$ and $D = f(0)\alpha^{-1}(1-\alpha)^{-1}V$, where f is the density of U (assumed to be continuous at 0). Formulae (6)–(8) assume simpler form and estimation of D reduces to estimation of the scalar quantity $f(0)$. Usual kernel estimators can do the job. Methods of this kind were proposed by Koenker (1987), McKean and Schrader (1987), Schrader and McKean (1987), who developed LAD-type tests. Welsh (1987), Babu (1986) and many others considered estimation of $f(0)$ or $f(0)^{-1}$ in this context. Unfortunately, this approach breaks down if the model assumptions are violated. Condition (9) is often not realistic. For instance, imagine that $Y = r(X) + U$, where r is a “slightly non-linear” function. Although model (9) fails, the M-functional t_0 suggested by the wrong model still makes sense and may be quite informative. Another example where (9) clearly fails is the two-class discrimination model mentioned above. Therefore, we should be interested in deriving inference procedures independent of the simplifying assumption (9). To this end, we must cope with estimation of the matrix D.

2. Consistency.

Let us make the following assumptions.

C1. The function Q is well defined by (2), it is twice differentiable at t_0.
C2. The matrix $D = \nabla^2 Q(t_0)$ is positive definite.

We will also need a mild assumption about continuity of the joint distribution of (Y, X) and a strong moment condition on X:

C3. $\mathbf{P}(Y - t^T X = s) = 0$ for t in a neighbourhood of t_0 and all s.
C4. $\mathbf{E}|X|^4 < \infty$.

Conditions **C1**, **C2** and $\mathbf{E}|X|^2 < \infty$ are sufficient for asymptotic normality (4); see Niemiro (1993). For our purposes, it is enough to remember that

$$t_n \to_p t_0, \qquad t_n = t_0 + O_p(n^{-1/2}).$$

Assumption **C1** is in fact redundant and it was introduced merely for clarity. To avoid problems with infinite $\mathbf{E}|Y|$, we could have used a simple trick and replaced (2) by $Q(t) = \mathbf{E}\varrho(Y - t^T X) - \varrho(Y)$. Two-fold differentiability of Q will always follow from other assumptions of our theorems. However, let us retain **C1** just to reassure ourselves that we know what we are trying to estimate! In this section, we treat **C1**–**C4** as standing assumptions. Consider a kernel function $K : \mathbb{R} \to \mathbb{R}$ such that

K. $0 \le K(s)$, $\int K(s)\,ds = 1$, $K(s)$ is increasing and right continuous for $s \le 0$, decreasing and left continuous for $s \ge 0$.

Note that the kernel is bounded, $K(s) \le K(0)$. Let h_n be a sequence of positive reals, $h_n \to 0$. Put

$$\widehat{D}_n(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\biggl(\frac{Y_i - t^T X_i}{h_n}\biggr) X_i X_i^T, \tag{10}$$

for $t \in \mathbb{R}^d$. Our estimator of D will be $\widehat{D}_n(t_n)$. Write $\widetilde{D}_n(t)$ for $\mathbf{E}\widehat{D}_n(t)$, so

$$\widetilde{D}_n(t) = \frac{1}{h_n} \mathbf{E} K\biggl(\frac{Y - t^T X}{h_n}\biggr) X X^T. \tag{11}$$

To prove that $\widehat{D}_n(t_n) \to_p D$, it is enough to show that for some $\varepsilon, \eta > 0$,

I. $\sup_{|t-t_0|\le\varepsilon} |\widehat{D}_n(t) - \widetilde{D}_n(t)| \to_p 0$,

II. $\sup_{|t-t_0|\le\eta} |\widetilde{D}_n(t) - D(t)| \to 0$ for some $D(t)$ and

III. $D(t) \to D$ as $t \to t_0$.

Let us begin with statements **II** and **III**, that is, with the "deterministic term" analysis.

LEMMA 1. *If* **II** *is true, then* $D(t) = \nabla G(t) = \nabla^2 Q(t)$ *and* $D(t)$ *is continuous for* $|t - t_0| < \eta$. *In particular,* **III** *holds.*

Proof. Each $\widetilde{D}_n(t)$ is a continuous function, by dominated convergence in view of the **C3**, **C4** and **K**. Hence $D(t)$ is continuous as the limit of the locally uniformly convergent sequence $\widetilde{D}_n(t)$.

Let $\delta_n(s) = h_n^{-1} K(h_n^{-1} s)$, $\sigma_n(s) = \int_0^s \delta_n(u)\,du$, $\nu_n(s) = \int_0^s \sigma_n(u)\,du$. Put $\varrho_n(s) = \nu_n(s) + (\alpha - 1/2)s$ and $\psi_n(s) = \sigma_n(s) + \alpha - 1/2$. Of course, $\varrho_n(s) \to \varrho(s)$ and $\psi_n(s) \to \psi(s)$, because $h_n \to 0$. By assumption,

$$\mathbf{E}\delta_n(Y - t^T X) X X^T = \widetilde{D}_n(t) \to D(t) \tag{12}$$

uniformly in t, for $|t - t_0| \le \eta$. We have $\mathbf{E}\delta_n(Y - t^T X) X X^T = \nabla \mathbf{E}\psi_n(Y - t^T X)X$, because differentiation under the expectation sign can be justified in a standard way (Th. 115 in Schwartz 1967). Dominated convergence gives

$$\mathbf{E}\psi_n(Y - t^T X)X \to \mathbf{E}\psi(Y - t^T X)X. \tag{13}$$

Uniform convergence of derivatives (12) allows us to strengthen pointwise convergence in (13) to almost uniform convergence. Moreover, we obtain $D(t) = \nabla\mathbf{E}\psi(Y - t^T X)X$ (Th. 111 in Schwartz 1967). Now, we can repeat the same reasoning once more. Notice that $\mathbf{E}\psi_n(Y - t^T X)X = \nabla\mathbf{E}\varrho_n(Y - t^T X)$ and

$$\mathbf{E}\varrho_n(Y - t^T X) \to \mathbf{E}\varrho(Y - t^T X), \tag{14}$$

and hence uniform convergence in (13) implies that $\mathbf{E}\psi(Y - t^T X)X = \nabla\mathbf{E}\varrho(Y - t^T X) = \nabla Q(t) = G(t)$. Putting things together, we get $D(t) = \nabla G(t) = \nabla^2 Q(t)$. ■

Let us focus on **I** now. It is a statement about uniform convergence of empirical means to expectations. The powerful modern empirical processes theory makes verification of **I** easy. Specifically, we will use maximal inequalities for manageable classes of functions, due to Pollard (1989). We need not invoke the notion of manageability here, because we are going to apply the inequalities only to subclasses of a Vapnik–Cervonenkis (VC) subgraph class of functions. Suppose $\mathcal{F}$ is a class of functions $f : \mathcal{Z} \to \mathbb{R}$. We say $\mathcal{F}$ is a *VC subgraph class* if the sets $\{(z, u) : 0 \le u \le f(z) \text{ or } f(z) \le u \le 0\}$, for all $f \in \mathcal{F}$, form a VC class of subsets of $\mathcal{Z} \times \mathbb{R}$. Assume $|f(z)| \le F(z)$ for all $f \in \mathcal{F}$, that is, F is an envelope of $\mathcal{F}$. Remark 1 on p. 200 in Kim and Pollard (1990) asserts that subclasses of a VC subgraph class are uniformly manageable for F. Consequently, the following fact is a corollary of Theorems 4.2 and 4.4 of Pollard (1989). Suppose $\mathcal{Z}$ is a measurable space and let $Z, Z_1, \ldots, Z_n, \ldots$ be a sequence of i.i.d. random elements of $\mathcal{Z}$.

Assume $\mathcal{F}$ is a VC subgraph class with envelope F such that $\mathbf{E}F(Z)^2 < \infty$ and $\mathcal{F}_n$ are subclasses of $\mathcal{F}$ containing the zero function. If

$$\sup_{f\in\mathcal{F}_n} \mathbf{E}|f(Z)| \to 0, \qquad n \to \infty, \tag{15}$$

then

$$\mathbf{E} \sup_{f\in\mathcal{F}_n} n\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbf{E}f(Z)\right|^2 \to 0, \qquad n \to \infty. \tag{16}$$

Moreover, if F_n are envelopes of $\mathcal{F}_n$, then there is a constant $C < \infty$ such that

$$\mathbf{E} \sup_{f\in\mathcal{F}_n} n\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbf{E}f(Z)\right|^2 \le C\mathbf{E}F_n(Z)^2. \tag{17}$$

To be precise, some conditions are needed to ensure measurability of the suprema above. In our application no measurability problems arise, since the classes of functions to be considered will clearly be permissible, in the sense defined by Pollard (1984, Appendix C). We are going to apply Pollard's theorem to

$$\mathcal{F} = \left\{ f_{t,h}(y,x) = x^j x^m K\left(\frac{y - t^T x}{h}\right) : t \in \mathbb{R}^d,\ h > 0 \right\} \cup \{0\}. \tag{18}$$

Here $(y,x) \in \mathbb{R} \times \mathbb{R}^d$ plays the role of z, and x^j and x^m denote two (fixed) components of x.

LEMMA 2. *The class $\mathcal{F}$ defined by* (18) *has the VC subgraph property.*

Proof. We are to show that subsets of $\mathbb{R}^{d+1} \times \mathbb{R}$ of the form

$$\{(y,x,u) : 0 \le u \le x^j x^m K(h^{-1}(y - t^T x)) \text{ or } x^j x^m K(h^{-1}(y - t^T x)) \le u \le 0\}$$

with $t \in \mathbb{R}^d$ and $h > 0$ belong to a VC class of sets. It is enough to check that the intersections of these sets with the three constant sets $\{(y,x,u) : x^j x^m > 0, u > 0\}$, $\{x^j x^m > 0, u > 0\}$ and $\{u = 0\}$ form VC classes. Consider only the first class of intersections, since the second one is quite similar and the third one is trivial. For $0 < u \le K(0)$ write $K^+(u) = \sup\{s \ge 0 : K(s) \ge u\}$ and $K^-(u) = \inf\{s \le 0 : K(s) \ge u\}$. Put $K^+(u) = -1$ and $K^-(u) = 1$ for $u > K(0)$. We have $0 < u \le K(s)$ iff $K^-(u) \le s \le K^+(u)$. Now, we can write the intersections under consideration as

$$\begin{aligned}&\{(y,x,u) : x^j x^m > 0, u > 0\}\\ &\qquad \cap \{(y,x,u) : hK^-(u(x^j x^m)^{-1}) \le y - t^T x \le hK^+(u(x^j x^m)^{-1})\}.\end{aligned}$$

Each such set is the result of union and intersection operations applied to two sets of the form $\{g(y,x,u) \ge 0\}$, where g runs through a finite-dimensional vector space of functions and a third, constant set. Consequently, the class of such sets is VC. ■

THEOREM 1. *Under assumptions* **C1**, **C2**, **C3**, **C4** *and* **K**, *if* $h_n = o(1)$ *and* $h_n^{-1} = O(n^{1/2})$ *then* $\widehat{D}_n(t_n) \to_p D$, *provided that* **II** *holds.*

Proof. In view of **C4**, we can use $F(y,x) = K(0)|x^j x^m|$ as a square integrable envelope of $\mathcal{F}$. It is easy to see that our condition **II** implies, for $\varepsilon < \eta$,

$$\sup_{|t - t_0| \le \varepsilon} \mathbf{E}|X^j X^m| K\left(\frac{Y - t^T X}{h_n}\right) \to 0. \tag{19}$$

In fact, $|X^j X^m| \le |X^j|^2 + |X^m|^2$, while $\mathbf{E}|X^j|^2 K(h_n^{-1}(Y - t^T X))$ is equal to h_n times a diagonal element of $\widetilde{D}_n(t)$. Thus, formula (19) is tantamount to (15) for subclasses $\mathcal{F}_n = \{f_{t,h_n} : |t - t_0| \le \varepsilon\} \cup \{0\}$. Therefore, (16) holds for these classes $\mathcal{F}_n$, that is,

$$\mathbf{E} \sup_{|t - t_0| \le \varepsilon} n h_n^2 |\widehat{D}_n(t) - \widetilde{D}_n(t)|^2 \to 0. \tag{20}$$

Since $n^{-1}h_n^{-2} = O(1)$, we get **I**. Lemma 1 ensures **III**. Of course, we have applied Pollard's theorem componentwise and for the matrix norm in (20) we can take the usual norm in d^2-dimensional euclidean space. ■

The condition $h_n^{-1} = O(n^{1/2})$ is certainly not necessary for consistency, but we think it is quite satisfactory for reasonable applications. Let us recall that our objective is to estimate $D = D(t_0)$; for $h_n = o(n^{-1/2})$ we would try to reduce the bias of an estimate of $D(t_n)$, where $t_n - t_0$ is of order $n^{-1/2}$. Note that we used assumption **C3** only in the proof of Lemma 1 to justify continuity of $\widetilde{D}_n(t)$. If the kernel K is continuous, this condition becomes unnecessary.

Convergence in probability can be strengthened to L^2 convergence, if we know that

IV. $\sup_n \sup_{t\in\mathbb{R}^d} |\widetilde{D}_n(t)| < \infty$.

Corollary. *Under the assumptions of Theorem* 1, $\widehat{D}_n(t_n) \to_{L^2} D$, *provided that* **II** *and* **IV** *hold.*

Proof. Application of (17) to $\mathcal{F}_n = \mathcal{F}$ yields

$$\mathbf{E}\sup_{t\in\mathbb{R}^d} nh_n^2|\widehat{D}_n(t) - \widetilde{D}_n(t)|^2 \le C\mathbf{E}|X|^4.$$

If we combine this with **IV**, we obtain an integrable random variable which dominates $\sup_{t\in\mathbb{R}^d} |\widehat{D}_n(t)|^2$. ■

Finally, we are left with the task of verifying **II** and **IV**. Explicit assumptions which imply these statements will be given in the forthcoming section.

3. Regularity conditions and rates of convergence. To ensure consistency of our estimator $\widetilde{D}_n(t_n)$ and to obtain bounds on the rate of convergence, we have to assume some regularity properties of the joint probability distribution of (Y, Z). The form of plausible regularity conditions depends on the way we look at this random vector: we may prefer either to consider the conditional distributions of Y given X, or the other way round. Thus, we have to examine separately two cases.

3.1. *Regression model.* If we regard Y as a function of X plus some random error (not necessarily independent of X), it is natural to assume the following.

R1. For every x, the conditional distribution of Y given $X = x$ has density $f(y|x)$, which is bounded and uniformly continuous as a function of y, uniformly in x.

In other words, we assume that $f(x|y) \le M$ for all x, y and to each ε there corresponds δ such that $|y_1 - y_2| < \delta$ implies $|f(y_1|x) - f(y_2|x)| < \varepsilon$ for all x.

R2. The matrix $D = \mathbf{E} f(t_0^T X|X) X X^T$ is positive definite.

Theorem 2. *Under assumptions* **R1**, **R2**, **C4** *and* **K**, *if* $h_n = o(1)$ *and* $h_n^{-1} = O(n^{1/2})$ *then* $\widehat{D}_n(t_n) \to_p D$.

Proof. In view of Theorem 1 and Lemma 1, it is enough to check that **II** holds, with

$$D(t) = \mathbf{E} f(t^T X|X) X X^T.$$

We have

$$\begin{aligned}
&|\widetilde{D}_n(t) - D(t)| \\
&\quad \le \mathbf{E} \int \frac{1}{h_n} K\biggl(\frac{y - t^T X}{h_n}\biggr) |f(y|X) - f(t^T X|X)|\, dy\, X X^T \\
&\quad \le \mathbf{E} \biggl[\varepsilon \int\limits_{|y - t^T X| < \delta} K\biggl(\frac{y - t^T X}{h_n}\biggr) dy + M \int\limits_{|y - t^T X| > \delta} K\biggl(\frac{y - t^T X}{h_n}\biggr) dy\biggr] |X|^2 \\
&\quad \le 2\varepsilon \mathbf{E}|X|^2,
\end{aligned}$$

for large n, if $|f(y|x)| \le M$ and δ corresponds to ε in the definition of uniform continuity of $f(y|x)$. ■

Corollary. *Under the assumptions of Theorem* 2, *we have* $\widehat{D}_n(t_n) \to_{L^2} D$.

To see this, note that $D(t) \le M \mathbf{E}|X|^2$ and $\widetilde{D}_n(t) \to D(t)$ uniformly. Apply the Corollary to Theorem 1.

Now, consider a stronger set of assumptions:

R3. For every x, the conditional distribution of Y given $X = x$ has density $f(y|x)$ with two uniformly bounded derivatives $\frac{d}{dy} f(y|x)$ and $\frac{d^2}{dy^2} f(y|x)$.

C5. $\mathbf{E}|X|^5 < \infty$.

KS. The kernel is symmetric: $K(-s) = K(s)$.

Theorem 3. *Under assumptions* **R2**, **R3**, **C5**, **K** *and* **KS**, *if* $h_n = o(1)$ *and* $h_n^{-1} = O(n^{1/2})$ *then* $\widehat{D}_n(t_n) = D + O_p(n^{-1/2} h_n^{-1/2} \vee h_n^2)$.

Proof. The two entries in $O_p(\cdot)$ correspond to bounds on "variance" and "bias" terms, respectively. Let us begin with the bound on $|\widehat{D}_n(t) - \widetilde{D}_n(t)|$. Since our present assumptions imply **C1** and **C2**, asymptotic normality (4) obtains, $|t_n - t_0| = O_p(n^{-1/2})$ and so we need a bound which is uniform in

t for $|t-t_0| = O(n^{-1/2})$. The maximal inequality (17) gives

$$\mathbf{E} \sup_{|t-t_0|\le Ln^{-1/2}} nh_n^2 |\widehat{D}_n(t) - \widetilde{D}_n(t)|^2 \le C\mathbf{E} \sup_{|t-t_0|\le Ln^{-1/2}} K^2\Big(\frac{Y - t^T X}{h_n}\Big)|X|^4. \tag{21}$$

The right hand side of (21) can be bounded by

$$C\mathbf{E} \int \sup_{|t-t_0|\le Ln^{-1/2}} MK^2\Big(\frac{y - t^T X}{h_n}\Big)\, dy\, |X|^4.$$

Now, $|y - t^T X| \ge ||y - t_0^T X| - |t-t_0||X||$. If $|y - t_0^T X| \ge \frac{1}{2}Ln^{-1/2}|X|$, then we have $K(h_n^{-1}|y - t^T X|) \le K(h_n^{-1}|y - t_0^T X|/2)$. Using the inequality $K(h_n^{-1}(y - t^T X)) \le K(0)$ for $|y - t_0^T X| \le \frac{1}{2}Ln^{-1/2}|X|$, we obtain

$$\begin{aligned}\int \sup_{|t-t_0|\le Ln^{-1/2}} K^2\Big(\frac{y - t^T X}{h_n}\Big)\, dy &= \int_{|y-t_0^T X|\ge\frac{1}{2}Ln^{-1/2}|X|} + \int_{|y-t_0^T X|\le\frac{1}{2}Ln^{-1/2}|X|} \\ &\le \int K^2\Big(\frac{y - t_0^T X}{2h_n}\Big)\, dy + LK(0)n^{-1/2}|X| \\ &= h_n \int K^2(s)\, ds + LK(0)n^{-1/2}|X|.\end{aligned}$$

Since **C5** is assumed, we can multiply the above inequality by $|X|^4$ and take expectation. It follows that the right hand side of (21) is at most

$$CM\mathbf{E}\Big[h_n \int K^2(s)\, ds + LK(0)n^{-1/2}|X|\Big]|X|^4 = O(h_n) + O(n^{-1/2}).$$

Therefore, (21) yields

$$\mathbf{E} nh_n^2 \sup_{|t-t_0|\le Ln^{-1/2}} |\widehat{D}_n(t) - \widetilde{D}_n(t)|^2 = O(h_n).$$

Consequently,

$$\sup_{|t-t_0|\le Ln^{-1/2}} |\widehat{D}_n(t) - \widetilde{D}_n(t)| = O_p(n^{-1/2}h_n^{-1/2}). \tag{22}$$

Now, let us turn to the bias term. Use the Taylor expansion of $f(y|x)$ around $y = t^T X$ with the second order remainder written in the Lagrange form. We have

$$\sup_t |\widetilde{D}_n(t) - D(t)| = O(h_n^2), \tag{23}$$

because $|\widetilde{D}_n(t) - D(t)|$ is equal to

$$\begin{aligned}\Big|\mathbf{E}\int \frac{1}{h_n}K\Big(\frac{y-t^TX}{h_n}\Big)\Big[&f(t^TX|X)+\frac{d}{dy}f(t^TX|X)(y-t^TX)\\&+\frac{1}{2}\frac{d^2}{dy^2}f(y^*|X)(y-t^TX)^2-f(t^TX|X)\Big]\,dy\,XX^T\Big|\\&\le \mathbf{E}\int \frac{1}{h_n}K\Big(\frac{y-t^TX}{h_n}\Big)N(y-t^TX)^2\,dy\,|X|^2\\&\le Nh_n^2\int u^2K(u)\,du\,\mathbf{E}|X|^2.\end{aligned}$$

In the above formula, y^* is some point between y and t^TX, and the constant N bounds d^2f/dy^2. Of course, the integral of the first order term cancels out due to **KS**.

We have yet to bound $|D(t_n)-D(t_0)|$ in probability. Note that

$$\sup_{|t-t_0|\le Ln^{-1/2}}|D(t)-D(t_0)|=O(n^{-1/2}),\tag{24}$$

because

$$\begin{aligned}|D(t)-D(t_0)|&=\Big|\mathbf{E}\Big[f(t_0^TX|X)+\frac{d}{dy}f(y^*|X)(t-t_0)^TX\Big]XX^T\Big|\\&\le N|t-t_0|\mathbf{E}|X|^3,\end{aligned}$$

by Lagrange's formula (here N bounds df/dy). Formulae (22)–(24) imply the conclusion. ∎

Corollary. *Under the assumptions of Theorem* 3, *the best bound on the rate of convergence obtains for $h_n=hn^{-1/5}$. For such h_n,*

$$\widehat{D}_n(t_n)=D+O_p(n^{-2/5}).$$

It is interesting to compare the rate of convergence asserted above with the magnitude of the remainder term in Bahadur's representation of $n^{1/2}(t_n–t_0)$. Kiefer (1967) proved this remainder is of precise order $O(n^{-1/4})$ in probability in the univariate case, when $X=1$ and t_0 is an α-quantile of Y. For linear models, the analogous result was obtained by Jurečková and Sen (1987; see also 1989). In our setting, results of Niemiro (1992) show the remainder is $O(n^{-1/4}(\log n)^{1/2}(\log\log n)^{1/4})$ almost surely. Let us take the $O(n^{-1/4})$ rate in probability as a very plausible conjecture. Since the representation of t_n as a sum of independent random vectors is used to derive (4)–(6), we can expect the left hand side of (6) is $\chi^2(d)$-distributed up to an $O(n^{-1/4})$ error. The additional error, introduced when we substitute $\widehat{D}_n(t_n)$ in place of D, is of order $O(n^{-2/5})$. The same remark applies to (7) and (8). Needless to say, the whole above discussion is rather crude.

3.2. *Discrimination model.* Let us explain how the function Q, defined by (2) in the Introduction, can be used in two-class discriminant analysis.

Consider a random vector (J,Z) with values in $\{-1,1\}\times\mathbb{R}^d$. The binary random variable J is interpreted as the indicator of the class to which an object belongs. Components of Z describe some measurable or observable features of the object. For various reasons, it is often desirable to look for a linear discriminant function $g(Z)=s+t^TZ$, $s\in\mathbb{R}$, $t\in\mathbb{R}^d$. The goal is to select g such that, roughly speaking, the conditional distributions of $g(Z)$ given $J=1$ and $J=-1$ are well separated from each other. The separability can be quantified in many ways, of course. The following criterion is suggested by analogy with linear regression. Let

$$Q(s,t)=\mathbf{E}\varrho(1-J(s+t^TZ)). \tag{25}$$

This criterion function is of the form considered in the Introduction. To see this, put $Y=1$ and $X^T=J(1,Z^T)$; note that we included explicitly the "interception term" s, thus X is now of dimension $d+1$ and $Q:\mathbb{R}^{d+1}\to\mathbb{R}$. Assume ϱ appearing in (25) is given by (1). For $\varrho(u)=|u|/2$, $Q(s,t)$ is small if $\operatorname{sign}(s+t^TZ)$ is close to J. Some thought shows that the intuitive meaning of Q remains clear also for, say, $\varrho(u)=u\vee 0=(u+|u|)/2$. This is why (25) is more judicious than the (perhaps more familiarly looking) criterion $\mathbf{E}\varrho(J-(s+t^TZ))$.

To state regularity conditions taylored for the model of discrimination, we will need the following quantities:

$$\begin{aligned} f_+(s,t)&=\frac{d}{ds}\mathbf{P}(t^TZ\le s\mid J=1),\\ m_+(s,t)&=\mathbf{E}(Z\mid t^TZ=s,\ J=1),\\ V_+(s,t)&=\mathbf{E}(ZZ^T\mid t^TZ=s,\ J=1). \end{aligned}$$

Note that m_+ is vector-valued, and V_+ is a matrix-valued function. Define f_-, m_- and V_- in the same way, replacing $J=1$ by $J=-1$. Write also

$$\pi_+=\mathbf{P}(J=1),\qquad \pi_-=1-\pi_+.$$

Put

$$\begin{aligned} D_+(s,t)&=\begin{pmatrix}1 & m_+(1-s)^T\\ m_+(1-s) & V_+(1-s)\end{pmatrix}f_+(1-s),\\ D_-(s,t)&=\begin{pmatrix}1 & -m_-(-1-s)^T\\ -m_-(-1-s) & V_-(-1-s)\end{pmatrix}f_-(-1-s). \end{aligned}$$

Assume

D1. The matrices $D_+(s,t)$ and $D_-(s,t)$ are well-defined, bounded and uniformly continuous as functions of s, uniformly in t in a neighbourhood of t_0.

D2. The matrix $D=\pi_+D_+(s_0,t_0)+\pi_-D_-(s_0,t_0)$ is positive definite.

The function $\widetilde{D}_n$, given by (11), now becomes

$$\begin{aligned}(26)\qquad \widetilde{D}_n(s,t) &= \mathbf{E}\frac{1}{h_n}K\biggl(\frac{1-J(s+t^TZ)}{h_n}\biggr)\begin{pmatrix}1 & JZ^T\\ JZ & ZZ^T\end{pmatrix}\\ &= \pi_+\int\frac{1}{h_n}K\biggl(\frac{u-s}{h_n}\biggr)D_+(u,t)\,du\\ &\quad+\pi_-\int\frac{1}{h_n}K\biggl(\frac{-u+s}{h_n}\biggr)D_-(u,t)\,du.\end{aligned}$$

Theorem 4. *Under assumptions* **D1**, **D2**, **C4** *and* **K**, *if* $h_n=o(1)$ *and* $h_n^{-1}=O(n^{1/2})$ *then* $\widehat{D}_n(t_n)\to_p D$.

Proof. In view of (26), assumption **D1** implies that $\widetilde{D}_n(s,t)\to\pi_+D_+(s,t)+\pi_-D_-(s,t)$ uniformly in a neighbourhood of (s_0,t_0). Thus **II** holds. The result follows from Theorem 1. Let us omit the details, because the argument is standard and quite similar to that in the proof of Theorem 2. ■

Condition **D2** clearly forces (s_0,t_0) to be the unique minimizer of $Q(s,t)$, because $D=\nabla^2Q(s_0,t_0)$ by Lemma 1.

Since the matrix functions appearing in **D1** look rather forbidding, let us give an example of a model in which $f_\pm$, $m_\pm$ and $V_\pm$ can be written explicitly and their regularity properties are easily seen. This will be the model with elliptically contoured class-conditional distributions.

A random vector Z is said to have *elliptically contoured* (e.c.) distribution if it has density of the form

$$(\det\Sigma)^{-1/2}(2\pi)^{-d/2}f_d\bigl(\tfrac12(z-\mu)^T\Sigma^{-1}(z-\mu)\bigr),$$

where d is the dimension of Z, $\mu\in\mathbb{R}^d$ and Σ is a symmetric, positive definite $d\times d$ matrix; we consider only absolutely continuous e.c. distributions. It is easily seen that the standardized random variable $W=\Sigma^{-1/2}(Z-\mu)$ has density $f_d(|w|^2/2)$, which is spherically symmetric. Each component of W, say W_1, has one-dimensional density $(d/du)\mathbf{P}(W_1\le u)=(2\pi)^{-1/2}f_1(u^2/2)$, where

$$f_1(v)=\frac{1}{\Gamma\bigl(\frac{d-1}{2}\bigr)}\int_v^\infty(s-v)^{d/2-3/2}f_d(s)\,ds$$

(see Szabłowski 1990). It will be convenient to consider μ, Σ and f_1 (but *not* f_d) as parameters of an e.c. distribution; thus write $Z\sim EC(\mu,\Sigma,f_1)$. Consider the following assumption:

EC. The conditional distributions of Z given $J=1$ and $J=-1$ are $EC(\mu_+,\Sigma,f_{1+})$ and $EC(\mu_-,\Sigma,f_{1-})$, respectively, with $\mu_+\ne\mu_-$. The second moments of Z exist. The functions $f_{1+}(u)$ and $f_{1-}(u)$ are continuous and decreasing.

Note, in passing, that for $d \geq 3$ the last part of the assumption is spurious, because the $d-2$-dimensional marginal densities of an e.c. distribution are necessarily continuous and decreasing functions of a quadratic form.

PROPOSITION. *Condition* **EC** *implies* **D1**.

Proof. Straightforward computation will lead to explicit formulae for $f_\pm$, $m_\pm$ and $V_\pm$. To avoid awkward repetitions, fix J, suppress conditioning on J in our notation and omit the subscripts $\pm$ for a moment. Begin with simple expressions for conditional moments of the standardized variable $W = \Sigma^{-1/2}(Z-\mu)$:

$$\begin{aligned}\mathbf{E}(W \mid W_1 = u) &= (u, 0, \ldots, 0),\\ \mathrm{Var}(W \mid W_1 = u) &= \mathrm{diag}(0, \sigma^2(u^2/2), \ldots, \sigma^2(u^2/2)),\end{aligned}$$

where W_1 is the first component of W, Var denotes the variance-covariance matrix and the function $\sigma^2(v)$ is given by

$$\sigma^2(v) = \frac{1}{f_1(v)} \int\limits_v^\infty f_1(s)\,ds$$

(see Szabłowski 1990 or Niemiro 1989). A simple change of variables yields

$$\begin{aligned}\mathbf{E}(Z \mid t^T Z = u) &= \mu + \frac{\Sigma t}{t^T \Sigma t}(u - t^T\mu),\\ \mathrm{Var}(Z \mid t^T Z = u) &= \left(\Sigma - \frac{\Sigma t t^T \Sigma}{t^T \Sigma t}\right)\sigma^2\left(\frac{(u - t^T\mu)^2}{2t^T \Sigma t}\right),\\ f(u,t) = \frac{d}{du}\mathbf{P}(t^T Z \leq u) &= \frac{1}{\sqrt{2\pi t^T \Sigma t}} f_1\left(\frac{(u-t^T\mu)^2}{2t^T\Sigma t}\right).\end{aligned}$$

It is easily seen that $\mu_+ \neq \mu_-$ implies $t_0 \neq 0$. Condition **D1** follows from the above formulae, applied to the class-conditional distributions of Z. ■

Thus, **EC** implies **D1** easily, with plenty to spare. In fact, the regularity properties of e.c. distributions ensure good behaviour of higher conditional moments (provided that the higher moments exist, of course). We again refer to Szabłowski (1990). Therefore, under **EC** and **C4**, it is not hard to verify sufficient conditions for the $O_p(n^{-2/5})$-rate of convergence of our estimator $\widehat{D}_n(t_n)$. We will not pursue this point.

To conclude, let us mention that under **EC**, the linear discriminant function $t_0^T Z$, where (s_0, t_0) is the minimizer of $Q(s,t)$, has the following optimality property. There exists some $s_1 \in \mathbb{R}$ such that $\mathbf{P}(\mathrm{sign}(s_1 + t_0^T Z) \neq J) = \inf_{s,t} \mathbf{P}(\mathrm{sign}(s + t^T Z) \neq J)$. In other words, the decision rule $\mathrm{sign}(g(Z))$ predicts J with minimum probability of error among linear g, if $g(Z) = s_1 + t_0^T Z$. This fact follows from simple considerations based on equivariance of $(Q(s,t),J)$ with respect to affine transformations of Z (see Bobrowski 1986 or Niemiro 1987, 1989). Unfortunately, in general the optimal "threshold" or "intercept" term s_1 is *not* equal to s_0.

References

G. J. Babu (1986), *Efficient estimation of the reciprocal of the density quantile function at a point*, Statist. Probab. Letters 4, 133–139.

P. Bloomfield and W. L. Steiger (1983), *Least Absolute Deviations*, *Theory*, *Applications*, *Algorithms*, Birkhäuser, Boston.

L. Bobrowski (1986), *Linear discrimination with symmetrical models*, Pattern Recognition 19, 101–109.

Y. Dodge (ed.) (1987), *Statistical Data Analysis Based on* L_1*-norm and Related Methods*, North-Holland.

— (ed.) (1992), L_1*-Statistical Analysis and Related Methods*, North-Holland.

D. J. Hand (1981), *Discrimination and Classification*, Wiley, New York.

J. Jurečková (1989), *Consistency of M-estimators in a linear model*, *generated by nonmonotone and discontinuous* ψ*-functions*, Probab. Math. Statist. 10, 1–10.

J. Jurečková and P. K. Sen (1987), *A second-order asymptotic distributional representation of M-estimators with discontinuous score functions*, Ann. Probab. 15, 814–823.

—, — (1989), *Uniform second order asymptotic linearity of M-statistics in linear models*, Statist. Decisions 7, 263–276.

J. Kiefer (1967), *On Bahadur's representation of sample quantiles*, Ann. Math. Statist. 38, 1323–1342.

J. Kim and D. Pollard (1990), *Cube root asymptotics*, Ann. Statist. 18, 191–219.

R. Koenker (1987), *A comparison of asymptotic testing methods for* ℓ_1*-regression*, in: Statistical Data Analysis Based on L_1-norm and Related Methods, Y. Dodge (ed.), North-Holland, 287–295.

R. Koenker and G. Basset (1978), *Regression quantiles*, Econometrica 46, 33–50.

J. W. McKean and R. M. Schrader (1987), *Least Absolute Errors Analysis of Variance*, in: Statistical Data Analysis Based on L_1-norm and Related Methods, Y. Dodge (ed.), North-Holland, 297–305.

W. Niemiro (1987), *Statistical properties of the method of minimization of perceptron criterion function in linear discriminant analysis*, Prace IBIB PAN 23 (in Polish).

— (1989), L^1*-optimal statistical discrimination procedures and their asymptotic properties*, Mat. Stos. 31, 57–89 (in Polish).

— (1992), *Asymptotics for M-estimators defined by convex minimization*, Ann. Statist. 20, 1514–1533.

— (1993), *Least empirical risk procedures in statistical inference*, Applicationes Math. 22, 55–67.

D. Pollard (1984), *Convergence of Stochastic Processes*, Springer.

— (1989), *Asymptotics via empirical processes*, Statist. Sci. 4, 341–366.

— (1991), *Asymptotics for least absolute deviation regression estimators*, Econom. Theory 7, 186–199.

C. R. Rao (1988), *Methodology based on the* L_1*-norm in statistical inference*, Sankhyā Ser. A 50, 289–313.

R. M. Schrader and J. W. McKean (1987), *Small sample properties of Least Absolute Errors Analysis of Variance*, in: Statistical Data Analysis Based on L_1-norm and Related Methods, Y. Dodge (ed.), North-Holland, 307–321.

L. Schwartz (1967), *Analyse Mathématique*, Hermann.

P. J. Szabłowski (1990), *Elliptically contoured random variables and their application to the extension of Kalman filter*, Comput. Math. Appl. 19, 61–72.

A. H. Welsh (1987), *Kernel estimates of the sparsity function*, in: Statistical Data Analysis Based on L_1-norm and Related Methods, Y. Dodge (ed.), North-Holland, 369–377.

WOJCIECH NIEMIRO
INSTITUTE OF APPLIED MATHEMATICS
WARSAW UNIVERSITY
BANACHA 2
02-097 WARSZAWA, POLAND
E-mail: WNIEM@APPLI.MIMUW.EDU.PL

INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
P.O. BOX 137
00-950 WARSZAWA, POLAND

Received on 6.4.1994