Conditions for Consistency of Minimum Contrast Estimators and M-Estimators[1]

By I. Vajda

Abstract: Minimum contrast estimators based on depending observations are considered. A condition is found which is necessary and sufficient for consistency of all asymptotically minimum contrast estimators. Specification of this condition for M-estimators of parameters of nonlinear and linear regression with random regressors is studied and the consistency is established under very general assumptions.

Key words and phrases: Minimum contrast estimator, M-estimator, nonlinear regression, linear regression, consistency, inconsistency.

AMS 1990 subject classification: 62 F 12, 62 J 02

1 Introduction

The influential paper of Pfanzagl (1969) introduced theoretically important classes of minimum contrast estimators (MCE's) and asymptotically minimum contrast estimators (AMCE's) as generalizations of maximum likelihood and asymptotically maximum likelihood estimators (MLE's and AMLE's). Using this paper, Strasser (1981) formulated simpler reasonably weak regularity assumptions on estimation models guaranteeing the existence of AMLE's and found a condition equivalent with the strong consistency of all AMLE's. Under weaker regularity assumptions Vajda (1992) found a condition equivalent with the usual (weak) consistency of all AMLE's.

In this paper we consider AMCE's for general depending observations. We formulate regularity assumptions under which AMCE's exist and introduce a condition equivalent with the weak consistency of all AMCE's. Easily verifiable conditions sufficient for consistency of all AMCE's and for inconsistency of all AMCE's are established as well. These conditions are used to prove consistency or inconsistency of some MCE's and AMCE's.

The sufficient condition is then specified for M-estimators of parameters of linear and nonlinear regression with stationary and ergodic errors and regressors. It is shown to yield sufficient conditions for consistency of M-estimators in nonlinear regression analogical

[1] Supported by CSAS grant N. 17503.

sufficient conditions for consistency of M-estimators in nonlinear regression analogical to those of Jennrich (1969), Richardson and Bhattacharyya (1986), 1987). In the linear model with i. i. d. errors and regressors these conditions reduce to the conditions presented by Chen and Wu (1988).

2 Regularity assumptions

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and (Θ, d) a locally compact separable metric space. We shall be interested in estimation of an unknown parameter $\theta_0 \in \Theta$ on the basis of random observations $\mathbf{X} = (X_1, \ldots, X_n)$ defined on a basic probability space $(\Omega, \mathcal{S}, \mathsf{P})$ and taking on values in sample probability spaces $(\mathcal{X}^n, \mathcal{A}^n, P_{\theta_0,n})$, $n \in \mathbb{N}$. The unknown distribution $P_{\theta_0,n}$ is assumed to belong to a known family $\mathcal{P}_n = \{P_{\theta,n} : \theta \in \Theta\}$. The value of θ_0 is assumed to be fixed and the quantifier "for all $\theta_0 \in \Theta$" is systematically omitted. Note also that the sample size index n is omitted in the symbols for observation variables, i. e. that in fact we admit triangular observation schemes

$$\mathbf{X}^{(n)} = \left(X_1^{(n)}, \ldots, X_n^{(n)}\right).$$

Throughout the paper we use the open spheres $S_a = \{\theta \in \Theta : d(\theta, \theta_0) < a\}$. The interiors of their complements $S_a^c = \Theta - S_a$ are denoted by $S_{a,c}$. The attention is restricted to $a > 0$ such that $S_a^c \neq \emptyset$. The standard topology and arithmetics of the extended real line $\overline{\mathbb{R}}$, introduced e.g. in Prerequisities of Halmos (1964), are used without specific references.

An *estimator* is a sequence of $(\mathcal{A}^n, \mathcal{B})$-measurable mappings $\theta_n : \mathcal{X} \to \Theta$, where $\mathcal{B}$ denotes the Borel σ-field of Θ. The estimator is said *consistent* if $\theta_n \to \theta_0$ in P. This means that, for all sufficiently small $a > 0$,

$$\lim_{n\to\infty} \mathsf{P}(\theta_n \in S_a) = 1. \tag{1}$$

Under present assumptions about Θ, this is equivalent with

$$\lim_{n\to\infty} \mathsf{P}(\theta_n \in S_{a,c}) = 0.$$

The definition of MCE and AMCE is based on the so-called contrast function $f(x, \theta) : \mathcal{X} \otimes \Theta \to \overline{\mathbb{R}}$ proposed by Pfanzagl (1969). First we introduce some useful notation. Put for $\theta \in \Theta$, $\emptyset \neq B \subset \Theta$, and $n \in \mathbb{N}$

$$f_n(\mathbf{x}, \theta) = \frac{1}{n} \sum_{i=1}^{n} f_n(x_i, \theta), \quad \mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n \qquad (-\infty + \infty = -\infty),$$

and

$$\begin{aligned} f_n(\mathbf{x}, B) &= \inf_{\theta \in B} f(\mathbf{x}, \theta), \quad \mathbf{x} \in \mathcal{X}^n, & f(x, B) = f_1(x, B), \quad x \in \mathcal{X}, \\ f_n(\theta) &= f_n(\mathbf{X}, \theta), & f_n(B) = f_n(\mathbf{X}, B). \end{aligned}$$

In general, $f_n(\theta)$ and $f_n(B)$ need not be random variables. If, however, $x \to f(x,\theta)$ is $\mathcal{A}$-measurable then $\mathrm{x} \to f_n(\mathrm{x},\theta)$ is $\mathcal{A}^n$-measurable, so that $f_n(\theta)$ is an $\overline{\mathbb{R}}$-valued random variable. Following Strasser (1981) we assume that $f(x,\theta)$ is $\mathcal{A} \otimes \mathcal{B}$-measurable and that the random function $\theta \to f_n(\theta)$ is separable (see § 2 in Chap. II of Doob (1953) for $\Theta \subset \mathbb{R}$ and Borges (1966) or p. 266 in Pfanzagl (1969) for the general Θ under consideration; a systematic theory is studied e. g. in §§ 2, 3 of Chap. III in Gikhman and Skorokhod (1971)). The measurability implies that $f_n(\theta_n)$ is an $\overline{\mathbb{R}}$-valued random variable for every estimator θ_n. The separability implies for each the existence of $A_n \in \mathcal{A}^n$ and with $\mathrm{P}(\mathrm{X} \in A_n) = 1$ such that for every nonempty open $B \subset \Theta$ there is an at most countable subset $B_n \subset B$ with the property

$$f_n(\mathrm{x}, B) = f_n(\mathrm{x}, B_n), \quad \mathrm{x} \in A_n.$$

Hence, for every nonempty open $B \subset \Theta$, the restriction of $f_n(\mathrm{x}, B)$ on A_n is $\mathcal{A}^n \cap A_n$-measurable and $f_n(B)$ may thus be viewed as an $\overline{\mathbb{R}}$-valued random variable.

It is known (cf. the above cited references) that if all functions $x \to f(x,\theta)$, $\theta \in \Theta$, are $\mathcal{A}$-measurable then the separability of the random function $\theta \to f_n(\theta)$ follows from its P-a. s. continuity or stochastic continuity.

Throughout the paper we restrict ourselves to functions $f(x,\theta)$ satisfying the following two regularity assumptions.

(A1) The function $f(x,\theta)$ is $\mathcal{A} \otimes \mathcal{B}$-measurable and, for every $n \in \mathbb{N}$, the random function $\theta \to f_n(\theta)$ is separable.

(A2) For every $n \in \mathbb{N}$, the random function $\theta \to f_n(\theta)$ is P-a. s. lower semicontinuous on Θ.i. e., for $P_{\theta_0,n}$-almost all $\mathrm{x} \in \mathcal{X}^n$, $\liminf_{k\to\infty} f_n(\mathrm{x},\theta_k) \geq f_n(\mathrm{x},\theta)$ if $\theta_k \to \theta$.

The observation model is supposed to satisfy the following assumption.

(A3) For every $n \in \mathbb{N}$, the family $\mathcal{P}_n$ consists of measure-theoretically equivalent distributions.

As indicated above, these assumptions are inspired by the paper of Strasser (1981). That paper was restricted to the special case with independent observations, i. e. with

$$P_{\theta,n} = P_\theta^n, \qquad \theta \in \Theta, \tag{2}$$

where the marginal distributions P_θ, $\theta \in \Theta$, do not depend on $n \in \mathbb{N}$ and are dominated by a σ-finite measure μ on $\mathcal{A}$, and the contrast function is of log-likelihood type,

$$f(x,\theta) = -\ln\left(\left(\frac{dP_\theta}{d\mu}\right)(x)\right) \qquad (-\ln 0 = \infty). \tag{3}$$

Our assumptions are simpler than those of Strasser (1981) and are more easily verifiable than those of Pfanzagl (1969) who was also restricted to the i. i. d. case (2).

Remark 1

(A3) can be replaced by the weaker assumption that, for every $n \in \mathbf{N}$, $\mathcal{P}_n$ is dominated by a probability measure Q_n, provided that at the same time the semicontinuty and separability of $\theta \to f_n(\mathrm{X}, \theta)$ in (A1) and (A2) are replaced by the semicontinuity and separability of $\theta \to f_n(\mathrm{Y}, \theta)$ for Y with the sample space $(\mathcal{X}^n, \mathcal{A}^n, Q_n)$. In this case Lemma 1 below can be proved simply by replacing $P_{\theta_0^*,n}$ by Q_n. Proof of Lemma 3,which is is the only remaining assertion of this paper where (A3) is employed,can be modified analogically. In this manner one easily obtains slightly extended versions of all the results that follow.

3 Basic definitions

Now we can complete the definition of contrast function and present the definitions of MCE and AMCE. Our definitions differ from those of Pfanzagl (1969) only in technical details, connected with different regularity assumptions and with our aim to study the weak consistency of estimators.

A *contrast function* is a mapping $f: \mathcal{X} \otimes \Theta \to \overline{\mathbb{R}}$ satisfying the assumptions (A1), (A2). A *minimum contrast estimator* (MCE) is an estimator θ_n such that

$$f_n(\theta_n) = f_n(\Theta) \qquad \text{P-a.s.}$$

An *asymptotically minimum contrast estimator* (AMCE) is an estimator θ_n such that, for some $\varepsilon_n \downarrow 0$,

$$\lim_{n \to \infty} \mathsf{P}\left(f_n^0(\theta_n) \leq f_n^0(\Theta) + \varepsilon_n\right) = 1 \tag{4}$$

where, here and in the sequel,

$$f_n^0(\cdot) = \varphi \circ f_n(\cdot) \tag{5}$$

for $\varphi: \overline{\mathbb{R}} \longrightarrow [-1, 1]$ defined by $\varphi(x) = x/(1 + |x|)$. AMCE's are also called *approximate minimum contrast estimators* (cf. Perlman (1972), Strasser (1981)).

Note that Pfanzagl (1969) defined in the special case (2) AMCE by the condition

$$\lim_{n \to \infty} \left(f_n^0(\theta_n) - f_n^0(\Theta)\right) = 0 \qquad \text{P-a.s.}$$

Similar definition has been considered under (2) and (3) by Strasser (1981). Replacing here the a.s. convergence by the convergence in probability we obtain (4). Hence our class of AMCE's is wider irrespectively of whether (2) or (3) is assumed or not.

Note also that the identifiability property of contrast functions

$$\mathsf{E}f(\theta) > \mathsf{E}f(\theta_0), \qquad \theta \in \Theta,\ \theta \neq \theta_0, \tag{6}$$

assumed under (2) by Pfanzagl (1969), is not required by our definition but will be considered below (cf. Theorem 3).

Example 1

Let Θ be a Borel subset of $\mathbf{R}$ and let us consider (2), with a dominated family $\{P_\theta : \theta \in \Theta\}$ of probability distributions on the Borel σ-field of $\mathbf{R}$ having the first absolute moments finite. Then the *least square error*

$$f(x,\theta) = (x-\theta)^2, \qquad x \in \mathbf{R},$$

is a contrast function. Indeed, the process

$$\theta \to \frac{1}{n}\sum_{i=1}^{n}(X_i-\theta)^2$$

is continuous for every $n \in \mathbf{N}$, so that (A1) and (A2) hold. The sample mean

$$\theta_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is obviously thus MCE provided $\Theta = \mathbf{R}$. The *least absolute error*

$$\hat{f}(x,\theta) = |x-\theta|, \qquad x \in \mathbf{R},$$

is another possible contrast function, for which the sample median

$$\hat{\theta}_n = \begin{cases} X_{(k+1)} & \text{if } n = 2k+1 \\ \frac{1}{2}\left(X_{(k)} + X_{(k+1)}\right) & \text{if } n = 2k \end{cases}$$

is MCE provided $\Theta = \mathbf{R}$. If Θ is a closed subset of Θ then the projections $\bar{\theta}_n$, $\tilde{\theta}^*_n$ of θ_n, $\hat{\theta}_n$ on Θ are the respective MCE's. If Θ is a proper nonclosed subset of Θ then the corresponding MCE's may not exist. For example, if Θ is the set of rational numbers and $P_\theta \equiv N(\theta,1)$ then neither the least square error nor the least absolute error estimator exist. We shall see (cf. Lemma 1 below) that the corresponding AMCE's $\bar{\theta}_n$ and $\tilde{\theta}_n$ do exist.

Example 2

This is a simplification and generalization of Example 3 on pp. 410–412 of Lehman (1983). Let $(\mathcal{X},\mathcal{A})$ be the real line $\mathbf{R}$ with the Borel σ-field of $\mathbf{R}$. Put $\Theta = (0,1)$, and consider the Lebesgue measure μ on $\mathcal{A}$, and an increasing function $\psi : (0,1) \to \mathbf{R}_+$ such that

$$\psi(\theta) > \ln\frac{1}{1-\theta}, \qquad \theta \in (0,1). \tag{7}$$

Under this condition the closed intervals

$$A_\theta = \left[\theta, \theta + e^{-\psi(\theta)}\right], \qquad \theta \in (0,1),$$

are contained in the parameter space $(0,1)$. Consider (2) for P_θ, $\theta \in (0,1)$, defined on $\mathcal{A}$ by the densities

$$p_\theta(x) = \left(\frac{dP_\theta}{d\mu}\right)(x) = (1-c)\,\mathbf{I}_{(0,1)}(x) + c\,e^{\psi(\theta)}\,\mathbf{I}_{A_\theta}(x), \qquad x \in \mathbf{R},$$

where $\frac{1}{2} < c < 1$ and where $\mathbf{I}_A$ is the indicator function of $A \subset \mathbf{R}$. This means that P_θ is a stochastic mixture of the uniform distribution on $(0,1)$ and the uniform distribution on A_θ.

Now we are going to introduce a family of contrast functions. Consider for $\alpha \in [0,1)$ the mappings $\varphi_\alpha : [0,\infty) \longrightarrow [-\infty,\infty)$ defined by

$$\varphi_\alpha(x) = \begin{cases} \frac{1-x^\alpha}{\alpha} & \text{if } \alpha \in (0,1) \\ \lim_{\alpha \downarrow 0} \varphi_\alpha(x) = -\ln x & \text{if } \alpha = 0 \ (\ln 0 = -\infty). \end{cases}$$

The expressions

$$f(x,\theta) = \varphi_\alpha(p_\theta(x)), \qquad x \in \mathbf{R},\ \theta \in (0,1),$$

will be called *α-functions*. Analogically as in the previous example, one can verify that the process

$$\theta \longrightarrow \frac{1}{n}\sum_{i=1}^{n} \varphi_\alpha(p_\theta(X_i))$$

is stochastically continuous. Consequently (A1) holds. The functions $\theta \longrightarrow p_\theta(x)$, $x \in \mathbf{R}$, are upper semicontinuous so that $\theta \longrightarrow \varphi_\alpha(p_\theta(x))$, $x \in \mathbf{R}$, are lower semicontinuous. This implies that (A2) holds. Therefore all α-functions are the contrast functions.

MCE corresponding to $\alpha = 0$ is the MLE. MCE's corresponding to $\alpha \in (0,1)$ have been introduced in Vajda (1984) and called *α-estimators*. It can be shown that in the present example the MLE, as well as all α-estimators, exist. However, they cannot be explicitely evaluated as the estimators of Example 1.

4 Condition equivalent with consistency of AMCE's

Our first result is concerning the existence of AMCE's. Lemma 1 shows in fact more, namely that the approximate minimum contrast condition (4) can be satisfied not only in the stated asymptotical stochastic sense, but also deterministically, for all sample sizes $n \in \mathbf{N}$

Lemma 1

For every $\varepsilon_n \downarrow 0$ there exists an estimator θ_n such that

$$f_n^0(\theta_n) \leq f_n^0(\Theta) + \varepsilon_n \qquad \mathrm{P}-a.s., \quad n \in \mathbf{N}.$$

This estimator is AMCE.

Proof

Let $n \in \mathbb{N}$ and $\theta_0^* \in \Theta$ be arbitrary fixed. By (A2) there exists $A_n \in \mathcal{A}^n$ of unit $P_{\theta_0^*,n}$-probability such that the functions $\theta \longrightarrow f_n^0(\mathrm{x},\theta)$, $\mathrm{x} \in A_n$, are lower semicontinuous. Hence the sets

$$B_n(\mathrm{x}) = \left\{\theta \in \Theta :\ f_n^0(\mathrm{x},\theta) \le f_n^0(\mathrm{x},\Theta) + \varepsilon_n\right\}, \qquad \mathrm{x} \in A_n,$$

are nonempty and closed. Define for $C \subset \Theta$ a subset $D_n = D_n(C)$ of A_n by

$$D_n = \{\mathrm{x} \in A_n :\ C \cap B_n(\mathrm{x}) = \emptyset\}.$$

If for every compact C if holds $D_n \in \mathcal{A}^n \cap A_n$ then, by Theorem 3.9 of Pfanzagl (1969), there exists an $(\mathcal{A}^n \cap A_n, \mathcal{B})$-measurable mapping $\hat{\theta}_n :\ A_n \to \Theta$ such that

$$\hat{\theta}_n(\mathrm{x}) \in B_n(\mathrm{x}), \qquad \mathrm{x} \in A_n.$$

Since by (A3) $P_{\theta_0^*,n}(A_n) = 1$ implies $P_{\theta_0,n}(A_n) = 1$, we see that in this case an arbitrary extension θ_n of $\hat{\theta}_n$ constant on $\mathcal{X}^n - A_n$ possesses the desired property $f_n^0(\theta_n) \le f_n^0(\Theta) + \varepsilon_n$ P-a.s.

Let C be compact. In order to prove the $\mathcal{A}^n \cap A_n$-measurability of $D_n = D_n(C)$ take into account that if $\mathrm{x} \in A_n$, then $C \cap B_n(\mathrm{x}) = \emptyset$ is equivalent with the condition

$$f_n^0(\mathrm{x},\theta) > f_n^0(\mathrm{x},\Theta) + \varepsilon_n, \qquad \theta \in C.$$

But if $\mathrm{x} \in A_n$, the the lower semicontinuous function $f_n^0(\mathrm{x},\theta)$ attains on compact C its infimum $f_n^0(\mathrm{x},C)$. Therefore the following formula holds

$$D_n = \left\{\mathrm{x} \in A_n :\ f_n^0(\mathrm{x},C) > f_n^0(\mathrm{x},\Theta) + \varepsilon_n\right\}.$$

By (A1), for every at most countable class $\mathcal{U}$ of open subsets $U \subset \Theta$ the restrictions $f_n^0(\mathrm{x},U)$ on A_n may be assumed $\mathcal{A}^n \cap A_n$-measurable (otherwise it suffices to replace A_n by a convenient $\mathcal{A}^n$-measurable subset $A_n^* \subset A_n$ of unit $\mathrm{P}_{\theta_0^*,n}$-probability). Since Θ can be assumed to be contained in every $\mathcal{U}$ under consideration, the restriction $f_n^0(\mathrm{x},\Theta)$ on A_n is $\mathcal{A}^n \cap A_n$-measurable. By Lemma 3.5 of Pfanzagl (1969), the $\mathcal{A}^n \cap A_n$-measurability of the restrictions of $f^0(\mathrm{x},U)$ on A_n for all $U \in \mathcal{U}$ implies the $\mathcal{A}^n \cap A_n$-measurability of the restriction of $f^0(\mathrm{x},C)$ on A_n. Thus it follows from the last formula for D_n that $D_n \in \mathcal{A}^n \cap A_n$ Q.E.D.

Lemma 2

If

$$\lim_{n\to\infty} \mathrm{P}\left(f_n^0(S_{\bullet,c}) > f_n^0(\Theta) + \delta_n\right) = 1 \qquad \text{for every} \quad \delta_n \downarrow 0, \tag{8}$$

then every AMCE satisfies (1).

Proof

Let us consider AMCE θ_n and $\varepsilon_n \downarrow 0$ satisfying (4). By (8) there exists $\delta_n \downarrow 0$, $\delta_n \geq \varepsilon_n$, such that

$$\lim_{n \to \infty} \mathsf{P}\left(f_n^0(S_{a,c}) > f_n^0(\Theta) + \delta_n\right) = 1.$$

It holds

$$\mathsf{P}\left(\theta_n \in S_{a,c}\right) \leq P\left(\theta_n \in S_{a,c}, f_n^0(\theta_n) \leq f_n^0(\Theta) + \delta_n\right) + \mathsf{P}\left(f_n^0(\theta_n) > f_n^0(\Theta) + \delta_n\right)$$

The first right-hand term is bounded above by

$$\mathsf{P}\left(f_n^0(S_{a,c}) \leq f_n^0(\Theta) + \delta_n\right)$$

and thus tends to zero. The second right-hand terms tends to zero by (4). Thus (1) holds Q. E. D.

Lemma 3

If (8) is not satisfied then there exists AMCE θ_n not satisfying (1).

Proof

By Lemma 1 there exists AMCE $\hat{\theta}_n$ and $\hat{\varepsilon}_n \downarrow 0$ such that

$$\lim_{n \to \infty} \mathsf{P}\left(f_n^0(\hat{\theta}_n) \leq f_n^0(\Theta) + \hat{\varepsilon}_n\right) = 1.$$

The statistical model $(\mathcal{X}^n, \mathcal{A}^n, P_{\theta,n} : \theta \in S_{a,c})$ and the restriction of the contrast function $f(x,\theta)$ on $\mathcal{X} \otimes S_{a,c}$ satisfy (A1) - (A3). Therefore, by Lemma 1 and by the equivalence assumption in (A3), there exists an estimator $\theta_n^c : \mathcal{X}^n \to S_{a,c}$ such that

$$f_n^0(\theta_n^c) \leq f_n^0(S_{a,c}) + \hat{\varepsilon}_n \qquad \mathsf{P}\text{-a. s.}$$

If (8) is not satisfied then there exists a sequence $\delta_n \downarrow 0$ such that

$$\limsup_{n \to \infty} \mathsf{P}\left(f_n^0(S_{a,c}) \leq f_n^0(\Theta) + \delta_n\right) = \gamma > 0.$$

Putting

$$\theta_n = \begin{cases} \theta_n^c & \text{if } f_n^0(S_{a,c}) \leq f_n^0(\Theta) + \delta_n \\ \hat{\theta}_n & \text{otherwise} \end{cases}$$

we obtain an estimator with the probability

$$\mathsf{P}\left(f_n^0(\theta_n) \leq f_n^0(\Theta) + \hat{\varepsilon}_n + \delta_n\right)$$

equal

$$\begin{aligned} & \mathsf{P}\left(f_n^0(\theta_n^\varepsilon) \leq f_n^0(\Theta) + \hat{\varepsilon}_n + \delta_n\right) \\ \geq \; & \mathsf{P}\left(f_n^0(S_{a,\varepsilon}) + \hat{\varepsilon}_n \leq f_n^0(\Theta) + \hat{\varepsilon}_n + \delta_n\right) = 1 \end{aligned}$$

under the condition $f_n^0(S_{a,\varepsilon}) \leq f_n^0(\Theta) + \delta_n$, and equal

$$\mathsf{P}\left(f_n^0(\hat{\theta}_n) \leq f_n^0(\Theta) + \hat{\varepsilon}_n + \delta_n\right) \longrightarrow 1$$

otherwise. In other words, θ_n is AMCE. On the other hand, taking into account the obvious inequalities

$$\mathsf{P}\left(\theta_n \in S_{a,\varepsilon}\right) \geq \mathsf{P}\left(\theta_n = \theta_n^\varepsilon\right) \geq \mathsf{P}\left(f_n^0(S_{a,\varepsilon}) \leq f_n^0(\Theta) + \delta_n\right),$$

we see that (1) cannot be satisfied by θ_n Q. E. D.

Theorem 1

All AMCE's are consistent if and only if (8) holds for all sufficiently small $a > 0$.

Proof

Clear from Lemmas 2, 3.

5 Conditions sufficient for consistency or inconsistency of all AMCE's

If (8) is not satisfied then there exists an inconsistent AMCE. But it might happen that, at the same time, the MCE exists and is consistent. Such an example was presented in Vajda (1992). The following result offers a condition sufficient for inconsistency of all AMCE's. This condition is obviously stronger than the contrary to (8).

Lemma 4

The existence of a consistent AMCE implies that, for all sufficiently small $a > 0$,

$$\lim_{n \to \infty} \mathsf{P}\left(f_n^0(\Theta) > f_n^0(S_a) - \varepsilon\right) = 1, \qquad \varepsilon > 0. \tag{9}$$

Proof

Let θ_n be a consistent AMCE and let $\varepsilon_n \downarrow 0$ satisfy (4). Then, for every $\varepsilon > 0$,

$$\left\{f_n^0(\Theta) \leq f_n^0(S_a) - \varepsilon\right\} \subset A_n \cup \left\{\theta_n \notin S_n\right\}$$

where

$$\begin{aligned}
A_n &= \left\{f_n^0(\Theta) \le f_n^0(S_a) - \varepsilon,\ \theta_n \in S_a\right\} \\
&\subset B_n \cup \left\{f_n^0(\theta_n) > f_n^0(\Theta) + \varepsilon_n\right\}, \\
B_n &= \left\{f_n^0(\Theta) \le f_n^0(S_a) - \varepsilon,\ f_n^0(\theta_n) \le f_n^0(\Theta) + \varepsilon_n\right\} \\
&\subset \left\{f_n^0(\Theta) \le f_n^0(S_a) - \varepsilon,\ f_n^0(S_a) \le f_n^0(\Theta) + \varepsilon_n\right\} \\
&\subset \left\{f_n^0(S_a) \le f_n^0(S_a) - \varepsilon + \varepsilon_n\right\}.
\end{aligned}$$

The last event is empty for $\varepsilon_n < \varepsilon$. Hence it follows from (5) and from the consistency of θ_n that

$$\lim_{n\to\infty} \mathsf{P}\left(f_n^0(\Theta) \le f_n^0(S_a) - \varepsilon\right) = \lim_{n\to\infty} \left[\mathsf{P}(f_n^0(\theta_n) > f_n^0(\Theta) + \varepsilon_n) + \mathsf{P}(\theta_n \notin S_a)\right] = 0 \quad \text{Q.E.D.}$$

Theorem 2

The condition

$$\lim_{n\to\infty} \mathsf{P}\left(f_n^0(S_{a,c}) > f_n^0(S_a) + \delta_n\right) = 1, \quad \delta_n \downarrow 0,\ a > 0, \tag{10}$$

is sufficient for consistency of all AMCE's. The weaker condition

$$\lim_{n\to\infty} \mathsf{P}\left(f_n^0(S_{a,c}) > f_n^0(S_a) - \varepsilon\right) = 1, \quad \varepsilon > 0,\ a > 0, \tag{11}$$

is necessary for consistency of at least one AMCE.

Proof

It follows from the inequality $f_n^0(S_a) \ge f_n^0(\Theta)$ that (10) implies (8) for all sufficiently small $a > 0$. Thus the first assertion follows from Lemma 2. It follows from the inequality $f_n^0(S_{a,c}) \ge f_n^0(\Theta)$ that (9) valid for all sufficiently small $a > 0$ implies (11). Therefore the second assertion follows from Lemma 4.

Now we present conditions under which (10) holds (Theorem 3) and conditions under which (11) does not hold (Theorem 4). Thus, in a combination with Theorem 2, these results present conditions sufficient for consistency of all AMCE's, or for inconsistency of all AMCE's, respectively. All results that follow are restricted to observation models satisfying the following regularity assumtions.

(A4) The observations $\mathbf{X} = (X_1, \ldots, X_n)$ are segments of an infinite stationary sequence $X_1, X_2, \ldots$ of random variables satisfying the law of large numbers in the sense that, for every $\mathcal{A}$-measurable function $\psi : \mathcal{X} \to \overline{\mathbb{R}}$ with $\min\{\mathsf{E}\,\psi(X)^+,\ \mathsf{E}\,\psi(X)^-\} < \infty$

(where $X = X_1$ and the integrands are the positive and negative parts of random variable $\psi(X)$),

$$\lim_{n\to\infty} \mathsf{P}\left(\left|\varphi \circ \frac{1}{n}\sum_{i=1}^{n}\psi(X_i) - \varphi \circ \mathsf{E}\,\psi(X_i)\right| > \varepsilon\right), \quad \varepsilon > 0 \quad \text{(cf. (5))}.$$

(A5) Every $\theta \in \Theta$ has an open neighborhood $B_\theta \subset \Theta$ such that

$$\mathsf{E}\, f(B_\theta) > -\infty, \quad \theta \in \Theta,$$

where $f(B_\theta) = f_1(B_\theta) = f(X, B_\theta)$ for $X = X_1$ as in (A4).

Remark 2.

It follows from (A5) and from the inequality $f(B_\theta) \leq f(\theta)$ that $\mathsf{E}\, f(\theta) > -\infty$, i.e. $\mathsf{E}\, f(\theta)^- < \infty$. Consequently both expectations considered in (6) are under (A5) well-defined integrals with values in $(-\infty, \infty]$ and (6) implies that $\mathsf{E}\, f(\theta_0)$ is finite. Analogically if the minimum considered in (A4) is finite then $\mathsf{E}\,\psi(X)$ is well-defined with values in $\overline{\mathbb{R}}$. If both $\mathsf{E}\,\psi(X)^+$ and $\mathsf{E}\,\psi(X)^-$ are finite then

$$\mathsf{E}\,\psi(X) = \mathsf{E}\,\psi(X)^+ - \mathsf{E}\,\psi(X)^-$$

is finite too and the limit relation in (A4) can be reduced to the common form

$$\lim_{n\to\infty} \mathsf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi(X_i) - \mathsf{E}\,\psi(X)\right| > \varepsilon\right) = 0, \quad \varepsilon > 0.$$

Example 3

Let the stationary sequence $X_1, X_2, \ldots$ considered in (A4) be independent or, more generally, ergodic. It is well known that then the last relation of Remark 2 holds as soon as $\mathsf{E}\,\psi(X)$ is finite. The validity of (A4) can be extended to the case $\mathsf{E}\,\psi(X) \in \{-\infty, \infty\}$ considered there as follows. Put for every $c > 0$ and $x \in \mathbb{R}$

$$\psi_+^c(x) = \psi(x)\,\mathbb{I}_{[-\infty,c]}(\psi(x)) \quad \text{and} \quad \psi_-^c(x) = \psi(x)\,\mathbb{I}_{[-c,\infty]}(\psi(x)).$$

Then

$$\frac{1}{n}\sum_{i=1}^{n}\psi_+^c(X_i) \leq \frac{1}{n}\sum_{i=1}^{n}\psi(X_i) \leq \frac{1}{n}\sum_{i=1}^{n}\psi_-^c(X_i)$$

where the left-hand side tends in P to $\mathsf{E}\,\psi_+^c(X)$ when $\mathsf{E}\,\psi_+^c(X)$ is finite and the right-hand side tends in P to $\mathsf{E}\,\psi_-^c(X)$ when $\mathsf{E}\,\psi_-^c(X)$ is finite. The first case takes place iff $\mathsf{E}\,\psi^-(X) > -\infty$ in which case the monotonicity theorem for integrals implies

$$\mathsf{E}\,\psi_+^c(X) \uparrow \mathsf{E}\,\psi(X) = \infty \quad \text{for } c \uparrow \infty.$$

Analogically the second case takes place iff $\mathsf{E}\,\psi(X)^+ < \infty$ in which case

$$\mathsf{E}\,\psi^c_-(X) \downarrow \mathsf{E}\,\psi(X) = -\infty \quad \text{for } c \uparrow \infty.$$

It follows from here that

$$\varphi \circ \frac{1}{n}\sum_{i=1}^{n} \psi(X_i)$$

tends in P to $\mathsf{E}\,\psi(X)$ Q. E. D.

Theorem 3.

Let the observations satisfy (A4), (A5) and let there exist a compact neighborhood $\Theta_0 \subset \Theta$ of θ_0 such that

$$\lim_{n\to\infty} \mathsf{P}(f^0_n(\Theta^c_0) > f^0_n(\theta_0) + \varepsilon) = 1 \quad \text{for some } \varepsilon > 0. \tag{12}$$

Then (6) implies (10), i. e. also the consistency of all AMCE's.

Proof.

Taking into account Theorem 2 and the inequality $f^0_n(S_a) \le f^0_n(\theta_0)$ we see that it suffices to prove that (6) implies (10) with $f^0_n(S_a)$ replaced by $f^0_n(\theta_0)$. Define

$$\Theta_a = \Theta_0 - S_a, \quad a > 0.$$

It follows from (12) that $\Theta_0 \neq \emptyset$. Hence there exists $a_0 > 0$ such that $\Theta_a \neq \emptyset$ for all $a \in (0, a_0)$. Let us consider an arbitrary $a \in (0, a_0)$. It obviously holds

$$S_{a,c} \subset S^c_a = \Theta - S_a = (\Theta^c_0 - S_a) \cup (\Theta_0 - S_a) \subset \Theta^c_0 \cup \Theta_a.$$

Therefore

$$f^0_n(S_{a,c}) \ge \min\left\{f^0_n(\Theta^c_0), f^0_n(\Theta_a)\right\}.$$

It follows from here and from (12) that the desired result will be proved if we prove that

$$\lim_{n\to\infty} \mathsf{P}(f^0_n(\Theta_a) > f^0_n(\theta_0) + \varepsilon_a) = 1 \quad \text{for some } \varepsilon_a > 0. \tag{13}$$

To this end consider the function

$$\Phi(\theta) = \mathsf{E}\, f(\theta), \quad \theta \in \Theta.$$

By Remark 2 this function takes on values in $(-\infty, \infty]$ and (6) implies that

$$\Delta\,\Phi(\theta) = \Phi(\theta) - \Phi(\theta_0)$$

is positive on Θ_a. By (A2) and Fatou's lemma

$$\liminf_{k\to\infty} \Phi(\theta_k) \ge \mathsf{E}\,\liminf f(\theta_k) \ge \mathsf{E}\, f(\theta_*) = \Phi(\theta_*)$$

for a sequence $\theta_k \in \Theta$ converging to $\theta_* \in \Theta$. Therefore Φ and also $\Delta\Phi$ are lower semicontinuous on Θ. Since Θ_ϵ is compact, the assumption

$$\inf_{\theta\in\Theta_\epsilon} \Delta\Phi(\theta) = 0$$

implies the existence of a sequence $\theta_k \in \Theta_\epsilon$ converging to some $\theta_* \in \Theta_\epsilon$ such that

$$\lim_{k\to\infty} \Delta\Phi(\theta_k) = 0.$$

This and the lower semicontinuity of $\Delta\Phi$ leads to $\Delta\Phi(\theta_*) = 0$ which contradicts the positivity of $\Delta\Phi$ on Θ_ϵ. Consequently we proved that for every $\theta \in \Theta$ it holds

$$\Phi(\theta) \geq \Phi(\theta_0) + 2\varepsilon_\epsilon, \quad \text{where } \varepsilon_\epsilon = \frac{1}{2}\inf_{\theta\in\Theta_\epsilon} \Delta\Phi(\theta) > 0.$$

Let us now consider an arbitrary $\theta \in \Theta$ and a sequence of open monotonically shrinking neighborhoods $B_\theta^k \subset \Theta$ of θ (i.e. $B_\theta^1 \supset B_\theta^2 \supset \ldots$ and $\bigcap_k B_\theta^k = \{\theta\}$). By (A2) it holds

$$\liminf_{k\to\infty} f(\theta_k) \geq f(\theta) \quad P\text{-a.s.}$$

for every sequence $\theta_k \in B_\theta^k$. This implies

$$\liminf_{k\to\infty} f(B_\theta^k) \geq f(\theta) \quad P\text{-a.s.}$$

On the other hand the monotonicity of the sequence $B_{\theta k}$ implies

$$f(B_\theta^1) \leq f(B_\theta^2) \leq \cdots \leq f(\theta).$$

Therefore it holds $f(B_\theta^k) \uparrow f(\theta)$ P-a.s. when $k \to \infty$. It follows from here, from (A5) and from the monotone convergence theorem for integrals that

$$\lim_{k\to\infty} \mathsf{E}\, f(B_\theta^k) = \Phi(\theta).$$

Hence there exist an open neighborhood B_θ of θ such that

$$\mathsf{E}\, f(B_\theta) > \Phi(\theta) - \varepsilon_\epsilon \geq \Phi(\theta_0) + \varepsilon_\epsilon, \quad \theta \in \Theta_\epsilon.$$

Since $\{B_\theta : \theta \in \Theta_\epsilon\}$ is a covering of the compact Θ_ϵ, there exists a finite subcovering $\{B_{\theta_1}, \ldots, B_{\theta_m}\}$. Put

$$\Theta^j = B_{\theta_j} \cap \Theta_\epsilon.$$

It holds for every $\Theta^* \subset \Theta$

$$f_n(\Theta^*) \geq \frac{1}{n}\sum_{i=1}^{n} f(X_i, \Theta^*).$$

Thus it holds

$$f_n(\Theta_a) = \min_{1\le j\le m} f_n(\Theta^j) \ge \min_{1\le j\le m} \frac{1}{n}\sum_{i=1}^{n} f(X_i,\Theta^j).$$

By the law of large numbers assumed in (A4) it holds for every $1 \le j \le m$

$$\lim_{n\to\infty} \mathsf{P}\left(\varphi \circ \frac{1}{n}\sum_{i=1}^{n} f(X_i,\Theta^j) > \varphi \circ \mathsf{E}\, f(X,\Theta^j) - \varepsilon_a^*\right) = 1$$

where $\varepsilon_a^* > 0$ is defined by the condition that, for some $\varepsilon_a^0 > 0$,

$$\varphi \circ (\Phi(\theta_0) + \varepsilon_a) - \varphi \circ \Phi(\theta_0) - \varepsilon_a^* = \varepsilon_a^0 > 0.$$

Since at the same time

$$\mathsf{E}\, f(X,\Theta^j) \ge \mathsf{E}\, f(X,B_{\theta_j}) \ge \Phi(\theta_0) + \varepsilon_a, \quad 1 \le j \le m,$$

it holds

$$\lim_{n\to\infty} \mathsf{P}\left(\min_{1\le j\le m} \varphi \circ \frac{1}{n}\sum_{i=1}^{n} f(X_i,\Theta^j) > \varphi \circ \Phi(\theta_0) + \varepsilon_a^0\right) = 1.$$

It is clear from these results that (13) holds   Q. E. D.

If the observations $X_1, X_2, \ldots$ are i. i. d. and the parameter space Θ is compact then Theorem 3 reduces to Theorem 1.12 of Pfanzagl (1969) with strong consistency replaced by the weak.

Theorem 4

Let the observations satisfy (A4), (A5). If there exists $a > 0$, $y < \mathsf{E}\, f(\theta_0)$ and an estimator with $\theta_n^* \in S_{a,c}$ P-a. s. for all $n \in \mathbb{N}$ such that

$$\limsup_{n\to\infty} \mathsf{P}(f_n(\theta_n^*) < y) > 0 \tag{14}$$

then (11) does not hold, so that all AMCE's are incosistent.

Proof.

It holds

$$f^0(S_a) = \varphi \circ f_n(S_a) \ge \varphi \circ \frac{1}{n}\sum_{i=1}^{n} f(X_i,S_a).$$

By the law of large numbers assumed in (A4), the right-hand expression tends in probability to $\varphi \circ \mathsf{E}\, f(X,S_a)$. By the same argument as in the previous proof,

$$\lim_{a\downarrow 0} \mathsf{E}\, f(X,S_a) = \mathsf{E}\, f(X,\theta_0).$$

Therefore if (14) holds then (11) cannot be satisfied   Q. E. D.

The next two examples and one corollary illustrate the applicability of Theorems 3 and 4. For simplicity we restrict ourselves to i. i. d. observations.

Example 4

Consider the statistical model of Example 1, under the additional assumption that, for all $\theta \in \Theta$, the first moment of distribution P_θ is θ, and the second moment is finite. The least square error contrast function satisfies in this case the relation

$$\Phi(\theta) = \mathsf{E}(X - \theta)^2 = \Phi(\theta_0) + (\theta - \theta_0)^2, \qquad \theta \in \Theta,$$

so that (6) holds. However, the compactness assumption in Theorem 3 does not allow to establish the consistency of the sample mean in the case $\Theta = \mathbf{R}$ by using Theorem 3. We shall return to this Example in the next section where the compactness will be replaced by a weaker assumption satisfied by this model.

Example 5

The assumption (6) alone is not sufficient for (10). Indeed, in the statistical model of Example 2, the functions $\Phi(\theta) = \mathsf{E}\, f(X, \theta)$ can be explicitely evaluated for all contrast α-functions, $\alpha \in [0, 1)$, introduced there. It turns out that, for all α under consideration, $\Phi(\theta)$ is continuous increasing in the interval A_{θ_0}, decreasing in the interval A_{t_0} where t_0 is the solution of the equation

$$t + e^{-\psi(t)} = \theta_0$$

in the domain $t \geq 0$, and constant in the remaining two intervals $[0, t_0]$ and $\left[\theta_0 + e^{-\psi(\theta_0)}, 1\right)$. Thus (6) holds. Nevertheless, as follows from the next result, the corresponding AMCE's are inconsistent.

Notice that the following condition (15) is stronger than the condition (7) in Example 2.

Corollary

If in the statistical model of Example 2

$$\psi(\theta) > \frac{1}{4(1 - \theta)^2}, \qquad \theta \in (0, 1), \tag{15}$$

then no contrast α-function, $\alpha \in [0, 1)$, satisfies (11), i.e. all corresponding AMCE's are inconsistent.

Proof

We shall prove that the estimator θ_n^* satisfying (14) is the n-th order statistics $X_{(n)}$. Since for every $\alpha \in (0, 1)$

$$\frac{1 - x^\alpha}{\alpha} \leq -\ln x, \qquad x > 0,$$

it suffices to prove

$$\lim_{n\to\infty} \mathsf{P}\left(-\frac{1}{n}\sum_{i=1}^{n} \ln p_{X_{(n)}}(X_i) < y\right) = 1, \qquad y < 0. \tag{16}$$

By the definition of $p_\theta(x)$ in Example 2, it holds

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \ln p_{X_{(n)}}(X_i) &= \frac{n-1}{n}\ln(1-c) + \frac{1}{n}\ln\left(1-c+c\,e^{\psi(X_{(n)})}\right)\\ &\geq \ln(1-c) + \frac{1}{n}\psi(X_{(n)}).\end{aligned}$$

Hence for $x = -y - \ln(1-c) > 0$

$$\begin{aligned}\mathsf{P}\left(-\frac{1}{n}\sum_{i=1}^{n} \ln p_{X_{(n)}}(X_i) < y\right) &\geq \mathsf{P}\left(\frac{1}{n}\psi(X_{(n)}) > x\right)\\ &= \mathsf{P}\left(X_{(n)} > \psi^{-1}(n\,x)\right).\end{aligned}$$

Consider now a probability measure $Q \ll \mu$ on $\mathbf{R}$ with the uniform density

$$q(x) = \left(\frac{dQ}{d\mu}\right)(x) = \mathbb{I}_{[-c/(1-c),1]}(x), \qquad x \in \mathbf{R}.$$

Denote by $\mathbf{Y} = (Y_1, \ldots, Y_n)$ the random vector obtained by replacing P_{θ_0} in the sample probability space by Q. It holds for every $z > 0$

$$\mathsf{P}\left(X_{(n)} > z\right) \geq \mathsf{P}\left(Y_{(n)} > z\right) = 1 - [z(1-e)+c]^n.$$

Therefore

$$\mathsf{P}\left(-\frac{1}{n}\sum_{i=1}^{n} p_{X_{(n)}}(X_i) < y\right) \geq 1 - \left[\psi^{-1}(n\,x)\,(1-c)+c\right]^n.$$

But

$$\psi^{-1}(n\,x)\,(1-c)+c = 1-(1-c)\,\left(1-\psi^{-1}(n\,x)\right)$$

and, by (15),

$$1-\psi^{-1}(n\,x) > \frac{1}{2\sqrt{nx}}.$$

Thus

$$\psi^{-1}(n\,x)\,(1-c)+c < 1 - \frac{1-c}{2\sqrt{nx}}$$

and, consequently,

$$\lim_{n\to\infty}\left[\psi^{-1}(n\,x)\,(1-c)+c\right]^n = 0.$$

This implies (16) Q. E. D.

6 Applications to M-estimators

Let us consider a topological space $\mathcal{R}$ of possible values of regressors and denote by $\mathcal{B}_{\mathcal{R}}$ and $\mathcal{B}_{\mathbf{R}}$ the Borel σ-fields of $\mathcal{R}$ and $\mathbf{R}$. Let $R_1, R_2, \ldots$ or $E_1, E_2, \ldots$ be mutually independent stationary and ergodic sequences of $\mathcal{R}$-valued or $\mathbf{R}$-valued random variables defined ou the basic probability space $(\Omega, \mathcal{S}, \mathrm{P})$. Finally, let us consider the following general regression model

$$Y_i = g(R_i, \theta_0) + E_i, \qquad i \in \mathbf{N}, \tag{17}$$

where θ_0 is an unknown parameter from Θ defined in Sec. 2 and $g : \mathcal{R} \otimes \Theta \to \mathbf{R}$ an arbitrary continuous function.

This *nonlinear regression model* is a particular case of the general statistical model of Sec. 2 with

$$(\mathcal{X}, \mathcal{A}) = (\mathbf{R} \otimes \mathcal{R},\ \mathcal{B}_{\mathbf{R}} \otimes \mathcal{B}_{\mathcal{R}}), \qquad X_i = (Y_i, R_i), \qquad i \in \mathbf{N},$$

and

$$P_{\theta,n}(A) = \int_A dP_n(\mathbf{x} - \mathbf{g}(\mathbf{r}, \theta))\, dQ_n(\mathbf{r}), \quad A \in \mathcal{A}^n,\ \theta \in \theta, \tag{18}$$

where P_n and Q_n are sample distributions of $\mathbf{E} = (E_1, \ldots, E_n)$ and $\mathbf{R} = (R_1, \ldots, R_n)$ on $(\mathbf{R}^m \mathcal{B}^n_{\mathbf{R}})$ and $(\mathcal{R}^n, \mathcal{B}^n_{\mathcal{R}})$ and

$$\mathbf{g}(\mathbf{r}, \theta) = (g(r_1, \theta), \ldots, g(r_n, \theta)), \quad \mathbf{r} = (r_1, \ldots, r_n) \in \mathcal{R}^n,\ \theta \in \Theta.$$

The observations $\mathbf{X} = (X_1, \ldots, X_n) = ((Y_1, R_1), \ldots, (Y_n, R_n))$ are obviously satisfying (A4) and (A5) (cf. Example 3). In particular, the marginal probability measure $P_\theta = P_{\theta,1}$ induced by the random variable $X = (Y, R) = (Y_1, R_1)$ on the marginal observation space $(\mathcal{X}, \mathcal{A})$ is given by the formula

$$P_\theta(A) = \int_A dP(x - g(r, \theta))\, dQ(r), \quad A \in \mathcal{A},\ \theta \in \Theta \tag{19}$$

for $P = P_1$ and $Q = Q_1$ introduced above. If both the errors $E_1, E_2, \ldots$ as well as the regressors $R_1, R_2, \ldots$ are i. i. d. then (2) holds for P_θ defined by (19).

If P_n and the distribution $P_{n,\mathbf{x}}$ defined on $(\mathbf{R}^n, \mathcal{B}^n_{\mathbf{R}})$ by

$$P_{n,\mathbf{x}}(S) = P_n(\{\mathbf{x} + \mathbf{y} :\ \mathbf{y} \in S\}$$

are measurre-theoretically equivalent for every $\mathbf{x} \in \mathbf{R}^n$, then (A3) holds as well. Let us restrict ourselves in the sequel to the case where either the equivalence between P_n and $P_{n,\mathbf{x}}$ takes place, or where all $P_{n,\mathbf{x}}$, $\mathbf{x} \in \mathbf{R}^n$, are dominated by σ-finite measure μ_n. In this case all $P_{n,\theta}$, $\theta \in \Theta$, are dominated by $\mu_n \otimes Q_n$ and, in view of Remark 1, (A3) may be considered satisfied as well.

Consider a continuous function $\rho: \mathbf{R} \to [0,\infty)$, and put

$$f(x,\theta) = \rho(y - g(r,\theta)), \qquad x = (y,r) \in \mathcal{X}, \qquad \theta \in \Theta. \tag{20}$$

Then, for every $n \in \mathbb{N}$, $f_n(\mathrm{x},\theta)$ is continuous on $\mathcal{X}^n \odot \Theta$. Consequently (A1) and (A2) hold so that (20) defines a contrast function.

The MCE for the contrast function (20) is called an *M-estimator* (ME), and the AMCE for (20) is called an *asymptotic* (or *approximate*) *M-estimator* (AME).

In Example 1 one can find in fact two M-estimators for the linear regression function $g(r,\theta) = \theta$, namely the *least square error estimator* defined by $M(x) = x^2$, and the *least absolute error estimator* defined by $M(x) = |x|$. Both these M-estimators can be applied to an arbitrary nonlinear regression model under consideration. Other theoretically and practically interesting examples of M-estimators of the present type can be found e.g. in Jurečková (1989).

Replacing the general $f_n(\theta)$ in (8),(10),(11) by

$$f_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i - g(R_i,\theta)) = \frac{1}{n}\sum_{i=1}^{n} \rho(E_i - \Delta g(R_i,\theta)), \tag{21}$$

where

$$\Delta g(r,\theta) = g(r,\theta) - g(r,\theta_0), \quad r \in \mathcal{R},\ \theta \in \Theta, \tag{22}$$

one easily obtains corollaries to Theorems 1 and 2 presenting a necessary and sufficient condition for consistency of all AME's and a necessary condition for consistency of at least one AME. We are now goging to formulate a corollary to Theorem 3.

First of all, notice that in the present model the identifiability property (6) reduces to

$$\mathsf{E}\,\rho(E - \Delta g(R,\theta)) > \mathsf{E}\,\rho(E), \quad \theta \in \Theta,\ \theta \neq \theta_0. \tag{23}$$

Further, restrict ourselves in the sequel to ρ nondecreasing in the domain $[0,\infty)$ and nonincreasing in $(-\infty,0)$. Such function ρ are typical for applications. Define $\rho(\infty)$, $\rho(-\infty)$ as the corresponding limits and consider a continuous function $\overline{\rho}: [0,\infty] \to [0,\infty]$ defined by

$$\overline{\rho}(y) = \min\{\rho(y), \rho(-y)\}.$$

Clearly, if $\overline{\rho} = \infty$ and $\mathsf{E}\,\rho(E) < \infty$ then

$$\frac{\mathsf{E}\,\rho(E)}{\overline{\rho}(y)} \downarrow 0 \quad \text{and also } \mathsf{P}(|E| > y) \downarrow 0 \quad \text{as } y \uparrow \infty.$$

Hence for every $\delta > 0$ there exist $y > 0$ and $\varepsilon > 0$ such that

$$\frac{2\varepsilon + \mathsf{E}\,\rho(E)}{\overline{\rho}(y)} + \mathsf{P}(|E| > y) < \delta. \tag{24}$$

Finally, consider the random field

$$Z_n(y,\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{[2y,\infty)}(|\Delta g(R_i,\theta)|), \quad y > 0,\ \theta \in \Theta,$$

where Δg is defined by (22).

Theorem 4

Let ρ be as above with $\bar{\rho} = \infty$ and let there exist a compact neighborhood $\Theta_0 \subset \Theta$ of θ_0 and $\delta > 0$ such that

$$Z_n^0(y) = \inf_{\theta \notin \Theta_0} Z_n(y, \theta)$$

satisfies the condition

$$\lim_{n \to \infty} \mathsf{P}(Z_n^0(y) > \delta) = 1, \quad y > 0. \tag{25}$$

Then the identifiability (23) implies (10), i.e. also the consistency of all AME's.

Proof.

In view of Theorem 3 it suffices to prove (12). To this end take into account that for every $y_1, y_2 \in \mathbb{R}$

$$\rho(y_1 - y_2) \geq \begin{cases} \bar{\rho}(y) & \text{if } |y_2| > 2y,\ |y_1| \leq y \\ 0 & \text{otherwise.} \end{cases}$$

It follows from here for every $y > 0$ and $\theta \in \Theta$

$$\begin{aligned} f_n(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \rho(E_i - \Delta g(R_i, \theta)) \qquad \text{(cf. (21))} \\ &\geq \bar{\rho}(y) \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{(2y,\infty)}(|\Delta g(R_i, \theta)|)\, \mathbf{I}_{[0,y]}(|E_i|) \\ &\geq \bar{\rho}(y) [Z_n(y, \theta) - W_n(y)], \end{aligned}$$

where $Z_n(y, \theta)$ is the above defined random field and

$$W_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{(y,\infty)}(|E_i|).$$

Therefore the infimum $f_n(\Theta_0^c)$ of $f_n(\theta)$ for $\theta \notin \Theta_0$ satisfies the relation

$$f_n(\Theta_0^c) \geq \bar{\rho}(y) [Z_n^0(y) - W_n(y)], \quad y > 0.$$

It follows from here, from (25) and from the fact that for every $y > 0$

$$\lim_{n \to \infty} \mathsf{P}(|W_n(y) - \mathsf{P}(|E| > y)| \leq \varepsilon) = 1, \quad \varepsilon > 0, \qquad \text{(cf. (A4))}$$

that (12) will be proved if we prove that there exist $y > 0$ and $\varepsilon > 0$ for which

$$\lim_{n \to \infty} \mathsf{P}(\bar{\rho}(y) [\delta - \mathsf{P}(|E| > y)] > f_n(\theta_0) + \varepsilon) = 1.$$

By (23) it holds $\mathsf{E}\, \rho(E) < \infty$ which implies that there exist $y > 0$ and $\varepsilon > 0$ satisfying (24). But (24) is equivalent with the inequality

$$\bar{\rho}(y) [\delta - \mathsf{P}(|E| > y)] \geq \mathsf{E}\, \rho(E) + 2\varepsilon$$

so that it suffices to prove

$$\lim_{n\to\infty} \mathsf{P}(\mathsf{E}\,\rho(E)+\varepsilon > f_n(\theta_0)) = 1.$$

This relation is clear from the identity

$$f_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\rho(E_i) \quad \text{(cf. (21))}$$

and from the law of large numbers (cf. (A4)) Q. E. D.

For compact Θ it follows from Theorem 4 that the identifiability (23) above already implies the consistency of all AME's. It follows from Theorem 3 and from the proof of Theorem 4 that in this case the additional restrictions on ρ considered in Theorem 4 are in fact superfluous. Thus for i. i. d. errors $E_1, E_2, \ldots$ this result may be considered as an alternative to the results of Jennrich (1969), Richardson and Bhattacharyya (1986, 1987), and others cited there, concerning consistency of M-estimators of parameters of nonlinear regression for the case where the regressors are chosen at random, in a stationary ergodic manner (noncompact Θ considered in these papers were in fact obtained only by a compactification). A rigorous extension of consistency of general M-estimators to typical noncompact parameter spaces appeared first in Chen and Wu (1988). Relation of our results to this paper will be discussed later (cf. Example 6 below).

Example 5

Let $\rho(x) = x^2$, $\mathsf{E}\,E = 0$, and $\mathsf{E}\,E^2 = \sigma^2 < \infty$. Then the identifiability reduces to the condition

$$\mathsf{P}(|\Delta g(R,\theta)| > 0) > 0 \quad \theta\in\Theta,\ \theta\neq\theta_0.$$

It is interesting that if Θ is not compact then a stronger condition on the complement Θ_0^c, namely

$$\mathsf{P}(|\Delta g(R,\theta)| > y) > \delta \quad \theta\in\Theta_0^c, \text{ for some } \delta>0 \text{ and all } y>0,$$

is still too weak to imply (25). It implies only

$$\lim_{n\to\infty} \mathsf{P}(Z_n(y,\theta) > \delta) = 1, \quad \theta\in\Theta_0^c, \text{ for some } \delta>0 \text{ and all } y>0. \tag{26}$$

However in some special cases, e. g. in the linear case considered in Example 6 below, (26) implies (25).

Now we formulate an alternative of Theorem 4 for the case where $\bar{\rho}(\infty) < \infty$.

Theorem 5

Let ρ be as above with $\mathsf{E}\,\rho(E) < \overline{\rho}(\infty) < \infty$ and let there exist a compact neighborhood $\Theta_0 \subset \Theta$ of θ_0 such that $Z_n^0(y)$ defined in Theorem 4 satisfies the condition

$$\lim_{n\to\infty} \mathsf{P}(Z_n^0(y) > \delta) = 1, \quad y > 0,\ 0 < \delta < 1. \tag{27}$$

Then the identifiability (23) implies (10), i.e. also the consistency of all AME's.

Proof.

The only point which differs from the previous case is that for $\overline{\rho}(\infty) < \infty$ one cannot prove (24) for every $\delta > 0$. Indeed, in this case it holds

$$\frac{\mathsf{E}\rho(E)}{\overline{\rho}(y)} \downarrow \frac{\mathsf{E}\rho(E)}{\overline{\rho}(\infty)} \quad \text{as } y \uparrow \infty, \text{ where } 0 \le \frac{\mathsf{E}\rho(E)}{\overline{\rho}(\infty)} < 1,$$

so that (24) can be established only for

$$\frac{\mathsf{E}\rho(E)}{\overline{\rho}(\infty)} < \delta < 1.$$

Thus, one can assert only that there exist $y > 0$, $\varepsilon > 0$ and $0 < \delta < 1$ such that (24) holds. However, if the stronger condition (27) replaces (25) then the proof can be carried out exactly the same way as in the previous case.

In Theorem 4 and 5 we have used an idea of Chen and Wu (1988). They considered the linear regression considered in the next example with i.i.d. errors and i.i.d. random regressors, and were able to establish consistency results for noncompact parameter spaces Θ by using similar relations as (24). In the framework of the general sufficient condition formulated by Theorem 3, their idea finds a prepared general context and can thus be presented with an extremal simplicity and transparency, and with a greater universality.

Note also that the results for functions ρ considered in Theorems 4 and 5 seem to be much more valuable than the recent results of Bai, Rao and Wu (1992) formulated for convex functions ρ since the nonconvex functions ρ are quite common in the theory as well as in applications (cf. e.g. Jurečková (1989)).

Example 6

Put $\Theta = \mathbb{R}^{p+1}$, $\mathcal{R} = \mathbb{R}^p$, and consider the points $\theta \in \mathbb{R}^{p+1}$ and $r \in \mathbb{R}^p$ as row vectors. Let $g(r,\theta)$ be linear in the sense

$$g(r,\theta) = \alpha + r\beta,$$

where α denotes the first coordinate of θ and β the transposed vector of the remaining coordinates (i.e. $r\,\beta$ denotes the scalar product of vectors r and β). In this case it suffices to consider θ_0 ad the zero vector and the closed spheres Θ_0 centered at zero. The linearity allows to replace the minimization over Θ_0^c in the definition of $Z_n^0(y)$ by the minimization over the surface of unit sphere. Since this surface is compact, (25) is in this case equivalent with (26). The condition (25) can thus be replaced by

$$\mathsf{P}(\alpha + \beta R \neq 0) > 0 \quad \text{for } (\alpha, \beta) \neq 0$$

and the condition (27) by

$$\mathsf{P}(\alpha + \beta R \neq 0) = 1 \quad \text{for } (\alpha, \beta) \neq 0.$$

These conditions are figuring as (2.1) and (2.2) in Theorem 1 of Chen and Wu (1988). Thus in the linear case our Theorems 4 and 5 reduce to a generalization of that Theorem to stationary and ergodic errors and regressors.

References

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

Billingsley P. (1966) Ergodic Theory and Information, Wiley, New York.

Borges R. (1966) Zur Existenz von separabeln stochastischen Prozessen. Zeitschr. Wahrsch. verw. Geb. 6: 125–128

Chen X. R., Wu Y. H. (1988) Strong consistency of M-estimates in linear models, Jour. Multivar. Analysis 27, 116–130.

Doob J. L. (1953) Stochastic Processes. Wiley, New York

Gikhman I. I., Skorokhod A. V. (1971) Theory of Random Processes (in Russian), Vol. I. Nauka, Moskva

Halmos P. R. (1964) Measure Theory. Academic Press, New York

Jennrich R. I. (1969) Asymptotic properties of non-linear least squares estimators. Ann. Math. Statist. 40: 633-643

Jurečková J. (1989) Consistency of M-estimators in linear model generated by a non-monotone and discontinuous ψ-functions. Probability and Statistics 10: 1–10

Lehman E. L. (1983) Theory of Point Estimation, Wiley, New York

Perlman M. D. (1972) On the strong consistency of approximate maximum likelihood estimators. Proc. IV-th Berkeley Symp. Prob. Math. Statist., pp 263–281

Pfanzagl J. (1969) On measurability and consistency of minimum contrast estimators. Metrika 14: 249-272

Richardson G. D., Bhattacharyya B. B. (1986) Consistent estimators in nonlinear regression for a noncompact parameter space. Ann. Statist. 14: 1591–1596

Richarson G. D., Bhattacharyya B. B. (1987) Consistent L_1-estimators in nonlinear regression for a noncompact parameter space, Sankhya 49, Ser. A, 377–398.

Strasser H. (1981) Consistency of maximum likelihood and Bayes estimates. Ann. S-tatist. 9: 1107–1113

Vajda I. (1984) Minimum divergence principle in statistical estimation. Statistics and Decisions, Suppl. Issue No. 1, pp 239–262

Vajda I. (1992) Conditions equivalent with consistency of approximate maximum likelihood estimates (submitted)