NONLINEAR LEAST-SQUARES ESTIMATION

DAVID POLLARD AND PETER RADCHENKO

ABSTRACT. The paper uses empirical process techniques to study the asymptotics of the least-squares estimator for the fitting of a nonlinear regression function. By combining and extending ideas of Wu and Van de Geer, it establishes new consistency and central limit theorems that hold under only second moment assumptions on the errors. An application to a delicate example of Wu's illustrates the use of the new theorems, leading to a normal approximation to the least-squares estimator with unusual logarithmic rescalings.

1. INTRODUCTION

Consider the model where we observe y_i for $i = 1, \dots, n$ with

(1) $$y_i = f_i(\theta) + u_i \qquad \text{where } \theta \in \Theta.$$

The unobserved f_i can be random or deterministic functions. The unobserved errors u_i are independent random variables with zero means and finite variances. The index set Θ might be infinite dimensional. Later in the paper it will prove convenient to also consider triangular arrays of observations.

Think of $f(\theta) = (f_1(\theta), \dots, f_n(\theta))'$ and $u = (u_1, \dots, u_n)'$ as points in $\mathbb{R}^n$. The model specifies a surface $M_\Theta = \{f(\theta) : \theta \in \Theta\}$ in $\mathbb{R}^n$. The vector of observations $y = (y_1, \dots, y_n)'$ is a random point in $\mathbb{R}^n$. The least squares estimator (LSE) $\widehat{\theta}_n$ is defined

1991 *Mathematics Subject Classification.* Primary 62E20. Secondary: 60F05, 62G08, 62G20.

Key words and phrases. Nonlinear least squares, empirical processes, subgaussian, consistency, central limit theorem.

to minimize the distance of y to M_Θ,

$$\widehat{\theta}_n = \operatorname{argmin}_{\theta\in\Theta} |y - f(\theta)|^2,$$

where $|\cdot|$ denotes the usual Euclidean norm on $\mathbb{R}^n$. Many authors have considered the behavior of $\widehat{\theta}_n$ as $n \to \infty$ when the y_i are generated by the model for a fixed θ_0 in Θ.

When the f_i are deterministic, it is natural to express assertions about convergence of $\widehat{\theta}_n$ in terms of the n-dimensional Euclidean distance $\kappa_n(\theta_1,\theta_2) := |f(\theta_1) - f(\theta_2)|$. For example, Jennrich (1969) took Θ to be a compact subset of $\mathbb{R}^p$, the errors $\{u_i\}$ to be iid with zero mean and finite variance, and the f_i to be continuous functions in θ. He proved strong consistency of the least squares estimator under the assumption that $n^{-1}\kappa_n(\theta_1,\theta_2)^2$ converges uniformly to a continuous function that is zero if and only if $\theta_1 = \theta_2$. He also gave conditions for asymptotic normality.

Under similar assumptions Wu (1981, Theorem 1) proved that existence of a consistent estimator for θ_0 implies that

$$\kappa_n(\theta) := \kappa_n(\theta,\theta_0) \to \infty \qquad \text{at each } \theta \neq \theta_0. \tag{2}$$

If Θ is finite, the divergence (2) is also a sufficient condition for the existence of a consistent estimator (Wu 1981, Theorem 2). His main consistency result (his Theorem 3) may be reexpressed as a general convergence assertion.

Theorem 1. *Suppose the $\{f_i\}$ are deterministic functions indexed by a subset Θ of $\mathbb{R}^p$. Suppose also that $\sup_i \operatorname{var}(u_i) < \infty$ and $\kappa_n(\theta) \to \infty$ at each $\theta \neq \theta_0$. Let S be a bounded subset of $\Theta\backslash\{\theta_0\}$ and let $R_n := \inf_{\theta\in S}\kappa_n(\theta)$. Suppose there exist constants $\{L_i\}$ such that*

(i) $\sup_{\theta\in S}|f_i(\theta) - f_i(\theta_0)| \le L_i$ *for each i;*
(ii) $|f_i(\theta_1) - f_i(\theta_2)| \le L_i|\theta_1 - \theta_2|$ *for all $\theta_1,\theta_2 \in S$;*
(iii) $\sum_{i\le n} L_i^2 = O(R_n^\alpha)$ *for some $\alpha < 4$.*

Then $\mathbb{P}\{\widehat{\theta}_n \notin S \text{ eventually }\} = 1$.

Remark. Assumption (i) implies $\sum_{i\le n} L_i^2 \ge \kappa_n(\theta)^2 \to \infty$ for each θ in S, which forces $R_n \to \infty$.

If Θ is compact and if for each $\theta \neq \theta_0$ there is a neighborhood $S = S_\theta$ satisfying the conditions of the Lemma then $\widehat{\theta}_n \to \theta_0$ almost surely.

Wu's paper was the starting point for several authors. For example, both Lai (1994) and Skouras (2000) generalized Wu's consistency results by taking the functions $f_i(\theta) = f_i(\theta, \omega)$ as random processes indexed by θ. They took the $\{u_i\}$ as a martingale difference sequence, with $\{f_i\}$ a predictable sequence of functions with respect to a filtration $\{\mathcal{F}_i\}$.

Another line of development is typified by the work of Van de Geer (1990) and Van de Geer and Wegkamp (1996). They took $f_i(\theta) = f(x_i, \theta)$, where $\mathcal{F} = \{f_\theta : \Theta\}$ is a set of deterministic functions (in fact they identified Θ with the index set $\mathcal{F}$) and the x_i are either fixed points in $\mathbb{R}^d$ or iid random variables that are independent of the errors. Van de Geer and Wegkamp (1996) gave necessary and sufficient conditions for the convergence $n^{-1}\kappa_n^2(\widehat{\theta}_n) \to 0$, which corresponds to consistency with respect to the $\mathcal{L}^2(P_n)$ pseudometric on the functional class $\mathcal{F}$. Under a stronger assumption about the errors, Van de Geer (1990) established sharper stochastic bounds for $\kappa_n(\widehat{\theta}_n)$ in terms of $\mathcal{L}^2$ entropy conditions on $\mathcal{F}$, using empirical process methods that were developed after Wu's work.

The stronger assumption was that the errors are uniformly subgaussian. In general, we say that a random variable W has a subgaussian distribution if there exists some finite τ such that

$$\mathbb{P}\exp(tW) \le \exp\left(\tfrac{1}{2}\tau^2 t^2\right) \qquad \text{for all } t \in \mathbb{R}.$$

We write $\tau(W)$ for the smallest such τ. Van de Geer assumed that $\sup_i \tau(u_i) < \infty$.

> **Remark.** Notice that we must have $\mathbb{P}W = 0$ when W is subgaussian because the linear term in the expansion of $\mathbb{P}\exp(tW)$ must vanish. When $\mathbb{P}W = 0$, subgaussianity is equivalent to existence of a finite constant β for which $\mathbb{P}\{|W| \ge x\} \le 2\exp(-x^2/\beta^2)$ for all $x \ge 0$.

In our paper we try to bring together the two lines of development. Our main motivation for working on nonlinear least squares was an example presented by Wu (1981, page 507).

He noted that his consistency theorem has difficulties with a simple model,

$$f_i(\theta) = \lambda i^{-\mu} \qquad \text{for } \theta = (\lambda, \mu) \in \Theta, \text{ a compact subset of } \mathbb{R} \times \mathbb{R}^+. \tag{3}$$

For example, condition (2) does not hold for $\theta_0 = (0, 0)$ at any θ with $\mu > 1/2$. When $\theta_0 = (\lambda_0, 1/2)$, Wu's method fails in a more subtle way, and the results of Van de Geer and Wegkamp (1996) do not yield consistency in the parametric sense. Van de Geer (1990)'s method would work if the errors satisfied the subgaussian assumption. In Section 4, under only second moment assumptions on the errors, we establish weak consistency and a central limit theorem.

The main idea behind all the proofs—ours, as well as those of Wu and Van de Geer—is quite simple. The LSE also minimizes the random function

$$\begin{aligned} G_n(\theta) &:= |y - f(\theta)|^2 - |u|^2 = \kappa_n(\theta)^2 - 2Z_n(\theta) \\ &\qquad\qquad \text{where } Z_n(\theta) := u'f(\theta) - u'f(\theta_0). \end{aligned} \tag{4}$$

In particular, $G_n(\widehat{\theta}_n) \le G_n(\theta_0) = 0$, that is, $\frac{1}{2}\kappa_n(\widehat{\theta}_n)^2 \le Z_n(\widehat{\theta}_n)$. For every subset S of Θ,

$$\mathbb{P}\{\widehat{\theta}_n \in S\} \le \mathbb{P}\{\exists \theta \in S : Z_n(\theta) \ge \tfrac{1}{2}\kappa_n(\theta)^2\} \le 4\mathbb{P} \sup_{\theta \in S} |Z_n(\theta)|^2 / \inf_{\theta \in S} \kappa_n(\theta)^4. \tag{5}$$

The final bound calls for a maximal inequality for Z_n.

Our methods for controlling Z_n are similar in spirit to those of Van de Geer. Under her subgaussian assumption, for every class of real functions $\{g_\theta : \theta \in \Theta\}$, the process

$$X(\theta) = \sum\nolimits_{i \le n} u_i g_i(\theta) \tag{6}$$

has subgaussian increments. Indeed, by the definition of $\tau(u_i)$

$$\begin{aligned} \mathbb{P} \exp\Big(t[X(\theta_1) - X(\theta_2)]\Big) &= \prod_i \mathbb{P} \exp(t u_i h_i) \quad \text{where } h_i = g_i(\theta_1) - g_i(\theta_2) \\ &\le \prod_i \exp\left(\tfrac{1}{2} t^2 \tau^2(u_i) h_i^2\right). \end{aligned}$$

Consequently, if $\tau(u_i) \le \tau_0$ for all i then

$$\tau^2\Big(X(\theta_1) - X(\theta_2)\Big) \le \sum\nolimits_{i \le n} \tau^2(u_i) \Big(g_i(\theta_1) - g_i(\theta_2)\Big)^2 \le \tau_0^2 |g(\theta_1) - g(\theta_2)|^2.$$

That is, the tails of $X(\theta_1) - X(\theta_2)$ are controlled by the n-dimensional Euclidean distance between the vectors $g(\theta_1)$ and $g(\theta_2)$. This property allowed her to invoke a chaining bound (similar to our Theorem 2) for the tail probabilities of $\sup_{\theta\in S}|Z_n(\theta)|$ for various annuli $S = \{\theta : R \le \kappa_n(\theta) < 2R\}$.

Under the weaker second moment assumption on the errors, we apply symmetrization arguments to transform to a problem involving a new process $Z_n^\circ(\theta)$ with conditionally subgaussian increments. We avoid Van de Geer's subgaussianity assumption at the cost of extra Lipschitz conditions on the $f_i(\theta)$, analogous to Assumption (ii) of Theorem 1, which lets us invoke chaining bounds for conditional second moments of $\sup_{\theta\in S}|Z_n^\circ(\theta)|$ for various S.

In Section 3 we prove a new consistency theorem (Theorem 3) and a new central limit theorem (Theorem 4) for nonlinear LSEs. More precisely, our consistency theorem corresponds to an explicit bound for $\mathbb{P}\{\kappa_n(\widehat{\theta}_n) \ge R\}$, but we state the result in a form that makes comparison with Theorem 1 easier. Our Theorem does not imply almost sure convergence, but our techniques could easily be adapted to that task. We regard the consistency as a preliminary to the next level of asymptotics and not as an end in itself. We describe the local asymptotic behavior with another approximation result, Theorem 4, which can easily be transformed into a central limit theorem under a variety of mild assumptions on the $\{u_i\}$ errors.

Theorem 4 generalizes the CLT proved by Wu. It covers all the examples in Wu, excluding only his examples of inconsistency. Our theorem does not cover the nonparametric result in Section 6.1 of Wegkamp (1998). In Section 4 we illustrate our new CLT by applying it to the model (3) to sharpen the consistency result at $\theta_0 = (1, 1/2)$ into the approximation

$$\left(\ell_n^{1/2}(\widehat{\lambda}_n - 1), \ell_n^{3/2}(1 - 2\widehat{\mu}_n)\right) = \textstyle\sum_{i\le n} u_i\zeta_{i,n}' + o_p(1) \tag{7}$$

where $\ell_n := \log n$ and

$$\zeta_{i,n} = i^{-1/2}\ell_n^{-1/2}\begin{pmatrix} 2 & -6 \\ -6 & 24 \end{pmatrix}\begin{pmatrix} 2 \\ \ell_i/\ell_n \end{pmatrix}.$$

The sum on the right-hand side of (7) is of order $O_p(1)$ when $\sup_i \text{var}(u_i) < \infty$. If the $\{u_i\}$ are also identically distributed, the sum has a limiting multivariate normal distribution.

This example may appear contrived. It was offered by Wu (Example 4, p. 507) as a case in which his consistency result did not apply: "The really interesting (or disappointing) case is [the case $\theta_0 = (1, 1/2)$ in our model (3)] for which [his consistency condition, our Theorem 1] is not satisfied." In the context of his CLT he cited the same example, noting that "This again demonstrates the difficulty of the asymptotic theory when [$\kappa_n^2(\theta)$] goes to infinity at a rate different from n". We feel that this example is therefore a good illustration of how our methods improve on Wu's results.

2. Maximal inequalities

Assumption (ii) of Theorem 1 ensures that the increments $Z_n(\theta_1) - Z_n(\theta_2)$ are controlled by the ordinary Euclidean distance in Θ; we allow for control by more general metrics. Wu invoked a maximal inequality for sums of random continuous processes, a result derived from a bound on the covering numbers for M_θ as a subset of $\mathbb{R}^n$ under the usual Euclidean distance; we work with covering numbers for other metrics.

Definition 1. *Let (T, d) be a pseudometric space. The covering number $N(\delta, T, d)$ is defined as the size of the smallest δ-net for T, that is, the smallest N for which there are points $t_1, \ldots, t_N$ in T with $\min_i d(t, t_i) \leq \delta$ for every t in T.*

> **Remark.** Allowing pseudometric rather than metric spaces is a slight increase in generality that is sometimes convenient when dealing with metrics defined by $\mathcal{L}^p$ norms on functions.

Standard chaining arguments give maximal inequalities for processes with subgaussian increments controlled by a pseudometric on the index set.

Theorem 2. *Let $\{W_t : t \in T\}$ be a stochastic process, indexed by a pseudometric space (T, d), with subgaussian increments. Let T_δ be a δ-net for T. Suppose:*

(i) *there is a constant K such that $\tau(W_s - W_t) \leq K d(s, t)$ for all $s, t \in T$;*

(ii) $J_\delta := \int_0^\delta \rho(N(y,T,d))\,dy < \infty$, *where* $\rho(N) := \sqrt{1+\log N}$.

Then there is a universal constant c_1 *such that*

$$\frac{1}{c_1}\sqrt{\mathbb{P}\sup_t |W_t|^2} \le K J_\delta + \rho(N(\delta,T,d))\max_{s\in T_\delta}\tau(W_s).$$

Remark. We should perhaps work with outer expectations because, in general, there is no guarantee that a supremum of uncountably many random variables is measurable. For concrete examples, such as the one discussed in Section 4, measurability can usually be established by routine separability arguments. Accordingly, we will ignore the issue in this paper.

Proof. Upper bound the L_2 norm of $\sup_t |W_t|$ by the sum of the L_2 norms of $\max_{s\in T_\delta}|W_s|$ and $\sup_{d(s,t)\le\delta}|W_s - W_t|$. The latter can be bounded above by a multiple of J_δ using Theorem 2.2.4 (and the display given above it) of van der Vaart and Wellner (1996). This theorem is stated for a general Orlicz norm $\|\cdot\|_\psi$ and should be applied for $\psi(x) = e^{x^2} - 1$, using the fact that the Orlicz norm corresponding to this function upper bounds the L_2 norm. The constant η can be taken equal to 2δ. Bound the L_2 norm of $\max_{s\in T_\delta}|W_s|$ above by a multiple of $\rho(N(\delta,T,d))\max_{s\in T_\delta}\tau(W_s)$ using Lemma 2.2.2 of the same book. □

Under the assumption that var$(u_i) \le \sigma^2$, the X process from (6) need not have subgaussian increments. However, it can be bounded in a stochastic sense by a symmetrized process $X^\circ(\theta) := \sum_{i\le n}\epsilon_i u_i g_i(\theta)$, where the $2n$ random variables $\epsilon_1,\dots,\epsilon_n,u_1,\dots,u_n$ are mutually independent with $\mathbb{P}\{\epsilon_i = +1\} = 1/2 = \mathbb{P}\{\epsilon_i = -1\}$. In fact, for each subset S of the index set Θ,

$$\mathbb{P}\sup_{\theta\in S}|X(\theta)|^2 \le 4\mathbb{P}\sup_{\theta\in S}|X^\circ(\theta)|^2. \tag{8}$$

For a proof see, for example, van der Vaart and Wellner (1996, Lemma 2.3.1). Moreover, (van der Vaart and Wellner 1996, Lemma 2.2.7)

$$\begin{aligned}\mathbb{P}_u \exp\left(t[X^\circ_{\theta_1} - X^\circ_{\theta_2}]\right) &= \prod_i \mathbb{P}_u \exp(\epsilon_i t u_i h_i) \quad \text{where } h_i = g_i(\theta_1) - g_i(\theta_2) \\ &= \prod_i \tfrac{1}{2}[\exp(tu_ih_i) + \exp(-tu_ih_i)] \\ &\le \prod_i \exp\left(\tfrac{1}{2}t^2u_i^2h_i^2\right).\end{aligned}$$

The subscript u indicates the conditioning on u. It follows from the above display that the process X° has conditionally subgaussian increments with

$$\tau_u^2\left(X^\circ_{\theta_1} - X^\circ_{\theta_2}\right) \le \sum\nolimits_{i\le n} u_i^2\left(g_i(\theta_1) - g_i(\theta_2)\right)^2. \tag{9}$$

We use this property of the symmetrized process to produce a maximal inequality for X.

Corollary 1. *Let S_δ be a δ-net for S and let X be as in (6). Suppose*

(i) $\mathbb{P}u_i = 0$ *and* $\text{var}(u_i) \le \sigma^2$ *for* $i = 1, \dots, n$

(ii) *there is a metric d for which $J_\delta := \int_0^\delta \rho(N(y, S, d))\, dy < \infty$*

(iii) *there are constants $L_1, \dots, L_n$ for which*

$$|g_i(\theta_1) - g_i(\theta_2)| \le L_i d(\theta_1, \theta_2) \qquad \text{for all } i \text{ and all } \theta_1, \theta_2 \in S$$

(iv) *there are constants $b_1, \dots, b_n$ for which $|g_i(\theta)| \le b_i$ for all i and all θ in S.*

Then there is a universal constant c_2 such that

$$\mathbb{P}\sup_{\theta\in S}|X_\theta|^2 \le c_2^2\sigma^2\left(LJ_\delta + B\rho(N(\delta, S, d))\right)^2$$

where $L := \sqrt{\sum_i L_i^2}$ and $B := \sqrt{\sum_i b_i^2}$.

Proof. It follows from inequality (9) and the derivation preceding it that

$$\tau_u(X^\circ_{\theta_1} - X^\circ_{\theta_2}) \le L_u d(\theta_1, \theta_2) \qquad \text{where } L_u := \sqrt{\sum\nolimits_{i\le n} L_i^2 u_i^2}$$

and

$$\tau_u(X^\circ_\theta) \le B_u := \sqrt{\sum\nolimits_{i\le n} b_i^2 u_i^2}$$

Apply Theorem 2 conditionally to the process X° to derive

$$\mathbb{P}_u \sup_{\theta\in S} |X^\circ_\theta|^2 \le c_1^2(L_u J_\delta + B_u \rho(N(\delta, T, d)))^2.$$

Invoke inequality (8), using the fact that $\mathbb{P}L_u^2 \le \sigma^2 L^2$ and $\mathbb{P}B_u^2 \le \sigma^2 B^2$. □

3. LIMIT THEOREMS

Inequality (5) and Corollary 1, with $g_i(\theta) = f_i(\theta) - f_i(\theta_0)$, give us some probabilistic control over $\widehat{\theta}_n$.

Theorem 3. *Let S be a subset of Θ equipped with a pseudometric d. Let $\{L_i : i = 1, \dots, n\}$, $\{b_i : i = 1, \dots, n\}$, and δ be positive constants such that*

(i) $|f_i(\theta_1) - f_i(\theta_2)| \le L_i d(\theta_1, \theta_2)$ *for all* $\theta_1, \theta_2 \in S$
(ii) $|f_i(\theta) - f_i(\theta_0)| \le b_i$ *for all* $\theta \in S$
(iii) $J_\delta := \int_0^\delta \rho\left(N(y, S, d)\right)\, dy < \infty$

Then

$$\mathbb{P}\{\widehat{\theta}_n \in S\} \le 4c_2^2\sigma^2\left(B\rho\left(N(\delta, S, d)\right) + LJ_\delta\right)^2 / R^4,$$

where $R := \inf\{\kappa_n(\theta) : \theta \in S\}$, *and* $L^2 = \sum_i L_i^2$, *and* $B^2 := \sum_i b_i^2$.

The Theorem becomes more versatile in its application if we partition S into a countable union of subsets S_k, each equipped with its own pseudometric and Lipschitz constants. We then have $\mathbb{P}\{\widehat{\theta}_n \in \cup_k S_k\}$ smaller than a sum over k of bounds analogous to those in the Theorem. As shown in Section 4, this method works well for the Wu example if we take $S_k = \{\theta : R_k \le \kappa_n(\theta) < R_{k+1}\}$, for an $\{R_k\}$ sequence increasing geometrically.

A similar appeal to Corollary 1, with the $g_i(\theta)$ as partial derivatives of $f_i(\theta)$ functions, gives us enough local control over Z_n to go beyond consistency. To accommodate the application in Section 4, we change notation slightly by working with a triangular array: for each n,

$$y_{in} = f_{in}(\theta_0) + u_{in}, \qquad \text{for } i = 1, 2, \dots, n,$$

where the $\{u_{in} : i = 1, \dots, n\}$ are unobserved independent random variables with mean zero and variance bounded by σ^2.

Theorem 4. *Suppose $\widehat{\theta}_n \to \theta_0$ in probability, with θ_0 an interior point of Θ, a subset of $\mathbb{R}^p$. Suppose also:*

(i) *Each f_{in} is continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0 with derivatives $D_{in}(\theta) = \partial f_{in}(\theta) / \partial\theta$.*

(ii) $\gamma_n^2 := \sum_{i \le n} |D_{in}(\theta_0)|^2 \to \infty$ *as* $n \to \infty$.

(iii) *There are constants $\{M_{in}\}$ with $\sum_{i \le n} M_{in}^2 = O(\gamma_n^2)$ and a metric d on $\mathcal{N}$ for which $|D_{in}(\theta_1) - D_{in}(\theta_2)| \le M_{in} d(\theta_1, \theta_2)$ for $\theta_1,\ \theta_2 \in \mathcal{N}$.*

(iv) *The smallest eigenvalue of the matrix $V_n = \gamma_n^{-2} \sum_{i \le n} D_{in}(\theta_0) D_{in}(\theta_0)'$ is bounded away from zero for n large enough.*

(v) $\int_0^1 \rho\left(N(y, \mathcal{N}, d)\right)\, dy < \infty$

(vi) $d(\theta, \theta_0) \to 0$ *as* $\theta \to \theta_0$.

Then $\kappa_n(\widehat{\theta}_n) = O_p(1)$ and

$$\gamma_n(\widehat{\theta}_n - \theta_0) = \sum\nolimits_{i \le n} \xi_{i,n} u_{in} + o_p(1) = O_p(1).$$

where $\xi_{i,n} = \gamma_n^{-1} V_n^{-1} D_{in}(\theta_0)$.

Proof. Let D be the $p \times n$ matrix with ith column $D_{in}(\theta_0)$, so that $\gamma_n^2 = \mathrm{trace}(DD')$ and $V_n = \gamma_n^{-2} DD'$. The main idea of the proof is to replace $f(\theta)$ by $f(\theta_0) + D'(\theta - \theta_0)$, thereby approximating $\widehat{\theta}_n$ by the least-squares solution

$$\overline{\theta}_n := \theta_0 + (DD')^{-1} Du = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} |y - f(\theta_0) - D'(\theta - \theta_0)|.$$

To simplify notation, assume with no loss of generality, that $f(\theta_0) = 0$ and $\theta_0 = 0$. Also, drop extra n subscripts when the meaning is clear. The assertion of the Theorem is that $\widehat{\theta}_n = \overline{\theta}_n + o_p(\gamma_n^{-1})$.

Without loss of generality, suppose the smallest eigenvalue of V_n is larger than a fixed constant $c_0^2 > 0$. Then

$$\gamma_n^2 = \text{trace}(DD') \geq \sup_{|t|\leq 1} |D't|^2 \geq \inf_{|t|\leq 1} |D't|^2 = c_0^2\gamma_n^2,$$

from which it follows that

$$c_0|t| \leq |D't|/\gamma_n \leq |t| \qquad \text{for all } t \in \mathbb{R}^p. \tag{10}$$

Similarly, $\mathbb{P}|Du|^2 = \text{trace}\left(D\mathbb{P}(uu')D'\right) \leq \sigma^2\gamma_n^2$, implying that $|Du| = O_p(\gamma_n)$ and

$$\overline{\theta}_n = \gamma_n^{-2}V_n^{-1}Du = O_p(\gamma_n^{-1}).$$

In particular, $\mathbb{P}\{\overline{\theta}_n \in \mathcal{N}\} \to 1$, because θ_0 is an interior point of Θ. Note also that

$$\mathbb{P}|\textstyle\sum_{i\leq n}\xi_i u_i|^2 \leq \sigma^2\text{trace}(\sum_{i\leq n}\xi_i\xi_i') = \sigma^2\text{trace}(V_n^{-1}) = O(1) \quad \text{by (iv)}.$$

Consequently $\sum_{i\leq n}\xi_i u_i = O_p(1)$.

From the assumed consistency, we know that there is a sequence of balls $\mathcal{N}_n \subseteq \mathcal{N}$ that shrink to $\{0\}$ for which $\mathbb{P}\{\widehat{\theta}_n \in \mathcal{N}_n\} \to 1$. From (vi) and (v), it follows that both $r_n := \sup\{d(\theta, 0) : \theta \in \mathcal{N}_n\}$ and $J_{r_n} = \int_0^{r_n} \rho\left(N(y, \mathcal{N}, d)\right) dy$ converge to zero as $n \to \infty$.

The $n \times 1$ remainder vector $R(\theta) := f(\theta) - D'\theta$ has ith component

$$R_i(\theta) = f_i(\theta) - D_i(0)'\theta = \theta' \int_0^1 D_i(t\theta) - D_i(0)\, dt. \tag{11}$$

Uniformly in the neighborhood $\mathcal{N}_n$ we have

$$|R(\theta)| \leq |\theta| \left(\sum_{i\leq n} M_{in}^2\right)^{1/2} r_n = o(|\theta|\gamma_n),$$

which, together with the upper bound from inequality (10), implies

$$|f(\theta)|^2 = |D'\theta|^2 + o(\gamma_n^2|\theta|^2) = O(\gamma_n^2|\theta|^2) \qquad \text{as } |\theta| \to 0. \tag{12}$$

In the neighborhood $\mathcal{N}_n$, via (11) we also have,

$$|u'R(\theta)| \leq |\theta| \sup_{s\in\mathcal{N}_n} |\sum_i u_i \left(D_i(s) - D_i(0)\right)|.$$

From Corollary 1 with $g_i(\theta) = D_i(\theta) - D_i(0)$ deduce that

$$\mathbb{P}\sup_{s\in\mathcal{N}_n} |\sum\nolimits_i u_i \left(D_i(s) - D_i(0)\right)|^2 \le c_2^2\sigma^2 J_{r_n}^2 \sum\nolimits_i M_{in}^2 = o(\gamma_n^2),$$

which implies

$$|u'R(\theta)| = o_p(\gamma_n|\theta|) \qquad \text{uniformly for } \theta \in \mathcal{N}_n. \tag{13}$$

Approximations (12) and (13) give us uniform approximations for the criterion functions in the shrinking neighborhoods $\mathcal{N}_n$:

$$\begin{aligned} G_n(\theta) &= |u - f(\theta)|^2 - |u|^2 \\ &= -2u'f(\theta) + |f(\theta)|^2 \\ &= -2u'D'\theta + |D'\theta|^2 + o_p(\gamma_n|\theta|) + o_p(\gamma_n^2|\theta|^2) \\ &= |u - D'\overline{\theta}_n|^2 - |u|^2 + |D'(\theta - \overline{\theta}_n)|^2 + o_p(\gamma_n|\theta|) + o_p(\gamma_n^2|\theta|^2). \end{aligned} \tag{14}$$

The uniform smallness of the remainder terms lets us approximate G_n at random points that are known to lie in $\mathcal{N}_n$.

The rest of the argument is similar to that of Chernoff (1954). When $\widehat{\theta}_n \in \mathcal{N}_n$ we have $G_n(\widehat{\theta}_n) \le G_n(0)$, implying

$$|D'(\widehat{\theta}_n - \overline{\theta}_n)|^2 + o_p(\gamma_n|\widehat{\theta}_n|) + o_p(\gamma_n^2|\widehat{\theta}_n|^2) \le |D'\overline{\theta}_n|^2.$$

Invoke (10) again, simplifying the last approximation to

$$c_0^2|\gamma_n\widehat{\theta}_n - \gamma_n\overline{\theta}_n|^2 \le O_p(1) + o_p\left(|\gamma_n\widehat{\theta}_n| + |\gamma_n\widehat{\theta}_n|^2\right).$$

It follows that $|\widehat{\theta}_n| = O_p(\gamma_n^{-1})$ and, via (12),

$$\kappa_n(\widehat{\theta}_n) = |f(\widehat{\theta}_n)| = O_p(1).$$

We may also assume that $\mathcal{N}_n$ shrinks slowly enough to ensure that $\mathbb{P}\{\overline{\theta}_n \in \mathcal{N}_n\} \to 1$. When both $\widehat{\theta}_n$ and $\overline{\theta}_n$ lie in $\mathcal{N}_n$ the inequality $G_n(\widehat{\theta}_n) \le G_n(\overline{\theta}_n)$ and approximation (14) give

$$|D'(\widehat{\theta}_n - \overline{\theta}_n)|^2 + o_p(1) \le o_p(1).$$

It follows that $\widehat{\theta}_n = \overline{\theta}_n + o_p(\gamma_n^{-1})$. □

Remark. If the errors are iid and $\max|\xi_{i,n}| = o(1)$ then the distribution of $\sum_{i\le n}\xi_{i,n}u_{in}$ is asymptotically $N\big(0, \sigma^2 V_n^{-1}\big)$.

4. Analysis of the important test case

The results in this section illustrate the work of our limit theorems in a particular case where Wu's method fails, namely model (3):

$$f_i(\theta) = \lambda i^{-\mu} \qquad \text{for } \theta = (\lambda, \mu) \in \Theta, \text{ a compact subset of } \mathbb{R} \times \mathbb{R}^+.$$

We prove both consistency and a central limit theorem for the case $\theta_0 = (\lambda_0, 1/2)$. In fact, without loss of generality, $\lambda_0 = 1$.

As before, let $\ell_n = \log n$. Remember $\theta = (\lambda, \mu)$ with $a \in \mathbb{R}$ and $0 \le \mu \le C_\mu$ for a finite constant C_μ greater than $1/2$, which ensures that $\theta_0 = (1, 1/2)$ is an interior point of the parameter space. Taking $C_\mu = 1/2$ would complicate the central limit theorem only slightly. The behavior of $\widehat{\theta}_n$ is determined by the behavior of the function

$$G_n(\gamma) := \textstyle\sum_{i\le n} i^{-1+\gamma} \qquad \text{for } \gamma \le 1,$$

or its standardized version

$$g_n(\beta) := G_n(\beta/\ell_n)/G_n(0) = \textstyle\sum_{i\le n}\Big(i^{-1}/G_n(0)\Big)\exp\Big(\beta\ell_i/\ell_n\Big),$$

which is the moment generating function of the probability distribution that puts mass $i^{-1}/G_n(0)$ at ℓ_i/ℓ_n, for $i = 1, \ldots, n$. For large n, the function g_n is well approximated by the increasing, nonnegative function

$$g(\beta) = \begin{cases} (e^\beta - 1)/\beta & \text{for } \beta \ne 0 \\ 1 & \text{for } \beta = 0 \end{cases},$$

the moment generating function of the uniform distribution on $(0,1)$. More precisely, comparison of the sum with the integral $\int_1^n x^{-1+\gamma}\,dx$ gives

$$G_n(\gamma) \;= \ell_n g(\gamma\ell_n) + r_n(\gamma) \qquad \text{with } 0 \le r_n(\gamma) \le 1 \text{ for } \gamma \le 1. \tag{15}$$

The distributions corresponding to both g_n and g are concentrated on $[0,1]$. Both functions have the properties described in the following lemma.

Lemma 1. *Suppose* $h(\gamma) = P\exp(\gamma x)$, *the moment generating function of a probability distribution concentrated on* $[0,1]$. *Then*

(i) $\log h$ *is convex*

(ii) $h(\gamma)^2/h(2\gamma)$ *is unimodal: increasing for* $\gamma < 0$, *decreasing for* $\gamma > 0$, *achieving its maximum value of* 1 *at* $\gamma = 0$

(iii) $h'(\gamma) \le h(\gamma)$

Proof. Assertion (i) is just the well known fact that the logarithm of a moment generating function is convex. Thus h'/h, the derivative of $\log h$, is an increasing function, which implies (ii) because

$$\frac{d}{d\gamma}\log\left(\frac{h(\gamma)^2}{h(2\gamma)}\right) = 2\frac{h'(\gamma)}{h(\gamma)} - 2\frac{h'(2\gamma)}{h(2\gamma)}.$$

Property (iii) comes from the representation $h'(\gamma) = P\left(xe^{\gamma x}\right)$. □

Remark. Direct calculation shows that $g(\gamma)^2/g(2\gamma)$ is a symmetric function.

Reparametrize by putting $\beta = (1-2\mu)\ell_n$, with $(1-2C_\mu)\ell_n \le \beta \le \ell_n$, and $\alpha = \lambda\sqrt{G_n(\beta/\ell_n)}$. Notice that $|f(\theta)| = |\alpha|$ and that θ_0 corresponds to $\alpha_0 = \sqrt{G_n(0)} \approx \sqrt{\ell_n}$ and $\beta_0 = 0$. Also

$$f_i(\theta) = \alpha\nu_i(\beta/\ell_n) \qquad \text{where} \quad \nu_i(\gamma) := i^{-1/2}\exp(\gamma\ell_i/2)/\sqrt{G_n(\gamma)},$$

and

$$\kappa_n(\theta)^2 = G_n(0)\left(\lambda^2 g_n(\beta) - 2\lambda g_n(\beta/2) + 1\right). \tag{16}$$

We define $\nu_i := \sup_{\gamma\le 1}\nu_i(\gamma)$.

Lemma 2. *For all* (α, β) *corresponding to* $\theta = (\lambda, \mu) \in \mathbb{R} \times [0, C_\mu]$*:*

(i) $\kappa_n(\theta) - \sqrt{G_n(0)} \le |\alpha| \le \kappa_n(\theta) + \sqrt{G_n(0)}$

(ii) $\sum_{i \le n} \nu_i^2 = O\Big(\log\log n\Big)$

(iii) $|d\nu_i(\beta/\ell_n)/d\beta| \le \frac{1}{2}\nu_i(\beta/\ell_n)$

(iv) $|f_i(\alpha_1, \beta_1) - f_i(\alpha_2, \beta_2)| \le \Big(|\alpha_1 - \alpha_2| + \frac{1}{2}|\alpha_2||\beta_1 - \beta_2|\Big)\,\nu_i$

(v) $|f_i(\theta) - f_i(\theta_0)| \le i^{-1/2} + |\alpha|\nu_i$

Proof. Inequalities (i) and (v) follow from the triangle inequality.

For inequality (ii), first note that $\nu_1^2 \le 1$. For $i \ge 2$, separate out contributions from three ranges:

$$\nu_i^2 = \max\left(\sup_{1 \ge \gamma \ge 1/\ell_n} \nu_i(\gamma)^2, \sup_{|\gamma| < 1/\ell_n} \nu_i(\gamma)^2, \sup_{\gamma \le -1/\ell_n} \nu_i(\gamma)^2\right).$$

For $\gamma \ge 1/\ell_n$, invoke (15) to get a tractable upper bound:

$$\nu_i(\gamma)^2 \le i^{-1}\frac{\exp(\gamma\ell_i)}{\ell_n g(\gamma\ell_n)} \le i^{-1}\gamma\frac{\exp(\gamma\ell_i)}{\exp(\gamma\ell_n) - 1} \le i^{-1}\frac{\exp\Big(\log\gamma + \gamma\log(i/n)\Big)}{1 - e^{-1}}.$$

The last expression achieves its maximum over $[1/\ell_n, 1]$ at

$$\gamma_0 := \begin{cases} 1/\log(n/i) & \text{if } 1 \le i \le n/e \\ 1 & \text{if } n/e \le i \le n \end{cases},$$

which gives

(17)
$$\sup_{1 \ge \gamma \ge 1/\ell_n} \nu_i(\gamma)^2 \le \frac{(e-1)^{-1}}{n} H\left(\frac{i \wedge (n/e)}{n}\right) \qquad \text{where } H(x) := 1/\Big(x\log(1/x)\Big).$$

Similarly, if $-1 < \gamma\ell_n < 1$,

$$\nu_i(\gamma)^2 \le \frac{\exp(\gamma\ell_i)}{i\ell_n g(\gamma\ell_n)} \le \frac{\exp(\ell_i/\ell_n)}{i\ell_n g(-1)} \le \frac{e/g(-1)}{i\ell_n}.$$

The last term is smaller than a constant multiple of the bound from (17). Finally, if $-\gamma = \delta \ge 1/\ell_n$ and $i \ge 2$ then

$$\nu_i(\gamma)^2 \le i^{-1}\delta\frac{\exp(-\delta\ell_i)}{1 - \exp(-\delta\ell_n)} \le i^{-1}\frac{\exp\Big(\log\delta - \delta\ell_i\Big)}{1 - e^{-1}} \le \frac{e^{-1}/(1 - e^{-1})}{i\ell_i}.$$

In summary, for some universal constant C,

$$\nu_i^2 \le C \max\left(n^{-1} H\left(\frac{i \wedge (n/e)}{n}\right), \frac{1}{i \log i}\right) \qquad \text{if } 2 \le i \le n.$$

Bounding sums by integrals we thus have

$$C^{-1}\sum_{i=2}^{n}\nu_i^2 \le \int_{1/n}^{1/e} H(x)\,dx + H(1/e)/n + \int_2^n \Big(x\log x\Big)^{-1}dx = O\Big(\log\log n\Big).$$

For (iii) note that

$$2\frac{d}{d\beta}\nu_i(\beta/\ell_n) = 2\frac{d}{d\beta}\exp\left(\tfrac{1}{2}\beta\ell_i/\ell_n\right)\Big(G_n(0)g_n(\beta)\Big)^{-1/2} = \left(\frac{\ell_i}{\ell_n} - \frac{g_n'(\beta)}{g_n(\beta)}\right)\nu_i(\beta),$$

which is bounded in absolute value by $\nu_i(\beta)$ because $0 \le g_n'(\beta) \le g_n(\beta)$.

For (iv)

$$\begin{aligned} |f_i(\alpha_1,\beta_1) - f_i(\alpha_2,\beta_2)| &\le |(\alpha_1-\alpha_2)\nu_i(\beta_1/\ell_n)| + |\alpha_2||\nu_i(\beta_1/\ell_n) - \nu_i(\beta_2/\ell_n)| \\ &\le |(\alpha_1-\alpha_2)|\nu_i + |\alpha_2||(\beta_1-\beta_2)|\tfrac{1}{2}\nu_i, \end{aligned}$$

the bound for the second term coming from the mean-value theorem and (iii). □

Lemma 3. *For $\epsilon > 0$, let $\mathcal{N}_\epsilon = \{\theta : \max\Big(|\lambda - 1|, |\beta|\Big) \ge \epsilon$. If ϵ is small enough, there exists a constant $C_\epsilon > 0$ such that $\inf\{\kappa_n(\theta) : \theta \notin \mathcal{N}_\epsilon\} \ge C_\epsilon\sqrt{\ell_n}$ when n is large enough.*

Proof. Suppose $|\beta| \ge \epsilon$. Remember that $G_n(0) \ge \ell_n$. Minimize over λ the lower bound (16) for $\kappa_n(\theta)^2$ by choosing $\lambda = g_n(\beta/2)/g_n(\beta)$, then invoke Lemma 1(ii).

$$\frac{\kappa_n(\theta)^2}{\ell_n} \ge 1 - \frac{g_n(\beta/2)^2}{g_n(\beta)} \ge 1 - \max\left(\frac{g_n(\epsilon/2)^2}{g_n(\epsilon)}, \frac{g_n(-\epsilon/2)^2}{g_n(-\epsilon)}\right) \to 1 - \frac{g(\epsilon/2)^2}{g(\epsilon)} > 0.$$

If $|\beta| \le \epsilon$ and ϵ is small enough to make $(1-\epsilon)e^{\epsilon/2} < 1 < (1+\epsilon)e^{-\epsilon/2}$, use

$$\kappa_n(\theta)^2 = \textstyle\sum_{i\le n} i^{-1}\Big(\lambda\exp(\beta\ell_i/2\ell_n) - 1\Big)^2.$$

If $\lambda \ge 1+\epsilon$ bound each summand from below by $i^{-1}((1+\epsilon)e^{-\epsilon/2} - 1)^2$. If $\lambda \le 1-\epsilon$ bound each summand from below by $i^{-1}(1-(1-\epsilon)e^{\epsilon/2})^2$. □

4.1. **Consistency.** On the annulus $S_R := \{R \le \kappa_n(\theta) < 2R\}$ we have

$$\begin{aligned} |a| &\le K_R := 2R + \sqrt{G_n(0)} \\ |f_i(\theta_1) - f_i(\theta_2)| &\le K_R \nu_i d_R(\theta_1, \theta_2) \\ &\qquad\text{where } d_R(\theta_1,\theta_2) := |\alpha_1 - \alpha_2|/K_R + \tfrac{1}{2}|\beta_1 - \beta_2| \\ |f_i(\theta) - f_i(\theta_0)| &\le b_i := i^{-1/2} + K_R \nu_i. \end{aligned}$$

Note that

$$\sum_{i \le n} \left(i^{-1/2} + K_R \nu_i \right)^2 = O(\ell_n + K_R^2 \log \ell_n) = O(K_R^2 \mathcal{L}_n) \qquad \text{where } \mathcal{L}_n := \log\log n.$$

The rectangle $\{|\alpha| \le K_R, |\beta| \le c\ell_n\}$ can be partitioned into $O(y^{-1}\ell_n/y)$ subrectangles of d_R-diameter at most y. Thus $N(y, S_R, d_R) \le C_0 \ell_n / y^2$ for a constant C_0 that depends only on C_μ, which gives

$$\int_0^1 \rho\left(N(y, S_R, d_R) \right) \, dy = O\left(\sqrt{\mathcal{L}_n} \right).$$

Apply Theorem 3 with $\delta = 1$ to conclude that

$$\mathbb{P}\{\widehat{\theta}_n \in S_R\} \le C_1 K_R^2 \mathcal{L}_n^2 / R^4 \le C_2 (R^2 + \ell_n) \mathcal{L}_n^2 / R^4.$$

Put $R = C_3 2^k (\ell_n \mathcal{L}_n^2)^{1/4}$ then sum over k to deduce that

$$\mathbb{P}\{\kappa_n(\widehat{\theta}_n) \ge C_3 (\ell_n \mathcal{L}_n^2)^{1/4}\} \le \epsilon \qquad \text{eventually}$$

if the constant C_3 is large enough. That is $\kappa_n(\widehat{\theta}_n) = O_p\left((\ell_n \mathcal{L}_n^2)^{1/4} \right)$ and, via Lemma 3,

$$|\widehat{\lambda}_n - 1| = o_p(1) \qquad \text{and} \qquad 2\ell_n |\widehat{\mu}_n - \mu_0| = |\widehat{\beta}| = o_p(1).$$

4.2. **Central limit theorem.** This time work with the (λ, β) reparametrization, with

$$\begin{aligned} f_i(\lambda,\beta) &= \lambda i^{-1/2 + \beta/2\ell_n} \\ D_i(\lambda,\beta)' = \left(\tfrac{\partial f_i(\lambda,\beta)}{\partial \lambda}, \tfrac{\partial f_i(\lambda,\beta)}{\partial \beta} \right) &= \left(1/\lambda, \ell_i / 2\ell_n \right) f_i(\lambda,\beta) \end{aligned}$$

and $\theta_0 = (\lambda_0, \beta_0) = (1, 0)$. Take d as the usual two-dimensional Euclidean distance in the (λ, β) space. For simplicity of notation, we omit some n subscripts, even though the relationship between θ and (λ, β) changes with n.

We have just shown that the LSE $(\widehat{\lambda}_n, \widehat{\beta}_n)$ is consistent.

Comparison of sums with analogous integrals gives the approximations

$$\textstyle\sum_{i\le n} i^{-1}\ell_i^{p-1} = \ell_n^p/p + r_p \qquad \text{with } |r_p| \le 1 \text{ for } p = 0,1,2,\dots. \tag{18}$$

In consequence,

$$\gamma_n^2 = \sum\nolimits_{i\le n} |D_i(\lambda_0,\beta_0)|^2 = \sum\nolimits_{i\le n} i^{-1}\left(1+\ell_i^2/4\ell_n^2\right) = \tfrac{13}{12}\ell_n + O(1)$$

and

$$V_n = \gamma_n^{-2}\textstyle\sum_{i\le n} i^{-1}\begin{pmatrix} 1 & \ell_i/2\ell_n \\ \ell_i/2\ell_n & \ell_i^2/4\ell_n^2 \end{pmatrix} = V + O(1/\ell_n) \qquad \text{where } V = \frac{1}{13}\begin{pmatrix} 12 & 3 \\ 3 & 1 \end{pmatrix}.$$

The smaller eigenvalue of V_n converges to the smaller eigenvalue of the positive definite matrix V, which is strictly positive.

Within the neighborhood $\mathcal{N}_\epsilon := \{\max\big(|\lambda-1|, |\beta|\big) \le \epsilon\}$, for a fixed $\epsilon \le 1/2$, both $|f_i(\lambda,\beta)|$ and $|D_i(\lambda,\beta)|$ are bounded by a multiple of $i^{-1/2}$. Thus

$$|D_i(\theta_1) - D_i(\theta_1)| \le \left|\lambda_1^{-1} - \lambda_2^{-1}\right| \, |f_i(\theta_1)| + 3|f_i(\theta_1) - f_i(\theta_2)| \le C_\epsilon i^{-1/2} d(\theta_1,\theta_2).$$

That is, we may take M_i as a multiple of $i^{-1/2}$, which gives $\sum_{i\le n} M_i^2 = O(\ell_n)$.

All the conditions of Theorem 4 are satisfied. We have

$$\sqrt{\ell_n}(\widehat{\lambda_n} - 1, \widehat{\beta_n}) = \tfrac{12}{13}\textstyle\sum_{i\le n} u_i i^{-1/2}\ell_n^{-1/2}(1, \ell_i/2\ell_n)V^{-1} + o_p(1).$$

Acknowledgements. We thank the referees for their constructive comments.

REFERENCES

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics 25*, 573–578.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics 40*, 633–643.

Lai, T. L. (1994). Asymptotic properties of nonlinear least-squares estimates in stochastic regression models. *Annals of Statistics 22*, 1917–1930.

Skouras, K. (2000). Strong consistency in nonlinear stochastic regression models. *Annals of Statistics 28*, 871–879.

Van de Geer, S. (1990). Estimating a regression function. *Annals of Statistics 18*, 907–924.

Van de Geer, S. and M. Wegkamp (1996). Consistency for the least squares estimator in nonparametric regression. *Annals of Statistics 24*, 2513–2523.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag.

Wegkamp, M. (1998). *Entropy Methods in Statistical Estimation*. Center for Mathematics and Computer Science. CWI Tract 125.

Wu, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics 9*, 501–513.

STATISTICS DEPARTMENT, YALE UNIVERSITY, BOX 208290 YALE STATION, NEW HAVEN, CT 06520-8290.

E-mail address: `david.pollard@yale.edu; radchenko@galton.uchicago.edu`

URL: `http://www.stat.yale.edu/˜pollard/; http://galton.uchicago.edu/˜radchenko`