Sankhyā : The Indian Journal of Statistics
2003, Volume 65, Part 2, pp 440-463

STRONG CONSISTENCY OF MINIMUM CONTRAST ESTIMATORS WITH APPLICATIONS

By ARUP BOSE
and
DEBAPRIYA SENGUPTA
Indian Statistical Institute, Kolkata

SUMMARY. We prove a strong consistency result for minimum contrast estimators for general regression problems with independent errors using technically transparent proofs. This unifies the study of strong consistency of least squares estimators in nonlinear regression models and maximum likelihood estimators in generalized linear models. We give new examples from nonlinear regression and generalized linear models where strong consistency can be established from our result. We also demonstrate that in many situations our result is significantly close to the best existing results.

1. Introduction

Let $\{Y_j : j \geq 1\}$ be an independent sequence where $Y_j \sim F_j(y, \theta_0), \theta_0$ belongs to interior of Θ which is a *compact, convex* subset of $I\!R^p$. Let $\rho_j : I\!R \times \Theta \to I\!R, \quad j \geq 1$ be a sequence of measurable maps. A sequence of *minimum contrast estimators* (MCE) is any measurable sequence of estimators $\{\theta_n\}$ which minimizes

$$Q_n(\theta) = \sum_{j=1}^{n} \rho_j(Y_j, \theta). \tag{1.1}$$

The study of MCE and related M-estimators originated in Huber (1964, 1967) while studying robust estimation. An excellent account of related literature can be found in Hampel *et. al.* (1986). For the purpose of the

Paper received November 2000; revised July 2001.
AMS (2000) *subject classification.* Primary 62F12; secondary 62E20, 62J02, 62J12.
Keywords and phrases. Minimum contrast estimator, nonlinear regression, least square estimate, generalized regression, link function, maximum likelihood estimate, law of large numbers, strong consistency.

current article we assume ρ_j's are absolutely continuous with Lipschitz continuous Radon-Nikodym derivatives with respect to θ. Barring few special instances like regression quantiles, which fail to have continuous Radon-Nikodym derivatives, most of the standard contrast functions satisfy this property.

Two important examples of minimum contrast estimators are the least squares estimators in nonlinear regression models and the maximum likelihood estimators in generalized regression models. There has been a volume of research to answer the question of *strong* consistency in these two models.

In the *nonlinear regression model*,

$$Y_j = f(j, \theta) + \epsilon_j, \quad j \geq 1, \tag{1.2}$$

where $\{\epsilon_j, \ \ j \geq 1\}$ is an independent sequence of mean zero random variables. Often $f(j, \theta) = f(x_j, \theta)$ to incorporate regressor variables. The *least squares estimator* (LSE) $\theta_{n,l}$ is any θ which minimizes

$$Q_n(\theta) = \sum_{j=1}^{n} (Y_j - f(j, \theta))^2. \tag{1.3}$$

In the *generalized linear model* with *natural link function*,

$$Y_j \ \text{ has density } \ f(y|x_j'\theta), \tag{1.4}$$

where $f(\cdot \mid \beta)$ is assumed to belong to some exponential family and mean of Y_j equals $x_j^{'}\theta$. The maximum likelihood estimator (MLE) $\theta_{n,m}$ is any θ which minimizes $Q_n(\theta) = \sum_{j=1}^{n} - \log f(Y_j|x_j'\theta)$.

With reference to (1.2), Pfanzagl (1969) studied strong consistency of the MLE of a location parameter when the scales are varying. For proper nonlinear regression problems we refer to works of Lai, Robbins and Wei (1978, 1979), Christopeit and Helmes (1980), Wu (1981), Lai and Wei (1982, 1987), Chen and Wu (1988) and Lai (1994) among others. In these works several aspects of the problem have been studied such as dependent error structures and stochastic regressors.

For generalized regression models of the form (1.4), fairly comprehensive answers related to strong consistency of the MLE can be found in Fahrmeir and Kaufmann (1985) and Chen, Hu and Ying (1999).

The main goal of the current paper is to unify the study of *strong* consistency in (1.2) and (1.4) when the observations are independent and the regressors are nonstochastic. Under this framework we derive strong consistency of MCE with minimal growth/decay restrictions on the regressor

sequence, $\{x_j\}$ and model variances, $\{\mathrm{Var}(Y_j)\}$. Besides, the proof of the main theorem also identifies some necessary logical steps which must be verified to establish any strong consistency result in an abstract sense. The apparent structural simplicity of the proof suggests that our technique could be useful for strong consistency problems in more general situations.

It is quite understandable that a very abstract generalization is not expected to produce the best result in a specific application since it cannot afford to use problem specific information. Nevertheless, we are able to show that in a variety of applications, our technique produces nearly the best existing answer. Below we give a few examples of such situations and also of instances where our results appear to be nontrivial addition to existing strong consistency results.

Our first example is that of the *linear regression* models with *independent* errors. Consider the model $Y_j = x_j'\theta + \epsilon_j$, where $\{\epsilon_j\}$ are independent, $E\epsilon_j = 0$, $E\epsilon_j^2 < \infty$ for all j and $\{x_j\}$ are $p \times 1$ vectors. The best strong consistency result for these models is due to Lai, Robbins and Wei (1979) (LRW).

Let the minimum and the maximum eigenvalues of any matrix A be denoted by $\underline{\lambda}(A)$ and $\overline{\lambda}(A)$ respectively. Also, let $\Gamma_n = \sum_{j=1}^n x_j x_j'$, which we assume to be non-singular for large n. Under the two conditions given below, the LSE is strongly consistent.

LRW 1 $\{\epsilon_j\}$ is such that $\sum c_i\epsilon_i < \infty$ almost surely for all $\{c_i\}$ such that $\sum c_i^2 < \infty$.

LRW 2 $\Gamma_n^{-1} \to 0$.

Note that the first condition is satisfied if $\{\epsilon_i\}$ are independent and $\sup_j E\epsilon_j^2 < \infty$. Let

$$D_n = \mathrm{Diag}(\Gamma_{n,11}, \ldots, \Gamma_{n,pp}) \text{ and } \quad \mu_n = \underline{\lambda}(D_n^{-1/2}\Gamma_n D_n^{-1/2}),$$

where $\Gamma_{n,jj}$'s denote diagonal elements of Γ_n.

To keep matters simple, let us assume that the design matrices $\{\Gamma_n\}$ are *well conditioned* in the sense that

$$\liminf_n \mu_n > 0. \tag{1.5}$$

Let also for some $\alpha > 1$,

$$L(x) = (\log(x))(\log\log(x))^\alpha.$$

Then, our result implies strong consistency of the LSE if

$$\underline{\lambda}(\Gamma_n) \to \infty$$

and

$$\frac{C_n^2 L(\max(C_n^2, \overline{\lambda}(\Gamma_n)))}{\underline{\lambda}(\Gamma_n)} \to 0, \tag{1.6}$$

where $C_n^2 = \max_{1 \le j \le n} E\epsilon_j^2$.

Since $\underline{\lambda}^{-1}(\Gamma_n) = O(\max_{1 \le j \le p} \Gamma_n^{jj})$, where Γ_n^{jj}'s are diagonal elements of Γ_n^{-1}, it may be noted that (1.6) allows for a trade-off between the magnitude of the error moments and the design variables. This is a relatively mild additional assumption compared to the condition of LRW. In many important applications of regression analysis, heteroscedastic models with unbounded regressors are encountered and we hope our result is a nontrivial addition to the available strong consistency results in such scenarios.

It may be noted that (1.6) is also a condition similar to Lai and Wei (1982). They have shown that for linear regressions with *stochastic* regressors, strong consistency holds if $\sup_j E\epsilon_j^2 < \infty$, $\underline{\lambda}(\Gamma_n) \to \infty$ almost surely, and for some $\alpha > 1$

$$\frac{(\log \overline{\lambda}(\Gamma_n))^\alpha}{\underline{\lambda}(\Gamma_n)} \to 0. \tag{1.7}$$

Our next example is on nonlinear regressions. A very general result obtained in this area is that of Wu (1981). As we will see later, the central idea behind our method is based on Wu (1981). As a specific example where Wu (1981) does not apply but our results do, consider the model (1.2) with $f(j, \theta) = \alpha \exp(\beta x_j)$. With an appropriate parameter space and a decay condition on $\{x_j\}$, the least squares estimator is strongly consistent but Wu's conditions are not satisfied. We compare our results with his in detail while discussing Example 2.2.

In a related work, Bai and Wu (1997) obtained a very general result for *weak consistency* of MCE with convex $\{\rho_j\}$-functions in (1.2). It has been demonstrated that for weak consistency, the rate conditions can be relaxed significantly. When specialised to the linear regression case, it reduces to $\Gamma_n^{-1} \to 0$. However, more assumptions are necessary for strong consistency in general.

Our final example is from generalized linear models. Fahrmeir and Kaufmann (1985) (FK) proved results for the strong consistency of the MLE in generalized linear models by techniques suggested by Wu (1981). As a specific case, consider the *Poisson regression* model, $P(Y_j = y) = \exp(\beta_j y - \exp(\beta_j))/y!$, and $\beta_j = x_j'\theta$. Let $\Gamma_n(\theta) = \sum_{j=1}^n \exp(x_j'\theta) x_j x_j'$. Suppose that θ_0 is the true parameter value and $||\theta - \theta_0|| < \delta$ for $\theta \in \Theta$ for some $\delta > 0$ ($||\cdot||$ denotes the Euclidean norm). For the sake of simplicity, suppose that $\{\Gamma_n(\theta_0)\}$ is well conditioned so that (1.5) holds.

The results of FK implies that the MLE in this model is strongly consistent if $\overline{\lambda}(\Gamma_n(\theta_0)) = O(\underline{\lambda}(\Gamma_n(\theta_0))^c)$ for some $c < 2$. On the other hand, our result implies strong consistency if for every $\theta \neq \theta_0 \in \Theta$, we can find $\eta > 0$, such that

$$\frac{\exp\{\eta|x|_n\}\, L(\,\overline{\lambda}(\Gamma_n(\theta_0))\,)}{\max\{\underline{\lambda}(\,\Gamma_n(\theta)\,),\ \underline{\lambda}(\,\Gamma_n(\theta_0)\,)\}} \to 0. \tag{1.8}$$

Here $|x|_n = \max_{1\le i\le n}\ |x_i|$.

In a closely related work, Chen, Hu and Ying (1999) have extended the methods of LRW to generalized linear models with natural link function where $\{x_i\}$ may be adaptive. The estimator β_n of β is obtained by solving

$$\sum_i^n x_i[y_i - \mu(x_i'\beta)] = 0.$$

They prove that if LRW 1 and LRW 2 hold and $\sup_n |x|_n < \infty$, then β_n is strongly consistent.

In Section 2 we give the basic idea, the main results and some examples. In Section 3 we provide the proofs along with a discussion on the basic ideas behind the proofs. An auxiliary result (Lemma 2 in Section 3) on the lower bound of the infimum of a linear plus a quadratic form is a key tool and may be of independent interest.

2. Main results and Examples

Throughout, $||\ \cdot\ ||$ denotes the Euclidean norm. For given sequences $\{a_n\}$ and $\{b_n\}$ of positive real numbers write, $a_n \approx b_n$ if $c\,a_n \le b_n \le C\,a_n$ for two positive constants c and C for sufficiently large n.

We shall assume that the parameter space is compact and convex. If it is not compact, one proves the existence of a sequence of *local minimizers* of $Q_n(\theta)$, which converges to θ_0 almost surely. Under this convention, we can restrict the original parameter space to an arbitrarily small (but fixed *a priori*) neighbourhood of θ_0. We shall adopt this convention whenever necessary, without loss of generality.

From (1.1) let us write

$$\begin{aligned} &Q_n(\theta) - Q_n(\theta_0) \\ &\quad = \ \{\,(Q_n(\theta) - EQ_n(\theta)) - (Q_n(\theta_0) - EQ_n(\theta_0))\} + \{EQ_n(\theta) - EQ_n(\theta_0)\}. \end{aligned}$$

In order to establish strong consistency, one shows that outside any neighbourhood of θ_0, the left side of the above equation is strictly positive. Often, a necessary condition for strong consistency is that the second term must be

large positive whenever $\theta \neq \theta_0$. See for example Theorem 3 of Wu (1981). Then it is necessary to show that the first term on the right side is dominated by the second term of the right side uniformly in θ belonging to the complement of any neighborhood of θ_0. As a consequence, a suitable assumption on the behavior of $E(Q_n(\theta) - Q_n(\theta_0))$ is crucial. We distinguish between *three* different types of behavior of this quantity.

(i) We say that (1.1) has a *parameter independent uniform rate* if for some $\gamma_n \to \infty$ and a suitable distance measure $d(\cdot,\ \cdot)$,

$$E(Q_n(\theta) - Q_n(\theta_0)) \approx \gamma_n\, d(\theta_0, \theta).$$

This encompasses the traditional set up for strong consistency results such as those of Drygas(1976).

(ii) We say that (1.1) has a *parameter independent non-uniform rate* if for a sequence of positive definite matrices Γ_n such that $\underline{\lambda}(\Gamma_n) \to \infty$,

$$E(Q_n(\theta) - Q_n(\theta_0)) \approx (\theta - \theta_0)^{'} \Gamma_n (\theta - \theta_0).$$

(iii) We say that (1.1) has *parameter dependent rate* if $E(Q_n(\theta) - Q_n(\theta_0))$ cannot be approximated by a fixed quadratic form unlike as in (ii) above.

Even though it is possible to state one general result which covers all the three cases, for the sake of technical transparency, results for Case (ii) and Case (iii) are stated separately. Section 2.1 deals with Case (ii) where the main result is Theorem 2.1. This result when specialised to nonlinear regression models takes the form of Theorem 2.2. Examples 2.1 and 2.2 discuss in details respectively, the general linear regression model and a specific nonlinear regression model. The general consistency result for Case (iii) is Theorem 2.3 in Section 2.2 and when specialised to generalised linear models it takes the form given in Theorem 2.4. Example 2.2 is continued in Section 2.2 to show how Case (iii) arises. The proofs are given in Section 3 where we also comment on our method of proof. An auxiliary result on a lower bound of a function which is the sum of a linear and a quadratic form may be of general interest.

2.1 *Models with parameter independent non-uniform rate.* Let us first state the five basic assumptions that we will use in this case.

Assumption A1. There exists $\Psi_j = (\psi_{j1}, \ldots, \psi_{jp}) : I\!R \times \Theta \to I\!R^p$ such that $E_{\theta_0}|\Psi_j(Y_j, \theta)|^2 < \infty$ and, for each $y \in I\!R$, $\theta_1, \theta_2 \in \Theta$, $\theta_\alpha = \alpha\theta_1 + (1-\alpha)\theta_2$,

$$\rho_j(y, \theta_1) - \rho_j(y, \theta_2) = (\theta_1 - \theta_2)' \int_0^1 \Psi_j(y, \theta_\alpha)\, d\alpha,$$

where $\theta_\alpha = \alpha\theta_1 + (1-\alpha)\theta_2$. Define $\bar{\Psi}_j(y, \theta) = \int_0^1 \Psi_j(y, \theta_0 + \alpha(\theta - \theta_0))\, d\alpha = (\bar{\psi}_{j1}, \ldots, \bar{\psi}_{jp})$, $\bar{b}_{jk}(\theta) = E_{\theta_0}\bar{\psi}_{jk}(Y_j, \theta)$, and let $\bar{B}_j(\theta) = (\bar{b}_{j1}(\theta), \ldots, \bar{b}_{jp}(\theta))$.

Assumption A2. For each j, k there exists square integrable random variable M_{jk} such that for some $\sigma > 0$, and for every $\theta, \theta' \in \Theta$, almost surely, $|\{\psi_{jk}(Y_j, \theta) - E_{\theta_0}(\psi_{jk}(Y_j, \theta))\} - \{\psi_{jk}(Y_j, \theta') - E_{\theta_0}(\psi_{jk}(Y_j, \theta'))\}| \le M_{jk}\|\theta - \theta'\|^\sigma$.

Assumption A3. For each n, there exists a $p \times p$ nonnegative definite matrix Γ_n such that uniformly in $\theta \in \Theta$ for sufficiently large n,

$$\sum_{j=1}^{n} \{E_{\theta_0}(\rho_j(Y_j, \theta) - \rho_j(Y_j, \theta_0))\} \ge (\theta - \theta_0)' \Gamma_n (\theta - \theta_0). \tag{2.1}$$

Next define for large n and $1 \le k \le p$,

$$d_{nk}^2 = \sum_{j=1}^{n} \max \left\{ E_{\theta_0} M_{jk}^2, \mathrm{Var}_{\theta_0}(\psi_{jk}(Y_j, \theta_0)) \right\}, \quad D_n = \mathrm{Diag}(d_{n1}^2, \ldots, d_{np}^2). \tag{2.2}$$

Also let, $\gamma_n = \underline{\lambda}(\Gamma_n)$, $\mu_n = \underline{\lambda}(D_n^{-1/2}\Gamma_n D_n^{-1/2})$. Assume the following

Assumption A4. $\gamma_n = \underline{\lambda}(\Gamma_n) \to \infty$ as $n \to \infty$.

Assumption A5. $d_{nk}^2 \to \infty$ as $n \to \infty$ for $1 \le k \le p$, and

$$\omega_n := \max_{1 \le k \le p} L(d_{nk}^2)/\mu_n\gamma_n \to 0,$$

where $L(x) = \log x \{\log\log x\}^\alpha$ for some $\alpha > 1$.

Remark on the assumptions. In A1 we impose restriction on $\{\rho_j\}$ that they must be absolutely continuous but they need not be too smooth. In particular, $\{\rho_j\}$ need not be twice differentiable. In the context of robust estimation it is not uncommon to have contrast functions that are not twice differentiable. For example, one common choice of ρ comes from Huber's ψ, namely, $(\partial/\partial\theta)\rho(y, \theta) = (y - \theta)$ for $|y - \theta| \le c$ and $= c\,\mathrm{sign}(y - \theta)$ otherwise.

The condition A2 guarantees that we are in a suitable Banach space equipped with L_∞ norm where a strong law for independent summands is

available. See Lemma 1. This restriction unfortunately eliminates the scope for direct application of our result to regression quantiles. Nevertheless, L_1-regression can be handled by using our approach in the following way. The contrast function (the *modulus* function) does not satisfy either A1 and A2. But note that for fixed $y \neq 0$, $(|y-t|-|y|)/t$ is a continuous function in t satisfying A2 with square integrable M_{jk}'s. One can use this property to successfully study strong consistency properties of L_1-regression estimators. This will be reported elsewhere.

A3 captures standard unbiasedness assumption (namely, $E\Psi_j(Y_j, \theta_0) = 0$). It also serves to define the key quantity needed for the identifiability condition in A4. This condition is quite straightforward to verify in standard applications. Note that $\{b_j\}$, being expectations, may happen to be smooth even when the $\{\rho_j\}$ are not so. This would ensure the existence of the sequence $\{\Gamma_n\}$, which acts like the *design matrix* if we adopt regression terminology. In a linear regression set-up we can choose $\Gamma_n = \sum_{i=1}^n x_i x_i'$.

The last two assumptions are regarding eigenvalues of essential constructs for this problem, namely, $\{\Gamma_n\}$ and $\{D_n\}$. A4 is a standard identifiability condition stated in terms of Γ_n. For non-linear regression models, A4 is equivalent to $\sum_{i=1}^n (f(j,\theta) - f(j,\theta_0))^2 \to \infty$ for $\theta \neq \theta_0$. Under additional mild restriction on $\{\epsilon_j\}$, it is a necessary condition for strong consistency of MCE (see Wu, 1981). A5 is the most crucial condition which imposes restrictions on the growth of the eigenvalues of $\{\Gamma_n\}$ and elements of $\{D_n\}$.

Computation of the essential rate constants $\{\gamma_n\}$ *and* $\{\mu_n\}$. The computation/estimates of these constants is crucial to verify A5. This can be simplified as follows. For a sequence of $p \times p$ positive definite matrices $\{A_n\}$, we know from the theory of quadratic forms (*cf.* Rao, 1974) that $\overline{\lambda}(A_n) \approx \max_{1\le i\le p} A_{n,ii}$ and $1/\underline{\lambda}(A_n) \approx \max_{1\le i\le p} A_n^{ii}$, where $A_{n,ii}$ and A_n^{ii} are diagonal elements of A_n and A_n^{-1} respectively. Using these, one can check that $1/\gamma_n \approx \max_{1\le i\le p} \Gamma_n^{ii}$. Similarly, it can be shown that $1/\mu_n \approx \max_{1\le i\le p} d_{ni}^2 \Gamma_n^{ii}$.

Note that in view of this, A5 is implied by

$$L\Big(\max_{1\le i\le p} d_{ni}^2\Big)\left(\max_{1\le i\le p} \Gamma_n^{ii}\right)\left(\max_{1\le i\le p} d_{ni}^2 \Gamma_n^{ii}\right) \to 0. \tag{2.3}$$

Hence we can replace A5 by imposing (2.3). This also establishes a close connection between A5 and conditions obtained by LRW and Chen, Hu and Ying (1999).

We now state our general theorem on the consistency of MCE for the parameter independent nonuniform case. The parameter dependent case is

dealt in Theorem 2.3. The proof of this and all other results are given in Section 3.

THEOREM 2.1. *Under* A1–A5, $\theta_n - \theta_0 = O(\omega_n^{1/2})$ *almost surely.*

Let us now show how our results apply to specific models such as nonlinear regression models and generalized linear models.

Application to linear and nonlinear regression models. Consider Model (1.2). We show how our general conditions translate directly into conditions on the functions $f(j,\theta)$ and offer a step-by-step verification of our assumptions in this class of models.

First observe that $\rho_j(Y_j,\theta) = (Y_j - f(j,\theta))^2$. If $f(j,\cdot)$ is once differentiable we can verify A1 and obtain

$$\Psi_j(\theta) = -2(Y_j - f(j,\theta))\, \bigtriangledown f(j,\theta), \tag{2.4}$$

where $\bigtriangledown f(j,\theta) = (f_1(j,\theta),\ldots,f_p(j,\theta))'$, denotes the first derivative of f. In view of (2.4) we make the following assumption.

Assumption S1. $f(j,\theta)$ is continuously differentiable in θ. Moreover, for each j,k, there exists a positive constant $\tau_{jk} > 0$, such that for some $\sigma > 0$ and for every $\theta, \theta^{'} \in \Theta$,

$$|f_k(j,\theta) - f_k(j,\theta^{'})| \leq \tau_{jk}||\theta - \theta^{'}||^{\sigma}.$$

In view of (2.4), a simple algebra shows that A2 will be satisfied in this case with

$$M_{j,k} = 2|\epsilon_j|\tau_{jk}.$$

Next we observe that by Taylor's theorem on the line segment joining θ_0 and θ, $E\,(\rho_j(Y_j,\theta) - \rho_j(Y_j,\theta_0)) = (f(j,\theta) - f(j,\theta_0))^2$
$= (\theta - \theta_0)^{'} \left(\int_0^1 \bigtriangledown f(j,\theta_\alpha)\, d\,\alpha\right) \left(\int_0^1 \bigtriangledown f(j,\theta_\alpha)\, d\,\alpha\right)^{'} (\theta - \theta_0)$. Let us define

$$\Gamma_n(\theta) = \sum_{i=1}^{n} \left(\int_0^1 \bigtriangledown f(j,\theta_\alpha)\, d\,\alpha\right) \left(\int_0^1 \bigtriangledown f(j,\theta_\alpha)\, d\,\alpha\right)^{'}. \tag{2.5}$$

To obtain a parameter independent Γ_n, assume that

Assumption S2.. There exists a sequence of positive definite matrices G_n and for each θ, there exists positive definite $H(\theta)$ such that, as $n \to \infty$,

$$G_n^{-1/2}\Gamma_n(\theta)G_n^{-1/2} \to H(\theta) \quad \text{uniformly in } \theta.$$

Note that $\Gamma_n(\theta)$ is continuous by S1. Thus, $H(\theta)$ is also continuous because of uniform convergence. Define $\Gamma_n = \inf_\theta \underline{\lambda}(H(\theta)G_n)$. Also, let for each k,

$$d_{nk}^2 = 4\sum_{j=1}^{n} \max\left(f_k^2(j,\theta_0), \tau_{jk}^2 \right) E(\epsilon_j^2).$$

Theorem 2.1 translates into the following theorem for nonlinear regression models. We omit its proof.

THEOREM 2.2 (*Nonlinear regression*). Under S1, S2, A4 and A5, the least squares estimator $\theta_{n,l}$ converges to θ_0 almost surely.

We now discuss how this result compares with existing results on nonlinear regression models.

REMARK 2.1. (*Partial comparison with the results of Wu (1981)*). The best strong consistency result in this class of models *with independent errors* appears to be Wu (1981) who proves the strong consistency of the LSE under the following assumptions and also under weaker local versions of these.

Assumption W. (i) For any $\delta > 0$, for some $c > 0$,

$$\limsup_{n\to\infty} \frac{\sum_{j=1}^n \sup_{|\theta-\theta_0|\geq\delta}[(f(j,\theta)-f(j,\theta_0))^2]^{(1+c)/2}}{\sum_{j=1}^n \inf_{|\theta-\theta_0|\geq\delta}(f(j,\theta)-f(j,\theta_0))^2} < \infty.$$

(ii) $f(j,\theta)$ are Lipschitz functions on θ and

$$\sup_{\theta_1\neq\theta_2} \frac{|\, f(j,\theta_1) - f(j,\theta_2)\,|}{|\,\theta_1-\theta_2\,|} \leq M \sup_{|\theta-\theta_0|\geq\delta} |\, f(j,\theta) - f(j,\theta_0)\,|$$

for some $\delta > 0$ and for all j, where M is independent of j.

(iii) $\sum_{j=1}^n (f(j,\theta)-f(j,\theta_0))^2 \to \infty$ and $\sup_j E(\epsilon_j^2) < \infty$.

There are two main points to be noted. First, this result requires $\sup_j E(\epsilon_j^2) < \infty$. Our condition A5 incorporates possible unbounded nature of the variance sequence. Secondly, Wu's condition attempts to incorporate regression functions that are not even differentiable. We assume in S1 that $f(j,\theta)$'s are continuously differentiable. The trade-off shows in the rate condition. While Wu's assumption W(i) requires $\overline{\lambda}(\Gamma_n) = O(\underline{\lambda}(\Gamma_n)^c)$ for some

$c < 2$, we can relax it to a great extent; to get an idea compare the above conditions with (1.5) and (1.6) in the linear regression case. We now discuss the linear regression model and a specific nonlinear regression model.

EXAMPLE 2.1 (*Linear models*). Consider the model $Y_j = x_j^{'}\theta + \epsilon_j$, where $\{x_j\}$ is a fixed regressor sequence and $\{\epsilon_j\}$ is an independent mean zero sequence with finite variances. Because we are dealing with the linear model, S1 is automatically satisfied with $\tau_{jk} = 0$ for all j, k. Moreover, we can choose $\Gamma_n = \sum_{j=1}^{n} x_j\, x_j^{'}$. Therefore, $\gamma_n = \underline{\lambda}(\sum_{j=1}^{n} x_j\, x_j^{'})$, which we need to assume diverges to infinity from A4. Next we can choose,

$$d_{nk}^2 = \sum_{j=1}^{n} x_{jk}^2\, E(\epsilon_j^2).$$

From the above facts and some elementary matrix computation it can be shown that $\mu_n \approx C_n^{-2}\underline{\lambda}(F_n)$ where $C_n^2 = \max_{1\le j\le n} E\epsilon_j^2$ and $F_{n,kl} = \Gamma_{n,kl}\,(\Gamma_{n,kk}\Gamma_{n,ll})^{-1/2}$. From the standard theory of linear models (see, Rao, 1974) it follows that for well conditioned design matrix (see (1.5)),

$$\underline{\lambda}(F_n) \approx \{\min_k (1 - R_{n,k.\bar{k}}^2)\},$$

where $R_{n,k.\bar{k}}^2$ is the multiple correlation of the k-th regressor variable on the rest based on the first n observations. Therefore, $\mu_n \approx C_n^{-2}$. Hence, A5 is equivalent to

$$\frac{C_n^2\, L(\, C_n^2\, \{\max_k \Gamma_{n,kk}\}\,)}{\{\min_k (1 - R_{n,k.\bar{k}}^2)\}\, \underline{\lambda}(\Gamma_n)} \to 0.$$

In conclusion, for linear models with independent errors and well conditioned design matrix (see (1.5)) if the error variances grow at a rate slightly less than the minimum eigenvalue of the design matrix, the LSE is strongly consistent.

EXAMPLE 2.2 (*A specific nonlinear regression model*). Consider the exponential regression model

$$Y_j = \alpha \exp(\beta x_j) + \epsilon_j,$$

where the ϵ_j are independent with mean zero and variance σ_j^2 (say). This model has applications in areas such as systems analysis. We will impose restrictions on the model so that we are in the parameter independent rate situation. This example will be continued in Section 2.3 where we will see

that when $\{x_j\}$ is unbounded, we may obtain the parameter dependent rate situation.

The parameter space for $\theta = (\alpha,\ \beta)$ is given by $\Theta = [a,\ A] \times [b,\ B] \subset (0,\ \infty) \times (-\infty,\ 0)$. The true values of the parameters are denoted by α_0 and β_0.

We first assume that $\{x_j\}$ are nonnegative and converges to zero as $j \to \infty$. Also assume that, $\sum_{j=1}^n x_j^2$ and $s_n^2 = \sum_{j=1}^n (x_j - \bar{x}_n)^2$ both diverge to ∞ as $n \to \infty$. Moreover, $r_n = (\sum_i^n x_j) \,/\, (n \sum_1^n x_j^2)^{1/2}$, is assumed to be bounded away from 1. Sequences like $x_j = j^{-u}$ will satisfy this property.

Simple calculations show that as $j \to \infty$,

$$f_1(j,\theta) = \exp(\beta x_j) \quad \text{and} \quad f_2(j,\theta) = \alpha\, x_j\, \exp(\beta x_j).$$

Also, for a suitable constant $C > 0$,

$$|f_1(j,\theta) - f_1(j,\theta')| \le C\,|x_j|\,||\theta - \theta'|| \text{ and } |f_2(j,\theta) - f_2(j,\theta')| \le C\,|x_j|\,||\theta - \theta'||.$$

Therefore, S1 is satisfied with $\tau_{j1} = \tau_{j2} = C\,x_j$. Some algebra yields that in (2.5)

$$\Gamma_n(\theta) \approx \begin{pmatrix} n(1+o(1)) & \{(\alpha+\alpha_0)/2\}\,(\sum_1^n x_j)\,(1+o(1)) \\ \{(\alpha+\alpha_0)/2\}\,(\sum_1^n x_j)\,(1+o(1)) & \{(\alpha+\alpha_0)/2\}^2\,(\sum_1^n x_j^2) \end{pmatrix} \tag{2.6}$$

Moreover, the $o(1)$ terms appearing in each cell of the above matrix are so uniformly in θ because Θ is compact. However, we need a parameter independent Γ_n. To obtain it, define $G_n = \mathrm{Diag}(n,\ \sum_1^n x_j^2)$. Consider the matrix $G_n^{-1/2}\Gamma_n(\theta)G_n^{-1/2}$ and recall that $r_n = (\sum_1^n x_j)/(n\,\sum_1^n x_j^2)^{1/2}$. It is not difficult to check that

$$\liminf_{n\to\infty} \inf_\theta \underline{\lambda}\left(\begin{bmatrix} 1 & \{(\alpha+\alpha_0)/2\}\,r_n \\ \{(\alpha+\alpha_0)/2\}\,r_n & \{(\alpha+\alpha_0)/2\}^2 \end{bmatrix} \right) \ge \liminf_{n\to\infty} C(1-r_n^2) > \eta > 0.$$

Thus (2.1) is satisfied with $\Gamma_n = \eta G_n$. Also, A4 is satisfied as $\gamma_n = \sum_1^n x_j^2 \to \infty$, by supposition. Next, since $x_j \to 0$, we can choose $d_{n1}^2 \approx \sum_{i=1}^n \max(1, x_j^2)\sigma_j^2 \approx \sum_{i=1}^n \sigma_j^2$ and $d_{n2}^2 \approx \sum_{i=1}^n x_j^2\sigma_j^2$. Also, as in Example 2.1, we can claim, $\mu_n \ge C_n^2$, with C_n as before.

Finally, A5 is satisfied if

$$C_n^2 L(\sum_1^n \sigma_j^2) / \left(\sum_1^n x_j^2\right) \to 0.$$

By choosing $x_j = j^{-u}$ for some $u > 0$ we see that $n = O([\sum_1^n x_j^2]^c)$ for some $c < 2$ iff $u > 1/4$. In our case, for the above choice, $s_n^2 = [(1-2u)u^2/(1-u)^2] \sum_1^n x_j^2$. Therefore, Theorem 2.2 is applicable for any $0 < u < 1/2$. For $u = 1/2$, it can be separately verified that $s_n^2 \approx \sum_1^n x_j^2 \approx \log n$. Hence strong consistency holds in this case if $\sup_n C_n^2 < \infty$ and $\sum_{j=1}^n \sigma_j^2 = o(n)$. Whether strong consistency holds with $\sum_{j=1}^n \sigma_j^2 \approx n$ cannot be settled by Theorem 2.2.

REMARK 2.2. (*Random regressors*). A very relevant issue is to what extent is it possible to relax the assumption that the regressors are non-random. The nonlinear regression model results of Lai (1994) allows some dependent structure and proves strong consistency under conditions similar to Wu (1981) with appropriate changes. Unfortunately, our method of proof relies heavily on a strong law of large numbers and it requires the independence of the $\{\epsilon_j\}$. It is not at all trivial to extend this lemma to the dependence case. The problem is that the underlying function space needs to be sufficiently smooth for such a dependent strong law to be valid. Lai (1994) works with a set of conditions so that this space is an appropriate Hilbert space on which a martingale strong law is available. It will be of interest to see if our method can be combined with his, to extend both our and Lai's results to an appropriate dependent set up. It may be noted that when $\{\epsilon_j\}$ is a martingale difference sequence with respect to the σ-fields G_j, with $\sup_j E(|\ \epsilon_j\ |^r \mid G_{n-1}) < \infty$, Lai and Wei (1982, 1987) have shown that strong consistency holds for the *stochastic linear regression* (then x_j is G_{j-1} measurable) if $\underline{\lambda}(\sum_i^n x_i x_i') \to \infty$ and (1.7) holds.

2.2 *Models with Parameter Dependent Rate.* Models with parameter dependent rate arise naturally. Consider the model given in Example 2.2 where $\alpha(=1)$ is known, $\beta \in [b, B] \subset (0, \infty)$ and $x_j = j$. Using some calculus, one can show that $\Gamma_n(\beta) \approx \sum_1^n \exp(2j\beta_0)\{1 - \exp((\beta - \beta_0)j)\}^2$. Thus, $\gamma_n \approx \sum_1^n \exp(2j\beta_0)\{1 - \exp(-(\beta_0 - b)j)\}^2 \approx \exp(2n\beta_0)$. On the other hand, by definition $d_n^2 \approx \sum_1^n j^2 \exp(2Bj) \approx n^2 \exp(2Bn)$. It is now obvious that $\omega_n \approx n \exp(2n(B - 2\beta_0))$. Therefore, A5 cannot hold unless we assume $B < 2b$. This is an unnatural restriction on the parameter space. To avoid this unnatural restriction, we will work with suitable local versions of the earlier assumptions.

Assumption A1 remains the same. We modify the remaining assumptions as follows. Let $B_\eta(\theta) = \{\theta' \in \Theta : ||\theta' - \theta|| \leq \eta\}$ for $\eta > 0$ and $\theta \in \Theta$.

Assumption X2. For each $\theta \in \Theta$ there exists $\eta > 0$ and square integrable

random variables $M_{jk}(\theta)$ such that for some $\sigma > 0$, and any $\theta' \in B_\eta(\theta)$ and for all $j \geq 1$ and $1 \leq k \leq p$,

$$|\{\psi_{jk}(Y_j, \theta') - E(\psi_{jk}(Y_j, \theta')\} - \{\psi_{jk}(Y_j, \theta) - E(\psi_{jk}(Y_j, \theta)\}| \leq M_{jk}(\theta)||\theta' - \theta||^\sigma,$$

Assumption X3. For each $\theta \in \Theta$ there exists $\eta > 0$ and a $p \times p$ nonnegative definite matrix $\Gamma_n(\theta)$ such that uniformly in $\theta_1 \in B_\eta(\theta)$, for sufficiently large n,

$$\sum_{i=1}^{n} \{E(\rho_j(Y_j, \theta_1) - \rho_j(Y_j, \theta_0))\} \geq (\theta_1 - \theta_0)' \, \Gamma_n(\theta) \, (\theta_1 - \theta_0). \tag{2.7}$$

Next define for large n,

$$\begin{aligned} d_{nk}^2(\theta) &\approx \sum_{j=1}^{n} \max(\, EM_{jk}^2(\theta), \mathrm{Var}(\psi_{jk}(Y_j, \theta))\,), \\ D_n(\theta) &= \mathrm{Diag}(d_{n1}^2(\theta), \ldots, d_{np}^2(\theta)). \end{aligned} \tag{2.8}$$

Also let, $\gamma_n(\theta) = \underline{\lambda}(\Gamma_n(\theta))$, $\mu_n(\theta) = \underline{\lambda}(D_n^{-1/2}(\theta)\Gamma_n(\theta)D_n^{-1/2}(\theta))$.

Note that the above choices may be influenced by the choice of η. To keep matters simple, this dependence is not explicitly shown in the notations. Also, it can be readily seen that the same $\Gamma_n(\theta)$ in X3 works for any $\eta' < \eta$. Next, suppose there is a choice of $\{\Gamma_n(\theta)\}$ such that

Assumption X4. $\gamma_n(\theta) = \underline{\lambda}(\Gamma_n(\theta)) \to \infty$ as $n \to \infty$, for every $\theta \in \Theta$, and

Assumption X5. For every $\theta \neq \theta_0 \in \Theta$, we can find $\eta > 0$, such that with this choice, $d_{nk}^2(\theta) \to \infty$ as $n \to \infty$ for $1 \leq k \leq p$, and

$$\omega_n(\theta) := \max_{1 \leq k \leq p} L(d_{nk}^2(\theta))/\mu_n(\theta)\gamma_n(\theta) \to 0.$$

As mentioned earlier, X2–X5 are implied by A2–A5. As a consequence the next theorem is more general than Theorem 2.1.

THEOREM 2.3. *Under* A1 *and* X2–X5, $\theta_n - \theta_0 \to 0$ *almost surely.*

REMARK 2.3. As in Theorem 2.1, it is possible to derive a (local) rate for almost sure convergence in Theorem 2.3 as well. For any $\delta > 0$, let $\omega_{n,\delta} = \sup_{\theta \in \bar{B}_\delta} \omega_n(\theta)$, where $\bar{B}_\delta = B_\delta^c(\theta_0) \cap \Theta$. Then one can show that $\theta_n - \theta_0 = O(\omega_{n,\delta}^{1/2})$ almost surely, for any fixed $\delta > 0$. We will not give a proof of this fact. As pointed out earlier, Wu (1981) states a result using a

local version of his conditions W(i) etc. However, his and our results are not quite comparable here.

EXAMPLE 2.2 (*Continued*). We now continue our discussion of Example 2.2. We restrict to the scenario described in the beginning of the section to positive regressors $\{x_j\}$ increasing to ∞. Assume that $\{\sigma_j^2\}$ and $\{x_j\}$ are such that the approximations below are valid. We can verify that for fixed $\beta \in [b, B]$ and for some $\eta < (1/4)b$, $|\exp(\beta' x) - \exp(\beta x)| \leq Cx \exp((\beta+\eta)x)\,|\beta' - \beta|$ for any $\beta - \eta < \beta' < \beta + \eta$, for any $x > 0$. Therefore we can choose $M_j(\beta) \approx |x_j| \exp((\beta+\eta)x_j)$. Thus, $d_n^2(\beta) \approx \sum_1^n \max\{ x_j^2 \exp(2\beta x_j), x_j^2 \exp(2(\beta+\eta)x_j \}\sigma_j^2 \approx \sum_1^n x_j^2 \exp(2(\beta+\eta)x_j)) \approx x_n^2 \exp(2(\beta+\eta)x_n)$. In a similar vein, $\gamma_n(\beta) \approx \sum_1^n x_j^2 \exp(2\beta x_j)$ $\{(\exp((\beta-\eta)x_j)-1)/(\beta-\eta)x_j\}^2 \approx \sum_1^n \exp(2(\beta-\eta)x_j) \approx \exp(2(\beta-\eta)x_n)$. From these estimates it follows that

$$\frac{L(d_n^2(\beta))}{\mu_n(\beta)\gamma_n(\beta)} \approx \{\max_{1\leq i\leq n} x_j\}^3 \exp(-2(\beta-3\eta)x_n),$$

and due to the restriction on η, this tends to 0 for every $\beta \in [b, B]$.

Consider now this model with $x_j = \log j$ and the parameter space as $[a,\ A] \times [-1/2,\ 0]$. This model was considered by Wu (1981, Example 4). If $\beta_0 = -1/2$, the issue of strong consistency of its LSE could not be settled by his results.

It can be checked that for any $\beta > -1/2$,

$$\mu_n \approx n^{-\eta}/(\log n)^6, \quad \text{and} \quad \gamma_n(\beta) \approx n^{1+2\beta-\eta}/(\log n)^4$$

where η is sufficiently small. These relations are established by using calculations similar to those given earlier and the integral approximation $\int_2^n x^{-2\beta} (\log x)^{-p}\, dx \approx (1-2\beta)^{-1}\, n^{1-2\beta}/(\log n)^p \ \{1 + p/(1-2\beta)\,(\log n)^{-1} + (p(p+1))/(1-2\beta)\,(\log n)^{-2} + O(\,(\log n)^{-3}\,)\}$ for $p > 0$.

Using the above estimates, the conditions of Theorem 2.3 can be verified and thus, the LSE of β is strongly consistent when $\beta_0 = -1/2$.

Application to Generalized linear models. These models were introduced by Nelder and Wedderburn (1972). We shall now see how the parameter dependent rates arise in these models. Suppose that the observations $\{Y_j : j \geq 1\}$ are independent with density $f(y, \beta_j)$, $j \geq 1$ respectively. Also assume that the density is from an *exponential* family and the *link function* is *linear*, that is, for all $j \geq 1$, $\beta_j = x_j'\theta$. But we emphasize that in our approach, the exponentiality and differentiability of the density or the linearity of the

link function are not essential. For simplicity we will also assume further that for some twice continuously differentiable function μ,

$$Y_j \text{ has density } \exp\{x_j'\theta - \mu(x_j'\theta)\},$$

with respect to some sigma-finite measure ν. We further assume that $\mu''(s) > 0$ over the natural parameter space. We first identify some of the key quantities.

$$\rho_j(y_j, \theta) = -(y_j\,(x_j'\theta) \;-\; \mu(x_j'\theta)),$$
$$\psi_{jk}(\theta) = -(y_j - \mu'(x_j'\theta))x_{jk}.$$

From above, we conclude $M_{jk} = 0$. Moreover, A1 and A2 are automatically satisfied. Note that for $\theta_1 \in B_\eta(\theta)$,

$$\begin{aligned} A_n(\theta_1) &= \textstyle\sum_{i=1}^n \{E(\rho_j(Y_j, \theta_1) - \rho_j(Y_j, \theta_0))\} \\ &= (\theta_1 - \theta_0)' \left\{ \textstyle\sum_{i=1}^n (\int_0^1 \alpha\mu''(x_j'\theta_{1\alpha}) d\,\alpha)\, x_j x_j' \right\} (\theta_1 - \theta_0), \end{aligned}$$

where $\theta_{1\alpha} = \theta_0 + \alpha(\theta_1 - \theta_0)$. Let

$$c(u) = \inf_{|s-t| \le u} \frac{\mu''(s)}{\mu''(t)}.$$

Note that $c(u)$ is a decreasing function of u. To estimate the integral above, we first split it into two integrals on $[0, \epsilon]$ and $[1-\epsilon, 1]$ respectively. In the next step depending on whether $\mu''(x_j'\theta_0) \ge \mu''(x_j'\theta_1)$ or not, we consider either $[0, \epsilon]$ or $[1-\epsilon, 1]$ and ignore the other one. As a result we obtain a lower bound of A_n because μ is convex. Also, we use some standard inequalities to show that for some constant M,

$$\begin{aligned} &A_n(\theta_1) \\ &\ge (1/2)\epsilon^2\; c(M\eta|x|_n)\; (\theta_1 - \theta_0)' \left\{ \sum_{j=1}^n \max\{\mu''(x_j'\theta),\, \mu''(x_j'\theta_0)\}\, x_j x_j' \right\} (\theta_1 - \theta_0) \end{aligned}$$

uniformly in $\theta_1 \in B_\eta(\theta)$. Here, $|x|_n = \max_{1 \le j \le n} |x_j|$. Define,

$$\Gamma_n(\theta) = c(M\eta|x|_n) \sum_{j=1}^n \max\{\mu''(x_j'\theta),\; \mu''(x_j'\theta_0)\}\, x_j x_j'.$$

Then X3 will hold with the above choice. Also, we may define

$$\gamma_n(\theta) \approx c(M\eta|x|_n)\;\; \max\{\tau_n(\theta),\quad \tau_n(\theta_0)\}.$$

where

$$\tau_n(\theta) = \underline{\lambda}\left(\sum_{j=1}^{n} \mu''(x_j'\theta)x_j x_j'\right).$$

Note that here again the dependence on η is implied. By defining $\tilde{\mu}_n = \underline{\lambda}(D_n^{-1/2}(\theta_0)\Gamma_n(\theta_0)D_n^{-1/2}(\theta_0))$ it can be readily seen that

$$\mu_n(\theta) \geq c(M\eta|x|_n)\,\tilde{\mu}_n.$$

Putting all these together we have,

THEOREM 2.4 (*Generalized Linear Models*). Assume that the parameter space Θ is compact. Suppose further that

(i) $\gamma_n(\theta) \to \infty$ for every $\theta \in \Theta$ and,

(ii) For every $\theta \neq \theta_0 \in \Theta$,

$$\frac{L\left(\bar{\lambda}(\Gamma_n(\theta_0))\right)}{c^2(M\eta|x|_n)\,\tilde{\mu}_n\,\max\{\tau_n(\theta),\ \tau_n(\theta_0)\}} \to 0.$$

Then $\theta_{n,m} \to \theta_0$ almost surely.

REMARK 2.4. Fahrmeir and Kaufmann (1985) have proved the strong consistency of the maximum likelihood estimator in generalized linear model. Their main assumption is an eigenvalue assumption and is very similar to condition W(i) of Wu (1981) mentioned earlier. We have already discussed and compared this condition vis a vis our conditions in the context of non-linear regression. Similar comparison holds here too. Moreover, the results of Fahrmeir and Kaufmann (1985) tackle essentially parameter independent cases. The strong consistency results of Chen, Hu and Ying (1999) are in the spirit of LRW. However, they assume bounded regressors.

3. Proofs

As already mentioned, the main goal of this work has been to unify major results on strong consistency in two different contexts, namely, (1.2) and (1.4), under the minimum contrast estimation umbrella. If we carefully go through the basic arguments of LRW, Wu (1981) and FK we identify two common logical steps in their proofs. The first step finds out an appropriate upper bound of a centered random process defined on Θ. This is the *linear* term in the Taylor expansion of $Q_n(\theta)$. In the second step a lower bound of

the deterministic *bias/quadratic* term is obtained. Finally, it is shown that the quadratic term dominates the random term of the first step.

In our method of proof we retain the first step of Wu (1981), *i.e*, we use essentially the same strong law for Banach space valued random elements. FK also use a similar strong law for generalized linear models case. On the other hand, Chen, Hu and Ying (1999) chose an indirect form of the strong law which was originally stated in LRW. The choice of the Banach space depends on the smoothness assumptions of the underlying model.

The novelty of our approach lies in the handling of the deterministic bias term in a unified manner. Firstly, we are able to classify different scenarios in terms of parameter independent/dependent rates. Usefulness of this classification is demonstrated by our success in settling the critical case in Example 2.2, which was left out in Wu (1981) as undecided. Moreover, we are able to incorporate two important situations, namely the unbounded regressors case in generalized linear models and the case of unbounded error variances in nonlinear least squares problems.

A couple of technical aspects of our method of proof are perhaps worth mentioning. Firstly in Lemma 2 below, an accurate lower bound is derived for the sum of a linear and a quadratic function. This lemma enables us to automatically obtain a rate of convergence for strong consistency (which is not usually the case) and lies at the heart of the argument for our main theorem. Secondly, the Taylor series approximation has been applied in the proof of Theorem 2.1 in a slightly modified form than we usually find.

Next we state two Lemmas needed in the proof of Theorem 2.1. We shall skip the proofs of Theorem 2.2 and 2.4 (as they easily follow from Theorem 2.1 and Theorem 2.3) and give only a brief outline of the proof of Theorem 2.3.

Let $\{Z_i(t) : i \geq 1, t \in T\}$ be independent mean zero random processes on T which is a compact metric space. Suppose that there exists a $\sigma > 0$ and nonnegative random variables M_i such that $|Z_i(t) - Z_i(s)| \leq M_i||t - s||^\sigma$. Let $\tau_i^2 = \max[E[Z_i(t_0)]^2, E(M_i^2)]$ and $d_n^2 = \sum_{i=1}^n \tau_i^2$.

LEMMA 3.1. Suppose that $d_n^2 \to \infty$ and $f : [0, \infty) \to [0, \infty)$ is a strictly increasing function. If $\int_0^\infty u^{-1} f^{-2}(u)du < \infty$, we have

$$\frac{1}{d_n f(d_n^2)} \sum_{i=1}^n Z_i(t) \to 0 \text{ in } ||\cdot||_\infty \text{ norm almost surely.}$$

PROOF. Note that T is compact with the metric $d(s,t) = ||t - s||^\sigma$. Let $N(\epsilon, d, T)$ be the minimum number of d - balls of radius at most ϵ that covers

T. Then $N(\epsilon, d, T)=O(\epsilon^{-p/\sigma})$ as $\epsilon \to 0$. Hence $\int_{0+}^{1}[\log N(\epsilon, d, T)]^{1/2} d\epsilon < \infty$. Consider the Banach space $Lip(d)=\{x \in C(T) : L(x)=\sup_{t \neq s} \frac{|x(t)-x(s)|}{d(s,t)} < \infty\}$ equipped with the norm $||x||_d = |x(t_0)| + L(x)$. By our assumptions, the identity map $i : Lip(d) \to C(S)$ is Type 2. Hence, for any sequence of independent, mean zero $Lip(d)$ valued random elements (Y_i), $E||Y_1 + \cdots + Y_n||_\infty^2 \leq A \sum_{i=1}^n E||Y_i||_d^2$ for some universal constant A. The remaining part of the proof follows the same way as in Wu (1981). We omit the details. □

The next Lemma is on the minimum of a function which is the sum of a linear and a quadratic form.

LEMMA 3.2. Let $a = (a_1 \ldots a_p)^{'}$ be a p-vector and B be a $p \times p$ positive definite matrix.

(i) If $p = 1$, $\min_{|u|=1} [au + bu^2] \geq \min(b - a, b + a)$.

(ii) If $p > 1$, suppose for some $0 < \eta < 1$,

$$a^{'} B^{-1} a \leq \frac{(p-1)(1-\eta)}{p^2} \underline{\lambda}(B).$$

Then,

$$\min_{||u||=1} [a^{'} u + u^{'} B u] \geq \underline{\lambda}(B) \, [\frac{1}{p} - (\frac{a^{'} B^{-1} a}{p \underline{\lambda}(B)})^{1/2} - \frac{1}{4} \frac{a^{'} B^{-1} a}{\underline{\lambda}(B)}].$$

REMARK 3.1. In applications to follow we apply Lemma 3.2 on a sequence $\{a_n, B_n\}$, where $a_n' B_n^{-1} a_n \,/\, \underline{\lambda}(B_n) \to 0$. Thus, the condition in (ii) is trivially satisfied. The extra precision over the usual upper bound for $a' B^{-1} a$ (namely, $||a||^2 \,/\, \underline{\lambda}(B)$) is necessary in the proof of Theorem 2.1. This fact also translates into an improved rate conditions for the theorem.

PROOF OF LEMMA 3.2. The case for $p = 1$ is trivial. For (ii) define

$$Q(u) = a^{'} u + u^{'} B u.$$

First assume that $B = \text{Diag}(b_1, \ldots, b_p)$ is a diagonal matrix. Consider the Lagrangian $\{a^{'} u + u^{'} B u + \lambda(u^{'} u - 1)\}$ where λ is the undetermined multiplier. Since $\{||u|| = 1\}$ is compact and does not have any boundary point the global minimum must be a local minimum as well. Thus, if the global minimum is attained at some u then (i) $||u|| = 1$ and (ii) u satisfies the condition of a stationary point for the Lagrangian, namely,

$$a_i + 2(b_i + \lambda) u_i = 0 \text{ for all } i. \tag{3.1}$$

Let $I = \{i : b_i + \lambda = 0\}$. Note that for any $i \in I$, $a_i = 0$. Now consider three cases:

CASE 1. $\sum_{i\in I} u_i^2 = 1$. In this case, $Q(u) = u'Bu \geq \underline{\lambda}(B)$ and the Lemma is trivially true since B is positive definite.

CASE 2. $\sum_{i\in I} u_i^2 = t$, where $1/p \leq t < 1$.

In this case noting that $a_i = 0$ for all $i \in I$ and using equation (3.1), for any extremum u,

$$Q(u) = \sum_{i\in I} b_i u_i^2 + \sum_{i\in I^c} [a_i(\frac{-a_i}{2(b_i+\lambda)}) + \frac{a_i^2 b_i}{4(b_i+\lambda)^2}].$$

Using the facts that $b_i \geq \underline{\lambda}(B)$, and $\sum_{i\in I} u_i^2 = t$ in the above expression, we get

$$Q(u) \geq \underline{\lambda}(B)[t - \frac{1}{4\underline{\lambda}(B)} \sum_{i\in I^c} \frac{a_i^2(b_i+2\lambda)}{(b_i+\lambda)^2}] = R(u), \text{ say.} \tag{3.2}$$

Since $t \geq 1/p$ and $(b+2\lambda)/(b+\lambda)^2 \leq 1/b$ for any λ, the right side (3.2) is at least as large as

$$\underline{\lambda}(B)[\frac{1}{p} - \frac{1}{4\underline{\lambda}(B)} \sum_{i\in I^c} \frac{a_i^2}{b_i}] = \underline{\lambda}[\frac{1}{p} - \frac{1}{4}\frac{a'B^{-1}a}{\underline{\lambda}(B)}]$$

establishing the Lemma in Case 2.

CASE 3. $\sum_{i\in I} u_i^2 = t$, where $t < 1/p$.

Fix $\epsilon > 0$ sufficiently small. Note that I^c must be nonempty and the cardinality of $I^c \leq p$). Hence, using (3.1) and the fact that $\sum_{i\in I^c} u_i^2 = 1 - t > (p-1)/p$, we can find $j \in I^c$ such that

$$\frac{a_j^2}{(b_j+\lambda)^2} \geq \frac{4(1-\epsilon)(1-t)}{p} = \delta^2, \text{ say,}$$

It now follows that

$$b_j + 2\lambda \leq 2\delta^{-1}|a_j| - b_j \leq 0. \tag{3.3}$$

The first inequality above follows trivially from the previous step. To obtain the last inequality above, first use the condition of the Lemma to obtain the upper bound $a_j^2/b_j^2 \le a_j^2/(\, b_j \underline{\lambda}(B)\,) \le a^{'} B^{-1} a \,/\, \underline{\lambda}(B) \le (p-1)(1-\eta)/p^2$. implied by the condition of the Lemma. On the other hand, $4/\delta^2 = p/\{(1-\epsilon)(1-t)\} < p^2/\{(p-1)(1-\epsilon)\}$ in Case 3. Hence, $(4/\delta^2)\,(a_j^2/b_j^2) \le (1-\eta)/(1-\epsilon) \le 1$, if ϵ is chosen sufficiently small.

Therefore, by using (3.2) and (3.3) and the above estimates we get

$$R(u) \ge \underline{\lambda}(B)\, t + \frac{1}{4}\frac{a_j^2(b_j - 2\delta^{-1}|a_j|)}{\delta^{-2}a_j^2} - \frac{1}{4}\sum_{i \in I^c - \{j\}} \frac{a_i^2(b_i + 2\lambda)}{(b_i+\lambda)^2}$$

$$\ge \underline{\lambda}(B)[t + \frac{1}{4}\delta^2(1 - 2\delta^{-1}(\frac{a^{'}B^{-1}a}{\underline{\lambda}(B)})^{1/2}) - \frac{1}{4}\frac{a^{'}B^{-1}a}{\underline{\lambda}(B)}].$$

Since $\delta^2 \to \frac{4}{p}(1-t)$ as $\epsilon \to 0$, $t + (1/4)\delta^2(1 - 2\delta^{-1}(a^{'}B^{-1}a/\underline{\lambda}(B))^{1/2}) \to t + \{(1-t)/p\}(1 - \{p/(1-t)\}^{1/2}(a^{'}B^{-1}a/\underline{\lambda}(B))^{1/2}) \ge 1/p - \{(1-t)/p\}^{1/2}(a^{'}B^{-1}a/\underline{\lambda}(B))^{1/2})$. The Lemma follows in Case 3.

For the general case when B is not necessarily diagonal, let P be an orthonormal matrix such that $PBP^{'} = \mathrm{Diag}(\lambda_1 \dots \lambda_p) = B_1$. The Lemma then follows from the observations that
$\underline{\lambda}(B_1) = \underline{\lambda}(B)$, $a^{'}u + u^{'}Bu = (Pa)^{'}Pu + (Pu)^{'}B_1(Pu)$ and $a^{'}B^{-1}a = (Pa)^{'}B_1^{-1}(Pa)$. □

Proof of Theorem 2.1. Recall the notations appearing in Assumptions A1–A5. We present here the proof in the case when $p > 1$. When $p = 1$, steps are exactly the same but much simpler. First, notice that it suffices to verify

$$\liminf_{n\to\infty} \inf_{|\theta - \theta_0| \ge \omega_n} [Q_n(\theta) - Q_n(\theta_0)] > 0 \text{ almost surely.}$$

Let $Z_{nk}(\theta) = \sum_{j=1}^n [\overline{\psi}_{jk}(Y_j,\theta) - \overline{b}_{jk}(\theta)]$, $Z_n(\theta) = (Z_{n1}(\theta), \dots Z_{np}(\theta))^{'}$, and $Z_{jk}(\theta) = \overline{\psi}_{jk}(Y_j,\theta) - \overline{b}_{jk}(\theta)$. By A2 and A4, Lemma 3.1 applies and hence for every k

$$\frac{1}{d_{nk}L(d_{nk}^2)}||\sum_{j=1}^n Z_{jk}(\theta)||_\infty \to 0 \text{ almost surely.} \tag{3.4}$$

By A1 and A3,

$$Q_n(\theta) - Q_n(\theta_0) \;=\; \sum_{j=1}^n \{\rho_j(Y_j,\theta) - \rho_j(Y_j,\theta_0)\}$$

$$
\begin{aligned}
&= (\theta-\theta_0)' \sum_{j=1}^{n}\{\bar{\Psi}_j(Y_j,\theta)-\bar{B}_j(\theta)\}+\sum_{j=1}^{n} E\{\rho_j(Y_j,\theta)-\rho_j(Y_j,\theta_0)\} \\
&\geq (\theta-\theta_0)' Z_n(\theta) + (\theta-\theta_0)' \Gamma_n (\theta-\theta_0),
\end{aligned}
$$

(using Fubini's theorem it can be readily seen from A1 that $E\{\rho_j(Y_j,\theta) - \rho_j(Y_j,\theta_0)\} = (\theta-\theta_0)' \bar{B}_j(\theta)$).

It now follows that

$$
\inf_{|\theta-\theta_0|\geq\delta}[Q_n(\theta)-Q_n(\theta_0] \geq \inf_{\theta_1\in\Theta}\inf_{|\theta-\theta_0|\geq\delta}[(\theta-\theta_0)' Z_n(\theta_1)+(\theta-\theta_0)' \Gamma_n (\theta-\theta_0)]. \tag{3.5}
$$

The second infimum on the right side of the above inequality can be re-expressed as

$$
\inf_{l\geq\delta}\inf_{||u||=1}[l\, u' Z_n(\theta_1) + l^2\, u'\Gamma_n u]. \tag{3.6}
$$

In the next step of our argument we apply Lemma 3.2(ii) to obtain an effective lower bound of (3.6). In order to do so we first verify that

$$
\gamma_n^{-1} Z_n'(\theta_1)\Gamma_n^{-1} Z_n(\theta_1 \to 0
$$

almost surely uniformly in θ_1. Writing $V_n = D_n^{1/2} L(D_n)$, $F_n = D_n^{-1/2}\Gamma_n D_n^{-1/2}$ and $R_n(\theta_1) = V_n^{-1} Z_n(\theta_1)$, we get, for every $\theta_1 \in \Theta$,

$$
\begin{aligned}
\gamma_n^{-1} Z_n'(\theta_1)\Gamma_n^{-1} Z_n(\theta_1) &\leq \gamma_n^{-1}\, R_n' \left(L(D_n)F_n^{-1}L(D_n)\right) R_n \\
&\leq \gamma_n^{-1}\, \bar{\lambda}(L(D_n)F_n^{-1}L(D_n))\, ||R_n||_\infty^2 \qquad (3.7) \\
&\leq \omega_n ||R_n||_\infty^2,
\end{aligned}
$$

after some elementary matrix algebra (here, $||(f_1(t),\ldots,f_p(t))||_\infty = (\sum_{i=1}^{p} ||f_i||_\infty^2)^{1/2}$). Now, by A5 and (3.4) both ω_n and $||R_n||_\infty$ tend to zero in respective senses. Thus, the above claim is verified.

In view of the above claim the condition of Lemma 3.2 is easily satisfied with any $\eta \in (0,1)$ for the inner infimum in (3.6), almost surely as $n \to \infty$, uniformly in θ_1, θ_2. As a consequence, we can conclude that (3.6) is at least as large as

$$
\delta^2\gamma_n \left[\frac{1}{p} - \frac{1}{\delta}\left(\frac{Z_n'(\theta_1)\,\Gamma_n^{-1}\, Z_n(\theta_1)}{p\gamma_n}\right)^{1/2} - \frac{1}{4\delta^2}\left(\frac{Z_n'(\theta_1)\,\Gamma_n^{-1}\, Z_n(\theta_1)}{\gamma_n}\right)\right] \tag{3.8}
$$

almost surely as $n \to \infty$.

Therefore, from (3.5)-(3.8) it follows that for any $\theta_1, \theta_2 \in \theta$,

$$\inf_{\theta_1 \in \Theta} \inf_{|\theta-\theta_0| \geq \delta} [(\theta-\theta_0)' Z_n(\theta_1) + (\theta-\theta_0)' \Gamma_n (\theta-\theta_0)]$$

$$\geq \delta^2 \gamma_n \left[\frac{1}{p} - \frac{1}{\delta p^{1/2}} \omega_n^{1/2} ||R_n||_\infty - \frac{1}{4\delta^2} \omega_n ||R_n||_\infty^2 \right].$$

Hence, we conclude

$$\inf_{|\theta-\theta_0| \geq \omega_n^{1/2}} [Q_n(\theta) - Q_n(\theta_0)] > 0 \text{ almost surely eventually.} \qquad \square$$

Proof of Theorem 2.3. We give only a brief outline of the proof of the convergence, but do not verify the rate of convergence claim in Remark 2.3. Fix any $\delta > 0$ and θ satisfying $||\theta - \theta_0|| \geq \delta$. In view of the assumptions X2–X5 and following the steps of the previous proof we can show that

$$\inf_{\theta \in B_\eta(\theta) \cap \Theta} [Q_n(\theta) - Q_n(\theta_0)] > 0 \text{ almost surely eventually,}$$

provided η is chosen so small that $\theta_0 \notin B_\eta(\theta)$. Now, using the fact that Θ is compact we can cover $\{\theta \in \Theta : ||\theta - \theta_0|| > 2\delta\}$ by finitely many neighborhoods $B_{\eta_1}(\theta_1), \ldots, B_{\eta_r}(\theta_r)$ such that $B_\delta(\theta_0) \cap B_{\eta_i}(\theta_i) = \phi$ for $1 \leq i \leq r$. Hence the theorem follows. The claimed rate of almost sure convergence in Remark 2.3 can be established by extending the argument given in the proof of Theorem 2.1. $\square$

Acknowledgement. The Referee has made very constructive comments which have led to a substantial improvement in clarity of presentation. He has also pointed out a few important references which we had overlooked. We are grateful to him.

References

Bai, Z.D. and Wu, Y. (1997). General M-estimation. *J. Multivariate Anal.* **63**, 119-135.

Bose, A. and Sengupta, D. (1991). Asymptotic Properties of Estimators in Nonlinear Regression and Generalized Regression Models. *Tech. Report, No. 20-91, Statistics and Mathematics Unit*, Indian Statistical Institute, Kolkata.

Chen, K., Hu, I. and Ying, Z. (1999). Strong Consistency of Maximum Quasi-likelihood Estimators in Generalized Linear Models with Fixed and Adaptive Designs. *Ann. Statist.* **27**, 1155-1163.

Chen, X.R. and Wu, Y.H. (1988). Strong Consistency of M-estimates in Linear Models. *J. Multivariate Anal.* **27**, 116-130.

Christopeit, N. and Helmes, K. (1980). Strong Consistency of Least Squares Estimators in Linear Regression Models. *Ann. Statist.* **8**, 778-788.

Drygas, H. (1976). Weak and Strong Consistency of the Least Square Estimator in Regression Models. *Z. Wahrsch. Verw. Gebiete* **34**, 119-127.

Fahrmeir, L. and Kaufmann, H. (1987). Consistency and Asymptotic Normality of the Maximum Likelihood Estimate in Generalized Linear Models. *Ann. Statist.* **13**, 342-368.

Hampel, F. R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics — The Approach based on Influence Functions.* Wiley, New York.

Huber, P.J. (1964). Robust Estimation of a Location Parameter. *Ann. Math. Statist.* **35**, 73-101.

Huber, P.J. (1967). The Behavior of the Maximum Likelihood Estimates under Nonstandard Conditions. *Proc. Fifth Berkeley Symp. Math. Stat. Probab.* **1**. Univ. Calif. Press, Berkeley, 221-233.

Lai, T,L. (1994). Asymptotic Properties of Nonlinear Least Squares Estimates in Stochastic Regression Models. *Ann. Statist.* **22**, 1917-1930.

Lai, T.L., Robbins. H. and Wei, C.Z. (1978). Strong Consistency of Least Squares Estimators in Multiple Regression. *Proc. Nat. Acad. Sci. U.S.A.* **75**, 3034-3036.

Lai, T.L., Robbins. H. and Wei, C.Z. (1979). Strong Consistency of Least Squares Estimates in Multiple Regression II. *J. Multivariate Anal.* **9**, 343-361.

Lai, T.L. and Wei. C.Z. (1982). Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems.*Ann. Statist.* **10**, 154-166.

Lai, T.L. and Wei. C.Z. (1987). Asymptotically Efficient Self Tuning Regulators.*J. Contrl. Optimiz.* **25**, 466-481.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *J.R. Stat. Soc. A*, **135**, 370-384

Pfanzagl, J. (1969). Consistent Estimation of a Location Parameter in the Presence of an Incidental Parameter. *Ann. Math. Statist.* **40**, 1353-1357.

Rao, C.R. (1974). *Linear Statistical Inference and Its Applications.* Wiley Eastern, New Delhi.

Wu, C.F. (1981). Asymptotic Theory of Nonlinear Least Squares Estimation. *Ann. Statist.* **9**, 501-513.

Arup Bose and Debapriya Sengupta

Stat-Math Unit

Indian Statistical Institute

203 B.T. Road

Kolkata 700108, INDIA

Email: abose@isical.ac.in

dps@isical.ac.in