APPLICATIONES MATHEMATICAE
22,1 (1993), pp. 55–67

W. NIEMIRO (Warszawa)

LEAST EMPIRICAL RISK PROCEDURES IN STATISTICAL INFERENCE

Abstract. We consider the empirical risk function

$$Q_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \cdot f(\alpha, Z_i)$$

(for *iid* Z_i's) under the assumption that $f(\alpha, z)$ is convex with respect to α. Asymptotics of the minimum of $Q_n(\alpha)$ is investigated. Tests for linear hypotheses are derived. Our results generalize some of those concerning *LAD* estimators and related tests.

0. Introduction. There is a general scheme, comprising such statistical procedures as: least absolute deviations (*LAD*), least squares (*LS*), least distances (*LD*), maximum likelihood (*ML*), discrimination based on perceptron-like criteria—to name but a few best known examples. This general scheme will be referred to as least empirical risk (*LER*) method.

Haberman (1989) and Niemiro (1992) examined asymptotic behavior of *LER* estimators, assuming that the underlying loss function is convex. (Here and throughout we slightly abuse the terminology of statistical decision theory. Speaking of "criterion function" would perhaps be pedantic, but certainly more correct.) Pollard (1991) pointed out that *the convexity argument is an idea whose time has come* and gave an excellent example of its application to *LAD* estimators. In this paper, we derive tests of significance for linear hypotheses, under the same basic assumption of convexity. We therefore provide a framework for obtaining, in a more general setting, results such as those described by Rao (1988), Bai, Rao and Yin (1990) or

1991 *Mathematics Subject Classification*: Primary 62F12.

Key words and phrases: convex minimization, asymptotics, least absolute deviations, least distances, tests of significance.

This research was partially supported by KBN grants no. 21168/91/01 and 80492/91/01.

Bai, Rao and Wu (1992). Although our conditions are fulfilled by a large class of *LER* procedures, we focus our attention on *LAD*- or *LD*-type ones. In particular, we discuss applications of *LAD*-related methods to discriminant analysis. This topic certainly deserves more attention than it has received in the literature so far (Niemiro, 1989).

1. Definitions and assumptions. Let Z be a random variable with values in a measurable space $\mathbf{Z}$ and $f : \mathbb{R}^d \times \mathbf{Z} \to \mathbb{R}$. We will regard $f(\alpha, Z)$ as a loss depending on the random quantity Z and on $\alpha \in \mathbb{R}^d$ chosen by the statistician. Accordingly, define $Q : \mathbb{R}^d \to \mathbb{R}$ by

$$Q(\alpha) = \mathbb{E} f(\alpha, Z)\,, \tag{1}$$

and call $Q(\alpha)$ the risk. Suppose the goal is to minimize the risk. Let $\alpha_* \in \mathbb{R}^d$ be such that

$$Q(\alpha_*) = \inf_\alpha Q(\alpha)\,. \tag{2}$$

If the probability distribution of Z is unknown but an *iid* sample $Z_1, \ldots, Z_n$ is available, then we can consider the empirical risk function Q_n, defined as

$$Q_n(\alpha) = \frac{1}{n} \sum_{i=1}^{n} f(\alpha, Z_i)\,, \tag{3}$$

and minimize Q_n instead of Q. Denote by α_n a point, depending on the sample, such that

$$Q_n(\alpha_n) = \inf_\alpha Q_n(\alpha)\,. \tag{4}$$

We will regard α_n as an estimate of α_*. Our basic assumptions are the following:

- (A) $f(\cdot, z) : \mathbb{R}^d \to \mathbb{R}$ is convex for each fixed $z \in \mathbf{Z}$.
- (B) Q is twice differentiable at α_*, with positive definite second derivative $\nabla^2 Q(\alpha_*)$.
- (C) $\partial f(\cdot, z)$ is a subgradient of $f(\cdot, z)$ such that $\mathbb{E}|\partial f(\alpha, Z)|^2 < \infty$ for each α.

Calling $\partial f(\cdot, z)$ a subgradient we mean that the inequality

$$f(\alpha, z) - f(\alpha_0, z) \geq (\alpha - \alpha_0)^T \partial f(\alpha_0, z) \tag{5}$$

holds for all $\alpha, \alpha_0 \in \mathbb{R}^d$ and $z \in \mathbf{Z}$. Here and in the sequel, $|\cdot|$ stands for the euclidean norm, $|\alpha| = (\alpha^T \alpha)^{1/2}$.

To conclude this section, let us briefly comment on existence and uniqueness problems. Conditions implicit in (1–4) (needed in order that these formulae make sense) can be justified using $(A$–$C)$. Let us only list basic facts, referring to Niemiro (1992) for a more comprehensive discussion of those details, which are not really important here. To begin with, assume $Q(\alpha)$

is well defined. (In fact, (C) implies $\mathbb{E}|f(\alpha,Z)-f(\alpha_0,Z)|<\infty$ for all α and α_0. To show this, note that $(\alpha-\alpha_0)^T\partial f(\alpha_0,Z)\le f(\alpha,Z)-f(\alpha_0,Z)\le(\alpha-\alpha_0)^T\partial f(\alpha,Z)$, by definition of subgradient. Replacing, if necessary, $f(\cdot,z)$ by $f(\cdot,z)-f(\alpha_0,z)$ with fixed α_0, we can assume that the expectation in (1) exists.) Convexity of Q follows from (A). If α_* satisfying (B) exists, it must be the unique minimizer of Q. Under our assumptions, α_n satisfying (4) can be shown to exist (at least for large n, with probability one). On the other hand, α_n may not be unique; in the case of ambiguity, α_n can be chosen arbitrarily, subject to (4). The same remark applies to ∂f. A subgradient exists, because $f(\cdot,z)$ is convex, but it is not uniquely determined at points of nondifferentiability of $f(\cdot,z)$. Assume ∂f is selected, subject to (5), in an arbitrary but fixed way. (In fact, we need *measurable* selections of $\partial f(\alpha,\cdot)$ and α_n; see Niemiro (1992) for a way of handling measurability problems.)

2. Asymptotic representations. In this section we give the basic approximation theorems. Most of the proofs are omitted or only sketched, because they can be found in Niemiro (1992). The proof of Theorem 1(b), which is new, will be relegated to the Appendix. Let $(A$–$C)$ be standing assumptions.

Write $\partial Q_n(\alpha)=\frac{1}{n}\sum_{i=1}^n\partial f(\alpha,Z_i)$, to fix a subgradient of Q_n. Let

$$\begin{aligned}\gamma_n &= \partial Q_n(\alpha_*)\,,\\ D &= \nabla^2 Q(\alpha_*)\,,\\ V &= \mathbb{E}\partial f(\alpha_*,Z)\partial f(\alpha_*,Z)^T\,.\end{aligned}\tag{6}$$

The last definition is correct in view of (C) (in fact, $V=\operatorname{Var}\partial f(\alpha_*,Z)$, the covariance matrix, because $\mathbb{E}\partial f(\alpha_*,Z)=\nabla Q(\alpha_*)=0$; see Niemiro, 1992).

Our assumptions allow us to approximate Q_n uniformly by a quadratic function and ∂Q_n by a linear function, near α_* (despite the fact that ∂Q_n may well be discontinuous!).

THEOREM 1. *For every M,*

$$\text{(a)}\quad \sup_{|\alpha-\alpha_*|\le Mn^{-1/2}} |Q_n(\alpha)-Q_n(\alpha_*)-(\alpha-\alpha_*)^T\partial Q_n(\alpha_*) -\tfrac{1}{2}(\alpha-\alpha_*)^T D(\alpha-\alpha_*)| = o_p(n^{-1})\,,$$

$$\text{(b)}\quad \sup_{|\alpha-\alpha_*|\le Mn^{-1/2}} |\partial Q_n(\alpha)-\partial Q_n(\alpha_*)-D(\alpha-\alpha_*)| = o_p(n^{-1/2})\,.$$

The proof is given in the Appendix. In fact, only part (b) has to be proved, since part (a) was established in the course of the proof of Theorem 4 in Niemiro (1992).

As a consequence of Theorem 1 we obtain the following analog of Ghosh's (1970) classical representation:

THEOREM 2. $\alpha_n = \alpha_* - D^{-1}\gamma_n + o_p(n^{-1/2})$.

P r o o f. From Theorem 1 we can deduce that $\alpha_n = \mu_n + o_p(n^{-1/2})$, where μ_n is the minimum point of the quadratic function $\gamma_n^T(\alpha - \alpha_*) + \frac{1}{2}(\alpha - \alpha_*)^T D(\alpha - \alpha_*)$. Of course, $\mu_n = \alpha_* - D^{-1}\gamma_n$. For details, see Niemiro (1992). ∎

Asymptotic normality of γ_n follows from the central limit theorem:

PROPOSITION 1. $n^{1/2}\gamma_n \to_d \mathrm{N}(0, V)$. ∎

As an immediate consequence, we get asymptotic normality of α_n:

PROPOSITION 2. $n^{1/2}(\alpha_n - \alpha_*) \to_d \mathrm{N}(0, D^{-1}VD^{-1})$. ∎

3. Linear hypotheses. Now let us turn to the slightly more general case of minimization with linear constraints. Suppose H is a $p \times d$ matrix of full rank p and $c \in \mathbb{R}^p$ is such that

(H) $\quad H\alpha_* = c.$

Denote by $\dot\alpha_n$ a point such that

$$H\dot\alpha_n = c, \qquad Q_n(\dot\alpha_n) = \inf_{H\alpha = c} Q_n(\alpha)\,. \tag{7}$$

Under (H), we have the following representation of $\dot\alpha_n$, similar to that of α_n:

THEOREM 3. $\dot\alpha_n = \alpha_* + (D^{-1}H^T(HD^{-1}H^T)^{-1}HD^{-1} - D^{-1})\gamma_n + o_p(n^{-1/2})$.

P r o o f. The argument given in the proof of Theorem 2 also applies to the affine subspace $\{\alpha : H\alpha = c\}$, instead of the whole $\mathbb{R}^d$. In consequence, $\alpha_n = \nu_n + o_p(n^{-1/2})$, where ν_n minimizes the quadratic function $\gamma_n^T(\alpha - \alpha_*) + \frac{1}{2}(\alpha - \alpha_*)^T D(\alpha - \alpha_*)$, subject to $H\alpha = c$. To find ν_n, it is enough to solve for α the following equations:

$$\begin{cases} \gamma_n + D(\alpha - \alpha_*) = H^T\lambda\,, \\ H\alpha = c\,. \end{cases}$$

Taking (H) into account, write these equations in the form

$$\begin{pmatrix} -D & \vdots & H^T \\ \cdots & & \cdots \\ H & \vdots & 0 \end{pmatrix} \begin{pmatrix} \alpha - \alpha_* \\ \cdots \\ \lambda \end{pmatrix} = \begin{pmatrix} \gamma_n \\ \cdots \\ 0 \end{pmatrix}\,.$$

The solution is $\nu_n = \alpha_* + (D^{-1}H^T(HD^{-1}H^T)^{-1}HD^{-1} - D^{-1})\gamma_n$. ∎

Assume V is nonsingular. Let us adopt the following notation:

$$\begin{aligned} A &= D^{-1}H^T(HD^{-1}VD^{-1}H^T)^{-1}HD^{-1}\,, \\ B &= D^{-1}H^T(HD^{-1}H^T)^{-1}HD^{-1}\,. \end{aligned} \tag{8}$$

THEOREM 4. *Under* (H) *we have*

$$\partial Q_n(\dot\alpha_n)^T A \partial Q_n(\dot\alpha_n) = \gamma_n^T A \gamma_n + o_p(n^{-1})\,, \tag{a}$$

$$(\alpha_n - \dot\alpha_n)^T DAD(\alpha_n - \dot\alpha_n) = \gamma_n^T A \gamma_n + o_p(n^{-1})\,, \tag{b}$$

$$2(Q_n(\dot\alpha_n) - Q_n(\alpha_n)) = \gamma_n^T B \gamma_n + o_p(n^{-1})\,. \tag{c}$$

P r o o f. To begin with, $\alpha_n = O_p(n^{-1/2})$ and $\dot\alpha_n = O_p(n^{-1/2})$.

To show (a), combine the representations given in Theorems 1(b) and 3. From $\partial Q_n(\dot\alpha_n) = \gamma_n + D(\dot\alpha_n - \alpha_*) + o_p(n^{-1/2})$ and $\dot\alpha_n - \alpha_* = (B - D^{-1})\gamma_n + o_p(n^{-1/2})$ we get $\partial Q_n(\dot\alpha_n) = DB\gamma_n + o_p(n^{-1/2})$. Check that $BDADB = A$ to complete the proof.

Part (b) follows immediately from Theorem 2.

To show (c), substitute the representations given in Theorems 2 and 3 into that of Theorem 1(a): the left-hand side of (c) is $(\dot\alpha_n - \alpha_n)^T D(\dot\alpha_n - \alpha_n) + o_p(n^{-1}) = \gamma_n^T BDB\gamma_n + o_p(n^{-1})$. Of course, $BDB = B$ and the result follows. ∎

For completeness, let us mention another representation, similar to those in Theorem 4, but with different interpretation. The quantity $Q(\alpha_n) - Q(\dot\alpha_n)$ can be regarded as an amount we lose, in terms of risk, when using the unconstrained estimate, α_n, instead of the constrained one, $\dot\alpha_n$. Assume, as before, that (H) is true.

PROPOSITION 3. $2(Q(\alpha_n) - Q(\dot\alpha_n)) = \gamma_n^T B \gamma_n + o_p(n^{-1})$.

P r o o f. Use Theorems 2, 3 and the obvious fact that

$$\sup_{|\alpha - \alpha_*| \le M n^{-1/2}} |Q(\alpha) - Q(\alpha_*) - \tfrac{1}{2}(\alpha - \alpha_*)^T D(\alpha - \alpha_*)| = o(n^{-1})\,. \quad ∎$$

Let us regard (H) as a statistical hypothesis. Suppose we have consistent estimators for the matrices D and V. Then we can use the following statistics to test (H):

$$R_n = n\partial Q_n(\dot\alpha_n)^T \widehat D^{-1} H^T (H\widehat D^{-1}\widehat V \widehat D^{-1} H^T)^{-1} H \widehat D^{-1} \partial Q_n(\dot\alpha_n)\,, \tag{LM}$$

$$W_n = n(H\alpha_n - c)^T (H\widehat D^{-1}\widehat V \widehat D^{-1} H^T)^{-1} (H\alpha_n - c)\,, \tag{W}$$

$$\Lambda_n = n(Q_n(\dot\alpha_n) - Q_n(\alpha_n))\,. \tag{LR}$$

Of course, they are analogs of the classical Lagrange multipliers (LM), Wald (W) and likelihood ratio (LR) tests for maximum likelihood (ML). It is straightforward to derive the asymptotic distributions for R_n and W_n from Theorem 4. Just take into account the fact that $AVA = A$ and use the Cochran theorem. Asymptotic distribution of Λ_n is not, in general, so simple. We have $2\Lambda_n \to_d \chi^2(p)$ iff $BVB = B$. However, there are situations, which are interesting from the viewpoint of practice (e.g. examples in the next section), when $\lambda V = D$ for some $\lambda \in \mathbb{R}$. In the last part of the following theorem we assume that this is the case and we have a consistent estimator for λ.

THEOREM 5. *Under the null hypothesis* (H) *we have*

(a) $$R_n \to_d \chi^2(p)\,,$$

(b) $$W_n \to_d \chi^2(p)\,,$$

provided that $\widehat{V} \to_p V$ *and* $\widehat{D} \to_p D$. *Moreover*,

(c) $$2\widehat{\lambda}\Lambda_n \to_d \chi^2(p)\,,$$

provided that $\lambda V = D$ *and* $\widehat{\lambda} \to_p \lambda$. ∎

Of course, when using the tests, it is crucial to have good estimates $\widehat{V}$ and $\widehat{D}$ or $\widehat{\lambda}$. Some consistent estimates of these "nuisance parameters" can be shown to exist in a quite general setting. Nevertheless, it is much more reasonable to look for better estimates, which take into account specific features of particular models. For instance, much work has been devoted to estimation of λ (or its reciprocal) in linear regression models with LAD-type loss function (Rao, 1988, Welsh, 1987 and many others). These important, interesting and difficult problems go beyond the scope of this paper.

To conclude this section, let us comment on the classical asymptotic theory of ML from the viewpoint of our Theorem 5. Consider a parametric family $\{p(\alpha,\cdot) : \alpha \in \mathbb{R}^d\}$ of probability densities. Let $Z_1, \ldots, Z_n$ be a sample from a density $p(\cdot)$. Setting

$$f(\alpha, z) = -\log p(\alpha, z)\,,$$

we get ML estimators as special cases of the LER method. The usual assumption is that

(L) $\quad p(\cdot) = p(\alpha_*, \cdot)$ for some α_*.

If (L) holds, then necessarily (2) is true. However, the hypothesis (H) makes sense also without assumption (L). The density $p(\alpha_*, \cdot)$ can be interpreted as the member of the parametric family $\{p(\alpha,\cdot)\}$ which is closest to $p(\cdot)$ in the sense of minimum Kullback–Leibler information. Assume the log-likelihood is concave and $(A$–$C)$ hold or (which is more frequently the case) other regularity properties imply the representations of Theorem 1. The conclusions of Theorem 5 are then also in force, no matter whether (L) holds or not. On the other hand, condition (L) does simplify the three tests, because it implies that $D = V = I(\alpha_*)$ (the Fisher information matrix). In this case the tests LM, W and LR assume their usual, simpler form and $2\Lambda_n$ is asymptotically distributed as $\chi^2(p)$.

4. Examples. First three of the examples to be given are well known and were discussed e.g. by Rao (1988), McKean and Schrader (1987). Our aim is to show that our theorems provide a general framework for obtaining such kind of results. In the last example we will be concerned with applications of

the general theory to discriminant analysis. Many discrimination procedures are based on minimization of some convex criteria. Although this technique is widely used (cf. Devijver and Kittler, 1982, Hand, 1981), the asymptotic theory has not been sufficiently developed yet. In particular, this remark applies to the case of nonsmooth criteria of *LAD* type.

EXAMPLE 1 (One-way classification). Let us consider objects, belonging to d distinct classes, with values of some measurement assigned to all of them. If the objects are drawn at random, we can assume that a single observation consists of a pair $Z = (X, Y)$ of random variables, where X takes values $1, \dots, d$ (it is an indicator of class), Y is real. Let us make the standard assumption:

$$Y = \alpha_*^k + U \text{ if } X = k;\ U \text{ is independent of } X\,.$$

We will use the following loss function:

$$f(\alpha, k, y) = |\alpha^k - y| - |y|\,, \tag{9}$$

where $z = (k, y)$, $\alpha = (\alpha^1, \dots, \alpha^d)^T \in \mathbb{R}^d$ (components of vectors will be indexed by superscripts throughout this section). Assume $\operatorname{med} U = 0$, so that α_* minimizes the risk, corresponding to (9). A sample $Z_1, \dots, Z_n$ can be regarded as an array:

$$\begin{array}{cccl} Y_1^1 & \dots & Y_{n_1}^1, & \text{class } 1, \\ \dots & \dots & \dots & \dots \\ Y_1^d & \dots & Y_{n_d}^d, & \text{class } d\,. \end{array}$$

Consider the null hypothesis: $\alpha_*^1 = \dots = \alpha_*^d$, which can be written as $H\alpha = 0$, with $(d-1) \times d$ matrix H defined as

$$H = \begin{pmatrix} 1 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & -1 \end{pmatrix}.$$

The unconstrained minimum of the empirical risk is at the vector of sample medians within classes, i.e. $\alpha_n = (m^1, \dots, m^d)^T$, where

$$m^k = \operatorname{med}(Y_1^k, \dots, Y_{n_k}^k)\,.$$

The constrained minimum is at the median of the pooled sample, $\dot{\alpha}_n = (m, \dots, m)^T$, where

$$m = \operatorname{med}(Y_1^1, \dots, Y_{n_1}^1, \dots, Y_1^d, \dots, Y_{n_d}^d)\,.$$

Of course, $\partial f(\alpha, k, y) = e^k \operatorname{sign}(\alpha^k - y)$, where $e^k = (0, \dots, 1, \dots, 0)^T$ is the kth versor (as usual, we set $\operatorname{sign} 0 = 0$ and so we choose a fixed version of

subgradient). Consequently,

$$\partial Q_n(\alpha) = \frac{1}{n}\sum_{k=1}^{d} e^k(\#\{i : Y_i^k < \alpha^k\} - \#\{i : Y_i^k > \alpha^k\})\,,$$

the symbol # standing for cardinality. In the case under consideration,

$$V = \operatorname{diag}(\pi^1, \ldots, \pi^d)\,,$$

where $\pi^k = \mathbb{P}(X = k)$. If U has a density $p(\cdot)$, continuous at 0, the median, then

$$D = 2p(0)V\,.$$

To see this, compute the kth partial derivative of Q:

$$\nabla_k Q(\alpha) = \pi^k(1 - 2\mathbb{P}(U > \alpha^k - \alpha_*^k))\,.$$

To derive the formulae for R_n and W_n, we need the matrix A given by (8), an estimate of which appears in (LM) and (W). The computation is standard and leads to a familiar result:

$$A = \operatorname{diag}(\pi)^{-1} - \mathbf{1}\mathbf{1}^T\,,$$

where $\pi = (\pi^1, \ldots, \pi^d)^T$, $\mathbf{1} = (1, \ldots, 1)^T$. Let us use the obvious estimate for V: $\widehat{V} = \operatorname{diag}(n_1/n, \ldots, n_d/n)^T$. Assume we have a consistent estimate $\widehat{p}$ for $p(0)$ (a kernel estimate, perhaps) and let $\widehat{D} = 2\widehat{p}\widehat{V}$. The three tests statistics, derived in Section 3, now become:

$$(LM) \qquad R_n = \sum_{k=1}^{d} \frac{(n_k^+ - n_k^-)^2}{n_k}\,,$$

where $n_k^+ = \#\{i : Y_i^k > m\}$, $n_k^- = \#\{i : Y_i^k < m\}$;

$$(W) \qquad W_n = 4\widehat{p}^2 \sum_{k=1}^{d} n_k(m^k - \overline{m})^2\,,$$

where $\overline{m} = n^{-1}\sum_{k=1}^{d} n_k m^k$;

$$(LR) \qquad 4\widehat{p}\Lambda_n = 4\widehat{p}\Big(\sum_{k=1}^{d}\sum_{i=1}^{n_k} |Y_i^k - m| - \sum_{k=1}^{d}\sum_{i=1}^{n_k} |Y_i^k - m^k|\Big)\,.$$

Under (H), in view of Theorem 5, all the three statistics are asymptotically distributed as $\chi^2(d-1)$.

Adamczyk (1993) discusses one-way classification of multivariate observations along similar lines.

EXAMPLE 2 (Location).

• *Marginal medians.* Let $\alpha, z \in \mathbb{R}^d$. Components of these and other vectors will be indexed by superscripts, as in the previous example. Set

$$f(\alpha, z) = \sum_{j=1}^{d} |\alpha^j - z^j| - |z^j| \,.$$

Now, α_*^j is a median of marginal distribution of the random variable Z^j, while α_n^j is a sample median of $Z_1^j, \ldots, Z_n^j$, $j = 1, \ldots, d$. Assume that each component Z^j has a density $p_j(\cdot)$, continuous and nonzero at α_*^j. Clearly

$$\partial f^j(\alpha, z) = \operatorname{sign}(\alpha^j - z^j) \,,$$
$$\operatorname{cov}(\operatorname{sign}(\alpha_*^j - Z^j), \operatorname{sign}(\alpha_*^k - Z^k)) = 4\mathbb{P}(\alpha_*^j < Z^j, \alpha_*^k < Z^k) - 1 \,,$$
$$\nabla_j Q(\alpha) = 1 - 2\mathbb{P}(Z^j > \alpha^j) \,,$$
$$\nabla_{jj}^2 Q(\alpha) = 2p_j(\alpha^j), \qquad \nabla_{jk}^2 Q(\alpha) = 0, \quad j \neq k$$

(∇_j and ∇_{jk}^2 standing for partial derivatives). One can verify that conditions (*A*–*C*) hold. The (j,k)th entry of the matrix $D^{-1}VD^{-1}$, appearing in Proposition 2, becomes

$$\frac{\mathbb{P}(\alpha_*^j < Z^j, \alpha_*^k < Z^k) - 1/4}{p_j(\alpha_*^j) p_k(\alpha_*^k)} \,.$$

• *Spatial median of Haldane* (1948). Let $\alpha, z \in \mathbb{R}^d$ and set

$$f(\alpha, z) = |\alpha - z| - |z| \,,$$

where $|z| = (z^T z)^{1/2}$, as usual. If the probability distribution of Z is not concentrated on any affine subspace of $\mathbb{R}^d$, then the risk function Q has a unique minimum α_* (Milasevic and Ducharme, 1987). This is, by definition, the spatial median. Let us consider the asymptotic behavior of α_n, its sample analogue. Assume that Z has a density, bounded in a neighborhood of α_*. Clearly

$$\partial f(\alpha, z) = \frac{\alpha - z}{|\alpha - z|}, \qquad \alpha \neq z \,.$$

Setting additionally $\partial f(\alpha, \alpha) = 0$ we define a subgradient. Conditions (*A*–*C*) hold, with

$$D = \nabla^2 Q(\alpha_*) = \mathbb{E}|\alpha_* - Z|^{-1}(I - |\alpha_* - Z|^{-2}(\alpha_* - Z)(\alpha_* - Z)^T)$$

being a positive definite matrix (see Niemiro (1992) for a proof). Of course,

$$V = \mathbb{E}|\alpha_* - Z|^{-2}(\alpha_* - Z)(\alpha_* - Z)^T$$

and the conclusion of Proposition 2 holds with D and V as above.

EXAMPLE 3 (Regression). Our results are directly applicable only to regression models with random design. Let us consider an *iid* sequence of

random vectors in $\mathbb{R}^d \times \mathbb{R}$:

$$Z_i = (X_i, Y_i)$$

and set

$$f(\alpha, x, y) = |y - \alpha^T x| - |y| \,. \tag{10}$$

The *LAD* estimate of linear regression coefficients is α_n, which minimizes the empirical risk (3), corresponding to (10). Let α_* be the minimum point of the risk function (1). Assume the probability distribution of (X, Y) satisfies the following regularity conditions. Let $\mathbb{E}|X|^2 < \infty$, and suppose that the density $p(\alpha, \cdot)$ of the random variable $T = Y - \alpha^T X$ and the matrix-valued function

$$V(\alpha, t) = \mathbb{E}(XX^T \mid Y - \alpha^T X = t)$$

are continuous in a neighborhood of $(\alpha_*, 0)$. Moreover, let $p(\alpha_*, 0) > 0$ and $V(\alpha_*, 0)$ be positive definite. Then conditions (*A*–*C*) hold and the conclusion of Proposition 2 is in force, with

$$D = 2p(\alpha_*, 0)V(\alpha_*, 0), \qquad V = \mathbb{E}XX^T \,.$$

The standard assumption is that $Y = \alpha_*^T X + U$, where the error U is independent of X and has a density $p(\cdot)$, continuous at 0, the unique median. In this case the matrix $D^{-1}VD^{-1}$ becomes equal to $V^{-1}(2p(0))^{-2}$, so we get the classical result of Basset and Koenker (1978) (in the random-design version proved by Bloomfield and Steiger, 1983).

Example 4 (Discrimination). Let us look at the previous example in another way. Suppose Y is a binary random variable with values, say, $y = 1$ and $y = -1$, indicating from which of two subpopulations the random vector X comes (note that the parts played by X and Y have been reversed, as compared to Example 1). Instead of regression, we can speak of discrimination. The risk function $Q(\alpha)$ becomes a criterion evaluating the quality of a linear discriminant function $\alpha^T x$. Incidentally, although the loss (10) remains a reasonable choice, another loss function is more natural for discrimination, namely

$$f(\alpha, x, y) = \begin{cases} (1 - \alpha^T x)^+ & \text{if } y = 1 \,, \\ (1 + \alpha^T x)^+ & \text{if } y = -1 \,. \end{cases} \tag{11}$$

The empirical risk corresponding to this loss function is called the perceptron criterion. Let us refer to Hand (1981) for general information on this. A nice example of application can be found in Bobrowski *et al.* (1987). The asymptotic behavior of the discriminant function $\alpha_n^T x$ which minimizes the perceptron criterion was investigated in Niemiro (1989), under strong assumptions on the underlying probability distributions. Now we are in a

position to obtain these results as simple corollaries of the results of Section 2. Considerations from the previous example are still in force, with slight modifications in case we use (11) instead of (10). Let us denote *a priori* probabilities of the two subpopulations by $\pi_\pm = \mathbb{P}(Y = \pm 1)$, write $p_\pm(\cdot)$ for the conditional densities of $T = \alpha_*^T X$ given $Y = \pm 1$ and let

$$V_\pm(t) = \mathbb{E}(XX^T \mid \alpha_*^T X = t,\ Y = \pm 1)\,.$$

Now the asymptotic normality, asserted in Proposition 2, holds with

$$D = \pi_+ p_+(1)V_+(1) + \pi_- p_-(-1)V_-(-1)\,,$$
$$V = \pi_+ \int_{-\infty}^{1} p_+(t)V_+(t)\,dt + \pi_- \int_{-1}^{\infty} p_-(t)V_-(t)\,dt\,.$$

Explicit formulae for D and V were derived by Niemiro (1989) in the case when the conditional distributions of X given Y are elliptically contoured.

Let us conclude this example with the following remark. Several convex criteria can be used to design linear discriminant functions in the case of more than two subpopulations as well. Devijver and Kittler (1982) review some of them, including the well-known *MSE* (mean squared error) criterion. The results of this paper can also be applied in this more general situation.

Appendix. To prove Theorem 1(b) we will need the following lemma, which is a strengthened version of Theorem 25.7 of Rockafellar (1970).

LEMMA 1. *Let $q_n : \mathbb{R}^d \to \mathbb{R}$ be convex functions and $q : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. If for every $\alpha \in \mathbb{R}^d$,*

$$q_n(\alpha) \to q(\alpha)\,,$$

then for every M,

$$\sup_{|\alpha| \le M} |\partial q_n(\alpha) - \nabla q(\alpha)| \to 0\,,$$

where ∂q_n stands for an arbitrary selection of subgradient.

Proof. It is enough to check that the proof given by Rockafellar still goes when we drop the assumption that the q_n are differentiable. ∎

Of course, differentiability of q is essential.

Proof of Theorem 1. Let us simplify notation, assuming without loss of generality that $\alpha_* = 0$ (to achieve this, it is enough to replace $f(\alpha, z)$ by $f(\alpha_* + \alpha, z)$). Set

$$q_n(\alpha) = n(Q_n(n^{-1/2}\alpha) - Q_n(0) - n^{-1/2}\alpha^T \partial Q_n(0))\,,$$
$$q(\alpha) = \tfrac{1}{2}\alpha^T D\alpha\,.$$

Now, we can rewrite part (a) of the theorem in the following, equivalent way:

$$(a') \qquad \sup_{|\alpha|\le M} |q_n(\alpha) - q(\alpha)| \to_p 0\,.$$

To prove (a), it is enough to notice that inequality (4.13) in Niemiro (1992) is tantamount to (a′). It remains to deduce part (b) of the theorem from part (a).

By Lemma 1, if $\sup_{|\alpha|\le M} |q_n(\alpha) - q(\alpha)| \to 0$ a.s., then $\sup_{|\alpha|\le M'} |\partial q_n(\alpha) - \nabla q(\alpha)| \to 0$ a.s. for $M' < M$. The standard technique of subsequences allows us to replace almost sure convergence by convergence in probability. Consequently, (a′) implies

$$(b') \qquad \sup_{|\alpha|\le M'} |\partial q_n(\alpha) - \nabla q(\alpha)| \to_p 0\,,$$

which is equivalent to part (b) of the theorem. ∎

References

K. Adamczyk (1993), *Asymptotic properties of ANOVA test under general loss functions*, Mat. Stos., to appear.

Z. D. Bai, C. R. Rao and Y. Q. Yin (1990), *Least absolute deviations analysis of variance*, Sankhyā A 52, 166–177.

Z. D. Bai, C. R. Rao and Y. H. Wu (1992), *M-estimation of multivariate linear regression parameters under a convex discrepancy function*, Statist. Sinica 2 (1), 237–254.

G. Basset and R. Koenker (1978), *Asymptotic theory of least absolute error regression*, J. Amer. Statist. Assoc. 73, 618–622.

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

L. Bobrowski, H. Wasyluk and W. Niemiro (1987), *Some technique of linear discrimination with application to analysis of thyroid diseases diagnosis*, Biocybernetics Biomed. Engrg. 7, 23–32.

P. A. Devijver and J. Kittler (1982), *Pattern Recognition*: *A Statistical Approach*, Prentice-Hall, London.

J. K. Ghosh (1971), *A new proof of the Bahadur representation of quantiles and an application*, Ann. Math. Statist. 42, 1957–1961.

S. J. Haberman (1989), *Concavity and estimation*, Ann. Statist. 17, 1631–1661.

J. B. S. Haldane (1948), *Note on the median of a multivariate distribution*, Biometrika 25, 414–415.

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

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.

P. Milasevic and G. R. Ducharme (1987), *Uniqueness of the spatial median*, Ann. Statist. 15, 1332–1333.

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

W. Niemiro (1992), *Asymptotics for M-estimators defined by convex minimization*, Ann. Statist., to appear.
D. Pollard (1991), *Asymptotics for least absolute deviation regression estimators*, Econometric Theory 7, 186–199.
C. R. Rao (1988), *Methodology based on the L_1-norm in statistical inference*, Sankhyā A 50, 289–313.
R. T. Rockafellar (1970), *Convex Analysis*, Princeton University Press.
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.

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

Received on 24.9.1992