Consistency of M estimates for separable nonlinear regression models

María Victoria Fasano[1] (virfeather@yahoo.com.ar)

Ricardo Maronna[1] (rmaronna@retina.ar)

[1]Departamento de Matemática, Facultad de Ciencias Exactas,

Universidad de La Plata, C.C. 172, La Plata 1900, Argentina.

Abstract

Consider a nonlinear regression model : $y_i = g(\mathbf{x}_i, \theta) + e_i$, $i = 1, ..., n$, where the $\mathbf{x}_i$ are random predictors $\mathbf{x}_i$ and θ is the unknown parameter vector ranging in a set $\mathbf{\Theta} \subset R^p$. All known results on the consistency of the least squares estimator and in general of M estimators assume that either $\mathbf{\Theta}$ is compact or g is bounded, which excludes frequently employed models such as the Michaelis-Menten, logistic growth and exponential decay models. In this article we deal with the so-called *separable* models, where $p = p_1 + p_2$, $\theta = (\alpha, \beta)$ with $\alpha \in A \subset R^{p_1}$, $\beta \in B \subset R^{p_2}$, and g has the form $g(\mathbf{x}, \theta) = \beta^T \mathbf{h}(\mathbf{x}, \alpha)$ where $\mathbf{h}$ is a function with values in R^{p_2}. We prove the strong consistency of M estimators under very general assumptions, assuming that $\mathbf{h}$ is a bounded function of α, which includes the three models mentioned above.

Key words and phrases: Nonlinear regression, separable models, consistency, robust estimation.

1 Introduction

Consider i.i.d. observations $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, given by the nonlinear model with random predictors:

$$y_i = g(\mathbf{x}_i, \theta_0) + e_i, \tag{1}$$

where $\mathbf{x}_i \in R^q$ and e_i are independent, and the unknown parameter vector θ_0 ranges in a set $\Theta \subset R^p$. An important case, usually called *separable*, are models where $p = p_1 + p_2$ and $\theta_0 = (\alpha_0, \beta_0)$ with $\alpha_0 \in A \subset R^{p_1}$ and $\beta_0 \in B \subset R^{p_2}$, and g of the form

$$g(\mathbf{x}, \theta) = g(\mathbf{x}, \alpha, \beta) = \sum_{j=1}^{p_2} \beta_j h_j(\mathbf{x}, \alpha), \tag{2}$$

where h_j $(j = 1, ..., p_2)$ are functions of $X \times R^{p_2} \to R$. Usually B is the whole of R^{p_2} or an unbounded subset of it. Examples are the Michaelis-Menten model, with

$$p_1 = p_2 = q = 1,\ x \geq 0,\ \alpha, \beta > 0,\ h(x, \alpha) = \frac{x}{x + \alpha}, \tag{3}$$

the logistic growth model, with

$$q = 1,\ p2 = 1,\ \ p_2 = 1,\ x \geq 0,\ \alpha_j > 0,\ \beta > 0,\ h(\mathbf{x}, \alpha) = \frac{e^{\alpha_2 x}}{1 + \alpha_1 (e^{\alpha_2 x} - 1)}, \tag{4}$$

the exponential decay model, with

$$q = 1,\ p_2 = p_1 + 1,\ \ x \geq 0,\ \alpha_j < 0,\ \beta_j \geq 0,\ g(\mathbf{x}, \alpha, \beta) = \beta_0 + \sum_{j=1}^{p_1} \beta_j e^{\alpha_j x}, \tag{5}$$

and the exponential growth model, like (5) but with $\alpha_j > 0$.

The classical least squares estimate (LSE) is given by

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (y_i - g(\mathbf{x}_i, \theta))^2.$$

The consistency of the LSE assuming $\mathrm{E}(e_i) = 0$ and $\mathrm{Var}(e_i) = \sigma^2 < \infty$ has been proved by several authors under the assumption of a compact Θ; in particular Amemiya (1983), Jennrich (1969) and Johansen (1984). Wu (1981) assumes that Θ is a finite set.

Richardson and Bhattacharyya (1986) do not require the compactness of Θ, but they assume $g(\mathbf{x}, \theta)$ to be a bounded function of θ, which excludes most separable models.

Shao (1992) showed the consistency of the LSE without requiring the compacity of Θ nor the boundedness of g, but requires assumptions on g that exclude the simplest separable models. For example, in the case $g(x, \theta) = \beta e^{\alpha x}$, for any $x_0 > 0$ one can make $g(x_0, \theta)$ =constant with $\alpha \to -\infty$ and $\beta \to 0$. This fact violates both "Condition 1" and "Condition 2" in page 427 of his paper.

The well-known fact that the LSE is sensitive to outliers has led to the development of *robust estimates* that are simultaneously highly efficient for normal errors and resistant to perturbations of the model. One of the most important families of robust estimates are the *M-estimates* proposed by Huber (1973) for the linear model. For nonlinear models they are defined by

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \rho\left(\frac{y_i - g(\mathbf{x}_i, \theta)}{\widehat{\sigma}}\right), \tag{6}$$

where ρ is a loss function whose properties will be described in the next section and $\widehat{\sigma}$ is an estimate of the error's scale. However, at this stage of our research we deal with the simpler case of known σ. Then it may be assumed without loss of generality that $\sigma = 1$ and therefore we shall deal with estimates of the form

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \rho(y_i - g(\mathbf{x}_i, \theta)). \tag{7}$$

All published results on the consistency of robust estimates for nonlinear

models require the compacity of Θ. Oberhofer (1982) deals with the L_1 estimator. Vainer and Kukush (1998) and Liese and Vajda (2003, 2004) deal with M estimates. The latter deal with $O\left(n^{-1/2}\right)$ consistency and asymptotic normality of M estimates in more general models. Stromberg (1995) proved the consistency of the Least Median of Squares estimate (Rousseeuw, 1984), and Čížek (2005) dealt with the consistency and asymptotic normality of the Least Trimmed Squares estimate. Fasano et al. (2012) study the functionals related to M estimators in linear and nonlinear regression; in the latter case, they also assume a compact $\boldsymbol{\Theta}$.

In this article we will prove the consistency of M estimates for separable models without assuming the compactness of Θ, but assuming the boundedness of the h_js; this case includes the exponential decay, logistic growth and Michaelis-Menten models. It can thus be considered as a generalization of (Richardson and Bhattacharyya, 1986).

2 The assumptions

It will be henceforth assumed that ρ is a "ρ–function" in the sense of (Maronna et al, 2006). i.e., $\rho(u)$ is a continuous nondecreasing function of $|u|$, such that $\rho(0) = 0$ and that if $\rho(u) < \sup_u \rho(u)$ and $0 \leq u < v$ then $\rho(u) < \rho(v)$. We shall consider two cases: unbounded ρ and bounded ρ. The first includes convex function, in particular the LSE with $\rho(x) = x^2$ and the well-known Huber function

$$\rho_k(x) = \begin{cases} x^2 & \text{if} \quad |x| \leq k \\ 2k|x| - k^2 & \text{if} \quad |x| > k \end{cases} \tag{8}$$

and the second includes the bisquare function $\rho(x) = \min\left\{1 - \left(1 - (x/k)^2\right)^3, 1\right\}$, where k is in both cases a constant that controls the estimator's efficiency.

Let $\mathbf{h}(\mathbf{x},\alpha) = (h_1(\mathbf{x},\alpha),...,h_{p_2}(\mathbf{x},\alpha))'$ where in general $\mathbf{a}'$ denotes the transpose of $\mathbf{a}$.The necessary assumptions are:

A B is a closed set such that $t\beta \in B$ for all $\beta \in B$ and $t > 0$.

B $\sup_{\alpha\in A} \mathrm{E}|\rho(y - \beta'\mathbf{h}(\mathbf{x},\alpha))| < \infty$ for all $\beta \in B$.

C The function $\mathrm{E}\rho(e-t)$ –where e denotes any copy of e_i– has a unique minimum at $t = 0$. Put $\lambda_0 = \mathrm{E}\rho(e)$.

D $\mathbf{h}$ is continuous in α a.s. and

$$\alpha \neq \alpha_0 \Rightarrow \sup_{\beta\in B} \mathrm{P}\{\beta'\mathbf{h}(\mathbf{x},\alpha) = \beta_0'\mathbf{h}(\mathbf{x},\alpha_0)\} < 1 \tag{9}$$

E Let $S = \sup_t \rho(t)$ (which may be infinite). Then

$$\delta =: \sup_{\beta\neq\mathbf{0},\ \alpha\in A} \mathbf{P}(\beta'\mathbf{h}(\mathbf{x},\alpha) = 0) < 1 - \frac{\lambda_0}{S}. \tag{10}$$

F Call $\mathcal{U}$ the family of all open neighborhoods of α_0. Then

$$\sup_{\beta} \inf_{U\in\mathcal{U}} \sup_{\alpha\notin U} \mathrm{P}\{\beta'\mathbf{h}(\mathbf{x},\alpha) = \beta_0'\mathbf{h}(\mathbf{x},\alpha_0)\} < 1.$$

G $\mathbf{h}$ is bounded as a function of α, i.e., $\sup_{\alpha\in A} \|\mathbf{h}(\mathbf{x},\alpha)\| < \infty$ a.s.

We now comment on the assumptions.

For (A) to hold in examples (3)-(4)-(5) we must enlarge the range of β_js to $\beta_j \geq 0$. However, to ensure the validity of (D) and (F), it will be assumed that the elements of the "true" vector β_0 are all positive.

If ρ is bounded, (B) holds without further conditions. Sufficient conditions for Huber's ρ and for the LSE are finite moments of e and of $\mathbf{h}(\mathbf{x},\alpha)$, of orders one and two, respectively.

A sufficient condition for (C) is that the distribution of e has an even density $f(u)$ that is nonincreasing for $u \geq 0$ and is decreasing in a neighborhood of $u = 0$ (see Lemma 3.1 of Yohai (1987)). If ρ is strictly convex with a derivative ψ, then a sufficient condition is $\mathrm{E}\psi(e) = 0$, which for the LSE reduces to $\mathrm{E}e = 0$.

Assumption (D) is required for ensure uniqueness of solutions. For examples (3)-(4) it is very easy to verify. For (5) it follows from the well-known linear independence of exponentials.

If $S = \infty$, (E) just means that $\delta < 1$ (since $\lambda_0 < \infty$ by (B)). Otherwise it puts a bound on δ. In our examples we have $\delta = 0$, since $\beta'\mathbf{h} > 0$ if β has a single nonnull (positive) element.

Assumption (F) is required in the case of non-compact A, to prevent the estimator $\widehat{\alpha}$ from "escaping to the border". In our examples the border for the α_js is either zero of infinity, and (F) is easily verified by a detailed but elementary calculation (taking into account the remark above that all elements of β_0 are positive). For example, in (3) it suffices to consider neighborhoods of the form $(\alpha_0/K, K\alpha_0)$ with K sufficiently large.

Finally, (G) is easily verified for models (3)-(4)-(5).

3 The results

For separable models the M-estimate is given by

$$\hat{\theta}_n = \left(\widehat{\alpha}_n, \widehat{\beta}_n\right) = \arg \min_{\alpha \in A,\ \beta \in B} \frac{1}{n} \sum_{i=1}^{n} \rho\left(y_i - \beta'\mathbf{h}\left(x_i, \alpha\right)\right).$$

We now state our main result.

Theorem 1 *Assume model (2) with conditions A-B-C-D-E-F-G. Then the M estimate* $\left(\widehat{\alpha}_n, \widehat{\beta}_n\right)$ *is strongly consistent for* θ_0.

We shall first need an auxiliary result, based on a proof in (Bianco and Yohai, 1996).

Lemma 2 *Assume model (2) with conditions A-B-C-D-E and A* compact. *Then* $\left\|\hat{\beta}_n\right\|$ *is ultimately bounded with probability one.*

Proof of the Lemma: Put

$$\lambda(\alpha,\beta) = \mathrm{E}\rho\left(y - \beta'\mathbf{h}(\mathbf{x},\alpha)\right).$$

It follows from (C) that $\lambda(\alpha,\beta)$ attains its minimum only when $\beta'\mathbf{h}(\mathbf{x},\alpha) = \beta_0'\mathbf{h}(\mathbf{x},\alpha_0)$ a.s. and by (9) this happens when $(\alpha,\beta) = (\alpha_0,\beta_0)$. Therefore

$$(\alpha,\beta) \neq (\alpha_0,\beta_0) \Rightarrow \lambda(\alpha,\beta) > \lambda(\alpha_0,\beta_0) = \lambda_0. \tag{11}$$

Let $\Gamma = \{\gamma \in B : \|\gamma\| = 1\}$. Then we may write $\beta = t\gamma$ with $t = \|\beta\| \in R_+$ and $\gamma \in \Gamma$.

We divide the proof into two cases.

Case I: bounded ρ : Assume that $S = \sup_u \rho(u) < \infty$. To simplify notation it will be assumed without loss of generality that $S = 1$. For each $(\alpha,\gamma) \in A \times \Gamma$ we have

$$\lim_{t\to\infty} \mathrm{E}\rho\left(\mathbf{y} - t\gamma'\mathbf{h}(\mathbf{x},\alpha)\right) \geq 1 - \delta > \lambda_0,$$

where δ is defined in (10). Let

$$\xi = 1 - \delta - \lambda_0 > 0, \quad \varepsilon = \frac{\xi}{4} < \frac{1-\delta}{4}.$$

Since (10) implies that $\mathrm{P}\left(|\gamma'\mathbf{h}(\mathbf{x},\alpha)| > 0\right) \geq 1 - \delta$ for $\gamma \in \Gamma$, then for each $(\alpha,\gamma) \in A \times \Gamma$ there are positive a, b such that

$$\mathrm{P}\left(|y| \leq a, |\gamma'\mathbf{h}(\mathbf{x},\alpha)| \geq b\right) \geq 1 - \delta - \varepsilon. \tag{12}$$

Then by (12) there exists $T > 0$ such that $t > T$ implies

$$\mathrm{E} \inf_{t>T} \rho\left(\mathbf{y} - t\gamma' \mathbf{h}\left(\mathbf{x}, \alpha\right)\right) > 1 - \delta - 2\varepsilon. \tag{13}$$

Therefore (13) implies that for each $(\alpha, \gamma) \in A \times \Gamma$ there exist a neighborhood $U(\alpha, \gamma) \subset A \times \Gamma$ and $T(\alpha, \gamma) \in R_+$ such that

$$\mathrm{E} \inf_{(\alpha_1, \gamma_1) \in U(\alpha, \gamma)} \inf_{t > T(\alpha, \gamma)} \rho\left(\mathbf{y} - t\gamma_1' \mathbf{h}\left(\mathbf{x}, \alpha_1\right)\right) > 1 - \delta - 2\varepsilon = \lambda_0 + \frac{\xi}{2}. \tag{14}$$

The neighborhoods $\{U(\alpha, \gamma) : \alpha \in A, \gamma \in \Gamma\}$ are a covering of the compact set $A \times \Gamma$, and therefore there exists a finite subcovering thereof: $\{U_j = U(\alpha_j, \gamma_j)\}_{j=1}^N$. Let $T_0 = \max_j T(\alpha_j, \gamma_j)$.

We shall show that $\limsup_{n\to\infty} \left\|\hat{\beta}_n\right\| \leq T_0$ a.s. Put for brevity

$$\lambda_n(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^n \rho\left(y_i - \beta' \mathbf{h}\left(x_i, \alpha\right)\right).$$

Then

$$\begin{aligned} \inf_{\|\beta\| > T_0} \inf_{\alpha \in A} \lambda_n(\alpha, \beta) &\geq \frac{1}{n} \sum_{i=1}^n \inf_{\alpha \in A, \gamma \in \Gamma} \inf_{t > T_0} \rho\left(y_i - t\gamma' \mathbf{h}\left(x_i, \alpha\right)\right) \\ &= \min_{j=1,\ldots,N} \frac{1}{n} \sum_{i=1}^n \inf_{(\alpha,\gamma) \in U_j} \inf_{t > T_0} \rho\left(y_i - t\gamma' \mathbf{h}\left(x_i, \alpha\right)\right), \end{aligned}$$

and therefore (14) and the Law of Large Numbers imply

$$\lim \inf_{n\to\infty} \inf_{\|\beta\| > T_0} \inf_{\alpha \in A} \lambda_n(\alpha, \beta) \geq \lambda_0 + \frac{\xi}{2} \text{ a.s.},$$

while

$$\lambda_n\left(\widehat{\alpha}_n, \widehat{\beta}_n\right) = \inf_{\beta \in B} \inf_{\alpha \in A} \lambda_n(\alpha, \beta) \leq \lambda_n(\alpha_0, \beta_0) \to \lambda_0 \text{ a.s.}$$

which shows that ultimately $\left\|\hat{\beta}_n\right\| \leq T_0$ with probability one.

Case II: unbounded ρ : Here an analogous but simpler procedure shows the existence of T_0 and neighborhoods $U(\alpha, \gamma)$ such that the left-hand member of (14) is larger than $2\lambda_0$, and the rest of the proof is similar.■

Proof of the Theorem: If A is not compact, we employ the same approach as in (Richardson and Bhattacharyya, 1986): the Čech-Stone compactification yields a compact set $\widetilde{A} \supset A$ such that each bounded continuous function on A has a unique continuous extension to $\widetilde{A}$. We have to ensure that (B), (D) and (E) continue to hold for $\alpha \in \widetilde{A}$. Since each element of $\widetilde{A}$ is the limit of a sequence of elements of A, (B) and (E) are immediate; and (D) follows from assumption (F). Therefore we can apply the Lemma to conclude that $\left(\widehat{\alpha}_n, \widehat{\beta}_n\right)$ remains ultimately in a compact a.s. The Theorem then follows from Theorem 1 of Huber (1967).■

4 Acknowledgements:

This research was partially supported by grants PID 5505 from CONICET and PICTs 21407 and 00899 from ANPCYT, Argentina.

References

Bianco, A., Yohai, V.J., 1996. Robust estimation in the logistic regression model, in Robust Statistics, Data Analysis and Computer Intensive Methods, Proceedings of the workshop in honor of Peter J. Huber, editor H. Rieder, Lecture Notes in Statistics 109, 17-34 Springer-Verlag, New York.

Čížek, P., 2006. Least trimmed squares in nonlinear regression under dependence. Jr. Statist. Plann. & Inf., 136, 3967-3988.

Fasano, M.V., Maronna, R.A., Sued, M., Yohai, V.J., 2012. Continuity and differentiability of regression M functionals. Bernouilli (to appear).

Huber, P. J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions, in Proceedings of the Fifth Berkeley Symposium

in Mathematical Statistics and Probability, Berkeley: University of California Press, Vol. 1, 221-233.

Jennrich, R. I., 1969. Asymptotic properties of nonlinear least squares estimators. Ann. Math. Statist., 40, 633-643.

Liese, F., Vajda, I., 2003. A general asymptotic theory of M-estimators I. Math. Meth. Statist., 12, 454-477.

Liese, F. Vajda, I., 2004. A general asymptotic theory of M-estimators II. Math. Meth. Statist., 13, 82-95.

Maronna, R.A., Martin, R.D., Yohai, V.J., 2006. Robust Statistics: Theory and Methods, John Wiley and Sons, New York.

Oberhofer, W., 1982. The consistency of nonlinear regression minimizing the L_1 norm. Ann. Statist., 10, 316-319.

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

Rousseeuw, P., 1984. Least median of squares regression. Jr.Amer. Statist. Assoc., 79, 871-880.

Shao, J., 1992. Consistency of Least-Squares Estimator and Its Jackknife Variance Estimator in Nonlinear Models. Can. Jr. Statist., 20, 415-428.

Stromberg, A. J., 1995. Consistency of the least median of squares estimator in nonlinear regression. Commun. Statist.: Th. & Meth., 24, 1971-1984.

Tabatabai M. A.,Argyros I. K., 1993. Robust estimation and testing for general nonlinear regression models. Appl. Math. & Comp., 58, 85-101.

Vainer, B. P., Kukush, A. G., 1998. The consistency of M-estimators constructed from a concave weight function. Th. Prob. & Math. Statist., 57, 11-18.

Wu, C. F., 1981. Asymptotic theory of nonlinear least squares estimation.

Ann. Statist., 9, 501-513.

Yohai, V. J., 1987. High breakdown-point and high efficiency estimates for regression. Ann. Statist., 15, 642-656.