Journal of Econometrics 32 (1986) 219–251. North-Holland

A CLASS OF PARTIALLY ADAPTIVE ONE-STEP M-ESTIMATORS FOR THE NON-LINEAR REGRESSION MODEL WITH DEPENDENT OBSERVATIONS*

Benedikt M. PÖTSCHER

University of Technology of Vienna, 1040 Vienna, Austria

Ingmar R. PRUCHA

University of Maryland, College Park, MD 20742, USA

Received November 1983, final version received January 1986

In this paper we consider a class of partially adaptive one-step M-estimators for the non-linear regression model with dependent observations. Those estimators adapt themselves with respect to a measure of the tailthickness of the disturbance distribution (as well as to a measure of the scale). The large-sample behavior of those estimators is examined theoretically for general disturbance distributions and numerically for various specific ones. The estimators considered are motivated by the Student-t maximum-likelihood estimator. Given appropriate specifications of the adaptation parameter the estimators are asymptotically efficient on the family of Student-t distributions including the normal distribution.

1. Introduction

For the standard regression model with normally distributed disturbances the least-squares estimator is the maximum-likelihood estimator and hence asymptotically efficient. If, however, the actual disturbance distribution differs from the normal distribution it is well known that this estimator may behave very poorly.[1] For this reason econometricians and statisticians have, in recent years, become increasingly interested in alternative estimators that have good properties over a wide range of distributions. Since Huber (1964) a large body

*An earlier version of this paper was presented at the European Econometric Society Meeting in Pisa, 1983. The authors are grateful to Clopper Almon, Herman Bierens, Chris Clague, Manfred Deistler, Harry Kelejian, Samuel Kotz, Paul Wonnacott and this journal's editor and referees for helpful comments. All remaining errors are our responsibility. We would like to thank Nestor Dominguez for computational assistance. The financial support of the Fonds zur Foerderung der Wissenschaftlichen Forschung, Grant 4393, and the support of computer time through the facilities of the Computer Science Center of the University of Maryland is gratefully acknowledged.

[1] See, e.g., Huber (1981) and the references cited therein.

of literature on robust estimation procedures has emerged. Many robust estimators proposed in the literature belong to the class of M-estimators.[2]

Besides robustness, another goal of any estimation procedure is efficiency. Fully adaptive estimators are, from a theoretical point of view, ideal in that respect. However, Bickel (1982, p. 664) states in his seminal work on adaptive estimation:

> 'The difficulty of nonparametric estimation of score functions suggests that a more practical goal is partial adaptation, the construction of estimates which are (i) always $\sqrt{T}$-consistent, and (ii) efficient over a large parametric subfamily of $\mathscr{F}$ [the space of distributions]. Our results indicate that... this goal should be achievable by using a one-step Newton approximation to the maximum likelihood estimate for the parametric subfamily by starting with an estimate which is $\sqrt{T}$-consistent for all of $\mathscr{F}$.'

Hogg and Lenth (1984) made a similar point in a recent paper that reviews adaptive statistical techniques.[3] Of course, partially adaptive estimators should also be reasonably robust.

The present paper considers a class of partially adaptive one-step M-estimators for the non-linear regression model with (possibly) stochastic time-dependent regressors. The members of this class of estimators can be viewed as being generated as one-step Gauss–Newton approximations to the normal equations of the (pseudo) maximum-likelihood estimator corresponding to the Student-t family. The score functions associated with our estimators are indexed by an adaptation parameter that depends not only on the scale but also on the tailthickness of the distribution and that is estimated from the data. The Student-t distribution has been characterized as an important dimension in the space of distributions.[4] It contains the Cauchy distribution as a special and the normal distribution as a limiting case. For suitably defined adaptation parameters (and given certain assumptions on the disturbance process and process of the exogenous variables) our estimators are asymptotically optimal on the Student-t family including the normal distribution.

[2] A recent econometric review of the literature on robust estimation is given in Koenker (1982). For recent econometric work on robust estimators and M-estimators in general (including pseudo maximum-likelihood estimators), see, e.g., Amemiya (1982), Bierens (1981, 1982, 1984a), Burguete, Gallant and Souza (1982), Domowitz and White (1982), Gallant and Holly (1980), Gilstein and Leamer (1983), Goldfeld and Quandt (1981), Gourieroux, Monfort and Trognon (1984), Koenker and Bassett (1978), Powell (1983), Prucha and Kelejian (1984), White (1982) and White and Domowitz (1984) – to mention a few.

[3] Basic concepts concerning adaptive estimation have been introduced by Stein (1956). For a recent econometric review and extension of the literature on adaptive estimation, see Manski (1984).

[4] See, e.g., Hall and Joiner (1982).

While our estimators can be viewed as approximations to the Student-t maximum-likelihood estimator it is not maintained that the actual disturbance distribution is Student-t. We note that our results suggest that our estimators are robust against deviations from the Student-t. Heuristically this seems to follow from the fact that the so-called ψ-functions (defined below) are bounded and redescending in all cases except when the adaptation parameter is zero; this latter case corresponds to the ψ-function of the least-squares estimator.

Estimators of the above type have been considered previously by Prucha and Kelejian (1984). The analysis of that paper pertains, however, only to linear models with fixed exogenous variables. Therefore it does not generally apply to the non-linear model with dependent observations considered here. Also, due to a different focus, Prucha and Kelejian maintain much stronger assumptions on the disturbance distribution than those maintained in the present paper. For a robust interpretation of the estimators it seems essential to reduce the moment requirements. Furthermore it seems of interest to relax the assumption that the disturbances are symmetrically distributed.

The plan of the paper is as follows: Section 2 contains some preliminaries. In section 3 we define our class of estimators for the non-linear regression model and investigate their asymptotic properties under weak conditions. We also give two specific definitions for the adaptation parameter, define corresponding estimators for the adaptation parameter and investigate the statistical properties of the latter. Section 4 contains numerical comparisons of the asymptotic efficiency of our estimators with alternative estimators, including robust ones. Concluding remarks are given in section 5. All proofs are relegated to the appendices.

2. Preliminaries

Consider the stochastic data-generating process (y_t, x_t) with $y_t \in \mathbb{R}$ and $x_t \in \mathbb{R}^L$ described by the non-linear regression model

$$y_t = g(x_t, \beta) + u_t, \tag{2.1}$$

where y_t, x_t and u_t denote, respectively, the dependent variable, the vector of explanatory variables and the disturbance term; g is the response function and $\beta = (\beta_1, \dots, \beta_K)' \in \Theta_B$ is the parameter vector. Clearly any data-generating process (y_t, x_t) can be written in the form (2.1) if the class of disturbance processes is not restricted. A complete set of our assumptions on the data-generating process and the model is given in section 3.

Given observations $(y_t, x_t)_{t=1,\dots,T}$ the classical non-linear least-squares problem is to estimate β by the minimizing value over Θ_B of $\sum_{t=1}^{T}(y_t - g(x_t, \beta))^2$. The corresponding estimator will be referred to as the non-linear least-squares estimator (leaving questions of existence and uniqueness aside). Generalizing a concept of Huber (1964) for the location problem, Relles (1968) and Huber (1973) introduced and investigated, in order to obtain more robust estimators, the class of so-called M-estimators for the linear regression model with fixed regressors and i.i.d. disturbances. Investigations in this vein for the non-linear regression model have been conducted by Grossmann (1976) for fixed regressors and Bierens (1981) for stochastic regressors. For a general analysis of M-estimators in non-linear models, see Burguete, Gallant and Souza (1982) and the references cited therein; recent references include Bierens (1982, 1984a), Domowitz and White (1982), Gourieroux, Monfort and Trognon (1984), White (1982) and White and Domowitz (1984).

We generally refer to an estimator for β that minimizes an objective function of the form

$$\sum_{t=1}^{T} \rho[y_t - g(x_t, \beta)] \tag{2.2}$$

over Θ_B as an M-estimator of type I. If, for instance, ρ and g are smooth and Θ_B is open, we can write the first-order condition for a minimum of (2.2) as

$$\sum_{t=1}^{T} \psi[y_t - g(x_t, \beta)](\partial/\partial\beta_k) g(x_t, \beta) = 0, \qquad k = 1, \ldots, K, \tag{2.3}$$

where $\psi = \rho'$. In the following, we shall refer to solutions of equations of the form (2.3) as M-estimators of type II. Clearly, in the linear case and $\Theta_B = \mathbb{R}^k$ minimizing (2.2) or solving (2.3) is equivalent for differentiable and, e.g., convex ρ. Suppose the disturbances are distributed i.i.d. with common distribution function F. Given F has a smooth density f (and given appropriate orthogonality conditions between regressors and disturbances), then the normal equations of the corresponding maximum-likelihood (ML) estimator are of the form (2.3) with $\psi = -f'/f$. Various functional forms of ψ have been proposed in order to obtain robust estimators. Of particular importance is Huber's (1964) ψ-function: $\psi^c(z) = z$ if $|z| < c$ and $\psi^c(z) = c \cdot \operatorname{sign}(z)$ if $|z| \geq c$. This leads for $c < \infty$ to estimators with certain minimax robustness properties over certain classes of distributions. The case $c = \infty$ corresponds to the non-linear least-squares estimator.

Estimators obtained from (2.2) and (2.3) are typically not invariant against a multiplication of (2.1) by a constant factor. Invariance can be achieved by solving, instead of (2.3),

$$\sum_{t=1}^{T} \psi[(y_t - g(x_t, \beta))/\hat{\sigma}](\partial/\partial\beta_k) g(x_t, \beta) = 0, \tag{2.4}$$

$$k = 1, \ldots, K,$$

where $\hat{\sigma}$ is a scale-equivariant estimator of a measure of scale, say σ. Under certain general conditions M-estimators have been shown to be asymptotically normal. Furthermore, whether σ is known or estimated $\sqrt{T}$-consistently yields the same asymptotic distribution if essentially $\mathrm{E}\{(\partial^2/\partial\sigma\partial u)\rho(u/\sigma)\}$ is zero;[5] see Huber (1967, 1973, 1981), Relles (1968), Maronna and Yohai (1981), Bierens (1981), Burguete, Gallant and Souza (1982) and Grossman (1982).

For linear models with fixed regressors and i.i.d. disturbances, Bickel (1975) considered the class of one-step M-estimators obtained as a first step in a Gauss–Newton approximation of (2.4) for given $\hat{\sigma}$. He showed that under suitable conditions a one-step M-estimator initialized by a $\sqrt{T}$-consistent estimator for β is asymptotically normal even if the corresponding M-estimator is not and is asymptotically equivalent to the corresponding M-estimator when that estimator is asymptotically normal. Furthermore, he showed that the asymptotic distribution of the one-step M-estimator does not depend on the particular choice of the estimator $\hat{\sigma}$ as long as that estimator is $\sqrt{T}$-consistent.

3. A class of partially adaptive estimators

3.1. Definition of the estimators

As remarked above the class of M-estimators considered in this paper is motivated by the maximum-likelihood estimator corresponding to Student-t distributed disturbances. Assume (for a moment) that the disturbances u_t are distributed identically with zero mean and are independent of lagged disturbances, lagged dependent variables and all exogenous regressors. In case of Student-t distributed disturbances the log-likelihood function is then given by

$$\begin{aligned}&\mathscr{L}(y_1,\ldots,y_T|v,H)\\&\quad=\text{const.}+\frac{T}{2}\ln H-\frac{v+1}{2}\sum_{t=1}^{T}\ln\left[1+\frac{H}{v}(y_t-g(x_t,\beta))^2\right],\end{aligned}\qquad(3.1)$$

where H has the interpretation of a measure of the inverse scale of the distribution and the degrees of freedom parameter v has the interpretation of a measure of the tailthickness. Large values of v correspond to thin tails. For given parameters v and H the normal equations for the ML estimator for β are of the form (2.3) with ψ replaced by

$$\psi_a(z)=z/(1+az^2),\qquad 0\leq a<\infty,\qquad(3.2)$$

[5] For example, this will be the case if ρ and the disturbance distribution are symmetric.

and $a = v^{-1}H$. Following Prucha and Kelejian (1984) we can use those normal equations to define for general disturbance distributions the following class of M-estimators of type II as the solution $\bar{\beta}$ of

$$\sum_{t=1}^{T} \psi_{\tilde{\mu}\tilde{h}}(y_t - g(x_t, \bar{\beta}))(\partial/\partial\beta_k)g(x_t, \bar{\beta}) = 0, \qquad k = 1,\dots,K. \quad (3.3)$$

Since for general disturbance distributions v^{-1} and H loose their original meaning, they have been replaced by estimators $\tilde{\mu} \geq 0$ and $\tilde{h} \geq 0$. In principle we are free in the choice of those estimators. However, the cases where $\tilde{\mu}$ and $\tilde{h}$ are defined such that the corresponding theoretical quantities $\mu \geq 0$ and $h \geq 0$ can again be interpreted as measures of the tailthickness and the inverse scale of the disturbance distribution (or where $\tilde{\mu}$ is set equal to some non-negative constant) seem of particular interest. Note that $\psi_{\tilde{\mu}\tilde{h}}(z) = \psi_{\tilde{\mu}}(z\tilde{h}^{1/2})/\tilde{h}^{1/2}$ as long as $\tilde{h} > 0$. Consequently, we can interpret the effect c^{f} $\tilde{h}$ as simply to standardize the data and the effect of $\tilde{\mu}$ as to choose an appropriate ψ-function according to the tailthickness of the distribution. For $\tilde{\mu}\tilde{h} \equiv 0$, eq. (3.3) gives the non-linear least-squares estimator. This is, of course, to be expected since the normal density can be obtained as a limiting case from the Student-t family for $v \to \infty$. Also the Cauchy ML estimator and Ohlsen's robust estimator considered in Andrews et al. (1972) are members of the class of estimators defined by (3.3). The ψ-functions defined in (3.2) are bounded and redescending for $a > 0$ and unbounded for $a = 0$.

In case of a linear model, i.e., $y_t = x_t\beta + u_t$, we can write (3.3) as

$$\bar{\beta} = \left[\sum_{t=1}^{T} x_t'\bar{w}_t x_t\right]^{-1} \sum_{t=1}^{T} x_t'\bar{w}_t y_t, \qquad (3.4)$$

with

$$\bar{w}_t = \left(1 + \tilde{\mu}\tilde{h}\bar{u}_t^2\right)^{-1} \quad \text{and} \quad \bar{u}_t = y_t - x_t\bar{\beta}.$$

The estimator $\bar{\beta}$ can then be interpreted as a weighted least-squares estimator, where for $\tilde{\mu}\tilde{h} > 0$ the weights $\bar{w}_t$ have the effect of giving more (less) weight to observations with relatively small (large) residuals.

Eq. (3.3) is non-linear and has to be solved iteratively. In the rest of the paper we concentrate on the one-step Gauss–Newton approximation. It will be shown below that the resulting estimator is not inferior to the root of (3.3). For a formal definition of the estimator we will maintain in the following always either Assumption 1 or 1′.

Assumption 1. There is an open neighborhood Θ_B^* of the parameter space $\Theta_B \subseteq \mathbb{R}^K$ such that $g(x, \beta)$, $(\partial/\partial\beta)g(x, \beta)$ and $(\partial^2/\partial\beta\partial\beta')g(x, \beta)$ are continuous on $\mathbb{R}^L \times \Theta_B^*$.

Assumption 1′. There is an open neighborhood Θ_B^* of the parameter space $\Theta_B \subseteq \mathbb{R}^K$ such that $g(x,\beta)$, $(\partial/\partial\beta)g(x,\beta)$ and $(\partial^2/\partial\beta\partial\beta')g(x,\beta)$ are continuous on Θ_B^* for all $x \in \mathbb{R}^L$ and measurable in x for each $\beta \in \Theta_B^*$.

Given those assumptions we can now define a class of one-step M-estimators corresponding to (3.3). Note that $\tilde{\mu}$ and $\tilde{h}$ enter those normal equations only in terms of their product. It hence proves convenient to define $\tilde{a} = \tilde{\mu}\tilde{h}$ and $a = \mu h$; furthermore $\tilde{\theta} = (\tilde{a}, \tilde{\beta}')'$ and $\theta = (a, \beta')'$.

Definition 1. Let $\tilde{\beta} \in \mathbb{R}^K$ be some initial estimator and let the estimator $\tilde{a} \in [0, \infty)$. Define

$$A_T(\theta) = (1/T)\sum_{t=1}^{T} A(t,\theta),$$

and

$$r_T(\theta) = (1/T)\sum_{t=1}^{T} r(t,\theta),$$

where

$$A(t,\theta) = -\psi_a'(y_t - g(x_t,\beta))\frac{\partial g(x_t,\beta)}{\partial\beta'}\frac{\partial g(x_t,\beta)}{\partial\beta} + \psi_a(y_t - g(x_t,\beta))\frac{\partial^2 g(x_t,\beta)}{\partial\beta\partial\beta'}, \tag{3.5a}$$

and

$$r(t,\theta) = \psi_a(y_t - g(x_t,\beta))\frac{\partial g(x_t,\beta)}{\partial\beta'}, \tag{3.5b}$$

for

$$\theta \in [0,\infty) \times \Theta_B^* \quad \text{and} \quad \psi_a'(z) = (\partial/\partial z)\psi_a(z).$$

Then the one-step M-estimator $\hat{\beta}$ corresponding to $\psi_{\tilde{a}}$ is defined as

$$\hat{\beta} = \tilde{\beta} - [A_T(\tilde{\theta})]^{-1} r_T(\tilde{\theta}), \tag{3.5c}$$

if $A_T(\tilde{\theta})$ is non-singular and $\tilde{\beta} \in \Theta_B^*$; otherwise $\hat{\beta} = \tilde{\beta}$.

Note that $r_T(\theta)$ and $A_T(\theta)$ correspond (up to a proportionality factor) to the first- and second-order derivatives of the (pseudo) Student-t log-likelihood function. For the linear model, (3.5) simplifies to

$$\hat{\beta} = \tilde{\beta} + \left[(1/T) \sum_{t=1}^{T} x_t' \left(\tilde{w}_t - 2\tilde{\mu}\tilde{h}\tilde{w}_t^2 \tilde{u}_t^2 \right) x_t \right]^{-1} (1/T) \sum_{t=1}^{T} x_t' \tilde{w}_t \tilde{u}_t, \tag{3.6}$$

where

$$\tilde{w}_t = \left(1 + \tilde{\mu}\tilde{h}\tilde{u}_t^2\right)^{-1} \quad \text{and} \quad \tilde{u}_t = y_t - x_t \tilde{\beta}.$$

The following remark only pertains to the linear model: If $\tilde{\beta}$ is scale- and shift-equivariant [i.e., $\tilde{\beta}(\gamma y + X\delta) = \gamma\tilde{\beta}(y) + \delta$ for $\gamma \in \mathbb{R}$ and $\delta \in \mathbb{R}^K$], if $\tilde{h}$ is scale-equivariant and shift-invariant [i.e., $\tilde{h}(\gamma y + X\delta) = \gamma^{-2}\tilde{h}(y)$], and if $\tilde{\mu}$ is scale- and shift-invariant [i.e., $\tilde{\mu}(\gamma y + X\delta) = \tilde{\mu}(y)$], then it is readily seen that $\hat{\beta}$ is also scale- and shift-equivariant.

3.2. Asymptotic properties

In this subsection we derive the asymptotic properties of $\hat{\beta}$ assuming that $\tilde{\beta}$ converges to some $\beta^\circ \in \Theta_B$ and $\tilde{a}$ converges to some $a^\circ \in [0, \infty)$. In our analysis we consider the following three alternative assumptions concerning the stochastic nature of the data-generating process.[6]

Assumption 2. The process (y_t, x_t) is of the form $y_t = \tau_y(\xi_t)$ and $x_t = \tau_x(\xi_t)$ where $\tau = [\tau_y, \tau_x]$: $\mathbb{R}^S \to \mathbb{R}^{L+1}$ is a continuous function and ξ_t is a stochastically stable process with respect to an α-mixing base. Furthermore, if Λ_t is the distribution function of ξ_t, the $(1/T)\sum_{t=1}^T \Lambda_t \to \Lambda$ properly.[7]

Assumption 2′. The process (y_t, x_t) is strictly stationary and ergodic.

We note that every strictly stationary process satisfying Assumption 2 is ergodic. The following condition is stronger than Assumption 2 and is used for a strong consistency result.

Assumption 2″. The process (y_t, x_t) is for some $r \geq 1$ [$r > 1$] and some $\delta > 0$ ϕ-mixing with $\phi(m) = O(m^{-\lambda})$ for $\lambda > r/(2r-1)$ [α-mixing with $\alpha(m) = O(m^{-\lambda})$ for $\lambda > r/(r-1)$]. Furthermore, if Δ_t is the distribution function of (y_t, x_t), then $(1/T)\sum_{t=1}^T \Delta_t \to \Delta$ properly.

The coefficients $\alpha(m)$ and $\phi(m)$ are defined as follows. Let (z_t) be a stochastic process, let $\mathscr{F}_{-\infty}^n$ be the σ-algebra generated by $z_n, z_{n-1}, \ldots$, and

[6] The index set of the data-generating process is either $\mathbb{N}$ or $\mathbb{Z}$.

[7] We use the term proper convergence as defined in Feller (1971, p. 248). Let Δ_t be the distribution function of $z_t = (y_t, x_t)$, then Assumption 2 implies that $(1/T)\sum_{t=1}^T \Delta_t \to \Delta$ properly with $\Delta = \Lambda \circ \tau^{-1}$. Hence we have for any integral $\int f(\tau(\xi))\mathrm{d}\Lambda(\xi) = \int f(z)\mathrm{d}\Delta(z)$.

let $\mathscr{F}_k^\infty$ be the σ-algebra generated by $z_k, z_{k+1}, \ldots$, then

$$\alpha(m) = \sup_k \sup\{|P(F \cap G) - P(F)P(G)|: F \in \mathscr{F}_{-\infty}^k, \quad G \in \mathscr{F}_{k+m}^\infty\},$$

and

$$\phi(m) = \sup_k \sup\{|P(G|F) - P(G)|: F \in \mathscr{F}_{-\infty}^k, \quad G \in \mathscr{F}_{k+m}^\infty, \quad P(F) > 0\}.$$

The coefficients $\alpha(m)$ and $\phi(m)$ are measures for the memory of the process (z_t). If $\phi(m)$ $[\alpha(m)]$ goes to zero, then the process is called ϕ-mixing [α-mixing]. Clearly every ϕ-mixing process is also α-mixing. Examples of ϕ-mixing processes are independent and m-dependent processes. Examples of α-mixing processes are strictly invertible Gaussian ARMA processes. For additional discussion of these concepts, see, e.g., White and Domowitz (1984). The concept of stochastically stable processes with respect to a base was introduced by Bierens (1981, 1984a). Since every process is stochastically stable with respect to itself the class of stochastically stable processes with respect to an α-mixing base is even wider than the class of α-mixing processes.

The following assumption represents the usual orthogonality condition between regressors and disturbances $u_t = y_t - g(x_t, \beta^\circ)$.

Assumption 3. The regressors x_t and disturbances u_t are stochastically independent for all t.

Corresponding to, respectively, Assumptions 2, 2′ and 2″ we will furthermore employ one of the moment conditions listed below. Those conditions are based on the expressions

$$G_t(\gamma) = \mathrm{E} \sup_{\beta \in U(\beta^\circ)} \|[(\partial/\partial\beta')g(x_t, \beta)][(\partial/\partial\beta)g(x_t, \beta)]\|^\gamma, \tag{3.7a}$$

$$D_t(\gamma) = \mathrm{E} \sup_{\beta \in U(\beta^\circ)} \|(\partial^2/\partial\beta\partial\beta')g(x_t, \beta)\|^\gamma, \tag{3.7b}$$

$$F_t(\gamma) = \mathrm{E} \sup_{\beta \in U(\beta^\circ)} \|[g(x_t, \beta) - g(x_t, \beta^\circ)][(\partial^2/\partial\beta\partial\beta')g(x_t, \beta)]\|^\gamma, \tag{3.7c}$$

where $U(\beta^\circ) \subseteq \Theta_B^*$ is some compact neighborhood of β° on which the following conditions are assumed to hold. [Note that in the linear model $G_t(\gamma) = \mathrm{E}\|x_t'x_t\|^\gamma$, $D_t(\gamma) = 0$ and $F_t(\gamma) = 0$.]

Assumption 4. $\sup_{T\geq 1}(1/T)\sum_{t=1}^{T}G_t(1+\delta)<\infty$ and $\sup_{T\geq 1}(1/T)\sum_{t=1}^{T}D_t(1+\delta)<\infty$ for some $\delta>0$; if $a^{\circ}=0$, we require furthermore that $\sup_{T\geq 1}(1/T)\sum_{t=1}^{T}F_t(1+\delta)<\infty$ and $\sup_{t\geq 1}\mathrm{E}|u_t|^{1+\delta}<\infty$.

Assumption 4′. $G_t(1)<\infty$ and $D_t(1)<\infty$; if $a^{\circ}=0$, we require furthermore that $F_t(1)<\infty$ and $\mathrm{E}|u_t|<\infty$.

The following assumption is again stronger than Assumption 4.

Assumption 4″. $\sup_{t\geq 1}G_t(r+\delta)<\infty$ and $\sup_{t\geq 1}D_t(r+\delta)<\infty$ for the same r as in Assumption 2″ and some $\delta>0$; if $a^{\circ}=0$, we require furthermore that $\sup_{t\geq 1}F_t(r+\delta)<\infty$ and $\sup_{t\geq 1}\mathrm{E}|u_t|^{r+\delta}<\infty$.

The above sets of assumptions cover a wide class of non-linear regression models with and without lagged dependent variables; fixed regressors are of course included as a special case. As a point of interest we note that under Assumption 1 the process (y_t, x_t) satisfies Assumptions 2 or 2″ iff the same is true for the process (u_t, x_t). An analogous statement can be made with respect to Assumption 2′ given Assumption 1′ holds. Together with the above sets of assumptions either one of the following two conditions can be used to ensure the consistency of $\hat{\beta}$ (at least for the 'slope' parameters):

Assumption A. The disturbances u_t are distributed symmetrically around zero.

Assumption B. (a) The disturbances u_t and/or the regressors x_t are distributed identically over time. (b) The response function is of the form $g(x,\beta)=\beta_1+g_*(x,\beta_*)$ with $\beta=(\beta_1,\beta_*')'$.

In the subsequent we will need the following asymptotic analogs of $A_T(\theta)$, $B_T(\theta)=(1/T)\sum_{t=1}^{T}r(t,\theta)r(t,\theta)'$ and $r_T(\theta)$:

$$A(\theta)=\int\left\{-\psi_a'(y-g(x,\beta))\frac{\partial g(x,\beta)}{\partial\beta'}\frac{\partial g(x,\beta)}{\partial\beta}\right.$$
$$\left.+\psi_a(y-g(x,\beta))\frac{\partial^2 g(x,\beta)}{\partial\beta\partial\beta'}\right\}\mathrm{d}\Delta(y,x), \tag{3.8a}$$

$$B(\theta)=\int[\psi_a(y-g(x,\beta))]^2\frac{\partial g(x,\beta)}{\partial\beta'}\frac{\partial g(x,\beta)}{\partial\beta}\mathrm{d}\Delta(y,x), \tag{3.8b}$$

$$r(\theta)=\int\psi_a(y-g(x,\beta))\frac{\partial g(x,\beta)}{\partial\beta'}\mathrm{d}\Delta(y,x). \tag{3.8c}$$

The proof of the following theorem is given in appendix A. For ease of presentation we refer in the subsequent to Assumption 3 also as 3′ and 3″ and to Assumption 1 also as 1″.

Theorem 1 (*Consistency*). *Let $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta}')'$ converge in probability* [*almost surely*] *to $\theta^{\circ} = (a^{\circ}, \beta^{\circ\prime})'$ with $a^{\circ} \geq 0$ and $\beta^{\circ} \in \Theta_B$. Given Assumptions 1–4 or 1′–4′* [*Assumptions 1′–4′ or 1″–4″*] *are satisfied, given the matrix $A(\theta^{\circ})$ is non-singular and*

(*a*) *Assumption A holds, then $\hat{\beta}$ converges in probability* [*almost surely*] *to β°,*

(*b*) *Assumption B holds, then $\hat{\beta}_*$ converges in probability* [*almost surely*] *to β°_*.*

Remark 1. In appendix A we actually derive, as a by-product to the proof of Theorem 1, a more general consistency result for a wider class of ψ-functions under somewhat weaker but more complex assumptions.[8] In the context of Theorem 1, but without Assumptions A or B, the general expression for the limit of $\hat{\beta}$ is given by $\beta^{\circ} - A(\theta^{\circ})^{-1} r(\theta^{\circ})$. Clearly the asymptotic bias is zero iff β° satisfies the asymptotic normal equations corresponding to (3.3) i.e., $r(\theta^{\circ}) = 0$; under the assumption of the theorem this is the case if either $\mathrm{E}[\psi_{a^{\circ}}(u_t)] = 0$ or $\mathrm{E}[(\partial/\partial\beta) g(x_t, \beta^{\circ})] = 0$ for all t. Part (a) of the theorem follows now from the fact that the former condition is satisfied in the case of symmetrically distributed disturbances as postulated in Assumption A [since the ψ-function (3.2) is antisymmetric]. Part (b) of the theorem follows from the fact that under Assumption B the bias turns out to be a multiple of $(1, 0, \ldots, 0)'$.

Remark 2. For $a^{\circ} > 0$, Theorem 1 does not require the existence of any moments of the disturbance process since in this case the ψ-function and its derivative are bounded. For $a^{\circ} = 0$, the function $\psi_{\tilde{a}}(z)$ converges to $\psi_0(z) = z$ which is the ψ-function of the non-linear least-squares estimator. The condition $\mathrm{E}|u_t| < \infty$ postulated in Assumption 4′ for stationary and ergodic processes is hence minimal. The somewhat higher moment requirements postulated in Assumptions 4 and 4″ may be viewed as trade-offs to the more general provisions in terms of the heterogeneity and interdependence of the corresponding data-generating processes.

Remark 3. Theorem 1 does not maintain an explicit identifiability condition. This point deserves some discussion. Note that if a data-generating process (y_t, x_t) can be written in the form $y_t = g_t(x_t) + u_t$ where u_t and x_t are

[8] The convergence result in appendix A is such that it also allows for an interpretation of estimation under misspecification.

independent, then the response function $g_t(x_t)$ is within these representations unique up to additive constants c_t (almost surely). Furthermore, if it is assumed that the disturbances are distributed symmetrically around zero, then $c_t = 0$;[9] if the disturbances are assumed to be identically distributed, then $c_t = c$. Hence in the context of model (2.1) Assumptions 3 and A and a separating condition like

$$\int (g(x,\beta) - g(x,\beta^{\circ}))^2 \,\mathrm{d}\Delta(x) = 0 \quad \text{iff} \quad \beta = \beta^{\circ}, \tag{3.9}$$

ensure the identifiability of β°. Similarly, Assumptions 3 and B and the above separating condition ensure the identifiability of β_*°. Theorem 1 can be formulated without such a separating condition since the assumptions that the starting estimator $\tilde{\beta}$ converges to β° and that $A(\theta^{\circ})$ is non-singular are essentially substitutes for such a condition.

To derive the asymptotic normality of $\hat{\beta}$ we employ the following additional assumptions. Assumption 5 is a strengthening of Assumption 3.

Assumption 5. The disturbances u_t are stochastically independent of $u_{t-1}, u_{t-2}, \dots, x_t, x_{t-1}, \dots,$ for all t.

Similarly as before we will refer to Assumption 5, for ease of presentation, also as 5′ and 5″. The following assumptions extend Assumptions 4, 4′ and 4″.

Assumption 6. If $a^{\circ} = 0$, we require additionally that $\sup_{t \geq 1} \mathrm{E}|u_t|^{3+\delta} < \infty$ for some $\delta > 0$ and $\sup_{T \geq 1}(1/T)\sum_{t=1}^{T} G_t(2+\delta) < \infty$.

Assumption 6′. If $a^{\circ} = 0$, we require additionally that $\mathrm{E}|u_t|^3 < \infty$ and $G_t(2) < \infty$.

Assumption 6″. If $a^{\circ} = 0$, we require additionally for the same r as in Assumption 2″ that $\sup_{t \geq 1} \mathrm{E}|u_t|^{3r+\delta} < \infty$ for some $\delta > 0$ and $\sup_{t \geq 1} G_t(2r+\delta) < \infty$.

The proof of the following theorem is given in appendix A.

Theorem 2 (Asymptotic Normality). Let $\tilde{\theta} = \theta^{\circ} + \mathrm{O}_p(T^{-1/2})$, where $\tilde{\theta} = (\tilde{a}, \tilde{\beta}')'$ and $\theta^{\circ} = (a^{\circ}, \beta^{\circ\prime})'$ with $a^{\circ} \geq 0$ and $\beta^{\circ} \in \Theta_B$. Given either Assumptions

[9] This would also be true if the symmetry assumption is replaced by an assumption such as, e.g., $\mathrm{E}(u_t) = 0$.

1–6 or 1′–6′ or 1″–6″ are satisfied, given the matrix $A(\theta^{\circ})$ is non-singular, and

(*a*) *Assumption A holds, then $\sqrt{T}(\hat{\beta}-\beta^{\circ})$ is asymptotically normal with mean zero and variance–covariance matrix $\Phi = A(\theta^{\circ})^{-1}B(\theta^{\circ})A(\theta^{\circ})^{-1}$.*

(*b*) *Assumption B holds and* $\mathrm{E}[\psi_{a^{\circ}}(u_t)]=0$, *then $\sqrt{T}(\hat{\beta}_{*}-\beta_{*}^{\circ})$ is asymptotically normal with zero mean and variance–covariance matrix $\Phi_{*} = [\partial\beta_{*}/\partial\beta]\Phi[\partial\beta_{*}/\partial\beta']$.*

Remark 4. In appendix A we derive again, as a by-product to the proof of Theorem 2, a more general asymptotic normality result for a wider class of ψ-functions under somewhat weaker but more complex assumptions. Given $\mathrm{E}[\psi_{a^{\circ}}(u_t)]=0$, we show in the context of Theorem 2 that $\sqrt{T}(\hat{\beta}-\beta^{\circ})$ is asymptotically equivalent to a normally distributed random vector with zero mean and variance–covariance matrix Φ plus a term of the form $\sqrt{T}(\tilde{a}-a^{\circ})A(\theta^{\circ})^{-1}p(\theta^{\circ})$, where

$$p(\theta^{\circ}) = \int [(\partial/\partial a)\psi_{a^{\circ}}(y-g(x,\beta^{\circ}))][(\partial/\partial\beta')g(x,\beta^{\circ})]\mathrm{d}\Delta(y,x)$$

$$= \lim (1/T)\sum_{t=1}^{T} \mathrm{E}[(\partial/\partial a)\psi_{a^{\circ}}(u_t)]\mathrm{E}[(\partial/\partial\beta')g(x_t,\beta^{\circ})]. \tag{3.10}$$

The estimator $\tilde{a}$ has no effect on the joint asymptotic distribution of the elements of $\hat{\beta}$ if $p(\theta^{\circ})=0$.[10] Within the framework of pseudo ML estimation this corresponds to the familiar orthogonality condition $\mathrm{E}[\partial^2\mathscr{L}/\partial a\partial\beta']=0$, where $\mathscr{L}$ is the pseudo log-likelihood function. Part (a) of the theorem follows since for symmetrically distributed disturbances as postulated in Assumption A we have $\mathrm{E}[(\partial/\partial a)\psi_{a^{\circ}}(u_t)]=0$ and hence $p(\theta^{\circ})=0$. Part (b) of the theorem follows since under Assumption B the vector $A(\theta^{\circ})^{-1}p(\theta^{\circ})$ turns out to be a multiple of $(1,0,\ldots,0)'$.

Remark 5. (i) For $a^{\circ}>0$, Theorem 2 does not postulate the existence of any moments of the disturbance process. For $a^{\circ}=0$, we have $\mathrm{E}[(\partial/\partial a)\psi_{a^{\circ}}(u_t)] = \mathrm{E}(u_t^3)$. Because of (3.10) it then follows that the condition $\mathrm{E}|u_t|^3<\infty$ postulated in Assumption 6′ for stationary and ergodic processes is minimal. The somewhat higher moment requirements postulated in Assumption 6 and 6″ may again be viewed as trade-offs to the more general provisions in terms of heterogeneity and interdependence of the corresponding data-generating pro-

[10] This is, of course, also true if $\tilde{a}\equiv a^{\circ}$.

cess. (ii) As a referee pointed out, for symmetric disturbances it is possible to relax the assumption that $\tilde{a}$ is $\sqrt{T}$-consistent to $\tilde{a}$ is consistent at the expense of higher moment requirements.

Remark 6. The results in appendix A imply that $\tilde{\Phi} = [A_T(\tilde{\theta})]^{-1} \cdot B_T(\tilde{\theta})[A_T(\tilde{\theta})]^{-1}$ is, under the assumptions of Theorem 2, a consistent estimator for Φ. Given the disturbances are identically and symmetrically distributed the matrix Φ reduces to $\{\mathrm{E}(\psi^2_{a^\circ}(u_t))/[\mathrm{E}(\psi'_{a^\circ}(u_t))]^2\}Q(\beta^\circ)^{-1}$, where $Q(\beta^\circ) = \int[(\partial/\partial\beta')g(x,\beta)][(\partial/\partial\beta)g(x,\beta)]\,\mathrm{d}\Delta$;[11] the estimator $\tilde{\Phi}$ simplifies analogously.

Remark 7. (i) The assumption $\mathrm{E}[\psi_{a^\circ}(u_t)] = 0$ weakens the assumption that the disturbances are symmetrically distributed. This condition is together with Assumption 5 essentially the condition that the score of the pseudo log-likelihood function is a martingale difference. Such a condition is often used to prove the asymptotic normality of the pseudo ML estimator. The pseudo ML estimator satisfies in a trivial way the definitorial equation of the class of one-step *M*-estimators considered here. Given the pseudo ML estimator is $\sqrt{T}$-consistent, Theorem 2 implies that it has the same asymptotic properties as the one-step *M*-estimators $\hat{\beta}$ covered by the theorem. (ii) The assumption $\mathrm{E}[\psi_{a^\circ}(u_t)] = 0$ can be dropped in part (b) of Theorem 2 in the case of a linear model with i.i.d. disturbances if the estimator $\hat{\beta}$ in (3.6) is modified to $\hat{\hat{\beta}} = \tilde{\beta} + \{[(1/T)\sum(\tilde{w}_t - 2\tilde{a}\tilde{w}_t^2\tilde{u}_t^2)][(1/T)\sum x_t'x_t]\}^{-1}(1/T)\sum x_t'\tilde{w}_t\tilde{u}_t$ and the matrix Φ is changed to $\{\mathrm{E}[\psi_{a^\circ}(u_t) - \mathrm{E}\psi_{a^\circ}(u_t)]^2/[\mathrm{E}\psi'_{a^\circ}(u_t)]^2\}Q(\beta^\circ)^{-1}$. Clearly the estimators $\hat{\hat{\beta}}$ and $\hat{\beta}$ have the same asymptotic distribution if $\mathrm{E}[\psi_{a^\circ}(u_t)] = 0$.

Remark 8. *M*-estimators of type I that are based on non-convex ρ-functions need not be consistent for general disturbance distributions. See Freedman and Diaconis (1982); their example is formulated within the simple context of a location model, $y_t = \beta + u_t$. Given a non-convex symmetric ρ-function they show that for a certain class of symmetric distributions the 'asymptotic' criterion function $\mathrm{E}[\rho(y-\beta)]$ has two absolute minima, and a local maximum at the true parameter value, yielding inconsistency of the *M*-estimator even though $\mathrm{E}[\rho'(u)] = 0$. Contrarily, we note that one-step *M*-estimators, as considered in this paper, remain consistent given they are started consistently. If we use an *M*-estimator of type I as a starting estimator, then the above remark suggests further that that estimator should be based on a convex ρ-function which will yield (under weak regularity conditions) consistency and asymptotic normality – see Burguete, Gallant and Souza (1982) and Grossman (1976).

[11] The matrix $A(\theta^\circ)$ reduces to $\mathrm{E}[\psi'_{a^\circ}(u_t)]Q(\beta^\circ)$. Given the non-singularity of $Q(\beta^\circ)$ the matrix $A(\theta^\circ)$ is non-singular iff $\mathrm{E}[\psi'_{a^\circ}(u_t)] \neq 0$. It follows from Bierens (1981, p. 75) that this is the case for all symmetric disturbance distributions with unimodal differentiable densities.

3.3. Specification and estimation of the adaptation parameter

In the following we consider two alternative specifications for the estimator of the adaptation parameter. We consider in particular specifications of the form $\tilde{a} = \tilde{\mu}\tilde{h}$ where $\tilde{\mu}$ and $\tilde{h}$ are estimators of, respectively, measures of the tailthickness and the (inverse) spread of the disturbance distribution. The following analysis maintains that the disturbances are identically distributed with common distribution function F.[12] The approach taken here is to express the parameters of the Student-t distribution H and v in terms of estimable characteristics such as moments or quantiles of the disturbance distribution. We then use the obtained expressions for H and v^{-1} to define functionals $h(F)$ and $\mu(F)$ that measure the (inverse) scale and tailthickness for *general* distributions F. Suppose we have $\sqrt{T}$-consistent estimators for those functionals. Since by construction $h(F)$ and $\mu(F)$ will coincide with H and v^{-1} on the family of Student-t distributions, it then follows from the above analysis that the corresponding one-step M-estimator $\hat{\beta}$ is asymptotically efficient on the family of Student-t distributions and (as a limiting case) for the normal distribution.

In the following we write $\sigma_\alpha = \sigma_\alpha(F) = \mathrm{E}(|u_t|^\alpha)$. Consider for a moment the Student-t distribution that underlies (3.1). For $v > 2$ we have the following relationship between its moments:

$$\frac{\sigma_2}{\sigma_1^2} = \frac{\pi}{v-2}\,\frac{\Gamma[v/2]^2}{\Gamma[(v-1)/2]^2} = p_1(v). \tag{3.11}$$

For $v > 1$, we further have

$$H = \frac{1}{\pi}\,\frac{v}{\sigma_1^2}\,\frac{\Gamma[(v-1)/2]^2}{\Gamma[v/2]^2} = q_1(v, \sigma_1), \tag{3.12}$$

$$\frac{\sigma_1}{\sigma_{1/2}^2} = \frac{\pi^{1/2}}{\Gamma[3/4]^2}\cdot\frac{\Gamma[v/2]\Gamma[(v-1)/2]}{\Gamma[(2v-1)/4]^2} = p_2(v), \tag{3.13}$$

$$H = \frac{\Gamma[3/4]^4}{\pi^2}\cdot\frac{v}{\sigma_{1/2}^4}\cdot\frac{\Gamma[(2v-1)/4]^4}{\Gamma[v/2]^4} = q_2(v, \sigma_{1/2}). \tag{3.14}$$

The above formulas can be readily obtained from results given, e.g., in Johnson and Kotz (1970). In appendix B we prove the following lemma which

[12] The estimators considered remain well defined even if that assumption is not satisfied. Degenerate distributions concentrated at one point are excluded from the subsequent discussion.

ensures that the subsequent definitions of functionals $h(F)$ and $\mu(F)$ are proper.

Lemma 1. The functions $p_1(v)$ and $p_2(v)$ are analytic and monotonically decreasing on, respectively, $(2, \infty)$ and $(1, \infty)$ with $p_1(2+) = p_2(1+) = \infty$ and $p_1(\infty) = \pi/2$, $p_2(\infty) = \sqrt{\pi}/\Gamma(3/4)^2$. The functions $q_1(v, \sigma_1)$ and $q_2(v, \sigma_{1/2})$ are analytic, respectively, in an open neighborhood of $[2, \infty) \times (0, \infty)$ and $[1, \infty) \times (0, \infty)$ with $q_1(\infty, \sigma_1) = 2/(\pi\sigma_1^2)$, $q_2(\infty, \sigma_{1/2}) = 2\Gamma(3/4)^4/(\pi\sigma_{1/2}^4)$.

We now use the relationship (3.11) to define, for general distributions F, the functional

$$\begin{aligned} \mu_1(F) &= \left[p_1^{-1}(\sigma_2/\sigma_1^2)\right]^{-1} \quad \text{if} \quad \sigma_2/\sigma_1^2 > \pi/2, \\ &= 0 \qquad\qquad\qquad\quad \text{otherwise}. \end{aligned} \tag{3.15}$$

That is, for distributions with ratios σ_2/σ_1^2 bigger than the value $\pi/2$, corresponding to the normal distribution, we set $\mu_1(F)$ equal to the inverse of the – because of Lemma 1 – unique solution of (3.11); otherwise we set $\mu_1(F)$ equal to zero. Analogously we can use (3.13) to define the alternative functional

$$\begin{aligned} \mu_2(F) &= \left[p_2^{-1}(\sigma_1/\sigma_{1/2}^2)\right]^{-1} \quad \text{if} \quad \sigma_1/\sigma_{1/2}^2 > \sqrt{\pi}\,\Gamma(3/4)^2, \\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise}. \end{aligned} \tag{3.16}$$

We adopt the convention $\sigma_\alpha/\sigma_{\alpha/2}^2 = \infty$ for $\alpha = 1,2$ if $\sigma_\alpha = \infty$, even if $\sigma_{\alpha/2} = \infty$, since then in all cases of interest – as is shown in appendix B – also the corresponding ratios of moment estimators converge to infinity. Hence $\mu_1(F)$ and $\mu_2(F)$ are well defined for all distributions. Using (3.12) and (3.14) we define functionals

$$h_1(F) = q_1(1/\mu_1(F), \sigma_1), \qquad h_2(F) = q_2(1/\mu_2(F), \sigma_{1/2}). \tag{3.17}$$

Note from Lemma 1 that $h_i(F)$ is well defined also for $\mu_i(F) = 0$, $i = 1,2$. Clearly, by design the functionals $\mu_i(F)$ and $h_i(F)$, $i = 1,2$, correspond to v^{-1} and H on the Student-t family.

Estimators of the above functionals are obtained by replacing in (3.15)–(3.17) the true moments σ_α by the estimators $\tilde{s}_\alpha = T^{-1}\sum_{t=1}^{T}|\tilde{u}_t|^\alpha$ with $\tilde{u}_t = y_t - g(x_t, \tilde{\beta})$. We denote those estimators by $\tilde{\mu}_i$ and $\tilde{h}_i$, $i = 1,2$. Note that due to the monotonicity of $p_i(\cdot)$ it is not difficult to compute those estimators numerically. The following theorem shows under which conditions $\tilde{\mu}_i \tilde{h}_i$, $i = 1,2$, satisfy the requirements of, respectively, Theorem 1 and Theorem 2. The proof is given in appendix B.

Theorem 3. Given Assumptions 1, 2, 4 or 1′, 2′, 4′ [Assumptions 1′, 2′, 4′, or 1″, 2″, 4″] are satisfied, given $\tilde{\beta}$ converges in probability [almost surely] to β° and the disturbance process (u_t) is strictly stationary and ergodic.

(*a*) *If $\sigma_{\gamma} < \infty$ for some $\gamma > 0$, then $\tilde{\mu}_1\tilde{h}_1$ and $\tilde{\mu}_2\tilde{h}_2$ are consistent [strongly consistent] for $\mu_1(F)h_1(F)$ and $\mu_2(F)h_2(F)$. (Note that the products $\mu_i(F)h_i(F)$ are always well defined.)*

(*b*) *Suppose $\tilde{\beta} = \beta^{\circ} + O_p(T^{-1/2})$ and the disturbances (u_t) are i.i.d. If $\sigma_4 < \infty$, then $\tilde{\mu}_1\tilde{h}_1 = \mu_1(F)h_1(F) + O_p(T^{-1/2})$; if $\sigma_2 < \infty$, F has a density that is essentially bounded in a neighborhood of zero and*

$$\sup_{t \geq 1}\sup_{\beta \in U(\beta^{\circ})}\{\|(\partial/\partial\beta)g(x_t,\beta)\|\} < M$$

where M is some fixed constant, then $\tilde{\mu}_2\tilde{h}_2 = \mu_2(F)h_2(F) + O_p(T^{-1/2})$.

From a practical point of view the estimators $\tilde{h}_i$ and $\tilde{\mu}_i$, $i = 1,2$, are appealing because they are not difficult to compute. A more robust choice of functionals could be obtained by expressing the parameters h and v of the Student-t distribution in terms of quantiles instead of moments and then by proceeding as above. However, computationally this approach is more involved. Alternatively, we could define $\tilde{\mu}\tilde{h}$ as the minimizing values of an estimator of the asymptotic variance–covariance matrix of $\hat{\beta}$ (or respectively $\hat{\beta}_*$) – however, this is left for future research.[13]

4. Comparison in terms of asymptotic efficiency

In the following we compare the asymptotic variance–covariance matrix of $\hat{\beta}$ for the parametrizations $\tilde{\mu}\tilde{h} = \tilde{\mu}_1\tilde{h}_1$ and $\tilde{\mu}\tilde{h} = \tilde{\mu}_2\tilde{h}_2$, say $\hat{\beta}_{S(1)}$ and $\hat{\beta}_{S(2)}$, with that of the Huber M-estimator based on the ψ-function ψ^c, say $\hat{\beta}_{H(c)}$, with that of the least-absolute-deviation (LAD) estimator, say $\hat{\beta}_{LAD}$, and with that of the least-squares estimator, say $\hat{\beta}_{LS}$.[14] The asymptotic distribution for the LAD estimator has, to the best of our knowledge, only been established for the linear regression model with fixed regressors and i.i.d. disturbances. For ease of presentation we maintain that assumption for all of the subsequent comparisons. We maintain furthermore that the regressors are bounded and consider a variety of symmetric distributions F with density f. The asymptotic variance–covariance matrix is then for all estimators and distributions considered proportional to the matrix Q^{-1}. We can, hence, concentrate on the

[13] For a related approach for a different class of M-estimators, see Yohai (1974) and Bierens (1981). In case of asymmetric distributions we may also consider to minimize the variance–covariance matrix of $\hat{\beta}_*$ with respect to the intercept, which, under asymmetry, typically has no natural meaning.

[14] The Huber ψ-function ψ^c is defined in section 2.

respective proportionality factor.[15] More specifically, $\sqrt{T}(\hat{\beta}_d - \beta^o)$ converges for $d = S(1)$, $S(2)$, $H(c)$, LAD and LS in distribution to a normal random vector with zero mean and variance–covariance matrix $\kappa_d Q^{-1}$, where

$$\kappa_{S(i)} = \mathrm{E}\left(w_{(i)}^2 u^2\right) / \left\{\mathrm{E}(w_{(i)}) - 2\mathrm{E}\left[\mu_i(F) h_i(F) w_{(i)}^2 u^2\right]\right\}^2, \tag{4.1a}$$

$$w_{(i)} = \left[1 + \mu_i(F) h_i(F) u^2\right]^{-1}, \qquad i = 1, 2,$$

$$\kappa_{H(c)} = \left\{\int_{-c}^{c} u^2 f(u)\,\mathrm{d}u + c^2\left[1 - \int_{-c}^{c} f(u)\,\mathrm{d}u\right]\right\} \Big/ \left[\int_{-c}^{c} f(u)\,\mathrm{d}u\right]^2, \tag{4.1b}$$

$$\kappa_{LAD} = [2f(0)]^{-2}, \qquad \kappa_{LS} = \sigma_2. \tag{4.1c}$$

In computing, respectively, $\kappa_{S(1)}$ and $\kappa_{S(2)}$ for a particular distribution F we first have to calculate the moments σ_1, σ_2 and $\sigma_{1/2}, \sigma_1$. Substituting the obtained moments into, respectively, (3.15), (3.17) and (3.16), (3.17), we obtain the corresponding values for $\mu_1(F), h_1(F)$ and $\mu_2(F), h_2(F)$. Given those values, we can then calculate the respective expectations in (4.1a).

Rather than to evaluate the integrals (expectations) appearing in the expressions of the above proportionality factors analytically, we calculated them by numerical integration techniques. The accuracy of the numerical procedure was checked for the Student-t distribution with $v > 2$ by comparing the obtained value for $\kappa_{S(i)}/\sigma_2$ with the analytically implied value $(v-2)(v-3)/(v(v+1))$. It was found that the numerical results were accurate up to the first five decimal places.

In table 1 we compare the asymptotic efficiency of all of the above mentioned estimators relative to the estimator $\hat{\beta}_{S(2)}$ in terms of the ratios of the corresponding proportionality factors, i.e., we report the ratios $\kappa_d/\kappa_{S(2)}$ for $d = S(1)$, $H(c)$, LAD and LS. We consider a variety of disturbance distributions. Let $f_S(u|v, \sigma_1)$, $f_N(u|\sigma_1)$, $f_{LAP}(u|\sigma_1)$ and $f_{LOG}(u|\sigma_1)$ denote, respectively, the densities of the Student-t distribution with v degrees of freedom, the normal distribution, the Laplace distribution and the logistic distribution with zero mean and first absolute moment σ_1. We then consider in particular disturbance distributions F with densities of the form $f(u) = (1-\varepsilon) f_*(u|\sigma_1) + \varepsilon f_N(u|q\sigma_1)$ where f_* stands for any of the above densities, ε is the percentage of contamination in terms of a normal distribution, and q is the factor

[15] The asymptotic distribution of the Huber M-estimator and the LAD estimator is given in Huber (1973) and Basset and Koenker (1978), respectively, for the linear model.

Table 1

Aymptotic efficiency of respective estimators relative to the one-step M-estimator, $\hat{\beta}_{S(2)}$.

Characteristics of the contamination		Ratios of respective asymptotic variances to asymptotic variance of $\hat{\beta}_{S(2)}$					
				$\kappa_{H(c)}/\kappa_{S(2)}$			
ε	q	$\kappa_{S(1)}/\kappa_{S(2)}$	$\kappa_{LAD}/\kappa_{S(2)}$	$c = 1.0\sigma_1^F$	$c = 1.5\sigma_1^F$	$c = 2.0\sigma_1^F$	$\kappa_{LS}/\kappa_{S(2)}$
Contaminated normal distribution $f(u) = (1-\varepsilon)f_N(u\|\sigma_1) + \varepsilon f_N(u\|q\sigma_1)$							
0.0	—	1.00	1.57	1.16	1.07	1.03	1.00
0.1	2	1.00	1.44	1.07	1.01	1.00	1.08
0.1	4	1.00	1.37	1.02	1.01	1.05	1.87
0.1	10	0.91	1.32	1.01	1.I4	1.37	7.17
0.3	2	1.00	1.33	1.02	1.00	1.02	1.17
0.3	4	1.05	1.28	1.11	1.31	1.58	2.70
0.3	10	1.44	1.47	2.45	4.03	5.72	14.92
Contaminated Student-t distribution with $v = 2.5$ degrees of freedom[a] $f(u) = (1-\varepsilon)f_S(u\|v = 2.5, \sigma_1) + \varepsilon f_N(u\|q\sigma_1)$							
0.0	—		1.22	1.04	1.14	1.25	3.00
0.1	2		1.18	1.07	1.20	1.34	2.80
0.1	4		1.21	1.19	1.42	1.67	4.29
0.1	10		1.28	1.56	2.08	2.66	14.95
0.3	2		1.08	1.08	1.24	1.39	2.25
0.3	4		1.09	1.51	2.03	2.52	4.65
0.3	10		1.17	3.82	6.23	8.74	23.01
Contaminated Student-t distribution with $v = 5$ degrees of freedom $f(u) = (1-\varepsilon)f_S(u\|v = 5, \sigma_1) - \varepsilon f_N(u\|q\sigma_1)$							
0.0	—	1.00	1.30	1.01	1.01	1.09	1.25
0.1	2	1.00	1.26	1.00	1.02	1.08	1.34
0.1	4	1.00	1.26	1.05	1.13	1.26	2.33
0.1	10	0.99	1.30	1.21	1.48	1.82	9.67
0.3	2	1.01	1.18	0.99	1.05	1.13	1.35
0.3	4	1.12	1.18	1.29	1.55	1.88	3.17
0.3	10	1.61	1.32	2.91	4.71	6.63	17.51
Contaminated Laplace distribution $f(u) = (1-\varepsilon)f_{LAP}(u\|\sigma_1) + \varepsilon f_N(u\|q\sigma_1)$							
0.0	—	1.32	0.80	1.05	1.17	1.27	1.59
0.1	2	1.06	0.78	1.07	1.19	1.30	1.64
0.1	4	1.08	0.80	1.21	1.47	1.61	2.91
0.1	10	1.16	0.87	1.60	2.02	2.47	12.53
0.3	2	1.07	0.74	1.06	1.19	1.29	1.53
0.3	4	1.22	0.72	1.43	1.83	2.20	3.61
0.3	10	1.80	0.77	3.30	5.26	7.34	19.27
Contaminated logistic distribution $f(u) = (1-\varepsilon)f_{LOG}(u\|\sigma_1) + \varepsilon f_N(u\|q\sigma_1)$							
0.0	—	1.00	1.33	1.03	1.00	1.01	1.10
0.1	2	1.00	1.28	1.01	1.01	1.04	1.20
0.1	4	1.00	1.28	1.03	1.09	1.18	2.12
0.1	10	0.97	1.29	1.15	1.37	1.65	8.86
0.3	2	1.01	1.21	0.99	1.03	1.24	1.26
0.3	4	1.09	1.19	1.19	1.46	1.76	2.96
0.3	10	1.54	1.34	2.72	4.41	6.23	16.59

[a] Since $\sqrt{T}$-consistency of $\tilde{\mu}_1$ and $\tilde{h}_1$ is not implied by Theorem 3, in case of a Student-t distribution with $v = 2.5$ degrees of freedom no corresponding results for $\kappa_{S(1)}/\kappa_{S(2)}$ are reported.

by which the first absolute moment of the contaminating normal distribution exceeds that of f_*. We denote the first absolute moment of F as σ_1^F.[16]

The results reported in table 1 are quite encouraging. Apart from the Laplace distribution (where the LAD estimator is the ML estimator) and ε-contaminated versions of that distribution, we find in nearly all cases considered that the one-step M-estimator $\hat{\beta}_{S(2)}$ is asymptotically more efficient relative to all other estimators. The gain in efficiency associated with, in particular, the use of $\hat{\beta}_{S(2)}$ can be quite substantial. Suppose we consider the case of a contaminated normal distribution. Then the results of table 1 show gains in relative efficiency of up to 47 percent as compared to the LAD estimator, of up to 145, 303 and 472 percent as compared to the Huber M-estimator (depending on the choice of c) and of up to 1392 percent as compared to the LS estimator. It is interesting to note that the gain in efficiency typically increases with the degree of the contamination not only as compared to the LS estimator but also as compared to the LAD estimator and Huber's M-estimators. This suggests that the one-step M-estimators considered in this paper are not only adaptive on but also robust against deviations from the family of Student-t distributions including the normal distribution.

5. Conclusion

The present paper considers a class of one-step M-estimators for the non-linear regression model with dependent observations that are partially adaptive in that the shape of the ψ-function is influenced by the data. More specifically, let h and μ be, respectively, measures of the (inverse) scale and the tailthickness of the distribution. Then the ψ-function is indexed by estimators for h and μ, say $\tilde{h}$ and $\tilde{\mu}$. We note that only the product of the two enters the ψ-function. Given appropriate choices for $\tilde{h}$ and $\tilde{\mu}$ the corresponding one-step M-estimators are asymptotically efficient on the family of Student-t distributions and the normal distribution.

The paper proves, under alternative assumptions on the stochastic law of the data-generating process and under weak-moment requirement, strong (weak) consistency of the one-step M-estimators given $\tilde{\mu}\tilde{h}$ converges strongly (weakly) to μh, and asymptotic normality given $\tilde{\mu}\tilde{h} = \mu h + O_p(T^{-1/2})$ with $\mu h \geq 0$.

The paper considers furthermore two specific choices for $\tilde{h}$ and $\tilde{\mu}$. We prove that those estimates are strongly (weakly) consistent and converge of $O_p(T^{-1/2})$. The choices are such that the corresponding one-step M-estimators

[16] The normal, Laplace and logistic distributions with zero mean are characterized by one parameter which can be expressed in terms of σ_1. Hence, by specifying σ_1 the respective densities are completely characterized. Analogously, we can characterize the density of the Student-t distribution completely in terms of v and σ_1. For a definition of the respective densities, see, e.g., Johnson and Kotz (1970).

are asymptotically efficient on the family of Student-t distributions and the normal distribution.

Using those choices for $\tilde{h}$ and $\tilde{\mu}$ we compare the corresponding one-step M-estimators with the LAD estimator, Huber's M-estimators and the LS estimator. The comparisons are made in terms of asymptotic covariances for a variety of disturbance distributions. The results are quite encouraging in that the estimators not only perform well in comparison to the LS estimator, but also in comparison to the robust LAD estimator and Huber's M-estimators. This is especially the case for distributions with thick tails which suggests that the estimators are not only partially adaptive but also robust. This result is, of course, not surprising since for $\mu h > 0$ our ψ-function is redescending and bounded. Concerning the LS estimator the results show once again that this estimator may be very unrobust. We report cases where the asymptotic covariances of the least squares estimator are several hundred times bigger than those of the one-step M-estimators considered in this paper.

Clearly further research of, in particular, the small sample properties of the present one-step M-estimators is needed. Also, an exploration of other choices for $\tilde{h}$ and $\tilde{\mu}$, in particular those based on quantiles or on a minimization of the variance–covariance matrix, seems of interest. Furthermore, future research may consider a generalization of the present one-step M-estimators to non-linear simultaneous-equations models.

Appendix A: Proof of Theorems 1 and 2

In this appendix we first derive results analogous to Theorems 1 and 2 for a more general class of ψ-functions. Theorems 1 and 2 follow as special cases. The following lemmata are needed:[17]

Lemma A.1. Let (z_t) be a stochastic process with values in a Euclidean space Z which is stochastically stable with respect to an α-mixing base. Let C be a compact subset of a Euclidean space, let $f(z, c)$ be a continuous real function on $Z \times C$ and let H_t be the distribution function of z_t. If there exists a distribution function H such that

$$(1/T) \sum_{t=1}^{T} H_t \to H \textit{ properly}, \tag{A.1}$$

and if for some $\delta > 0$,

$$\sup_{T \geq 1} (1/T) \sum_{t=1}^{T} \mathrm{E} \sup_{c \in C} |f(z_t, c)|^{1+\delta} < \infty, \tag{A.2}$$

[17] The index set for (z_t) is either $\mathbf{N}$ or $\mathbf{Z}$.

then $\int f(z,c)\,\mathrm{d}H(z)$ *is finite and continuous on* C *and*

$$\operatorname*{plim}_{T\to\infty}\sup_{c\in C}\left|(1/T)\sum_{t=1}^{T}f(z_t,c)-\int f(z,c)\,\mathrm{d}H(z)\right|=0. \tag{A.3}$$

Proof. Bierens (1983, 1984a, b).[18]

Lemma A.2. Let (z_t) *be a stochastic process with values in a Euclidean space* Z*. Let* C *be a compact subset of a Euclidean space, let* $f(z,c)$ *be a real function on* $Z\times C$ *and let* H_t *be the distribution function of* z_t*.*

(*a*) *If* f *is continuous on* $Z\times C$, *if* $(1/T)\sum_{t=1}^{T}H_t\to H$ *properly and*

$$\sup_{t\geq 1}\mathrm{E}\sup_{c\in C}|f(z_t,c)|^{r+\delta}<\infty, \tag{A.4}$$

for some $r\geq 1$ $[r>1]$ *and some* $\delta>0$ *and if* (z_t) *is* ϕ*-mixing with* $\phi(m)=\mathrm{O}(m^{-\lambda})$ *for* $\lambda>r/(2r-1)$ [α*-mixing with* $\alpha(m)=\mathrm{O}(m^{-\lambda})$ *for* $\lambda>r/(r-1)$],

(*b*) *or if* $f(z,c)$ *is Borel measurable on* Z *for all* $c\in C$ *and continuous on* C *for all* $z\in Z$, (z_t) *is strictly stationary and ergodic and*

$$\mathrm{E}\sup_{c\in C}|f(z_t,c)|<\infty, \tag{A.5}$$

then $\int f(z,c)\,\mathrm{d}H(z)$ *is finite and continuous on* C *and*

$$\sup_{c\in C}\left|(1/T)\sum_{t=1}^{T}f(z_t,c)-\int f(z,c)\,\mathrm{d}H(z)\right|\overset{\text{a.s.}}{\to}0. \tag{A.6}$$

Proof. (a) It follows from the assumptions that for every compact set $A\subseteq C$ the processes $\sup_{c\in A}f(z_t,c)$ and $\inf_{c\in A}f(z_t,c)$ satisfy the strong law of large numbers given in McLeish (1975, Theorem 2.10). Now proceeding similar as in the proof of theorem 2.3.3 in Bierens (1981) the result follows. (b) Completely analogous, exploiting ergodicity instead of McLeish's theorem. Q.E.D.

Note that part (a) of the above lemma is closely related to theorem 2.3 in White and Domowitz (1984). A careful inspection of that theorem shows, however, that its assumptions do not generally cover the functional forms of $f(z,c)$ and stochastic processes considered in the following. In particular, the

[18] We note that the assumption of measurability in the first argument and continuity in the second argument of f is too weak for the lemma to hold. Inspection of the proof in the above reference reveals that the lemma holds if we make the above assumption of continuity of f.

assumption of continuity of $f(z_t, c)$ on C uniformly in t almost surely is not always satisfied.

In the following (except for the actual proofs of Theorems 1 and 2)$\psi_a(z)$, or equivalently $\psi(z, a)$ denotes a more general ψ-function than the one considered in the main body of the paper. The nuisance parameter, a, now varies in a set $\Theta_A \subseteq \mathbb{R}^p$. The one-step estimator $\hat{\beta}$ as well as $A(t, \theta)$, $B(t, \theta)$, $r(t, \theta)$, $A(\theta)$, $B(\theta)$ and $r(\theta)$ are defined as before with the general function ψ_a replacing the more special one. The conditions $\tilde{a} \in [0, \infty)$ and $\theta \in [0, \infty) \times \Theta_B^*$ in Definition 1 are to be replaced by $\tilde{a} \in \Theta_A$ and $\theta \in \Theta_A \times \Theta_B^*$, respectively.

The following three assumptions will replace Assumptions 4,4′,4″. A further assumption defining the feasible ψ-functions is also necessary. The following conditions are assumed to hold on some compact neighborhood $U(\theta^0) \subseteq \Theta_A \times \Theta_B^*$.

Assumption A.4. $\sup_{T \geq 1}(1/T)\Sigma_{t=1}^T \mathrm{E} \sup_{\theta \in U(\theta^0)} \|A(t, \theta)\|^{1+\delta} < \infty$ and $\sup_{T \geq 1}(1/T)\Sigma_{t=1}^T \mathrm{E} \sup_{\theta \in U(\theta^0)} \|r(t, \theta)\|^{1+\delta} < \infty$ for some $\delta > 0$.

Assumption A.4′. $\mathrm{E} \sup_{\theta \in U(\theta^0)} \|A(t, \theta)\| < \infty$ and $\mathrm{E} \sup_{\theta \in U(\theta^0)} \|r(t, \theta)\| < \infty$.

Assumption A.4″. For the same r as in Assumption 2″ and for some $\delta > 0$, we have $\sup_{t \geq 1} \mathrm{E} \sup_{\theta \in U(\theta^0)} \|A(t, \theta)\|^{r+\delta} < \infty$ and $\sup_{t \geq 1} \mathrm{E} \sup_{\theta \in U(\theta^0)} \times \|r(t, \theta)\|^{r+\delta} < \infty$.

Assumption A.7. The function $\psi_a(z) = \psi(z, a)$ is defined on $\mathbb{R} \times \Theta_A \subseteq \mathbb{R}^{p+1}$, where Θ_A is the permissible parameter space for the nuisance parameter. The functions $\psi(z, a)$ and $(\partial/\partial z)\psi(z, a)$ are continuous on $\mathbb{R} \times \Theta_A$.

Lemma A.3. Let $\tilde{\theta} = (\tilde{a}, \tilde{\beta}')'$ converge in probability [almost surely] to $\theta^0 = (a^\circ, \beta^{\circ\prime})'$, where $a^\circ \in \Theta_A$ and $\beta^\circ \in \Theta_B$, and let Assumption A.7 hold. Given Assumptions 1, 2, A.4 or 1′, 2′, A.4′ [Assumptions 1′, 2′, A.4′, or 1″, 2″, A.4″] are satisfied and $A(\theta^\circ)$ is non-singular, then $\hat{\beta} \to \beta^0 - A(\theta^0)^{-1} r(\theta^0)$ in probability [almost surely].

Proof. It is readily seen that under the above assumptions the functions $A_T(\theta)$ and $r_T(\theta)$ converge uniformly to $A(\theta)$ and $r(\theta)$ on $U(\theta^0)$ in probability [almost surely] as a consequence of Lemmata A.1 and A.2. Note that $A(\theta)$ and $r(\theta)$ are continuous on $U(\theta^0)$. Consequently the convergence of $\tilde{\theta}$ to θ^0 in probability [almost surely] implies also $A_T(\tilde{\theta}) \to A(\theta^0)$ and $r_T(\tilde{\theta}) \to r(\theta^0)$ in probability [almost surely]. [Note that $A_T(\tilde{\theta})$ and $r_T(\tilde{\theta})$ are well defined whenever $\tilde{\theta} \in \Theta_A \times \Theta_B^*$.] Q.E.D.

Lemma A.4. Given the assumptions of Lemma A.3 and Assumption 3 are satisfied. (a) If $\mathrm{E}[\psi_{a^\circ}(u_t)] = 0$ *or* $\mathrm{E}[(\partial/\partial\beta')g(x_t, \beta^\circ)] = 0$ *for all $t \geq 1$, then* *the bias* $A(\theta^\circ)^{-1}r(\theta^\circ)=0$. (*b*) *If Assumption B is satisfied, then the bias* $A(\theta^\circ)^{-1}r(\theta^\circ)=(\lambda,0,\ldots,0)'$ *with* $\lambda=[-\int\psi_{a^\circ}(y-g(x,\beta^\circ))\,\mathrm{d}\Delta]/[\int\psi'_{a^\circ}(y-g(x,\beta^\circ))\,\mathrm{d}\Delta]$.

Proof. By Assumptions 2 and 2″ we have uniform integrability and hence Assumptions A.4 and A.4″ imply that $\bar{r}_T\rightarrow r(\theta^\circ)$ and $\bar{A}_T\rightarrow A(\theta^\circ)$, where $\bar{r}_T=(1/T)\sum_{t=1}^T\mathrm{E}[r(t,\theta^\circ)]$ and $\bar{A}_T=(1/T)\sum_{t=1}^T\mathrm{E}[A(t,\theta^\circ)]$ (in the stationary case this is trivially satisfied). To prove part (a) note that $\bar{r}_T=(1/T)\sum_{t=1}^T\mathrm{E}[\psi_{a^\circ}(u_t)]\mathrm{E}[(\partial/\partial\beta')g(x_t,\beta^\circ)]=0$ by assumption, which gives the result. Part (b) can be shown as follows: Consider first the case where u_t is identically distributed. Then $\bar{r}_T=\mathrm{E}[\psi_{a^\circ}(u_t)](1/T)\sum_{t=1}^T\mathrm{E}[(\partial/\partial\beta')g(x_t,\beta^\circ)]$ where both expectations clearly exist. [The case $\psi_{a^\circ}(u_t)\equiv 0$ is trivial.] By Assumption B the first element of $(\partial/\partial\beta')g(x_t,\beta^\circ)$ is one and the first row and column of $(\partial^2/\partial\beta\partial\beta')g(x_t,\beta^\circ)$ is zero. Since $\bar{A}_T$ is non-singular for large T, we must have $\mathrm{E}[\psi'_{a^\circ}(u_t)]\neq 0$. But now $\bar{A}_T(\lambda,0,\ldots,0)'=\bar{r}_T$ for $\lambda=-\mathrm{E}[\psi_{a^\circ}(u_t)]/\mathrm{E}[\psi'_{a^\circ}(u_t)]$ which implies the result. Next consider the case where the regressors are identically distributed. Similarly as above it follows that $\bar{A}_T(\lambda_T,0,\ldots,0)'=\bar{r}_T$ where now $\lambda_T=-(1/T)\sum_{t=1}^T\mathrm{E}[\psi_{a^\circ}(u_t)]/(1/T)\sum_{t=1}^T\mathrm{E}[\psi'_{a^\circ}(u_t)]$ and the denominator is non-zero for large T. It is readily seen that λ_T converges to λ under the present assumptions. Q.E.D.

Proof of Theorem 1. We first show that Assumptions 4, 4′, 4″ imply Assumptions A.4, A.4′, A.4″. In the case $a^\circ>0$ choose $U(\theta^\circ)=[a^\circ-c,a^\circ+c]\times U(\beta^\circ)$ with c such that $a^\circ-c>0$. Note from (3.2) that $\psi_a(z)$ and $\psi'_a(z)$ are bounded for $z\in\mathbb{R}$ and $|a-a^\circ|\leq c$. Because of this it is readily seen that Assumptions 4, 4′, 4″ imply A.4, A.4′, A.4″ if $a^\circ>0$. In the case $a^\circ=0$ choose $U(\theta^\circ)=[0,c]\times U(\beta^\circ)$ for some $c>0$. Since $\psi'_a(z)$ is still bounded on $\mathbb{R}\times[0,c]$ by some constant M, since $|\psi_a(z)|\leq|z|$ and $\psi_0(z)=z$, we have

$$\mathrm{E}\sup_{\theta\in U(\theta^\circ)}\|A(t,\theta)\|^{1+\delta}$$

$$\leq 2^{\delta}\big[MG_t(1+\delta)+2^{\delta}\big\{F_t(1+\delta)+\mathrm{E}|u_t|^{1+\delta}D_t(1+\delta)\big\}\big],$$

and

$$\mathrm{E}\sup_{\theta\in U(\theta^\circ)}\|r(t,\theta)\|^{1+\delta}\leq 2^{\delta}\big[\mathrm{E}|u_t|^{1+\delta}G_t((1+\delta)/2)+G_t(1+\delta)\big].$$

In the last step we have applied the mean value theorem to $g(x_t,\beta)-g(x_t,\beta^\circ)$. Note that $U(\beta^\circ)$ can be assumed to be convex without loss of generality and that $G_t(\gamma/2)\leq[G_t(\gamma)]^{1/2}$. Hence Assumption 4 implies Assumption A.4. The remaining implications are proved analogously. Clearly Assumption A.7 is satisfied by the ψ-function (3.2). Consequently Lemma A.3

applies. Using Assumptions A and B and Assumption 3, the theorem follows now from Lemma A.4. Q.E.D.

For the asymptotic normality result the following assumptions will replace Assumptions 6, 6′, 6″. Let $p(t,\theta)=(\partial/\partial\beta')g(x_t,\beta)(\partial/\partial a)\psi(y_t-g(x_t,\beta),a)$ and $p(\theta)=\int(\partial/\partial\beta')g(x,\beta)(\partial/\partial a)\psi(y-g(x,\beta),a)\,\mathrm{d}\Delta(y,x)$.

Assumption A.6. $\sup_{T\geq 1}(1/T)\sum_{t=1}^{T}\mathrm{E}\sup_{\theta\in U(\theta^\circ)}\|p(t,\theta)\|^{1+\delta}<\infty$ and $\sup_{T\geq 1}(1/T)\sum_{t=1}^{T}\mathrm{E}\|r(t,\theta^\circ)\|^{2+\delta}<\infty$ for some $\delta>0$.

Assumption A.6′. $\mathrm{E}\sup_{\theta\in U(\theta^\circ)}\|p(t,\theta)\|<\infty$ and $\mathrm{E}\|r(t,\theta^\circ)\|^2<\infty$.

Assumption A.6″. $\sup_{t\geq 1}\mathrm{E}\sup_{\theta\in U(\theta^\circ)}\|p(t,\theta)\|^{r+\delta}<\infty$ and $\sup_{T\geq 1}(1/T)\sum_{t=1}^{T}\mathrm{E}\|r(t,\theta^\circ)\|^{2+\delta}<\infty$ for the same r as in Assumption 2″ and some $\delta>0$.

The following assumption is stronger than Assumption A.7.

Assumption A.8. The function $\psi(z,a)$ is defined on an open neighborhood of $\mathbb{R}\times\Theta_A\subseteq\mathbb{R}^{p+1}$. The set Θ_A is convex. The functions $\psi(z,a)$, $(\partial/\partial z)\psi(z,a)$ and $(\partial/\partial a)\psi(z,a)$ are continuous on $\mathbb{R}\times\Theta_A$.

Lemma A.5. Let $\tilde\theta=\theta^\circ+\mathrm{O}_p(T^{-1/2})$ where $\tilde\theta=(\tilde a,\tilde\beta')'$, $\theta^\circ=(a^\circ,\beta^{\circ\prime})'$ with $a^\circ\in\Theta_A$ and $\beta^\circ\in\Theta_B$ and let Assumptions 5 and A.8 be satisfied. If Assumptions 1, 2, A.4, A.6 or 1′, 2′, A.4′, A.6′ or 1″, 2″, A.4″, A.6″ hold, then given $\mathrm{E}[\psi(u_t,a^\circ)]=0$ *and $A(\theta^\circ)$ is non-singular,*

$$T^{1/2}(\hat\beta-\beta^\circ)=-A(\theta^\circ)^{-1}\left[T^{-1/2}\sum_{t=1}^{T}r(t,\theta^\circ)\right]$$
$$-A(\theta^\circ)^{-1}p(\theta^\circ)T^{1/2}(\tilde a-a^\circ)+\mathrm{o}_p(1),$$

where the first term is asymptotically normal, $\mathrm{N}(0,\Phi)$ *with* $\Phi=A(\theta^\circ)^{-1}B(\theta^\circ)A(\theta^\circ)^{-1}$.

Proof. From the mean-value theorem we have

$$\begin{aligned}T^{1/2}(\hat\beta-\beta^\circ)&=T^{1/2}(\tilde\beta-\beta^\circ)-T^{1/2}A_T(\tilde\theta)^{-1}r_T(\tilde\theta)\\&=\left[I-A_T(\tilde\theta)^{-1}A_T(\bar\theta)\right]T^{1/2}(\tilde\beta-\beta^\circ)\\&\quad-A_T(\tilde\theta)^{-1}\left[T^{-1/2}\sum_{t=1}^{T}r(t,\theta^\circ)\right]\\&\quad-A_T(\tilde\theta)^{-1}p_T(\bar\theta)T^{1/2}(\tilde a-a^\circ),\end{aligned}$$

with

$$p_T(\bar{\theta}) = T^{-1} \sum_{t=1}^{T} p(t, \bar{\theta}),$$

and where (in abuse of notation) $A_T(\bar{\theta})$ and $p_T(\bar{\theta})$ are the respective quantities evaluated row-wise at appropriate mean values. [Note that for large T we have $\tilde{\theta} \in U(\theta^\circ)$ on ω-sets of measure tending to one.] Given the above assumptions we have $[I - A_T(\tilde{\theta})^{-1} A_T(\bar{\theta})] \to 0$ and $A_T(\tilde{\theta})^{-1} p_T(\bar{\theta}) \to A(\theta^\circ)^{-1} p(\theta^\circ)$ in probability.[19] Because of Assumption 5 and $\mathrm{E}[\psi(u_t, a^\circ)] = 0$ it follows that $r(t, \theta^\circ)$ is a martingale difference w.r.t. the σ-algebra generated by past u_t's and current and past x_t's. In the stationary case the result follows now from the Lindeberg–Lévy CLT for martingale differences [see Billingsley (1968, theorem 23.1)], in the other cases from Brown's CLT [see Bierens (1984a, lemma 4), which holds also if in his lemma $\sigma^2 = 0$], using the Cramér–Wold device. Note that the third condition of Lemma 4 in Bierens (1984a) is satisfied because of Lemmata A.1 and A.2 and uniform integrability. Q.E.D.

The proof of the following lemma is similar to the proof of Lemma A.4.

Lemma A.6. Given the assumptions of Lemma A.5 are satisfied. (*a*) *If furthermore* $\mathrm{E}[(\partial/\partial a)\psi(u_t, a^\circ)] = 0$ *for all* $t \geq 1$, *then* $T^{1/2}(\hat{\beta} - \beta^\circ) \xrightarrow{\text{i.d.}} \mathrm{N}(0, \Phi)$. (*b*) *If Assumption B holds, then* $T^{1/2}(\hat{\beta}_* - \beta^\circ{}_*) \xrightarrow{\text{i.d.}} \mathrm{N}(0, \Phi_*)$ *with* $\Phi_* = [\partial\beta_*/\partial\beta]\Phi[\partial\beta_*/\partial\beta']$.

Note that under the assumptions of Lemma A.5 clearly $A_T(\tilde{\theta}) \to A(\theta^\circ)$. If one adds the conditions $\sup_{T \geq 1}(1/T)\sum_{t=1}^{T} \mathrm{E}\sup_{\theta \in U(\theta^\circ)} \|r(t, \theta)\|^{2+\delta} < \infty$ $[\mathrm{E}\sup_{\theta \in U(\theta^\circ)} \|r(t, \theta)\|^2 < \infty,\ \sup_{t \geq 1} \mathrm{E}\sup_{\theta \in U(\theta^\circ)} \|r(t, \theta)\|^{2r+\delta} < \infty]$ to Assumption A.6 [A.6′, A.6″], one gets also $B_T(\tilde{\theta}) \to B(\theta^\circ)$. Hence $\tilde{\Phi} = A_T(\tilde{\theta})^{-1} B_T(\tilde{\theta}) A_T(\tilde{\theta})^{-1}$ consistently estimates Φ. Note that in the context of Theorem 2 these additional conditions are automatically implied by the other conditions.

Proof of Theorem 2. Note that Assumption A.8 is satisfied. Assumptions A.6, A.6′, A.6″ are implied by Assumptions 6, 6′, 6″. This is clear for $a^\circ > 0$

[19] Precisely speaking the mean values need not be measurable, but $[I - A_T(\tilde{\theta})^{-1} A_T(\bar{\theta})] T^{1/2} \cdot (\tilde{\beta} - \beta^\circ) - A_T(\tilde{\theta})^{-1} p_T(\bar{\theta}) T^{1/2} (\tilde{a} - a^\circ)$ is measurable by construction; hence the above argument can be made precise by standard argumentation.

and follows for $a^{\circ} = 0$ from

$$\mathrm{E} \sup_{\theta \in U(\theta^{\circ})} \|p(t, \theta)\|^{1+\delta} \leq 2^{\delta}\left[\mathrm{E}|u_t|^{3+3\delta} G_t((1+\delta)/2) + G_t(2+2\delta)\right],$$

$$\mathrm{E}\|r(t, \theta^{\circ})\|^{2+\delta} \leq \mathrm{E}|u_t|^{2+\delta} G_t(1+\delta/2),$$

and the analogous relations for the other cases. The theorem then follows from Lemmata A.5 and A.6 making use of Assumptions A and B. [Note that $(\partial/\partial a)\psi(z, a)$ is an antisymmetric function of z.] Q.E.D.

Appendix B: Proof of Lemma 1 and Theorem 3

Proof of Lemma 1

Analyticity of p_1, p_2, q_1, q_2 is obvious. The first derivatives of p_1 and p_2 are respectively $p_1'(v) = p_1(v)[-(v-2)^{-1} + \phi(v/2) - \phi(v/2 - \frac{1}{2})]$ and $p_2'(v) = p_2(v)[\frac{1}{2}\phi(v/2) + \frac{1}{2}\phi(v/2 - \frac{1}{2}) - \phi(v/2 - \frac{1}{4})]$, where $\phi(x) = \Gamma'(x)/\Gamma(x)$. Now $p_1'(v) = p_1(v)[-(v-2)^{-1} + 2\beta(v-1)] = p_1(v)[-2\int_0^{\infty} x((v-2)^2 + x^2)^{-1} (sh\pi x)^{-1}\mathrm{d}x] < 0$, using formulas 8.370 and 3.522.2 in Gradshteyn and Ryzhik (1965) [where also $\beta(x)$ is defined] and observing that the integrand is positive. From their formula 8.363.8 we see that $\phi''(x) < 0$; hence ϕ is concave which implies $p_2'(v) < 0$. This shows the monotonicity of p_1 and p_2 on their respective intervals of definition. $p_1(2+) = \infty$ is obvious, $p_2(1+) = \infty$ since $\Gamma(0) = \infty$. $p_1(\infty)$, $p_2(\infty)$, q_1 (∞, σ_1) and $q_2(\infty, \sigma_{1/2})$ are easily calculated using formula 8.328.2 in Gradshteyn and Ryzhik (1965). Q.E.D.

Recall that we set $\sigma_{\alpha}/\sigma_{\alpha/2}^2 = \infty$ whenever $\sigma_{\alpha} = \infty$ and that degenerate distributions are excluded from the discussion. Define $s_{\alpha} = T^{-1}\sum|u_t|^{\alpha}$. For the proof of Theorem 3 we need the following two lemmata.

Lemma B.1. Assume $\tilde{\beta} \to \beta^{\circ}$ in probability [almost surely] and Assumptions 1, 2, 4 or 1′, 2′, 4′ [1′, 2′, 4′ or 1″, 2″, 4″] hold. Let (u_t) be strictly stationary and ergodic. (a) Then $\tilde{s}_{1/2} \to \sigma_{1/2}$, $\tilde{s}_1 \to \sigma_1$ in probability [almost surely]. If additionally $\sigma_{\gamma} < \infty$ for some $\gamma > 0$, then $\tilde{s}_1/\tilde{s}_{1/2}^2 \to \sigma_1/\sigma_{1/2}^2$ and $\tilde{s}_2/\tilde{s}_1^2 \to \sigma_2/\sigma_1^2$ in probability [almost surely]. (b) Assume $\tilde{\beta} = \beta^{\circ} + \mathrm{O}_p(T^{-1/2})$ and (u_t) to be i.i.d. If $\sigma_4 < \infty$, we have $\tilde{s}_2/\tilde{s}_1^2 = \sigma_2/\sigma_1^2 + \mathrm{O}_p(T^{-1/2})$; if $\sigma_2 < \infty$ we have $\tilde{s}_1 = \sigma_1 + \mathrm{O}_p(T^{-1/2})$. If $\sigma_2 < \infty$, F has a density that is (essentially) bounded in a neighborhood of zero and for some constant M we have $\sup\{\|(\partial/\partial\beta)g(x_t, \beta)\|: t \in \mathbb{N}, \beta \in U(\beta^{\circ})\} \leq M$ *a.s., then $\tilde{s}_1/\tilde{s}_{1/2}^2 = \sigma_1/\sigma_{1/2}^2 + \mathrm{O}_p(T^{-1/2})$ and $\tilde{s}_{1/2} = \sigma_{1/2} + \mathrm{O}_p(T^{-1/2})$.*

Proof. (a) We prove the strong convergence part only. The convergence in probability part follows from a standard subsequence argument. For large T and $0 < \alpha \leq 1$, we have from the mean-value theorem $|\tilde{s}_\alpha - s_\alpha| \leq T^{-1} \times \sum |\tilde{u}_t - u_t|^\alpha = T^{-1} \sum |g(x_t, \beta^\circ) - g(x_t, \tilde{\beta})|^\alpha \leq T^{-1} \sum \|(\partial/\partial\beta) g(x_t, \bar{\beta}_t)\|^\alpha \times \|\tilde{\beta} - \beta^\circ\|^\alpha \leq T^{-1} \sum \sup_{\beta \in U(\beta^\circ)} \|(\partial/\partial\beta) g(x_t, \beta)\|^\alpha \|\tilde{\beta} - \beta^\circ\|^\alpha \to 0$ since $\beta \to \beta^\circ$, and the sum converges to some finite value by Lemma A.2 and Assumptions 1′, 2′, 4′ or 1″, 2″, 4″ since $\alpha(r + \delta)/2 \leq (r + \delta)$. Since the u_t's are ergodic it follows that $s_\alpha \to \sigma_\alpha$ regardless if σ_α is finite or not, as is easily seen by a truncation argument. Hence $\tilde{s}_\alpha \to \sigma_\alpha$ almost surely. As long as $\sigma_{1/2} < \infty$ this implies also that $\tilde{s}_1/\tilde{s}_{1/2}^2$ converges to $\sigma_1/\sigma_{1/2}^2$. Now if $\sigma_{1/2} = \infty$, we obtain from Ljapunov's inequality [see, e.g., Loève (1963, p. 172); note that $\tilde{s}_\alpha$ is the αth moment of a discrete distribution] $\tilde{s}_{1/2}^{1-\gamma} \leq \tilde{s}_\gamma^{1/2} \tilde{s}_1^{1/2-\gamma}$ (without loss of generality $\gamma < \frac{1}{2}$) which implies $\tilde{s}_{1/2}^\gamma \leq \tilde{s}_\gamma^{1/2} (\tilde{s}_1/\tilde{s}_{1/2}^2)^{1/2-\gamma}$. Now $\tilde{s}_{1/2}$ tends to infinity and $\tilde{s}_\gamma$ to $\sigma_\gamma < \infty$ as just shown, hence $\tilde{s}_1/\tilde{s}_{1/2}^2$ tends to infinity.

Next consider $\tilde{s}_2/\tilde{s}_1^2$. If $\sigma_1 = \infty$, we get using Ljapunov's inequality $\tilde{s}_1^\gamma \leq \tilde{s}_\gamma (\tilde{s}_2/\tilde{s}_1^2)^{1-\gamma}$ which shows that $\tilde{s}_2/\tilde{s}_1^2$ goes to infinity since $\tilde{s}_1$ does and $\tilde{s}_\gamma \to \sigma_\gamma$ as shown above. For $\sigma_1 < \infty$, we have to show that $\tilde{s}_2 \to \sigma_2$. If also $\sigma_2 < \infty$, we have $|\tilde{s}_2 - s_2| \leq (2/T) |\sum u_t (g(x_t, \tilde{\beta}) - g(x_t, \beta^\circ))| + (1/T) \sum \|(\partial/\partial\beta) g(x_t, \bar{\beta})\|^2 \|\tilde{\beta} - \beta^\circ\|^2 \leq 2 \|T^{-1} \sum u_t (\partial/\partial\beta) g(x_t, \bar{\beta})\| \|\tilde{\beta} - \beta^\circ\| + \mathrm{o}(1) \leq (T^{-1} \sum u_t^2)^{1/2} (T^{-1} \sum \sup_{\beta \in U(\beta^\circ)} \|(\partial/\partial\beta) g(x_t, \beta)\|^2)^{1/2} \|\tilde{\beta} - \beta^\circ\| + \mathrm{o}(1)$ by similar arguments as before. Now $T^{-1} \sum u_t^2 \to \sigma_2 < \infty$; hence the r.h.s. is $\mathrm{o}(1)$ almost surely. Next if $\sigma_2 = \infty$, we have $\tilde{s}_2^2 = T^{-1} \sum u_t^2 - (2/T) \times \sum u_t (g(x_t, \tilde{\beta}) - g(x_t, \beta^\circ)) + T^{-1} \sum (g(x_t, \tilde{\beta}) - g(x_t, \beta^\circ))^2$. From the mean-value theorem and the assumptions it follows that the last sum is bounded (actually it goes to zero). The absolute value of the second sum on the r.h.s. is bounded by $(T^{-1} \sum u_t^2)^{1/2} (T^{-1} \sum (g(x_t, \tilde{\beta}) - g(x_t, \beta^\circ))^2)^{1/2}$ and $T^{-1} \sum u_t^2 \to \sigma_2 = \infty$, which again follows from ergodicity (after truncation). Hence $\tilde{s}_2^2 \to \sigma_2 = \infty$.

(b) $T^{1/2} |\tilde{s}_1 - \sigma_1| \leq T^{1/2} |\tilde{s}_1 - s_1| + T^{1/2} |s_1 - \sigma_1| \leq T^{1/2} \|\tilde{\beta} - \beta^\circ\| \times T^{-1} \sum \sup_{\beta \in U(\beta^\circ)} \|(\partial/\partial\beta) g(x_t, \beta)\| + \mathrm{O}_p(1) = \mathrm{O}_p(1)$ since $s_1 - \sigma_1$ satisfies a CLT because $\sigma_2 < \infty$. These inequalities hold at least on ω-sets where $\tilde{\beta} \in U(\beta^\circ)$ and whose probabilities tend to one. Similarly under $\sigma_4 < \infty$ we have $T^{1/2} |\tilde{s}_2 - \sigma_2| \leq T^{1/2} |\tilde{s}_2 - s_2| + T^{1/2} |s_2 - \sigma_2| = \mathrm{O}_p(1)$. This proves $\tilde{s}_2/\tilde{s}_1^2 = \sigma_2/\sigma_1^2 + \mathrm{O}_p(T^{-1/2})$. Finally consider $T^{1/2} |\tilde{s}_{1/2} - \sigma_{1/2}|$. By assumption $T^{1/2} |s_{1/2} - \sigma_{1/2}| = \mathrm{O}_p(1)$ since it satisfies a CLT; hence we concentrate on $T^{1/2} |\tilde{s}_{1/2} - s_{1/2}|$. This can be decomposed into $T^{-1/2} \sum \| \tilde{u}_t|^{1/2} - |u_t|^{1/2}| 1_V(|u_t|) + T^{-1/2} \sum \| \tilde{u}_t|^{1/2} - |u_t|^{1/2}| 1_W(|u_t|) = A_T + B_T$, where $V = [0, f(T)^{-1}]$, $W = (f(T)^{-1}, \infty)$ and $f(T)$ is to be defined below. Now for every $\delta > 0$ there exists a $K(\delta) > 0$ such that the event $E = \{ M T^{1/2} \|\tilde{\beta} - \beta^\circ\| \geq K(\delta)$ or $\tilde{\beta} \notin U(\beta^\circ)\}$ has $P(E) < \delta$ for all large T; the constant M is defined above. Choosing $f(T) = T^{1/2}/2K(\delta)$ we obtain the relations: $P(B_T > N) \leq P(E) + P(E^c$ and $B_T > N)$. Now on the event E^c we have that $|u_t| > f(T)^{-1}$ implies $|\tilde{u}_t| > |u_t| - f(T)^{-1} > 0$. By expanding $|\tilde{u}_t|^{1/2}$ we obtain $|\tilde{u}_t|^{1/2} = |u_t|^{1/2} + \frac{1}{2} \bar{u}_t^{-1/2} (|\tilde{u}_t|$

$-|u_t|)$ where $\bar{u}_t$ lies between $|\tilde{u}_t|$ and $|u_t|$; this implies $||\tilde{u}_t|^{1/2}-|u_t|^{1/2}| \leq \frac{1}{2}\bar{u}_t^{-1/2}M\|\tilde{\beta}-\beta^{\circ}\|$ using the mean-value theorem. Hence on the above event B_T is dominated by $(M/2)T^{-1/2}\|\tilde{\beta}-\beta^{\circ}\|\sum\bar{u}_t^{-1/2}1_W(|u_t|)$ and clearly $\bar{u}_t > |u_t| - f(T)^{-1}$, and this last expression can again be bounded by $\frac{1}{2}K(\delta)T^{-1}\sum(|u_t| - f(T)^{-1})^{-1/2}1_W(|u_t|)$. Putting this together we obtain $P(B_T > N) \leq \delta + P(\frac{1}{2}K(\delta)T^{-1}\sum(|u_t|-f(T)^{-1})^{-1/2}1_W(|u_t|) > N) \leq \delta + \frac{1}{2}K(\delta)\mathrm{E}[(|u_t| - f(T)^{-1})^{-1/2}1_W(|u_t|)]/N$, which can be made arbitrarily small by choosing N large, since the expectation remains bounded when F has a density that is essentially bounded in a neighborhood of zero [note that $f(T)^{-1} \to 0$].

Secondly, $A_T \leq (MT^{1/2}\|\tilde{\beta}-\beta^{\circ}\|)^{1/2}T^{-3/4}\sum 1_V(|u_t|)$ on $\{\tilde{\beta} \in U(\beta^{\circ})\}$ by arguments already used at the beginning of this proof. Since $\tilde{\beta} = \beta^{\circ} + \mathrm{O}_p(T^{-1/2})$ it suffices to show that $P(T^{-3/4}\sum 1_V(|u_t|) > N)$ can be made arbitrarily small. This probability is bounded by $N^{-1}T^{1/4}P(|u_t| \leq f(T)^{-1}) \leq N^{-1}T^{1/4}2Cf(T)^{-1}$, where C is a bound for the density of F in a neighborhood of zero, hence by using the definition of $f(T)$ we obtain as a final bound $4CK(\delta)N^{-1}T^{-1/4}$. This proves that $\tilde{s}_{1/2} = \sigma_{1/2} + \mathrm{O}_p(T^{-1/2})$. The result for $\tilde{s}_1/\tilde{s}_{1/2}^2$ follows immediately. Q.E.D.

Lemma B.2. The functions defined by (3.15) and (3.16) are differentiable everywhere except in one point and have left- and right-hand derivatives which are bounded in every compact interval. The functions $q_1(1/x, y)$ and $q_2(1/x, y)$ have partial derivatives with respect to x and y everywhere on $(0, \frac{1}{2}) \times (0, \infty)$ and $(0,1) \times (0, \infty)$. Furthermore $\lim_{x \to 0}(\partial q_i/\partial x)(1/x, y)$ is bounded uniformly whenever y varies in a compact interval, $i = 1, 2$.

Proof. Denote the function given by (3.15) by $\mu_1 = g(b_1)$ where $b_1 = \sigma_2/\sigma_1^2$ for ease of notation. Now if $b_1 < \pi/2$ we have $g'(b_1) = 0$ and for $b_1 > \pi/2$ clearly $g'(b_1)$ exists, is continuous and is given by $g'(b_1) = -\mu_1^2[p_1'(p_1^{-1}(b_1))]^{-1} = -\mu_1^2(p_1'(\mu_1^{-1}))^{-1}$. Obviously $\lim_{b_1 \uparrow \pi/2} g'(b_1) = 0$ and

$$\lim_{b_1 \downarrow \pi/2} g'(b_1) = \lim_{\mu_1 \downarrow 0}\left[-\mu_1^2\left(p_1'\left(\mu_1^{-1}\right)\right)^{-1}\right] = \lim_{v \to \infty}\left[-\left(v^2 p_1'(v)\right)^{-1}\right]$$

$$= -\lim_{v \to \infty}\left(p_1(v)\right)^{-1} \lim_{v \to \infty} v^{-2}\left[-(v-2)^{-1} + 2\beta(v-1)\right]^{-1}$$

$$= \frac{2}{\pi}\lim_{v \to \infty}\left[2v^2\int_0^{\infty} x\left((v-2)^2 + x^2\right)^{-1}(sh\pi x)^{-1}\,\mathrm{d}x\right]^{-1},$$

using results already established and Lemma 2. The last expression now equals

$(1/\pi)[\int_0^\infty x(sh\pi x)^{-1}\,\mathrm{d}x]^{-1}$ by dominated convergence [e.g., $4x(sh\pi x)^{-1}$ is a dominating function]. Using formula 3.521.1 of Gradshteyn and Ryzhik (1965) we obtain $\lim_{b_1 \downarrow \pi/2} g'(b_1) = 2/\pi$. Using the mean-value theorem we see that the right-hand derivative at $b_1 = \pi/2$ exists and equals $2/\pi$. The left-hand derivative is obviously zero. Summarizing we see that on every compact interval the left- and right-hand derivatives of g are bounded. Defining similarly $\mu_2 = f(b_2)$ as the function given by (3.16), where now $b_2 = \sigma_1/\sigma_{1/2}^2$ we proceed completely analogous [for short we set $c = \sqrt{\pi}/\Gamma(\frac{3}{4})^2$]:

$$\lim_{b_2 \downarrow c} f'(b_2)$$

$$= \lim_{v \to \infty} \left[-\left(v^2 p_2'(v)\right)^{-1} \right]$$

$$= -\lim_{v \to \infty} \left(p_2(v)\right)^{-1} \lim_{v \to \infty} v^{-2}\left[\tfrac{1}{2}\phi(v/2) + \tfrac{1}{2}\phi\left(v/2 - \tfrac{1}{2}\right) - \phi\left(v/2 - \tfrac{1}{4}\right)\right]^{-1}$$

$$= -c^{-1} \lim_{v \to \infty} \left\{ (v^2/2) \left[\sum_{k=0}^{\infty} \left[\left(v/2 - \tfrac{1}{4} + k\right)^{-1} - \left(v/2 + k\right)^{-1} \right] \right.\right.$$

$$\left.\left. - \sum_{k=0}^{\infty} \left[\left(v/2 - \tfrac{1}{2} + k\right)^{-1} - \left(v/2 - \tfrac{1}{4} + k\right)^{-1} \right] \right] \right\}^{-1}$$

$$= -c^{-1} \lim_{v \to \infty} \left[(-v^2/16) \sum_{k=0}^{\infty} \left(v/2 - \tfrac{1}{4} + k\right)^{-1} \right.$$

$$\left. \times \left(v/2 + k\right)^{-1} \left(v/2 - \tfrac{1}{2} + k\right)^{-1} \right]^{-1},$$

using formula 8.363.3. Calling the expression in brackets $A(v)$ we obtain

$$(-v^2/16)\left[\left(v/2 - \tfrac{1}{2}\right)^{-3} + \int_0^\infty \left(v/2 - \tfrac{1}{2} + x\right)^{-3} \mathrm{d}x \right]$$

$$\leq (-v^2/16) \sum_{k=0}^{\infty} \left(v/2 - \tfrac{1}{2} + k\right)^{-3} \leq A(v)$$

$$\leq (-v^2/16) \sum_{k=0}^{\infty} \left(v/2 + k\right)^{-3}$$

$$\leq (-v^2/16) \int_0^\infty \left(v/2 + x\right)^{-3} \mathrm{d}x,$$

which shows after an easy calculation that $\lim_{v \to \infty} A(v) = -\frac{1}{8}$. Putting this together, we obtain $\lim_{b_2 \downarrow c} f'(b_2) = 8c^{-1}$. Calculating $(\partial q_1/\partial x)(1/x, y)$, we obtain, setting $z = x^{-1}$, the expression

$$\frac{-1}{\pi y^2} z \frac{\Gamma\left(z/2 - \frac{1}{2}\right)^2}{\Gamma(z/2)^2} \left\{ z + z^2 \left[\phi\left(z/2 - \frac{1}{2}\right) - \phi(z/2) \right] \right\}.$$

Now for $z \to \infty$ the first factor of the product remains bounded by virtue of formula 8.328.2. Now using formula 8.363.3, we obtain

$$\phi\left(\frac{z}{2} - \frac{1}{2}\right) - \phi\left(\frac{z}{2}\right) = -\frac{1}{2} \sum_{k=0}^{\infty} \left(\frac{z}{2} + k\right)^{-1} \left(\frac{z-1}{2} + k\right)^{-1}.$$

Therefore we obtain

$$\frac{-2}{(z-1)^2} - \frac{1}{z-1} \leq \phi\left(\frac{z}{2} - \frac{1}{2}\right) - \phi\left(\frac{z}{2}\right) \leq -\frac{1}{z},$$

by estimating the sum by integrals in an obvious way. But this shows that second factor in the above product is also bounded. The result for q_2 is obtained in a completely analogous way. Q.E.D.

Proof of Theorem 3

(a) Is an immediate consequence of Lemma B.1 and the continuity of the mappings defining $\mu_i(F)$ and $h_i(F)$ which follows easily from Lemma 1.

(b) If $\sigma_4 < \infty$, then clearly $\sigma_2/\sigma_1^2 < \infty$. The mapping (3.15) attaching $\mu_1(F)$ to σ_2/σ_1^2 is differentiable everywhere except at $\sigma_2/\sigma_1^2 = \pi/2$ and has bounded left- and right-hand derivatives in every compact interval as follows from Lemma B.2. Then by a simple argument using the mean value theorem and Lemma B.1, we obtain $\tilde{\mu}_1 = \mu_1(F) + \mathrm{O}_p(T^{-1/2})$. Applying the mean-value theorem to h_1 and using Lemmata B.1 and B.2 we see that $\tilde{h}_1 = h_1(F) + \mathrm{O}_p(T^{-1/2})$.

Now if $\sigma_2 < \infty$ clearly $\sigma_1/\sigma_{1/2}^2 < \infty$ and by the same arguments as above we obtain the desired results for $\tilde{\mu}_2$ and $\tilde{h}_2$. Q.E.D.

References

Amemiya, T., 1982, Two stage least absolute deviations estimators, Econometrica 50, 689–711.

Andrews, D.F. et al., 1972, Robust estimates of location: Survey and advances (Princeton University Press, Princeton, NJ).

Bassett, G., Jr. and R. Koenker, 1978, Asymptotic theory of least absolute error regression, Journal of the American Statistical Association 73, 618–622.

Bickel, P.J., 1975, One-step Huber estimates in the linear model, Journal of the American Statistical Association 70, 428–434.

Bickel, P.J., 1982, On adaptive estimation, Annals of Statistics 10, 647–671.

Bierens, H.J., 1981, Robust methods and asymptotic theory in nonlinear econometrics, Lecture notes in economics and mathematical systems no. 192 (Springer-Verlag, Berlin).

Bierens, H.J., 1982, Consistent model specification tests, Journal of Econometrics 20, 105–134.

Bierens, H. J., 1983, Model specification testing of time series regression, Research memorandum no. 8321 (Department of Economics, University of Amsterdam, Amsterdam).

Bierens, H.J., 1984a, Model specification testing of time series regression, Journal of Econometrics 26, 323–353.

Bierens, H.J., 1984b, Moving average index modeling of economic time series and the Granger-causal structure of unemployment in the Netherlands, Research memorandum no. 8423 (Department of Economics, University of Amsterdam, Amsterdam).

Billingsley, P., 1968, Convergence of probability measures (Wiley, New York).

Burguete, J., A.R. Gallant and G. Souza, 1982, On unification of the asymptotic theory of nonlinear econometric models, Econometric Reviews 1, 151–190.

Domowitz, I. and H. White, 1982, Misspecified models with dependent observations, Journal of Econometrics 20, 35–58.

Feller, W., 1971, An introduction to probability theory and its applications, Vol. II (Wiley, New York).

Gallant, A.R. and A. Holly, 1980, Statistical inference in an implicit, nonlinear, simultaneous equation model in the context of maximum likelihood estimation, Econometrica 48, 697–720.

Gilstein, C.Z. and E.E. Leamer, 1983, Robust sets of regression estimates, Econometrica 51, 321–334.

Goldfeld, S.M. and R.E. Quandt, 1981, Econometric modelling with non-normal disturbances, Journal of Econometrics 17, 141–155.

Gourieroux, C., A. Monfort and A. Trognon, 1984, Pseudo maximum likelihood methods: Theory, Econometrica 52, 681–700.

Gradshteyn, I.S. and I.M. Ryzhik, 1965, Tables of integrals, series and products (Academic Press, New York).

Grossmann, W., 1976, Robust nonlinear regression, in: J. Gordesch and P. Naeve, eds., Proceedings of the 2nd COMPSTAT symposium (Physica Verlag, Vienna).

Grossmann, W., 1982, Statistical estimation of nonlinear regression functions, Math. Operationsforsch. Statist., Ser. Statistics 13, 455–471.

Hall, D.L. and B.L. Joiner, 1982, Representations of the space of distributions useful in robust estimation of location, Biometrika 69, 55–59.

Hogg, R.V. and R.V. Lenth, 1984, A review of some adaptive statistical techniques, Mimeo. (Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA).

Huber, P.J., 1964, Robust estimation of a location parameter, Annals of Mathematical Statistics 35, 73–101.

Huber, P.J., 1967, The behaviour of maximum likelihood estimates under nonstandard conditions, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. I (University of California Press, Berkeley, CA).

Huber, P.J., 1973, Robust regression: Asymptotics, conjectures and Monte Carlo, Annals of Statistics 1, 799–821.

Huber, P.J., 1981, Robust statistics (Wiley, New York).

Johnson, N.L. and S. Kotz, 1970, Distributions in statistics: Continuous univariate distributions (Wiley, New York).

Koenker, R., 1982, Robust methods in econometrics, Econometric Reviews 1, 213–255.

Koenker, R. and G. Bassett, Jr., 1978, Regression quantiles, Econometrica 46, 33–50.

Loève, M., 1963, Probability theory (Van Nostrand, New York).

Manski, C.F., 1984, Adaptive estimation in nonlinear regression models, Econometric Reviews 3, 145–194.

Maronna, R.A. and V.J. Yohai, 1981, Asymptotic behaviour of general *M*-estimates for regression and scale with random carriers, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 58, 7–20.

McLeish, D.L., 1975, A maximal inequality and dependent strong laws, Annals of Probability 3, 826–836.

Powell, J.L., 1983, The asymptotic normality of two-stage least absolute deviations estimators, Econometrica 51, 1569–1576.

Prucha, I.R. and H.H. Kelejian, 1984, The structure of simultaneous equation estimators: A generalization towards nonnormal disturbances, Econometrica 52, 721–736.

Relles, D., 1968, Robust regression by modified least squares, Thesis (Yale University, New Haven, CT).

Stein, C., 1956, Efficient nonparametric testing and estimation, in: Proceedings of the third Berkeley symposium on mathematical statistics and probability, Vol. I (University of California Press, Berkeley, CA).

White, H., 1982, Maximum likelihood estimation of misspecified models, Econometrica 50, 1–25.

White, H. and I. Domowitz, 1984, Nonlinear regression with dependent observations, Econometrica 52, 143–161.

Yohai, V.J., 1974, Robust estimation in the linear model, Annals of Statistics 2, 562–567.