

ESTIMATION WHEN A PARAMETER IS ON A BOUNDARY

BY

DONALD W. K. ANDREWS

COWLES FOUNDATION PAPER NO. 988



**COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY**

**Box 208281
New Haven, Connecticut 06520-8281**

2000

<http://cowles.econ.yale.edu/>

ESTIMATION WHEN A PARAMETER IS ON A BOUNDARY

BY DONALD W. K. ANDREWS¹

This paper establishes the asymptotic distribution of an extremum estimator when the true parameter lies on the boundary of the parameter space. The boundary may be linear, curved, and/or kinked. Typically the asymptotic distribution is a function of a multivariate normal distribution in models without stochastic trends and a function of a multivariate Brownian motion in models with stochastic trends. The results apply to a wide variety of estimators and models.

Examples treated in the paper are: (i) quasi-ML estimation of a random coefficients regression model with some coefficient variances equal to zero and (ii) LS estimation of an augmented Dickey-Fuller regression with unit root and time trend parameters on the boundary of the parameter space.

KEYWORDS: Asymptotic distribution, inequality restrictions, random coefficients regression, stochastic trends, unit root model.

1. INTRODUCTION

TO OBTAIN THE ASYMPTOTIC DISTRIBUTION of an estimator, a standard assumption in the literature is that the true parameter is in the interior of the parameter space. This assumption is convenient because it allows one to make use of the fact that first order conditions hold, at least asymptotically. There are numerous cases of interest, however, in which the true parameter is on the boundary of the parameter space. Examples are given below.

In this paper, we provide results that establish the asymptotic distribution of an extremum estimator when the true parameter may be on the boundary of the parameter space. In such cases, the first order conditions do not hold with positive probability for all sample sizes. We provide general high level assumptions under which the asymptotic results hold, we provide sufficient conditions for the high level assumptions, and we verify these conditions in two examples. A sequel to this paper, Andrews (1997a), provides additional sufficient conditions for the high level assumptions and considers six additional examples.

We start by outlining the steps used to obtain the asymptotic distribution. In the process, we describe the asymptotic distribution itself. Let $\hat{\theta}$ be an estimator

¹The author thanks Moshe Buchinsky, Arthur Lewbel, Whitney Newey, David Pollard, Chris Sims, the co-editor, and three referees for helpful comments, Glenna Ames for typing the manuscript, and Rosemarie Lewis and Carol Copeland for proofreading the manuscript. The author gratefully acknowledges the research support of the National Science Foundation via Grant Numbers SBR-9410675 and SBR-9730277.

that maximizes a function $\ell_T(\theta)$ over a parameter space $\Theta \subset R^s$, where T is the sample size. First, one establishes that $\hat{\theta}$ converges in probability to some value θ_0 as $T \rightarrow \infty$. The case of interest is when θ_0 is on the boundary of Θ . Second, for $\theta \in \Theta$ close to θ_0 , one approximates the estimator objective function $\ell_T(\theta)$ via $-1/2$ times a quadratic function of θ , denoted $q_T(B_T(\theta - \theta_0))$, whose coefficients are random and are normalized such that they converge in distribution to some limit. Here, B_T is a deterministic normalization matrix. For nontrending data, often $B_T = T^{1/2}I_s$. Third, one shows that the quadratic approximation implies that $B_T(\hat{\theta} - \theta_0) = O_p(1)$.

Fourth, one considers the shape of the parameter space for θ near θ_0 , because only such values are relevant asymptotically. We treat the case where the shifted and rescaled parameter space $B_T(\Theta - \theta_0)/b_T$ is a convex cone Λ centered at θ_0 (at least locally to θ_0), or can be approximated by Λ , where b_T is a sequence of scalar constants such that $b_T \rightarrow \infty$ as $T \rightarrow \infty$. For example, this includes cases in which θ_0 is on a boundary defined by linear and/or nonlinear equality and/or inequality constraints.

Fifth, one shows that maximizing $\ell_T(\theta)$ over $\theta \in \Theta$ is asymptotically equivalent to minimizing $q_T(\lambda)$ over $\lambda \in \Lambda$ in the sense that $B_T(\hat{\theta} - \theta_0) = \hat{\lambda}_T + o_p(1)$, where $\hat{\lambda}_T$ is defined to minimize $q_T(\lambda)$ over $\lambda \in \Lambda$.

Sixth, one obtains the asymptotic distribution of $\hat{\lambda}_T$ using the convergence in distribution of the coefficients of $q_T(\lambda)$ and the continuous mapping theorem. Let $q(\lambda) = (\lambda - Z)\mathcal{F}(\lambda - Z)$ denote the limit of $q_T(\lambda)$, where Z is a random s -vector, and \mathcal{F} is a (possibly random) $s \times s$ matrix. Let λ minimize $q(\lambda)$ over $\lambda \in \Lambda$. The random vector $\hat{\lambda}_T$ is a continuous function of the coefficients of $q_T(\cdot)$ and the latter converge in distribution. In consequence, $\hat{\lambda}_T$ converges in distribution to $\hat{\lambda}$. For example, in the case of nontrending data, Z typically has a normal distribution and \mathcal{F} is a positive definite nonrandom matrix.

Seventh, the results of the fifth and sixth steps combine to show that $B_T(\hat{\theta} - \theta_0)$ converges in distribution to $\hat{\lambda}$. We provide conditions under which certain subvectors of $\hat{\lambda}$ have simple expressions and we obtain closed form solutions for subvectors of $\hat{\lambda}$.

In sum, we find that the asymptotic distribution of $\hat{\theta}$ is given by that of a random vector that minimizes a stochastic quadratic function over a convex cone Λ that approximates the shifted and rescaled parameter space. The asymptotic distribution often depends on (estimable) nuisance parameters. It is easy to simulate.

As discussed below, there are numerous antecedents in the literature to the approach outlined above. For example, the use of a quadratic approximation to the estimator objective function, rather than the reliance on first order conditions, has been made by Chernoff (1954), LeCam (1960), Jeganathan (1982), Pollard (1985), Pakes and Pollard (1989), and van der Vaart and Wellner (1996), among others.

Our results are designed to cover a wide variety of estimators and models. The estimators covered by the results include least squares (LS), quasi-maxi-

imum likelihood (QML), generalized method of moments (GMM), minimum distance, two-step, and semiparametric estimators among others. The estimator objective function can be smooth or nonsmooth, so that simulated method of moment (MSM) and least absolute deviation estimators are covered. This feature is obtained by using stochastic equicontinuity or stochastic differentiability conditions, as in Pollard (1985), Pakes and Pollard (1989), Andrews (1994a, b), and van der Vaart and Wellner (1996).

The results apply when the estimator function is not necessarily defined in a neighborhood of the true parameter. In consequence, the results cover random coefficient models in which some coefficient variances are zero. This contrasts with many testing papers that consider tests when the true parameter is on the boundary of the maintained hypothesis, but the estimator objective function is assumed to be well-defined in a neighborhood of the true boundary point, such as Chernoff (1954), Gourieroux and Monfort (1989), and Andrews (1996, 1998a), among others. To obtain these results we use a generalization of Taylor's Theorem that does not require the function to be defined in a neighborhood of the point of expansion.

The models covered by the results include cross-sectional, panel, and time series models. The results allow for deterministic and stochastic trends in linear time series models. In consequence, the results can be applied to obtain the asymptotic distributions of estimators of unit root and cointegration models when there are binding equality and/or inequality restrictions on the parameters. The results also can be applied to least squares and other estimators in models with heavy tails when there are binding constraints on the parameters.

We note that the assumptions employed here are such that one often can use existing results in the literature (that are designed for the case where the true parameter is an interior point) to help verify the assumptions. This is particularly useful for semiparametric estimators. One does not need to re-prove results regarding the effect of preliminary nonparametric estimators on the properties of the estimator objective function.

By approximating the parameter space by a cone, we allow the boundary to be linear, curved, and/or kinked. The parameter space may have empty interior, as occurs when there are equality restrictions. This approach was used by Chernoff (1954) in the context of likelihood ratio tests. Our approximation condition extends that of Chernoff (1954) to allow for models with trends. In addition, we provide primitive sufficient conditions for the approximation to hold when the boundary is determined by nonlinear equality and/or inequality constraints.

Several papers in the literature consider the asymptotic properties of estimators when the true parameter lies on the boundary of the parameter space. Aitchison and Silvey (1958) consider ML estimators for iid models with smooth likelihoods when the parameter is subject to smooth equality constraints. Moran (1971) considers ML estimators for iid models with smooth likelihoods with one or two parameters restricted to be nonnegative when the true values of these

parameter(s) are zero. Chant (1974) generalizes Moran's results for the same model to cover more than two nonnegativity restrictions. Self and Liang (1987) generalize Chant's results for the same model, but there are problems with their results. Gourieroux and Monfort (1989, Ch. 21) consider an extremum estimator based on a smooth objective function when the true parameter is on a boundary defined by smooth inequality constraints. They provide the asymptotic distribution of some functions of this estimator, but not the asymptotic distribution of the entire estimator. Added in press: two additional references are Geyer (1994) and Wang (1996). Judge and Takayama (1966), Lovell and Prescott (1970), Rothenberg (1973, Ch. 3), and Liew (1976) consider the finite sample behavior of the LS estimator of the linear regression model when it is subject to linear equality and inequality restrictions. Rothenberg (1973, Ch. 3) also provides some finite sample efficiency results that apply to a general class of inequality restricted estimators.

The asymptotic results derived here are useful for a number of purposes: (i) They provide insight into the finite sample behavior of estimators when the true parameter is on the boundary of the parameter space. (ii) They establish conditions under which the asymptotic distribution of the estimator of a subvector of the parameter is not affected by the true values of another subvector being on a boundary of the parameter space; see Section 6.1. (iii) They provide conditions under which the usual formulae for the asymptotic standard errors of extremum estimators are conservative when the true parameter is on a boundary; see Section 6.3. (iv) The results can be used to formulate several methods of generating consistent estimators of the asymptotic standard errors and/or the whole asymptotic distribution of extremum estimators that apply whether or not the true parameter is on a boundary; see Section 6.4. (v) The results can be used to show that the standard bootstrap does not generate consistent estimators of the asymptotic standard errors of extremum estimators when the true parameter is on a boundary; see Andrews (1999). (vi) The estimation results of this paper are useful for constructing Wald-type tests when the null and alternative hypotheses are more complicated than just nonlinear equality restrictions and unrestricted parameters, respectively; see Andrews (1998b). (vii) A by-product of the estimation results is the determination of the asymptotic distribution of the estimator objective function. This can be used to determine the asymptotic distributions of quasi-likelihood ratio test statistics for nonstandard testing problems; see Andrews (1998b). (viii) The results can be used to analyze the properties of model selection procedures for general extremum estimators including cases where smaller models result from the specification of the parameter as a point on the boundary of the parameter space of a larger model. (ix) The results can be used to determine the asymptotic behavior of items that are of interest from a Bayesian perspective, including the (nonstandard) asymptotic distribution of the posterior distribution in likelihood contexts when a parameter is on a boundary. Research on several of the topics above is in progress.

We now discuss examples. This paper and its sequel, Andrews (1997a), consider eight examples. We treat the first two in this paper. The first example is a random coefficient regression model in which some random coefficient variances are zero. The second example is an augmented Dickey-Fuller regression model (i.e., an $AR(p)$ model with a time trend and a root that may equal one) with the largest root restricted to be less than or equal to one and the time trend parameter restricted to be nonnegative.

The third example is an iid regression model with nonlinear equality and/or inequality restrictions on the regression parameters. We note that nonlinear inequality restrictions arise in demand, utility, cost, and profit function estimation when Slutsky conditions, convexity, quasi-convexity, concavity, or quasi-concavity is imposed. The fourth example is the same as the third except that the regressors are integrated. The fifth example is an iid nonlinear median regression model with nonlinear equality and/or inequality restrictions. This model is estimated using the restricted least absolute deviations estimator. The sixth example is a multinomial discrete response model estimated via a MSM estimator. We consider the case where the model includes random coefficients, random effects, or measurement errors and the variances of some of these random terms are zero. The seventh example is a GARCH(1, q^*) or IGARCH(1, q^*) model in which the GARCH MA parameters are restricted to be nonnegative and some of the true GARCH MA parameters equal zero. The eighth example is a partially linear model estimated by the semiparametric LS estimator of Robinson (1988), but subject to nonlinear equality and/or inequality constraints.

The remainder of the paper is organized as follows. Section 2 describes the two examples considered in this paper. Section 3 considers the quadratic approximation of the estimator objective function. Section 4 provides conditions under which the parameter space, suitably shifted and rescaled, can be locally approximated by a cone. Section 5 establishes the asymptotic distribution of the extremum estimator. Section 6 introduces a partitioning of the parameter vector θ that yields a simplification of the asymptotic distribution of the extremum estimator, discusses LAN and LAMN conditions, and provides methods for obtaining consistent asymptotic standard error estimates. An Appendix of Proofs provides proofs of the results given in the paper.

All limits below are taken “as $T \rightarrow \infty$ ” unless stated otherwise. Let “wp $\rightarrow 1$ ” abbreviate “with probability that goes to one as $T \rightarrow \infty$.” Let “for all $\gamma_T \rightarrow 0$ ” abbreviate “for all sequences of positive scalar constants $\{\gamma_T: T \geq 1\}$ for which $\gamma_T \rightarrow 0$.” Let \xrightarrow{p} and \xrightarrow{d} denote convergence in probability and distribution respectively. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues, respectively, of a matrix A . Let ∂A denote the boundary and $\text{cl}(A)$ denote the closure of a set A . Let $S(\theta, \varepsilon)$ denote an open sphere centered at θ with radius ε . Let $C(\theta, \varepsilon)$ denote an open cube centered at θ with sides of length 2ε . Let $:=$ denote “equals by definition.”

2. EXAMPLES

2.1. *Random Coefficient Regression*

Example 1 is a random coefficient regression model. The variances of the random coefficients are nonnegative. We determine the asymptotic distribution of the Gaussian QML estimator when one or more of the random coefficient variances are zero.

The model is

$$\begin{aligned}
 (2.1) \quad Y_t &= \theta_5 + X_t' \gamma_t + \theta_3^{1/2} \varepsilon_t \\
 &= \theta_5 + X_t' \theta_4 + (\theta_3^{1/2} \varepsilon_t + X_t' \Omega^{1/2} (\theta_1, \theta_2) \eta_t), \quad \text{where} \\
 \gamma_t &:= \theta_4 + \Omega^{1/2} (\theta_1, \theta_2) \eta_t.
 \end{aligned}$$

The vector $\gamma_t \in R^b$ is the random coefficient vector. The observed variables are $\{(Y_t, X_t): t \leq T\}$. The regressors are $X_t := (X_{1t}', X_{2t}') \in R^b$, where $X_{1t} \in R^p$ and $X_{2t} \in R^{b_2}$. $\Omega(\theta_1, \theta_2)$ is a diagonal matrix with the random coefficient variance parameters (θ_1', θ_2') on the diagonal. The vector $\theta := (\theta_1', \theta_2', \theta_3, \theta_4', \theta_5')$ is the unknown parameter to be estimated. The random variables $\eta_t \in R^b$ and $\varepsilon_t \in R$ are unobserved errors that satisfy $E\varepsilon_t = 0$, $E\varepsilon_t^2 = 1$, $E(\eta_t | X_t) = \mathbf{0}$ a.s., $E(\eta_t \eta_t' | X_t) = I_b$ a.s., and $E(\eta_t \varepsilon_t | X_t) = \mathbf{0}$ a.s. The random variables $\{(Y_t, X_t, \varepsilon_t, \eta_t): t \leq T\}$ are iid.

The parameter $\theta_1 \in R^p$ contains the random coefficient variances that are on the boundary (i.e., the true value of θ_1, θ_{10} , is $\mathbf{0}$). The parameter $\theta_2 \in R^{b_2}$ contains the random coefficient variances that are not on the boundary (i.e., each element of the true value of θ_2, θ_{20} , is positive). The parameter θ_3 is the idiosyncratic error variance. The true value of θ_3, θ_{30} , is positive. The parameter $\theta_4 \in R^b$ is the deterministic part of the regression coefficients. The parameter $\theta_5 \in R$ is the intercept.

2.2. *Dickey-Fuller Regression Model*

Example 2 is a Dickey-Fuller time series regression model with estimated constant and time trend. This is an autoregressive model of order $b + 1$ that has at most one unit root and all other roots in the stationary region. We consider the case where the parameter space restricts the coefficient on the first lag of the time series (i.e., the potential unit root) to be less than or equal to one and the coefficient on the time trend to be greater than or equal to zero. Thus, the model precludes the possibility of an explosive series and/or of a series with negative growth.

We determine the asymptotic distribution of the LS estimator when the time series has a unit root and a zero coefficient on the time trend. In this case, two

parameters are on the boundary of the parameter space. The true process is the process that defines the null hypothesis of most unit root tests. Most unit root tests, however, impose at most one of the two restrictions on the parameters.

The model is

$$\begin{aligned}
 Y_t &= \theta_1 Y_{t-1} + \theta_2 t + \theta_3 + \Delta \bar{Y}'_{t-1} \theta_4 + \varepsilon_t, \\
 &\text{where } -1 < \theta_1 \leq 1, \quad \theta_2 \geq 0, \\
 (2.2) \quad \Delta \bar{Y}'_{t-1} &= (\Delta Y_{t-1}, \Delta Y_{t-2}, \dots, \Delta Y_{t-b})', \quad \Delta Y_t = Y_t - Y_{t-1}, \\
 E(\varepsilon_t | \mathcal{F}_{t-1}) &= 0 \quad \text{a.s.}, \quad E(\varepsilon_t^2 | \mathcal{F}_{t-1}) = \sigma^2 \quad \text{a.s.}, \\
 \mathcal{F}_t &= \sigma(\varepsilon_1, \dots, \varepsilon_t),
 \end{aligned}$$

$Y_t, \varepsilon_t, \theta_1, \theta_2, \theta_3 \in R$, and $\Delta \bar{Y}'_{t-1}, \theta_4 \in R^b$. The observed time series is $\{Y_t: -b \leq t \leq T\}$. The parameter vector to be estimated is $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)'$.

3. QUADRATIC APPROXIMATION OF THE OBJECTIVE FUNCTION

3.1. Definition of the Extremum Estimator and Consistency

Let Y_T denote the data matrix when the sample size is T for $T = 1, 2, \dots$. We consider an estimator objective function $\ell_T(\theta)$ that depends on Y_T . Maximization of $\ell_T(\theta)$ over a parameter space $\Theta \subset R^s$ yields the estimator $\hat{\theta}$ that we analyze in this paper. The estimator objective function $\ell_T(\theta)$ can be a log-likelihood function, a quasi-log likelihood function, a least squares criterion function, a GMM objective function, a minimum distance objective function, an objective function that depends on finite or infinite dimensional preliminary estimators, etc.

By definition, the extremum estimator $\hat{\theta}$ satisfies $\hat{\theta} \in \Theta$ and

$$(3.1) \quad \ell_T(\hat{\theta}) = \sup_{\theta \in \Theta} \ell_T(\theta) + o_p(1).$$

We only require that $\ell_T(\hat{\theta})$ be within $o_p(1)$ of the global maximum of $\ell_T(\theta)$ over $\theta \in \Theta$, rather than the exact global minimum, because this circumvents the question of existence and eases the computational burden.

Let θ_0 denote the pseudo-true value of the parameter θ . By assumption, $\theta_0 \in \text{cl}(\Theta)$.

We assume consistency of $\hat{\theta}$ for θ_0 :

$$\text{ASSUMPTION 1: } \hat{\theta} = \theta_0 + o_p(1).$$

A well-known sufficient condition for Assumption 1 that often holds when the data do not involve trending variables is the following:

ASSUMPTION 1*: (a) For some function $\ell(\theta): \Theta \rightarrow R$, $\sup_{\theta \in \Theta} |T^{-1} \ell_T(\theta) - \ell(\theta)| \xrightarrow{p} 0$. (b) For all $\varepsilon > 0$, $\sup_{\theta \in \Theta/S(\theta_0, \varepsilon)} \ell(\theta) < \ell(\theta_0)$, where $\Theta/S(\theta_0, \varepsilon)$ denotes all vectors θ in Θ but not in $S(\theta_0, \varepsilon)$.

Note that here and below a superscript *, 2*, or 3* on an assumption denotes that the assumption is sufficient (sometimes only in the presence of other specified assumptions) for the unsuperscripted assumption.

Assumption 1*(a) is a uniform convergence condition that can be verified by using a uniform law of large numbers; see Andrews (1992) and references therein. Assumption 1*(b) is an asymptotic identification condition. Sufficient conditions for Assumption 1*(b), which we call Assumption 1*(b*), are (i) $\ell(\theta)$ is uniquely maximized over Θ at θ_0 , (ii) $\ell(\theta)$ is continuous on Θ , and (iii) Θ is compact.

A second sufficient condition for Assumption 1 that allows for the case where $T^{-1} \ell_T(\theta) \rightarrow_p \ell(\theta)$ for all $\theta \in \Theta$ and $\ell(\theta) = -\infty$ for some $\theta \neq \theta_0$ is given by Pfanzagl (1969, Theorem 1.12). An extension of Pfanzagl's result is used in Andrews (1997a) for the IGARCH(1, q^*) Example.

When the data involve trending variables no generally applicable proof of consistency is available. Usually, one has to establish consistency on a case by case basis. For linear models this is often straightforward, but for nonlinear models it can be difficult. See Andrews and McDermott (1995) and Saikkonen (1995) for some results regarding the latter models.

3.2. Quadratic Approximation of the Objective Function

We consider the case where the estimator objective function $\ell_T(\theta)$ has a quadratic expansion in θ about θ_0 :

$$(3.2) \quad \ell_T(\theta) := \ell_T(\theta_0) + D\ell_T(\theta_0)'(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)' D^2\ell_T(\theta_0)(\theta - \theta_0) + R_T(\theta).$$

The remainder term $R_T(\theta)$ specifies the sense in which the expansion holds. When $\ell_T(\theta)$ has partial derivatives of order two with respect to (wrt) θ , $D\ell_T(\theta_0)$ and $D^2\ell_T(\theta_0)$ typically are the s -vector and $s \times s$ matrix of first and second partial derivatives, respectively, of $\ell_T(\theta)$ with respect to θ evaluated at θ_0 . We do not require $\ell_T(\theta)$ to have partial derivatives of order two wrt θ , however, for two reasons. First, $\ell_T(\theta)$ is not defined on a neighborhood of θ_0 for some of our applications of interest. Thus, at best, $D\ell_T(\theta_0)$ will consist of left or right partial derivatives for some of its elements. Second, $\ell_T(\theta)$ involves absolute value or sign functions in some applications of interest, so pointwise partial derivatives (or even left or right pointwise partial derivatives) do not exist in some cases. Nevertheless, $\ell_T(\theta)$ is often differentiable in a stochastic sense, which is the case considered here.

We introduce a norming matrix B_T for $D\ell_T(\theta_0)$ and $D^2\ell_T(\theta_0)$ so that each is $O_p(1)$ but not $o_p(1)$ (as indicated in Assumption 3 below). B_T is a deterministic $s \times s$ matrix. In most cases with nontrending data, $B_T = T^{1/2}I_s$. In some cases with nontrending data, however, it is useful to take $B_T = T^{1/2}M$, where M is a nonsingular nondiagonal matrix. By appropriate choice of M , one may be able to obtain a block diagonal normalized “quasi-information” matrix \mathcal{F}_T , defined below. This yields a simplified expression for the asymptotic distribution of the extremum estimator. This occurs with the GARCH(1, q^*) Example of Andrews (1997a).

With trending data, B_T is always more complicated than $T^{1/2}I_s$. For example, in the Dickey-Fuller Example 2, B_T is an asymmetric matrix of the form $B_T = \Upsilon_T M$, where Υ_T is diagonal, $\lambda_{\min}(\Upsilon_T) \rightarrow \infty$, and M is nonsingular.

Let

$$(3.3) \quad \mathcal{F}_T := -B_T^{-1} D^2\ell_T(\theta_0) B_T^{-1} \quad \text{and} \quad Z_T := \mathcal{F}_T^{-1} B_T^{-1} D\ell_T(\theta_0),$$

where B_T^{-1} denotes $(B_T^{-1})'$. The quadratic expansion of (3.2) can be rewritten as

$$(3.4) \quad \begin{aligned} \ell_T(\theta) &= \ell_T(\theta_0) + \frac{1}{2} Z_T' \mathcal{F}_T Z_T - \frac{1}{2} q_T(B_T(\theta - \theta_0)) + R_T(\theta), \quad \text{where} \\ q_T(\lambda) &:= (\lambda - Z_T)' \mathcal{F}_T (\lambda - Z_T) \quad \text{for } \lambda \in R^s. \end{aligned}$$

The terms in the quadratic expansion of $\ell_T(\theta)$ are assumed to satisfy the following assumptions:

ASSUMPTION 2: For all $0 < \gamma < \infty$, $\sup_{\theta \in \Theta: \|B_T(\theta - \theta_0)\| \leq \gamma} |R_T(\theta)| = o_p(1)$ for some nonrandom matrices B_T for which $\lambda_{\min}(B_T) \rightarrow \infty$.

ASSUMPTION 3: $(B_T^{-1} D\ell_T(\theta_0), \mathcal{F}_T) \rightarrow_d (G, \mathcal{F})$ for some random variables $G \in R^s$ and $\mathcal{F} \in R^{s \times s}$ for which \mathcal{F} is symmetric and nonsingular with probability one.

A sufficient condition for Assumption 2 that we often employ is the following:

ASSUMPTION 2*: For all $\gamma_T \rightarrow 0$, $\sup_{\theta \in \Theta: \|\theta - \theta_0\| \leq \gamma_T} |R_T(\theta)| / (1 + \|B_T(\theta - \theta_0)\|)^2 = o_p(1)$ for some nonrandom matrices B_T for which $\lambda_{\min}(B_T) \rightarrow \infty$.

In the next subsection, we give a sufficient condition for Assumption 2*. It relies on the existence of left and/or right partial derivatives of $\ell_T(\theta)$. Andrews (1997a) gives two additional sufficient conditions for Assumption 2*. The first relies on a stochastic differentiability condition that generalizes that of Pollard (1985) and van der Vaart and Wellner (1996, Theorem 3.2.16). The second applies specifically to GMM and minimum distance estimators and generalizes the stochastic equicontinuity condition of Pakes and Pollard (1989). Andrews (1997a) also gives a condition that is sufficient for Assumption 2, but not for

Assumption 2*. It covers the case where $\ell_T(\theta)$ is the sum of Lipschitz functions of θ . None of the sufficient conditions referred to above requires the parameter space Θ or the domain of $\ell_T(\theta)$ to include a neighborhood of θ_0 .

Assumption 3 allows the normalized “information” matrix \mathcal{F}_T to be random even in the limit as $T \rightarrow \infty$. This is necessary to cover models with stochastic trends, such as unit root and cointegration models. In models with no stochastic trends (but possibly with deterministic trends), \mathcal{F}_T converges to a nonstochastic limit \mathcal{F} . In this case, one can take \mathcal{F}_T to be the nonstochastic limit \mathcal{F} in the quadratic expansion of (3.4) and the remainder term $R_T(\theta)$ can absorb the difference. Thus, a sufficient condition for Assumption 3, that is applicable in models with no stochastic trends, is the following:

ASSUMPTION 3*: $B_T^{-1} D\ell_T(\theta_0) \rightarrow_d G$ for some random variable $G \in R^s$, $\mathcal{F}_T \in R^{s \times s}$ is nonrandom and does not depend on T , and $\mathcal{F} (= \mathcal{F}_T)$ is symmetric and nonsingular.

In quasi-log likelihood cases, Assumption 3 is implied by the convergence in distribution of the normalized score function and the convergence in distribution or probability of the “Hessian” of the likelihood. In such cases, Assumption 3 usually follows from the central limit theorem (CLT) and the law of large numbers (LLN) in models without stochastic trends and from an invariance principle in models with stochastic trends. In GMM cases, Assumption 3 usually follows from the CLT and several convergence in probability results. In some cases, such as with Han’s (1987) maximum rank correlation estimator (see Sherman (1993)), Assumption 3 follows from a CLT and LLN for U -statistics. In minimum distance, semiparametric, and other cases that rely on preliminary estimators, verification of Assumption 3 requires asymptotic results for the preliminary estimators. Results already in the literature often can be used.

We describe the limit quantities G and \mathcal{F} in more detail in Section 6.

3.3. A Sufficient Condition for the Quadratic Approximation of the Objective Function

Here, we provide a sufficient condition for Assumption 2* that relies on smoothness of $\ell_T(\theta)$. It uses a Taylor expansion of $\ell_T(\theta)$ about θ_0 , but does not require $\ell_T(\theta)$ to be defined in a neighborhood of θ_0 . The requisite Taylor’s theorem is established in the Appendix.

First, we introduce some terminology. Let f be a function whose domain includes $\mathcal{X} \subset R^s$. Let $a \in \mathcal{X}$. We want a Taylor expansion of $f(x)$ about $f(a)$ to hold for points $x \in \mathcal{X}$. We suppose $\mathcal{X} - a$ equals the intersection of a union of orthants and an open cube, $C(\mathbf{0}, \varepsilon)$, centered at $\mathbf{0}$ with edges of length 2ε for some $\varepsilon > 0$. (Thus, $\mathcal{X} - a$ is locally equal to a union of orthants.) As defined, \mathcal{X} is a cube centered at a with some “orthants” of the cube removed.

We say f has *left/right (l/r) partial derivatives* (of order 1) on \mathcal{X} if it has partial derivatives at each interior point of \mathcal{X} ; if it has partial derivatives at each

boundary point of \mathcal{X} with respect to (wrt) coordinates that can be perturbed to the left and right; and if it has left (right) partial derivatives at each boundary point of \mathcal{X} wrt coordinates that can be perturbed only to the left (right). Note that the shape of \mathcal{X} is such that for all $x \in \mathcal{X}$ and for all coordinates x_j of x it is possible to perturb x_j to the right or left or both and stay within \mathcal{X} . Thus, it is possible to define the left, the right, or the two-sided partial derivative of f wrt x_j at $x \forall j \leq s$ and $\forall x \in \mathcal{X}$.

We say f has l/r partial derivatives of order k on \mathcal{X} for $k \geq 2$ if f has l/r partial derivatives of order $k - 1$ on \mathcal{X} and each of the latter has l/r partial derivatives on \mathcal{X} . We say f has continuous l/r partial derivatives of order k on \mathcal{X} if f has l/r partial derivatives of order k on \mathcal{X} , each of which is continuous at all points in \mathcal{X} , where continuity is defined in terms of local perturbations only within \mathcal{X} .

A sufficient condition for Assumption 2* is the following:

ASSUMPTION 2^{2*}: (a) *The domain of $\ell_T(\theta)$ includes a set Θ^+ that satisfies (i) $\Theta^+ - \theta_0$ equals the intersection of a union of orthants and an open cube $C(\mathbf{0}, \varepsilon)$ for some $\varepsilon > 0$ and (ii) $\Theta \cap S(\theta_0, \varepsilon_1) \subset \Theta^+$ for some $\varepsilon_1 > 0$, where Θ is the parameter space of Assumption 2*.*

(b) *$\ell_T(\theta)$ has continuous l/r partial derivatives of order 2 on $\Theta^+ \forall T \geq 1$ with probability one.*

(c) *For all $\gamma_T \rightarrow 0$,*

$$\sup_{\theta \in \Theta: \|\theta - \theta_0\| \leq \gamma_T} \left\| B_T^{-1} \left(\frac{\partial^2}{\partial \theta \partial \theta'} \ell_T(\theta) - \frac{\partial^2}{\partial \theta \partial \theta'} \ell_T(\theta_0) \right) B_T^{-1} \right\| = o_p(1),$$

where $(\partial/\partial \theta)\ell_T(\theta)$ and $(\partial^2/\partial \theta \partial \theta')\ell_T(\theta)$ denote the s vector and $s \times s$ matrix of l/r partial derivatives of $\ell_T(\theta)$ of orders one and two respectively.

Assumption 2^{2*}(a) specifies a set Θ^+ with a special shape on which $\ell_T(\theta)$ must be defined. For each $\theta \in \Theta^+$, $\ell_T(\theta)$ has a quadratic approximation via the Taylor's theorem in the Appendix. On the other hand, Assumption 2^{2*} does not require that near θ_0 the parameter space Θ is a union of orthants centered at θ_0 . What Assumption 2^{2*} requires is that Θ be contained in such a set near θ_0 . If $\Theta - \theta_0$ happens to be a union of orthants local to $\mathbf{0}$, then one can take $\Theta^+ = \Theta \cap C(\theta_0, \varepsilon)$ in Assumption 2^{2*}.

Assumption 2^{2*}(b) is designed to hold in cases in which θ_0 is on the boundary of the set where the objective function can be defined, such as in the random coefficient regression example. Of course, it also holds in cases in which θ_0 is on the boundary of the parameter space, but the objective function can be defined on a neighborhood of θ_0 . In such cases, one can take Θ^+ to be an open cube $C(\theta_0, \varepsilon)$ for some $\varepsilon > 0$.

Assumption 2^{2*}(c) can be verified in the case of nontrending data as follows. Suppose $B_T = T^{1/2}M$, $(\partial^2/\partial \theta \partial \theta')\ell_T(\theta)/T \rightarrow_p (\partial^2/\partial \theta \partial \theta')\ell(\theta)$ uniformly over

$\theta \in \Theta \cap S(\theta_0, \varepsilon_2)$ for some $\varepsilon_2 > 0$ and some nonrandom function $(\partial^2/\partial\theta\partial\theta')\ell(\theta)$ that is continuous at θ_0 . Then Assumption 2^{2*}(c) holds. The uniform convergence of $(\partial^2/\partial\theta\partial\theta')\ell_T(\theta)/T$ can be established via a uniform LLN; e.g., see Andrews (1992).

When stochastic or deterministic trends enter the objective function in a linear fashion, then part of the matrix $(\partial^2/\partial\theta\partial\theta')\ell_T(\theta)$ does not depend on θ and Assumption 2^{2*}(c) holds trivially for that part of $(\partial^2/\partial\theta\partial\theta')\ell_T(\theta)$.

LEMMA 1: (a) *Assumption 2^{2*} implies Assumption 2* with $D\ell_T(\theta_0)$ and $D^2\ell_T(\theta_0)$ of (3.2) given by $(\partial/\partial\theta)\ell_T(\theta_0)$ and $(\partial^2/\partial\theta\partial\theta')\ell_T(\theta_0)$ (i.e., by the 1/r partial derivatives of $\ell_T(\theta)$ at θ_0 of orders one and two) respectively.*

(b) *If Assumption 2^{2*} holds and $-B_T^{-1}(\partial^2/\partial\theta\partial\theta')\ell_T(\theta_0)B_T^{-1} \rightarrow_p \mathcal{F}$ for some nonrandom matrix \mathcal{F} , then Assumption 2* holds with $D\ell_T(\theta_0)$ of (3.2) given by $(\partial/\partial\theta)\ell_T(\theta_0)$ and with $D^2\ell_T(\theta_0)$ of (3.2) given by either $(\partial^2/\partial\theta\partial\theta')\ell_T(\theta_0)$ or $-B_T'\mathcal{F}B_T$.*

COMMENT: The proof of the Lemma and other results below are given in the Appendix.

3.4. Rate of Convergence of the Extremum Estimator

To obtain the asymptotic distribution of the extremum estimator, we first establish its rate of convergence.

ASSUMPTION 4: $B_T(\hat{\theta} - \theta_0) = O_p(1)$.

Sufficient conditions for Assumption 4 are given in the following theorem. Alternative sufficient conditions are given by Andrews (1997a) and van der Vaart and Wellner (1996, Theorems 3.2.5 and 3.2.10).

THEOREM 1: *Assumptions 1, 2*, and 3 imply Assumption 4.*

COMMENTS: 1. This Theorem shows why it is often useful to employ Assumption 2* rather than Assumption 2—Assumption 2* not only delivers the desired quadratic approximation of the estimator objective function, but it delivers Assumption 4 as well. There are some occasions, however, where it is preferable to employ Assumption 2 and use a different argument to verify Assumption 4; see Andrews (1997a).

2. The Theorem holds even if $o_p(1)$ is replaced by $O_p(1)$ in (3.1) and Assumption 3 is replaced by $B_T^{-1} D\ell_T(\theta_0) = O_p(1)$, $\lambda_{\max}(\mathcal{F}_T) = O_p(1)$, $\lambda_{\min}^{-1}(\mathcal{F}_T) = O_p(1)$, and \mathcal{F}_T is symmetric wp $\rightarrow 1$.

3. The proof of the Theorem is similar to numerous proofs in the literature; e.g., see the proof of Lemma 1 of Chernoff (1954).

3.5. *Quadratic Approximation of the Objective Function (Continued)*

Let $\hat{\theta}_q$ denote an (approximate) maximizer of the quadratic approximation to $\ell_T(\theta)$ or, equivalently, an (approximate) minimizer of $q_T(B_T(\theta - \theta_0))$. By definition, $\hat{\theta}_q$ satisfies $\hat{\theta}_q \in \text{cl}(\Theta)$ and

$$(3.5) \quad q_T\left(B_T\left(\hat{\theta}_q - \theta_0\right)\right) = \inf_{\theta \in \Theta} q_T\left(B_T\left(\theta - \theta_0\right)\right) + o_p(1).$$

Note that

$$(3.6) \quad \begin{aligned} \inf_{\theta \in \Theta} q_T\left(B_T\left(\theta - \theta_0\right)\right) &= \inf_{\lambda \in B_T\left(\Theta - \theta_0\right)} q_T\left(\lambda\right), \quad \text{where} \\ B_T\left(\Theta - \theta_0\right) &:= \left\{\lambda \in R^s: \lambda = B_T\left(\theta - \theta_0\right) \text{ for some } \theta \in \Theta\right\}. \end{aligned}$$

Our next result shows that $\hat{\theta}_q$ is B_T -consistent and the objective function at $\hat{\theta}$ is a simple shift of the quadratic function $-\frac{1}{2}q_T(B_T(\theta - \theta_0))$ evaluated at $\hat{\theta}_q$.

THEOREM 2: *Suppose Assumptions 2–4 hold. Then,*

- (a) $B_T(\hat{\theta}_q - \theta_0) = O_p(1)$,
- (b) $\ell_T(\hat{\theta}) = \ell_T(\theta_0) + \frac{1}{2}Z_T' \mathcal{F}_T Z_T - \frac{1}{2}q_T(B_T(\hat{\theta} - \theta_0)) + o_p(1)$,
- (c) $\ell_T(\hat{\theta}_q) = \ell_T(\theta_0) + \frac{1}{2}Z_T' \mathcal{F}_T Z_T - \frac{1}{2}q_T(B_T(\hat{\theta}_q - \theta_0)) + o_p(1)$,
- (d) $\ell_T(\hat{\theta}) = \ell_T(\hat{\theta}_q) + o_p(1)$,
- (e) $q_T(B_T(\hat{\theta} - \theta_0)) = q_T(B_T(\hat{\theta}_q - \theta_0)) + o_p(1)$, and
- (f) $\ell_T(\hat{\theta}) = \ell_T(\theta_0) + \frac{1}{2}Z_T' \mathcal{F}_T Z_T - \frac{1}{2}q_T(B_T(\hat{\theta}_q - \theta_0)) + o_p(1)$.

COMMENT: Part (a) holds even if $o_p(1)$ is replaced by $O_p(1)$ in (3.5), Assumption 3 is replaced by $B_T^{-1}D\ell_T(\theta_0) = O_p(1)$, $\lambda_{\max}(\mathcal{F}_T) = O_p(1)$, $\lambda_{\min}^{-1}(\mathcal{F}_T) = O_p(1)$, and \mathcal{F}_T is symmetric w.p $\rightarrow 1$, and Assumption 4 is replaced by the assumption that θ_0 is in the closure of Θ .

3.6. *Examples (Continued)*

In this section, for the two examples of Section 2, we specify the estimator objective function $\ell_T(\theta)$, the parameter space Θ , and assumptions that are sufficient for Assumptions 1, 2*, and 3 (which imply Assumptions 2–4). We verify Assumptions 2* and 3 for both examples. Verification of Assumption 1 for both examples is given in Andrews (1997b).

3.6.1. *Random Coefficient Regression*

In Example 1, we consider the Gaussian QML estimator, which is based on the supposition that ε_t and η_t are normally distributed and independent of X_t .

The Gaussian quasi-log likelihood function is

$$(3.7) \quad \begin{aligned} \ell_T(\theta) := & -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^T \ln(\theta_3 + X_t' \Omega(\theta_1, \theta_2) X_t) \\ & - \frac{1}{2} \sum_{t=1}^T (Y_t - \theta_5 - X_t' \theta_4)^2 / (\theta_3 + X_t' \Omega(\theta_1, \theta_2) X_t). \end{aligned}$$

The true parameter vector θ_0 is

$$(3.8) \quad \theta_0 := (\theta'_{10}, \theta'_{20}, \theta_{30}, \theta'_{40}, \theta_{50})' = (\mathbf{0}', \theta'_{20}, \theta_{30}, \theta'_{40}, \theta_{50})',$$

where $\theta_{20} > 0$ (element by element) and $\theta_{30} > 0$. The parameter space Θ is a bounded subset of R^s that restricts all elements of θ_1 and θ_2 to be nonnegative and that bounds θ_3 away from zero:

$$(3.9) \quad \begin{aligned} \Theta := \{ \theta \in R^s : \theta = (\theta'_1, \theta'_2, \theta_3, \theta'_4, \theta_5)' , \theta_1 \geq 0, \theta_2 \geq 0, \theta_3 \geq c, \\ \| \theta_j \| \leq M_j \quad \forall j \leq 5 \} \end{aligned}$$

for some $c > 0$ and $0 < M_j < \infty \quad \forall j \leq 5$.

The quadratic approximation of $\ell_T(\theta)$ at θ_0 is defined as follows. Let

$$(3.10) \quad \begin{aligned} X_t &:= (X_{t1}, \dots, X_{tb})', \quad X_t^2 := (X_{t1}^2, \dots, X_{tb}^2)', \\ W_t &:= (X_t', 1)', \quad W_t^2 := (X_t^{2'}, 1)', \\ \text{res}_t(\theta) &:= Y_t - \theta_5 - X_t' \theta_4, \quad \text{and} \quad \text{var}_t(\theta) := \theta_3 + X_t' \Omega(\theta_1, \theta_2) X_t. \end{aligned}$$

Define

$$(3.11) \quad \begin{aligned} D\ell_T(\theta_0) &:= \sum_{t=1}^T \left(\frac{\text{res}_t^2(\theta_0) - \text{var}_t(\theta_0)}{2 \text{var}_t^2(\theta_0)} W_t^{2'}, \frac{\text{res}_t(\theta_0)}{\text{var}_t(\theta_0)} W_t' \right)', \\ D^2\ell_T(\theta_0) &:= -T\mathcal{F}, \\ \mathcal{F} := \mathcal{F}_T &:= \begin{bmatrix} \frac{1}{2} E W_t^2 W_t^{2'} / \text{var}_t^2(\theta_0) & \mathbf{0} \\ \mathbf{0} & E W_t W_t' / \text{var}_t(\theta_0) \end{bmatrix}, \\ B_T &:= T^{1/2} I_s, \quad \text{and} \quad Z_T := \mathcal{F}^{-1} T^{-1/2} D\ell_T(\theta_0). \end{aligned}$$

With these definitions, the quadratic approximations of (3.2) and (3.4) hold (in particular, Assumption 2^{2*} holds) under the assumptions above and the moment conditions below.

We assume that

$$(3.12) \quad E\|\varepsilon_t X_t\|^4 < \infty, \quad E\|\eta_t\|^4 \|X_t\|^8 < \infty,$$

$$(3.13) \quad EW_t^2 W_t^{2'} / \text{var}_t^2(\theta_0) > 0 \quad \text{and} \quad EW_t W_t' / \text{var}_t(\theta_0) > 0,$$

where “> 0” denotes “is positive definite.”

We verify Assumption 2* for this example using Assumption 2^{2*} and Lemma 1(b). Let $\Theta^+ = \Theta \cap C(\theta_0, \varepsilon)$ for some $0 < \varepsilon < \min\{M_j : j \leq 5\}$ (where the M_j are specified in the definition of Θ). Then, $\Theta^+ - \theta_0$ equals the intersection of the orthant $\Lambda := (R^+)^p \times R^{s-p}$ and the open cube $C(\mathbf{0}, \varepsilon)$, as required by Assumption 2^{2*}(a)(i). Also, $\Theta \cap S(\theta_0, \varepsilon_1) \subset \Theta^+$ for $0 < \varepsilon_1 < \varepsilon$, as required by Assumption 2^{2*}(a)(ii). The quasi-likelihood function $\ell_T(\theta)$ of (3.7) has continuous l/r partial derivatives of order two on Θ^+ , as required by Assumption 2^{2*}(b).

The matrix of l/r partial derivatives of order two of $\ell_T(\theta)$ is

$$(3.14) \quad \frac{\partial^2}{\partial \theta \partial \theta'} \ell_T(\theta) := - \sum_{t=1}^T \begin{pmatrix} \frac{2 \text{res}_t^2(\theta) - \text{var}_t(\theta)}{\text{var}_t^3(\theta)} W_t^2 W_t^{2'} & \frac{\text{res}_t(\theta)}{\text{var}_t^2(\theta)} W_t W_t^{2'} \\ \frac{\text{res}_t(\theta)}{\text{var}_t^2(\theta)} W_t^2 W_t' & \frac{1}{\text{var}_t(\theta)} W_t W_t' \end{pmatrix}.$$

By a uniform LLN (e.g., see Andrews (1992, Theorem 4) using Assumption TSE-1C), $\sup_{\theta \in \Theta} |T^{-1}(\partial^2 / \partial \theta \partial \theta') \ell_T(\theta) - T^{-1}E(\partial^2 / \partial \theta \partial \theta') \ell_T(\theta)| \rightarrow_p 0$. Also, $T^{-1}E(\partial^2 / \partial \theta \partial \theta') \ell_T(\theta)$ is continuous at θ_0 . In consequence, Assumption 2^{2*}(c) holds. By a LLN, $-T^{-1}(\partial^2 / \partial \theta \partial \theta') \ell_T(\theta_0) \rightarrow_p \mathcal{I}$, where \mathcal{I} is defined in (3.11). In consequence, Lemma 1(b) is applicable and Assumption 2* holds with $D\ell_T(\theta_0)$ and $D^2\ell_T(\theta_0)$ of (3.2) as defined in (3.11).

In this example, \mathcal{F}_T does not depend on T and $\mathcal{F}(=\mathcal{F}_T)$ is symmetric and positive definite by (3.11) and (3.13). Thus, Assumption 3* holds provided $T^{-1/2}D\ell_T(\theta_0) \rightarrow_d G$ for some G . By the definition of $D\ell_T(\theta_0)$ in (3.11) and the moment assumptions of (3.12), the CLT for iid mean zero finite variance random variables yields

$$(3.15) \quad T^{-1/2}D\ell_T(\theta_0) \xrightarrow{d} G \sim N(\mathbf{0}, \mathcal{I}), \quad \text{where} \quad \mathcal{I} := \begin{bmatrix} \frac{1}{4}E \frac{(\text{res}_t^2(\theta_0) - \text{var}_t(\theta_0))^2}{\text{var}_t^4(\theta_0)} W_t^2 W_t^{2'} & \frac{1}{2}E \frac{\text{res}_t^3(\theta_0)}{\text{var}_t^3(\theta_0)} W_t^2 W_t' \\ \frac{1}{2}E \frac{\text{res}_t^3(\theta_0)}{\text{var}_t^3(\theta_0)} W_t W_t^{2'} & EW_t W_t' / \text{var}_t(\theta_0) \end{bmatrix}.$$

3.6.2. *Dickey-Fuller Regression*

In Example 2, we consider the LS estimator. The estimator objective function is

$$(3.16) \quad \ell_T(\theta) := -\frac{1}{2} \sum_{t=1}^T (Y_t - X_t' \theta)^2, \quad \text{where} \quad X_t := (Y_{t-1}, t, 1, \Delta \bar{Y}_{t-1})'.$$

The parameter space Θ is given by

$$(3.17) \quad \Theta := \{\theta \in R^s: -1 < \theta_1 \leq 1, \theta_2 \geq 0, g(z) := 1 - \theta_{41}z - \dots - \theta_{4b}z^b \text{ has roots outside the unit circle, where } \theta_4 := (\theta_{41}, \dots, \theta_{4b})'\}.$$

The true parameter θ_0 corresponds to a unit root model with nonnegative drift:

$$(3.18) \quad \theta_0 := (\theta_{10}, \theta_{20}, \theta_{30}, \theta'_{40})' = (1, 0, \theta_{30}, \theta'_{40})',$$

where $\theta_{30} \geq 0$ and θ_{40} has a characteristic equation with roots outside the unit circle. Note that the latter implies that $1 - \mathbf{1}'\theta_{40} > 0$, where $\mathbf{1} := (1, \dots, 1)' \in R^b$. We could consider the case of negative drift (i.e., $\theta_{30} < 0$) with little extra work. But, this case is not of great practical importance. We assume that $\sigma^2 > 0$. As defined, θ_{10} and θ_{20} are, and θ_{30} and θ_{40} are not, on the boundary of Θ .

The quadratic approximation of (3.2) and (3.4) holds with

$$(3.19) \quad \begin{aligned} D\ell_T(\theta_0) &:= \sum_{t=1}^T \varepsilon_t X_t, & D^2\ell_T(\theta_0) &:= -\sum_{t=1}^T X_t X_t', \\ R_T(\theta) &:= 0, & B_T &:= \mathcal{T}_T M, \\ M &:= \begin{bmatrix} 1 & 0 & 0 & \mathbf{0}' \\ \mu_0 & 1 & 0 & \mathbf{0}' \\ -\mu_0 & 0 & 1 & \mu_0 \mathbf{1}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I_b \end{bmatrix}, \\ L &:= M^{-1} = \begin{bmatrix} 1 & -\mu_0 & \mu_0 & \mathbf{0}' \\ 0 & 1 & 0 & \mathbf{0}' \\ 0 & 0 & 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & -\mu_0 \mathbf{1} & I_b \end{bmatrix}, \\ \mu_0 &:= \theta_{30}/(1 - \mathbf{1}'\theta_{40}), & \mathcal{T}_T &:= \text{Diag}(T, T^{3/2}, T^{1/2}, \dots, T^{1/2}), \quad \text{and} \\ \mathcal{F}_T &:= -B_T^{-1} D^2\ell_T(\theta_0) B_T^{-1} = \mathcal{T}_T^{-1} \sum_{t=1}^T L X_t (L X_t)' \mathcal{T}_T^{-1}. \end{aligned}$$

(The matrix M is determined via $M=L^{-1}$, where L is chosen such that $(\mathcal{T}_T^{-1}\sum_{t=1}^T \varepsilon_t \times LX_t, \mathcal{T}_T^{-1}\sum_{t=1}^T LX_t(LX_t)'\mathcal{T}_T^{-1})$ converges in distribution to (G, \mathcal{F}) in (3.21) below.) Assumption 2* holds because $R_T(\theta) = 0$.

To verify Assumption 3, we impose the following mild tail condition on the errors: For some random variable ε , some $0 < c < \infty$, and some $\eta > 0$,

$$(3.20) \quad P(|\varepsilon_t| \geq x) \leq cP(|\varepsilon| \geq x) \quad \forall x > 0 \quad \text{and} \quad E|\varepsilon|^{2+\eta} < \infty.$$

Under the assumptions given, we have

$$(3.21) \quad \begin{aligned} & (B_T^{-1}D\ell_T(\theta_0), \mathcal{F}_T) \\ & := \left(\mathcal{T}_T^{-1} \sum_{t=1}^T \varepsilon_t LX_t, \mathcal{T}_T^{-1} \sum_{t=1}^T LX_t(LX_t)'\mathcal{T}_T^{-1} \right) \\ & \xrightarrow{d} (G, \mathcal{F}), \quad \text{where} \\ & G := \begin{pmatrix} \frac{1}{2}\sigma\lambda(W^2(1) - 1) \\ \sigma(W(1) - \int_0^1 W(r) dr) \\ \sigma W(1) \\ G_4 \end{pmatrix}, \quad G_4 \sim N(\mathbf{0}, V), \\ & \mathcal{F} := \begin{pmatrix} \lambda^2 \int_0^1 W^2(r) dr & \lambda \int_0^1 rW(r) dr & \lambda \int_0^1 W(r) dr & \mathbf{0}' \\ \lambda \int_0^1 rW(r) dr & 1/3 & 1/2 & \mathbf{0}' \\ \lambda \int_0^1 W(r) dr & 1/2 & 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & V \end{pmatrix}, \\ & \lambda := \sigma/(1 - \mathbf{1}'\theta_{40}), \quad \gamma_j = \text{Cov}(\Delta Y_t, \Delta Y_{t-j}) \quad \forall j = 0, \dots, b-1, \\ & V := \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{b-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{b-2} \\ \vdots & & \vdots & & \vdots \\ \gamma_{b-1} & \gamma_{b-2} & \gamma_{b-3} & \cdots & \gamma_0 \end{pmatrix}, \end{aligned}$$

and $W(\cdot)$ is a standard scalar Brownian motion on $[0, 1]$ that is independent of G_4 . Note that γ_j is the j th order autocovariance of a b th order autoregressive process with autoregressive parameter θ_{40} and error variance σ^2 . Thus, γ_j depends only on θ_{40} and σ^2 . The matrix V is nonsingular and \mathcal{F} is nonsingular with probability one. Thus, Assumption 3 holds.

The proof of (3.21) is given in Exercise 17.6 of Hamilton (1994, p. 540) extended to allow for errors $\{\varepsilon_t; t \geq 1\}$ that form a martingale difference sequence, rather than an iid sequence, using the invariance principle for linear processes in Theorem 3.15 of Phillips and Solo (1992, p. 983) in place of the one used by Hamilton.

4. THE PARAMETER SPACE

4.1. *Local Approximation to the Shifted and Rescaled Parameter Space*

This section provides conditions on the parameter space under which we can derive the asymptotic distribution of $\hat{\theta}$. It is apparent from Assumption 4 that the asymptotic distribution of $\hat{\theta}$ depends on the features of the parameter space Θ only near θ_0 . In particular, we find that the asymptotic distribution of $\hat{\theta}$ depends on a local approximation to the shifted and rescaled parameter space $B_T(\Theta - \theta_0)/b_T$, where $\{b_T: T \geq 1\}$ is some sequence of scalar constants for which $b_T \rightarrow \infty$.

If Θ includes a neighborhood of θ_0 , then

$$(4.1) \quad \inf_{\lambda \in B_T(\Theta - \theta_0)} q_T(\lambda) = \inf_{\lambda \in \Lambda} q_T(\lambda) + o_p(1),$$

where $\Lambda = R^s$. This follows because $B_T(\Theta - \theta_0) \rightarrow R^s$ (provided $\lambda_{\min}(B_T) \rightarrow \infty$).

Our interest lies in the case where Θ does not include a neighborhood of θ_0 . Thus, we do not require (4.1) to hold with $\Lambda = R^s$. Rather, we find sufficient conditions for (4.1) to hold with Λ given by a cone. By definition, a set $\Lambda \subset R^s$ is a *cone* if $\lambda \in \Lambda$ implies $a\lambda \in \Lambda \forall a \in R$ with $a > 0$. Examples of cones include R^s , linear subspaces, orthants, unions of orthants, and sets defined by linear equalities and/or inequalities of the form $\Gamma_a \lambda = \mathbf{0}$ and $\Gamma_b \lambda \leq \mathbf{0}$, where Γ_j is a $k_j \times s$ matrix for $j = a, b$.

Define the distance between a point $y \in R^s$ and a set $\Lambda \subset R^s$ by

$$(4.2) \quad \text{dist}(y, \Lambda) := \inf_{\lambda \in \Lambda} \|y - \lambda\|.$$

We say that a sequence of sets $\{\Phi_T \subset R^s: T \geq 1\}$ is *locally approximated* (at the origin) by a cone $\Lambda \subset R^s$ if

$$(4.3) \quad \begin{aligned} \text{dist}(\phi_T, \Lambda) &= o(\|\phi_T\|) && \forall \{\phi_T \in \Phi_T: T \geq 1\} \text{ such that } \|\phi_T\| \rightarrow 0 \\ \text{and} &&& \\ \text{dist}(\lambda_T, \Phi_T) &= o(\|\lambda_T\|) && \forall \{\lambda_T \in \Lambda: T \geq 1\} \text{ such that } \|\lambda_T\| \rightarrow 0. \end{aligned}$$

This definition extends a definition of Chernoff (1954), who considers the local approximation of a single set by a cone. The extension is necessary to cover cases where the normalization matrix B_T is *not* of the form $\omega_T M$ for $\omega_T \in R$. Thus, the extension is necessary to cover cases where some variables possess deterministic and/or stochastic trends. We note that condition (4.3) is the same as requiring that the Hausdorff distance between $\Phi_T \cap S(\mathbf{0}, \varepsilon_T)$ and $\Lambda \cap S(\mathbf{0}, \varepsilon_T)$ goes to zero at a faster rate than ε_T , where $\varepsilon_T \rightarrow 0$ as $T \rightarrow \infty$.

ASSUMPTION 5: *For some sequence of scalar constants $\{b_T: T \geq 1\}$ for which $b_T \rightarrow \infty$ and $b_T \leq c \lambda_{\min}(B_T)$ for some $0 < c < \infty$, $\{B_T(\Theta - \theta_0)/b_T: T \geq 1\}$ is locally approximated by a cone Λ .*

LEMMA 2: *Suppose Assumptions 3 and 5 hold. Then, $\inf_{\lambda \in B_T(\Theta - \theta_0)} q_T(\lambda) = \inf_{\lambda \in \Lambda} q_T(\lambda) + o_p(1)$.*

COMMENTS: 1. Assumption 5 holds with $\Lambda = R^s$ if Θ contains a neighborhood of θ_0 , which is the standard case considered in the literature, provided $\lambda_{\min}(B_T) \rightarrow \infty$. This follows because $(B_T(\Theta - \theta_0)/\lambda_{\min}(B_T)) \cap S(0, \varepsilon) = S(0, \varepsilon) = \Lambda \cap S(0, \varepsilon)$ for some $\varepsilon > 0$.

2. Theorem 2(f) and Lemma 2 give

$$(4.4) \quad \ell_T(\hat{\theta}) = \ell_T(\theta_0) + \frac{1}{2}Z_T' \mathcal{F}_T Z_T - \frac{1}{2} \inf_{\lambda \in \Lambda} q_T(\lambda) + o_p(1).$$

4.2. Sufficient Conditions for Assumption 5

We now give two easily verifiable sufficient conditions for Assumption 5. Andrews (1997a) provides alternative sufficient conditions. We specify the conditions in terms of the parameter space Θ shifted to be centered at the origin rather than at θ_0 , i.e., in terms of $\Theta - \theta_0$. We say that a set $\Gamma \subset R^s$ is locally equal to a set $\Lambda \subset R^s$ if $\Gamma \cap C(\mathbf{0}, \varepsilon) = \Lambda \cap C(\mathbf{0}, \varepsilon)$ for some $\varepsilon > 0$.

ASSUMPTION 5*: (a) $\Theta - \theta_0$ is locally equal to a cone $\Lambda \subset R^s$.

(b) $B_T = b_T I_s$ for some scalar constants $\{b_T: T \geq 1\}$ for which $b_T \rightarrow \infty$.

Assumption 5*(a) covers many cases of interest. For example, it covers the common case where for some $\varepsilon > 0$

$$(4.5) \quad \Theta \cap C(\theta_0, \varepsilon) = \left\{ \theta \in R^s: \theta - \theta_0 \in \prod_{j=1}^s I_j, \theta \in C(\theta_0, \varepsilon) \right\} \quad \text{and}$$

$$\Lambda := \prod_{j=1}^s I_j, \quad \text{where } I_j = \{0\}, R, R^+, \text{ or } R^- \text{ for } j \leq s.$$

Here, $R^+ := \{x \in R: x \geq 0\}$ and $R^- := \{x \in R: x \leq 0\}$. Assumption 5* also allows for parameter spaces $\Theta - \theta_0$ that are defined by multivariate equality and/or inequality constraints. For example, one could have

$$(4.6) \quad \Theta := \{ \theta \in R^s: \Gamma_a \theta = r_1, \Gamma_b \theta \leq r_2, \|\theta\| \leq c < \infty \},$$

$\Gamma_a \theta_0 = r_1$, and $\Gamma_b \theta_0 \leq r_2$ with equality for zero or more elements of r_2 , where Γ_j is an $\ell_j \times s$ matrix, r_j is an ℓ_j -vector, and $0 \leq \ell_j \leq s$ for $j = a, b$. In this example,

$$(4.7) \quad \Lambda := \{ \lambda \in R^s: \Gamma_a \lambda = \mathbf{0}, \Gamma_{b1} \lambda \leq \mathbf{0} \},$$

where Γ_{b1} denotes the submatrix of Γ_b that consists of the rows of Γ_b for which $\Gamma_b \theta_0 \leq r_2$ holds as an equality. In most cases where Assumption 5* is applicable, $B_T = T^{1/2} I_s$.

Assumption 5* is not applicable in dynamic models with deterministic and/or stochastic trends, such as in the Dickey-Fuller Regression Example 2, because $B_T \neq b_T I_s$ in these models. Assumption 5* also is not applicable in the GARCH(1, q^*) Example of Andrews (1997a) for which $B_T = T^{1/2} M$ with M nondiagonal. For such cases, we introduce a more general sufficient condition for Assumption 5.

A cone is uniquely determined by the elements of the unit sphere that it contains. The maximal distance between two cones can be defined as the

maximal distance between the subsets of the unit sphere that correspond to the two cones. That is, for two cones A_1 and A_2 , we define

$$(4.8) \quad \text{dist}_c(A_1, A_2) := \max \left\{ \sup_{\lambda_1 \in A_1} \inf_{\lambda_2 \in A_2} \|\lambda_1 / \|\lambda_1\| - \lambda_2 / \|\lambda_2\|\|, \right. \\ \left. \sup_{\lambda_2 \in A_2} \inf_{\lambda_1 \in A_1} \|\lambda_1 / \|\lambda_1\| - \lambda_2 / \|\lambda_2\|\| \right\}.$$

Note that $\text{dist}_c(A_1, A_2)$ is the Hausdorff distance between the subsets of the unit sphere contained in A_1 and A_2 .

- ASSUMPTION 5^{2*}: (a) $\Theta - \theta_0$ is locally equal to a cone $\Lambda^* \subset R^s$.
 (b) $B_T = T_T M$, where T_T is diagonal, $\lambda_{\min}(T_T) \rightarrow \infty$, and M is nonsingular.
 (c) For some cone $\Lambda \subset R^s$, $\text{dist}_c(T_T M \Lambda^*, \Lambda) \rightarrow 0$.

For example, Assumption 5^{2*}(a) holds with Θ defined via equality and/or inequality constraints, as in (4.6). The verification of part (c) of Assumption 5^{2*} is typically straightforward, though it can be somewhat tedious.

LEMMA 3: Each of Assumptions 5* and 5^{2*} is sufficient for Assumption 5.

COMMENT: Assumptions 5* and 5^{2*} do not allow for any curvature in the boundary of Θ near θ_0 . See Andrews (1997a) for sufficient conditions for Assumption 5 that allow for curvature.

4.3. Examples (Continued)

4.3.1. Random Coefficient Regression

Assumptions 5* holds in Example 1 with

$$(4.9) \quad \Lambda := (R^+)^p \times R^{s-p}.$$

4.3.2. Dickey-Fuller Regression

We verify Assumption 5 in this example using Assumption 5^{2*}. Assumption 5^{2*} holds because $\Theta - \theta_0$ is locally equal to the cone

$$(4.10) \quad \Lambda^* = \{\lambda^* \in R^s : \lambda^* = (\lambda_1^*, \lambda_2^*, \lambda_3^*, \lambda_4^{*'})', \lambda_1^* \leq 0, \lambda_2^* \geq 0, \lambda_3^* \in R, \lambda_4^* \in R^b\}.$$

Assumption 5^{2*}(b) holds because $B_T = T_T M$. Assumption 5^{2*}(c) requires $\text{dist}_c(T_T M \Lambda^*, \Lambda) \rightarrow 0$ for some cone Λ . In the present case, we have

$$B_T := T_T M := \begin{pmatrix} T & 0 & 0 & \mathbf{0}' \\ T^{3/2}\mu_0 & T^{3/2} & 0 & \mathbf{0}' \\ -T^{1/2}\mu_0 & 0 & T^{1/2} & T^{1/2}\mu_0 \mathbf{1}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & T^{1/2}I_b \end{pmatrix} \quad \text{and}$$

$$(4.11) \quad \begin{aligned} T_T M \Lambda^* &= \{ \lambda \in R^s : \lambda_1 = T\lambda_1^*, \lambda_2 = T^{3/2}\mu_0 \lambda_1^* + T^{3/2}\lambda_2^*, \\ &\lambda_3 = -T^{1/2}\mu_0 \lambda_1^* + T^{1/2}\lambda_3^* + T^{1/2}\mu_0 \mathbf{1}'\lambda_4^*, \text{ and} \\ &\lambda_4 = T^{1/2}\lambda_4^* \text{ for } \lambda^* \in \Lambda^* \} \\ &= \{ \lambda \in R^s : \lambda_1 \leq 0, \lambda_2 \geq T^{1/2}\mu_0 \lambda_1, \lambda_3 \in R, \lambda_4 \in R^b \}. \end{aligned}$$

From (4.11), Λ depends on $\mu_0 = \theta_{30}/(1 - \mathbf{1}'\theta_{40})$. That is, it depends on the value of the drift parameter θ_{30} of the unit root process. If $\theta_{30} = 0$, then

$$(4.12) \quad \Lambda := B_T \Lambda^* = \{ \lambda \in R^s : \lambda_1 \leq 0, \lambda_2 \geq 0, \lambda_3 \in R, \lambda_4 \in R^b \}.$$

If $\theta_{30} > 0$, then $\text{dist}_c(T_T M \Lambda^*, \Lambda) \rightarrow 0$ for

$$(4.13) \quad \Lambda := \{ \lambda \in R^s : \lambda_1 \leq 0, \lambda_2 \in R, \lambda_3 \in R, \lambda_4 \in R^b \}.$$

In consequence, when the true unit root process has positive drift, the limit distribution of $B_T(\hat{\theta} - \theta_0)$ is the same whether or not the time trend parameter is restricted by Θ to be nonnegative or not.

5. ASYMPTOTIC DISTRIBUTION OF THE EXTREMUM ESTIMATOR

5.1. Asymptotic Distribution

In this section, we determine the asymptotic distribution of $B_T(\hat{\theta} - \theta_0)$.

We show that $B_T(\hat{\theta} - \theta_0)$ is asymptotically equivalent to $\hat{\lambda}_T$ provided Λ is convex. By definition, $\hat{\lambda}_T \in \text{cl}(\Lambda)$ and

$$(5.1) \quad q_T(\hat{\lambda}_T) = \inf_{\lambda \in \Lambda} q_T(\lambda).$$

The random variable $\hat{\lambda}_T$ is a version of the projection of Z_T onto the cone Λ with respect to the norm $\|\lambda\|_T = (\lambda' \mathcal{F}_T \lambda)^{1/2}$; see Perlman (1969, Sec. 4). If Λ is convex, $\hat{\lambda}_T$ is uniquely defined. Whether or not Λ is convex, the following orthogonality property holds: $\hat{\lambda}_T' \mathcal{F}_T (\hat{\lambda}_T - Z_T) = 0$; see Perlman (1969, Lem. 4.1).

For example, if Λ is a linear subspace of R^s , as occurs with linear or nonlinear equality constraints, then $\hat{\lambda}_T$ is a linear function of Z_T : $\hat{\lambda}_T = P_{T\Lambda} Z_T$, where $P_{T\Lambda}$ is the projection matrix onto Λ with respect to the norm $\|\cdot\|_T$. For instance, if $\Lambda := \{ \lambda \in R^s : \Gamma \lambda = \mathbf{0} \}$, where Γ is full row rank, then $P_{T\Lambda} := I_s - \mathcal{F}_T^{-1} \Gamma' (\Gamma \mathcal{F}_T^{-1} \Gamma')^{-1} \Gamma$. (We note that for most of our examples, Λ is not a linear subspace.)

ASSUMPTION 6: Λ is convex.

Assumption 6 holds for all examples in this paper and Andrews (1997a).

The asymptotic distribution of $\hat{\lambda}_T$ and, hence, of $B_T(\hat{\theta} - \theta_0)$ is given by that of $\hat{\lambda}$. By definition, $\hat{\lambda} \in \text{cl}(\Lambda)$ and

$$(5.2) \quad q(\hat{\lambda}) = \inf_{\lambda \in \Lambda} q(\lambda),$$

where $q(\lambda) := (\lambda - Z)' \mathcal{F}(\lambda - Z)$ and $Z := \mathcal{F}^{-1}G$.

As with $\hat{\lambda}_T$, $\hat{\lambda}$ is not necessarily uniquely defined. It is unique, however, under Assumption 6.

The asymptotic distribution of $B_T(\hat{\theta} - \theta_0)$ is given in the following theorem.

THEOREM 3: (a) Suppose Assumptions 2–6 hold. Then, $B_T(\hat{\theta} - \theta_0) = \hat{\lambda}_T + o_p(1)$.
 (b) Suppose Assumptions 2–6 hold. Then, $\hat{\lambda}_T \rightarrow_d \hat{\lambda}$ and $B_T(\hat{\theta} - \theta_0) \rightarrow_d \hat{\lambda}$.
 (c) Suppose Assumptions 2–5 hold. Then,

$$\ell_T(\hat{\theta}) - \ell_T(\theta_0) \rightarrow_d \frac{1}{2} \left(Z' \mathcal{F} Z - \inf_{\lambda \in \Lambda} q(\lambda) \right) = \frac{1}{2} \hat{\lambda}' \mathcal{F} \hat{\lambda}.$$

COMMENTS: 1. In the classical case in which θ_0 is not on a boundary, $\Lambda = R^s$ and $\hat{\lambda} = \mathcal{F}^{-1}G$. Thus, if G is Gaussian and \mathcal{F} is nonrandom (as typically occurs in models without stochastic trends), then $B_T(\hat{\theta} - \theta_0)$ has a Gaussian distribution. The case of primary interest in this paper is when θ_0 is on a boundary and $\Lambda \neq R^s$. In this case, the distribution of $\hat{\lambda}$ is more complex. Section 6 analyzes its distribution in detail.

2. The proof of Theorem 3(a) is easy if $\Lambda = R^s$, which is the standard case considered in the literature and which corresponds to the case where θ_0 is not on a boundary. The proof is as follows. By Theorem 2(e) and Lemma 2, $q_T(B_T(\hat{\theta} - \theta_0)) = \inf_{\lambda \in \Lambda} q_T(\lambda) + o_p(1)$. If $\Lambda = R^s$, then $\hat{\lambda}_T = Z_T$, $\inf_{\lambda \in \Lambda} q_T(\lambda) = 0$, and

$$(5.3) \quad q_T(B_T(\hat{\theta} - \theta_0)) = (B_T(\hat{\theta} - \theta_0) - \hat{\lambda}_T)' \mathcal{F}_T (B_T(\hat{\theta} - \theta_0) - \hat{\lambda}_T) = o_p(1).$$

In view of Assumption 3, this gives the result of Theorem 3(a). When $\Lambda \neq R^s$, the proof of Theorem 3(a) is more complicated.

3. Theorem 3(a) still holds when Assumption 3 is replaced by $B_T^{-1} D \ell_T(\theta_0) = O_p(1)$, $\lambda_{\max}(\mathcal{F}_T) = O_p(1)$, $\lambda_{\min}^{-1}(\mathcal{F}_T) = O_p(1)$, and \mathcal{F}_T is symmetric wp $\rightarrow 1$.

4. The result of Theorem 3(c) can be used to obtain the asymptotic distribution of a quasi-likelihood ratio statistic, as is done in Andrews (1998b).

5.2. Examples (Continued)

5.2.1. Random Coefficient Regression

Assumption 6 holds in this example by (4.9). By Theorem 3, $T^{1/2}(\hat{\theta} - \theta_0) \rightarrow_d \hat{\lambda}$, where $\hat{\lambda}$ satisfies (5.2) with (G, \mathcal{F}) defined in (3.15) and (3.11) and Λ defined in (4.9).

5.2.2. Dickey-Fuller Regression

In this example, Assumption 6 holds for all values of the drift parameter θ_{30} by (4.12) and (4.13). By Theorem 3, $B_T(\hat{\theta} - \theta_0) \rightarrow_d \hat{\lambda}$, where $\hat{\lambda}$ satisfies (5.2) with (G, \mathcal{I}) defined in (3.21) and with A defined in (4.12) or (4.13) depending on the value of θ_{30} .

6. ASYMPTOTIC DISTRIBUTIONS OF SUBVECTORS OF THE EXTREMUM ESTIMATOR

6.1. A Partitioning of θ into (β, δ, ψ)

In this section, we simplify the asymptotic distribution of $B_T(\hat{\theta} - \theta_0)$ by partitioning θ into three subvectors and providing separate expressions for each of the three corresponding subvectors of $\hat{\lambda}$. We partition θ as follows:

$$(6.1) \quad \theta = (\theta'_*, \psi')' = (\beta', \delta', \psi')' \quad \text{and} \quad \theta_* = (\beta', \delta')',$$

where $\beta \in R^p, \delta \in R^q, \psi \in R^r, 0 \leq p, q, r \leq s$, and $p + q + r = s$.

Below we assume that the asymptotic “quasi-information matrix” \mathcal{I} is block diagonal between θ_* and ψ . We also assume that δ_0 is a parameter that is not on a boundary (where $\theta_0 = (\beta'_0, \delta'_0, \psi'_0)'$). These features characterize the subvectors β, δ , and ψ . The results given below cover cases where no parameters δ and/or ψ appear simply by setting q and/or r equal to 0.

We partition $\hat{\theta}, \theta_0, B_T, G, \mathcal{I}, Z, \hat{\lambda}_T, \hat{\lambda}$, and $D\mathcal{L}_T(\theta_0)$ conformably with θ . Let

$$(6.2) \quad \hat{\theta} = \begin{pmatrix} \hat{\theta}_* \\ \hat{\psi} \end{pmatrix} = \begin{pmatrix} \hat{\beta} \\ \hat{\delta} \\ \hat{\psi} \end{pmatrix}, \quad \theta_0 = \begin{pmatrix} \theta_{*0} \\ \psi_0 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \delta_0 \\ \psi_0 \end{pmatrix},$$

$$B_T = \begin{bmatrix} B_{*T} & B_{*\psi T} \\ B_{\psi * T} & B_{\psi T} \end{bmatrix} = \begin{bmatrix} B_{\beta T} & B_{\beta\delta T} & B_{\beta\psi T} \\ B_{\delta\beta T} & B_{\delta T} & B_{\delta\psi T} \\ B_{\psi\beta T} & B_{\psi\delta T} & B_{\psi T} \end{bmatrix},$$

$$G = \begin{pmatrix} G_* \\ G_\psi \end{pmatrix} = \begin{pmatrix} G_\beta \\ G_\delta \\ G_\psi \end{pmatrix}, \quad \mathcal{I} = \begin{bmatrix} \mathcal{I}_* & \mathcal{I}_{*\psi} \\ \mathcal{I}_{\psi*} & \mathcal{I}_\psi \end{bmatrix} = \begin{bmatrix} \mathcal{I}_\beta & \mathcal{I}_{\beta\delta} & \mathcal{I}_{\beta\psi} \\ \mathcal{I}_{\delta\beta} & \mathcal{I}_\delta & \mathcal{I}_{\delta\psi} \\ \mathcal{I}_{\psi\beta} & \mathcal{I}_{\psi\delta} & \mathcal{I}_\psi \end{bmatrix},$$

$$Z = \begin{pmatrix} Z_* \\ Z_\psi \end{pmatrix} = \begin{pmatrix} Z_\beta \\ Z_\delta \\ Z_\psi \end{pmatrix}, \quad \hat{\lambda}_T = \begin{pmatrix} \hat{\lambda}_{*T} \\ \hat{\lambda}_{\psi T} \end{pmatrix} = \begin{pmatrix} \hat{\lambda}_{\beta T} \\ \hat{\lambda}_{\delta T} \\ \hat{\lambda}_{\psi T} \end{pmatrix},$$

$$\hat{\lambda} = \begin{pmatrix} \hat{\lambda}_* \\ \hat{\lambda}_\psi \end{pmatrix} = \begin{pmatrix} \hat{\lambda}_\beta \\ \hat{\lambda}_\delta \\ \hat{\lambda}_\psi \end{pmatrix}, \quad \text{and}$$

$$D\ell_T(\theta_0) = \begin{pmatrix} D_* \ell_T(\theta_0) \\ D_\psi \ell_T(\theta_0) \end{pmatrix} = \begin{pmatrix} D_\beta \ell_T(\theta_0) \\ D_\delta \ell_T(\theta_0) \\ D_\psi \ell_T(\theta_0) \end{pmatrix}.$$

The defining features of the parameters ψ and δ , respectively, are the following:

ASSUMPTION 7: (a) \mathcal{F} is block diagonal between θ_* and ψ . That is, $\mathcal{F}_{*\psi} = \mathcal{F}'_{\psi*} = \mathbf{0}$. (b) The cone Λ of Assumption 5 is a product set $\Lambda_\beta \times \Lambda_\delta \times \Lambda_\psi$, where $\Lambda_\beta \subset R^p$, $\Lambda_\delta \subset R^q$, and $\Lambda_\psi \subset R^r$ are cones.

ASSUMPTION 8: $\Lambda_\delta = R^q$.

Assumptions 7 and 8 require that the asymptotic information matrix is block diagonal between θ_* and ψ and that δ_0 is not on a boundary respectively. We note that Assumption 7(a) often can be made to hold by reparameterization when \mathcal{F} is nonrandom. Suppose we start with a parameter $\tilde{\theta} = (\theta'_*, \eta')$, where $\theta_* \in R^{p+q}$ and $\eta \in R^r$, and θ_* is the parameter of interest. Suppose that Assumptions 2 and 3 hold with \mathcal{F} equal to

$$\begin{bmatrix} \mathcal{F}_* & \mathcal{F}_{\eta*} \\ \mathcal{F}_{*\eta} & \mathcal{F}_\eta \end{bmatrix}.$$

Let $\delta = \eta - \mathcal{F}_{\eta*} \mathcal{F}_*^{-1} \theta_*$ and $\theta = (\theta'_*, \delta')$. Then, Assumptions 2, 3, and 7(a) typically hold for the parameter θ with \mathcal{F} equal to $\text{Diag}(\mathcal{F}_*, \mathcal{F}_\eta - \mathcal{F}_{\eta*} \mathcal{F}_*^{-1} \mathcal{F}_{*\eta})$.

Under Assumption 7,

$$\begin{aligned} Z_* &= \mathcal{F}_*^{-1} G_*, & Z_\psi &= \mathcal{F}_\psi^{-1} G_\psi, & \text{and} \\ (6.3) \quad Z_\beta &= HZ_* = \mathcal{F}_\beta^{-1} G_\beta + \mathcal{F}_\beta^{-1} \mathcal{F}_{\beta\delta} (\mathcal{F}_\delta - \mathcal{F}_{\delta\beta} \mathcal{F}_\beta^{-1} \mathcal{F}_{\beta\delta})^{-1} (\mathcal{F}_{\delta\beta} \mathcal{F}_\beta^{-1} G_\beta - G_\delta), \\ &\text{where } H := [I_p : \mathbf{0}] \in R^{p \times (p+q)}. \end{aligned}$$

Define

$$\begin{aligned} (6.4) \quad q_\beta(\lambda_\beta) &:= (\lambda_\beta - Z_\beta)' (H \mathcal{F}_*^{-1} H')^{-1} (\lambda_\beta - Z_\beta) & \text{and} \\ q_\psi(\lambda_\psi) &:= (\lambda_\psi - Z_\psi)' \mathcal{F}_\psi (\lambda_\psi - Z_\psi). \end{aligned}$$

Given Assumptions 7 and 8, we can split the terms of the quadratic approximation to $\ell_T(\hat{\theta})$, and, in consequence, $\hat{\lambda}$ into separate terms involving β , δ ,

and ψ :

THEOREM 4: *Suppose Assumptions 3, 7, and 8 hold. Then,*

- (a) $q_\beta(\hat{\lambda}_\beta) = \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta)$,
- (b) $\hat{\lambda}_\delta = \mathcal{F}_\delta^{-1} G_\delta - \mathcal{F}_\delta^{-1} \mathcal{F}_{\delta\beta} \hat{\lambda}_\beta$,
- (c) $q_\psi(\hat{\lambda}_\psi) = \inf_{\lambda_\psi \in \Lambda_\psi} q_\psi(\lambda_\psi)$,
- (d) $Z'Z = Z'_\beta(H\mathcal{T}_*^{-1}H')^{-1}Z_\beta + G'_\delta\mathcal{F}_\delta^{-1}G_\delta + Z'_\psi\mathcal{F}_\psi Z_\psi$,
- (e) $\inf_{\lambda \in \Lambda} q(\lambda) = \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta) + \inf_{\lambda_\psi \in \Lambda_\psi} q_\psi(\lambda_\psi)$, and
- (f) $Z'Z - \inf_{\lambda \in \Lambda} q(\lambda) = Z'_\beta(H\mathcal{T}_*^{-1}H')^{-1}Z_\beta - \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta) + G'_\delta\mathcal{F}_\delta^{-1}G_\delta + Z'_\psi\mathcal{F}_\psi Z_\psi - \inf_{\lambda_\psi \in \Lambda_\psi} q_\psi(\lambda_\psi) = \hat{\lambda}'_\beta(H\mathcal{T}_*^{-1}H')^{-1}\hat{\lambda}_\beta + G'_\delta\mathcal{F}_\delta^{-1}G_\delta + \hat{\lambda}'_\psi\mathcal{F}_\psi\hat{\lambda}_\psi$.

COMMENTS: 1. If $\Lambda_\beta = R^p$, which holds if β_0 is not on a boundary, then $\inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta) = 0$ and $\hat{\lambda}_\beta = Z_\beta$. Similarly, if $\Lambda_\psi = R^r$, then $\inf_{\lambda_\psi \in \Lambda_\psi} q_\psi(\lambda_\psi) = 0$ and $\hat{\lambda}_\psi = Z_\psi = \mathcal{F}_\psi^{-1}G_\psi$. Our interest here is in cases where one or the other or both of these simplifications does not hold.

2. If Λ_β is a linear subspace of R^p , which holds in the case of linear or nonlinear equality constraints as considered by Aitchison and Silvey (1958), then $\hat{\lambda}_\beta = P_{\Lambda_\beta}Z_\beta$, where P_{Λ_β} is the projection matrix onto Λ_β with respect to the norm $\|\lambda_\beta\|_\beta^2 := \lambda'_\beta(H\mathcal{T}_*^{-1}H')^{-1}\lambda_\beta$. For example, if $\Lambda_\beta = \{\lambda_\beta \in R^p: \Gamma_a\lambda_\beta = \mathbf{0}\}$, then $P_{\Lambda_\beta} := I_p - H\mathcal{T}_*^{-1}H'\Gamma'_a(\Gamma_aH\mathcal{T}_*^{-1}H'\Gamma'_a)^{-1}\Gamma_a$.

Theorems 3 and 4 combine to give the following corollary:

COROLLARY 1: (a) *Suppose Assumptions 2–8 hold. Then,*

$$\begin{aligned}
 B_{\beta T}(\hat{\beta} - \beta_0) + B_{\beta\delta T}(\hat{\delta} - \delta_0) + B_{\beta\psi T}(\hat{\psi} - \psi_0) &\xrightarrow{d} \hat{\lambda}_\beta, \\
 \text{where } \hat{\lambda}_\beta \text{ solves } q_\beta(\hat{\lambda}_\beta) &= \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta), \\
 B_{\delta\beta T}(\hat{\beta} - \beta_0) + B_{\delta T}(\hat{\delta} - \delta_0) + B_{\delta\psi T}(\hat{\psi} - \psi_0) &\xrightarrow{d} \mathcal{F}_\delta^{-1}G_\delta - \mathcal{F}_\delta^{-1}\mathcal{F}_{\delta\beta}\hat{\lambda}_\beta, \\
 B_{\psi\beta T}(\hat{\beta} - \beta_0) + B_{\psi\delta T}(\hat{\delta} - \delta_0) + B_{\psi T}(\hat{\psi} - \psi_0) &\xrightarrow{d} \hat{\lambda}_\psi, \\
 \text{where } \hat{\lambda}_\psi \text{ solves } q_\psi(\hat{\lambda}_\psi) &= \inf_{\lambda_\psi \in \Lambda_\psi} q_\psi(\lambda_\psi),
 \end{aligned}$$

and the convergence of these three terms holds jointly.

(b) *Suppose Assumptions 2–8 hold. Then,*

$$\begin{aligned}
 B_{\beta T}(\hat{\beta} - \beta_0) &\xrightarrow{d} \hat{\lambda}_\beta \quad \text{provided } B_{\beta\delta T} = \mathbf{0} \text{ and } B_{\beta\psi T} = \mathbf{0}, \\
 B_{\delta T}(\hat{\delta} - \delta_0) &\xrightarrow{d} \mathcal{F}_\delta^{-1}G_\delta - \mathcal{F}_\delta^{-1}\mathcal{F}_{\delta\beta}\hat{\lambda}_\beta \quad \text{provided } B_{\delta\beta T} = \mathbf{0} \text{ and } B_{\delta\psi T} = \mathbf{0}, \\
 B_{\psi T}(\hat{\psi} - \psi_0) &\xrightarrow{d} \hat{\lambda}_\psi \quad \text{provided } B_{\psi\beta T} = \mathbf{0} \text{ and } B_{\psi\delta T} = \mathbf{0},
 \end{aligned}$$

and the convergence holds jointly, where $\hat{\lambda}_\beta$ and $\hat{\lambda}_\psi$ are as in part (a).

(c) Suppose Assumptions 2–5, 7, and 8 hold. Then,

$$\begin{aligned} \ell_T(\hat{\theta}) - \ell_T(\theta_0) &\xrightarrow{d} \frac{1}{2} \left(Z'_\beta (H\mathcal{T}_*^{-1}H')^{-1} Z_\beta - \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta) \right) \\ &\quad + \frac{1}{2} G'_\delta \mathcal{T}_\delta^{-1} G_\delta + \frac{1}{2} \left(Z'_\psi \mathcal{T}_\psi Z_\psi - \inf_{\lambda_\psi \in \Lambda_\psi} q_\psi(\lambda_\psi) \right) \\ &= \frac{1}{2} \left(\hat{\lambda}'_\beta (H\mathcal{T}_*^{-1}H')^{-1} \hat{\lambda}_\beta + G'_\delta \mathcal{T}_\delta^{-1} G_\delta + \hat{\lambda}'_\psi \mathcal{T}_\psi \hat{\lambda}_\psi \right). \end{aligned}$$

COMMENTS: 1. Each of the three results of Corollary 1(b) is applicable in the examples of this paper and Andrews (1997a) except in the Dickey-Fuller Regression Example considered in this paper and the GARCH(1, q^*) Example of Andrews (1997a). In these two examples, only the first and third results of Corollary 1(b) are applicable.

2. Corollary 1(b) shows that the asymptotic distributions of $\hat{\beta}$ and $\hat{\delta}$ do not depend on whether ψ_0 is on a boundary. Similarly, the asymptotic distribution of $\hat{\psi}$ does not depend on whether β_0 is on a boundary. For example, in the Random Coefficients Regression Example, the Gaussian QML estimator of the regression slope coefficients does not depend on whether the variances of the random coefficients are positive or zero.

3. Corollary 1(b) shows that the asymptotic distribution of $\hat{\delta}$ depends on whether β_0 is on a boundary if and only if $\mathcal{T}_{\delta\beta} \neq \mathbf{0}$. For example, in the Regression with Restricted Parameters Example of Andrews (1997a), where some slope coefficients are restricted, the asymptotic distribution of the LS estimator of slope coefficients that are unrestricted does not depend on whether the true restricted coefficients are on a boundary if and only if the asymptotic “information” matrix is block diagonal between the restricted and unrestricted slope coefficients.

4. Corollary 1 reduces the dimensionality of the minimization problem $\inf_{\lambda \in \Lambda} q(\lambda)$ by splitting it up into three separate minimization problems of lower dimensions, one of which is solved analytically. This facilitates the solution of the minimization problem whether one uses analytics or simulation.

6.2. LAN and LAMN Conditions for $(\hat{\beta}, \hat{\delta})$

We now concentrate on the asymptotic distributions of $\hat{\beta}$ and $\hat{\delta}$. The parameter ψ is considered to be a nuisance parameter. The following results for $\hat{\beta}$ and $\hat{\delta}$ can be applied to $\hat{\psi}$ by re-labeling ψ as $\theta_* = (\beta', \delta')$.

We specify three conditions that indicate the form that the limit random variables (G_*, \mathcal{T}_*) (which determine the asymptotic distributions of $\hat{\beta}$ and $\hat{\delta}$), take in typical cases. The first condition is applicable in models in which $B_{*T}^{-1'} D_* \ell_T(\theta_0)$ and \mathcal{T}_{*T} may depend on deterministic and stochastic trends, but none of the elements of $\theta_{*0} = (\beta'_0, \delta'_0)'$ are unit roots. This includes the Regression with Restricted Parameters and Integrated Regressors Example of

Andrews (1997a). It excludes the Dickey-Fuller Regression Example. (Note that ψ_0 may contain unit roots.) Models covered by the first condition are *locally asymptotically mixed normal* (LAMN) models (with respect to the parameters (β, δ)).

ASSUMPTION 3^{2*}: (a) *Assumption 3 holds.*
 (b) $G_* \sim N(\mu, \mathcal{I}_*)$ conditional on some σ -field \mathcal{F} , for some nonrandom $(p + q)$ -vector μ and some (possibly) random $(p + q) \times (p + q)$ -matrix \mathcal{I}_* that is \mathcal{F} measurable.

The second condition covers the *locally asymptotically normal* (LAN) case (again, with respect to the parameters (β, δ)). It is applicable in cross-sectional contexts and in time series contexts in which $B_{*T}^{-1'} D_* \ell_T(\theta_0)$ and \mathcal{F}_{*T} may depend on deterministic trends but not on stochastic trends.

ASSUMPTION 3^{3*}: (a) *Assumption 3 holds.*
 (b) $G_* \sim N(\mathbf{0}, \mathcal{I}_*)$ for some nonrandom $(p + q) \times (p + q)$ -matrix \mathcal{I}_* .
 (c) \mathcal{F}_* is nonrandom.

Next, we consider the case where \mathcal{I}_* of Assumption 3^{2*} or Assumption 3^{3*} is proportional to \mathcal{F}_* .

ASSUMPTION 3^{4*}: (a) *Assumption 3^{2*} holds.*
 (b) $\mathcal{I}_* = c\mathcal{F}_*$ for some scalar constant $c > 0$.

It is apparent that Assumption 3^{3*} \Rightarrow 3^{2*} \Rightarrow 3 and Assumption 3^{4*} \Rightarrow 3^{2*} \Rightarrow 3. If $\ell_T(\theta)$ is a correctly specified log-likelihood function and Assumption 3^{2*} holds, then the information matrix equality implies that Assumption 3^{4*} holds with $c = 1$. Assumption 3^{4*} holds for LS estimators of regression models with $c = \sigma^2$ provided Assumption 3^{4*} holds and the regression errors are homoskedastic conditional on the regressors with variance σ^2 .

6.3. A Closed Form Expression for $\hat{\lambda}_\beta$

We now consider an assumption on A_β under which we have a simple closed form expression for $\hat{\lambda}_\beta$ and, hence, for $\hat{\lambda}_\beta$ as well.

ASSUMPTION 9: $A_\beta = \{\lambda_\beta \in R^p: \Gamma_a \lambda_\beta = \mathbf{0}, \Gamma_b \lambda_\beta \leq \mathbf{0}\}$, where $\Gamma := [\Gamma'_a; \Gamma'_b]$ is a full row rank matrix.

Note that Assumption 9 allows for the case where Γ_a or Γ_b does not appear. Assumption 9 holds in all of the examples of this paper and Andrews (1997a). For A_β as in Assumption 9, $\hat{\lambda}_\beta$ is the solution to a quadratic programming (QP) problem with mixed linear equality and inequality constraints.

The following lemma provides a characterization of $\hat{\lambda}_\beta$ when Assumption 9 holds.

LEMMA 4: *Suppose that Assumptions 3 and 7–9 hold. Then, $\hat{\lambda}_\beta = P_L Z_\beta$ for some linear subspace L of the form $L := \{\ell \in R^p : \Gamma_a \ell = \mathbf{0}, \Gamma_{b1} \ell = \mathbf{0}\}$, where Γ_{b1} is comprised of some (possibly zero) rows of Γ_b and P_L is the projection matrix onto L with respect to the norm $\|\lambda_\beta\|_\beta^2 = \lambda'_\beta (H\mathcal{T}_*^{-1}H')^{-1}\lambda_\beta$. That is,*

$$P_L = I_p - H\mathcal{T}_*^{-1}H'\Gamma'_1(\Gamma_1 H\mathcal{T}_*^{-1}H'\Gamma'_1)^{-1}\Gamma_1,$$

where $\Gamma_1 := \begin{bmatrix} \Gamma'_a \\ \Gamma'_{b1} \end{bmatrix}$.

COMMENTS: 1. The number of different linear subspaces of the form L is 2^{p_b} , where p_b is the number of inequality constraints in Λ_β , i.e., the number of rows of Γ_b .

2. Lemma 4 still holds if Γ is not full row rank provided one replaces Γ_1 in the definition of P_L with a matrix that equals Γ_1 but has any redundant rows deleted.

Lemma 4 yields the following closed form expression for $\hat{\lambda}_\beta$.

THEOREM 5: *Suppose that Assumptions 3 and 7–9 hold. Then:*

(a) $\hat{\lambda}_\beta = P_{L(\hat{j})} Z_\beta$, where \hat{j} minimizes $CF_j := Z'_\beta \Gamma'_j (\Gamma_j H\mathcal{T}_*^{-1}H'\Gamma'_j)^{-1} \Gamma_j Z_\beta$ over $j = 1, \dots, 2^{p_b}$ for which $P_{L(j)} Z_\beta \in \Lambda_\beta$. Here, $L(j) := \{\ell \in R^p : \Gamma_a \ell = \mathbf{0}, \Gamma_{bj} \ell = \mathbf{0}\}$, $\Gamma_j := \begin{bmatrix} \Gamma'_a \\ \Gamma'_{bj} \end{bmatrix}$, $P_{L(j)} = I_p - H\mathcal{T}_*^{-1}H'\Gamma'_j(\Gamma_j H\mathcal{T}_*^{-1}H'\Gamma'_j)^{-1}\Gamma_j$, and $\{\Gamma_{bj} : j = 1, \dots, 2^{p_b}\}$ consists of all the different matrices comprised of some (possibly zero) rows of Γ_b .

(b) $\hat{\lambda}_\beta = \sum_{j=1}^{2^{p_b}} P_{L(j)} Z_\beta \times 1(P_{L(j)} Z_\beta \in \Lambda_\beta) \times \prod_{k=1}^{2^{p_b}} 1(CF_j \leq CF_k \text{ or } P_{L(k)} Z_\beta \notin \Lambda_\beta)$.

(c) For any $p \times p$ (possibly random) matrix A that is symmetric and nonsingular with probability one, $\hat{\lambda}_\beta = AP_{L_A(\hat{j})} Z_{\beta A}$, where \hat{j} is as in part (a), $Z_{\beta A} := A^{-1} Z_\beta$, and $P_{L_A(j)} := I_p - A^{-1}H\mathcal{T}_*^{-1}H'\Gamma'_j(\Gamma_j H\mathcal{T}_*^{-1}H'\Gamma'_j)^{-1}\Gamma_j A$.

COMMENTS: 1. In part (a), \hat{j} indexes the constraints, $\Gamma_{\hat{j}}$, that are binding, given Z_β and \mathcal{T}_* . Given the constraints $\Gamma_{\hat{j}}$, $\hat{\lambda}_\beta$ is obtained simply by an oblique projection of Z_β onto the linear subspace, $L(\hat{j})$, defined by the constraints. Part (b) provides a closed form expression for $\hat{\lambda}_\beta$ based on the characterization of $\hat{\lambda}_\beta$ given in part (a).

2. Part (c) shows that $\hat{\lambda}_\beta$ can be expressed in terms of a vector $Z_{\beta A}$ rather than Z_β . One can choose A such that $Z_{\beta A}$ has fewer nuisance parameters than Z_β . If Z_β has a normal distribution, this is done by taking A to be the inverse of the square root of the covariance matrix of Z_β . By expressing $\hat{\lambda}_\beta$ in this way, one can minimize the number of nuisance parameters on which λ_β depends.

As an example of Theorem 5, suppose $\Lambda_\beta = R^+ \times R^{p-1}$. Then,

$$\hat{\lambda}_\beta = \begin{cases} AZ_{\beta A} & \text{if } Z_{\beta A1} \geq 0, \\ A(0, Z_{\beta A2} - \rho_{12}Z_{\beta A1}, \dots, Z_{\beta Ap} - \rho_{p1}Z_{\beta A1})' & \text{otherwise,} \end{cases}$$

(6.5) where
 $A = \text{Diag}^{1/2}(H\mathcal{T}_*^{-1}H')$, $Z_{\beta A} := (Z'_{\beta A1}, \dots, Z'_{\beta Ap})' := A^{-1}Z_\beta$,
 and
 $\rho_{ij} = [A^{-1}H\mathcal{T}_*^{-1}H'A^{-1}]_{ij}$, for $i, j = 1, \dots, p$.

Our choice of A minimizes the number of nuisance parameters. The formula for $\hat{\lambda}_\beta$ also is valid for any positive definite diagonal matrix $A > 0$. When $\Lambda_\beta = R^- \times R^{p-1}$, the inequality in (6.5) is reversed.

Results of Lovell and Prescott (1970, Sec. 4) for the normal linear regression model imply that the mean squared error of each element of $\hat{\lambda}_\beta$ as an estimator of 0 is less than or equal to the mean squared error of each corresponding element of $Z_{\beta A}$ when $\Lambda_\beta = R^+ \times R^{p-1}$. This implies that the conventional asymptotic standard errors that are based on the assumption that no parameters are on a boundary are conservative estimators (i.e., estimators whose probability limits are greater than or equal to the true asymptotic standard errors) when one element of β is on a boundary and Assumption 3^{3*} holds (or Assumption 3^{2*} holds with $\mu = 0$ and $E\mathcal{I}_* < \infty$). It also implies that the estimator $\hat{\theta}$ has a smaller mean squared error of its asymptotic distribution than does an unrestricted version of the estimator that is based on a parameter space that contains a full neighborhood of θ_0 .

Rothenberg (1973, p. 57) conjectures that Lovell and Prescott's (1970) result for the normal linear regression model with one parameter on a boundary extends to the general case where the parameter is on the boundary of a convex set. We agree that this is probably true, but we do not have a proof. If true, then the conventional asymptotic standard errors that are based on the assumption that no parameters are on a boundary are conservative estimators whenever Assumptions 3^{3*} and 6 hold (or Assumptions 3^{2*} and 6 hold with $\mu = 0$ and $E\mathcal{I}_* < \infty$), which covers the vast majority of cases in the literature.

As a second example, suppose $\Lambda_\beta = (R^+)^2 \times R^{p-2}$. Then,

$$\hat{\lambda}_\beta = AP_{L_A(\hat{\beta})}Z_{\beta A}, \quad \text{where}$$

$$P_{L_A(\hat{\beta})}Z_{\beta A} := 1(Z_{\beta A1} > 0, Z_{\beta A2} > 0)Z_{\beta A} + 1(Z_{\beta A1} - \rho_{21}Z_{\beta A2} > 0, Z_{\beta A2} \leq 0) \times (Z_{\beta A1} - \rho_{21}Z_{\beta A2}, 0, Z_{\beta A3} - \rho_{23}Z_{\beta A2}, \dots, Z_{\beta Ap} - \rho_{2p}Z_{\beta A2})' + 1(Z_{\beta A1} \leq 0, Z_{\beta A2} - \rho_{12}Z_{\beta A1} > 0) \times (0, Z_{\beta A2} - \rho_{12}Z_{\beta A1}, \dots, Z_{\beta Ap} - \rho_{1p}Z_{\beta A1})'$$

(6.6)

where A and ρ_{ij} are as in (6.5). The formula for $\hat{\lambda}_\beta$ also is valid for any positive definite diagonal matrix A . For the case where $\Lambda_\beta = R^- \times R^+ \times R^{p-2}$ (as occurs in the Dickey-Fuller Regression Example with $p = 2$), (6.6) holds but with the first of the two inequalities reversed in each of the indicator functions in the definition of $P_{L_A(\hat{j})}Z_{\beta A}$. Adjustments of (6.6) for the cases where $\Lambda_\beta = R^+ \times R^- \times R^{p-2}$ and $\Lambda_\beta = (R^-)^2 \times R^{p-2}$ are analogous.

For the case where Λ_β is of the form $\Lambda_\beta = \{\lambda_\beta \in R^p: \lambda_{\beta 1} \geq 0, \Gamma_a \lambda_\beta = \mathbf{0}\}$, $\hat{\lambda}_\beta$ is as defined in (6.5), but with $Z_{\beta A}$ replaced by $P_{\Gamma_a A}Z_{\beta A}$, where $P_{\Gamma_a A} := I_p - A^{-1}H\mathcal{T}_*^{-1}H'\Gamma_a'(\Gamma_a H\mathcal{T}_*^{-1}H'\Gamma_a')^{-1}\Gamma_a A$. For the case where Λ_β is of the form $\Lambda_\beta = \{\lambda_\beta \in R^p: \lambda_{\beta 1} \geq 0, \lambda_{\beta 2} \geq 0, \Gamma_a \lambda_\beta = \mathbf{0}\}$, $\hat{\lambda}_\beta$ is as defined in (6.6), but with $Z_{\beta A}$ replaced by $P_{\Gamma_a A}Z_{\beta A}$.

One can simulate the distribution of $\hat{\lambda}_\beta$ when Λ_β is as in Assumption 9 by simulating Z_β or $Z_{\beta A}$ and computing $\hat{\lambda}_\beta$ using a standard quadratic programming algorithm; e.g., see Gill, Murray, and Wright (1981). The programs GAUSS and Matlab have built-in procedures for doing so, called QPROG and QP respectively. The GAUSS procedure QPROG is very quick. For example, 10,000 simulation repetitions with $p = 15$, four equality constraints, and ten inequality constraints take about 63 seconds using a PC with Pentium 90 processor.

Alternatively, one can use the formulae of Theorem 5 or the equations above. These are easy to program because they only involve computing CF_j for $j = 1, \dots, 2^{p_b}$, finding the value \hat{j} that maximizes CF_j , and then computing $\hat{\lambda}_\beta = P_{L(\hat{j})}Z_\beta$ or $\hat{\lambda}_\beta = AP_{L_A(\hat{j})}Z_{\beta A}$. This method is not to be recommended if p_b is large, but for small values of p_b it works well. It is easy to program and is quick.

6.4. Consistent Standard Error Estimates

In this section, we describe three procedures for obtaining standard error estimators that are consistent whether or not the true parameter is on a boundary. Each actually provides a consistent estimator of the whole asymptotic distribution of $B_T(\hat{\theta} - \theta_0)$.

The first method is described as follows. Suppose the parameter space Θ is

$$(6.7) \quad \Theta = \{\theta \in R^s: g_a(\theta) = \mathbf{0}, m(\theta) \leq \mathbf{0}\}.$$

Assume that $m(\cdot): \Theta \rightarrow R^J$ is continuously differentiable at θ_0 . Let $m(\theta) = (m_1(\theta), \dots, m_J(\theta))'$. For $j = 1, \dots, J$, let $\{\eta_{Tj}: T \geq 1\}$ be a sequence of random variables (possibly constants) that satisfies $\eta_{Tj} \lambda_{\min}(B_T) \xrightarrow{p} \infty$. We specify a rule based on $m_j(\hat{\theta})$ and η_{Tj} to determine which (if any) of the inequality constraints are binding at the true parameter. If

$$(6.8) \quad m_j(\hat{\theta}) > -\eta_{Tj},$$

then we conclude that the j th constraint is binding. (Because this rule is essentially a one dimensional one-sided Wald test for some significance level α_T such that $\alpha_T \rightarrow 0$, the η_{Tj} 's could be chosen to be the critical values for such tests multiplied by an estimate of the standard error of $m_j(\hat{\theta})$ based on the usual formulae that assume that $m_j(\hat{\theta})$ is not on a boundary.)

Let j_1, \dots, j_k index the constraints that are found to be binding. Let $g_b(\theta) = (m_{j_1}(\theta), \dots, m_{j_k}(\theta))'$. Our estimate of the asymptotic distribution of $B_T(\hat{\theta} - \theta_0)$ is based on the supposition that the constraints that are found to be binding, i.e., $g_a(\theta) = \mathbf{0}$ and $g_b(\theta) \leq \mathbf{0}$, actually are binding. Then, we can obtain standard error estimators by simulating the asymptotic distribution with any unknown parameters replaced by consistent estimators.

This method is consistent given Assumptions 2–4, because

- (i) $m_j(\hat{\theta}) = m_j(\theta_0) + ((\partial/\partial\theta')m_j(\theta_0)B_T^{-1})B_T(\hat{\theta} - \theta_0) + o(\|\hat{\theta} - \theta_0\|)$,
- (ii) if $m_j(\theta_0) = 0$, $P(m_j(\hat{\theta}) > -\eta_{Tj}) = P((\partial/\partial\theta')m_j(\theta_0)B_T^{-1}O_p(1)\lambda_{\min}(B_T) + o_p(1) > -\eta_{Tj}\lambda_{\min}(B_T)) \rightarrow 1$, and
- (iii) if $m_j(\theta_0) < 0$, $P(m_j(\hat{\theta}) > -\eta_{Tj}) = P(m_j(\theta_0) + o_p(1) > -\eta_{Tj}) \rightarrow 0$.

The second method is a subsample method introduced by Wu (1990) and extended by Politis and Romano (1994) to cover cases where the statistic of interest has *some* asymptotic distribution, not necessarily normal, such as those considered in this paper. The method is applicable in iid contexts (see Politis and Romano (1994, Sec. 2)), as well as in stationary time series contexts (see Politis and Romano (1994, Sec. 3; 1996, Sec. 3)). A random subsampling variant of the procedure is also available; see Politis and Romano (1994, Sec. 2.2).

The third method is a version of the bootstrap in which bootstrap samples of size $T_1 (< T)$, rather than T , are employed. One uses the bootstrap distribution of $B_{T_1}(\hat{\theta}_{T_1}^* - \hat{\theta}_T)$ to estimate the distribution of $B_T(\hat{\theta}_T - \theta_0)$, where $\hat{\theta}_T$ denotes the estimator $\hat{\theta}$ constructed using T observations and $\hat{\theta}_{T_1}^*$ denotes the bootstrap estimator of $\hat{\theta}$ constructed from T_1 observations. In an iid context, $\hat{\theta}_{T_1}^*$ is constructed from T_1 iid draws with replacement from the original sample of T observations. This version of the bootstrap is consistent when $B_T = T^{1/2}M$ (for any matrix M), if $T_1/T \rightarrow 0$ as $T \rightarrow \infty$. Typically, one approximates the distribution of $B_{T_1}(\hat{\theta}_{T_1}^* - \hat{\theta}_T)$ by taking a number of simulation draws of it. Consistency of this procedure and the others above rely on the existence of an asymptotic distribution for $B_T(\hat{\theta} - \theta_0)$, which is established in this paper.

6.5. Examples (Continued)

6.5.1. Random Coefficient Regression

In Example 1, we partition θ as in Section 6.1 with

$$(6.9) \quad \begin{aligned} \theta_* &:= (\theta'_1, \theta'_2, \theta'_3)', & \psi &:= (\theta'_4, \theta'_5)', & \beta &:= \theta_1, & \text{and} \\ \delta &:= (\theta'_2, \theta'_3)'. \end{aligned}$$

With this partitioning, Assumptions 7 and 8 hold. In particular, by (3.11), \mathcal{F} is block diagonal between θ_* and ψ . The set Λ is a product set $\Lambda_\beta \times \Lambda_\delta \times \Lambda_\psi$ with

$$(6.10) \quad \Lambda_\beta := (R^+)^p, \quad \Lambda_\delta := R^{b_2+1}, \quad \text{and} \quad \Lambda_\psi := R^{b+1}.$$

Thus, Assumption 9 also holds.

With this partitioning, from (3.11) and (3.15), we have

$$(6.11) \quad \begin{aligned} \mathcal{F}_* &:= \frac{1}{2}EW_t^2W_t^{2'}/\text{var}_t^2(\theta_0), & \mathcal{F}_\psi &:= EW_tW_t'/\text{var}_t(\theta_0), \\ G &:= (G'_*, G'_\psi) \sim N(\mathbf{0}, \mathcal{I}), & G_* &\sim N(\mathbf{0}, \mathcal{I}_*), & G_\psi &\sim N(\mathbf{0}, \mathcal{I}_\psi), \\ \mathcal{I}_* &:= \frac{1}{4}E \frac{(\text{res}_t^2(\theta_0) - \text{var}_t(\theta_0))^2}{\text{var}_t^4(\theta_0)} W_t^2W_t^{2'}, & \text{and} & \\ \mathcal{I}_\psi &:= EW_tW_t'/\text{var}_t(\theta_0). \end{aligned}$$

Assumption 3^{3*} holds with \mathcal{I}_* as above. Assumption 3^{4*} holds with $c = 1$ if the errors ε_t and η_t are normally distributed.

By Theorem 3, $T^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} \hat{\lambda}$, where $\hat{\lambda} = (\hat{\lambda}'_\beta, \hat{\lambda}'_\delta, \hat{\lambda}'_\psi)'$. By Theorem 4(c), $q_\psi(\hat{\lambda}_\psi) = \inf_{\lambda_\psi \in \Lambda_\psi} q_\psi(\lambda_\psi)$, where $q_\psi(\lambda_\psi) := (\lambda_\psi - Z_\psi)' \mathcal{F}_\psi (\lambda_\psi - Z_\psi)$. Because $\Lambda_\psi = R^{b+1}$, this gives

$$(6.12) \quad \begin{aligned} \hat{\lambda}_\psi = Z_\psi &:= \mathcal{F}_\psi^{-1}G_\psi \sim N(\mathbf{0}, \mathcal{F}_\psi^{-1}) \quad \text{and} \\ T^{1/2} \left(\left(\hat{\theta}'_4, \hat{\theta}'_5 \right)' - \left(\theta'_{40}, \theta'_{50} \right)' \right) &\xrightarrow{d} \hat{\lambda}_\psi \sim N(\mathbf{0}, (EW_tW_t'/\text{var}_t(\theta_0))^{-1}). \end{aligned}$$

Thus, the QML regression parameter estimators $\hat{\theta}_4$ and $\hat{\theta}_5$ are asymptotically $N(\mathbf{0}, \mathcal{F}_\psi^{-1})$ whether or not some random coefficient variances are zero.

The matrix $B_T = T^{1/2}I_s$ obviously is block diagonal. Hence, by Corollary 1(b),

$$(6.13) \quad \begin{aligned} T^{1/2}(\hat{\theta}_1 - \theta_{10}) &\xrightarrow{d} \hat{\lambda}_\beta, \quad \text{where } \hat{\lambda}_\beta \text{ satisfies} \\ q_\beta(\hat{\lambda}_\beta) &= \inf_{\lambda_\beta \in (R^+)^p} q_\beta(\lambda_\beta), \\ q_\beta(\lambda_\beta) &:= (\lambda_\beta - Z_\beta)' (H\mathcal{F}_*^{-1}H')^{-1}(\lambda_\beta - Z_\beta), \quad \text{and} \\ Z_\beta &:= H\mathcal{F}_*^{-1}G_* \sim N(\mathbf{0}, H\mathcal{F}_*^{-1}\mathcal{I}_*\mathcal{F}_*^{-1}H'). \end{aligned}$$

For example, if $p = 1$ (i.e., there is one random coefficient with zero variance), then by (6.5), $\hat{\lambda}_\beta = AZ_{\beta A}1(Z_{\beta A} \geq 0) = Z_\beta 1(Z_\beta \geq 0)$ and $\hat{\lambda}_\beta$ has a half-normal distribution. If $p > 1$, then $\hat{\lambda}_\beta$ is given in closed form by (6.6) or Theorem 5.

Also by Corollary 1(b),

$$\begin{aligned}
 & T^{1/2} \left((\hat{\theta}'_2, \hat{\theta}'_3)' - (\theta'_{20}, \theta'_{30})' \right) \xrightarrow{d} \hat{\lambda}_\delta, \\
 & \hat{\lambda}_\delta = \mathcal{F}_\delta^{-1} G_\delta - \mathcal{F}_\delta^{-1} \mathcal{F}_{\delta\beta} \hat{\lambda}_\beta, \\
 & \mathcal{F}_\delta := \frac{1}{2} E \begin{pmatrix} X_{2t}^2 \\ 1 \end{pmatrix} \begin{pmatrix} X_{2t}^2 \\ 1 \end{pmatrix}' / \text{var}_t^2(\theta_0), \quad X_{2t}^2 := (X_{tp+1}^2, \dots, X_{tb}^2)' \in R^{b_2}, \\
 (6.14) \quad & \mathcal{F}_{\delta\beta} := \frac{1}{2} E \begin{pmatrix} X_{2t}^2 \\ 1 \end{pmatrix} X_{1t}', \quad X_{1t}^2 := (X_{t1}^2, \dots, X_{tp}^2)', \\
 & G_* := \begin{pmatrix} G_\beta \\ G_\delta \end{pmatrix} \sim N(\mathbf{0}, \mathcal{I}_*), \quad \text{and} \\
 & G_\delta \sim N \left(\mathbf{0}, \frac{1}{4} E \frac{(\text{res}_t^2(\theta_0) - \text{var}_t(\theta_0))^2}{\text{var}_t^4(\theta_0)} \begin{pmatrix} X_{2t}^2 \\ 1 \end{pmatrix} \begin{pmatrix} X_{2t}^2 \\ 1 \end{pmatrix}' \right).
 \end{aligned}$$

6.5.2. Dickey-Fuller Regression

In this example, we partition θ as in Section 6.1 above with

$$(6.15) \quad \theta_* = (\theta_1, \theta_2, \theta_3)', \quad \psi = \theta_4, \quad \beta = (\theta_1, \theta_2)', \quad \text{and} \quad \delta = \theta_3.$$

With this partitioning, Assumptions 7 and 8 hold by (3.21), (4.12), and (4.13).

The set A is a product set $A_\beta \times A_\delta \times A_\psi$ with

$$(6.16) \quad A_\beta = \begin{cases} R^- \times R^+ & \text{if } \theta_{30} = 0, \\ R^- \times R & \text{if } \theta_{30} > 0, \end{cases} \quad A_\delta = R, \quad \text{and} \quad A_\psi = R^b.$$

With the above partitioning, from (3.21), we have

$$\begin{aligned}
 (6.17) \quad & \mathcal{F}_* := \begin{pmatrix} \lambda^2 \int_0^1 W^2(r) dr & \lambda \int_0^1 r W(r) dr & \lambda \int_0^1 W(r) dr \\ \lambda \int_0^1 r W(r) dr & 1/3 & 1/2 \\ \lambda \int_0^1 W(r) dr & 1/2 & 1 \end{pmatrix}, \quad \mathcal{F}_\psi = V, \\
 & G := \begin{pmatrix} G_* \\ G_\psi \end{pmatrix}, \quad G_* := \begin{pmatrix} \frac{1}{2} \sigma \lambda (W^2(1) - 1) \\ \sigma (W(1) - \int_0^1 W(r) dr) \\ \sigma W(1) \end{pmatrix}, \quad \text{and} \\
 & G_\psi := G_4,
 \end{aligned}$$

where G_ψ is independent of G_* and \mathcal{F}_* .

By Theorem 3, $\Upsilon_T M(\hat{\theta} - \theta_0) \xrightarrow{d} \hat{\lambda}$, where $\hat{\lambda} = (\hat{\lambda}_\beta, \hat{\lambda}_\delta, \hat{\lambda}'_\psi)'$. By Theorem 4(c) and the fact that $A_\psi = R^b$, we find that

$$(6.18) \quad \hat{\lambda}_\psi = Z_\psi := \mathcal{F}_\psi^{-1} G_\psi \sim N(\mathbf{0}, V^{-1}).$$

By Theorem 4(a), $\hat{\lambda}_\beta$ solves $q_\beta(\hat{\lambda}_\beta) = \inf_{\lambda_\beta \in A_\beta} q_\beta(\lambda_\beta)$, where A_β is defined in (6.16). Closed form expressions for $\hat{\lambda}_\beta$ are given in (6.6) (with the first inequality reversed in each indicator function) when $\theta_{30} = 0$ and by $\hat{\lambda}_\beta = Z_\beta \mathbf{1}(Z_\beta \leq 0)$ when $\theta_{30} > 0$. Given $\hat{\lambda}_\beta$, Theorem 4(b) gives a closed form expression for $\hat{\lambda}_\delta$:

$$(6.19) \quad \hat{\lambda}_\delta = \mathcal{F}_\delta^{-1} G_\delta - \mathcal{F}_\delta^{-1} \mathcal{F}_{\delta\beta} \hat{\lambda}_\beta := \sigma W(1) - \left(\lambda \int_0^1 W(r) dr, 1/2 \right) \hat{\lambda}_\beta, \quad \text{where}$$

$$\mathcal{F}_\delta := 1, \quad \mathcal{F}_{\delta\beta} := \left(\lambda \int_0^1 W(r) dr, 1/2 \right), \quad \text{and} \quad G_\delta := \sigma W(1).$$

Note that $(\hat{\lambda}_\beta, \hat{\lambda}_\delta)$ is independent of $\hat{\lambda}_\psi$.
 We have

$$(6.20) \quad T_T M(\hat{\theta} - \theta_0) = \begin{pmatrix} T(\hat{\theta}_1 - \theta_{10}) \\ T^{3/2} \mu_0(\hat{\theta}_1 - \theta_{10}) + T^{3/2}(\hat{\theta}_2 - \theta_{20}) \\ -T^{1/2} \mu_0(\hat{\theta}_1 - \theta_{10}) + T^{1/2}(\hat{\theta}_3 - \theta_{30}) + T^{1/2} \mu_0 \mathbf{1}(\hat{\theta}_4 - \theta_{40}) \\ T^{1/2}(\hat{\theta}_4 - \theta_{40}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \hat{\lambda}_{\beta 1} \\ \hat{\lambda}_{\beta 2} \\ \hat{\lambda}_\delta \\ \hat{\lambda}_\psi \end{pmatrix},$$

where $\hat{\lambda}_\beta := (\hat{\lambda}_{\beta 1}, \hat{\lambda}_{\beta 2})'$. Equation (6.20) provides the asymptotic distribution of the unit root estimator $\hat{\theta}_1$ and of the short-run dynamics parameter estimator $\hat{\theta}_4$ directly. Note that the latter is asymptotically normal even though the unit root and time trend parameters, θ_{10} and θ_{20} , are on the boundary of the parameter space.

Equation (6.20) also provides the asymptotic distributions of nondegenerate linear combinations of the estimators $\hat{\theta}_1, \dots, \hat{\theta}_4$ that include the time trend and intercept parameter estimators $\hat{\theta}_2$ and $\hat{\theta}_3$. From these, the asymptotic distributions of $\hat{\theta}_2$ and $\hat{\theta}_3$ can be determined. First, the second row of (6.20) implies that $T\mu_0(\hat{\theta}_1 - \theta_{10}) + T(\hat{\theta}_2 - \theta_{20}) \xrightarrow{p} 0$ and the first row implies that $T\mu_0(\hat{\theta}_1 - \theta_{10}) \xrightarrow{d} \mu_0 \hat{\lambda}_{\beta 1}$. Hence,

$$(6.21) \quad T(\hat{\theta}_2 - \theta_{20}) \xrightarrow{d} -\mu_0 \hat{\lambda}_{\beta 1}.$$

Thus, the asymptotic joint distribution of $(T(\hat{\theta}_1 - \theta_{10}), T(\hat{\theta}_2 - \theta_{20}))'$ is $(\hat{\lambda}_{\beta 1}, -\mu_0 \hat{\lambda}_{\beta 1})'$, which is singular. Second, by the first, third, and fourth rows of (6.20),

$$(6.22) \quad T^{1/2}(\hat{\theta}_3 - \theta_{30}) \xrightarrow{d} \hat{\lambda}_\delta - \mu_0 \mathbf{1} \hat{\lambda}_\psi,$$

because $-T^{1/2} \mu_0(\hat{\theta}_1 - \theta_{10}) = o_p(1)$. Hence, (6.20) yields the asymptotic distributions of all the elements of $\hat{\theta}$ and their convergence holds jointly.

Cowles Foundation, P.O. Box 208281, New Haven, CT 06520-8281, U.S.A.

Manuscript received June, 1997; final revision received October, 1998.

APPENDIX

A. *A Taylor Expansion for a Function with Left/Right Partial Derivatives*

The following Taylor's Theorem is used to prove Lemma 1. Let f be as in Section 3.3. For $x \in \mathcal{X}$, let $(\partial/\partial x_j)f(x)$ denote the l/r partial derivative with respect to x_j (the j th element of x) of f at x . Let $(\partial^k/\partial x_{i_1}, \dots, \partial x_{i_k})f(x)$ denote the k th order l/r partial derivative of f at x with respect to x_{i_1}, \dots, x_{i_k} , where i_ℓ is a positive integer less than $s + 1 \forall \ell \leq k$.

THEOREM 6: *Let f be a function whose domain includes $\mathcal{X} \subset R^s$. Let $a \in \mathcal{X}$. Suppose $\mathcal{X} - a$ equals the intersection of a union of orthants and an open cube $C(\mathbf{0}, \varepsilon)$ for some $\varepsilon > 0$. Suppose f has continuous l/r partial derivatives of order $n + 1$ on \mathcal{X} for some integer $n \geq 0$. Then, for any $x \in \mathcal{X}$, there exists a point c on the line segment joining x and a such that*

$$f(x) = \sum_{k=0}^n \frac{1}{k!} D^k f(a)(x - a, \dots, x - a) + \frac{1}{(n + 1)!} D^{n+1} f(c)(x - a, \dots, x - a),$$

where $D^0 f(a)(x - a, \dots, x - a) := f(a)$ and for $k = 1, \dots, n + 1$ $D^k f(a)(x - a, \dots, x - a)$ denotes the k -linear map $D^k f(a)$ applied to the k -tuple $(x - a, \dots, x - a)$ defined by

$$D^k f(a)(x - a, \dots, x - a) = \sum_{i_1, \dots, i_k=1}^s \frac{\partial^k f(a)}{\partial x_{i_1}, \dots, \partial x_{i_k}} (x_{i_1} - a_{i_1}) \times \dots \times (x_{i_k} - a_{i_k}).$$

COMMENT: If the l/r partial derivatives of f of order k are continuous with respect to \mathcal{X} at a (i.e., they are continuous where continuity is defined in terms of local perturbations only within \mathcal{X}), then they are symmetric (i.e., $(\partial^2/\partial x_1 \partial x_2)f(a) = (\partial^2/\partial x_2 \partial x_1)f(a)$ for $k = 2$, etc.). This holds by the same argument as used to prove the symmetry of mixed (two-sided) partial derivatives; e.g., see Courant (1988, Ch. II, Sec. 3.3, pp. 55–56).

PROOF OF THEOREM 6: When $s = 1$, \mathcal{X} is either an open interval that contains a or a half-closed interval with a at the closed end. The Theorem holds in the former case by the standard one dimensional Taylor's Theorem. It holds in the latter case because standard proofs of the one dimensional Taylor's Theorem (e.g., see Apostol (1961, p. 366)) go through with x allowed to be an endpoint of \mathcal{X} provided the derivative of order k of f is redefined to be the l/r derivative of order k of f . The reason is that Rolle's Theorem (or the mean value theorem), upon which the proof depends, does not require f to be differentiable at the endpoints of \mathcal{X} .

When $s > 1$, standard proofs of Taylor's Theorem (e.g., see Courant (1988, Ch. II, Sec. 6, pp. 78–82)) apply Taylor's Theorem for $s = 1$ to the function $F(\lambda) = f(a + \lambda(x - a))$ for $\lambda \in [0, 1]$ and use the chain rule for multi-variable functions to verify the necessary differentiability conditions on F and to yield the form of the Taylor expansion.

The main condition of the chain rule is that the functions involved are differentiable at the appropriate points. In place of the condition of differentiability, we use the condition of l/r differentiability. We say that a function f is l/r differentiable at x if it can be approximated at x by a linear function and the approximation holds for all perturbations within \mathcal{L} . That is, $f(x+h) = f(x) + A'h + \varepsilon'_h h$ and $\|\varepsilon'_h\| \rightarrow 0$ as $\|h\| \rightarrow 0 \forall x+h \in \mathcal{L}$ for some vector A that is independent of h . Now, standard proofs of the chain rule (e.g., see Courant (1988, Ch. II, Sec. 5.1, pp. 69–73)) go through straightforwardly with partial derivatives and differentiable functions replaced by l/r partial derivatives and l/r differentiable functions.

To show that the functions $F(\lambda)$, $dF(\lambda)/d\lambda, \dots, d^n F(\lambda)/d\lambda^n$ are l/r differentiable for $\lambda \in [0, 1]$ (which is needed to apply our generalized chain rule), we use a generalization of the result that a function with continuous partial derivatives at a point is differentiable at that point. Standard proofs of this result (e.g., see Courant (1988, Ch. II, Sec. 4.1, pp. 59–62)) go through straightforwardly to show that a function with continuous l/r partial derivatives at a point is l/r differentiable at that point. In consequence, under the assumptions of the Theorem, the chain rule for l/r differentiable functions is applicable and the proof of Taylor's Theorem for continuous l/r partially differentiable functions is the same as that for continuous partially differentiable functions, which is referenced above.

B. Proofs for Quadratic Approximation Section 3

PROOF OF LEMMA 1: We prove part (a) first. By the Taylor expansion of Theorem 6, $\ell_T(\theta)$ satisfies (3.2) with

$$(7.1) \quad R_T(\theta) = \frac{1}{2}(\theta - \theta_0)' \left(\frac{\partial^2}{\partial \theta \partial \theta'} \ell_T(\theta^\dagger) - \frac{\partial^2}{\partial \theta \partial \theta'} \ell_T(\theta_0) \right) (\theta - \theta_0),$$

where θ^\dagger lies between θ and θ_0 , when $\theta \neq \theta_0$ and $R_T(\theta) = 0$ when $\theta = \theta_0$. Thus,

$$(7.2) \quad \begin{aligned} & \sup_{\theta \in \Theta: \|\theta - \theta_0\| \leq \gamma_T} |R_T(\theta_0)| / (1 + \|B_T(\theta - \theta_0)\|)^2 \\ & \leq \sup_{\theta \in \Theta: \|\theta - \theta_0\| \leq \gamma_T} \frac{1}{2} \left| (B_T(\theta - \theta_0))' B_T^{-1} \left(\frac{\partial^2}{\partial \theta \partial \theta'} \ell_T(\theta^\dagger) - \frac{\partial^2}{\partial \theta \partial \theta'} \ell_T(\theta_0) \right) B_T^{-1} \right. \\ & \quad \left. \times B_T(\theta - \theta_0) \right| / \|B_T(\theta - \theta_0)\|^2 \\ & \leq \sup_{\theta \in \Theta: \|\theta - \theta_0\| \leq \gamma_T} \frac{1}{2} \left\| B_T^{-1} \left(\frac{\partial^2}{\partial \theta \partial \theta'} \ell_T(\theta) - \frac{\partial^2}{\partial \theta \partial \theta'} \ell_T(\theta_0) \right) B_T^{-1} \right\| \\ & = o_p(1), \end{aligned}$$

where the equality holds by Assumption 2^{*}(c).

Part (b) follows from part (a) because the difference between the third summand on the right-hand side of (3.2) when defined with $D^2 \ell_T(\theta_0) = (\partial^2 / \partial \theta \partial \theta') \ell_T(\theta_0)$ and when defined with $D^2 \ell_T(\theta_0) = -B_T' \mathcal{S} B_T$ can be absorbed in the $R_T(\theta)$ term without affecting Assumption 2^{*}, due to the $1/(1 + \|B_T(\theta - \theta_0)\|)^2$ factor in Assumption 2^{*}.

PROOF OF THEOREM 1: Let $\kappa_T := \mathcal{S}_T^{-1/2} B_T(\hat{\theta} - \theta_0)$. θ_0 is in the closure of Θ (by Assumption 1). Thus, by (3.1), (3.2), (3.3), and Assumptions 1, 2^{*}, and 3,

$$(7.3) \quad \begin{aligned} o_p(1) & \leq \ell_T(\hat{\theta}) - \ell_T(\theta_0) \\ & = \kappa_T' \mathcal{S}_T^{-1/2} Z_T - \frac{1}{2} \|\kappa_T\|^2 + R_T(\hat{\theta}) \\ & = O_p(\|\kappa_T\|) - \frac{1}{2} \|\kappa_T\|^2 + (1 + \|\mathcal{S}_T^{-1/2} \kappa_T\|)^2 o_p(1) \\ & = O_p(\|\kappa_T\|) - \frac{1}{2} \|\kappa_T\|^2 + o_p(\|\kappa_T\|) + o_p(\|\kappa_T\|^2) + o_p(1). \end{aligned}$$

Rearranging this equation gives $\|\kappa_T\|^2 \leq 2\|\kappa_T\|O_p(1) + o_p(1)$. Let ξ_T denote the $O_p(1)$ term. Then,

$$(7.4) \quad (\|\kappa_T\| - \xi_T)^2 \leq \xi_T^2 + o_p(1) = O_p(1).$$

Taking square roots gives $\|\kappa_T\| \leq O_p(1)$. Given Assumption 3, this establishes Assumption 4.

PROOF OF THEOREM 2: Let $\kappa_{qT} := \mathcal{F}_T^{1/2}B_T(\hat{\theta}_q - \theta_0)$. By (3.5) and Assumption 3, we have

$$(7.5) \quad \begin{aligned} \|\kappa_{qT} - \mathcal{F}_T^{1/2}Z_T\|^2 &= q_T(B_T(\hat{\theta}_q - \theta_0)) \leq q_T(0) + o_p(1) \\ &= Z_T^T \mathcal{F}_T Z_T + o_p(1) = O_p(1). \end{aligned}$$

Thus, $\kappa_{qT} = \mathcal{F}_T^{1/2}Z_T + O_p(1) = O_p(1)$. By Assumption 3, this establishes part (a).

Parts (b) and (c) hold by (3.4), Assumptions 2 and 4, and part (a).

Parts (d) and (e) hold by parts (b) and (c), (3.1), and (3.5):

$$(7.6) \quad \begin{aligned} o_p(1) &\leq \ell_T(\hat{\theta}) - \ell_T(\hat{\theta}_q) \\ &= \frac{1}{2}q_T(B_T(\hat{\theta}_q - \theta_0)) - \frac{1}{2}q_T(B_T(\hat{\theta} - \theta_0)) + o_p(1) \leq o_p(1). \end{aligned}$$

Part (f) holds by parts (b) and (e).

C. Proofs for Parameter Space Section 4

PROOF OF LEMMA 2: Let $Z_{Tb} = Z_T/b_T$. By Assumption 3, $\|Z_{Tb}\| = O_p(b_T^{-1})$. For any set $\Gamma \subset R^s$ and $z \in R^s$, let

$$(7.7) \quad \text{dist}_T(z, \Gamma) := \inf_{\lambda \in \Gamma} ((\lambda - z)^T \mathcal{F}_T (\lambda - z))^{1/2}.$$

Note that $\text{dist}_T(Z_T, \Lambda) = \inf_{\lambda \in \Lambda} q_T^{1/2}(\lambda)$. Because Λ is a cone,

$$\text{dist}_T(Z_{Tb}, \Lambda) = b_T^{-1} \inf_{\lambda \in \Lambda} q_T^{1/2}(\lambda).$$

Also,

$$(7.8) \quad \begin{aligned} \text{dist}_T(Z_{Tb}, B_T(\Theta - \theta_0)/b_T) &= \inf_{\lambda \in B_T(\Theta - \theta_0)/b_T} (\lambda - Z_T/b_T)^T \mathcal{F}_T (\lambda - Z_T/b_T)^{1/2} \\ &= b_T^{-1} \inf_{\lambda \in B_T(\Theta - \theta_0)/b_T} (b_T \lambda - Z_T)^T \mathcal{F}_T (b_T \lambda - Z_T)^{1/2} \\ &= b_T^{-1} \inf_{\theta \in \Theta} q_T^{1/2}(B_T(\theta - \theta_0)). \end{aligned}$$

Let

$$(7.9) \quad C_T := \text{dist}_T(Z_{Tb}, \Lambda) - \text{dist}_T(Z_{Tb}, B_T(\Theta - \theta_0)/b_T).$$

By the results above, $C_T = b_T^{-1}(\inf_{\lambda \in \Lambda} q_T^{1/2}(\lambda) - \inf_{\lambda \in B_T(\Theta - \theta_0)} q_T^{1/2}(\lambda))$ and it suffices to show that $C_T = o_p(b_T^{-1})$.

Let $Z_{\Theta T b} \in B_T(\Theta - \theta_0)/b_T$ be such that $\text{dist}_T(Z_{Tb}, B_T(\Theta - \theta_0)/b_T) = \text{dist}(Z_{Tb}, \{Z_{\Theta T b}\}) + o_p(b_T^{-1})$. Define $Z_{\Lambda T b} \in \Lambda$ analogously with $B_T(\Theta - \theta_0)/b_T$ replaced by Λ . By Assumption 5, $\text{dist}(Z_{\Theta T b}, \Lambda) = o(\|Z_{\Theta T b}\|)$. This and Assumption 3 give $\text{dist}_T(Z_{\Theta T b}, \Lambda) = o_p(\|Z_{\Theta T b}\|)$. Analogously, $\text{dist}_T(Z_{\Lambda T b}, B_T(\Theta - \theta_0)/b_T) = o_p(\|Z_{\Lambda T b}\|)$. (To make the above argument utilizing Assumption 5 really precise, we need to use an almost sure representation argument based on the fact that $Z_{\Theta T b} = o_p(1)$, as proved below. For brevity, we do not give the details.)

By the triangle inequality,

$$(7.10) \quad \begin{aligned} C_T &\leq \text{dist}_T(Z_{Tb}, \{Z_{\Theta T b}\}) + \text{dist}_T(Z_{\Theta T b}, \Lambda) - \text{dist}(Z_{Tb}, B_T(\Theta - \theta_0)/b_T) \\ &= \text{dist}_T(Z_{\Theta T b}, \Lambda) + o_p(b_T^{-1}) \\ &= o_p(\|Z_{\Theta T b}\|) + o_p(b_T^{-1}). \end{aligned}$$

Analogously, $C_T \geq o_p(\|Z_{\Lambda T b}\|) + o_p(b_T^{-1})$.

By assumption, $\mathbf{0}$ belongs to the closure of $\Theta - \theta_0$ and, hence, to the closure of $B_T(\Theta - \theta_0)/b_T$. This gives

$$(7.11) \quad \begin{aligned} \text{dist}_T(Z_{Tb}, \{Z_{\Theta Tb}\}) &= \text{dist}_T(Z_{Tb}, B_T(\Theta - \theta_0)/b_T) + o_p(b_T^{-1}) \\ &\leq \|\mathcal{F}_T^{1/2} Z_{Tb}\| + o_p(b_T^{-1}). \end{aligned}$$

Using Assumption 3, we then obtain

$$(7.12) \quad \begin{aligned} \|Z_{\Theta Tb} - Z_{Tb}\| &\leq \text{dist}_T(Z_{Tb}, \{Z_{\Theta Tb}\})/\lambda_{\min}(\mathcal{F}_T^{1/2}) \\ &\leq (\|\mathcal{F}_T^{1/2} Z_{Tb}\| + o_p(b_T^{-1}))/\lambda_{\min}(\mathcal{F}_T^{1/2}) \\ &\leq \|Z_{Tb}\| \lambda_{\max}(\mathcal{F}_T^{1/2})/\lambda_{\min}(\mathcal{F}_T^{1/2}) + o_p(b_T^{-1}) = O_p(b_T^{-1}). \end{aligned}$$

Thus,

$$(7.13) \quad \|Z_{\Theta Tb}\| \leq \|Z_{\Theta Tb} - Z_{Tb}\| + \|Z_{Tb}\| = O_p(b_T^{-1}).$$

Analogously, $\|Z_{ATb}\| = O_p(b_T^{-1})$. Combining these results gives $C_T = o_p(b_T^{-1})$.

PROOF OF LEMMA 3: Assumption 5* implies Assumption 5 because (i) $B_T = b_T I_s$ implies that $B_T(\Theta - \theta_0)/b_T = \Theta - \theta_0$, (ii) for $\phi_T \in (\Theta - \theta_0) \cap S(0, \varepsilon)$, $\text{dist}(\phi_T, \Lambda) = 0$ for some $\varepsilon > 0$ by Assumption 5*(a), and (iii) for $\lambda_T \in \Lambda \cap S(0, \varepsilon)$, $\text{dist}(\lambda_T, \Theta - \theta_0) = 0$ for some $\varepsilon > 0$ by Assumption 5*(a).

We now show that Assumption 52* implies Assumption 5 with $b_T := \lambda_{\min}(\mathcal{T}_T)$. Assume Assumption 52* holds. A sequence $\{\phi_T \in R^s: T \geq 1\}$ with $\|\phi_T\| \rightarrow 0$ satisfies

$$(7.14) \quad \phi_T \in B_T(\Theta - \theta_0)/b_T \quad \forall T \text{ large} \quad \text{iff} \quad \phi_T \in B_T \Lambda^* \quad \forall T \text{ large}.$$

This holds because $T_{Tj}/b_T \geq 1 \quad \forall T \geq 1, \forall j \leq s$ (where $\mathcal{T}_T := \text{diag}(T_{T1}, \dots, T_{Ts})$) implies that $\|b_T M^{-1} T_T^{-1} \phi_T\| \leq \|M^{-1}\| \cdot \|\phi_T\| \rightarrow 0$. Suppose $\phi_T \in B_T(\Theta - \theta_0)/b_T \quad \forall T$ large; then $b_T M^{-1} T_T^{-1} \phi_T \in (\Theta - \theta_0) \cap S(0, \varepsilon) = \Lambda^* \cap S(0, \varepsilon) \subset \Lambda^* \quad \forall T$ large and $\phi_T \in B_T \Lambda^* \quad \forall T$ large. Conversely, suppose $\phi_T \in B_T \Lambda^* \quad \forall T$ large; then $b_T M^{-1} T_T^{-1} \phi_T \in \Lambda^* \cap S(0, \varepsilon) = (\Theta - \theta_0) \cap S(0, \varepsilon) \subset \Theta - \theta_0 \quad \forall T$ large and $\phi_T \in B_T(\Theta - \theta_0)/b_T \quad \forall T$ large.

Using (7.14), for any sequence $\{\phi_T \in B_T(\Theta - \theta_0)/b_T: T \geq 1\}$ with $\|\phi_T\| \rightarrow 0$, we have $\phi_T \in B_T \Lambda^* \quad \forall T$ large. For such a sequence,

$$(7.15) \quad \text{dist}(\phi_T, \Lambda) = \|\phi_T\| \text{dist}(\phi_T/\|\phi_T\|, \Lambda) \leq \|\phi_T\| \text{dist}_c(B_T \Lambda^*, \Lambda) = o(\|\phi_T\|),$$

where the first equality holds because Λ is a cone, the inequality holds by the definition of $\text{dist}_c(\cdot, \cdot)$ and the fact that $\phi_T \in B_T \Lambda^*$ and $B_T \Lambda^*$ is a cone, and the last equality holds by Assumption 52*(c).

For any sequence $\{\lambda_T \in \Lambda: T \geq 1\}$ for which $\|\lambda_T\| \rightarrow 0$,

$$(7.16) \quad \begin{aligned} \text{dist}(\lambda_T, B_T \Lambda^*) &= \|\lambda_T\| \text{dist}(\lambda_T/\|\lambda_T\|, B_T \Lambda^*) \leq \|\lambda_T\| \text{dist}_c(\Lambda, B_T \Lambda^*) \\ &= o(\|\lambda_T\|) \end{aligned}$$

by the same argument as above. Now, for some $\phi_T \in B_T \Lambda^* \quad \forall T \geq 1$,

$$(7.17) \quad \text{dist}(\lambda_T, B_T \Lambda^*) = \|\lambda_T - \phi_T\| + o(\|\lambda_T\|) \geq \text{dist}(\lambda_T, B_T(\Theta - \theta_0)/b_T)$$

$\forall T$ large, where the inequality holds because $\|\lambda_T\| \rightarrow 0$ implies $\|\phi_T\| \rightarrow 0$ implies $\phi_T \in B_T(\Theta - \theta_0)/b_T$ using (7.14). Equations (7.15)–(7.17) combine to verify Assumption 5.

D. Proofs for Asymptotic Distribution Section 5

PROOF OF THEOREM 3: First, we establish part (a). Let $\lambda_T^* \in \text{cl}(\Lambda)$ be such that $\|B_T(\hat{\theta} - \theta_0) - \lambda_T^*\| = \text{dist}(B_T(\hat{\theta} - \theta_0), \Lambda)$. λ_T^* is unique because Λ is a convex cone; see Perlman (1969, Sec. 4). By Assumptions 4 and 5, $\|B_T(\hat{\theta} - \theta_0)/b_T - \lambda_T^*/b_T\| = \text{dist}(B_T(\hat{\theta} - \theta_0)/b_T, \Lambda) = o(\|B_T(\hat{\theta} - \theta_0)/b_T\|) = o_p(b_T^{-1})$ and so

$$(7.18) \quad \|B_T(\hat{\theta} - \theta_0) - \lambda_T^*\| = o_p(1).$$

Thus, it suffices to show that $\|\lambda_T^* - \hat{\lambda}_T\| = o_p(1)$.

Define $\|\cdot\|_T$ by $\|\lambda\|_T := (\lambda^T \mathcal{F}_T \lambda)^{1/2}$. By Assumption 3, it suffices to show that $\|\lambda_T^* - \hat{\lambda}_T\|_T = o_p(1)$. By Assumption 3, (7.18) holds with $\|\cdot\|$ replaced by $\|\cdot\|_T$. This, the triangle inequality, and Lemma 2 give

$$(7.19) \quad \|\lambda_T^* - Z_T\|_T = \|B_T(\hat{\theta} - \theta_0) - Z_T\|_T + o_p(1) = \|\hat{\lambda}_T - Z_T\|_T + o_p(1).$$

In consequence,

$$(7.20) \quad \begin{aligned} \varepsilon_T &:= \|\lambda_T^* - Z_T\|_T - \|\hat{\lambda}_T - Z_T\|_T = o_p(1) \quad \text{and} \\ \varepsilon_T^* &:= \|\lambda_T^* - Z_T\|_T^2 - \|\hat{\lambda}_T - Z_T\|_T^2 = o_p(1). \end{aligned}$$

First, suppose $Z_T \in \text{cl}(\Lambda)$. Then, $\hat{\lambda}_T = Z_T$, $\|\lambda_T^* - \hat{\lambda}_T\|_T = \|\lambda_T^* - Z_T\|_T = \|\hat{\lambda}_T - Z_T\|_T + \varepsilon_T = \varepsilon_T = o_p(1)$.

Alternatively, suppose $Z_T \notin \text{cl}(\Lambda)$. (We now use a geometric argument that is most easily followed by drawing a picture.) $\hat{\lambda}_T$ is on the boundary of Λ , because $\hat{\lambda}_T$ minimizes $\|\lambda - Z_T\|_T$ over $\lambda \in \text{cl}(\Lambda)$ and $Z_T \notin \text{cl}(\Lambda)$. Let $L(\hat{\lambda}_T, Z_T)$ denote the line through $\hat{\lambda}_T$ and Z_T . $L(\hat{\lambda}_T, Z_T)$ is perpendicular (with respect to the norm $\|\cdot\|_T$) to the ray through $\hat{\lambda}_T$ starting at the origin. Let P_L denote the projection onto $L(\hat{\lambda}_T, Z_T)$ with respect to the norm $\|\cdot\|_T$. Because $\lambda_T^* \in \Lambda$ and Λ is convex, $P_L \lambda_T^* \in \Lambda$. By definition of $\hat{\lambda}_T$, $\|\hat{\lambda}_T - Z_T\|_T \leq \|P_L \lambda_T^* - Z_T\|_T$. In consequence, $\hat{\lambda}_T$ lies on the line segment joining $P_L \lambda_T^*$ and Z_T .

By the orthogonality of projections,

$$(7.21) \quad \|\lambda_T^* - \hat{\lambda}_T\|_T^2 = \|\lambda_T^* - P_L \lambda_T^*\|_T^2 + \|P_L \lambda_T^* - \hat{\lambda}_T\|_T^2.$$

We claim that (i) $\|\lambda_T^* - P_L \lambda_T^*\|_T^2 \leq \varepsilon_T^*$ and (ii) $\|P_L \lambda_T^* - \hat{\lambda}_T\|_T^2 \leq \varepsilon_T^2$. These two claims and (7.21) combine to yield $\|\lambda_T^* - \hat{\lambda}_T\|_T \leq \varepsilon_T^* + \varepsilon_T^2$ when $Z_T \notin \text{cl}(\Lambda)$, which gives the desired result.

Claim (i) follows from

$$(7.22) \quad \begin{aligned} \|\lambda_T^* - P_L \lambda_T^*\|_T^2 &= \|\lambda_T^* - Z_T\|_T^2 - \|P_L \lambda_T^* - Z_T\|_T^2 \\ &= \|\hat{\lambda}_T - Z_T\|_T^2 + \varepsilon_T^* - \|P_L \lambda_T^* - Z_T\|_T^2 \\ &\leq \varepsilon_T^*, \end{aligned}$$

because $\hat{\lambda}_T$ lies on the line segment joining $P_L \lambda_T^*$ and Z_T .

Claim (ii) is established as follows. The first equality of (7.22) implies that

$$(7.23) \quad \|P_L \lambda_T^* - Z_T\|_T \leq \|\lambda_T^* - Z_T\|_T = \|\hat{\lambda}_T - Z_T\|_T + \varepsilon_T.$$

This result and the fact that $\hat{\lambda}_T$ lies on the line segment joining $P_L \lambda_T^*$ and Z_T give

$$(7.24) \quad \begin{aligned} \|P_L \lambda_T^* - \hat{\lambda}_T\|_T &= \|P_L \lambda_T^* - Z_T\|_T - \|\hat{\lambda}_T - Z_T\|_T \\ &\leq \|\hat{\lambda}_T - Z_T\|_T + \varepsilon_T - \|\hat{\lambda}_T - Z_T\|_T \\ &= \varepsilon_T, \end{aligned}$$

which completes the proof of part (a).

Next, we establish part (b). In part (b), $\hat{\lambda}_T$ is uniquely defined because Λ is a convex cone. We can write $\hat{\lambda}_T = h(B_T^{-1} D \mathcal{L}_T(\theta_0), \mathcal{F}_T)$, where the function h is defined implicitly in (5.1). The function h is continuous at all points $(B_T^{-1} D \mathcal{L}_T(\theta_0), \mathcal{F}_T)$ for which \mathcal{F}_T is nonsingular. Because \mathcal{F} is nonsingular with probability one, the continuous mapping theorem gives $\hat{\lambda}_T = h(B_T^{-1} D \mathcal{L}_T(\theta_0), \mathcal{F}_T) \rightarrow_d h(G, \mathcal{F}) = \hat{\lambda}$. The second result of part (b) holds by the first result and part (a) of the Theorem.

The convergence result of part (c) holds by (4.4), Assumption 3, and the continuous mapping theorem. The equality in part (c) holds by the orthogonality property $\hat{\lambda}^T(\hat{\lambda} - Z) = 0$, which does not require Assumption 6; see Perlman (1969, Lemma 4.1), and some algebra.

E. Proofs for Asymptotic Distribution of Subvectors Section 6

PROOF OF THEOREM 4: First, we break up $q(\lambda)$ and $Z'\mathcal{Z}$ into terms involving θ_* and ψ . For $\lambda_* \in \Lambda_\beta \times \Lambda_\delta$, define

$$(7.25) \quad q_*(\lambda_*) := (\lambda_* - Z_*)'\mathcal{F}_*(\lambda_* - Z_*).$$

By Assumption 7,

$$(7.26) \quad \begin{aligned} q(\lambda) &= q_*(\lambda_*) + q_\psi(\lambda_\psi) \quad \text{for } \lambda = (\lambda'_*, \lambda'_\psi)', \\ \inf_{\lambda \in \Lambda} q(\lambda) &= \inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*) + \inf_{\lambda_\psi \in \Lambda_\psi} q_\psi(\lambda_\psi), \quad \text{and} \\ Z'\mathcal{Z} &= Z'_*\mathcal{F}_*Z_* + Z'_\psi\mathcal{F}_\psi Z_\psi. \end{aligned}$$

Next, we have

$$(7.27) \quad \begin{aligned} 0 &\leq q_*(\hat{\lambda}_*) - \inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*) \\ &\leq q_*(\hat{\lambda}_*) - \inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*) + q_\psi(\hat{\lambda}_\psi) - \inf_{\lambda_\psi \in \Lambda_\psi} q_\psi(\lambda_\psi) \\ &= q(\hat{\lambda}) - \inf_{\lambda \in \Lambda_\beta \times \Lambda_\delta \times \Lambda_\psi} q(\lambda) \\ &= 0, \end{aligned}$$

where the first equality uses (7.26) and the second holds by the definition of $\hat{\lambda}$. In consequence, we obtain

$$(7.28) \quad q_*(\hat{\lambda}_*) = \inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*).$$

Part (c) of the Theorem follows from (7.26) and (7.28).

We now use Assumption 8 to break $q_*(\lambda_*)$ and $Z'_*\mathcal{F}_*Z_*$ into terms involving β and δ . Let

$$(7.29) \quad \begin{aligned} A &:= \begin{bmatrix} I_p \\ -\mathcal{F}_\delta^{-1}\mathcal{F}_{\delta\beta} \end{bmatrix} \in R^{(p+q) \times p}, \quad P^\perp := AH \in R^{(p+q) \times (p+q)}, \quad \text{and} \\ P &:= I_{p+q} - P^\perp. \end{aligned}$$

Define the norm $\|\cdot\|_*$ on R^{p+q} by $\|h\|_* = (h'\mathcal{F}_*h)^{1/2}$ for $h \in R^{p+q}$. Let L be the linear subspace of R^{p+q} defined by $L := \{(0', \delta') : \text{for some } \delta \in R^q\}$. Let L^\perp denote the orthogonal complement of L with respect to $\|\cdot\|_*$. P and P^\perp project onto L and L^\perp , respectively, with respect to $\|\cdot\|_*$. Thus, $(Ph_1)\mathcal{F}_*P^\perp h_2 = 0 \forall h_1, h_2 \in R^{p+q}$. By some algebra,

$$(7.30) \quad A'\mathcal{F}_*A = (H\mathcal{F}_*^{-1}H')^{-1} \quad \text{and} \quad P\mathcal{F}_*^{-1}G_* = \begin{bmatrix} \mathbf{0} \\ \mathcal{F}_\delta^{-1}G_\delta \end{bmatrix}.$$

(For the second result, note that $I_{p+q} = \begin{bmatrix} H \\ F \end{bmatrix}$ for $F := [\mathbf{0} : I_q] \in R^{q \times (p+q)}$, $HP\mathcal{F}_*^{-1}G_* = \mathbf{0}$ because $HA = I_p$, and $FP\mathcal{F}_*^{-1}G_* = \mathcal{F}_\delta^{-1}G_\delta$ because $\mathcal{F}_\delta F\mathcal{F}_*^{-1}G_* = G_\delta$ by some algebra.)

The above results give

$$(7.13) \quad \begin{aligned} Z'_*\mathcal{F}_*Z_* &= (P^\perp Z'_*)'\mathcal{F}_*P^\perp Z_* + (PZ'_*)'\mathcal{F}_*PZ_* \\ &= Z'_\beta (H\mathcal{F}_*^{-1}H')^{-1} Z_\beta = G'_\beta \mathcal{F}_\delta^{-1} G_\delta. \end{aligned}$$

Equations (7.26) and (7.31) establish part (d) of the Theorem.

For $\lambda_* = (\lambda'_\beta, \lambda'_\delta)' \in \Lambda_\beta \times \Lambda_\delta$, we have

$$(7.32) \quad P\lambda_* = \begin{pmatrix} \mathbf{0} \\ \lambda_\delta + \mathcal{F}_\delta^{-1}\mathcal{F}_{\delta\beta}\lambda_\beta \end{pmatrix}.$$

For $\lambda_* = (\lambda'_\beta, \lambda'_\delta) \in \Lambda_\beta \times \Lambda_\delta$, define

$$(7.33) \quad q_\delta(\lambda_\beta, \lambda_\delta) := (\lambda_\delta + \mathcal{T}_\delta^{-1} \mathcal{T}_{\delta\beta} \lambda_\beta - \mathcal{T}_\delta^{-1} G_\delta)' \mathcal{T}_\delta (\lambda_\delta + \mathcal{T}_\delta^{-1} \mathcal{T}_{\delta\beta} \lambda_\beta - \mathcal{T}_\delta^{-1} G_\delta).$$

Then, using (7.30) and (7.32), we have

$$(7.34) \quad q_*(\lambda_*) = (P^\perp \lambda_* - P^\perp Z_*)' \mathcal{T}_* (P^\perp \lambda_* - P^\perp Z_*) + (P\lambda_* - PZ_*)' \mathcal{T}_* (P\lambda_* - PZ_*) \\ = q_\beta(\lambda_\beta) + q_\delta(\lambda_\beta, \lambda_\delta).$$

Under Assumption 8, for any $\lambda_\beta \in R^p$, we have

$$(7.35) \quad \inf_{\lambda_\delta \in \Lambda_\delta} q_\delta(\lambda_\beta, \lambda_\delta) = \inf_{\lambda_\delta \in R^q} q_\delta(\lambda_\beta, \lambda_\delta) = 0.$$

Thus, using (7.34) and (7.35), we obtain

$$(7.36) \quad \inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*) = \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta).$$

Equations (7.26) and (7.36) establish part (e) of the Theorem.

Part (a) of the Theorem follows from

$$(7.37) \quad 0 \leq q_\beta(\hat{\lambda}_\beta) - \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta) \leq q_\beta(\hat{\lambda}_\beta) + q_\delta(\hat{\lambda}_\beta, \hat{\lambda}_\delta) - \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta) \\ = q_*(\hat{\lambda}_*) - \inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*) \leq 0,$$

where the equality holds by (7.34) and (7.36) using Assumption 8.

By equation (7.34),

$$(7.38) \quad q_*(\hat{\lambda}_*) = q_\beta(\hat{\lambda}_\beta) + q_\delta(\hat{\lambda}_\beta, \hat{\lambda}_\delta).$$

By equations (7.28) and (7.36), $q_*(\hat{\lambda}_*) = \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta)$. This, (7.38), and part (a) of the Theorem give $q_\delta(\hat{\lambda}_\beta, \hat{\lambda}_\delta) = 0$. The latter and (7.33) yield part (b) of the Theorem.

The first equality of part (f) of the Theorem follows from parts (d) and (e). The second equality of part (f) holds by the orthogonality properties $\hat{\lambda}'_\beta (H \mathcal{T}_*^{-1} H')^{-1} (\hat{\lambda}_\beta - Z_\beta) = 0$ and $\hat{\lambda}'_\psi \mathcal{T}_\psi (\hat{\lambda}_\psi - Z_\psi) = 0$; see Perlman (1969, Lemma 4.1), and some algebra.

PROOF OF LEMMA 4: For any linear subspace $L \subset R^p$ and any $z \in R^p$, $\ell_z \in L$ is the projection of z onto L with respect to the norm $\|\cdot\|_\beta$ if and only if ℓ_z minimizes $\|z - \ell\|_\beta$ over $\ell \in L \cap S(\ell_z, \varepsilon)$ for some $\varepsilon > 0$. Necessity of the latter holds by the definition of a projection. To prove sufficiency of the latter, suppose the latter holds but the former does not. Then, $P_L z \neq \ell_z$, and every point on the line segment joining $P_L z$ and ℓ_z yields a smaller criterion function value than the endpoint ℓ_z . But this is a contradiction.

Now, given $\hat{\lambda}_\beta \in \Lambda_\beta$, we can construct two matrices Γ_{b1} and Γ_{b2} such that $\Gamma_{b1} \hat{\lambda}_\beta = \mathbf{0}$ and $\Gamma_{b2} \hat{\lambda}_\beta < \mathbf{0}$ (element by element), where Γ_{b1} and Γ_{b2} are comprised of different rows of Γ_b and together they include all the rows of Γ_b . In addition, $\Gamma_a \hat{\lambda}_\beta = \mathbf{0}$. Let $L := \{\ell \in R^p: \Gamma_a \ell = \mathbf{0}, \Gamma_{b1} \ell = \mathbf{0}\}$. For some $\varepsilon > 0$, $L \cap S(\hat{\lambda}_\beta, \varepsilon) = \Lambda_\beta \cap S(\hat{\lambda}_\beta, \varepsilon)$, because the restrictions $\Gamma_{b2} \ell < \mathbf{0}$ are satisfied for ℓ close to $\hat{\lambda}_\beta$. By definition, $\hat{\lambda}_\beta$ minimizes $\|\lambda_\beta - Z_\beta\|_\beta$ over $\lambda_\beta \in \Lambda_\beta \cap S(\hat{\lambda}_\beta, \varepsilon)$. Hence, $\hat{\lambda}_\beta$ minimizes the same function over $\lambda_\beta \in L \cap S(\hat{\lambda}_\beta, \varepsilon)$ as well. By the first paragraph of the proof, then, $\hat{\lambda}_\beta$ equals the projection of Z_β onto the linear subspace L .

REFERENCES

AITCHISON, J., AND S. D. SILVEY (1958): "Maximum Likelihood Estimation of Parameters Subject to Constraint," *Annals of Mathematical Statistics*, 29, 813-828.

- ANDREWS, D. W. K. (1992): "Generic Uniform Convergence," *Econometric Theory*, 8, 241–257.
- (1994a): "Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity," *Econometrica*, 62, 43–72.
- (1994b): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics, Vol. IV*, ed. by R. F. Engle and D. McFadden. New York: North-Holland.
- (1996): "Admissibility of the Likelihood Ratio Test When the Parameter Space Is Restricted Under the Alternative," *Econometrica*, 64, 705–718.
- (1997a): "Estimation When a Parameter Is on a Boundary of the Parameter Space: Part II," unpublished manuscript, Yale University.
- (1997b): "Estimation When a Parameter Is on a Boundary of the Parameter Space: Theory and Applications," Cowles Foundation Discussion Paper Number 1153, Yale University.
- (1998a): "Hypothesis Testing with a Restricted Parameter Space," *Journal of Econometrics*, 84, 155–199.
- (1998b): "Testing When a Parameter Is on the Boundary of the Maintained Hypothesis," unpublished manuscript, Cowles Foundation Discussion Paper No. 1229, Yale University.
- (1999): "Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space," *Econometrica*, 67, forthcoming.
- ANDREWS, D. W. K., AND C. J. MCDERMOTT (1995): "Nonlinear Econometric Models with Deterministically Trending Variables," *Review of Economic Studies*, 62, 343–360.
- APOSTOL, T. M. (1961): *Calculus*, Vol. I. New York: Blaisdell.
- CHANT, D. (1974): "On Asymptotic Tests of Composite Hypotheses in Nonstandard Conditions," *Biometrika*, 61, 291–298.
- CHERNOFF, H. (1954): "On the Distribution of the Likelihood Ratio," *Annals of Mathematical Statistics*, 54, 573–578.
- COURANT, R. (1988): *Differential and Integral Calculus*, Vol. II, Wiley Classics Library Edition. New York: Wiley. First published 1936.
- GEYER, C. J. (1994): "On the Asymptotics of Constrained M-estimation," *Annals of Statistics*, 22, 1993–2010.
- GILL, P. E., W. MURRAY, AND M. H. WRIGHT (1981): *Practical Optimization*. New York: Academic Press.
- GOURIEROUX, C., AND A. MONFORT (1989): *Statistique et Modeles Econometriques*, Vol. 2. Paris: Economica. English translation (1995): *Statistics and Econometric Models*, Vol. 2, translated by Q. Vuong. Cambridge, U.K.: Cambridge University Press.
- HAMILTON, J. D. (1994): *Time Series Analysis*. Princeton, NJ: Princeton University Press.
- HAN, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 35, 303–316.
- HUBER, P. J. (1967): "The Behaviour of Maximum Likelihood Estimates under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, 1. Berkeley: University of California, pp. 221–233.
- JEGANATHAN, P. (1982): "On the Asymptotic Theory of Estimation When the Limit of the Loglikelihood Is Mixed Normal," *Sankhya*, Series A, 44, Part 2, 172–212.
- JUDGE, G. G., AND T. TAKAYAMA (1966): "Inequality Restrictions in Regression Analysis," *Journal of the American Statistical Association*, 61, 166–181.
- LE CAM, L. (1960): "Locally Asymptotically Normal Families of Distributions," *University of California Publications in Statistics*, 3, 37–98.
- LIEW, C. K. (1976): "Inequality Constrained Least-squares Estimation," *Journal of the American Statistical Association*, 71, 746–751.
- LOVELL, M. C., AND E. PRESCOTT (1970): "Multiple Regression with Inequality Constraints: Pretesting Bias, Hypothesis Testing and Efficiency," *Journal of the American Statistical Association*, 65, 913–925.
- MORAN, P. A. P. (1971): "Maximum-likelihood Estimation in Non-standard Conditions," *Proceedings of the Cambridge Philosophical Society*, 70, 441–450.
- PAKES, A., AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

- PERLMAN, M. D. (1969): "One-sided Problems in Multivariate Analysis," *Annals of Mathematical Statistics*, 40, 549–567. Corrections in *Annals of Mathematical Statistics*, 42, 1777.
- PFANZAGL, J. (1969): "On the Measurability and Consistency of Minimum Contrast Estimates," *Metrika*, 14, 249–272.
- PHILLIPS, P. C. B., AND V. SOLO (1992): "Asymptotics for Linear Processes," *Annals of Statistics*, 20, 971–1001.
- POLITIS, D. N., AND J. P. ROMANO (1994): "Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions," *Annals of Statistics*, 22, 2031–2050.
- (1996): "Subsampling for Econometric Models: Comments on 'Bootstrapping Time Series Models,'" *Econometric Reviews*, 15, 169–176.
- POLLARD, D. (1985): "New Ways to Prove Central Limit Theorems," *Econometric Theory*, 1, 295–314.
- ROBINSON, P. M. (1988): "Root- N -Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.
- ROTHENBERG, T. J. (1973): *Efficient Estimation with A Priori Information*, Cowles Foundation Monograph No. 23. New Haven: Yale University Press.
- SAIKKONEN, P. (1995): "Problems with the Asymptotic Theory of Maximum Likelihood Estimation in Integrated and Cointegrated Systems," *Econometric Theory*, 11, 888–911.
- SELF, S. G., AND K.-Y. LIANG (1987): "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610.
- SHERMAN, R. P. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123–137.
- VAN DER VAART, A. W., AND J. WELLNER (1996): *Weak Convergence and Empirical Processes*. New York: Springer.
- WANG, J. (1996): "Asymptotics of Least-Squares Estimators for Constrained Nonlinear Regression," *Annals of Statistics*, 24, 1316–1326.
- WU, C. F. J. (1990): "On the Asymptotic Properties of the Jackknife Histogram," *Annals of Statistics*, 18, 1438–1452.