COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1153

ESTIMATION WHEN A PARAMETER IS ON A BOUNDARY: THEORY AND APPLICATIONS

Donald W. K. Andrews

June 1997

Estimation When a Parameter Is on a Boundary: Theory and Applications

Donald W. K. Andrews[1]
Cowles Foundation for Research in Economics
Yale University

January 1997
Revised: May 1997

Abstract

Examples treated explicitly in the paper are: (1) quasi-ML estimation of a random coefficients regression model with some coefficient variances equal to zero, (2) LS estimation of a regression model with nonlinear equality and/or inequality restrictions on the parameters and iid regressors, (3) LS estimation of an augmented Dickey-Fuller regression with unit root and time trend parameters on the boundary of the parameter space, (4) method of simulated moments estimation of a multinomial discrete response model with some random coefficient variances equal to zero, some random effect variances equal to zero, or some measurement error variances equal to zero, (5) quasi-ML estimation of a GARCH$(1, q^*)$ or IGARCH$(1, q^*)$ model with some GARCH MA parameters equal to zero, (6) semiparametric LS estimation of a partially linear regression model with nonlinear equality and/or inequality restrictions on the parameters, and (7) LS estimation of a regression model with nonlinear equality and/or inequality restrictions on the parameters and integrated regressors.

Keywords: Asymptotic distribution, boundary, equality restrictions, extremum estimator, GARCH$(1, q^*)$ model, generalized method of moments estimator, inequality restrictions, integrated regressors, least squares estimator, maximum likelihood estimator, locally asymptotically mixed normal, locally asymptotically normal, method of simulated moments estimator, nonlinear equality and inequality restrictions, parameter restrictions, partially linear model, random coefficients regression, quasi-maximum likelihood estimator, restricted estimator, semiparametric estimator, stochastic trends, unit root model.

JEL Classification No.: C13.

1. Introduction

To obtain asymptotic distributions of estimators, 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 extremum estimators when the true parameter may be on the boundary of the parameter space. 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. We provide general high level assumptions under which the results hold and we verify these assumptions in seven examples.

The approach used here is to approximate the estimator objective function by a quadratic function rather than to rely on first order conditions. This approach was used by Chernoff (1954) to establish the asymptotic distribution of the likelihood ratio test in iid models with smooth likelihoods when the true parameter may be on a boundary. It has also been used by various authors to obtain the asymptotic properties of estimators when the true parameter is in the interior of the parameter space; see LeCam (1960), Jeganathan (1982), Pollard (1985), and Pakes and Pollard (1989) 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-maximum likelihood (QML), generalized method of moments (GMM), minimum distance, two-step, and semi-parametric estimators among others. The estimator objective function can be smooth or non-smooth, so that simulated method of moment (MSM) and least absolute deviation estimators are covered. This feature is obtained by using stochastic equicontinuity conditions, as in Pollard (1985), Pakes and Pollard (1989), and Andrews (1994a, b).

The results apply when the estimator objective 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), Wolak (1989), and Andrews (1996, 1997b), 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. Given our treatment of models with trends, our results are derived using a deterministic normalization matrix B_T rather than the more common scalar normalization of $T^{1/2}$.

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.

The estimator asymptotic distributions that are determined here are functions of multivariate normal distributions, in the case of models with no stochastic trends, and of multivariate Brownian motions, in the case of models with stochastic trends. The particular function depends on the shape of the parameter space. The parameter space is approximated locally to the true parameter by a convex cone Λ and it is this set that determines the function of the multivariate normal (or Brownian motion) that gives the asymptotic distribution of the estimator. For example, in a simple case where the parameter is restricted by a single nonlinear inequality restriction and the model does not contain stochastic trends, the asymptotic distribution is a half normal distribution.

More generally, the asymptotic distribution is given by the distribution of the vector that minimizes a certain quadratic function over the set Λ, where the coefficients of the quadratic function have a multivariate normal distribution (or are a function of a multivariate Brownian motion). We provide a closed form solution to this minimization problem for the fairly general case where the boundary of the parameter space is determined by linear and/or nonlinear equality and/or inequality restrictions. The asymptotic distributions can be simulated quite easily and quickly.

The asymptotic results derived here are useful for several purposes. First and foremost, they provide insight into the finite sample behavior of estimators when the true parameter is on the boundary of the parameter space. Second, they establish conditions under which the asymptotic distribution of the estimator of a sub-vector 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. Third, 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. Fourth, 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. Fifth, 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 (1997a).

Sixth, 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 (1997c). Seventh, 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 likelihood ratio-like test statistics for nonstandard testing problems; see Andrews (1997c). Eighth, 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. Ninth, the results can be used to determine the asymptotic behaviour 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.

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 non-negative 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 non-negativity 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. 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.

We now discuss the examples considered in this paper. For each of the seven examples, primitive sufficient conditions are given under which the high level assumptions of the paper are satisfied. The examples are chosen for their own intrinsic interest and to illustrate the various novel features of the general results.

The first example is a random coefficient regression model in which some random coefficient variances are zero. We consider the QML estimator. This example is one in which the estimator objective function cannot be defined in a neighborhood of the true parameter independently of sample size given that the regressors may be unbounded.

The second example is a regression model with nonlinear equality and/or inequality restrictions on the regression parameters. The errors and regressors are assumed to be iid. We consider the LS estimator. This example exhibits curved and kinked boundaries. We note that nonlinear inequality restrictions arise in utility, cost, and profit function estimation when convexity, quasi-convexity, concavity, or quasi-concavity is imposed at some point(s) in the sample; see Gallant and Golub (1984).

The third example is an augmented Dickey–Fuller regression model with the largest root restricted to be less than or equal to one and the time trend parameter restricted to be positive. The true parameter is taken to be a parameter with a unit root and a zero time trend parameter. (A Dickey–Fuller regression model is a univariate autoregressive model with at most one unit root that is written in terms of the level of the first lag of the time series and the differences of higher order lags of the time series.) This example illustrates the case of a model with deterministic and stochastic trends that is not a locally asymptotically mixed normal (LAMN) model.

The fourth example is a multinomial discrete response model estimated via a MSM estimator as in McFadden (1989) and Pakes and Pollard (1989). We consider the case where the model includes random coefficients, as in Hausman and Wise (1978), random effects, as in McFadden (1989), or measurement error, as in McFadden (1989), and the variances of some of these random terms are zero. This example illustrates the case of an estimator objective function that is discontinuous. A related class of GMM estimators of discrete response models with random coefficients that is used in the industrial organization literature is that of Berry (1994) and Berry, Levinsohn, and Pakes (1995). The results given in this paper also could be applied to determine the asymptotic distribution of these estimators when some of the random coefficient variances are zero.

The fifth example is a GARCH$(1, q^*)$ or IGARCH$(1, q^*)$ model in which the GARCH MA parameters are restricted to be non-negative and some of the true GARCH MA parameters equal zero. We consider the QML estimator. We note that asymptotic results for the QML estimator are not available in the literature for this model even when the true parameter is not on a boundary. (The results given below cover this case as well.) Verification of the assumptions for this model requires one to establish moment bounds for the quasi-log likelihood, quasi-score, and quasi-Hessian function, as in the GARCH(1,1) model. We do so along the lines of Lumsdaine (1996). More specifically, we generalize results given in Lee and Hansen (1994).

The sixth example is a partially linear model estimated by the semiparametric LS estimator of Robinson (1988), but subject to nonlinear equality and/or inequality constraints. This example illustrates the application of the general results to a semiparametric estimator and to an estimator that depends on a preliminary estimator. The example shows that one can derive the limit distribution for the estimator when the parameter is on the boundary with very little additional work beyond that which is needed to establish its distribution when the true parameter is in the interior of the parameter space. In particular, the hard parts of the verification of the assumptions follow directly from the results of Robinson (1988) with no additional work.

The seventh example is a regression model with nonlinear equality and/or inequality restrictions and regressors that are integrated of order one. We consider the LS estimator. This example illustrates the application of the general results to a LAMN model with stochastic trends. Note that Moon (1997) considers minimum distance estimators of a linear regression model with nonlinear equality constraints.

We note that the results of the paper apply to parametric two-step estimators, although none of our examples are of this form. For example, consider the Heckman two-step estimator of a sample selection model. If the correlation between the errors in the two equations of the model is generated by a common random effect, then the coefficient on the selection bias correction term in the main equation is necessarily non-negative. In this case, the regression parameter of the main equation is on the boundary of the parameter space when true random effect variance is zero, which corresponds to the case where there is no selectivity bias. Our results apply to this case.

The remainder of the paper is organized as follows. Section 2 introduces the seven examples. Section 3 considers the quadratic approximation of the estimator objective function, the B_T-consistency of the extremum estimator, and the application of the results of the section to the first three examples. Section 4 provides conditions under which the parameter space, suitably shifted and rescaled, can be locally approximated by a convex cone, provides an asymptotic representation of the extremum estimator, and discusses the first three examples. 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 and applies the results to the first three examples.

Section 7 provides several alternative sufficient conditions for the assumption of Section 3 that guarantees that the estimator objective function is approximately quadratic. It also verifies this assumption for the first three examples. Section 8 provides proofs of consistency for the first three examples.

Section 9 treats GMM and minimum distance estimators as special cases of extremum estimators and provides sufficient conditions for the quadratic approximation of the estimator objective function for these cases. It also applies these results to the Multinomial Discrete Response Model Example.

Section 10 applies the results of the paper to the GARCH$(1, q^*)$ Example. Section 11 applies the general results to the Partially Linear Regression Model Example. Section 12 applies the general results to the Regression with Restricted Parameters and Integrated Regressors Example. Section 13 provides proofs for the GARCH$(1, q^*)$ example that are omitted from Section 10 because of their length.

All limits below are taken "as $T \to \infty$" unless stated otherwise. Let "wp $\to$ 1" abbreviate "with probability that goes to one as $T \to \infty$." Let "for all $\gamma_T \to 0$" abbreviate "for all sequences of positive scalar constants $\{\gamma_T : T \geq 1\}$ for which $\gamma_T \to 0$." 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 $\partial\Lambda$ denote the boundary and $\mathrm{cl}(\Lambda)$ denote the closure of a set Λ. Let $a \odot b$ denote the Hadamard product (i.e., the element by element product) of two vectors a and b. 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

In this section we introduce several examples that are used to illustrate the general results given below. The examples cover models with iid, deterministically trending, and stochastically trending data. They cover parameter spaces with linear and nonlinear boundaries. The examples cover estimators that are ML, QML, LS, and GMM estimators. They cover objective functions that are smooth functions of the parameter as well as non-smooth functions.

2.1. Random Coefficient Regression

Example 1 is a random coefficient regression model. The variances of the random coefficients are necessarily greater than or equal to zero. We determine the asymptotic distribution of the Gaussian QML estimator when one or more of the random coefficient variances is equal to zero and, hence, the true parameter is on the boundary.

The model is

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

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 $X_{2t} \in R^{b_2}$. $D(\theta_1, \theta_2)$ is a diagonal matrix with the random coefficient variance parameters $(\theta_1', \theta_2')'$ 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, mean zero, variance one, uncorrelated errors (i.e., $E\eta_t = \mathbf{0}$, $E\eta_t\eta_t' = I_b$, $E\varepsilon_t = 0$, $E\varepsilon_t^2 = 1$, and $E\eta_t\varepsilon_t = \mathbf{0}$). The random variables $\{(Y_t, X_t, \varepsilon_t, \eta_t) : t \leq T\}$ are iid.

The parameter $\theta_1 \in R^p$ includes the random coefficient variances that are on the boundary (i.e., the true value of θ_1, θ_{10}, is $\mathbf{0}$). The parameter $\theta_2 \in R^{b_2}$ includes 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. It is positive (i.e., θ_{30} is positive). The parameter $\theta_4 \in R^b$ is the deterministic part of the regression coefficients. It is not on a boundary. The parameter $\theta_5 \in R$ is the intercept. It is not on a boundary.

2.2. Regression with Restricted Parameters

Example 2 is a linear regression model with equality and/or inequality restrictions on the regression parameters. We consider both linear and nonlinear restrictions. We determine the asymptotic distribution of the least squares (LS) estimator (subject to the restrictions) when some of the inequality restrictions are satisfied as equalities. In this case, the true regression parameter is on a linear or nonlinear boundary of the parameter space.

The model is

$$Y_t = X_t'\theta + \varepsilon_t, \qquad (2.2)$$

where $\{(Y_t, X_t) : t \leq T\}$ are observed dependent and regressor variables, respectively, $\{\varepsilon_t : t \leq T\}$ are iid unobserved errors, $X_t \in R^s$, $\theta \in R^s$, $Y_t \in R$, and $\varepsilon_t \in R$. The regressors are iid.

The parameter θ is assumed to satisfy the restrictions:

$$g_a(\theta) = \mathbf{0},\ g_b(\theta) \leq \mathbf{0}, \text{ and } h(\theta) \leq \mathbf{0}, \tag{2.3}$$

where $g_j(\cdot) \in R^{c_j}$ for $j = a, b$ and $h(\cdot)$ is vector-valued.

We suppose that the true parameter θ_0 satisfies the restrictions of (2.3) with

$$g_a(\theta_0) = \mathbf{0},\ g_b(\theta_0) = \mathbf{0}, \text{ and } h(\theta_0) < \mathbf{0}. \tag{2.4}$$

Thus, θ_0 is on the part of the boundary of the parameter space, Θ, that is determined by $g_a(\cdot)$ and $g_b(\cdot)$.

2.3. Dickey–Fuller Regression Model

Example 3 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

$$\begin{aligned} &Y_t = \theta_1 Y_{t-1} + \theta_2 t + \theta_3 + \Delta \bar{Y}_{t-1}' \theta_4 + \varepsilon_t, \text{ where} \\ &\Delta \bar{Y}_{t-1} = (\Delta Y_{t-1}, \Delta Y_{t-2}, ..., \Delta Y_{t-b})',\ \Delta Y_t = Y_t - Y_{t-1}, \\ &E(\varepsilon_t | \mathcal{F}_{t-1}) = 0 \text{ a.s.},\ E(\varepsilon_t^2 | \mathcal{F}_{t-1}) = \sigma^2 \text{ a.s.},\ \mathcal{F}_t = \sigma(\varepsilon_1, ..., \varepsilon_t), \end{aligned} \tag{2.5}$$

Y_t, ε_t, θ_1, θ_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 $\theta = (\theta_1, \theta_2, \theta_3, \theta_4')'$.

2.4. Multinomial Response Model

This example is a multinomial response model estimated by the method of simulated models (MSM). We use the same notation as McFadden (1989) and Pakes and Pollard (1989). The t-th individual has m alternatives to choose between. The ℓ-th choice is associated with a utility (or profit) of $Z_{t\ell}' h(\eta_t, \theta_0)$ for $\ell = 1, ..., m$, where $Z_{t\ell}$ is a b-vector of covariates for the ℓ-th choice and the t-th individual, $\theta_0 \in R^s$ is an unknown parameter, $\eta_t \in R^r$ is a vector of errors with known distribution, and $h(\cdot, \cdot)$

is a known R^b-valued function. The t-th individual chooses the alternative with the greatest utility. Thus, the response vector $d_t \in \{0,1\}^m$ can be written as

$$d_t := D(Z_t h(\eta_t, \theta_0)), \text{ where } Z_t := [Z_{t1} \ \cdots \ Z_{tm}]' \in R^{m \times b} \quad (2.6)$$

and $D(\cdot) : R^m \to \{0,1\}^m$ puts a one in the position of the largest component and a zero elsewhere. The choice is indicated by the component with a one. We assume that there is zero probability of a tie.

The random variables $\{(Z_t, \eta_t) : t = 1, ..., T\}$ are assumed to be iid.

By specifying different covariate vectors $Z_{t\ell}$, error vectors η_t, parameter vectors θ, and functional forms $h(\cdot,\cdot)$, one obtains a variety of different models. For example, if η_t has a standard multivariate normal distribution, then the model is in the family of probit models. We consider several such models. The first is a random coefficient probit model considered by Hausman and Wise (1978), and more recently, by Horowitz (1993) among others. In this case, the element of $h(\eta_t, \theta)$ that corresponds to a covariate in $Z_{t\ell}$ whose coefficient is random is of the form $\theta_i + \theta_j^{1/2}\eta_t \sim N(\theta_i, \theta_j)$. We consider the case where p (≥ 1) random coefficient variances are zero and, hence, the parameter θ_0 is on the boundary of the parameter space. For notational convenience, we order the parameters such that the first p elements of θ are these parameters.

Second, we consider a binary probit panel data model with autocorrelated errors and random effects, see McFadden (1989, Sec. 5). Let the first element of θ be the proportion of the total error variance due to the random effect. We consider the case where the true proportion is zero and, hence, the parameter θ_0 is on the boundary of the parameter space.

Third, we consider a probit model with measurement error on some covariates, see McFadden (1989, Sec. 6). In this case, some of the elements of θ correspond to the variance parameters of the measurement errors. We analyze the case where p (≥ 1) of these variance parameters equal zero and θ_0 is on a boundary. As above, we order the elements of θ such that these are the first p elements of θ.

All of the above cases can be treated simultaneously by analyzing the general multinomial response model introduced above with a parameter vector θ whose first p or more elements must be non-negative and whose true value θ_0 has its first p elements equal to zero.

2.5. GARCH Model

This example is a time series model for conditional heteroskedasticity, viz., the generalized autoregressive conditional heteroskedasticity GARCH$(1, q^*)$ model. We consider the case where the GARCH parameters are restricted to be non-negative, as in Bollerslev (1986). These restrictions guarantee that the conditional variance process is non-negative. These restrictions are not necessary, however, for the conditional variance process to be non-negative, see Nelson and Cao (1992). But, if they are true, their use leads to more efficient estimators.

We determine the asymptotic distribution of the Gaussian QML estimator when the GARCH–AR parameter is in $(0,1)$, the GARCH–MA parameter on the first lag is positive, and one or more of the other $q^* - 1$ GARCH–MA parameters is zero. In this case, the true parameter vector θ_0 is on the boundary. We allow for covariance stationary GARCH and integrated IGARCH models. (We note, however, that two drawbacks of the IGARCH results given below are that we only establish consistency of a "local" QML estimator and we impose a p^*-th order moment condition on the conditional variance process for some $0 < p^* < 1$. The former is analogous to the consistency results that are available in the literature for the IGARCH(1,1) model, see Lumsdaine (1996) and Lee and Hansen (1994). The latter has been shown to hold in the IGARCH$(1,1)$ model by Nelson (1990), but it has not been verified for IGARCH$(1,q^*)$ models with $q^* > 1$.)

Our results can be applied when none of the GARCH parameters is zero. In this case too our results are novel. The only consistency and asymptotic normality results in the literature for the QML estimator of GARCH and IGARCH models are for GARCH(1,1) and IGARCH (1,1) models with intercept but no regression function, see Lumsdaine (1996) and Lee and Hansen (1994). Our results allow for a regression function and more than one GARCH–MA parameter. We use the same methods as in the above papers to bound the requisite moments to obtain the laws of large numbers and central limit theorem needed for the asymptotic theory. More specifically, we extend various results in Lee and Hansen (1994) to cover the GARCH$(1,q^*)$ model with regression function.

2.6. Partially Linear Regression Model

This example is a partially linear regression model with nonlinear equality and/or inequality restrictions on the parameter vector. The partially linear model is a semiparametric model. We consider estimation of the finite dimensional parameter of the model using a semiparametric LS method introduced by Robinson (1988), who considers the partially linear regression model without any restrictions. We define the model and use the same assumptions as in Robinson (1988). In fact, we are able to use Robinson's results to establish the only difficult parts of the proof of the asymptotic distribution of the semiparametric LS estimator under nonlinear equality and/or inequality restrictions.

The model is

$$Y_t = X_t'\theta + \mu(Z_t) + \varepsilon_t, \tag{2.7}$$

where $\{(Y_t, X_t, Z_t) : t = 1, 2, ..., T\}$ are the observed random variables, θ is the unknown parameter to be estimated, $\mu(\cdot)$ is an unknown function, and ε_t is an unobserved error. As in Robinson (1988), we assume that (Y_t, X_t, Z_t) are iid across t, $E\varepsilon_t = 0$ and ε_t is independent of (X_t, Z_t).

The parameter θ is assumed to satisfy the same nonlinear restrictions as in (2.3) of Example 2. In addition, the true parameter θ_0 is assumed to satisfy (2.4). In this case, the parameter θ_0 is on the boundary of the parameter space.

2.7. Regression with Restricted Parameters and Integrated Regressors

This example is the same as the Regression with Restricted Parameters Example 2 except that the regressors are integrated of order one rather than iid.

3. Quadratic Approximation of the Objective Function and B_T-Consistency of the Extremum Estimator

3.1. Quadratic Approximation of the Objective Function

Let $\boldsymbol{Y}_T$ denote the data matrix when the sample size is T for $T = 1, 2, ...$ We consider an estimator objective function $\ell_T(\theta)$ that depends on $\boldsymbol{Y}_T$. Maximization of $\ell_T(\theta)$ over a parameter space $\Theta \subset R^s$ yields the estimator $\widehat{\theta}$ 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, or some other objective function.

Let θ_0 denote the true value of the parameter θ. (Or, if model misspecification renders a "true value" to be meaningless, then θ_0 can be defined to be the probability limit of $\widehat{\theta}$.) By assumption, $\theta_0 \in \text{cl}(\Theta)$.

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

$$(3.1)\quad \ell_T(\theta) := \ell_T(\theta_0) + D\ell_T(\theta_0)'(\theta - \theta_0) + \tfrac{1}{2}(\theta - \theta_0)' D^2\ell_T(\theta_0)(\theta - \theta_0) + R_T(\theta).$$

The remainder term $R_T(\theta)$ specifies the sense in which the expansion holds. When $\ell_T(\theta)$ is twice partially differentiable in 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 be twice partially differentiable in 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 note that by a theorem of Lebesgue (e.g., see Royden (1968, Cor. 5.2.5, p. 100)) a real function that has bounded variation on a closed interval is differentiable almost everywhere Lebesgue on that interval. Thus, there are often obvious candidates for $D\ell_T(\theta_0)$ and $D^2\ell_T(\theta_0)$ even when $\ell_T(\theta)$ is not pointwise twice differentiable on Θ.

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 Assumptions 2 and 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 non-diagonal matrix. By appropriate choice of M, one may be able to obtain a block diagonal normalized "quasi-information" matrix $\mathcal{J}_T$, defined below. This yields a simplified expression for the asymptotic distribution of the extremum estimator. The GARCH$(1, q^*)$ Example is one in which we take $B_T = T^{1/2}M$ for $M \neq I_s$.

With trending data, B_T is always more complicated than $T^{1/2}I_s$. For example, it is diagonal with diagonal elements $T^{1/2}$ and T in a linear model with stationary regressors and integrated exogenous regressors of order one. In a dynamic regression model with stationary regressors, integrated regressors, and a time trend, B_T is an asymmetric matrix of the form $B_T = \Upsilon_T M$, where Υ_T is diagonal, $\lambda_{\min}(\Upsilon_T) \to \infty$, and M is nonsingular. Such cases are permitted here.

Let

$$\mathcal{J}_T := -B_T^{-1'} D^2 \ell_T(\theta_0) B_T^{-1} \text{ and } Z_T := \mathcal{J}_T^{-1} B_T^{-1'} D\ell_T(\theta_0), \tag{3.2}$$

where $B_T^{-1'}$ denotes $(B_T^{-1})'$. The quadratic expansion of (3.1) can be rewritten as

$$\begin{aligned} &\ell_T(\theta) = \ell_T(\theta_0) + \tfrac{1}{2} Z_T' \mathcal{J}_T Z_T - \tfrac{1}{2} q_T(B_T(\theta - \theta_0)) + R_T(\theta), \text{ where} \\ &q_T(\lambda) := (\lambda - Z_T)' \mathcal{J}_T (\lambda - Z_T) \text{ for } \lambda \in R^s. \end{aligned} \tag{3.3}$$

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

Assumption 1. *For all* $\gamma_T \to 0$, $\sup_{\theta \in \Theta: ||\theta - \theta_0|| \leq \gamma_T} |R_T(\theta)|/(1 + ||B_T(\theta - \theta_0)||)^2 = o_p(1)$.

Assumption 2. $B_T^{-1'} D\ell_T(\theta_0) = O_p(1)$.

Assumption 3. $\mathcal{J}_T$ *is symmetric* wp $\to 1$, $\lambda_{\max}(\mathcal{J}_T) = O_p(1)$, *and* $\lambda_{\min}^{-1}(\mathcal{J}_T) = O_p(1)$.

In Section 7 we give two sets of sufficient conditions for Assumption 1. The first relies on the existence of left and/or right partial derivatives of $\ell_T(\theta)$ and the second relies on a stochastic differentiability condition that generalizes that of Pollard (1985). The first verifies Assumption 1 with the term $1/(1 + ||B_T(\theta - \theta_0)||)^2$ replaced by $1/||B_T(\theta - \theta_0)||^2$ and the second with it replaced by $1/||B_T(\theta - \theta_0)||$. Neither condition requires the parameter space Θ or the domain of $\ell_T(\theta)$ to include a neighborhood of θ_0.

In addition, in Section 9 we provide a sufficient condition for Assumption 1 for the special case where $\ell_T(\theta)$ is of the GMM or minimum distance form. This condition generalizes that of Pakes and Pollard (1989).

The idea of incorporating a term like the $1/(1 + ||B_T(\theta - \theta_0)||)^2$ term in Assumption 1 was introduced by Huber (1967).

For the time being, we illustrate the plausibility of Assumption 1 by considering the standard case in which $\ell_T(\theta)$ is twice differentiable in a neighborhood of θ_0 with first and second derivatives given by $D\ell_T(\theta)$ and $D^2\ell_T(\theta)$ respectively. By a two term Taylor expansion, (3.1) holds with

$$
\begin{aligned}
|R_T(\theta)| &= \left|\tfrac{1}{2}(B_T(\theta-\theta_0))'[B_T^{-1\prime}(D^2\ell_T(\theta^\dagger) - D^2\ell_T(\theta_0))B_T^{-1}]B_T(\theta-\theta_0)\right| \\
&\leq ||B_T(\theta-\theta_0)||^2 \cdot ||B_T^{-1\prime}(D^2\ell_T(\theta^\dagger) - D^2\ell_T(\theta_0))B_T^{-1}||, \qquad (3.4)
\end{aligned}
$$

where $\theta^\dagger$ lies between θ and θ_0. Hence, Assumption 1 holds in this case if

$$
\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T} ||B_T^{-1\prime}(D^2\ell_T(\theta) - D^2\ell_T(\theta_0))B_T^{-1}|| = o_p(1) \qquad (3.5)
$$

for all $\gamma_T \to 0$. Note that in verifying Assumption 1 in this case we have used the "$||B_T(\theta-\theta_0)||$" term rather than the "1" term in the "$1/(1+||B_T(\theta-\theta_0)||)^2$" multiplicand of Assumption 1. Condition (3.5) holds by a uniform law of large numbers for $B_T^{-1\prime} D^2\ell_T(\theta)B_T^{-1}$ over some neighborhood of θ_0 plus continuity of the limit function at θ_0 or by the existence of third derivatives of $\ell_T(\theta)$ that are $O_p(1)$ uniformly over a neighborhood of θ_0.

Next, we note that the form of Assumption 1 is designed to allow one to easily replace the objective function $\ell_T(\theta)$ by a more tractable function, say $\mathcal{L}_T(\theta)$, that is a close approximation to $\ell_T(\theta)$. For example, in the GARCH$(1, q^*)$ Example, $\ell_T(\theta)$ is a sum of quasi-log likelihood contributions that depends on initial conditions and, hence, is not stationary and ergodic. We can define a more tractable function $\mathcal{L}_T(\theta)$ to be the stationary and ergodic analogue of $\ell_T(\theta)$ that replaces the initial conditions by terms that depend on the infinite history of the process. Now, suppose

$$
\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T} |\ell_T(\theta) - \ell_T(\theta_0) - \mathcal{L}_T(\theta) + \mathcal{L}_T(\theta_0)| = o_p(1) \qquad (3.6)
$$

for all $\gamma_T \to 0$. Also, suppose $\mathcal{L}_T(\theta)$ has an expansion of the form

$$
\mathcal{L}_T(\theta) = \mathcal{L}_T(\theta_0) + D\mathcal{L}_T(\theta_0)'(\theta-\theta_0) + \tfrac{1}{2}(\theta-\theta_0)'D^2\mathcal{L}_T(\theta_0)(\theta-\theta_0) + R_T^*(\theta), \qquad (3.7)
$$

where $R_T^*(\theta)$ satisfies Assumption 1 with $R_T(\theta)$ replaced by $R_T^*(\theta)$. Then, $\ell_T(\theta)$ satisfies (3.1) with

$$
\begin{aligned}
D\ell_T(\theta_0) &= D\mathcal{L}_T(\theta_0),\ D^2\ell_T(\theta_0) = D^2\mathcal{L}_T(\theta_0), \text{ and} \\
R_T(\theta) &= R_T^*(\theta) + (\ell_T(\theta) - \ell_T(\theta_0) - \mathcal{L}_T(\theta) + \mathcal{L}_T(\theta_0)). \qquad (3.8)
\end{aligned}
$$

Assumption 1 holds for $\ell_T(\theta)$ in this case by (3.6) and (3.7). Note that, in this case, we use the "1" term rather than the "$||B_T(\theta-\theta_0)||$" term in the "$1/(1+||B_T(\theta-\theta_0)||)^2$" multiplicand of Assumption 1 to bound the term in parentheses in (3.8).

Assumption 2 is implied by the convergence in distribution of the normalized score function in quasi-log likelihood cases. In such cases, Assumption 2 usually follows from a central limit theorem (CLT) in models without stochastic trends and from an invariance principle in models with stochastic trends. In GMM cases, Assumption 2 usually follows from a CLT and one or more convergence in probability results. In other cases, such as with Han's (1987) maximum rank correlation estimator (see

Sherman (1993)), Assumption 2 follows from a CLT for U-statistics. In minimum distance and other cases that rely on preliminary estimators, verification of Assumption 2 requires asymptotic results for the preliminary estimators. Results already in the literature often can be used in such cases.

Assumption 3 allows the normalized "information" matrix $\mathcal{J}_T$ to be random even in the limit as $T \to \infty$. 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), $\mathcal{J}_T$ converges to a non-stochastic limit $\mathcal{J}$. In this case, one can take $\mathcal{J}_T$ to be the non-stochastic limit $\mathcal{J}$ in the quadratic expansion of (3.3) and the remainder term $R_T(\theta)$ can absorb the difference without requiring any adjustment in Assumption 1 (due to the "$||B_T(\theta - \theta_0)||$" term that appears in Assumption 1). Thus, a sufficient condition for Assumption 3, that is applicable in models with no trends or with deterministic trends, is the following:

Assumption 3*. *$\mathcal{J}_T$ is non-random and does not depend on T. $\mathcal{J}$ $(= \mathcal{J}_T)$ is symmetric and $\lambda_{\min}(\mathcal{J}) > 0$.*

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 un-superscripted assumption.

3.2. B_T-Consistency of the Extremum Estimator

Next, we define the extremum estimator $\widehat{\theta}$ of θ. To ease the computational burden, to circumvent the question of existence, and to ease the verification of assumptions, we only require that $\ell_T(\widehat{\theta})$ is within $o_p(1)$ of the global maximum of $\ell_T(\theta)$ over $\theta \in \Theta$, rather than the exact global maximum. By definition, $\hat{\theta} \in \Theta$ and

$$\ell_T(\widehat{\theta}) = \sup_{\theta \in \Theta} \ell_T(\theta) + o_p(1). \tag{3.9}$$

We assume consistency of $\widehat{\theta}$ for the true value θ_0:

Assumption 4. $\widehat{\theta} = \theta_0 + o_p(1)$.

A sufficient condition for Assumption 4 that often holds when the data does not involve trending variables is the following:

Assumption 4*. (a) *For some function $\ell(\theta) : \Theta \to R$, $\sup_{\theta \in \Theta} |T^{-1}\ell_T(\theta) - \ell(\theta)| \xrightarrow{p} 0$.* (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 $S(\theta_0, \varepsilon)$.*

Assumption 4*(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 4*(b) is an asymptotic identification condition. Sufficient conditions for Assumption 4*(b), which we call Assumption 4*(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.

The sufficiency of Assumption 4* for Assumption 4 is well-known.[2]

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.

Let $\widehat{\theta}_q$ 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, $\widehat{\theta}_q$ satisfies $\widehat{\theta}_q \in \mathrm{cl}(\Theta)$ and

$$q_T(B_T(\widehat{\theta}_q-\theta_0)) = \inf_{\theta\in\Theta} q_T(B_T(\theta-\theta_0)) + o_p(1). \tag{3.10}$$

Note that

$$\begin{aligned} &\inf_{\theta\in\Theta} q_T(B_T(\theta-\theta_0)) = \inf_{\lambda\in B_T(\Theta-\theta_0)} q_T(\lambda), \text{ where} \\ &B_T(\Theta-\theta_0) := \{\lambda \in R^s : \lambda = B_T(\theta-\theta_0) \text{ for some } \theta\in\Theta\}. \end{aligned} \tag{3.11}$$

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

Theorem 1. *Suppose Assumptions* 1–4 *hold. Then,*
(a) $B_T(\widehat{\theta}-\theta_0) = O_p(1)$,
(b) $B_T(\widehat{\theta}_q-\theta_0) = O_p(1)$,
(c) $\ell_T(\widehat{\theta}) = \ell_T(\theta_0) + \frac{1}{2}Z_T'\mathcal{J}_T Z_T - \frac{1}{2}q_T(B_T(\widehat{\theta}-\theta_0)) + o_p(1)$,
(d) $\ell_T(\widehat{\theta}_q) = \ell_T(\theta_0) + \frac{1}{2}Z_T'\mathcal{J}_T Z_T - \frac{1}{2}q_T(B_T(\widehat{\theta}_q-\theta_0)) + o_p(1)$,
(e) $\ell_T(\widehat{\theta}) = \ell_T(\widehat{\theta}_q) + o_p(1)$,
(f) $q_T(B_T(\widehat{\theta}-\theta_0)) = q_T(B_T(\widehat{\theta}_q-\theta_0)) + o_p(1)$, *and*
(g) $\ell_T(\widehat{\theta}) = \ell_T(\theta_0) + \frac{1}{2}Z_T'\mathcal{J}_T Z_T - \frac{1}{2}q_T(B_T(\widehat{\theta}_q-\theta_0)) + o_p(1)$.

Comments. 1. Parts (a) and (b) hold even if $o_p(1)$ is replaced by $O_p(1)$ in (3.9) and (3.10).

2. Part (b) only requires Assumptions 2 and 3 and that θ_0 is in the closure of Θ (which is implied by Assumption 4).

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

3.3. Examples (Continued)

In this section, we introduce the objective function $\ell_T(\theta)$ and the parameter space Θ for the first three of the examples of Section 2. We also specify assumptions that are sufficient for Assumptions 1–4 and verify Assumptions 2 and 3 for each of these examples. For ease of reading, the verification of Assumptions 1 and 4 for each example is relegated to Sections 7 and 8 below respectively.

3.3.1. Random Coefficient Regression

In Example 1, we consider the Gaussian QML estimator. The Gaussian quasi-log likelihood function is

$$\begin{aligned}\ell_T(\theta) := &-\tfrac{T}{2}\ln(2\pi) - \tfrac{1}{2}\sum_{t=1}^{T}\ln(\theta_3 + X_t' D(\theta_1,\theta_2)X_t)\\ &-\tfrac{1}{2}\sum_{t=1}^{T}(Y_t - \theta_5 - X_t'\theta_4)^2/(\theta_3 + X_t' D(\theta_1,\theta_2)X_t).\end{aligned} \tag{3.12}$$

The true parameter vector θ_0 is

$$\theta_0 := (\theta_{10}', \theta_{20}', \theta_{30}, \theta_{40}', \theta_{50})' = (\mathbf{0}', \theta_{20}', \theta_{30}, \theta_{40}', \theta_{50})', \tag{3.13}$$

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 non-negative and that bounds θ_3 away from zero:

$$\begin{aligned}\Theta := \{&\theta \in R^s : \theta = (\theta_1', \theta_2', \theta_3, \theta_4', \theta_5)',\ \theta_1 \geq 0,\ \theta_2 \geq 0,\ \theta_3 \geq c,\\ &||\theta_j|| \leq M_j\ \forall j \leq 5\}\end{aligned} \tag{3.14}$$

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

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

$$\begin{aligned}&X_t := (X_{t1}, ..., X_{tb})',\ X_t^2 := (X_{t1}^2, ..., X_{tb}^2)',\\ &W_t := (X_t', 1)',\ W_t^2 := (X_t^{2'}, 1)',\\ &\text{res}_t(\theta) := Y_t - \theta_5 - X_t'\theta_4, \text{ and } \text{var}_t(\theta) := \theta_3 + X_t' D(\theta_1,\theta_2)X_t.\end{aligned} \tag{3.15}$$

Define

$$\begin{aligned}&D\ell_T(\theta_0) := \sum_{t=1}^{T}\begin{pmatrix}\dfrac{\text{res}_t^2(\theta_0) - \text{var}_t(\theta_0)}{2\text{var}_t^2(\theta_0)}W_t^2\\ \dfrac{\text{res}_t(\theta_0)}{\text{var}_t(\theta_0)}W_t\end{pmatrix},\ D^2\ell_T(\theta_0) := -T\mathcal{J},\\ &\mathcal{J} := \mathcal{J}_T := \begin{bmatrix}\frac{1}{2}EW_t^2W_t^{2'}/\text{var}_t^2(\theta_0) & \mathbf{0}\\ \mathbf{0} & EW_tW_t'/\text{var}_t(\theta_0)\end{bmatrix},\ B_T := T^{1/2}I_s, \text{ and}\\ &Z_T := \mathcal{J}^{-1}T^{-1/2}D\ell_T(\theta_0).\end{aligned} \tag{3.16}$$

With these definitions, the quadratic approximations of (3.1) and (3.3) hold with a remainder term $R_T(\theta)$ that satisfies Assumption 1 under the assumptions given above and the moment conditions given below. The latter is shown in Section 7.

Assumption 2 holds for Example 1 by the CLT provided

$$E\varepsilon_t^4 < \infty,\ E||\eta_t||^4 < \infty, \text{ and } E||X_t||^8 < \infty. \tag{3.17}$$

Assumption 3 holds for Example 1 provided

$$EW_t^2 W_t^{2\prime}/\mathrm{var}_t^2(\theta_0) > 0 \text{ and } EW_t W_t'/\mathrm{var}_t(\theta_0) > 0, \tag{3.18}$$

where "> 0" denotes "is positive definite."

Assumption 4 is verified in Section 8.

3.3.2. Regression with Restricted Parameters

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

$$\ell_T(\theta) := -\tfrac{1}{2}\sum_{t=1}^{T}(Y_t - X_t'\theta)^2. \tag{3.19}$$

The parameter space Θ is given by

$$\Theta := \{\theta \in R^s : g_a(\theta) = \mathbf{0},\ g_b(\theta) \leq \mathbf{0},\ h(\theta) \leq \mathbf{0}\}. \tag{3.20}$$

The quadratic approximation of (3.1) and (3.3) holds with

$$D\ell_T(\theta_0) := \sum_{t=1}^{T}\varepsilon_t X_t,\ D^2\ell_T(\theta_0) := -\sum_{t=1}^{T} X_t X_t', \text{ and } R_T(\theta) = 0. \tag{3.21}$$

Assumption 1 holds because $R_T(\theta) = 0$.

The errors and regressors $\{(\varepsilon_t, X_t) : t \leq T\}$ are iid with

$$E\varepsilon_t X_t = \mathbf{0},\ E||\varepsilon_t X_t||^2 < \infty, \text{ and } EX_t X_t' > 0. \tag{3.22}$$

In this case,

$$B_T = T^{1/2} I_s \text{ and } \mathcal{J}_T = T^{-1}\sum_{t=1}^{T} X_t X_t'. \tag{3.23}$$

Assumption 2 holds by the CLT for iid mean zero finite variance random variables using (3.22). Assumption 3 holds by the LLN for iid random variables with finite mean using (3.22).

Assumption 4 is verified under the assumptions above in Section 8 below.

3.3.3. Dickey–Fuller Regression

In example 3, we consider the LS estimator. The objective function is

$$\ell_T(\theta) := -\tfrac{1}{2}\sum_{t=1}^{T}(Y_t - X_t'\theta)^2, \text{ where } X_t := (Y_{t-1}, t, 1, \Delta\bar{Y}_{t-1}')'. \tag{3.24}$$

The parameter space Θ is given by

$$\begin{aligned}\Theta := \{\theta \in R^s : -1 < \theta_1 \le 1,\ \theta_2 \ge 0,\ g(z) := 1 - \theta_{41}z - \cdots - \theta_{4b}z^b \\ \text{has roots outside the unit circle, where } \theta_4 := (\theta_{41}, ..., \theta_{4b})'\}.\end{aligned} \tag{3.25}$$

The true parameter vector θ_0 corresponds to a unit root model with non-negative drift:

$$\theta_0 := (\theta_{10}, \theta_{20}, \theta_{30}, \theta_{40}')' = (1, 0, \theta_{30}, \theta_{40}')', \tag{3.26}$$

where $\theta_{30} \ge 0$ and θ_{40} has characteristic equation with roots outside the unit circle. Note that the latter implies that $1 - \mathbf{1}'\theta_{40} > 0$, where $\mathbf{1} := (1, ..., 1)' \in R^b$. 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 $\sigma^2 > 0$. Given the definitions of θ_0 and Θ, the parameters θ_{10} and θ_{20} are on the boundary of the parameter space and θ_{30} and θ_{40} are not on the boundary.

The quadratic approximation of (3.1) and (3.3) holds with

$$D\ell_T(\theta_0) := \sum_{t=1}^{T}\varepsilon_t X_t,\ D^2\ell_T(\theta_0) := -\sum_{t=1}^{T}X_tX_t',\ R_T(\theta) := 0,\ B_T := \Upsilon_T M,$$

$$M := \begin{bmatrix} 1 & 0 & 0 & \mathbf{0}' \\ \mu_0 & 1 & 0 & \mathbf{0}' \\ -\mu_0 & 0 & 1 & \mu_0\mathbf{1}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I_b \end{bmatrix},\ L := M^{-1'} = \begin{bmatrix} 1 & -\mu_0 & \mu_0 & \mathbf{0}' \\ 0 & 1 & 0 & \mathbf{0}' \\ 0 & 0 & 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & -\mu_0\mathbf{1} & I_b \end{bmatrix},$$

$\mu_0 := \theta_{30}/(1 - \mathbf{1}'\theta_{40})$, $\Upsilon_T := \mathrm{Diag}(T, T^{3/2}, T^{1/2}, ..., T^{1/2})$, and

$$\mathcal{J}_T := -B_T^{-1'}D^2\ell_T(\theta_0)B_T^{-1} = \Upsilon_T^{-1}\sum_{t=1}^{T}LX_t(LX_t)'\Upsilon_T^{-1}. \tag{3.27}$$

To verify Assumptions 2 and 3, we impose the following mild tail condition on the errors: For some random variable ε, some $0 < c < \infty$, and some $\eta > 0$,

$$P(|\varepsilon_t| \ge x) \le cP(|\varepsilon| \ge x)\ \ \forall x > 0 \text{ and } E|\varepsilon|^{2+\eta} < \infty. \tag{3.28}$$

Under the assumptions given, we have

$$\left(B_T^{-1'}D\ell_T(\theta_0), \mathcal{J}_T\right) := \left(\Upsilon_T^{-1}\sum_{t=1}^{T}\varepsilon_t LX_t, \Upsilon_T^{-1}\sum_{t=1}^{T}LX_t(LX_t)'\Upsilon_T\right) \xrightarrow{d} (G, \mathcal{J}), \text{ where}$$

$$G := \begin{pmatrix} \frac{1}{2}\sigma\lambda(W^2(1)-1) \\ \sigma(W(1) - \int_0^1 W(r)dr) \\ \sigma W(1) \\ G_4 \end{pmatrix}, \; G_4 \sim N(\mathbf{0}, V),$$

$$\mathcal{J} := \begin{pmatrix} \lambda^2 \int_0^1 W^2(r)dr & \lambda \int_0^1 rW(r)dr & \lambda \int_0^1 W(r)dr & \mathbf{0}' \\ \lambda \int_0^1 rW(r)dr & 1/3 & 1/2 & \mathbf{0}' \\ \lambda \int_0^1 W(r)dr & 1/2 & 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & V \end{pmatrix},$$

$$\lambda := \sigma/(1 - \mathbf{1}'\theta_{40}), \; \gamma_j = \mathrm{Cov}(\Delta Y_t, \Delta Y_{t-j}) \quad \forall j = 0, ..., b-1,$$

$$V := \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{b-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{b-2} \\ \vdots & & \vdots & & \vdots \\ \gamma_{b-1} & \gamma_{b-2} & \gamma_{b-3} & \cdots & \gamma_0 \end{pmatrix}, \tag{3.29}$$

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 $\mathcal{J}$ is nonsingular with probability one. Thus, Assumptions 2 and 3 hold.

The proof of (3.29) 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 given in Theorem 3.15 of Phillips and Solo (1992, p. 983) in place of the invariance principle utilized by Hamilton.

Assumption 4 is verified in Section 8.

3.4. Proofs

Proof of Theorem 1. Let $\kappa_T := \mathcal{J}_T^{1/2} B_T(\widehat{\theta} - \theta_0)$. θ_0 is in the closure of Θ (by Assumption 4). Thus, by (3.1), (3.2), (3.9), and Assumptions 1–4,

$$\begin{aligned} o_p(1) &\leq \ell_T(\widehat{\theta}) - \ell_T(\theta_0) \\ &= \kappa_T' \mathcal{J}_T^{1/2} Z_T - \tfrac{1}{2}||\kappa_T||^2 + R_T(\widehat{\theta}) \\ &= O_p(||\kappa_T||) - \tfrac{1}{2}||\kappa_T||^2 + \left(1 + ||\mathcal{J}_T^{-1/2}\kappa_T||\right)^2 o_p(1) \\ &= O_p(||\kappa_T||) - \tfrac{1}{2}||\kappa_T||^2 + o_p(||\kappa_T||) + o_p(||\kappa_T||^2) + o_p(1). \end{aligned} \tag{3.30}$$

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,

$$(||\kappa_T|| - \xi_T)^2 \leq \xi_T^2 + o_p(1) = O_p(1). \tag{3.31}$$

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

Let $\kappa_{qT} := \mathcal{J}_T^{1/2} B_T(\widehat{\theta}_q - \theta_0)$. By (3.10) and Assumptions 2 and 3, we have

$$\begin{aligned} ||\kappa_{qT} - \mathcal{J}_T^{1/2} Z_T||^2 &= q_T(B_T(\widehat{\theta}_q - \theta_0)) \leq q_T(0) + o_p(1) \\ &= Z_T' \mathcal{J}_T Z_T + o_p(1) = O_p(1). \end{aligned} \tag{3.32}$$

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

Parts (c) and (d) hold by (3.3), Assumption 1, and parts (a) and (b).

Parts (e) and (f) hold by parts (c) and (d), (3.9), and (3.10):

$$\begin{aligned} o_p(1) &\leq \ell_T(\widehat{\theta}) - \ell_T(\widehat{\theta}_q) \\ &= \tfrac{1}{2} q_T(B_T(\widehat{\theta}_q - \theta_0)) - \tfrac{1}{2} q_T(B_T(\widehat{\theta} - \theta_0)) + o_p(1) \leq o_p(1). \end{aligned} \tag{3.33}$$

Part (g) holds by parts (c) and (f). □

4. The Parameter Space

This section provides conditions on the parameter space under which we can derive the asymptotic distribution of the extremum estimator $\widehat{\theta}$.

4.1. Local Approximation to the Shifted and Rescaled Parameter Space

It is apparent from Theorem 1(a) that the asymptotic distribution of $\widehat{\theta}$ depends on the features of the parameter space Θ only near θ_0. In particular, we find that the asymptotic distribution of $\widehat{\theta}$ 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 $b_T \to \infty$.

If Θ includes a neighborhood of θ_0, then

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

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

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 some 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 $j = a, b$.

To provide sufficient conditions for (4.1) when Λ is a cone, we need to introduce some definitions. Define the distance between a point $y \in R^s$ and a set $\Lambda \subset R^s$ by

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

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

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

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 $\varepsilon_T \to 0$ as $T \to \infty$.

If the sets $\{\Phi_T : T \geq 1\}$ do not depend on T in the definition above, then we say that the set Φ $(:= \Phi_T)$ is locally approximated (at the origin) by a cone $\Lambda \subset R^s$ if (4.3) holds.

Assumption 5. *For some sequence of scalar constants* $\{b_T : T \geq 1\}$ *for which* $b_T \to \infty$ *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.

Sufficient conditions for Assumption 5 are given below. Assumption 5 allows for linear, kinked, and curved boundaries.

Note that 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 $\lambda_{\min}(B_T) \to \infty$. 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 $\varepsilon > 0$.

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

Comment. The requirement in Assumption 5 that $b_T \leq c\lambda_{\min}(B_T)$ for some $0 < c < \infty$ is not actually needed for Lemma 1 to hold. It is imposed because it is needed for Lemma 3 below, which yields a convenient method of verifying Assumption 5 when $\Theta - \theta_0$ is a product set, and because it sacrifices little or no generality of the results.

Theorem 1(g) and Lemma 1 give

$$\text{(4.4)} \qquad \ell_T(\widehat{\theta}) = \ell_T(\theta_0) + \tfrac{1}{2} Z_T' \mathcal{J}_T Z_T - \tfrac{1}{2} \inf_{\lambda \in \Lambda} q_T(\lambda) + o_p(1).$$

The two quadratic forms on the right-hand side of (4.4) can be re-expressed as a single quadratic form as follows.

Let $\widehat{\lambda}_T$ be a minimizer of $q_T(\lambda)$ over $\text{cl}(\Lambda)$. By definition, $\widehat{\lambda}_T \in \text{cl}(\Lambda)$ and

$$\text{(4.5)} \qquad q_T(\widehat{\lambda}_T) = \inf_{\lambda \in \Lambda} q_T(\lambda).$$

The random variable $\widehat{\lambda}_T$ is a version of the projection of Z_T onto the cone Λ with respect to the norm $||\lambda||_T = (\lambda' \mathcal{J}_T \lambda)^{1/2}$; see Perlman (1969, Sec. 4). If Λ is convex, $\widehat{\lambda}_T$ is uniquely defined.

For example, if Λ is a linear subspace of R^s, as occurs with linear or nonlinear equality constraints, then $\widehat{\lambda}_T$ is a linear function of $Z_T : \widehat{\lambda}_T = P_{T\Lambda} Z_T$, 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 $P_{T\Lambda} := I_s - \mathcal{J}_T^{-1}\Gamma'(\Gamma\mathcal{J}_T^{-1}\Gamma')^{-1}\Gamma$. (We note that for most of our examples, Λ is not a linear subspace.)

Whether or not Λ is convex, the following orthogonality property holds

$$\widehat{\lambda}_T' \mathcal{J}_T(\widehat{\lambda}_T - Z_T) = 0; \tag{4.6}$$

see Perlman (1969, Lem. 4.1). Some algebra then gives

$$Z_T' \mathcal{J}_T Z_T - \inf_{\lambda \in \Lambda} q_T(\lambda) = \widehat{\lambda}_T' \mathcal{J}_T \widehat{\lambda}_T. \tag{4.7}$$

(The right-hand side of (4.7) is uniquely defined whether or not $\widehat{\lambda}_T$ is.)

Combining (4.4) and (4.7) shows that under Assumptions 1–5,

$$\ell_T(\widehat{\theta}) = \ell_T(\theta_0) + \tfrac{1}{2}\widehat{\lambda}_T' \mathcal{J}_T \widehat{\lambda}_T + o_p(1). \tag{4.8}$$

4.2. An Asymptotic Representation of the Extremum Estimator

We now show that $B_T(\widehat{\theta} - \theta_0)$ is asymptotically equivalent to $\widehat{\lambda}_T$, where $\widehat{\lambda}_T$ is defined in (4.5), provided Λ is convex. If Λ is convex, then $\widehat{\lambda}_T$ is uniquely defined. If Λ is not convex, e.g., if Λ consists of the union of the positive and negative orthants, then $\widehat{\lambda}_T$ could take on multiple values that are not of distance $o_p(1)$ from each other. In this case, we do not have asymptotic equivalence of $B_T(\widehat{\theta} - \theta_0)$ and $\widehat{\lambda}_T$.

We assume:

Assumption 6. Λ *is convex.*

Assumption 6 holds for all the examples of Section 2.

We note two characterizations of convex cone. A set Λ is a convex cone iff Λ is closed under addition and positive scalar multiplication iff Λ contains all finite linear combinations with positive coefficients of its elements; see Rockafellar (1970, Thm. 2.6 and Cor. 2.6.1, p. 14).

Theorem 2. *Suppose Assumptions* 1–6 *hold. Then,* $B_T(\widehat{\theta} - \theta_0) = \widehat{\lambda}_T + o_p(1)$.

Comments. 1. Theorem 2 is used to determine the asymptotic distribution of $\widehat{\theta}$, because it is straightforward to obtain the asymptotic distribution of $\widehat{\lambda}_T$ by the continuous mapping theorem provided $(Z_T, \mathcal{J}_T)$ converges in distribution to some limit.

2. The proof of Theorem 2 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 1(f) and Lemma 1, $q_T(B_T(\widehat{\theta}-\theta_0)) = \inf_{\lambda \in \Lambda} q_T(\lambda) + o_p(1)$. If $\Lambda = R^s$, then $\widehat{\lambda}_T = Z_T$, $\inf_{\lambda \in \Lambda} q_T(\lambda) = 0$, and

$$q_T(B_T(\widehat{\theta}-\theta_0)) = (B_T(\widehat{\theta}-\theta_0) - \widehat{\lambda}_T)'\mathcal{J}_T(B_T(\widehat{\theta}-\theta_0) - \widehat{\lambda}_T) = o_p(1). \tag{4.9}$$

In view of Assumption 3, this gives the result of Theorem 2. When $\Lambda \neq R^s$, the proof of Theorem 2 is more difficult.

4.3. Sufficient Conditions for Assumption 5

We now give several easily verifiable sufficient conditions for Assumption 5. 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 $\Gamma \cap C(\mathbf{0}, \varepsilon) = \Lambda \cap C(\mathbf{0}, \varepsilon)$.

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 $b_T \to \infty$.*

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

$$\Theta \cap C(\theta_0, \varepsilon) = \{\theta \in R^s : \theta - \theta_0 \in \mathop{\times}_{j=1}^{s} I_j,\ \theta \in C(\theta_0, \varepsilon)\} \text{ and}$$

$$\Lambda := \mathop{\times}_{j=1}^{s} I_j, \text{ where } I_j = \{0\},\ R,\ R^+, \text{ or } R^- \text{ for } j \leq s. \tag{4.10}$$

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

$$\Theta := \{\theta \in R^s : \Gamma_a \theta = r_1,\ \Gamma_b \theta \leq r_2,\ ||\theta|| \leq c < \infty\}, \tag{4.11}$$

$\Gamma_a \theta_0 = r_1$, 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,

$$\Lambda := \{\lambda \in R^s : \Gamma_a \lambda = \mathbf{0},\ \Gamma_{b1} \lambda \leq \mathbf{0}\}, \tag{4.12}$$

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, $B_T = T^{1/2} I_s$. The Regression with Restricted Parameters and Integrated Regressors Example provides one example, however, where it may hold with $B_T = T I_s$.

Assumption 5* is not applicable in dynamic models with deterministic and/or stochastic trends, such as in the Dickey–Fuller Regression Example 3, because $B_T \neq b_T I_s$ in these models. Assumption 5* also is not applicable in the GARCH$(1, q^*)$ Example for which $B_T = T^{1/2} M$ with M non-diagonal. For such cases, we introduce a more general sufficient condition for Assumption 5 that allows for a non-diagonal B_T matrix.

The next condition uses a definition of the (maximal) distance between two cones, which we now define. 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 Λ_1 and Λ_2, we define

$$\text{dist}_c(\Lambda_1, \Lambda_2) := \sup_{\lambda_1 \in \Lambda_1} \inf_{\lambda_2 \in \Lambda_2} ||\lambda_1/||\lambda_1|| - \lambda_2/||\lambda_2||||. \tag{4.13}$$

Note that $\text{dist}_c(\Lambda_1, \Lambda_2)$ is the Hausdorf distance between the subsets of the unit sphere contained in $\Lambda_1$and Λ_2.

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

For example, Assumption 5^{2*}(a) holds with Θ defined via equality and/or inequality constraints, as in (4.11).

The verification of part (c) of Assumption 5^{2*} is typically straightforward, though it can be somewhat tedious. To illustrate its verification, suppose $s = 2$, $\Lambda^* = (R^+)^2$, $B_T = \text{Diag}(T^{1/2}, T)$, and M has elements M_{ij} for $i, j = 1, 2$ with $M_{11} > 0$ and $M_{22} > 0$. Then,

$$\begin{aligned} \Upsilon_T M \Lambda^* &= \Big\{ \lambda : \lambda = (\lambda_1, \lambda_2)',\ \lambda_1 = T^{1/2} M_{11} \lambda_1^* + T^{1/2} M_{12} \lambda_2^*, \\ &\qquad \lambda_2 = T M_{21} \lambda_1^* + T M_{22} \lambda_2^*,\ \lambda_1^* \geq 0,\ \lambda_2^* \geq 0 \Big\} \\ &= \Big\{ \lambda : \lambda_1 = M_{11} \lambda_1^* + M_{12} \lambda_2^*,\ \lambda_2 = T^{1/2} M_{21} \lambda_1^* + T^{1/2} M_{22} \lambda_2^*, \\ &\qquad \lambda_1^* \geq 0,\ \lambda_2^* \geq 0 \Big\} \end{aligned} \tag{4.14}$$

If $M_{12} \geq 0$, then $\Upsilon_T M \Lambda^* = \{\lambda : \lambda_1 \geq 0,\ \lambda_2 \geq T^{1/2} M_{21} \lambda_1\}$. Hence, if $M_{12} \geq 0$ and $M_{21} = 0$, then $\Upsilon_T M \Lambda^* = (R^+)^2$ and $\Lambda := (R^+)^2$. If $M_{12} \geq 0$ and $M_{21} > 0$, then $\text{dist}_c(\Upsilon_T M \Lambda^*, \Lambda) \to 0$, where $\Lambda := \{\lambda : \lambda_1 = 0,\ \lambda_2 \geq 0\}$. If $M_{12} \geq 0$ and $M_{21} < 0$, then $\text{dist}_c(\Upsilon_T M \Lambda^*, \Lambda) \to 0$, where $\Lambda := \{\lambda : \lambda_1 \geq 0,\ \lambda_2 \in R\}$. If $M_{12} < 0$, then $\Upsilon_T M \Lambda^* = \{\lambda : \lambda_1 \in R,\ \lambda_2 = M_{21} \lambda_1^* + M_{22} \lambda_2^*,\ \lambda_1^* \geq 0,\ \lambda_2^* \geq 0\}$. Hence, if $M_{12} < 0$ and $M_{21} < 0$, then $\Upsilon_T M \Lambda^* = R^2$ and $\Lambda := R^2$. If $M_{12} < 0$ and $M_{21} \geq 0$, then $\Upsilon_T M \Lambda^* = R \times R^+$ and $\Lambda := R \times R^+$.

Assumptions 5^* and 5^{2*} do not allow for any curvature in the boundary of Θ near θ_0. Such curvature arises in some examples, such as cases where Θ is a sphere, ellipse, cylinder, or a set defined by smooth nonlinear equality and/or inequality constraints, and θ_0 is on its boundary. Assumption 5 can be verified in these cases using the following conditions.

First we state a condition that is the same as that of Chernoff (1954) (except that Chernoff requires $b_T = 1/2$). It is a straightforward simplification of Assumption 5 that holds when B_T is proportional to I_s.

Assumption 5^{3*}. (a) *$\Theta - \theta_0$ is locally approximated by a cone $\Lambda \subset R^s$.*
(b) *$B_T = b_T I_s$ for some scalar constants $\{b_T : T \geq 1\}$ for which $b_T \to \infty$.*

The following sufficient condition for Assumption 5^{3*} considers the case where θ_0 is on the boundary of Θ and some smooth nonlinear equality and/or inequality constraints are binding at θ_0.

Assumption 5^{4*}. (a) *For some $\varepsilon > 0$, $\Theta \cap S(\boldsymbol{\theta}_0, \varepsilon) = \{\theta \in R^s : g_a(\theta) = \mathbf{0}, g_b(\theta) \leq \mathbf{0}$, $||\theta - \theta_0|| \leq \varepsilon\}$, where $g_j(\theta) \in R^{c_j}$ for $0 \leq c_j < \infty$ for $j = a, b$, $g_j(\theta_0) = \mathbf{0}$ for $j = a, b$, and $g(\cdot) = (g_a(\cdot)', g_b(\cdot)')'$ is continuously differentiable on some neighborhood of θ_0 with $\frac{\partial}{\partial \theta'} g(\theta_0)$ of full row rank.*
(b) *$B_T = b_T I_s$ for some scalar constants $\{b_T : T \geq 1\}$ for which $b_T \to \infty$.*

Note that if the true parameter vector θ_0 changes then the inequality constraints that are binding at θ_0, $g_b(\cdot)$, typically change too.

Lemma 2. *Each of Assumptions 5^*, 5^{2*}, 5^{3*}, and 5^{4*} is sufficient for Assumption 5. Under Assumption 5^{4*}, Assumption 5 holds with*
$\Lambda = \left\{\lambda \in R^s : \frac{\partial}{\partial \theta'} g_a(\theta_0)\lambda = \mathbf{0},\ \frac{\partial}{\partial \theta'} g_b(\theta_0)\lambda \leq \mathbf{0}\right\}$.

Next, we show that if Θ is a product set (at least locally to θ_0) and B_T is correspondingly block diagonal, then Assumption 5 can be verified by separately verifying it for each of the component sets. Thus, a different sufficient condition, Assumption 5^*, ..., 5^{4*}, can be used for each component set. This is convenient and it provides for a wider variety of sufficient conditions for Assumption 5.

Lemma 3. *Assume the following conditions hold:*
(a) *$\Theta - \theta_0$ is locally equal to a product set. That is, $(\Theta - \theta_0) \cap S(\mathbf{0}, \varepsilon) = (\times_{j=1}^{J}(\Theta_j - \theta_{j0})) \cap S(\mathbf{0}, \varepsilon)$ for some $\varepsilon > 0$, for some $\Theta_j \subset R^{d_j}$ $\forall j \leq J$, where $\sum_{j=1}^{J} d_j = s$ and $\theta_0 = (\theta'_{10}, ..., \theta'_{J0})'$.*
(b) *B_T is block diagonal with diagonal blocks $B_{jT} \subset R^{d_j \times d_j}$ $\forall j \leq J$.*
(c) *For some positive scalar constants $\{b_{jT} : T \geq 1\}$ for which $b_{jT} \to \infty$ and $b_{jT} \leq c_j \lambda_{\min}(B_{jT})$ for some $0 < c_j < \infty$, $\{B_{jT}(\Theta_j - \theta_{j0})/b_{jT} : T \geq 1\}$ satisfies Assumption 5 for some cone Λ_j $\forall j \leq J$.*
Then, Assumption 5 holds for $\{B_T(\Theta - \theta_0)/b_T : T \geq 1\}$ with $\Lambda := \times_{j=1}^{J} \Lambda_j$ and $b_T := \min_{j \leq J} b_{jT}$.

Comment. The proof of Lemma 3 shows that if Assumption 5 holds for some sequence $\{b_T : T \geq 1\}$, then it holds for any sequence $\{d_T : T \geq 1\}$ for which $d_T \to \infty$ and $d_T \leq b_T$ for T sufficiently large.

4.4. Examples (Continued)

4.4.1. Random Coefficient Regression

Assumptions 5^* and 6 hold in Example 1 with

$$\Lambda := (R^+)^p \times R^{s-p}, \tag{4.15}$$

where $R^+ := \{x \in R : x \geq 0\}$.

4.4.2. Regression with Restricted Parameters

Assumption 5^{4*} holds in Example 2 provided $g_a(\theta)$ and $g_b(\theta)$ are continuously differentiable on some neighborhood of θ_0 and $\frac{\partial}{\partial\theta'}g(\theta_0)$ is full row rank, where $g(\theta) = (g_a(\theta)', g_b(\theta)')'$. In consequence, by Lemma 2, Assumption 5 holds with

$$\Lambda := \left\{\lambda \in R^s : \frac{\partial}{\partial\theta'}g_a(\theta_0)\lambda = \mathbf{0},\ \frac{\partial}{\partial\theta'}g_b(\theta_0)\lambda \leq \mathbf{0}\right\}. \tag{4.16}$$

For Λ as such, Assumption 6 holds.

For example, suppose

$$g_j(\theta) := v_j'\theta - d_j \tag{4.17}$$

for some given $v_j \in R^s$ and $d_j \in R$ for $j = a, b$. Then,

$$\Lambda := \{\lambda \in R^s : v_a'\lambda = 0,\ v_b'\lambda \leq 0\}. \tag{4.18}$$

Alternatively, suppose

$$\begin{aligned} &g_j(\theta) := v_j'\theta^2 - d_j, \text{ where} \\ &\theta = (\theta_1, ..., \theta_s)' \text{ and } \theta^2 = (\theta_1^2, ..., \theta_s^2)' \end{aligned} \tag{4.19}$$

for v_j and c_j as above for $j = a, b$. Then, the boundary of Θ at θ_0 is elliptical and

$$\Lambda := \{\lambda \in R^s : (v_a \odot \theta_0)'\lambda = 0,\ (v_b \odot \theta_0)'\lambda \leq 0\}. \tag{4.20}$$

4.4.3. 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

$$\Lambda^* = \{\lambda^* \in R^s : \lambda^* = (\lambda_1^*, \lambda_2^*, \lambda_3^*, \lambda_4^{*'})',\ \lambda_1^* \leq 0,\ \lambda_2^* \geq 0,\ \lambda_3^* \in R,\ \lambda_4^* \in R^b\}. \tag{4.21}$$

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

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

$$\begin{aligned} \Upsilon_T M\Lambda^* = \Big\{&\lambda \in R^s : \lambda_1 = T\lambda_1^*,\ \lambda_2 = T^{3/2}\mu_0\lambda_1^* + T^{3/2}\lambda_2^*, \\ &\lambda_3 = -T^{1/2}\mu_0\lambda_1^* + T^{1/2}\lambda_3^* + T^{1/2}\mu_0\mathbf{1}'\lambda_4^*, \text{ and} \\ &\lambda_4 = T^{1/2}\lambda_4^* \text{ for } \lambda^* \in \Lambda^*\Big\} \\ = \Big\{&\lambda \in R^s : \lambda_1 \leq 0,\ \lambda_2 \geq T^{1/2}\mu_0\lambda_1,\ \lambda_3 \in R,\ \lambda_4 \in R^b\Big\}. \end{aligned} \tag{4.22}$$

From (4.22), 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

$$\Lambda := B_T\Lambda^* = \{\lambda \in R^s : \lambda_1 \leq 0,\ \lambda_2 \geq 0,\ \lambda_3 \in R,\ \lambda_4 \in R^b\}. \tag{4.23}$$

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

$$\Lambda := \{\lambda \in R^s : \lambda_1 \leq 0,\ \lambda_2 \in R,\ \lambda_3 \in R,\ \lambda_4 \in R^b\}. \tag{4.24}$$

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

Assumption 6 holds for all values of θ_{30}.

4.5. Proofs

Proof of Lemma 1. Let $Z_{Tb} = Z_T/b_T$. By Assumptions 2 and 3, $||Z_{Tb}|| = O_p(b_T^{-1})$. For any set $\Gamma \subset R^s$ and $z \in R^s$, let

$$\text{dist}_T(z, \Gamma) := \inf_{\lambda \in \Gamma}((\lambda - z)'\mathcal{J}_T(\lambda - z))^{1/2}. \tag{4.25}$$

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,

$$\begin{aligned}
\text{dist}_T(Z_{Tb}, B_T(\Theta - \theta_0)/b_T) &= \inf_{\lambda \in B_T(\Theta-\theta_0)/b_T} (\lambda - Z_T/b_T)'\mathcal{J}_T(\lambda - Z_T/b_T)^{1/2} \\
&= b_T^{-1} \inf_{\lambda \in B_T(\Theta-\theta_0)/b_T} (b_T\lambda - Z_T)'\mathcal{J}_T(b_T\lambda - Z_T)^{1/2} \\
&= b_T^{-1} \inf_{\theta \in \Theta} q_T^{1/2}(B_T(\theta - \theta_0)).
\end{aligned} \tag{4.26}$$

Let

$$C_T := \text{dist}_T(Z_{Tb}, \Lambda) - \text{dist}_T(Z_{Tb}, B_T(\Theta - \theta_0)/b_T). \tag{4.27}$$

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

By the triangle inequality,

$$\begin{aligned}
C_T &\leq \text{dist}_T(Z_{Tb}, \{Z_{\Theta Tb}\}) + \text{dist}_T(Z_{\Theta Tb}, \Lambda) - \text{dist}(Z_{Tb}, B_T(\Theta - \theta_0)/b_T) \\
&= \text{dist}_T(Z_{\Theta Tb}, \Lambda) + o_p(b_T^{-1}) \\
&= o_p(||Z_{\Theta Tb}||) + o_p(b_T^{-1}).
\end{aligned} \tag{4.28}$$

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

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

$$\begin{aligned} \text{dist}_T(Z_{Tb}, \{Z_{\Theta Tb}\}) &= \text{dist}_T(Z_{Tb}, B_T(\Theta - \theta_0)/b_T) + o_p(b_T^{-1}) \\ &\leq ||\mathcal{J}_T^{1/2} Z_{Tb}|| + o_p(b_T^{-1}). \end{aligned} \tag{4.29}$$

Using Assumptions 2 and 3, we then obtain

$$\begin{aligned} ||Z_{\Theta Tb} - Z_{Tb}|| &\leq \text{dist}_T(Z_{Tb}, \{Z_{\Theta Tb}\})/\lambda_{\min}(\mathcal{J}_T^{1/2}) \\ &\leq \left(||\mathcal{J}_T^{1/2} Z_{Tb}|| + o_p(b_T^{-1})\right)/\lambda_{\min}(\mathcal{J}_T^{1/2}) \\ &\leq ||Z_{Tb}||\lambda_{\max}(\mathcal{J}_T^{1/2})/\lambda_{\min}(\mathcal{J}_T^{1/2}) + o_p(b_T^{-1}) = O_p(b_T^{-1}). \end{aligned} \tag{4.30}$$

Thus,

$$||Z_{\Theta Tb}|| \leq ||Z_{\Theta Tb} - Z_{Tb}|| + ||Z_{Tb}|| = O_p(b_T^{-1}). \tag{4.31}$$

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

Proof of Theorem 2. Let $\lambda_T^* \in \text{cl}(\Lambda)$ be such that $||B_T(\widehat{\theta} - \theta_0) - \lambda_T^*|| = \text{dist}(B_T(\widehat{\theta} - \theta_0), \Lambda)$. λ_T^* is unique because Λ is a convex cone; see Perlman (1969, Sec. 4). By Assumption 5 and Theorem 1(a), $||B_T(\widehat{\theta} - \theta_0)/b_T - \lambda_T^*/b_T|| = \text{dist}(B_T(\widehat{\theta} - \theta_0)/b_T, \Lambda) = o(||B_T(\widehat{\theta} - \theta_0)/b_T||) = o_p(b_T^{-1})$ and so

$$||B_T(\widehat{\theta} - \theta_0) - \lambda_T^*|| = o_p(1). \tag{4.32}$$

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

Define $|| \cdot ||_T$ by $||\lambda||_T := (\lambda' \mathcal{J}_T \lambda)^{1/2}$. By Assumption 3, it suffices to show that $||\lambda_T^* - \widehat{\lambda}_T||_T = o_p(1)$.

By Assumption 3, (4.32) holds with $|| \cdot ||$ replaced by $|| \cdot ||_T$. This, the triangle inequality, and Lemma 1 give

$$||\lambda_T^* - Z_T||_T = \left|\left|B_T(\widehat{\theta} - \theta_0) - Z_T\right|\right|_T + o_p(1) = \left|\left|\widehat{\lambda}_T - Z_T\right|\right|_T + o_p(1). \tag{4.33}$$

In consequence,

$$\begin{aligned} \varepsilon_T &:= ||\lambda_T^* - Z_T||_T - \left|\left|\widehat{\lambda}_T - Z_T\right|\right|_T = o_p(1) \text{ and} \\ \varepsilon_T^* &:= ||\lambda_T^* - Z_T||_T^2 - \left|\left|\widehat{\lambda}_T - Z_T\right|\right|_T^2 = o_p(1). \end{aligned} \tag{4.34}$$

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

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

By the orthogonality of projections,

$$\left|\left|\lambda_T^* - \widehat{\lambda}_T\right|\right|_T^2 = ||\lambda_T^* - P_L\lambda_T^*||_T^2 + \left|\left|P_L\lambda_T^* - \widehat{\lambda}_T\right|\right|_T^2. \tag{4.35}$$

We claim that (i) $||\lambda_T^* - P_L\lambda_T^*||_T^2 \leq \varepsilon_T^*$ and (ii) $||P_L\lambda_T^* - \widehat{\lambda}_T||_T^2 \leq \varepsilon_T^2$. These two claims and (4.35) combine to yield $||\lambda_T^* - \widehat{\lambda}_T||_T \leq \varepsilon_T^* + \varepsilon_T^2$ when $Z_T \notin \text{cl}(\Lambda)$, which gives the desired result.

Claim (i) follows from

$$\begin{aligned} ||\lambda_T^* - P_L\lambda_T^*||_T^2 &= ||\lambda_T^* - Z_T||_T^2 - ||P_L\lambda_T^* - Z_T||_T^2 \\ &= \left|\left|\widehat{\lambda}_T - Z_T\right|\right|_T^2 + \varepsilon_T^* - ||P_L\lambda_T^* - Z_T||_T^2 \\ &\leq \varepsilon_T^*, \end{aligned} \tag{4.36}$$

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

Claim (ii) is established as follows. The first equality of (4.36) implies that

$$||P_L\lambda_T^* - Z_T||_T \leq ||\lambda_T^* - Z_T||_T = \left|\left|\widehat{\lambda}_T - Z_T\right|\right|_T + \varepsilon_T. \tag{4.37}$$

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

$$\begin{aligned} \left|\left|P_L\lambda_T^* - \widehat{\lambda}_T\right|\right|_T &= ||P_L\lambda_T^* - Z_T||_T - \left|\left|\widehat{\lambda}_T - Z_T\right|\right|_T \\ &\leq \left|\left|\widehat{\lambda}_T - Z_T\right|\right|_T + \varepsilon_T - \left|\left|\widehat{\lambda}_T - Z_T\right|\right|_T \\ &= \varepsilon_T. \quad \square \end{aligned} \tag{4.38}$$

Proof of Lemma 2. 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 5^{2*} implies Assumption 5 with $b_T := \lambda_{\min}(\Upsilon_T)$. Assume Assumption 5^{2*} holds. A sequence $\{\phi_T \in R^s : T \geq 1\}$ with $||\phi_T|| \to 0$ satisfies

$$\phi_T \in B_T(\Theta - \theta_0)/b_T \ \ \forall T \text{ large iff } \phi_T \in B_T\Lambda^* \ \ \forall T \text{ large.} \tag{4.39}$$

This holds because $\Upsilon_{Tj}/b_T \geq 1 \ \forall T \geq 1, \forall j \leq s$ (where $\Upsilon_T := \text{Diag}(\Upsilon_{T1}, ..., \Upsilon_{Ts})$) implies that $||b_T M^{-1\prime}\Upsilon_T^{-1}\phi_T|| \leq ||M^{-1\prime}|| \cdot ||\phi_T|| \to 0$. Suppose $\phi_T \in B_T(\Theta - \theta_0)/b_T \ \forall T$ large, then $b_T M^{-1\prime}\Upsilon_T^{-1}\phi_T \in (\Theta - \theta_0) \cap S(0, \varepsilon) = \Lambda^* \cap S(0, \varepsilon) \subset \Lambda^* \ \forall T$ large and $\phi_T \in$

$B_T\Lambda^*$ $\forall T$ large. Conversely, suppose $\phi_T \in B_T\Lambda^*$ $\forall T$ large, then $b_T M^{-1\prime}\Upsilon_T^{-1}\phi_T \in \Lambda^* \cap S(0,\varepsilon) = (\Theta - \theta_0) \cap S(0,\varepsilon) \subset \Theta - \theta_0$ $\forall T$ large and $\phi_T \in B_T(\Theta - \theta_0)/b_T$ $\forall T$ large.

Using (4.39), for any sequence $\{\phi_T \in B_T(\Theta - \theta_0)/b_T : T \geq 1\}$ with $||\phi_T|| \to 0$, we have $\phi_T \in B_T\Lambda^*$ $\forall T$ large. For such a sequence,

$$\text{dist}(\phi_T, \Lambda) = ||\phi_T||\text{dist}(\phi_T/||\phi_T||, \Lambda) \leq ||\phi_T||\text{dist}_c(B_T\Lambda^*, \Lambda) = o(||\phi_T||), \tag{4.40}$$

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 5^{2*}(c).

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

$$\begin{aligned} \text{dist}(\lambda_T, B_T\Lambda^*) &= ||\lambda_T||\text{dist}(\lambda_T/||\lambda_T||, B_T\Lambda^*) \leq ||\lambda_T||\text{dist}_c(\Lambda, B_T\Lambda^*) \\ &= o(||\lambda_T||) \end{aligned} \tag{4.41}$$

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

$$\text{dist}(\lambda_T, B_T\Lambda^*) = ||\lambda_T - \phi_T|| + o(||\lambda_T||) \geq \text{dist}(\lambda_T, B_T(\Theta - \theta_0)/b_T) \tag{4.42}$$

$\forall T$ large, where the inequality holds because $||\lambda_T|| \to 0$ implies $||\phi_T|| \to 0$ implies $\phi_T \in B_T(\Theta - \theta_0)/b_T$ using (4.39). Equations (4.40)–(4.42) combine to verify Assumption 5.

Assumption 5^{3*} implies Assumption 5 because $B_T = b_T I_s$ implies that $B_T(\Theta - \theta_0)/b_T = \Theta - \theta_0$.

Lastly, we show that Assumption 5^{4*} implies Assumption 5^{3*}. By assumption, $g_j(\theta_0) = \mathbf{0}$ for $j = a, b$. Let $\Gamma_j := \frac{\partial}{\partial\theta'} g_j(\theta_0) \in R^{c_j \times s}$ for $j = a, b$. Let

$$\Gamma := \begin{bmatrix} \Gamma_a \\ \Gamma_b \\ \Gamma_c \end{bmatrix} \text{ and } g^+(\theta) := \begin{pmatrix} g_a(\theta) \\ g_b(\theta) \\ \Gamma_c(\theta - \theta_0) \end{pmatrix}, \tag{4.43}$$

where $\Gamma_c \in R^{(s - c_a - c_b)\times s}$ is chosen such that Γ is nonsingular. Then, $g^+(\theta_0) = \mathbf{0}$ and $\frac{\partial}{\partial\theta'} g^+(\theta_0) = \Gamma$.

Let $\Phi := \Theta - \theta_0$. Given $\phi \in \Phi$, define

$$\lambda^* := \Gamma^{-1} g^+(\theta_0 + \phi). \tag{4.44}$$

Then, $\Gamma\lambda^* = g^+(\theta_0 + \phi)$, $\Gamma_a\lambda^* = g_a(\theta_0 + \phi) = g_a(\theta) = \mathbf{0}$, and $\Gamma_b\lambda^* = g_b(\theta_0 + \phi) = g_b(\theta) \leq \mathbf{0}$ for $\theta = \theta_0 + \phi \in \Theta$. Hence, $\lambda^* \in \Lambda$. Element by element mean value expansions give

$$\begin{aligned} \lambda^* &:= \Gamma^{-1} g^+(\theta_0 + \phi) = \Gamma^{-1} g^+(\theta_0) + \Gamma^{-1}\frac{\partial}{\partial\theta'} g^+(\theta_0)\phi + o(||\phi||) \\ &= \mathbf{0} + \Gamma^{-1}\Gamma\phi + o(||\phi||) = \phi + o(||\phi||). \end{aligned} \tag{4.45}$$

We conclude that $\text{dist}(\phi, \Lambda) \leq ||\phi - \lambda^*|| = o(||\phi||)$, as required by Assumption 5^{3*}.

Next, the function $m(\cdot) := g^+(\theta_0 + \cdot) : R^s \to R^s$ is continuously differentiable on a neighborhood of $\mathbf{0}$ with nonsingular Jacobian matrix at $\mathbf{0}$ and $m(\mathbf{0}) = \mathbf{0}$. Hence, by the inverse function theorem, there exists a function $m^{-1}(\cdot) : R^s \to R^s$ that satisfies $m^{-1}(\phi)$ is continuously differentiable and $m(m^{-1}(\phi)) = \phi$ for all ϕ in a neighborhood of $\mathbf{0}$, $m^{-1}(\mathbf{0}) = \mathbf{0}$, and $\frac{\partial}{\partial \phi'} m^{-1}(\mathbf{0}) = \left[\frac{\partial}{\partial \phi'} m(0)\right]^{-1}$ $(:= \Gamma^{-1})$.

Given $\lambda \in \Lambda$ with λ close to $\mathbf{0}$, define

$$\phi^* := m^{-1}(\Gamma\lambda). \tag{4.46}$$

Then, $g^+(\theta_0 + \phi^*) := m(\phi^*) = m(m^{-1}(\Gamma\lambda)) = \Gamma\lambda$, $g_a(\theta_0 + \phi^*) = \Gamma_a\lambda = \mathbf{0}$, and $g_b(\theta_0 + \phi^*) = \Gamma_b\lambda \leq \mathbf{0}$. Hence, $\phi^* \in \Phi$. Element by element mean value expansions give

$$\begin{aligned} \phi^* &:= m^{-1}(\Gamma\lambda) = m^{-1}(\mathbf{0}) + \frac{\partial}{\partial \phi'} m^{-1}(\mathbf{0})\Gamma\lambda + o(||\lambda||) \\ &= \mathbf{0} + \left[\frac{\partial}{\partial \phi'} m(\mathbf{0})\right]^{-1} \Gamma\lambda + o(||\lambda||) = \lambda + o(||\lambda||). \end{aligned} \tag{4.47}$$

Hence, $\text{dist}(\lambda, \Theta - \theta_0) \leq ||\lambda - \phi^*|| = o(||\lambda||)$ and Assumption 5^{3*} holds. □

Proof of Lemma 3. First, we show that if Assumptions (a) and (b) of the Lemma hold and (i) $\{B_{jT}(\Theta_j - \theta_{j0})/b_T : T \geq 1\}$ satisfies Assumption 5 with cone Λ_j $\forall j \leq J$, where $b_T := \min_{j \leq J} b_{jT}$ and $\{b_{jT} : T \geq 1\}$ are as in Assumption (c), then (ii) $\{B_T(\Theta - \theta_0)/b_T : T \geq 1\}$ satisfies Assumption 5 with cone $\Lambda := \times_{j=1}^J \Lambda_j$. Second, we show that if Assumption 5 holds for a sequence $\{b_T : T \geq 1\}$, then it holds for any sequence $\{d_T : T \geq 1\}$ for which $d_T \to \infty$ and $d_T \leq b_T$ $\forall T \geq 1$. The latter and Assumption (c) of the Lemma imply that condition (i) holds. Hence, condition (ii) holds, which is the desired result.

Assume condition (i) holds. First, note that $b_T \to 0$ and

$$b_T := \min_{j \leq J} b_{jT} \leq \min_{j \leq J} c_j \lambda_{\min}(B_{jT}) \leq \max_{\ell \leq J} c_\ell \lambda_{\min}(B_T). \tag{4.48}$$

Thus, $\{b_T : T \geq 1\}$ satisfies the requisite conditions of Assumption 5.

Consider a sequence $\{\phi_T \in B_T(\Theta - \theta_0)/b_T : T \geq 1\}$ for which $||\phi_T|| \to 0$. We have

$$b_T B_T^{-1'} \phi_T \in \Theta - \theta_0 \text{ and } ||b_T B_T^{-1'} \phi_T|| \leq ||b_T B_T^{-1'}|| \cdot ||\phi_T|| \to 0, \tag{4.49}$$

where the convergence to zero holds because (4.48) implies that $||b_T B_T^{-1'}|| = O_p(1)$. Equation (4.49) implies that $b_T B_T^{-1'} \phi_T \in \times_{j=1}^J (\Theta_j - \theta_{j0})$, $\phi_T \in \times_{j=1}^J B_{jT}(\Theta_j - \theta_{j0})/b_T$, and $\phi_{jT} \in B_{jT}(\Theta_j - \theta_{j0})/b_T$ $\forall j \leq J$, $\forall T$ large, where $\phi_T = (\phi_{1T}, ..., \phi_{JT})'$. Using these results, we obtain:

$$\begin{aligned} \text{dist}(\phi_T, \Lambda) &= \inf_{\lambda \in \Lambda} ||\phi_T - \lambda|| = \inf_{\lambda_j \in \Lambda_j : \forall j \leq J} \left(\sum_{j=1}^J ||\phi_{jT} - \lambda_j||^2\right)^{1/2} \\ &= \left(\sum_{j=1}^J o(||\phi_{jT}||^2)\right)^{1/2} = o(||\phi_T||), \end{aligned} \tag{4.50}$$

where the second equality holds because Λ is a product space and the second last equality holds by condition (i).

Next, consider a sequence $\{\lambda_T \in \Lambda : T \geq 1\}$ for which $||\lambda_T|| \to 0$. We want to show that $\text{dist}(\lambda_T, B_T(\Theta - \theta_0)/b_T) = o(||\lambda_T||)$. This will complete the proof that condition (ii) holds. Let $\lambda_T := (\lambda'_{1T}, ..., \lambda'_{JT})'$, where $\lambda_{jT} \in R^{d_j}$ $\forall j \leq J$. Let $\phi_{jT} \in B_{jT}(\Theta - \theta_0)/b_T$ be such that $\text{dist}(\lambda_{jT}, B_{jT}(\Theta - \theta_0)/b_T) = ||\lambda_{jT} - \phi_{jT}|| + o(||\lambda_{jT}||)$ $\forall j \leq J$. Note that $||\phi_{jT}|| \to 0$ $\forall j \leq J$ because the left-hand side of the last equation is $o(1)$ and $||\lambda_{jT}|| \to 0$. Define $\phi_T := (\phi'_{1T}, ..., \phi'_{JT})'$. We have $||\phi_T|| \to 0$. Hence, for some $\varepsilon > 0$ and all T large,

$$
\begin{aligned}
\phi_T &\in \left(\mathop{\times}_{j=1}^{J} B_{jT}(\Theta_j - \theta_{j0})/b_T \right) \cap S(0, \varepsilon) \\
&= (B_T(\Theta - \theta_0)/b_T) \cap S(\mathbf{0}, \varepsilon) \subset B_T(\Theta - \theta_0)/b_T) \qquad (4.51)
\end{aligned}
$$

using Assumptions (a) and (b). In consequence,

$$
\begin{aligned}
\text{dist}(\lambda_T, B_T(\Theta - \theta_0)/b_T) &\leq ||\lambda_T - \phi_T|| = \left(\sum_{j=1}^{J} ||\lambda_{jT} - \phi_{jT}||^2 \right)^{1/2} \\
&= \left(\sum_{j=1}^{J} (\text{dist}(\lambda_{jT}, B_{jT}(\Theta - \theta_0)/b_T) + o(||\lambda_{jT}||))^2 \right)^{1/2} \\
&= \left(\sum_{j=1}^{J} o(||\lambda_{jT}||)^2 \right)^{1/2} = o(||\lambda_T||), \qquad (4.52)
\end{aligned}
$$

where the second last equality uses condition (i).

Now we establish the second result stated in the first paragraph of this proof. Suppose Assumption 5 holds for the sequence of sets $\Phi_T := B_T(\Theta - \theta_0)/b_T$ for $T \geq 1$. Assume $d_T \leq b_T$ $\forall T \geq 1$ and $d_T \to \infty$. We want to show that Assumption 5 holds for the sets $b_T \Phi_T / d_T$ for $T \geq 1$. Consider a sequence $\{b_T \phi_T / d_T \in b_T \Phi_T / d_T : T \geq 1\}$ for which $||b_T \phi_T / d_T|| \to 0$. Then, $||\phi_T|| \leq ||b_T \phi_T / d_T|| \to 0$ and we have

$$
\text{dist}(b_T \phi_T / d_T, \Lambda) = (b_T/d_T)\text{dist}(\phi_T, \Lambda) = (b_T/d_T) o(||\phi_T||) = o(||b_T \phi_T / d_T||), \qquad (4.53)
$$

where the first equality holds because Λ is a cone.

Next, let $\{\lambda_T \in \Lambda : T \geq 1\}$ be a sequence for which $||\lambda_T|| \to 0$. We have

$$
\begin{aligned}
\text{dist}(\lambda_T, b_T \Phi_T / d_T) &= \inf_{\phi_T \in \Phi_T} ||\lambda_T - b_T \phi_T / d_T|| = (b_T/d_T) \inf_{\phi_T \in \Phi_T} ||d_T \lambda_T / b_T - \phi_T|| \\
&= (b_T/d_T)\text{dist}(d_T \lambda_T / b_T, \Phi_T) \\
&= (b_T/d_T) o(||d_T \lambda_T / b_T||) = o(||\lambda_T||), \qquad (4.54)
\end{aligned}
$$

where the second last inequality holds by Assumption 5 for the sets $\{\Phi_T : T \geq 1\}$ because $d_T \phi_T / b_T \in \Lambda$ and $||d_T \lambda_T / b_T|| \leq ||\lambda_T|| \to 0$. This concludes the proof that Assumption 5 holds for the sets $\{b_T \Phi_T / d_T : T \geq 1\}$. $\square$

5. Asymptotic Distribution of the Extremum Estimator

5.1. General Results

In this section, we determine the asymptotic distribution of $\widehat{\theta}$. For this, we assume:

Assumption 7. $(B_T^{-1'} D\ell_T(\theta_0), \mathcal{J}_T) \xrightarrow{d} (G, \mathcal{J})$ *for some random variables* $G \in R^s$ *and* $\mathcal{J} \in R^{s \times s}$ *for which* $\mathcal{J}$ *is symmetric and nonsingular with probability one.*

For models without stochastic trends, $\mathcal{J}$ is typically non-random. In this case, it suffices to have $B_T^{-1'} D\ell_T(\theta_0) \xrightarrow{d} G$ and $\mathcal{J}_T = \mathcal{J} + o_p(1)$. Typically, G is a mean zero Gaussian random variable and the convergence in distribution is established via the methods described in Section 3 to verify Assumption 2. We describe G and $\mathcal{J}$ in more detail in Section 6.

We note that Assumption 7 implies Assumptions 2 and 3.

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

$$q(\widehat{\lambda}) = \inf_{\lambda \in \Lambda} q(\lambda), \text{ where}$$

$$q(\lambda) := (\lambda - Z)' \mathcal{J} (\lambda - Z) \text{ and } Z := \mathcal{J}^{-1} G. \tag{5.1}$$

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

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

Theorem 3. (a) *Suppose Assumptions* 1 *and* 4–7 *hold. Then,* $\widehat{\lambda}_T \xrightarrow{d} \widehat{\lambda}$ *and* $B_T(\widehat{\theta} - \theta_0) \xrightarrow{d} \widehat{\lambda}$.
(b) *Suppose Assumptions* 1, 4, 5, *and* 7 *hold. Then,* $\ell_T(\widehat{\theta}) - \ell_T(\theta_0) \xrightarrow{d} \frac{1}{2}(Z'\mathcal{J}Z - \inf_{\lambda \in \Lambda} q(\lambda)) = \frac{1}{2}\widehat{\lambda}' \mathcal{J} \widehat{\lambda}$.

Comments. 1. In the classical case in which θ_0 is not on a boundary, $\Lambda = R^s$ and $\widehat{\lambda} = \mathcal{J}^{-1} G$. Thus, if G is Gaussian and $\mathcal{J}$ is non-random (as typically occurs in models without stochastic trends), then $B_T(\widehat{\theta} - \theta_0)$ has a Gaussian distribution. Alternatively, if G is Gaussian conditional on $\mathcal{J}$ and $\mathcal{J}$ is random (as occurs in some models with cointegration), then $B_T(\widehat{\theta} - \theta_0)$ has a mixed Gaussian distribution.

2. 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 $\widehat{\lambda}$ is more complex than in Comment 1. The following section, Section 6, analyzes its distribution in some detail.

3. In part (b), $\widehat{\lambda}' \mathcal{J} \widehat{\lambda}$ is uniquely defined, even though $\widehat{\lambda}$ need not be (because Assumption 6 is not imposed).

4. The result of Theorem 3(b) can be used to obtain the asymptotic distribution of a likelihood ratio-like statistic, as is done in Andrews (1997b). It is a by-product of the results needed to obtain the asymptotic distribution of $B_T(\widehat{\theta} - \theta_0)$.

5.2. Examples (Continued)

5.2.1. Random Coefficient Regression

In Example 1, $\mathcal{J}_T$ does not depend on T and $\mathcal{J}$ $(= \mathcal{J}_T)$ is symmetric and positive definite by (3.16) and (3.18). Thus, Assumption 7 holds provided $T^{-1/2}D\ell_T(\theta_0) \xrightarrow{d} G$ for some random variable G. By the definition of $D\ell_T(\theta_0)$ in (3.16) and the moment assumptions of (3.17), the CLT for iid mean zero finite variance random variables yields

$$T^{-1/2}D\ell_T(\theta_0) \xrightarrow{d} G \sim N(\mathbf{0}, \mathcal{I}), \text{ where}$$

$$\mathcal{I} := \begin{bmatrix} \frac{1}{4}E\dfrac{(\text{res}_t^2(\theta_0) - \text{var}_t(\theta_0))^2}{\text{var}_t^4(\theta_0)}W_t^2W_t^{2\prime} & \frac{1}{2}E\dfrac{\text{res}_t^3(\theta_0)}{\text{var}_t^3(\theta_0)}W_t^2W_t' \\ \\ \frac{1}{2}E\dfrac{\text{res}_t^3(\theta_0)}{\text{var}_t^3(\theta_0)}W_tW_t^{2\prime} & EW_tW_t'/\text{var}_t(\theta_0) \end{bmatrix}. \tag{5.2}$$

By Theorem 3, $T^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} \widehat{\lambda}$, where $\widehat{\lambda}$ satisfies (5.1) with $(G, \mathcal{J})$ defined in (5.2) and (3.16) and Λ defined in (4.15).

5.2.2. Regression with Restricted Parameters

In Example 2, Assumption 7 holds with

$$G \sim N(\mathbf{0}, \mathcal{I}), \mathcal{I} = E\varepsilon_t^2 X_t X_t', \text{ and } \mathcal{J} = EX_tX_t'. \tag{5.3}$$

This follows from the CLT for iid mean zero finite variance random variables and the LLN for iid finite mean random variables using (3.21) and (3.22).

Thus, by Theorem 3, $T^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} \widehat{\lambda}$, where $\widehat{\lambda}$ satisfies (5.1) with $(G, \mathcal{J})$ defined in (5.3) and Λ defined in (4.16).

5.2.3. Dickey–Fuller Regression

Assumption 7 holds in this example by (3.29) with $(G, \mathcal{J})$ defined in (3.29). In consequence, $B_T(\widehat{\theta} - \theta_0) \xrightarrow{d} \widehat{\lambda}$, where $\widehat{\lambda}$ is defined in (5.1) with Λ defined in (4.23) or (4.24) depending on the value of the drift parameter θ_{30}.

5.3. Proofs

Proof of Theorem 3. In part (a), $\widehat{\lambda}_T$ is uniquely defined because Λ is a convex cone. We can write $\widehat{\lambda}_T = h(B_T^{-1\prime} D\ell_T(\theta_0), \mathcal{J}_T)$, where the function h is defined implicitly in (4.5). The function h is continuous at all points $(B_T^{-1\prime} D\ell_T(\theta_0), \mathcal{J}_T)$ for which $\mathcal{J}_T$ is nonsingular. Because $\mathcal{J}$ is nonsingular with probability one, the continuous mapping theorem gives $\widehat{\lambda}_T = h(B_T^{-1\prime} D\ell_T(\theta_0), \mathcal{J}_T) \xrightarrow{d} h(G, \mathcal{J}) = \widehat{\lambda}$. The second result of part (a) holds by the first result and Theorem 2.

Part (b) holds by (4.4), (4.8), Assumption 7, and the continuous mapping theorem. □

6. Asymptotic Distributions of Subvectors of the Extremum Estimator

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

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

We partition θ as follows:

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

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" $\mathcal{J}$ 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 $\widehat{\theta}$, θ_0, B_T, G, $\mathcal{J}$, Z, $\widehat{\lambda}_T$, $\widehat{\lambda}$, and $D\ell_T(\theta_0)$ conformably with θ. Let

$$\widehat{\theta} = \begin{pmatrix} \widehat{\theta}_* \\ \widehat{\psi} \end{pmatrix} = \begin{pmatrix} \widehat{\beta} \\ \widehat{\delta} \\ \widehat{\psi} \end{pmatrix}, \ \theta_0 = \begin{pmatrix} \theta_{*0} \\ \psi_0 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \delta_0 \\ \psi_0 \end{pmatrix},$$

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

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

$$Z = \begin{pmatrix} Z_* \\ Z_\psi \end{pmatrix} = \begin{pmatrix} Z_\beta \\ Z_\delta \\ Z_\psi \end{pmatrix}, \ \widehat{\lambda}_T = \begin{pmatrix} \widehat{\lambda}_{*T} \\ \widehat{\lambda}_{\psi T} \end{pmatrix} = \begin{pmatrix} \widehat{\lambda}_{\beta T} \\ \widehat{\lambda}_{\delta T} \\ \widehat{\lambda}_{\psi T} \end{pmatrix}, \ \widehat{\lambda} = \begin{pmatrix} \widehat{\lambda}_* \\ \widehat{\lambda}_\psi \end{pmatrix} = \begin{pmatrix} \widehat{\lambda}_\beta \\ \widehat{\lambda}_\delta \\ \widehat{\lambda}_\psi \end{pmatrix}, \text{ and}$$

$$D\ell_T(\theta_0) = \begin{pmatrix} D_*\ell_T(\theta_0) \\ D_\psi\ell_T(\theta_0) \end{pmatrix} = \begin{pmatrix} D_\beta\ell_T(\theta_0) \\ D_\delta\ell_T(\theta_0) \\ D_\psi\ell_T(\theta_0) \end{pmatrix}. \tag{6.2}$$

The defining feature of the parameter ψ is the following:

Assumption 8. (a) *$\mathcal{J}$ is block diagonal between θ_* and ψ. That is, $\mathcal{J}_{*\psi} = \mathcal{J}_{\psi *}' = \mathbf{0}$.* (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.*

The defining feature of the parameter δ is the following:

Assumption 9. *$\Lambda_\delta = R^q$.*

Assumptions 8 and 9 require that the asymptotic information matrix is block diagonal between θ_* and ψ and that δ_0 is not on a boundary. Any subvector of θ that does not satisfy one or another of these conditions is lumped in with β.

Under Assumption 8,

$$
\begin{aligned}
&Z_* = \mathcal{J}_*^{-1}G_*,\ Z_\psi = \mathcal{J}_\psi^{-1}G_\psi, \text{ and}\\
&Z_\beta = HZ_* = \mathcal{J}_\beta^{-1}G_\beta + \mathcal{J}_\beta^{-1}\mathcal{J}_{\beta\delta}(\mathcal{J}_\delta - \mathcal{J}_{\delta\beta}\mathcal{J}_\beta^{-1}\mathcal{J}_{\beta\delta})^{-1}(\mathcal{J}_{\delta\beta}\mathcal{J}_\beta^{-1}G_\beta - G_\delta), \text{ where}
\end{aligned}
$$

(6.3) $H := [I_p \vdots \mathbf{0}] \in R^{p\times(p+q)}$.

Define

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

Given Assumptions 8 and 9, we can split the terms of the quadratic approximation to $\ell_T(\widehat{\theta})$, and in consequence $\widehat{\lambda}$, into separate terms involving β, δ, and ψ :

Theorem 4. *Suppose Assumptions* 7–9 *hold. Then,*
(a) $q_\beta(\widehat{\lambda}_\beta) = \inf_{\lambda_\beta\in\Lambda_\beta} q_\beta(\lambda_\beta)$,
(b) $\widehat{\lambda}_\delta = \mathcal{J}_\delta^{-1}G_\delta - \mathcal{J}_\delta^{-1}\mathcal{J}_{\delta\beta}\widehat{\lambda}_\beta$,
(c) $q_\psi(\widehat{\lambda}_\psi) = \inf_{\lambda_\psi\in\Lambda_\psi} q_\psi(\lambda_\psi)$,
(d) $Z'\mathcal{J}Z = Z_\beta'(H\mathcal{J}_*^{-1}H')^{-1}Z_\beta + G_\delta'\mathcal{J}_\delta^{-1}G_\delta + Z_\psi'\mathcal{J}_\psi Z_\psi$,
(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) $Z'\mathcal{J}Z - \inf_{\lambda\in\Lambda} q(\lambda) = Z_\beta'(H\mathcal{J}_*^{-1}H')^{-1}Z_\beta - \inf_{\lambda_\beta\in\Lambda_\beta} q_\beta(\lambda_\beta) + G_\delta'\mathcal{J}_\delta^{-1}G_\delta + Z_\psi'\mathcal{J}_\psi Z_\psi - \inf_{\lambda_\psi\in\Lambda_\psi} q_\psi(\lambda_\psi) = \widehat{\lambda}_\beta'(H\mathcal{J}_*^{-1}H')^{-1}\widehat{\lambda}_\beta + G_\delta'\mathcal{J}_\delta^{-1}G_\delta + \widehat{\lambda}_\psi'\mathcal{J}_\psi\widehat{\lambda}_\psi$.

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 $\widehat{\lambda}_\beta = Z_\beta$. Similarly, if $\Lambda_\psi = R^r$, then $\inf_{\lambda_\psi\in\Lambda_\psi} q_\psi(\lambda_\psi) = 0$ and $\widehat{\lambda}_\psi = Z_\psi = \mathcal{J}_\psi^{-1}G_\psi$. These simplifications correspond to the standard case considered in the literature. Our interest here is in cases where one or 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 $\widehat{\lambda}_\beta = P_{\Lambda_\beta}Z_\beta$, where P_{Λ_β} is the projection matrix onto Λ_β with respect to the norm $||\lambda_\beta||_\beta^2 := \lambda_\beta'(H\mathcal{J}_*^{-1}H')^{-1}\lambda_\beta$. For example, if $\Lambda_\beta = \{\lambda_\beta \in R^p : \Gamma_a\lambda_\beta = \mathbf{0}\}$, then $P_{\Lambda_\beta} := I_p - H\mathcal{J}_*^{-1}H'\Gamma_a'(\Gamma_a H\mathcal{J}_*^{-1}H'\Gamma_a')^{-1}\Gamma_a$.

Theorems 3 and 4 combine to give

Corollary 1. (a) *Suppose Assumptions* 1 *and* 4–9 *hold. Then,*

$$
\begin{aligned}
&B_{\beta T}(\widehat{\beta} - \beta_0) + B_{\beta\delta T}(\widehat{\delta} - \delta_0) + B_{\beta\psi T}(\widehat{\psi} - \psi_0) \xrightarrow{d} \widehat{\lambda}_\beta,\\
&\textit{where } \widehat{\lambda}_\beta \textit{ solves } q_\beta(\widehat{\lambda}_\beta) = \inf_{\lambda_\beta\in\Lambda_\beta} q_\beta(\lambda_\beta),
\end{aligned}
$$

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

and the convergence of these three terms holds jointly.
(b) *Suppose Assumptions* 1 *and* 4–9 *hold. Then,*

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

and the convergence of these three terms holds jointly, where $\widehat{\lambda}_\beta$ *and* $\widehat{\lambda}_\psi$ *are as in part* (a)*.*
(c) *Suppose Assumptions* 1*,* 4*,* 5*, and* 7–9 *hold. Then,*

$$
\begin{aligned}
\ell_T(\widehat{\theta})-\ell_T(\theta_0) &\xrightarrow{d} \tfrac{1}{2}(Z_\beta'(H\mathcal{J}_*^{-1}H')^{-1}Z_\beta-\inf_{\lambda_\beta\in\Lambda_\beta} q_\beta(\lambda_\beta))\\
&\quad+\tfrac{1}{2}G_\delta'\mathcal{J}_\delta^{-1}G_\delta+\tfrac{1}{2}(Z_\psi'\mathcal{J}_\psi Z_\psi-\inf_{\lambda_\psi\in\Lambda_\psi} q_\psi(\lambda_\psi))\\
&= \tfrac{1}{2}(\widehat{\lambda}_\beta'(H\mathcal{J}_*^{-1}H')^{-1}\widehat{\lambda}_\beta+G_\delta'\mathcal{J}_\delta^{-1}G_\delta+\widehat{\lambda}_\psi'\mathcal{J}_\psi\widehat{\lambda}_\psi).
\end{aligned}
$$

Comments. 1. All of the three results of Corollary 1(b) are applicable in the examples of Section 2 except in the Dickey–Fuller Regression Example and the GARCH$(1,q^*)$ Example. 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 $\widehat{\beta}$ and $\widehat{\delta}$ do not depend on whether ψ_0 is on a boundary. Similarly, the asymptotic distribution of $\widehat{\psi}$ does not depend on whether β_0 is on a boundary. For example, in the Random Coefficients Regression Example 1, 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 $\widehat{\delta}$ depends on whether β_0 is on a boundary if and only if $\mathcal{J}_{\delta\beta}\neq\mathbf{0}$. For example, in the Regression with Restricted Parameters Example 2, 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 $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\delta}})$

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

We specify three conditions that imply Assumption 7 and that indicate the form that the limit random variables $(G_*, \mathcal{J}_*)$, which determine the asymptotic distributions of $\widehat{\beta}$ and $\widehat{\delta}$, take in typical cases. The first condition is applicable in models in which $B_{*T}^{-1'} D_* \ell_T(\theta_0)$ and $\mathcal{J}_{*T}$ 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. It excludes the Dickey–Fuller Regression Example 2. (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 7*. (a) *Assumption* 7 *holds.*
(b) $G_* \sim N(\mu, \mathcal{I}_*)$ *conditional on some* σ-field $\mathcal{F}$, *for some non-random* $(p+q)$*-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 $\mathcal{J}_{*T}$ may depend on deterministic trends but not on stochastic trends.

Assumption 7^{2*}. (a) *Assumption* 7 *holds.*
(b) $G_* \sim N(\mathbf{0}, \mathcal{I}_*)$ *for some nonrandom* $(p+q) \times (p+q)$*-matrix* $\mathcal{I}_*$.
(c) $\mathcal{J}_*$ *is nonrandom.*

It is apparent that Assumption $7^{2*} \Rightarrow 7^* \Rightarrow 7$.

Next, we consider the case where $\mathcal{I}_*$ of Assumption 7^* or Assumption 7^{2*} is proportional to $\mathcal{J}_*$.

Assumption 7^{3*}. (a) *Assumption* 7^* *holds.*
(b) $\mathcal{I}_* = c\mathcal{J}_*$ *for some scalar constant* $c > 0$.

Clearly, Assumption $7^{3*} \Rightarrow 7^* \Rightarrow 7$.

If $\ell_T(\theta)$ is a correctly specified log-likelihood function and Assumption 7^* holds, then the information matrix equality implies that Assumption 7^{3*} holds with $c = 1$. Assumption 7^{3*} also holds in some likelihood cases where the log-likelihood is misspecified. (For example, it occurs in a regression model with Gaussian quasi-log likelihood function when θ_* is the regression parameter, the autocorrelation and heteroskedasticity is correctly specified, and the errors are not actually Gaussian but have finite variance.)

Assumption 7^{3*} holds for LS estimators of regression models with $c = \sigma^2$ provided Assumption 7^* holds and the regression errors are homoskedastic conditional on the regressors with variance σ^2.

Assumption 7^{3*} holds for GMM and minimum distance estimators with $c = 1$ provided an asymptotically optimal weight matrix is employed. (An asymptotically optimal weight matrix is one that is (asymptotically) block diagonal between the θ_* and ψ parameters and has a consistent estimator of the inverse of the asymptotic covariance matrix of the moment conditions or of the unrestricted parameter estimator as its upper block.)

When Assumption 7^{3*} holds, the number of nuisance parameters that appear in the asymptotic distribution of $\widehat{\lambda}_\beta$ is reduced substantially. The distribution of $\widehat{\lambda}_\beta$ depends on the nuisance parameters (or nuisance random variables) in the inverse of the weight matrix of $q_\beta(\lambda_\beta)$, viz., $H\mathcal{J}_*^{-1}H'$, (but only up to scale) and in the (conditional) covariance matrix of Z_β $(:= H\mathcal{J}_*^{-1}G_\beta)$, viz., $H\mathcal{J}_*^{-1}\mathcal{I}_*\mathcal{J}_*^{-1}H'$.

Suppose Assumption 7^{3*} holds. Then, the matrix $H\mathcal{J}_*^{-1}\mathcal{I}_*\mathcal{J}_*^{-1}H'$ equals $cH\mathcal{J}_*^{-1}H'$ and knowledge of the former implies knowledge of the $H\mathcal{J}_*^{-1}H'$ up to scale. Thus, the number of nuisance parameters in the distribution of $\widehat{\lambda}_\beta$ equals the number of nonredundant elements of $H\mathcal{J}_*^{-1}\mathcal{I}_*\mathcal{J}_*^{-1}H'$, which is $p(p+1)/2$. This is the same number of nuisance parameters as in the standard case considered in the literature where the true parameter θ_0 is not on a boundary. These nuisance parameters are all estimable.

Next, we consider the nuisance parameters that appear in the distribution of $\widehat{\lambda}_\beta'(H\mathcal{J}_*^{-1}H')^{-1}\widehat{\lambda}_\beta$. The latter is the part of the asymptotic distribution of the maximized objective function that depends on β. It is important for testing results.

Let D be a $p \times p$ (possibly random) matrix that is symmetric and nonsingular with probability one. Our leading choice for D is

$$D = Diag^{1/2}(H\mathcal{J}_*^{-1}H'). \tag{6.5}$$

Define $\widehat{\lambda}_{\beta D}$ such that $\widehat{\lambda}_{\beta D} \in$cl$(\Lambda_{\beta D})$ and

$$\begin{aligned} q_{\beta D}(\widehat{\lambda}_{\beta D}) &= \inf_{\lambda_\beta \in \Lambda_{\beta D}} q_{\beta D}(\lambda_\beta), \text{where } \Lambda_{\beta D} := D^{-1}\Lambda_\beta, \\ q_{\beta D}(\lambda_\beta) &:= (\lambda_\beta - Z_{\beta D})'(D^{-1}H\mathcal{J}_*^{-1}H'D^{-1})^{-1}(\lambda_\beta - Z_{\beta D}) \text{ for } \lambda_\beta \in R^p, \text{ and} \\ Z_{\beta D} &:= D^{-1}Z_\beta. \end{aligned} \tag{6.6}$$

Lemma 4. *For any $p \times p$ (possibly random) matrix D that is symmetric and nonsingular with probability one, $\widehat{\lambda}_\beta = D\widehat{\lambda}_{\beta D}$.*

Comments. 1. By Lemma 4, $\widehat{\lambda}_\beta'(H\mathcal{J}_*^{-1}H')^{-1}\widehat{\lambda}_\beta$ equals $\widehat{\lambda}_{\beta D}'(D^{-1}H\mathcal{J}_*^{-1}H'D^{-1})^{-1}\widehat{\lambda}_{\beta D}$. The distribution of this term depends on the nuisance parameters (or nuisance random variables) in $D^{-1}H\mathcal{J}_*^{-1}H'D^{-1}$ (but only up to scale) and in the (conditional) covariance matrix of $Z_{\beta D}$, viz., $D^{-1}H\mathcal{J}_*^{-1}\mathcal{I}_*\mathcal{J}_*^{-1}H'D^{-1}$, under Assumption 7^*. Suppose D is as in (6.5) and Assumption 7^{3*} holds. Then, the matrix $D^{-1}H\mathcal{J}_*^{-1}\mathcal{I}_*\mathcal{J}_*^{-1}H'D^{-1}$ equals $cD^{-1}H\mathcal{J}_*^{-1}H'D^{-1}$ and knowledge of the former implies knowledge of $D^{-1}H\mathcal{J}_*^{-1}H'D^{-1}$ up to scale. In this case, the total number of nuisance parameters reduces to $p(p-1)/2$.

2. The proof of Lemma 4 follows easily from the fact that $q_\beta(\lambda_\beta) = q_{\beta D}(D^{-1}\lambda_\beta)$.

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

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

Assumption 10. *$\Lambda_\beta = \{\lambda_\beta \in R^p : \Gamma_a \lambda_\beta = \mathbf{0},\ \Gamma_b \lambda_\beta \leq \mathbf{0}\}$, where $\Gamma := [\Gamma_a' \vdots \Gamma_b']'$ is a full row rank matrix.*

Assumption 10 holds in all of the examples considered in the paper. For Λ_β as in the Lemma 5, $\widehat{\lambda}_\beta$ is the solution to a quadratic programming (QP) problem with mixed linear equality and inequality constraints.

The following lemma provides a characterization of $\widehat{\lambda}_\beta$ when Assumption 10 holds.

Lemma 5. *Suppose that Assumptions 7–10 hold. Then, $\widehat{\lambda}_\beta = P_L Z_\beta$ 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 $\|\lambda_\beta\|_\beta^2 = \lambda_\beta'(H\mathcal{J}_*^{-1}H')^{-1}\lambda_\beta$. That is, $P_L = I_p - H\mathcal{J}_*^{-1}H'\Gamma_1'(\Gamma_1 H\mathcal{J}_*^{-1}H'\Gamma_1')^{-1}\Gamma_1$, where $\Gamma_1 := [\Gamma_a' \vdots \Gamma_{b1}']'$.*

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 5 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 5 yields the following closed form expression for $\widehat{\lambda}_\beta$.

Theorem 5. *Suppose that Assumptions 7–10 hold. Then,*
(a) *$\widehat{\lambda}_\beta = P_{L(\widehat{j})} Z_\beta$, where $\widehat{j}$ minimizes $CF_j := Z_\beta'\Gamma_j'(\Gamma_j H\mathcal{J}_*^{-1}H'\Gamma_j')^{-1}\Gamma_j Z_\beta$ over $j = 1, ..., 2^{p_b}$ for which $P_{L(j)} Z_\beta \in \Lambda_\beta$. Here, $L(j) := \{\ell \in R^p : \Gamma_a \ell = \mathbf{0},\ \Gamma_{bj}\ell = \mathbf{0}\}$, $\Gamma_j := [\Gamma_a' \vdots \Gamma_{bj}']'$, $P_{L(j)} = I_p - H\mathcal{J}_*^{-1}H'\Gamma_j'(\Gamma_j H\mathcal{J}_*^{-1}H'\Gamma_j')^{-1}\Gamma_j$, and $\{\Gamma_{bj} : j = 1, ...2^{p_b}\}$ consists of all the different matrices comprised of some (possibly zero) rows of Γ_b.*
(b) $\widehat{\lambda}_\beta = \sum_{j=1}^{2^{p_b}} P_{L(j)} Z_\beta \times 1(P_{L(j)} Z_\beta \in \Lambda_\beta) \times \prod_{k=1}^{2^{p_b}} 1\left(CF_j \leq CF_k \text{ or } P_{L(k)} Z_\beta \notin \Lambda_\beta\right).$
(c) *for any $p \times p$ (possibly random) matrix D that is symmetric and nonsingular with probability one, $\widehat{\lambda}_\beta = DP_{L_D(\widehat{j})} Z_{\beta D}$, where $\widehat{j}$ is as in part* (a)*, $Z_{\beta D} := D^{-1}Z_\beta$, and $P_{L_D(j)} := I_p - D^{-1}H\mathcal{J}_*^{-1}H'\Gamma_j'(\Gamma_j H\mathcal{J}_*^{-1}H'\Gamma_j')^{-1}\Gamma_j D$.*

As an example of Theorem 5, suppose $\Lambda_\beta = R^+ \times R^{p-1}$. Take $D = \mathrm{Diag}^{1/2}(H\mathcal{J}_*^{-1}H')$. Then,

$$\widehat{\lambda}_\beta = \begin{cases} DZ_{\beta D} & \text{if } Z_{\beta D1} \geq 0 \\ DP_{L_D} Z_{\beta D} & \text{otherwise,} \end{cases}$$

$$= \begin{cases} DZ_{\beta D} & \text{if } Z_{\beta D1} \geq 0 \\ D(0, Z_{\beta D2} - \rho_{12} Z_{\beta D1}, ..., Z_{\beta Dp} - \rho_{p1} Z_{\beta D1})' & \text{otherwise,} \end{cases} \quad \text{where}$$

$Z_{\beta D} := (Z_{\beta D1}', ..., Z_{\beta Dp}')' := D^{-1}Z_\beta,$

$\rho_{ij} = [D^{-1}H\mathcal{J}_*^{-1}H'D^{-1}]_{ij}$, for $i, j = 1, ...p,$ and

(6.7) $P_{L_D} := I_p - D^{-1}H\mathcal{J}_*^{-1}H'\Gamma_1'(\Gamma_1 H\mathcal{J}_*^{-1}H'\Gamma_1')^{-1}\Gamma_1 D,\ \Gamma_1 := (1, 0, ..., 0) \in R^{1\times p}.$

When $\Lambda_\beta = R^- \times R^{p-1}$, the inequality in (6.7) 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 $\widehat{\lambda}_\beta$ as an estimator of 0 is less than or equal to the mean squared error of each corresponding element of $Z_{\beta D}$ 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 7^{2*} holds (or Assumption 7^* holds with $\mu = 0$ and $E\mathcal{I}_* < \infty$).

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 6 and 7^{2*} hold (or Assumptions 6 and 7^* 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,

$$
\begin{aligned}
\widehat{\lambda}_\beta \; &= D P_{L_D(\widehat{\jmath})} Z_{\beta D}, \text{ where} \\
P_{L_D(\widehat{\jmath})} Z_{\beta D} := \; & Z_{\beta D} 1(Z_{\beta D1} > 0,\, Z_{\beta D2} > 0) \\
&+ \begin{pmatrix} Z_{\beta D1} - \rho_{21} Z_{\beta D2} \\ 0 \\ Z_{\beta D3} - \rho_{23} Z_{\beta D2} \\ \vdots \\ Z_{\beta Dp} - \rho_{2p} Z_{\beta D2} \end{pmatrix} 1(Z_{\beta D1} - \rho_{21} Z_{\beta D2} > 0,\, Z_{\beta D2} \le 0) \\
&+ \begin{pmatrix} 0 \\ Z_{\beta D2} - \rho_{12} Z_{\beta D1} \\ \vdots \\ Z_{\beta Dp} - \rho_{1p} Z_{\beta D1} \end{pmatrix} 1(Z_{\beta D1} \le 0, Z_{\beta D2} - \rho_{12} Z_{\beta D1} > 0), \qquad (6.8)
\end{aligned}
$$

where D and ρ_{ij} are as in (6.7). For the case where $\Lambda_\beta = R^- \times R^+ \times R^{p-2}$ (as occurs in the Dickey–Fuller Regression Example with $p = 2$), (6.8) holds but with the first of the two inequalities reversed in each of the indicator functions in the definition of $P_{L_D(\widehat{\jmath})} Z_{\beta D}$. Adjustments of (6.8) 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

$$\Lambda_\beta = \{\lambda_\beta \in R^p : \lambda_{\beta 1} \ge 0, \Gamma_a \lambda_\beta = \mathbf{0}\}, \qquad (6.9)$$

$\widehat{\lambda}_\beta$ is as defined in (6.7), but with $Z_{\beta D}$ replaced by $P_{\Gamma_a D} Z_{\beta D}$, where

$$P_{\Gamma_a D} := I_p - D^{-1} H \mathcal{J}_*^{-1} H' \Gamma_a' \left(\Gamma_a H \mathcal{J}_*^{-1} H' \Gamma_a' \right)^{-1} \Gamma_a D. \qquad (6.10)$$

For the case where Λ_β is of the form

$$\Lambda_\beta = \{\lambda_\beta \in R^p : \lambda_{\beta 1} \geq 0, \lambda_{\beta 2} \geq 0, \Gamma_a \lambda_\beta = \mathbf{0}\}, \tag{6.11}$$

$\widehat{\lambda}_\beta$ is as defined in (6.8), but with $Z_{\beta D}$ replaced by $P_{\Gamma_a D} Z_{\beta D}$.

One can simulate the distribution of $\widehat{\lambda}_\beta$ when Λ_β is as in Assumption 10 by simulating Z_β or $Z_{\beta D}$ and computing $\widehat{\lambda}_\beta$ 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 takes about 63 seconds using a PC with Pentium 90 processor. The procedure QPROG also appears to be quite accurate. Its solutions and the closed form solutions provided by Theorem 5 were found to differ by 10^{-14} or less across a number of trials.

Alternatively, one can use the formulae of Theorem 5 or the equations above to compute $\widehat{\lambda}_\beta$. These are easy to program because they only involve computing CF_j for $j = 1, ..., 2^{p_b}$, finding the value $\widehat{j}$ that maximizes CF_j, and then computing $\widehat{\lambda}_\beta = P_{L(\widehat{j})} Z_\beta$ or $\widehat{\lambda}_\beta = DP_{L_D(\widehat{j})} Z_{\beta D}$. 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 Estimators

In this section, we describe three different procedures for obtaining standard error estimators that are consistent whether or not the true parameter is on a boundary of the parameter space. Each of these methods actually provides a consistent estimator of the whole asymptotic distribution of the extremum estimator. As mentioned in the Introduction, the standard bootstrap does not yield consistent standard error estimators.

The first method is described as follows. Suppose the parameter space Θ is defined by equality and/or inequality constraints:

$$\Theta = \{\theta \in R^s : g_a(\theta) = \mathbf{0}, m(\theta) \leq \mathbf{0}\}. \tag{6.12}$$

Assume that the function $m(\cdot) : \Theta \to R^J$ that defines the inequality constraints is continuously differentiable at θ_0. Let $m(\theta) = (m_1(\theta), ..., m_J(\theta))'$.

For $j = 1, ..., J$, let $\{\eta_{Tj} : T \geq 1\}$ be a sequence of random variables (possibly constants) that satisfies

$$\eta_{Tj} \lambda_{\min}(B_T) \xrightarrow{p} \infty. \tag{6.13}$$

We specify a rule based on $m_j(\widehat{\theta})$ and η_{Tj} to determine which (if any) of the inequality constraints are binding at the true parameter. If

$$m_j(\widehat{\theta}) > -\eta_{Tj}, \tag{6.14}$$

then we presume that the $j-th$ constraint is binding. (Because this rule is essentially a one dimensional one-sided Wald test for some significance levels $\{\alpha_T : T \geq 1\}$ that converge to 0 as $T \to \infty$, the η_{Tj} 's could be chosen to be the critical values for such tests multiplied by an estimate of the standard error of $m_j(\widehat{\theta})$ based on the usual formulae that assume that $m_j(\widehat{\theta})$ is not on a boundary of the parameter space.)

Let $j_1, ..., j_k$ index the constraints that are presumed to be binding. We construct vectors $g_b(\theta)$ and $h(\theta)$ that correspond to the constraints that are determined to be binding and not binding, respectively, according to the rule above:

$$g_b(\theta) = (m_{j_1}(\theta), ..., m_{j_k}(\theta))' \text{ and } h(\theta) = (m_{j_{k+1}}(\theta), ..., m_J(\theta))'. \tag{6.15}$$

We calculate asymptotic standard errors (or any other feature of interest of the asymptotic distribution of $B_T(\widehat{\theta}_T - \theta_0)$) based on the assumption that the constraints that are presumed to be binding actually are binding. Thus, we presume that the true parameter is as in (2.4) of the Regression with Restricted Parameters Example 2. Then, the asymptotic distribution of the extremum estimator is as determined in the sections above and we can obtain standard error estimators by simulating the asymptotic distribution with any unknown parameters replaced by consistent estimators.

Consistency of the standard error estimators just described depends on whether the rule for determining which inequality constraints are binding is correct with probability that goes to one as $T \to \infty$. It is, given (6.13) and Assumptions 1–4, because

$$\begin{aligned}
P(m_j(\widehat{\theta}) > -\eta_{Tj}) &= P(m_j(\theta_0) + (\tfrac{\partial}{\partial\theta'} m_j(\theta_0) B_T^{-1}) B_T(\widehat{\theta} - \theta_0) + o(||\widehat{\theta} - \theta_0||) > -\eta_{Tj}) \\
&= \begin{cases} P(\frac{\partial}{\partial\theta'} m_j(\theta_0) B_T^{-1} O_p(1) \lambda_{\min}(B_T) + o_p(1) > -\eta_{Tj}\lambda_{\min}(B_T)) & \text{if } m_j(\theta_0) = 0 \\ P(m_j(\theta_0) + o_p(1) > -\eta_{Tj}) & \text{if } m_j(\theta_0) < 0 \end{cases} \\
&\to \begin{cases} 1 & \text{if } m_j(\theta_0) = 0 \\ 0 & \text{if } m_j(\theta_0) < 0. \end{cases}
\end{aligned} \tag{6.16}$$

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. More specifically, one uses the bootstrap distribution of $B_{T_1}(\widehat{\theta}^*_{T_1} - \widehat{\theta}_T)$ to estimate the distribution of $B_T(\widehat{\theta}_T - \theta_0)$, where $\widehat{\theta}_T$ denotes the estimator $\widehat{\theta}$ constructed using T observations and $\widehat{\theta}^*_{T_1}$ denotes the bootstrap estimator of $\widehat{\theta}$ constructed from T_1 observations. In an iid context, $\widehat{\theta}^*_{T_1}$ is constructed from T_1 iid draws 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 $T_1/T \to 0$ as $T \to \infty$. Typically, one approximates the distribution of $B_{T_1}(\widehat{\theta}^*_{T_1} - \widehat{\theta}_T)$

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 $B_T(\widehat{\theta}_T - \theta_0)$, 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

$$\theta_* := (\theta_1', \theta_2', \theta_3)',\ \psi := (\theta_4', \theta_5)',\ \beta := \theta_1, \text{ and } \delta := (\theta_2', \theta_3)'. \tag{6.17}$$

With this partitioning, Assumptions 8 and 9 hold. In particular, by (3.16), $\mathcal{J}$ is block diagonal between θ_* and ψ. The set Λ is a product set $\Lambda_\beta \times \Lambda_\delta \times \Lambda_\psi$ with

$$\Lambda_\beta := (R^+)^p,\ \Lambda_\delta := R^{b_2+1}, \text{ and } \Lambda_\psi := R^{b+1}. \tag{6.18}$$

Thus, Assumption 10 also holds.

With this partitioning, from (3.16) and (5.2), we have

$$\begin{aligned} &\mathcal{J}_* := \tfrac{1}{2} E W_t^2 W_t^{2\prime} / \text{var}_t^2(\theta_0),\ \mathcal{J}_\psi := E W_t W_t' / \text{var}_t(\theta_0), \\ &G := (G_*', G_\psi')' \sim N(\mathbf{0}, \mathcal{I}),\ G_* \sim N(\mathbf{0}, \mathcal{I}_*),\ G_\psi \sim N(\mathbf{0}, \mathcal{I}_\psi), \\ &\mathcal{I}_* := \tfrac{1}{4} E \frac{(\text{res}_t^2(\theta_0) - \text{var}_t(\theta_0))^2}{\text{var}_t^4(\theta_0)} W_t^2 W_t^{2\prime}, \text{ and } \mathcal{I}_\psi := E W_t W_t' / \text{var}_t(\theta_0). \end{aligned} \tag{6.19}$$

Assumption 7^{2*} holds with $\mathcal{I}_*$ as above. Assumption 7^{3*} holds if the errors ε_t and η_t are normally distributed.

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

$$\begin{aligned} &\widehat{\lambda}_\psi = Z_\psi := \mathcal{J}_\psi^{-1} G_\psi \sim N(\mathbf{0}, \mathcal{J}_\psi^{-1}) \text{ and} \\ &T^{1/2}((\widehat{\theta}_4', \widehat{\theta}_5)' - (\theta_{40}', \theta_{50})') \xrightarrow{d} \widehat{\lambda}_\psi \sim N(\mathbf{0}, (E W_t W_t' / \text{var}_t(\theta_0))^{-1}). \end{aligned} \tag{6.20}$$

Thus, the QML regression parameter estimators $\widehat{\theta}_4$ and $\widehat{\theta}_5$ are asymptotically normal with covariance matrix $\mathcal{J}_\psi^{-1}$ 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),

$$T^{1/2}(\widehat{\theta}_1 - \theta_{10}) \xrightarrow{d} \widehat{\lambda}_\beta, \tag{6.21}$$

where $\widehat{\lambda}_\beta$ satisfies

$$\begin{aligned} &q_\beta(\widehat{\lambda}_\beta) = \inf_{\lambda_\beta \in (R^+)^p} q_\beta(\lambda_\beta), \\ &q_\beta(\lambda_\beta) := (\lambda_\beta - Z_\beta)' \left(H \mathcal{J}_*^{-1} H' \right)^{-1} (\lambda_\beta - Z_\beta), \text{ and} \\ &Z_\beta := H \mathcal{J}_*^{-1} G_* \sim N(\mathbf{0},\ H \mathcal{J}_*^{-1} \mathcal{I}_* \mathcal{J}_*^{-1} H'). \end{aligned} \tag{6.22}$$

For example, if $p = 1$ (i.e., there is one random coefficient with zero variance), then Assumption 10 holds and by (6.7),

$$\widehat{\lambda}_\beta = DZ_{\beta D}1(Z_{\beta D} \geq 0) = Z_\beta 1(Z_\beta \geq 0). \tag{6.23}$$

Thus, $\widehat{\lambda}_\beta$ has a half-normal distribution. If $p > 1$, then $\widehat{\lambda}_\beta$ is given in closed form by (6.8) or Theorem 5.

Also by Corollary 1(b),

$$\begin{aligned}
&T^{1/2}((\widehat{\theta}_2', \widehat{\theta}_3)' - (\theta_{20}', \theta_{30})') \xrightarrow{d} \widehat{\lambda}_\delta,\\
&\widehat{\lambda}_\delta = \mathcal{J}_\delta^{-1} G_\delta - \mathcal{J}_\delta^{-1} \mathcal{J}_{\delta\beta} \widehat{\lambda}_\beta,\\
&\mathcal{J}_\delta := \tfrac{1}{2} E \begin{pmatrix} X_{2t}^2 \\ 1 \end{pmatrix} \begin{pmatrix} X_{2t}^2 \\ 1 \end{pmatrix}' / \mathrm{var}_t^2(\theta_0),\ X_{2t}^2 := (X_{tp+1}^2, ..., X_{tb}^2)' \in R^{b_2},\\
&\mathcal{J}_{\delta\beta} := \tfrac{1}{2} E \begin{pmatrix} X_{2t}^2 \\ 1 \end{pmatrix} X_{1t}^2,\ X_{1t}^2 := (X_{t1}^2, ..., X_{tp}^2)',\\
&G_* := \begin{pmatrix} G_\beta \\ G_\delta \end{pmatrix} \sim N(\mathbf{0}, \mathcal{I}_*), \text{ and } G_\delta \sim N\left(\mathbf{0}, \tfrac{1}{4} E \frac{(\mathrm{res}_t^2(\theta_0) - \mathrm{var}_t(\theta_0))^2}{\mathrm{var}_t^4(\theta_0)} \begin{pmatrix} X_{2t}^2 \\ 1 \end{pmatrix} \begin{pmatrix} X_{2t}^2 \\ 1 \end{pmatrix}' \right).
\end{aligned} \tag{6.24}$$

6.5.2. Regression with Restricted Parameters

Typically, the restrictions $g_a(\theta) = \mathbf{0}$ and $g_b(\theta) \leq \mathbf{0}$ of Example 2 only involve some of the elements of θ. In this case, the vector $\frac{\partial}{\partial \theta'} g(\theta_0)$, where $g(\theta) := (g_a(\theta)', g_b(\theta)')'$, that determines Λ contains some non-zero columns, say p of them, and some columns of zeros, say $s - p$ of them. Without loss of generality, assume that the first p columns of $\frac{\partial}{\partial \theta'} g(\theta_0)$ are non-zero vectors and the last $s - p$ columns are zero vectors for $1 \leq p \leq s$.

We partition X_t such that

$$\begin{aligned}
&X_t := \begin{pmatrix} X_{*t} \\ X_{\psi t} \end{pmatrix} := \begin{pmatrix} X_{\beta t} \\ X_{\delta t} \\ X_{\psi t} \end{pmatrix} \text{ and}\\
&\mathcal{J} = EX_t X_t' := \begin{bmatrix} \mathcal{J}_* & \mathbf{0} \\ \mathbf{0} & \mathcal{J}_\psi \end{bmatrix} := \begin{bmatrix} EX_{*t} X_{*t}' & \mathbf{0} \\ \mathbf{0} & EX_{\psi t} X_{\psi t}' \end{bmatrix},
\end{aligned} \tag{6.25}$$

where $X_{*t} \in R^{p+q}$, $X_{\psi t} \in R^r$, $X_{\beta t} \in R^p$, $X_{\delta t} \in R^q$, $\mathcal{J}_* \in R^{(p+q) \times (p+q)}$, and $\mathcal{J}_\psi \in R^{r \times r}$. Here p is the number of non-zero columns of $(\partial / \partial \theta') g(\theta_0)$ and (q, r) are taken such that $\mathcal{J}$ is block diagonal and $p + q + r = s$. Such a partitioning is always possible, because one could have $r = 0$.

We partition $\widehat{\theta}$, θ_0, and θ conformably with X_t. That is,

$$\widehat{\theta} := (\widehat{\beta}', \widehat{\delta}', \widehat{\psi}')',\ \theta_0 := (\beta_0', \delta_0', \psi_0')', \text{ and } \theta = (\beta', \delta', \psi')', \tag{6.26}$$

where $\widehat{\beta}, \beta_0, \beta \in R^p$, $\widehat{\delta}, \delta_0, \delta \in R^q$, and $\widehat{\psi}, \psi_0, \psi \in R^r$.

Now, with the above partitioning, Assumptions 8 and 9 hold. The matrix $\mathcal{J}$ is block diagonal by (6.25). The matrix B_T is diagonal. The set Λ is a product set $\Lambda_\beta \times \Lambda_\delta \times \Lambda_\psi$ with

$$\Lambda_\beta = \left\{\lambda_\beta \in R^p : \frac{\partial}{\partial \beta'} g_a(\theta_0)\lambda_\beta = \mathbf{0},\ \frac{\partial}{\partial \beta'} g_b(\theta_0)\lambda_\beta \leq \mathbf{0}\right\},\ \Lambda_\delta = R^q, \text{ and } \Lambda_\psi = R^r$$

$$\text{where } \frac{\partial}{\partial \beta'} g_j(\theta_0) \in R^{c_j \times p} \text{ for } j = a, b. \tag{6.27}$$

For example, if $g_j(\theta) = v_j'\theta - d_j$ for $j = a, b$, then

$$\Lambda_\beta = \{\lambda_\beta \in R^p : v_{a\beta}'\lambda_\beta = 0,\ v_{b\beta}'\lambda_\beta \leq 0\},$$
$$\text{where } v_j = (v_{j\beta}', \mathbf{0}')' \text{ and } v_{j\beta} \in R^p \text{ for } j = a, b. \tag{6.28}$$

Alternatively, if $g_j(\theta) = v_j'\theta^2 - d_j$ for $j = a, b$, then

$$\Lambda_\beta = \{\lambda_\beta \in R^p : (v_{a\beta} \odot \beta_0)'\lambda_\beta = 0,\ (v_{b\beta} \odot \beta_0)'\lambda_\beta \leq 0\},$$
$$\text{where } v_j = (v_{j\beta}', \mathbf{0}')' \text{ and } v_{j\beta} \in R^p \text{ for } j = a, b. \tag{6.29}$$

With conformable partitioning to that above, we have

$$\mathcal{I} := \begin{bmatrix} \mathcal{I}_* & \mathcal{I}_{*\psi} \\ \mathcal{I}_{\psi *} & \mathcal{I}_\psi \end{bmatrix} = \begin{bmatrix} E\varepsilon_t^2 X_{*t}X_{*t}' & E\varepsilon_t^2 X_{*t}X_{\psi t}' \\ E\varepsilon_t^2 X_{\psi t}X_{*t}' & E\varepsilon_t^2 X_{\psi t}X_{\psi t}' \end{bmatrix},$$
$$G := \begin{pmatrix} G_* \\ G_\psi \end{pmatrix} \sim N(\mathbf{0}, \mathcal{I}),\ G_* \sim N(\mathbf{0}, \mathcal{I}_*), \text{ and } G_\psi \sim N(\mathbf{0}, \mathcal{I}_\psi), \tag{6.30}$$

where $G_* \in R^{p+q}$ and $G_\psi \in R^r$. In this case, Assumption 7^{2*} holds. If the errors are homoskedastic conditional on X_{*t} with variance σ^2, i.e., $E(\varepsilon_t^2|X_{*t}) = \sigma^2$ a.s., then Assumption 7^{3*} holds.

By Corollary 1(b) and the fact that $\Lambda_\psi = R^r$, we have

$$T^{1/2}(\widehat{\psi} - \psi_0) \xrightarrow{d} \widehat{\lambda}_\psi = Z_\psi := \mathcal{J}_\psi^{-1} G_\psi \sim N(\mathbf{0}, \mathcal{J}_\psi^{-1}\mathcal{I}_\psi\mathcal{J}_\psi^{-1}). \tag{6.31}$$

Thus, the LS estimator of ψ_0 is asymptotically normal with covariance matrix $\mathcal{J}_\psi^{-1}\mathcal{I}_\psi\mathcal{J}_\psi^{-1}$ whether or not the restriction $g(\theta_0) \leq 0$ is satisfied as an equality.

By Corollary 1(b),

$$T^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{d} \widehat{\lambda}_\beta, \tag{6.32}$$

where $\widehat{\lambda}_\beta$ solves $q_\beta(\widehat{\lambda}_\beta) = \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta)$ with Λ_β as in (6.27) and $q_\beta(\lambda_\beta)$ defined using (6.25) and (6.30). In the simplest case where $p = 1$, which occurs when $g(\theta)$ places an upper or lower bound or an equality constraint on a single parameter at $\theta = \theta_0$, the closed form expression for $\widehat{\lambda}_\beta$ given in (6.7) is applicable. If $p > 1$, a closed form expression for $\widehat{\lambda}_\beta$ is given in (6.8) or Theorem 5.

By Corollary 1(b),

$$T^{1/2}(\widehat{\delta}-\delta_0) \xrightarrow{d} \widehat{\lambda}_\delta = \mathcal{J}_\delta^{-1}G_\delta - \mathcal{J}_\delta^{-1}\mathcal{J}_{\delta\beta}\widehat{\lambda}_\beta, \text{ where}$$

$$G_* := \begin{pmatrix} G_\beta \\ G_\delta \end{pmatrix},\ G_\delta \sim N(\mathbf{0}, \mathcal{I}_\delta),\ \mathcal{I}_\delta = E\varepsilon_t^2 X_{\delta t}X_{\delta t}',$$

$$\mathcal{J}_* := \begin{bmatrix} \mathcal{J}_\beta & \mathcal{J}_{\beta\delta} \\ \mathcal{J}_{\delta\beta} & \mathcal{J}_\delta \end{bmatrix} = \begin{bmatrix} EX_{\beta t}X_{\beta t}' & EX_{\beta t}X_{\delta t}' \\ EX_{\delta t}X_{\beta t}' & EX_{\delta t}X_{\delta t}' \end{bmatrix}, \tag{6.33}$$

$G_\delta \in R^q$, and $\mathcal{J}_\delta \in R^{q\times q}$.

6.5.3. Dickey–Fuller Regression

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

$$\theta_* = (\theta_1, \theta_2, \theta_3)',\ \psi := \theta_4,\ \ \beta := (\theta_1, \theta_2)',\ \text{and } \delta := \theta_3. \tag{6.34}$$

With this partitioning, Assumptions 8 and 9 hold by (3.29), (4.23), and (4.24). The set Λ is a product set $\Lambda_\beta \times \Lambda_\delta \times \Lambda_\psi$ with

$$\Lambda_\beta = \begin{cases} R^- \times R^+ & \text{if } \theta_{30} = 0 \\ R^- \times R & \text{if } \theta_{30} > 0, \end{cases} \quad \Lambda_\delta = R,\ \text{and } \Lambda_\psi = R^b. \tag{6.35}$$

With the above partitioning, from (3.29), we have

$$\mathcal{J}_* := \begin{pmatrix} \lambda^2\int_0^1 W^2(r)dr & \lambda\int_0^1 rW(r)dr & \lambda\int_0^1 W(r)dr \\ \lambda\int_0^1 rW(r)dr & 1/3 & 1/2 \\ \lambda\int_0^1 W(r)dr & 1/2 & 1 \end{pmatrix},\ \mathcal{J}_\psi = V,$$

$$G := \begin{pmatrix} G_* \\ G_\psi \end{pmatrix},\ G_* := \begin{pmatrix} \frac{1}{2}\sigma\lambda(W^2(1)-1) \\ \sigma(W(1) - \int_0^1 W(r)dr) \\ \sigma W(1) \end{pmatrix},\ \text{and } G_\psi := G_4, \tag{6.36}$$

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

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

$$\widehat{\lambda}_\psi = Z_\psi := \mathcal{J}_\psi^{-1}G_\psi \sim N(\mathbf{0}, V^{-1}). \tag{6.37}$$

By Theorem 4(a), $\widehat{\lambda}_\beta$ solves $q_\beta(\widehat{\lambda}_\beta) = \inf_{\lambda_\beta\in\Lambda_\beta} q_\beta(\lambda_\beta)$, where Λ_β is defined in (6.35). Closed form expressions for $\widehat{\lambda}_\beta$ are given in (6.8) (with the first inequality reversed in each indicator function) and (6.23) (with the inequality reversed) for the cases where $\theta_{30} = 0$ and $\theta_{30} > 0$ respectively. Given $\widehat{\lambda}_\beta$, Theorem 4(b) gives a closed form expression for $\widehat{\lambda}_\delta$:

$$\widehat{\lambda}_\delta = \mathcal{J}_\delta^{-1}G_\delta - \mathcal{J}_\delta^{-1}\mathcal{J}_{\delta\beta}\widehat{\lambda}_\beta := \sigma W(1) - (\lambda\textstyle\int_0^1 W(r)dr, 1/2)\widehat{\lambda}_\beta, \text{ where}$$

$$\mathcal{J}_\delta := 1,\ \mathcal{J}_{\delta\beta} := (\lambda\textstyle\int_0^1 W(r)dr, 1/2),\ \text{and } G_\delta := \sigma W(1). \tag{6.38}$$

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

We have

$$\Upsilon_T M(\widehat{\theta}-\theta_0) = \begin{pmatrix} T(\widehat{\theta}_1 - \theta_{10}) \\ T^{3/2}\mu_0(\widehat{\theta}_1 - \theta_{10}) + T^{3/2}(\widehat{\theta}_2 - \theta_{20}) \\ -T^{1/2}\mu_0(\widehat{\theta}_1 - \theta_{10}) + T^{1/2}(\widehat{\theta}_3 - \theta_{30}) + T^{1/2}\mu_0 \mathbf{1}'(\widehat{\theta}_4 - \theta_{40}) \\ T^{1/2}(\widehat{\theta}_4 - \theta_{40}) \end{pmatrix} \overset{d}{\to} \begin{pmatrix} \widehat{\lambda}_{\beta 1} \\ \widehat{\lambda}_{\beta 2} \\ \widehat{\lambda}_\delta \\ \widehat{\lambda}_\psi \end{pmatrix}, \tag{6.39}$$

where $\widehat{\lambda}_\beta := (\widehat{\lambda}_{\beta 1}, \widehat{\lambda}_{\beta 2})'$. Equation (6.39) provides the asymptotic distribution of the unit root estimator $\widehat{\theta}_1$ and of the short-run dynamics parameter estimator $\widehat{\theta}_4$ 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 (6.39) also provides the asymptotic distributions of nondegenerate linear combinations of the estimators $\widehat{\theta}_1, ..., \widehat{\theta}_4$ that include the time trend and intercept parameter estimators $\widehat{\theta}_2$ and $\widehat{\theta}_3$. From these, the asymptotic distributions of $\widehat{\theta}_2$ and $\widehat{\theta}_3$ can be determined quite easily. First, the second row of (6.39) implies that $T\mu_0(\widehat{\theta}_1 - \theta_{10}) + T(\widehat{\theta}_2 - \theta_{20}) \overset{p}{\to} 0$ and the first row implies that $T\mu_0(\widehat{\theta}_1 - \theta_{10}) \overset{d}{\to} \mu_0 \widehat{\lambda}_{\beta 1}$. Hence,

$$T(\widehat{\theta}_2 - \theta_{20}) \overset{d}{\to} -\mu_0 \widehat{\lambda}_{\beta 1}. \tag{6.40}$$

Thus, the asymptotic joint distribution of $(T(\widehat{\theta}_1 - \theta_{10}), T(\widehat{\theta}_2 - \theta_{20}))'$ is $(\widehat{\lambda}_{\beta 1}, -\mu_0 \widehat{\lambda}_{\beta 1})'$, which is singular. Second, by the first, third, and fourth rows of (6.39),

$$T^{1/2}(\widehat{\theta}_3 - \theta_{30}) \overset{d}{\to} \widehat{\lambda}_\delta - \mu_0 \mathbf{1}' \widehat{\lambda}_\psi, \tag{6.41}$$

because $-T^{1/2}\mu_0(\widehat{\theta}_1 - \theta_{10}) = o_p(1)$. Hence, (6.39) yields the asymptotic distributions of all of the elements of $\widehat{\theta}$ and their convergence holds jointly.

6.6. Proofs

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

$$q_*(\lambda_*) := (\lambda_* - Z_*)' \mathcal{J}_* (\lambda_* - Z_*). \tag{6.42}$$

By Assumption 8,

$$\begin{aligned} q(\lambda) &= q_*(\lambda_*) + q_\psi(\lambda_\psi) \text{ for } \lambda = (\lambda_*', \lambda_\psi')', \\ \inf_{\lambda \in \Lambda} q(\lambda) &= \inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*) + \inf_{\lambda_\psi \in \Lambda_\psi} q_\psi(\lambda_\psi), \text{ and} \\ Z'\mathcal{J}Z &= Z_*'\mathcal{J}_* Z_* + Z_\psi' \mathcal{J}_\psi Z_\psi. \end{aligned} \tag{6.43}$$

Next, we have

$$\begin{aligned} 0 &\le q_*(\widehat{\lambda}_*) - \inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*) \\ &\le q_*(\widehat{\lambda}_*) - \inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*) + q_\psi(\widehat{\lambda}_\psi) - \inf_{\lambda_\psi \in \Lambda_\psi} q_\psi(\lambda_\psi) \\ &= q(\widehat{\lambda}) - \inf_{\lambda \in \Lambda_\beta \times \Lambda_\delta \times \Lambda_\psi} q(\lambda) \\ &= 0, \end{aligned} \tag{6.44}$$

where the first equality uses (6.43) and the second holds by the definition of $\widehat{\lambda}$. In consequence, we obtain

$$q_*(\widehat{\lambda}_*) = \inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*). \tag{6.45}$$

Part (c) of the Theorem follows from (6.43) and (6.45) .

We now use Assumption 9 to break $q_*(\lambda_*)$ and $Z_*'\mathcal{J}_*Z_*$ into terms involving β and δ. Let

$$A := \begin{bmatrix} I_p \\ -\mathcal{J}_\delta^{-1}\mathcal{J}_{\delta\beta} \end{bmatrix} \in R^{(p+q)\times p},\ P^\perp := AH \in R^{(p+q)\times(p+q)},\text{ and } P := I_{p+q} - P^\perp. \tag{6.46}$$

Define the norm $||\cdot||_*$ on R^{p+q} by $||h||_* = (h'\mathcal{J}_*h)^{1/2}$ 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, $(Ph_1)'\mathcal{J}_*P^\perp h_2 = 0\ \forall h_1, h_2 \in R^{p+q}$. By some algebra,

$$A'\mathcal{J}_*A = (H\mathcal{J}_*^{-1}H')^{-1} \text{ and } P\mathcal{J}_*^{-1}G_* = \begin{bmatrix} \mathbf{0} \\ \mathcal{J}_\delta^{-1}G_\delta \end{bmatrix}. \tag{6.47}$$

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

The above results give

$$\begin{aligned} Z_*'\mathcal{J}_*Z_* &= (P^\perp Z_*)'\mathcal{J}_*P^\perp Z_* + (PZ_*)'\mathcal{J}_*PZ_* \\ &= Z_\beta'(H\mathcal{J}_*^{-1}H')^{-1}Z_\beta + G_\delta'\mathcal{J}_\delta^{-1}G_\delta. \end{aligned} \tag{6.48}$$

Equations (6.43) and (6.48) establish part (d) of the Theorem.

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

$$P\lambda_* = \begin{pmatrix} \mathbf{0} \\ \lambda_\delta + \mathcal{J}_\delta^{-1}\mathcal{J}_{\delta\beta}\lambda_\beta \end{pmatrix}. \tag{6.49}$$

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

$$q_\delta(\lambda_\beta, \lambda_\delta) := (\lambda_\delta + \mathcal{J}_\delta^{-1}\mathcal{J}_{\delta\beta}\lambda_\beta - \mathcal{J}_\delta^{-1}G_\delta)'\mathcal{J}_\delta(\lambda_\delta + \mathcal{J}_\delta^{-1}\mathcal{J}_{\delta\beta}\lambda_\beta - \mathcal{J}_\delta^{-1}G_\delta). \tag{6.50}$$

Then, using (6.47) and (6.49), we have

$$\begin{aligned} q_*(\lambda_*) &= (P^{\perp}\lambda_* - P^{\perp}Z_*)'\mathcal{J}_*(P^{\perp}\lambda_* - P^{\perp}Z_*) + (P\lambda_* - PZ_*)'\mathcal{J}_*(P\lambda_* - PZ_*) \\ &= q_\beta(\lambda_\beta) + q_\delta(\lambda_\beta, \lambda_\delta). \end{aligned} \tag{6.51}$$

Under Assumption 9, for any $\lambda_\beta \in R^p$, we have

$$\inf_{\lambda_\delta \in \Lambda_\delta} q_\delta(\lambda_\beta, \lambda_\delta) = \inf_{\lambda_\delta \in R^q} q_\delta(\lambda_\beta, \lambda_\delta) = 0. \tag{6.52}$$

Thus, using (6.51) and (6.52), we obtain

$$\inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*) = \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta). \tag{6.53}$$

Equations (6.43) and (6.53) establish part (e) of the Theorem.

Part (a) of the Theorem follows from

$$\begin{aligned} 0 &\le q_\beta(\widehat{\lambda}_\beta) - \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta) \le q_\beta(\widehat{\lambda}_\beta) + q_\delta(\widehat{\lambda}_\beta, \widehat{\lambda}_\delta) - \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta) \\ &= q_*(\widehat{\lambda}_*) - \inf_{\lambda_* \in \Lambda_\beta \times \Lambda_\delta} q_*(\lambda_*) \le 0, \end{aligned} \tag{6.54}$$

where the equality holds by (6.51) and (6.53) using Assumption 9.

By equation (6.51),

$$q_*(\widehat{\lambda}_*) = q_\beta(\widehat{\lambda}_\beta) + q_\delta(\widehat{\lambda}_\beta, \widehat{\lambda}_\delta). \tag{6.55}$$

By equations (6.45) and (6.53), $q_*(\widehat{\lambda}_*) = \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta)$. This, (6.55), and part (a) of the Theorem give $q_\delta(\widehat{\lambda}_\beta, \widehat{\lambda}_\delta) = 0$. The latter and (6.50) 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 applying the argument of (4.6) and (4.7) twice. □

Proof of Lemma 5. 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 $||\ell - z||_\beta$ 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 $\widehat{\lambda}_\beta \in \Lambda_\beta$, we can construct two matrices Γ_{b1} and Γ_{b2} such that $\Gamma_{b1}\widehat{\lambda}_\beta = \mathbf{0}$ and $\Gamma_{b2}\widehat{\lambda}_\beta < \mathbf{0}$ (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, $\Gamma_a\widehat{\lambda}_\beta = \mathbf{0}$. Let $L := \{\ell \in R^p : \Gamma_a \ell = \mathbf{0},\ \Gamma_{b1}\ell = \mathbf{0}\}$. For some $\varepsilon > 0$, $L \cap S(\widehat{\lambda}_\beta, \varepsilon) = \Lambda_\beta \cap S(\widehat{\lambda}_\beta, \varepsilon)$, because the restrictions $\Gamma_{b2}\ell < \mathbf{0}$ are satisfied for ℓ close to $\widehat{\lambda}_\beta$. By definition, $\widehat{\lambda}_\beta$ minimizes $\|\lambda_\beta - Z_\beta\|_\beta$ over $\lambda_\beta \in \Lambda_\beta \cap S(\widehat{\lambda}_\beta, \varepsilon)$. Hence, $\widehat{\lambda}_\beta$ minimizes the same function over $\lambda_\beta \in L \cap S(\widehat{\lambda}_\beta, \varepsilon)$ as well. By the first paragraph of the proof, then, $\widehat{\lambda}_\beta$ equals the projection of Z_β onto the linear subspace L. □

7. Sufficient Conditions for the Quadratic Approximation of the Objective Function

In this section we provide two sufficient conditions for Assumption 1. The first 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. No such Taylor expansions are available in the literature, that we are aware of, so our first task is to establish one for some arbitrary non-stochastic function f. Then, we apply this Taylor expansion to $\ell_T(\theta)$ to give the first sufficient condition for Assumption 1.

The second sufficient condition does not require pointwise smoothness of $\ell_T(\theta)$. Instead, it relies on a stochastic differentiability condition analogous to that of Pollard (1985) and on the smoothness of the probability limit $\ell(\theta)$ of $T^{-1}\ell_T(\theta)$. For the latter, we again use the Taylor expansion referred to above that does not require $\ell(\theta)$ to be defined on a neighborhood of θ_0.

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

Let f be a function whose domain includes $\mathcal{X} \subset R^s$. Let $a \in \mathcal{X}$. We want to derive a Taylor expansion of $f(x)$ about $f(a)$ 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 establish a Taylor expansion that only requires f to be defined on $\mathcal{X}$ and whose terms depend on left, right, and/or two-sided partial derivatives of f for points x in $\mathcal{X}$.

We now introduce some terminology. 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). For $x \in \mathcal{X}$, let $\frac{\partial}{\partial x_j} f(x)$ denote the l/r partial derivative wrt x_j (the j-th element of x) of f at x.

Note that the shape of $\mathcal{X}$ is such that $\forall x \in \mathcal{X}$ 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 $\forall x \in \mathcal{X}$. This is not the case for some other shapes for $\mathcal{X}$. For example, if $\mathcal{X}$ is the cone $\{x \in R^2 : x_1 \geq 0,\, x_1 \leq x_2\}$, then x_2 cannot be perturbed to the left or right within $\mathcal{X}$ at $x = \mathbf{0}$.

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 $\mathcal{X}$. Let $(\partial^k/\partial x_{i_1}, ..., \partial x_{i_k}) f(x)$ denote the k-th order l/r partial derivative of f at x wrt $x_{i_1}, ..., x_{i_k}$, where i_ℓ is a positive integer less than $s+1$ $\forall \ell \leq k$. 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 $\mathcal{X}$.

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*

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

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

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

Comment. If the l/r partial derivatives of f of order k are continuous wrt $\mathcal{X}$ 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).

7.2. A Sufficient Condition Via Smoothness

The first sufficient condition for Assumption 1 is the following:

Assumption 1*. (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* 1.
(b) $\ell_T(\theta)$ *has continuous l/r partial derivatives of order* 2 *on* Θ^+ $\forall T \geq 1$ *with probability one.*
(c) *For all* $\gamma_T \to 0$,

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

where $\frac{\partial}{\partial \theta} \ell_T(\theta)$ and $\frac{\partial^2}{\partial \theta \partial \theta'} \ell_T(\theta)$ 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 1*(a) specifies a set Θ^+ on which $\ell_T(\theta)$ must be defined. Near θ_0, it must be a union of orthants centered at θ_0. For each $\theta \in \Theta^+$, $\ell_T(\theta)$ has a quadratic approximation via Theorem 6. On the other hand, Assumption 1* does not require that near θ_0 the parameter space Θ *is* a union of orthants centered at θ_0. What Assumption 1* requires is that Θ is contained in such a set near θ_0. For example, we could have $\theta_0 = \mathbf{0}$, $\Theta^+ = \{\theta \in R^2 : \theta = (\theta_1, \theta_2)',\ \theta_1 \geq 0,\ \theta_2 \geq 0,\ ||\theta|| \leq b_1\}$ and $\Theta = \{\theta \in R^2 : \theta = (\theta_1, \theta_2)',\ \theta_1 \geq 0,\ \theta_1 \leq \theta_2,\ ||\theta|| \leq b_2\}$ for some constants $0 < b_1, b_2 \leq \infty$. In this case, the parameter space Θ is not a union of orthants near θ_0, but Assumption 1*(a) still holds. If Θ happens to be a union of orthants local to θ_0, then one can take $\Theta^+ = \Theta$ or $\Theta^+ = \Theta \cap C(\theta_0, \varepsilon)$ in Assumption 1*.

Assumption 1*(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 1*(c) can be verified in the case of non-trending data as follows. Suppose $B_T = T^{1/2}M$, $\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta)/T \xrightarrow{p} \frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta)$ uniformly over $\theta \in \Theta \cap S(\theta_0, \varepsilon_2)$ for some $\varepsilon_2 > 0$ and some non-random function $\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta)$ that is continuous at θ_0. Then Assumption 1*(c) holds. The uniform convergence of $\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta)/T$ can be established via a uniform law of large numbers, e.g., see Andrews (1992).

When stochastic or deterministic trends enter the objective function in a linear fashion, then part of the matrix $\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta)$ does not depend on θ and Assumption 1*(c) holds trivially for that part of $\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta)$.

Lemma 6. (a) *Assumption* 1* *implies Assumption* 1 *with* $D\ell_T(\theta_0)$ *and* $D^2\ell_T(\theta_0)$ *of* (3.1) *given by* $\frac{\partial}{\partial\theta}\ell_T(\theta_0)$ *and* $\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta_0)$ *(i.e., by the* ℓ/r *partial derivatives of* $\ell_T(\theta)$ *at* θ_0 *of orders one and two) respectively.*
(b) *If Assumption* 1* *holds and* $-B_T^{-1'}\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta_0)B_T^{-1} \xrightarrow{p} \mathcal{J}$ *for some non-random matrix* $\mathcal{J}$*, then Assumption* 1 *holds with* $D\ell_T(\theta_0)$ *of* (3.1) *given by* $\frac{\partial}{\partial\theta}\ell_T(\theta_0)$ *and* $D^2\ell_T(\theta_0)$ *of* (3.1) *given by either* $\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta_0)$ *or* $-B_T'\mathcal{J}B_T$*.*

7.3. A Sufficient Condition Via Stochastic Differentiability

Pollard (1985) introduced a concept of stochastic differentiability, showed how it could be used to obtain the asymptotic normality of extremum estimators, and gave sufficient conditions for it using a Huber-type bracketing condition, an empirical process bracketing condition, and a combinatorial Vapnik–Cervonenkis condition. The stochastic differentiability condition allows one to consider objective functions that are not pointwise differentiable. These include the objective functions for quantile regression estimators, censored quantile regression estimators, Huber regression M-estimators, method of simulated moments estimators, etc.

Here we extend Pollard's stochastic differentiability concept to cover differentiability at a boundary point of the parameter space Θ and to cover sequences of random functions that are not necessarily averages of iid functions. We say that a sequence of random functions $\{g_T(\theta) : T \geq 1\}$ is *stochastically differentiable* at θ_0 for $\Theta \subset R^s$ with random derivative s-vector $Dg_T(\theta_0)$ if

$$g_T(\theta) = g_T(\theta_0) + Dg_T(\theta_0)'(\theta - \theta_0) + r_T(\theta) \text{ and}$$

$$\text{(7.1)} \qquad \sup_{\theta \in \Theta : ||\theta - \theta_0|| \le \gamma_T} T|r_T(\theta)|/(1 + ||T^{1/2}(\theta - \theta_0)||)^2 = o_p(1)$$

for all $\gamma_T \to 0$. We apply this definition to random variables $g_T(\theta)$, such as samples averages, that are $O_p(1)$ $\forall \theta \in \Theta_0$ but not $o_p(1)$ $\forall \theta \in \Theta_0$, where Θ_0 denotes some neighborhood of θ_0, and for which $T^{1/2} D g_T(\theta_0) = O_p(1)$.

Note that the stochastic differentiability condition is weaker when θ_0 is on the boundary of Θ than when θ_0 is in the interior because the supremum is taken over a smaller set.

Also note that compared to Pollard's (1985) definition of stochastic differentiability, our definition is slightly more general even in the context that he considers. Specifically, his definition is obtained by taking $g_T(\theta)$ to be a sample average of iid random functions and by replacing the term $1/(1 + ||T^{1/2}(\theta - \theta_0)||)^2$ by $1/((1 + ||T^{1/2}(\theta - \theta_0)||)||T^{1/2}(\theta - \theta_0)||)$. Our formulation seems simpler. In addition, it allows one to easily obtain stochastic differentiability of a function $g_T(\theta)$ from the stochastic differentiability of an approximating function $g_T^*(\theta)$ provided $g_T^*(\theta)$ satisfies

$$\text{(7.2)} \qquad T \sup_{\theta \in \Theta : ||\theta - \theta_0|| \le \gamma_T} |g_T(\theta) - g_T(\theta_0) - g_T^*(\theta) + g_T^*(\theta_0)| = o_p(1)$$

for all $\gamma_T \to 0$. In this case, both $g_T(\theta)$ and $g_T^*(\theta)$ have the same random derivative s-vector $Dg_T^*(\theta_0)$. With Pollard's definition, (7.2) is not sufficient to yield stochastic differentiability of $g_T(\theta)$ given that of $g_T^*(\theta)$.

As mentioned above, empirical process results can be used to verify the stochastic differentiability condition. Pollard (1985) shows this for empirical processes based on iid random variables. For dependent data, the empirical process results for Doukhan, Massart, and Rio (1995) and Arcones and Yu (1994) can be used. For other references, see Andrews (1994b). Also, the Huber-type bracketing condition in Pollard (1985) applies with dependent random variables.

Our second sufficient condition for Assumption 1 is the following:

Assumption 1^{2*}. (a) $B_T = T^{1/2} I_s$.
(b) *For some non-random function* $\ell(\theta)$, $T^{-1}\ell_T(\theta) \xrightarrow{p} \ell(\theta)$ $\forall \theta \in \Theta \cap S(\theta_0, \varepsilon)$ *for some* $\varepsilon > 0$.
(c) *The domain of* $\ell(\theta)$ *includes a set* Θ^+ *that satisfies conditions* (i) *and* (ii) *of Assumption* 1^*(a).
(d) $\ell(\theta)$ *has continuous l/r partial derivatives* wrt θ *of order two on* Θ^+ *with l/r partial derivatives* $\frac{\partial}{\partial \theta}\ell(\theta)$ *and* $\frac{\partial^2}{\partial \theta \partial \theta'}\ell(\theta)$ *of orders one and two, respectively, that satisfy* $\frac{\partial}{\partial \theta}\ell(\theta_0) = \mathbf{0}$.
(e) $\{T^{-1}\ell_T(\theta) - \ell(\theta) : T \ge 1\}$ *is stochastically differentiable at* θ_0 *for* Θ *with random derivative vector denoted* $T^{-1}D\ell_T(\theta_0)$.

Assumption 1^{2*}(a) implies that Assumption 1^{2*} only applies to models without deterministically or stochastically trending variables. Assumptions 1^{2*}(b)–(d) require that the normalized objective function has a limit function that is smooth at θ_0, at least in the directions determined by Θ. This is a weaker condition than smoothness of the objective function itself because the limit function is typically an integral (by the law of large numbers) and integration is a smoothing operation. The assumption that $\frac{\partial}{\partial\theta}\ell(\theta_0) = \mathbf{0}$ in Assumption 1^{2*}(d) typically holds because the probability limit θ_0 of $\widehat{\theta}$ minimizes the probability limit $\ell(\theta)$ of $T^{-1}\ell_T(\theta)$ over Θ.

Lemma 7. *Assumption* 1^{2*} *implies Assumption* 1 *with* $D^2\ell_T(\theta_0) = -T\mathcal{J}$ *in* (3.1), *where* $\mathcal{J} := -\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta_0)$ *is the negative of the* $s \times s$ *matrix of l/r partial derivatives of* $\ell(\theta)$ *of order two at* θ_0.

7.4. Examples (Continued)

In this section, we use the results of the previous two subsections to verify Assumption 1 for the first three examples.

7.4.1. Random Coefficient Regression

We verify Assumption 1 using Assumption 1^* and Lemma 6(b). Let $\Theta^+ = \Theta \cap C(\theta_0, \varepsilon)$ for some $\varepsilon < \min\{M_j : j \leq 5\}$ (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 1^*(a)(i). Also, $\Theta \cap S(\theta_0, \varepsilon_1) \subset \Theta^+$ for $0 < \varepsilon_1 < \varepsilon$, as required by Assumption 1^*(a)(ii). The quasi-likelihood function $\ell_T(\theta)$ of (3.12) has continuous l/r partial derivatives of order two on Θ^+, as required by Assumption 1^*(b).

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

$$\text{(7.3)} \qquad \frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta) := -\sum_{t=1}^{T} \begin{pmatrix} \frac{2\text{res}_t^2(\theta) - \text{var}_t(\theta)}{\text{var}_t^3(\theta)} W_t^2 W_t^{2\prime} & \frac{\text{res}_t(\theta)}{\text{var}_t^2(\theta)} W_t W_t^{2\prime} \\ \frac{\text{res}_t(\theta)}{\text{var}_t^2(\theta)} W_t^2 W_t' & \frac{1}{\text{var}_t(\theta)} W_t W_t' \end{pmatrix}.$$

By a uniform LLN (e.g., see Andrews (1992, Theorem 4) using Assumption TSE–1C),

$$\text{(7.4)} \qquad \sup_{\theta \in \Theta} \left| T^{-1}\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta) - T^{-1}E\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta) \right| \xrightarrow{p} 0.$$

Also, $T^{-1}E\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta)$ is continuous at θ_0. In consequence, Assumption 1^*(c) holds. By (7.4), $-T^{-1}\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta_0) \xrightarrow{p} \mathcal{J}$, where $\mathcal{J}$ is defined in (3.16). In consequence, Lemma 6(b) is applicable and Assumption 1 holds with $D\ell_T(\theta_0)$ and $D^2\ell_T(\theta_0)$ of (3.1) as defined in (3.16).

7.4.2. Regression with Restricted Parameters

Assumption 1 holds trivially in this example because $R_T(\theta) = 0$.

7.4.3. Dickey–Fuller Regression

Assumption 1 holds trivially in this example because $R_T(\theta) = 0$.

7.5. Proofs

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 $\mathcal{X}$. That is, $f(x + h) = f(x) + A'h + \varepsilon_h' h$ and $||\varepsilon_h|| \to 0$ as $||h|| \to 0\ \forall x+h \in \mathcal{X}$ 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 $F(\lambda)$, $\frac{d}{d\lambda}F(\lambda)$, ..., $\frac{d^n}{d\lambda^n}F(\lambda)$ 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. □

Proof of Lemma 6. We prove part (a) first. By the Taylor expansion of Theorem 6, $\ell_T(\theta)$ satisfies (3.1) with

$$R_T(\theta) = \tfrac{1}{2}(\theta - \theta_0)' \left(\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta_0) \right) (\theta - \theta_0), \tag{7.5}$$

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

$$\begin{aligned} &\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T} |R_T(\theta)|/(1 + ||B_T(\theta - \theta_0)||)^2 \\ \leq &\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T} \tfrac{1}{2}|(B_T(\theta - \theta_0))'B_T^{-1\prime}(\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta_0))B_T^{-1} \end{aligned}$$

$$\times B_T(\theta-\theta_0)|/||B_T(\theta-\theta_0)||^2$$

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

$$= o_p(1), \tag{7.6}$$

where the equality holds by Assumption 1*(c).

Part (b) follows from part (a) because the difference between the third summand on the right-hand side of (3.1) defined with $D^2\ell_T(\theta_0)=\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta_0)$ and it defined with $D^2\ell_T(\theta_0)=-B_T'\mathcal{J}B_T$ can be absorbed in the $R_T(\theta)$ term without affecting Assumption 1, due to the $1/(1+||B_T(\theta-\theta_0)||)^2$ factor in Assumption 1. □

Proof of Lemma 7. Define $r_T(\theta)$ via

$$T^{-1}\ell_T(\theta)-\ell(\theta)=T^{-1}\ell_T(\theta_0)-\ell(\theta_0)+T^{-1}D\ell_T(\theta_0)'(\theta-\theta_0)+r_T(\theta). \tag{7.7}$$

By Assumption 1^{2*}(e), $r_T(\theta)$ satisfies (7.1).

By Theorem 6 and Assumptions 1^{2*}(c) and 1^{2*}(d), a Taylor expansion of order two of $\ell(\theta)$ about θ_0 gives: $\forall\theta\in\Theta\cap C(\theta_0,\varepsilon)$,

$$\begin{aligned}\ell(\theta) &= \ell(\theta_0)+\frac{\partial}{\partial\theta'}\ell(\theta_0)(\theta-\theta_0)+\tfrac{1}{2}(\theta-\theta_0)'\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta^\dagger)(\theta-\theta_0)\\ &= \ell(\theta_0)-\tfrac{1}{2}(\theta-\theta_0)'\mathcal{J}(\theta-\theta_0)+o(||\theta-\theta_0||^2),\end{aligned} \tag{7.8}$$

where $\theta^\dagger$ lies on the line segment joining θ and θ_0. The second equality holds because $\frac{\partial}{\partial\theta}\ell(\theta_0)=\mathbf{0}$ and $\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta)$ is continuous at θ_0 by Assumption 1^{2*}(d).

Combining (7.7) and (7.8) gives

$$\begin{aligned}T^{-1}\ell_T(\theta) &= T^{-1}\ell_T(\theta_0)+T^{-1}D\ell_T(\theta_0)'(\theta-\theta_0)-\tfrac{1}{2}(\theta-\theta_0)'\mathcal{J}(\theta-\theta_0)\\ &\quad+o(||\theta-\theta_0||^2)+r_T(\theta).\end{aligned} \tag{7.9}$$

Combining (7.9) with (3.1) multiplied by T^{-1} and with $D^2\ell_T(\theta_0)=-T\mathcal{J}$ yields

$$R_T(\theta)=Tr_T(\theta)+o(T||\theta-\theta_0||^2). \tag{7.10}$$

This and (7.1) yield Assumption 1. □

8. Proof of Consistency for the Examples

In this section we verify Assumption 4 (consistency) for each of the first three examples.

8.1. Random Coefficient Regression

Consistency of the QML estimator of Example 1 is established by verifying Assumptions 4*(a) and 4*(b*). Assumption 4*(a) holds by the uniform LLN of Andrews

(1992, Theorem 4) using Assumption TSE–1D and the standard pointwise LLN for iid random variables with finite mean. The function $\ell(\theta)$ of Assumption 4*(a) is

$$\begin{aligned}\ell(\theta) :=& -\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}E\ln(\theta_3 + X_t' D(\theta_1,\theta_2)X_t)\\ &- \tfrac{1}{2}E(Y_t - \theta_5 - X_t'\theta_4)^2/(\theta_3 + X_t'D(\theta_1,\theta_2)X_t).\end{aligned} \tag{8.1}$$

To verify Assumption 4*(b*), we note that the function $\ell(\theta)$ is continuous on Θ and Θ is compact. It remains to show that $\ell(\theta)$ has a unique maximum on Θ at θ_0. First we show that for any $(\theta_1, \theta_2, \theta_3)$ in the parameter space the third summand of $\ell(\theta)$ is uniquely maximized by $(\theta_4, \theta_5) = (\theta_{40}, \theta_{50})$. Note that the third summand of $\ell(\theta)$ can be written as $-\frac{1}{2}E(\theta_5 - \theta_{50} + X_t'(\theta_4 - \theta_{40}))^2/(\theta_3 + X_t'D(\theta_1,\theta_2)X_t)$. In consequence, the third summand is uniquely maximized at $(\theta_{40}, \theta_{50})$ if and only if (iff) $E(a'W_t)^2/(\theta_3 + X_t'D(\theta_1,\theta_2)X_t) > 0$ whenever $a \neq 0$ iff $E(a'W_t)^2/(\theta_{30} + X_t'D(\theta_{10},\theta_{20})X_t) > 0$ whenever $a \neq 0$ iff $EW_tW_t'/\text{var}_t(\theta_0) > 0$, where the second "iff" holds because $(\theta_{30} + X_t'D(\theta_{10},\theta_{20})X_t)/(\theta_3 + X_t'D(\theta_1,\theta_2)X_t)$ is positive with probability one. The last condition holds by (3.18).

Next, we show that, for any parameter $\theta = (\theta_1', \theta_2', \theta_3, \theta_{40}', \theta_{50})'$, $\ell(\theta)$ is uniquely maximized when $(\theta_1', \theta_2', \theta_3)' = (\theta_{10}', \theta_{20}', \theta_{30})'$. For θ as above, $\ell(\theta)$ can be written as

$$\ell(\theta) = -\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}E\ln(\theta_3 + X_t'D(\theta_1,\theta_2)X_t) - \tfrac{1}{2}E\frac{\theta_{30} + X_t'D(\theta_{10},\theta_{20})X_t}{\theta_3 + X_t'D(\theta_1,\theta_2)X_t}. \tag{8.2}$$

The function $\ln x + y/x$ is uniquely minimized over $x \in R$ at $x = y$. Thus, $\ell(\theta)$ is minimized by any vector $(\theta_1', \theta_2', \theta_3, \theta_{40}', \theta_{50})'$ for which $P(\theta_3 + X_t'D(\theta_1,\theta_2)X_t = \theta_{30} + X_t'D(\theta_{10},\theta_{20})X_t) = 1$. The latter holds only if $(\theta_1', \theta_2', \theta_3)' = (\theta_{10}', \theta_{20}', \theta_{30})'$, because $EW_t^2W_t^{2'}/\text{var}_t^2(\theta_0) > 0$ (by (3.18)) implies that $E(a'W_t^2)^2/\text{var}_t^2(\theta_0) > 0$ $\forall a \neq \mathbf{0}$, which implies that $P(a'W_t^2 = 0) = 1$ only if $a = \mathbf{0}$, which in turn implies that $P((\theta_1' - \theta_{10}', \theta_2' - \theta_{20}', \theta_3 - \theta_{30})W_t^2 = 0) = 1$ only if $(\theta_1', \theta_2', \theta_3)' = (\theta_{10}', \theta_{20}', \theta_{30})'$, as desired.

Thus, Assumption 4*(b*) holds and the verification of Assumption 4 is complete.

8.2. Regression with Restricted Parameters

Consistency of the restricted LS estimator $\widehat{\theta}$ (Assumption 4) is established as follows. Because $\widehat{\theta}$ maximizes $\ell_T(\theta)$ over Θ up to $o_p(1)$ and $\theta_0 \in \Theta$, we have

$$\sum_{t=1}^{T}(Y_t - X_t'\widehat{\theta})^2 \leq \sum_{t=1}^{T}(Y_t - X_t'\theta_0)^2 + o_p(1). \tag{8.3}$$

Some algebraic manipulations yield

$$(\widehat{\theta} - \theta_0)'\sum_{t=1}^{T}X_tX_t'(\widehat{\theta} - \theta_0) \leq 2\sum_{t=1}^{T}\varepsilon_tX_t'(\widehat{\theta} - \theta_0) + o_p(1). \tag{8.4}$$

Thus, using the Cauchy–Schwarz inequality, we have

$$\left\| \left(\sum_{t=1}^{T} X_t X_t' \right)^{1/2} (\widehat{\theta} - \theta_0) \right\|^2 \le 2 \left\| \left(\sum_{t=1}^{T} X_t X_t' \right)^{-1/2} \sum_{t=1}^{T} \varepsilon_t X_t \right\| \times \left\| \left(\sum_{t=1}^{T} X_t X_t' \right)^{1/2} (\widehat{\theta} - \theta_0) \right\| + o_p(1). \tag{8.5}$$

Completing the square yields

$$\left(\left\| \left(\sum_{t=1}^{T} X_t X_t' \right)^{1/2} (\widehat{\theta} - \theta_0) \right\| - \left\| \left(\sum_{t=1}^{T} X_t X_t' \right)^{-1/2} \sum_{t=1}^{T} \varepsilon_t X_t \right\| \right)^2 \le \left\| \left(\sum_{t=1}^{T} X_t X_t' \right)^{-1/2} \sum_{t=1}^{T} \varepsilon_t X_t \right\|^2 + o_p(1). \tag{8.6}$$

From (8.6), if

$$\left(\sum_{t=1}^{T} X_t X_t' \right)^{-1/2} \sum_{t=1}^{T} \varepsilon_t X_t = O_p(1), \tag{8.7}$$

then $\left(\sum_{t=1}^{T} X_t X_t' \right)^{1/2} (\widehat{\theta} - \theta_0) = O_p(1)$. The latter, the fact that $T^{-1} \sum_{t=1}^{T} X_t X_t' \xrightarrow{p} E X_t X_t' > 0$ implies that $\widehat{\theta} \xrightarrow{p} \theta_0$ and Assumption 4 holds.

Equation (8.7) holds by the CLT and the LLN because $T^{-1/2} \sum_{t=1}^{T} \varepsilon_t X_t' = O_p(1)$ and $\left(T^{-1} \sum_{t=1}^{T} X_t X_t' \right)^{-1/2} = (E X_t X_t')^{-1/2} + o_p(1)$. This complete the verification of Assumption 4.

8.3. Dickey–Fuller Regression

Consistency of $\widehat{\theta}$ for Example 3 is proved in a manner similar to the proof for Example 2. By replacing $X_t'(\theta - \theta_0)$ with $(X_t' B_T^{-1}) B_T (\theta - \theta_0)$ in equations (8.4)–(8.7), we obtain the following analogue of (8.7). If $\left(B_T^{-1\prime} \sum_{t=1}^{T} X_t X_t' B_T^{-1} \right)^{-1/2} B_T^{-1\prime} \sum_{t=1}^{T} \varepsilon_t X_t = O_p(1)$, then $\left(B_T^{-1\prime} \sum_{t=1}^{T} X_t X_t' B_T^{-1} \right)^{-1/2} B_T (\widehat{\theta} - \theta_0) = O_p(1)$. By (3.29), the former condition holds and $\left(B_T^{-1\prime} \sum_{t=1}^{T} X_t X_t' B_T^{-1} \right)^{1/2}$ converges in distribution to $\mathcal{J}^{1/2}$, a matrix that is nonsingular with probability one. Hence, $B_T(\widehat{\theta} - \theta_0) = O_p(1)$. Because $B_T = \Upsilon_T M$ with M nonsingular and $\lambda_{\min}(\Upsilon_T) \to \infty$, this yields $\widehat{\theta} \xrightarrow{p} \theta_0$ under θ_0.

9. Generalized Method of Moments and Minimum Distance Estimation

9.1. General Results

In this section, we show that the results of Section 3 are applicable to generalized method of moments (GMM) and minimum distance (MD) objective functions. We determine the components of the quadratic approximations of (3.1) and (3.3) for such objective functions and provide sufficient conditions for Assumptions 1–3. These sufficient conditions do not allow for stochastic or deterministic trends. They do allow for rescaled deterministic trends, however, as in Andrews and McDermott (1995).

The most general asymptotic results available for GMM/MD estimators are those of Pakes and Pollard (1989). Their results allow for a nondifferentiable objective function, as occurs with simulation estimators and with estimators for econometric models that exhibit kinks or discontinuities. Their asymptotic distributional results require the true parameter θ_0 to be an interior point. The results given here generalize theirs to allow θ_0 to be on a boundary. The conditions we provide are identical to those of Pakes and Pollard (1989) when θ_0 is an interior point. Furthermore, the stochastic equicontinuity condition that is required is slightly weaker when θ_0 is on a boundary than when it is an interior point. In consequence, the sufficient conditions for stochastic equicontinuity given in Pakes and Pollard (1989) are applicable when θ_0 is on a boundary. If the objective function is smooth, the stochastic equicontinuity condition is easy to verify, and primitive sufficient conditions for it are given below.

For the reader's convenience, we adopt the same notation for GMM/MD estimation as that of Pakes and Pollard (1989) (but with the sample size given by T rather than n). The GMM objective function is

$$\ell_T(\theta) := -T||A_T(\theta)G_T(\theta)||^2/2, \tag{9.1}$$

where $G_T(\theta) : \Theta \to R^k$ is a vector of sample moment conditions, such as a sample average, $A_T(\theta) : \Theta \to R^{k\times k}$ is a random weight matrix, and $||\cdot||$ denotes the Euclidean norm. (The division by two is strictly for convenience. It eliminates some constants in the formulae below.) The random variables $G_T(\theta)$ and $A_T(\theta)$ are normalized such that each is $O_p(1)$, but not $o_p(1)$ (except $G_T(\theta_0)$, which is $O_p(T^{-1/2})$). The MD objective function is exactly the same except that $G_T(\theta)$ is not a vector of moment conditions, but rather, the difference between an unrestricted estimator $\widehat{\xi}_T$ of a parameter ξ_0 and a vector of restrictions $h(\theta)$ on ξ_0. That is, $G_T(\theta) = \widehat{\xi}_T - h(\theta)$, where $\xi_0 = h(\theta_0)$.

In place of Assumptions 1–3, we use the following assumptions:

Assumption GMM1. (a) *For some non-random function* $G(\theta)$, $G_T(\theta) \xrightarrow{p} G(\theta)$ $\forall\theta \in \Theta \cap S(\theta_0, \varepsilon)$ *for some* $\varepsilon > 0$.
(b) $G(\theta) = G(\theta_0) + \Gamma(\theta - \theta_0) + o(||\theta - \theta_0||)$ *as* $||\theta - \theta_0|| \to 0$ *for* $\theta \in \Theta \cap S(\theta_0, \varepsilon)$ *for some* $\varepsilon > 0$, *where* Γ *is a non-random* $k \times s$ *matrix.*
(c) $G(\theta_0) = \mathbf{0}$.
(d) *For all* $\gamma_T \to 0$,

$\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T} T^{1/2}||G_T(\theta)-G(\theta)-G_T(\theta_0)||/(1+T^{1/2}||\theta-\theta_0||) = o_p(1)$ *under* θ_0.
(e) *For some non-random nonsingular matrix* A *and for all* $\gamma_T \to 0$,
$\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T} ||A_T(\theta) - A|| = o_p(1)$.

Assumption GMM2. $T^{1/2}G_T(\theta_0) = O_p(1)$.

Assumption GMM3. Γ *is full column rank* s $(\leq k)$.

Assumption GMM1(a) can be verified using a pointwise LLN in the GMM context or by showing that $\widehat{\xi}_T \xrightarrow{p} \xi_0$ in the MD context. (Assumption GMM1(a) is employed because it serves to define the limit function $G(\theta)$. It is not actually used in any of the proofs. Any function $G(\theta)$ that satisfies Assumption GMM1(d) could be used to define $G(\theta)$. There is little to be gained by this, however, because it is hard to imagine a case for which Assumption GMM1(d) holds for a function $G(\cdot)$ that does not satisfy Assumption GMM1(a).)

Assumption GMM1(b) holds if $G(\theta)$ is differentiable at θ_0. In this case, $\Gamma = \frac{\partial}{\partial\theta'}G(\theta_0)$. This requires that $G(\theta)$ is defined on a neighborhood of θ_0. In Section 9.2 below, we provide sufficient conditions for Assumptions GMM1(b) when $G(\theta)$ is not defined on a neighborhood of θ_0 using left/right partial derivatives.

Assumption GMM1(c) holds if the moment conditions are correctly specified in the GMM context and if the parameter ξ_0 satisfies the restrictions $\xi_0 = h(\theta_0)$ in the MD context.

Assumption GMM1(d) is a stochastic equicontinuity condition. It can be verified using the empirical process results of Pakes and Pollard (1989) or any of the empirical process results given or referenced in Andrews (1994b). In such cases, the condition actually is verified with the denominator "$1 + T^{1/2}||\theta - \theta_0||$" replaced by "1."

Alternatively, if $G_T(\theta)$ is smooth in θ, then Assumption GMM1(d) can be verified easily with the denominator "$1+T^{1/2}||\theta-\theta_0||$" replaced by "$T^{1/2}||\theta-\theta_0||$." To see this, suppose $G_T(\theta)$ is differentiable at θ_0 with derivative matrix $\frac{\partial}{\partial\theta'}G_T(\theta)$ that satisfies: For all $\gamma_T \to 0$,

$$\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T} \left|\left|\frac{\partial}{\partial\theta'}G_T(\theta) - \Gamma\right|\right| = o_p(1). \tag{9.2}$$

(This condition can be verified using a uniform LLN and continuity of the probability limit of $\frac{\partial}{\partial\theta'}G_T(\theta)$ at θ_0.) Then, applying the mean-value theorem element by element, we have

$$G_T(\theta) = G_T(\theta_0) + \frac{\partial}{\partial\theta'}G_T(\theta^\dagger)(\theta-\theta_0) = G_T(\theta_0) + \Gamma(\theta-\theta_0) + o_p(||\theta-\theta_0||) \tag{9.3}$$

uniformly over $\{\theta \in \Theta : ||\theta - \theta_0|| \leq \gamma_T\}$, where $\theta^\dagger$ lies between θ and θ_0 and may differ across rows. Combining this result with Assumptions GMM1(b) and GMM1(c) gives

$$||G_T(\theta) - G_T(\theta_0) - G(\theta)|| = o_p(||\theta - \theta_0||) \tag{9.4}$$

uniformly over $\{\theta \in \Theta : ||\theta - \theta_0|| \leq \gamma_T\}$. This immediately implies Assumption GMM1(d) using the "$T^{1/2}||\theta - \theta_0||$" part of the denominator in Assumption GMM1(d).

The above verification of Assumption GMM1(d) using smoothness relies on $G_T(\theta)$ being defined on a neighborhood of θ_0 (in order to define the derivative of $G_T(\theta)$ at θ_0). Below we show that Assumption GMM1(d) can be verified using smoothness even when $G_T(\theta)$ is not defined on a neighborhood of θ_0 by using left/right partial differentiability of $G_T(\theta)$.

Assumption GMM1(e) requires that the weight matrix $A_T(\theta)$ is well behaved. It can be verified using a uniform LLN and continuity of the probability limit of $A_T(\theta)$ at θ_0.

The stochastic equicontinuity condition of Assumption GMM1(d) differs somewhat from that given in Pakes and Pollard (1989). The following result, however, shows that it is equivalent to that of Pakes and Pollard (1989) given the other assumptions. We state Assumption GMM1(d) as is, because it is in the most convenient form for verification.

Lemma 8. *Under Assumptions* GMM1–GMM3 *except* GMM1(d)*, Assumption* GMM1(d) *is equivalent to each of the following two conditions: For all* $\gamma_T \to 0$*,*
(a) $\sup_{\theta \in \Theta : ||\theta - \theta_0|| \leq \gamma_T} ||G_T(\theta) - G(\theta) - G_T(\theta_0)||/(T^{-1/2} + ||G(\theta)||) = o_p(1)$ *and*
(b) $\sup_{\theta \in \Theta : ||\theta - \theta_0|| \leq \gamma_T} ||G_T(\theta) - G(\theta) - G_T(\theta_0)||/(T^{-1/2} + ||G_n(\theta)|| + ||G(\theta)||) = o_p(1)$.

Comment. Condition (b) of the Lemma is the same as that of Pakes and Pollard (1989, condition (iii) of Theorem 3.3).

The following result shows that under Assumptions GMM1–GMM3 the objective function $\ell_T(\theta)$ has a quadratic approximation given by

$$\begin{aligned} \ell_T(\theta) := & -T||A_T(\theta_0)G_T(\theta_0)||^2/2 - TG_T(\theta_0)'A'A\Gamma(\theta - \theta_0) \\ & + \tfrac{1}{2}(\theta - \theta_0)'(-T\Gamma'A'A\Gamma)(\theta - \theta_0) + R_T(\theta), \end{aligned} \tag{9.5}$$

where $R_T(\theta)$ satisfies Assumption 1. That is, (3.1) and (3.3) hold with

$$\begin{aligned} & D\ell_T(\theta_0) := -T\Gamma'A'AG_T(\theta_0),\ D^2\ell_T(\theta_0) := -T\Gamma'A'A\Gamma,\ \mathcal{J}_T := \mathcal{J} := \Gamma'A'A\Gamma, \text{ and} \\ & Z_T := (\Gamma'A'A\Gamma)^{-1}\Gamma'A'AT^{1/2}G_T(\theta_0). \end{aligned} \tag{9.6}$$

Theorem 7. (a) *Assumptions* GMM1 *and* GMM2 *imply that Assumption* 1 *holds with* $D\ell_T(\theta_0)$ *and* $D^2\ell_T(\theta_0)$ *as in* (9.6)*.*
(b) *Assumption* GMM2 *implies Assumption* 2*.*
(c) *Assumption* GMM3 *implies Assumption* 3*.*

Assumption 4 (consistency of $\widehat{\theta}$) can be established in the GMM/MD case using Assumption 4* as in Section 3. Alternatively, it an be established using Theorem 3.1 and Lemma 3.4 of Pakes and Pollard (1989), which is applicable even if θ_0 is on the boundary of Θ.

9.2. Sufficient Conditions for Assumption GMM1

Next, we provide simple smoothness conditions that are sufficient for Assumption GMM1 and that apply even when $G(\theta)$ and $G_T(\theta)$ are not defined on a neighborhood of θ_0.

Assumption GMM1*. (a) *Assumptions* GMM1(a), GMM1(c), *and* GMM1(e) *hold.*
(b) *The domain of $G(\theta)$ includes a set Θ^+ that satisfies conditions* (i) *and* (ii) *of Assumption* 1*(a). *Each element of the k-vector-valued function $G(\theta)$ has continuous l/r partial derivatives of order one on Θ^+.*
(c) *Each element of the k-vector-valued function $G_T(\theta)$ has continuous l/r partial derivatives of order one on Θ^+ $\forall T \geq 1$ with probability one.*
(d) *For all sequences of scalar constants $\{\gamma_T : T \geq 1\}$ for which $\gamma_T \to 0$,*

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

where $\frac{\partial}{\partial \theta'} G_T(\theta)$ denotes the $k \times s$ matrix of l/r partial derivatives of $G_T(\theta)$.
(e) $\frac{\partial}{\partial \theta'} G_T(\theta_0) \xrightarrow{p} \Gamma := \frac{\partial}{\partial \theta'} G(\theta_0)$, *where $\frac{\partial}{\partial \theta'} G(\theta_0)$ denotes the $k \times s$ matrix of l/r partial derivatives of $G(\theta)$ at θ_0.*

Lemma 9. *Assumption* GMM1* *implies Assumption* GMM1.

9.3. Multinomial Response Model

9.3.1. Method of Simulated Moments Estimator

For the Multinomial Response Model Example, we consider a method of simulated moments (MSM) estimator. This is a GMM estimator. The moment conditions are defined as follows. Let $\pi(Z_t, \theta)$ denote the conditional expectation of $D(Z_t h(\eta_t, \theta))$ given Z_t. Let $W(Z_t, \theta)$ denote an $s \times m$ matrix of instruments. See McFadden (1989) regarding the choice of instruments. The moment conditions, then, are

$$\frac{1}{T} \sum_{t=1}^{T} W(Z_t, \theta)(d_t - \pi(Z_t, \theta)). \tag{9.7}$$

These moment conditions have expectation zero when $\theta = \theta_0$, as desired. Following McFadden (1989) and Pakes and Pollard (1989), we take the number of moment conditions to equal the dimension s of θ. In this case, the choice of weight matrix is immaterial, so we take $A_T(\theta) = I_s$.

The conditional probability vector $\pi(Z_t, \theta)$ is computationally intractable in many cases because it is a vector of high dimensional integrals. In consequence, we replace it in the moment conditions by a simulated probability

$$\widehat{\pi}_S(Z_t, \theta) := \frac{1}{S} \sum_{j=1}^{S} D(Z_t h(\eta_{tj}, \theta)), \tag{9.8}$$

where $\eta_{t1},...,\eta_{tS}$ are simulated random variables each with the same distribution as η_t and $\xi_t := (Z_t, \eta_t, \eta_{t1}, ..., \eta_{tS})$ is iid across $t = 1, ..., T$. (With crude frequency simulators, $\eta_{t1}, ..., \eta_{tS}$ are iid. With variance reduced simulators, they are not independent.) The same simulated random variables are used for all θ.

The simulated moment conditions upon which the GMM estimator is based are

$$G_T(\theta) := \tfrac{1}{T}\sum_{t=1}^{T} W(Z_t,\theta)(d_t - \widehat{\pi}_S(Z_t,\theta)). \tag{9.9}$$

The true parameter vector is

$$\theta_0 := (\theta_{10}', \theta_{20}', \theta_{30}')' = (\mathbf{0}', \theta_{20}', \theta_{30}')', \tag{9.10}$$

where $\theta_{10} \in R^p$, $\theta_{20} > 0$ (element by element), (θ_1, θ_2) are the parameters that must be non-negative, and θ_3 contains the remaining parameters. The parameter space Θ is

$$\begin{aligned} \Theta := \{\theta \in R^s : \theta = (\theta_1', \theta_2', \theta_3')',\ \theta_1 \geq 0,\ \theta_2 \geq 0, \\ \theta_3 = (\theta_{31}, ..., \theta_{3J})',\ c_{\ell j} \leq \theta_{3j} \leq c_{uj}\ \forall j = 1, ..., J\} \end{aligned} \tag{9.11}$$

for some constants $-\infty \leq c_{\ell j} < c_{uj} \leq \infty$ for $j = 1, ..., J$. We assume that the true subvector θ_{30} of θ_0 is not on a boundary. The parameter space could incorporate additional restrictions without affecting the results given below, provided none are binding at θ_0. Note that Θ is not necessarily a bounded subset of R^s.

9.3.2. Quadratic Approximation of the GMM Objective Function

The function $G(\theta)$ and the matrix Γ that appear in Assumption GMM1 are

$$\begin{aligned} &G(\theta) = EG_T(\theta) = EW(Z_t,\theta)(\pi(Z_t,\theta_0) - \pi(Z_t,\theta)) \text{ and} \\ &\Gamma = \frac{\partial}{\partial\theta'}G(\theta_0), \end{aligned} \tag{9.12}$$

where $\frac{\partial}{\partial\theta'}G(\theta_0)$ denotes the matrix of right partial derivatives of $G(\theta)$ at θ_0, see Section 7.

The components of the quadratic expansion of the GMM criterion function given in (3.1) and (3.3) are as follows:

$$\begin{aligned} &D\ell_T(\theta_0) := T\Gamma' G_T(\theta_0),\ D^2\ell_T(\theta_0) := -T\Gamma'\Gamma, \\ &\mathcal{J}_T := \mathcal{J} := \Gamma'\Gamma, \text{ and } Z_T := (\Gamma'\Gamma)^{-1}T^{1/2}G_T(\theta_0). \end{aligned} \tag{9.13}$$

We verify Assumptions GMM1–GMM3 and 4 for this example using the approach of Pakes and Pollard (1989, Sec. 4.2). To do so, we use the assumptions stated above plus the following:

(i) $\inf_{\theta\in\Theta: ||\theta-\theta_0||>\varepsilon} ||G(\theta)|| > 0\ \forall\varepsilon > 0.$

(ii) $G(\theta)$ has continuous right partial derivatives with respect to θ at θ_0.

(iii) Γ is nonsingular.

(iv) $E \sup_{\theta \in \Theta} ||W(Z_t, \theta)|| < \infty$.

(v) $E \sup_{\theta \in \Theta : ||\theta - \theta_0|| \leq \varepsilon} ||W(Z_t, \theta)||^2 < \infty$ for some $\varepsilon > 0$.

(vi) $\{B(\theta) : \theta \in \Theta\}$ is a VC class of sets, where $B(\theta) := \{(z, \eta) \in R^b \times R^r : z' h(\eta, \theta) \geq 0\}$.

(vii) $\mathcal{F}_W := \{W(\cdot, \theta) : \theta \in \Theta\}$ is a Euclidean class of functions with envelope F that satisfies $EF^2(Z_t) < \infty$. (9.14)

VC and Euclidean classes are defined in Pakes and Pollard (1989, Sec. 2).

A sufficient condition for (i) is that Θ is bounded, $G(\theta)$ is continuous on Θ, and $G(\theta)$ has a unique zero at θ_0. A sufficient condition for (vi) is that $h(\eta, \theta)$ is of the form

$$h(\eta, \theta) = \beta_1(\theta) + \beta_2(\theta)\eta + \eta' \beta_3(\theta)\eta \tag{9.15}$$

for some functions $\beta_j(\theta)$, $j = 1, 2, 3$. This holds for the probit models considered by McFadden (1989) and those discussed above. Sufficient conditions for (vii) are that Θ is bounded, condition (v) holds, and $W(\cdot, \theta)$ satisfies the Lipschitz condition

$$||W(Z, \theta^*) - W(Z, \theta)|| \leq \phi(Z)||\theta^* - \theta|| \quad \forall Z \in R^{k \times m},\ \forall \theta^*, \theta \in \Theta \text{ and } E\phi^2(Z_t) < \infty. \tag{9.16}$$

Sufficiency of the above conditions is shown in Pakes and Pollard (1989, Sec. 4.2).

We now verify Assumptions GMM1–GMM3 and 4. Assumption GMM1(a) holds by a pointwise LLN for iid random variables with finite mean using (iv). GMM1(b) holds by the one-term Taylor expansion of Theorem 6 in Section 7 below. GMM1(c) holds because $Ed_t = E\pi(Z_t, \theta_0) = E\widehat{\pi}_S(Z_t, \theta_0)$ using the assumption of identical distributions of $\eta_t, \eta_{t1}, ..., \eta_{tS}$. GMM1(d) holds via empirical process results by the argument of Pakes and Pollard (1989, Sec. 4.2) using (v)–(vii). GMM1(e) holds because $A_T(\theta) = I_s$. GMM2 holds by the CLT for iid square integrable random variables using (v). GMM2 holds by (iii). Assumption 4 holds by verifying Assumption 4*. Assumption 4*(a) holds by an empirical process uniform LLN by the argument of Pakes and Pollard (1989, Sec. 4.2) using (iv), (vi), and (vii). Assumption 4*(b) holds by (i).

Given that Assumptions GMM1–GMM3 and 4 hold, we obtain $T^{1/2}(\widehat{\theta} - \theta_0) = O_p(1)$ by Theorems 1 and 7, where $\widehat{\theta}$ is the MSM estimator.

9.3.3. Parameter Space

Assumptions 5* and 6 hold in this example with $\Lambda := (R^+)^p \times R^{s-p}$.

9.3.4. Asymptotic Distribution of the MSM Estimator

In this example, $\mathcal{J} = \Gamma'\Gamma$ is non-random and nonsingular. Assumption 7 holds by the CLT for iid square integrable random variables with

(9.17) $G \sim N(\mathbf{0}, \mathcal{I})$ and $\mathcal{I} := EW(Z_t, \theta_0)(d_t - \widehat{\pi}_S(Z_t, \theta_0))(d_t - \widehat{\pi}_S(Z_t, \theta_0))'W(Z_t, \theta_0)'$.

If $\eta_t, \eta_{t1}, ..., \eta_{tS}$ are independent conditional on Z_t a.s., then $\mathcal{I}$ simplifies to

$$\mathcal{I} = \left(1 + \tfrac{1}{S}\right) EW(Z_t, \theta_0)(\text{Diag}(\pi(Z_t, \theta_0)) - \pi(Z_t, \theta_0)\pi(Z_t, \theta_0)')W(Z_t, \theta_0)'. \quad (9.18)$$

By Theorem 3, $T^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} \widehat{\lambda}$, where $\widehat{\lambda}$ satisfies (5.1) with $(G, \mathcal{J})$ defined in (9.17) and $\Lambda = (R^+)^p \times R^{s-p}$.

9.3.5. Asymptotic Distribution of Subvectors of the MSM Estimator

In this example, the matrix $\mathcal{J}$ $(= \Gamma'\Gamma)$ usually is not block diagonal. In consequence, we partition θ as in Section 6.1 with $r = 0$, where r is the dimension of ψ. Thus, no parameter ψ appears. In this case, we write

$$\theta = (\beta', \delta')', \ \beta = \theta_1, \text{ and } \delta = (\theta_2', \theta_3')'. \quad (9.19)$$

The set Λ is a product set $\Lambda_\beta \times \Lambda_\delta$ with

$$\Lambda_\beta := (R^+)^p \text{ and } \Lambda_\delta := R^{s-p}. \quad (9.20)$$

Assumption 7^{2*} holds with $G_* = G$, $\mathcal{I}_* = \mathcal{I}$, and $\mathcal{J}_* = \mathcal{J}$ using the fact that $r = 0$.

Recall that the parameter θ_1 $(= \beta)$ corresponds to the random coefficient variances that are zero in the random coefficient probit model, the proportion of the error variance due to the random effect in the binary probit panel data model, or the measurement error variances that are zero in the measurement error probit model. By Corollary 1(b), the MSM estimator of this parameter has asymptotic distribution given by

$$T^{1/2}(\widehat{\theta}_1 - \theta_{10}) \xrightarrow{d} \widehat{\lambda}_\beta, \quad (9.21)$$

where $\widehat{\lambda}_\beta$ is as in (6.23) when $p = 1$, as in (6.8) when $p = 2$, and as in Theorem 5 for $p > 2$.

Also, by Corollary 1(b), the asymptotic distribution of the remaining parameters is given by

$$\begin{aligned} & T^{1/2}\left(\left(\widehat{\theta}_2', \widehat{\theta}_3'\right)' - (\theta_{20}', \theta_{30}')'\right) \xrightarrow{d} \widehat{\lambda}_\delta, \text{ where} \\ & \widehat{\lambda}_\delta = \mathcal{J}_\delta^{-1} G_\delta - \mathcal{J}_\delta^{-1} \mathcal{J}_{\delta\beta} \widehat{\lambda}_\beta, \ G := \begin{pmatrix} G_\beta \\ G_\delta \end{pmatrix} \sim N(\mathbf{0}, \mathcal{I}), \\ & \mathcal{J}_\delta = \left(\frac{\partial}{\partial \delta'} G(\theta_0)\right)' \frac{\partial}{\partial \delta'} G(\theta_0), \ \mathcal{J}_{\delta\beta} = \left(\frac{\partial}{\partial \delta'} G(\theta_0)\right)' \frac{\partial}{\partial \beta'} G(\theta_0). \end{aligned} \quad (9.22)$$

and $\frac{\partial}{\partial \beta'} G(\theta_0)$ denotes the matrix of right partial derivatives of $G(\theta)$ with respect to β $(= \theta_1)$ at θ_0.

9.4. Proofs

Proof of Lemma 8. Condition (a) and Assumption GMM1(d) are easily shown to be equivalent given the assumption that $G(\theta) = \Gamma(\theta - \theta_0) + o(||\theta - \theta_0||)$.

Condition (a) obviously implies condition (b). To obtain the converse, we assume condition (b) holds and write: Uniformly over $\{\theta \in \Theta : ||\theta - \theta_0|| \leq \gamma_T\}$,

$$\begin{aligned} &||G_T(\theta) - G(\theta) - G_T(\theta_0)|| = o_p(T^{-1/2} + ||G_T(\theta)|| + ||G(\theta)||) \\ &\leq o_p(T^{-1/2} + ||G_T(\theta) - G(\theta) - G_T(\theta_0)|| + ||G_T(\theta_0)|| + 2||G(\theta)||), \end{aligned} \tag{9.23}$$

where the first equality uses condition (b) and the second uses Minkowski's inequality. Rearranging this equation and using the assumption that $||G_T(\theta_0)|| = O_p(T^{-1/2})$ yields

$$||G_T(\theta) - G(\theta) - G_T(\theta_0)|| = o_p(T^{-1/2} + ||G(\theta)||) \tag{9.24}$$

uniformly over $\{\theta \in \Theta : ||\theta - \theta_0|| \leq \gamma_T\}$. Hence, condition (a) of the Lemma holds. □

Proof of Theorem 7. Parts (b) and (c) of the Theorem are obvious. To prove part (a), we first consider the case where $A_T(\theta) = I_k$. We define the remainder term $R_T(\theta)$ and a close approximation to it, $R_T^*(\theta)$, that is obtained by replacing $\Gamma(\theta - \theta_0)$ by $G(\theta)$:

$$\begin{aligned} R_T(\theta) &:= -TG_T(\theta)'G_T(\theta)/2 + TG_T(\theta_0)'G_T(\theta_0)/2 + TG_T(\theta_0)'\Gamma(\theta - \theta_0) \\ &\quad + (\theta - \theta_0)'T\Gamma'\Gamma(\theta - \theta_0)/2, \\ R_T^*(\theta) &:= -TG_T(\theta)'G_T(\theta)/2 + T(G_T(\theta_0) - G(\theta))'(G_T(\theta_0) - G(\theta))/2, \text{ and} \\ \sup_{\theta \in \Theta: ||\theta - \theta_0|| \leq \gamma_T} & |R_T^*(\theta) - R_T(\theta)|/(1 + T^{1/2}||\theta - \theta_0||)^2 \\ &= \tfrac{1}{2} \sup_{\theta \in \Theta: ||\theta - \theta_0|| \leq \gamma_T} T|2G_T(\theta_0)'(G(\theta) - \Gamma(\theta - \theta_0)) + G(\theta)'G(\theta) \\ &\quad - (\theta - \theta_0)'\Gamma'\Gamma(\theta - \theta_0)|/(1 + T^{1/2}||\theta - \theta_0||)^2 \\ &= o_p(1), \end{aligned} \tag{9.25}$$

where the second equality in the last equation holds because $T^{1/2}(G(\theta) - \Gamma(\theta - \theta_0)) = o(T^{1/2}||\theta - \theta_0||)$ and $T^{1/2}G_T(\theta_0) = O_p(1)$. Thus, it suffices to show that Assumption 1 holds with $R_T(\theta)$ replaced by $R_T^*(\theta)$.

Define

$$\eta_T = \sup_{\theta \in \Theta: ||\theta - \theta_0|| \leq \gamma_T} \frac{T^{1/2}||G_T(\theta) - G(\theta) - G_T(\theta_0)||}{1 + T^{1/2}||\theta - \theta_0||}. \tag{9.26}$$

By Assumption GMM1(d), $\eta_T = o_p(1)$.

Let a, b, and c be k-vectors for which $a = b + c$. By the Cauchy–Schwarz inequality

$$|a'a - b'b| = |c'c + 2b'c| \leq c'c + 2||b|| \cdot ||c||. \tag{9.27}$$

Take $a = G_T(\theta)$ and $b = G(\theta) + G_T(\theta_0)$ in the equation above, multiply by $2T/(1+T^{1/2}||\theta-\theta_0||)^2$, and take the supremum over θ to obtain

$$\begin{aligned}
&\sup_{\theta\in\Theta:||\theta-\theta_0||\le\gamma_T} \frac{2|R_T^*(\theta)|}{(1+T^{1/2}||\theta-\theta_0||)^2} \\
&\le \eta_T^2 + 2 \sup_{\theta\in\Theta:||\theta-\theta_0||\le\gamma_T} \frac{||T^{1/2}G(\theta)+T^{1/2}G_T(\theta_0)||}{(1+T^{1/2}||\theta-\theta_0||)}\eta_T \\
&= \eta_T^2 + O_p(1)\eta_T \\
&= o_p(1), \qquad (9.28)
\end{aligned}$$

which establishes part (a) for the case where $A_T(\theta) = I_k$.

Next, we establish part (a) for the case where $A_T(\theta)$ is as in Assumption GMM1(e). The idea is to use the same proof as above but with $G_T(\theta)$, $G(\theta)$, and Γ replaced by $A_T(\theta)G_T(\theta)$, $AG(\theta)$, and $A\Gamma$ respectively. This method works provided Assumptions GMM1(b), GMM1(c), and GMM1(d), which are used in the proof above, hold with the same changes. Assumptions GMM1(b) and GMM1(c) obviously do. By Lemma 3.5 of Pakes and Pollard (1989), Assumption GMM1(e) and condition (b) of Lemma 8 imply that condition (b) of Lemma 8 holds with $G_T(\theta)$ and $G(\theta)$ replaced by $A_T(\theta)G_T(\theta)$ and $AG(\theta)$ respectively. In addition, Lemma 8 holds with these changes made to its conditions (a) and (b) and to Assumption GMM1(d) by the proof given for the Lemma. The last two results imply that Assumption GMM1(d) holds with the changes listed above, as desired. □

Proof of Lemma 9. Assumption GMM1*(b) implies Assumption GMM1(b) by Theorem 6. To establish Assumption GMM1*(d), we write:

$$\begin{aligned}
G_T(\theta) - G_T(\theta_0) - G(\theta) &= \frac{\partial}{\partial\theta'}G_T(\theta^\dagger)(\theta-\theta_0) - G(\theta) \\
&= \left(\frac{\partial}{\partial\theta'}G_T(\theta_0) + o_p(1)\right)(\theta-\theta_0) - \Gamma(\theta-\theta_0) + o(||\theta-\theta_0||) \\
&= o_p(||\theta-\theta_0||) \qquad (9.29)
\end{aligned}$$

uniformly over $\theta \in \Theta : ||\theta-\theta_0|| \le \gamma_T$, where the first equality holds for some $\theta^\dagger$ between θ and θ_0 by Theorem 6 ($\theta^\dagger$ may differ across rows of $\frac{\partial}{\partial\theta'}G_T(\theta^\dagger)$), the second equality holds by Assumptions GMM1*(d) and GMM1(b), and the last equality holds by Assumption GMM1*(e). Multiplying (9.29) by $T^{1/2}/||T^{1/2}(\theta-\theta_0)||$ and taking the supremum over $\theta \in \Theta : ||\theta-\theta_0|| \le \gamma_T$ establishes Assumption GMM1(d). □

10. GARCH(1,q^*) Example

10.1. Specification of the Model

First we describe the GARCH$(1,q^*)$ model. For notational convenience, we order the GARCH–MA parameters such that the first p (≥ 0) GARCH–MA parameters, $\theta_{10} \in R^p$, are those whose true values are zeros, rather than those that correspond to the first p lags of the conditional variance, and the last b (≥ 1) GARCH–MA parameters, $\theta_{20} \in R^b$, are those whose true values are non-zeroes.

The true process generating the data is

$$
\begin{aligned}
&Y_t := \theta_{50} + X_t'\theta_{60} + \varepsilon_t, \\
&\varepsilon_t := h_{0t}^{1/2} z_t, \ \{z_t : t = ..., 0, 1, ...\} \text{ are stationary and ergodic,} \\
&E(z_t|\mathcal{F}_{t-1}) = 0 \text{ a.s.}, \ E(z_t^2|\mathcal{F}_{t-1}) = 1 \text{ a.s.}, \ \mathcal{F}_t = \sigma(z_t, z_{t-1}, ...), \\
&h_{0t} := \theta_{40}(1 - \theta_{30}) + \varepsilon_{1t}^{2\prime}\theta_{10} + \varepsilon_{2t}^{2\prime}\theta_{20} + \theta_{30} h_{0t-1}, \\
&\varepsilon_{1t}^2 := (\varepsilon_{t-j_1}^2, ..., \varepsilon_{t-j_p}^2)' \text{ for } 2 \leq j_1 < j_2 < \cdots < j_p, \text{ and} \\
&\varepsilon_{2t}^2 := (\varepsilon_{t-k_1}^2, ..., \varepsilon_{t-k_b}^2)' \text{ for } 1 = k_1 < k_2 < \cdots < k_b,
\end{aligned} \tag{10.1}
$$

for $t = ..., 0, 1, ...,$ where $j_\ell \neq k_m$ for any $\ell = 1, ..., p$ and $m = 1, ..., b$; $\theta_{10}, \ \varepsilon_{1t}^2 \in R^p$; $\theta_{20}, \ \varepsilon_{2t}^2 \in R^b$; $q^* = \max\{j_p, k_b\}$; $\theta_{30}, \ \theta_{40}, \ \theta_{50} \in R$; and $\theta_{60}, \ X_t \in R^r$. The regressors $\{X_t : t = ..., 0, 1, ...\}$ are stationary and ergodic and independent of the innovations $\{z_t : t = ..., 0, 1, ...\}$. The GARCH–MA coefficients on ε_{1t}^2 are assumed to be zero. That is, $\theta_{10} = \mathbf{0}$. The GARCH–MA parameters on ε_{2t}^2 are assumed to be positive. That is, $\theta_{20} > 0$ (element by element). The GARCH–AR parameter θ_{30} is assumed to lie in (0,1). The intercept of the conditional variance equation θ_{40} is assumed to be positive. The only parameters on a boundary are those in θ_{10}.

The random variables $\{(Y_t, X_t) : 1 \leq t \leq T\}$ are observed. The true parameter vector is $\theta_0 := (\theta_{10}', \theta_{20}', \theta_{30}, \theta_{40}, \theta_{50}, \theta_{60}')' = (\mathbf{0}', \theta_{20}', \theta_{30}, \theta_{40}, \theta_{50}, \theta_{60}')' \in R^s$.

An equivalent expression for the conditional variance h_{0t} is $h_{0t} = \theta_{40} + \sum_{k=0}^{\infty} \theta_{30}^k \times(\varepsilon_{1t-k}^{2\prime}\theta_{10} + \varepsilon_{2t-k}^{2\prime}\theta_{20})$ for $t = ..., 0, 1, ...$ Note that by repeated substitution the true conditional variance h_{0t} is a weighted sum of the infinite sequence of past squared innovations $\{z_u^2 : u = t-1, t-2, ...\}$; see Bollerslev (1986, equations (A.2) and (A.3)) for details. The weighted sum is well-defined (though possibly infinite) for any parameter θ_0, because the weights are all non-negative.

The model used to generate a quasi-likelihood function is

$$
\begin{aligned}
&Y_t = \theta_5 + X_t'\theta_6 + e_t(\theta) \text{ for } t = 1, ..., T, \\
&e_t(\theta) = h_t^*(\theta)^{1/2} z_t \text{ for } t = 1, ..., T, \ \{z_t : t = 1, ..., T\} \text{ are iid } N(0,1), \\
&h_t^*(\theta) := \theta_4(1 - \theta_3) + e_{1t}^2(\theta)'\theta_1 + e_{2t}^2(\theta)'\theta_2 + \theta_3 h_{t-1}^*(\theta) \text{ for } t = q^* + 1, ..., T, \\
&e_{1t}^2(\theta) := (e_{t-j_1}^2(\theta), ..., e_{t-j_p}^2(\theta))', \ e_{2t}^2(\theta) := (e_{t-k_1}^2(\theta), ..., e_{t-k_b}^2(\theta))', \text{ and} \\
&\theta := (\theta_1', \theta_2', \theta_3, \theta_4, \theta_5, \theta_6')'.
\end{aligned} \tag{10.2}
$$

The initial conditions $h_1^*(\theta), ..., h_{q^*}^*(\theta)$ are arbitrary non-negative functions of θ and/or $\{(Y_t, X_t) : t = 1, ..., q^*\}$.

10.2. Quadratic Approximation of the Quasi-log Likelihood Function

We consider the Gaussian QML estimator of θ. The Gaussian quasi-log likelihood function is

$$\ell_T(\theta) := -\tfrac{T}{2}\ln(2\pi) - \tfrac{1}{2}\sum_{t=1}^{T}\ln h_t^*(\theta) - \tfrac{1}{2}\sum_{t=1}^{T} e_t^2(\theta)/h_t^*(\theta). \tag{10.3}$$

The parameter space Θ is a compact subset of R^s that bounds the GARCH–AR parameter, θ_3, away from zero and one, the GARCH–MA parameter on the first lag, θ_{21}, away from zero, and the conditional variance intercept parameter, θ_4, away from zero:

$$\begin{aligned}\Theta := \{\theta \in R^s : \theta = (\theta_1', \theta_2', \theta_3, \theta_4, \theta_5, \theta_6')',\ &\theta_{j\ell} \le \theta_j \le \theta_{ju}\ \forall j = 1, ..., b,\\ &\text{where } \theta_{j\ell} \text{ and } \theta_{ju} \text{ are finite constants or constant vectors}\\ &\text{that satisfy } \theta_{1\ell} = \mathbf{0},\ \theta_{2\ell} := (\theta_{21\ell}, 0, ..., 0)',\ \theta_{21\ell} > 0,\ \theta_{3\ell} > 0,\\ &\theta_{3u} < 1, \text{ and } \theta_{4\ell} > 0\}.\end{aligned} \tag{10.4}$$

(The vector inequalities above are element by element inequalities.) We assume that $\theta_0 \in \Theta$ and that each subvector of θ_0 satisfies the above inequalities strictly except θ_{10}, which equals $\mathbf{0}$ and causes θ_0 to be on the boundary of Θ. (To obtain results when all of the GARCH–MA parameters are positive, we take $p = 0$ and the subvector $\theta_1 \in R^p$ disappears.)

Note that the parameter space need not restrict the GARCH parameters to be values that generate a stationary process. Such restrictions could be added if desired. They would have no effect on the asymptotic results given below provided the parameter space is compact and none of these restrictions is binding at the true parameter vector θ_0. The true parameter vector θ_0, however, will be such that the true process is stationary under the assumptions given below.

Define

$$\begin{aligned} h_t(\theta) &:= \theta_4 + \sum_{k=0}^{\infty}\theta_3^k(e_{1t-k}^2(\theta)'\theta_1 + e_{2t-k}^2(\theta)'\theta_2) \text{ and}\\ \ell_{tt}(\theta) &:= -\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}\ln(h_t(\theta)) - \tfrac{1}{2}e_t^2(\theta)/h_t(\theta).\end{aligned} \tag{10.5}$$

(The double subscript on $\ell_{tt}(\theta)$ is used to distinguish $\ell_{tt}(\theta)$ from $\ell_T(\theta)$ when $t = T$.) Note that $h_t(\theta)$ is the unobserved conditional variance given the parameter θ with the initial conditions $h_1^*(\theta), ..., h_{q^*}^*(\theta)$ replaced by an infinite weighted sum of lagged values of $e_t^2(\theta)$. Also, $\ell_{tt}(\theta)$ is the corresponding unobserved t-th contribution to the quasi-log likelihood. The asymptotic behavior of the actual quasi-log likelihood formed using $h_t^*(\theta)$ is shown to be equivalent to that based on $h_t(\theta)$ using the argument of (3.6)-(3.8).

We take the normalization matrix B_T to be of the form $T^{1/2}M$, where M is a nonsingular non-diagonal matrix that is designed to yield block diagonality of the

Hessian quasi-information matrix $\mathcal{J}_T$ between the regression slope parameters and the remaining parameters. This block diagonality is achieved without assuming symmetry of the true (perhaps non-Gaussian) distribution of z_t.

By definition,

$$M := \begin{bmatrix} I_{p+b+2} & \mathbf{0} & \mathbf{0}' \\ \mathbf{0}' & 1 & EX_t' \\ \mathbf{0} & \mathbf{0} & I_r \end{bmatrix} \text{ and } L := M^{-1'} = \begin{bmatrix} I_{p+b+2} & \mathbf{0} & \mathbf{0}' \\ \mathbf{0}' & 1 & \mathbf{0}' \\ \mathbf{0}' & -EX_t & I_r \end{bmatrix}. \tag{10.6}$$

When L pre-multiplies a vector that has $(1, X_t')'$ as its last $r+1$ elements, it yields a vector that is the same except that its last r elements are the mean zero random variables $X_t - EX_t$. This is the key feature that yields block diagonality of $\mathcal{J}_T$. Note that $M\theta = (\theta_1', \theta_2', \theta_3, \theta_4, \theta_5 + EX_t'\theta_6, \theta_6')'$. We determine the asymptotic distribution of $B_T(\widehat{\theta} - \theta_0) = T^{1/2}(M\widehat{\theta} - M\theta_0)$. This directly gives the asymptotic distribution of $T^{1/2}(\widehat{\theta}_j - \theta_{j0})$ for all $j \neq 5$. The asymptotic distribution of $T^{1/2}(\widehat{\theta}_5 - \theta_{50})$ is given by $T^{1/2}v'(M\widehat{\theta} - M\theta_0)$, where $v := (\mathbf{0}', 1, -EX_t')'$.

The components of the quadratic approximation of $\ell_T(\theta)$ at θ_0 are defined as follows:

$$\begin{aligned} & D\ell_T(\theta_0) := \sum_{t=1}^{T} \tfrac{\partial}{\partial \theta}\ell_{tt}(\theta_0) = \sum_{t=1}^{T} \left(\tfrac{1}{2}(z_t^2 - 1)\tfrac{\partial}{\partial \theta}h_t(\theta_0)/h_{0t} - z_t \tfrac{\partial}{\partial \theta}e_t(\theta_0)/h_{0t}^{1/2} \right), \\ & D^2\ell_T(\theta_0) := -T\mathcal{J},\ B_T := T^{1/2}M,\ \ B_T^{-1'} := T^{-1/2}L, \\ & Z_T := \mathcal{J}^{-1}T^{-1/2}LD\ell_T(\theta_0), \text{ and} \\ & \mathcal{J} := \mathcal{J}_T = -EL\tfrac{\partial^2}{\partial\theta\partial\theta'}\ell_{tt}(\theta_0)L' = \tfrac{1}{2}EL\tfrac{\partial}{\partial\theta}h_t(\theta_0)\tfrac{\partial}{\partial\theta'}h_t(\theta_0)L'/h_{0t}^2 \\ & \qquad\qquad + EL\tfrac{\partial}{\partial\theta}e_t(\theta_0)\tfrac{\partial}{\partial\theta'}e_t(\theta_0)L'/h_{0t}. \end{aligned} \tag{10.7}$$

Some calculations yield

$$\begin{aligned} & L\frac{\partial}{\partial\theta}e_t(\theta_0) = (\mathbf{0}', \mathbf{0}', 0, 0, -1, -X_t^{*\prime})' \text{ and} \\ & L\frac{\partial}{\partial\theta}h_t(\theta_0) = \sum_{k=0}^{\infty}\theta_{30}^k \begin{pmatrix} \varepsilon_{1t-k}^2 \\ \varepsilon_{2t-k}^2 \\ h_{0t-k} - \theta_{40} \\ 1 - \theta_{30} \\ -2\varepsilon_{2t-k}'\theta_{20} \\ -2X_{2t-k}^*(\varepsilon_{2t-k} \odot \theta_{20}) \end{pmatrix}, \text{ where} \\ & X_t^* = X_t - EX_t,\ \varepsilon_{2t} = (\varepsilon_{t-k_1}, ..., \varepsilon_{t-k_b})' \text{ and } X_{2t}^* = [X_{t-k_1}^* \cdots X_{t-k_b}^*]. \end{aligned} \tag{10.8}$$

Note that $\frac{\partial}{\partial\theta}e_t(\theta_0)$ and $\frac{\partial}{\partial\theta}h_t(\theta_0)$ equal $L\frac{\partial}{\partial\theta}e_t(\theta_0)$ and $L\frac{\partial}{\partial\theta}h_t(\theta_0)$, respectively, with X_t^* and X_{2t-k}^* replaced by X_t and X_{2t-k}.

To verify Assumptions 1–4, we use the following additional assumptions:

$$\begin{aligned} & \text{(a) } E(z_t^4|\mathcal{F}_{t-1}) \leq \mathcal{K} < \infty \text{ a.s. and } P(z_t^2 = 1) \neq 1, \\ & \text{(b) } E\begin{pmatrix} 1 \\ X_t \end{pmatrix}\begin{pmatrix} 1 \\ X_t \end{pmatrix}' \text{ is positive definite and } E||X_t||^{10} < \infty, \text{ and} \\ & \text{(c) } Eh_{0t}^{p^*} < \infty \text{ for some } 0 < p^* \leq 1 \text{ for some } t. \end{aligned} \tag{10.9}$$

Assumptions (a) and (b) are not very restrictive. Assumption (c) holds with $p^* = 1$ when the true process is not integrated, i.e., when $\theta_{30} + \sum_{j=1}^{b} \theta_{2j0} < 1$, where $\theta_{20} := (\theta_{210}, ..., \theta_{2b0})'$. This follows by the proof of Theorem 1 in Bollerslev (1986) extended to cover the case of square integrable stationary innovations $\{z_t : t = ..., 0, 1, ...\}$. Assumption (c) holds in an IGARCH(1,1) model under our assumptions on the innovations by Lemma 2(3) of Lee and Hansen (1994), who extend results of Nelson (1990). We conjecture that Assumption (c) holds for an IGARCH$(1, q^*)$ process for $q^* > 1$ (i.e., a process with $\theta_{30} + \sum_{j=1}^{b} \theta_{2j0} = 1$), but we do not yet have a proof.

Under the assumptions above, $\{h_{0t} : t = ..., 0, 1, ...\}$, $\{Y_t : t = ..., 0, 1, ...\}$, and $\{\frac{\partial}{\partial\theta}\ell_{tt}(\theta_0) : t = ..., 0, 1, ...\}$ are strictly stationary and ergodic and Assumption 1 holds. This is proved in Section 13. In addition, $\frac{\partial}{\partial\theta}\ell_{tt}(\theta_0)$ is shown to be square integrable in Section 13. Assumption 2 holds by the central limit theorem for square integrable, stationary and ergodic, martingale difference sequences applied to $\{\frac{\partial}{\partial\theta}\ell_{tt}(\theta_0) : t = ..., 0, 1, ...\}$; see Billingsley (1968, Thm. 23.1). (Note that $E(\frac{\partial}{\partial\theta}\ell_{tt}(\theta_0) |\mathcal{G}_{t-1}) = \mathbf{0}$ a.s., where $\mathcal{G}_t = \sigma(X_{t+1}, X_t, ..., z_t, z_{t-1}, ...)$ for $t = ..., 0, 1, ...$ Hence, $\{\frac{\partial}{\partial\theta}\ell_{tt}(\theta_0) : t = ..., 0, 1, ...\}$ is a martingale difference sequence.) Assumption 3 holds because $\mathcal{J}$ is symmetric, positive definite, and independent of T. Positive definiteness of $\mathcal{J}$ is proved in Section 13.

Assumption 4 is shown to hold in Section 13 when $\theta_{30} + \sum_{j=1}^{b} \theta_{2j0} < 1$. When the true process is IGARCH$(1, q^*)$, we only show that a "local" QML estimator satisfies Assumption 4, see Section 13. The difficulty in proving consistency of the QML estimator is that $E\ell_{tt}(\theta)$ may equal minus infinity when the GARCH–AR parameter θ_3 is small relative to θ_{30}. In this case, $T^{-1}\ell_T(\theta)$ does not converge uniformly to a function $\ell(\theta)$, which complicates a proof of consistency. Note that it is the behavior of $T^{-1}\ell_T(\theta)$ at points that have small likelihood values (and, hence, points that are not likely to equal the QML estimator) that complicate the standard method of proof. In consequence, we conjecture that the QML estimator is in fact consistent in the IGARCH$(1, q^*)$ case.

10.3. Parameter Space

Assumptions 5^{2*} and 6 hold in this example with

$$\Lambda := (R^+)^p \times R^{s-p}. \tag{10.10}$$

Assumption 5^{2*} holds because part (a) holds with $\Lambda^* = \Lambda$, part (b) holds with $\Upsilon_T = T^{1/2} I_s$ and M as in (10.6), and part (c) holds because $\Upsilon_T M \Lambda^* = \Lambda$.

10.4. Asymptotic Distribution of the QML Estimator

In this example, $\mathcal{J}_T$ does not depend on T and $\mathcal{J}$ $(= \mathcal{J}_T)$ is symmetric and positive definite. (The latter is proved in Section 13.) Thus, Assumption 7 holds provided $T^{-1/2} L D\ell_T(\theta_0) \xrightarrow{d} G$ for some random variable G. By the CLT for square integrable, stationary and ergodic, martingale difference sequences (see Billingsley

(1968, Thm. 23.1)), we have

$$(10.11)\quad T^{-1/2}LD\ell_T(\theta_0) \xrightarrow{d} G \sim N(\mathbf{0},\mathcal{I}), \text{ where } \mathcal{I} := EL\frac{\partial}{\partial\theta}\ell_{tt}(\theta_0)\frac{\partial}{\partial\theta'}\ell_{tt}(\theta_0)L'$$

for $\frac{\partial}{\partial\theta}\ell_{tt}(\theta_0)$ defined in (10.7). Square integrability of $\frac{\partial}{\partial\theta}\ell_{tt}(\theta_0)$ is established in Section 13.

By Theorem 3, $T^{1/2}M(\widehat{\theta}-\theta_0) \xrightarrow{d} \widehat{\lambda}$, where $\widehat{\lambda}$ satisfies (5.1) with $(G,\mathcal{J})$ defined in (10.7) and (10.11) and $\Lambda = (R^+)^p \times R^{s-p}$

10.5. Asymptotic Distributions of Subvectors of the QML Estimator

We partition θ as in Section 6.1 with

$$(10.12)\quad \theta_* := (\theta_1', \theta_2', \theta_3, \theta_4, \theta_5)',\ \psi := \theta_6,\ \beta = \theta_1, \text{ and } \delta := (\theta_2', \theta_3, \theta_4, \theta_5)'.$$

With this partitioning, we find that $\mathcal{J}$ is block diagonal between the regression slope parameters and the other parameters because

$$(10.13)\qquad L_{6\cdot}E\frac{\partial^2}{\partial\theta\partial\theta'}\ell_{tt}(\theta_0)L_{j\cdot}' = \mathbf{0}\ \ \forall j = 1,...,5,$$

where $L_{j\cdot}$ denotes the rows of L corresponding to θ_j for $j = 1,...,6$. This holds because $L_{6\cdot}\frac{\partial}{\partial\theta}e_t(\theta_0)$ and $L_{6\cdot}\frac{\partial}{\partial\theta}h_t(\theta_0)$ are linear in X_{t-i}^* for $i \geq 0$, $L_{j\cdot}\frac{\partial}{\partial\theta}e_t(\theta_0)$ and $L_{j\cdot}\frac{\partial}{\partial\theta}h_t(\theta_0)$ do not depend on $\{X_t^* : t = ...,0,1,...\}$, $EX_t^* = \mathbf{0}\ \forall t$, $\{X_t^* : t = ...,0,1,...\}$ and $\{z_t : t = ...,0,1,...\}$ are independent, and the expectations are well-defined. (The latter is established in Section 13.) Thus, Assumption 8 holds.

The set Λ is a product set $\Lambda_\beta \times \Lambda_\delta \times \Lambda_\psi$ with

$$(10.14)\qquad \Lambda_\beta := (R^+)^p,\ \Lambda_\delta := R^{b+3}, \text{ and } \Lambda_\psi := R^r.$$

Thus, Assumptions 9 and 10(a) also hold.

With this partitioning, from (10.7), (10.8), and (10.10), we have

$$\begin{aligned}
\mathcal{J}_* &= \tfrac{1}{2}E\tfrac{\partial}{\partial\theta_*}h_t(\theta_0)\tfrac{\partial}{\partial\theta_*'}h_t(\theta_0)/h_{0t}^2 + E(1/h_{0t})\left(\begin{smallmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0}' & 1\end{smallmatrix}\right),\\
\mathcal{J}_\psi &= \tfrac{1}{2}EL_{6\cdot}\tfrac{\partial}{\partial\theta}h_t(\theta_0)\tfrac{\partial}{\partial\theta'}h_t(\theta_0)L_{6\cdot}'/h_{0t}^2 + EX_t^*X_t^{*\prime}/h_{0t},\\
L_{6\cdot}\tfrac{\partial}{\partial\theta}h_t(\theta_0) &= -2\sum_{k=0}^{\infty}\theta_{30}^k X_{2t-k}^*(\varepsilon_{2t-k}\odot\theta_{20})\\
G &:= (G_*', G_\psi')' \sim N(\mathbf{0},\mathcal{I}),\ G_* \sim N(\mathbf{0},\mathcal{I}_*),\ G_\psi \sim N(\mathbf{0},\mathcal{I}_\psi),\\
\mathcal{I}_* &:= E\left(\tfrac{1}{2}(z_t^2-1)\tfrac{\partial}{\partial\theta_*}h_t(\theta_0)/h_{0t} + z_t\left(\begin{smallmatrix}\mathbf{0}\\ 1\end{smallmatrix}\right)/h_{0t}^{1/2}\right)\\
&\qquad \times\left(\tfrac{1}{2}(z_t^2-1)\tfrac{\partial}{\partial\theta_*}h_t(\theta_0)/h_{0t} + z_t\left(\begin{smallmatrix}\mathbf{0}\\ 1\end{smallmatrix}\right)/h_{0t}^{1/2}\right)', \text{ and}\\
\mathcal{I}_\psi &:= E\left(\tfrac{1}{2}(z_t^2-1)L_{6\cdot}\tfrac{\partial}{\partial\theta}h_t(\theta_0)/h_{0t} + z_tX_t^*/h_{0t}^{1/2}\right)\\
&\qquad \times\left(\tfrac{1}{2}(z_t^2-1)L_{6\cdot}\tfrac{\partial}{\partial\theta}h_t(\theta_0)/h_{0t} + z_tX_t^*/h_{0t}^{1/2}\right)'. \qquad (10.15)
\end{aligned}$$

Thus, in this example, Assumption 7^{2*} holds. If the innovation z_t has third moment equal to zero conditional on $\mathcal{F}_{t-1}$ a.s., then $\mathcal{I}_*$ and $\mathcal{I}_\psi$ simplify somewhat. In particular, $\mathcal{I}_*$ can be written as

$$\mathcal{I}_* = \tfrac{1}{4}E(z_t^2 - 1)^2 \tfrac{\partial}{\partial \theta_*} h_t(\theta_0) \tfrac{\partial}{\partial \theta_*'} h_t(\theta_0)/h_{0t}^2 + E(1/h_{0t}) \left(\begin{smallmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & 1 \end{smallmatrix}\right). \tag{10.16}$$

In consequence, $\mathcal{I}_* = \mathcal{J}_*$ and Assumption 7^{3*} holds with $c = 1$ if z_t has conditional third moment equal to zero and conditional fourth moment equal to three given $\mathcal{F}_{t-1}$ a.s. The latter is true if $\{z_t : t = ..., 0, 1, ...\}$ are iid $N(0, 1)$.

By Theorem 3, $T^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} \widehat{\lambda}$, where $\widehat{\lambda} = (\widehat{\lambda}_\beta', \widehat{\lambda}_\delta', \widehat{\lambda}_\psi')'$. By Theorem 4(c), Corollary 1(a), and the fact that $\Lambda_\psi = R^r$, we have

$$\begin{aligned} &\widehat{\lambda}_\psi = Z_\psi := \mathcal{J}_\psi^{-1} G_\psi \sim N(\mathbf{0}, \mathcal{J}_\psi^{-1}) \text{ and} \\ &T^{1/2}(\widehat{\theta}_6 - \theta_{60}) \xrightarrow{d} \widehat{\lambda}_\psi \sim N(\mathbf{0}, \mathcal{J}_\psi^{-1}). \end{aligned} \tag{10.17}$$

In consequence, the regression slope parameter estimator $\widehat{\theta}_6$ is asymptotically normal with covariance matrix $\mathcal{J}_\psi^{-1}$ whether or not some GARCH–MA parameters are zero.

By Theorem 4(a) and Corollary 1(a),

$$\begin{aligned} &T^{1/2}(\widehat{\theta}_1 - \theta_{10}) \xrightarrow{d} \widehat{\lambda}_\beta, \text{ where } \widehat{\lambda}_\beta \text{ minimizes} \\ &(\lambda_\beta - Z_\beta)' \left(H \mathcal{J}_*^{-1} H'\right)^{-1} (\lambda_\beta - Z_\beta) \text{ over } \lambda_\beta \in (R^+)^p \text{ and} \\ &Z_\beta := H \mathcal{J}_*^{-1} G_* \sim N(\mathbf{0}, H \mathcal{J}_*^{-1} \mathcal{I}_* \mathcal{J}_*^{-1} H'). \end{aligned} \tag{10.18}$$

For example, if $p = 1$ (i.e., the GARCH model is over parametrized by one GARCH–MA parameter), then Assumption 10 holds and $\widehat{\lambda}_\beta$ is as in (6.23). If $p > 1$, $\widehat{\lambda}_\beta$ is given in closed form by (6.8) or Theorem 5.

By Corollary 1(a),

$$\begin{aligned} &T^{1/2}((\widehat{\theta}_2', \widehat{\theta}_3, \widehat{\theta}_4, \widehat{\theta}_5 + EX_t'\widehat{\theta}_6)' - (\theta_{20}', \theta_{30}, \theta_{40}, \theta_{50} + EX_t'\theta_6)') \xrightarrow{d} \widehat{\lambda}_\delta, \text{ and} \\ &T^{1/2}(\widehat{\theta}_5 - \theta_{50}) \xrightarrow{d} v_1'\widehat{\lambda}_\delta - EX_t'\widehat{\lambda}_\beta, \text{ where } v_1 = (\mathbf{0}', 1)', \\ &\widehat{\lambda}_\delta = \mathcal{J}_\delta^{-1} G_\delta - \mathcal{J}_\delta^{-1} \mathcal{J}_{\delta\beta} \widehat{\lambda}_\beta, \\ &\mathcal{J}_\delta := \tfrac{1}{2}E \tfrac{\partial}{\partial \delta} h_t(\theta_0) \tfrac{\partial}{\partial \delta'} h_t(\theta_0)/h_{0t}^2 + E(1/h_{0t}) \left(\begin{smallmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & 1 \end{smallmatrix}\right) \in R^{(b+3)\times(b+3)}, \\ &\mathcal{J}_{\delta\beta} := \tfrac{1}{2}E \tfrac{\partial}{\partial \delta} h_t(\theta_0) \tfrac{\partial}{\partial \beta'} h_t(\theta_0)/h_{0t}^2, \text{ and } G_* := \left(\begin{smallmatrix} G_\beta \\ G_\delta \end{smallmatrix}\right) \sim N(\mathbf{0}, \mathcal{I}_*). \end{aligned} \tag{10.19}$$

11. Partially Linear Regression Example

11.1. Semiparametric Least Squares Objective Function

In this example, the estimator objective function is a semiparametric LS function:

$$\ell_T(\theta) := -\tfrac{1}{2} \sum_{t=1}^{T} (Y_t - \widehat{E}(Y_t \mid X_t, Z_t) - (X_t - \widehat{E}(X_t \mid Z_t))'\theta)^2, \tag{11.1}$$

where $\widehat{E}(Y_t \mid X_t, Z_t)$ and $\widehat{E}(X_t \mid Z_t)$ are nonparametric bias-reducing kernel estimators of $\eta_1(X_t, Z_t) := E(Y_t \mid X_t, Z_t)$ and $\eta_2(Z_t) := E(X_t \mid Z_t)$, respectively, as defined in Robinson (1988).

The parameter space is the same as in (3.20) for Example 2.

The quadratic approximation of (3.1) and (3.3) holds with

$$D\ell_T(\theta_0) := -\tfrac{1}{2}\sum_{t=1}^{T}(Y_t - \widehat{E}(Y_t|X_t, Z_t) - (X_t - \widehat{E}(X_t|Z_t))'\theta_0)(X_t - \widehat{E}(X_t|Z_t)),$$

$$D^2\ell_T(\theta_0) := -\sum_{t=1}^{T}(X_t - \widehat{E}(X_t|Z_t))(X_t - \widehat{E}(X_t|Z_t))',$$

$$B_T := T^{1/2}I_s, \mathcal{J}_T := -\tfrac{1}{T}D^2\ell_T(\theta_0), Z_T = \mathcal{J}_T^{-1}T^{-1/2}D\ell_T(\theta_0), \text{ and } R_T(\theta) = 0. \quad (11.2)$$

Assumption 1 holds because $R_T(\theta) = 0$.

Assumptions 2 and 3 hold by the Propositions given in the Appendix of Robinson (1988) under the following conditions:

$$\Phi := E(X_t - E(X_t|Z_t))(X_t - E(X_t|Z_t))' > 0, E\varepsilon_t^2 := \sigma^2 < \infty, E\,|X_t|^4 < \infty, \quad (11.3)$$

Z_t has a density $f(\cdot)$ with respect to Lebesgue measure, the functions $\mu(\cdot)$, $\eta_1(\cdot)$, and $\eta_2(\cdot)$ satisfy the smoothness and boundedness conditions of Robinson (1988, Thm. 1), the bandwidth and trimming parameters and the kernel used in the kernel estimators $\widehat{E}(Y_t \mid X_t, Z_t)$ and $\widehat{E}(X_t \mid Z_t)$ satisfy the conditions of Robinson (1988, Thm. 1).

Assumption 4 holds under the assumptions given above by the proof of consistency for Example 2 in Section 8.2 with X_t replaced by $X_t - \widehat{E}(X_t|Z_t)$. Equation (8.7) holds in the present case by Assumptions 2 and 3, which hold under the given conditions by the Propositions in the Appendix of Robinson (1988).

11.2. Parameter Space

Assumptions 5^{4*} and 6 hold in this example under the same conditions on $g_1(\theta)$ and $g_2(\theta)$ and with the same Λ matrix as in the Regression with Restricted Parameters Example 2.

11.3. Asymptotic Distribution of the Semiparametric Least Squares Estimator

The Propositions in the Appendix of Robinson (1988) show that Assumption 7 holds in this example under the assumptions given above with

$$G \sim N(\mathbf{0}, \mathcal{I}),\ \mathcal{I} = \sigma^2\Phi, \text{ and } \mathcal{J} = \Phi. \quad (11.4)$$

Thus, by Theorem 3, $T^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} \widehat{\lambda}$, where $\widehat{\lambda}$ satisfies (5.1) with $(G, \mathcal{J})$ defined in (11.4) and Λ defined in (4.16).

11.4. Asymptotic Distributions of Subvectors of the Semiparametric Least Squares Estimator

We partition θ and X_t in this example in the same way as for Example 2, but with X_t replaced by $X_t - E(X_t|Z_t)$. Then, the asymptotic distributions of $T^{1/2}(\widehat{\beta} - \beta_0)$, $T^{1/2}(\widehat{\delta} - \delta_0)$, and $T^{1/2}(\widehat{\psi} - \psi_0)$ are the same as given in Example 2 except that X_t is replaced by $X_t - E(X_t|Z_t)$ throughout. In addition, the asymptotic distributions simplify somewhat in the present example because ε_t is assumed to be independent of the regressors. In particular, we find that $\mathcal{I} = \sigma^2\Phi = \sigma^2\mathcal{J}$, $\mathcal{I}_* = \sigma^2\mathcal{J}_*$, $\mathcal{I}_\beta = \sigma^2\mathcal{J}_\beta$, $\mathcal{I}_\delta = \sigma^2\mathcal{J}_\delta$, and $\mathcal{I}_\psi = \sigma^2\mathcal{J}_\psi$. This implies that Assumption 7^{3*} holds and that $T^{1/2}(\widehat{\psi} - \psi_0) \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathcal{J}_\psi^{-1})$.

12. Regression with Restricted Parameters and Integrated Regressors Example

12.1. Specification of the Integrated Regressors

In this example, the estimator, objective function, parameter space, and quadratic approximation to the objective function are the same as in the Regression with Restricted Parameters Example 2; see(3.19)-(3.21).

In contrast to Example 2, however, the regressors are assumed here to be integrated of order one. We assume the errors and regressors satisfy the conditions of Park and Phillips (1988, Section 2). In particular,

$$(12.1)\quad X_t := \sum_{s=1}^{t} U_s,\ W_t := \begin{pmatrix} \varepsilon_t \\ U_t \end{pmatrix},\ S_T := \sum_{t=1}^{T} W_t, \text{ and } S_T(r) := T^{-1/2}S_{[Tr]},$$

where $[Tr]$ denotes the integer part of Tr. We assume that

$$(12.2)\quad \begin{aligned} &S_T(\cdot) \Rightarrow B(\cdot), \\ &T^{-1}\sum_{t=1}^{T} S_t W_t' \xrightarrow{d} \textstyle\int_0^1 B(r)dB(r)' + \Pi, \end{aligned}$$

and the convergence holds jointly, where "$\Rightarrow$" denotes weak convergence and $B(r)$ is a $(s+1)$-vector Brownian motion with positive definite covariance matrix Ω. By definition,

$$(12.3)\quad \Omega := \lim_{T\to\infty} T^{-1}ES_TS_T' \text{ and } \Pi := \lim_{T\to\infty} T^{-1}\sum_{t=2}^{T}\sum_{j=1}^{t} EW_jW_t' := \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix},$$

where $\Pi_{11} \in R$, $\Pi_{21} \in R^s$, and $\Pi_{22} \in R^{s\times s}$. Primitive conditions under which (12.2) holds are referenced in Park and Phillips (1988, Section 2). For example, mean zero asymptotically weakly dependent processes $\{W_t : t \geq 1\}$, such as linear processes and strong mixing processes, satisfy (12.2) under suitable moment and dependence assumptions.

The norming matrix B_T in this case equals TI_s.

Assumption 1 holds trivially in this example because $R_T(\theta) = 0$.

Assumption 2 holds by (12.2) because $B_T^{-1'} D\ell_T(\theta_0)$ equals the first column of $T^{-1}\sum_{t=1}^T S_t W_t'$ with the first element eliminated. Assumption 3 holds because (12.2) implies (see Park and Phillips (1988, Lemma 2.1(c))) that

$$(12.4) \qquad \mathcal{J}_T := T^{-2}\sum_{t=1}^T X_t X_t' \xrightarrow{d} \int_0^1 B_2(r)B_2(r)'dr, \text{ where } B(r) := \begin{pmatrix} B_1(r) \\ B_2(r) \end{pmatrix},$$

$B_1(r) \in R$, $B_2(r) \in R^s$, and the limiting integral is positive definite with probability one. (The convergence in (12.4) is joint with that of (12.2).)

Assumption 4 holds under the assumptions above. The proof is the same as that given in (8.3)-(8.7) for Example 2. Equation (8.7) holds in the present case by (12.2) and (12.4) because $T^{-1}\sum_{t=1}^T \varepsilon_t X_t = O_p(1)$ and $\left(T^{-2}\sum_{t=1}^T X_t X_t'\right)^{-1/2} \xrightarrow{d} \left(\int_0^1 B_2(r)B_2(r)'dr\right)^{-1/2}$. In consequence, $\left(\sum_{t=1}^T X_t X_t'\right)^{1/2}(\widehat{\theta} - \theta_0) = O_p(1)$. The latter and the fact that $T^{-2}\sum_{t=1}^T X_t X_t' \xrightarrow{d} \int_0^1 B_2(r)B_2(r)'dr$, where $\int_0^1 B_2(r)B_2(r)'dr > 0$ with probability one, imply that $\widehat{\theta} \xrightarrow{p} \theta_0$ and Assumption 4 holds.

12.2. Parameter Space

Assumption 5^{4*} holds in this example provided $g_a(\theta)$ and $g_b(\theta)$ are continuously differentiable on some neighborhood of θ_0 and $\frac{\partial}{\partial\theta'}g(\theta_0)$ is full row rank, where $g(\theta) = (g_a(\theta)', g_b(\theta)')'$. In consequence, by Lemma 2, Assumption 5 holds with

$$(12.5) \qquad \Lambda := \left\{\lambda \in R^s : \frac{\partial}{\partial\theta'}g_a(\theta_0)\lambda = \mathbf{0},\ \frac{\partial}{\partial\theta'}g_b(\theta_0)\lambda \leq \mathbf{0}\right\}.$$

For Λ as such, Assumption 6 holds.

12.3. Asymptotic Distribution of the Least Squares Estimator

In this example, Assumption 7 holds with

$$(12.6) \qquad G = \int_0^1 B_2(r)dB_1(r) + \Pi_{21} \text{ and } \mathcal{J} = \int_0^1 B_2(r)B_2(r)'dr.$$

This follows from (3.21) and (12.1)–(12.4).

By Theorem 3, $T(\widehat{\theta} - \theta_0) \xrightarrow{d} \widehat{\lambda}$, where $\widehat{\lambda}$ satisfies (5.1) with $(G, \mathcal{J})$ defined in (12.6) and Λ defined in (4.16).

12.4. Asymptotic Distributions of Subvectors of the Least Squares Estimator

Typically, the restrictions $g_a(\theta) = \mathbf{0}$ and $g_b(\theta) \leq \mathbf{0}$ in this example only involve some of the elements of θ. In this case, the vector $\frac{\partial}{\partial\theta'}g(\theta_0)$, where $g(\theta) :=$

$(g_a(\theta)', g_b(\theta)')'$, that determines Λ contains some non-zero columns, say p of them, and some columns of zeros, say $s-p$ of them. Without loss of generality, assume that the first p columns of $\frac{\partial}{\partial\theta'}g(\theta_0)$ are non-zero vectors and the last $s-p$ columns are zero vectors for $1 \leq p \leq s$.

We partition X_t such that

$$X_t := \begin{pmatrix} X_{*t} \\ X_{\psi t} \end{pmatrix} := \begin{pmatrix} X_{\beta t} \\ X_{\delta t} \\ X_{\psi t} \end{pmatrix},\ B_2(r) := \begin{pmatrix} B_*(r) \\ B_\psi(r) \end{pmatrix},\ \text{and}$$

$$\mathcal{J} = \int_0^1 B_2(r)B_2(r)'dr := \begin{bmatrix} \mathcal{J}_* & \mathbf{0} \\ \mathbf{0} & \mathcal{J}_\psi \end{bmatrix} := \begin{bmatrix} \int_0^1 B_*(r)B_*(r)'dr & 0 \\ 0 & \int_0^1 B_\psi(r)B_\psi(r)'dr \end{bmatrix}, \tag{12.7}$$

where $X_{*t} \in R^{p+q}$, $X_{\psi t} \in R^r$, $X_{\beta t} \in R^p$, ..., $B_*(r) \in R^{p+q}$, $B_\psi(r) \in R^r$, ... Such a partitioning is always possible, because one could have $r = 0$.

We partition $\widehat{\theta}$, θ_0, and θ conformably with X_t. That is,

$$\widehat{\theta} := (\widehat{\beta}', \widehat{\delta}', \widehat{\psi}')',\ \theta_0 := (\beta_0', \delta_0', \psi_0')',\ \text{and}\ \theta = (\beta', \delta', \psi')', \tag{12.8}$$

where $\widehat{\beta}, \beta_0, \beta \in R^p$, $\widehat{\delta}, \delta_0, \delta \in R^q$, and $\widehat{\psi}, \psi_0, \psi \in R^r$.

Now, with the above partitioning, Assumptions 8 and 9 hold. The matrix $\mathcal{J}$ is block diagonal by (12.7). The matrix B_T is diagonal. The set Λ is a product set $\Lambda_\beta \times \Lambda_\delta \times \Lambda_\psi$ with

$$\Lambda_\beta = \left\{ \lambda_\beta \in R^p : \frac{\partial}{\partial\beta'}g_a(\theta_0)\lambda_\beta = \mathbf{0},\ \frac{\partial}{\partial\beta'}g_b(\theta_0)\lambda_\beta \leq \mathbf{0} \right\},\ \Lambda_\delta = R^q,\ \text{and}\ \Lambda_\psi = R^r$$
$$\text{where}\ \frac{\partial}{\partial\beta'}g_j(\theta_0) \in R^{c_j \times p}\ \text{for}\ j = a, b. \tag{12.9}$$

With conformable partitioning to that above, we have

$$G := \begin{pmatrix} G_* \\ G_\psi \end{pmatrix},\ G_* := \int_0^1 B_*(r)dB_1(r) + \Pi_*,$$
$$G_\psi := \int_0^1 B_\psi(r)dB_1(r) + \Pi_\psi,\ \text{and}\ \Pi_{21} := \begin{pmatrix} \Pi_* \\ \Pi_\psi \end{pmatrix}, \tag{12.10}$$

where G_*, $\Pi_* \in R^{p+q}$ and G_ψ, $\Pi_\psi \in R^r$. Conditional on $B_*(.)$, G_* has a normal distribution and, hence, Assumption 7* holds. If the regressors X_{*t} are (asymptotically) strictly exogenous, i.e., $B_*(r)$ and $B_1(r)$ are independent and $\Pi_* = \mathbf{0}$, then

$$G_* \sim N(\mathbf{0}, \Omega_{11}\mathcal{J}_*)\ \text{conditional on}\ B_*(.), \tag{12.11}$$

where Ω_{11} is the (1,1) element of the covariance matrix Ω of the Brownian motion $B(\cdot)$. In this case, Assumption 7^{3*} holds with $c = \Omega_{11}$.

By Corollary 1(b) and the fact that $\Lambda_\psi = R^r$, we obtain

$$T(\widehat{\psi} - \psi_0) \xrightarrow{d} \widehat{\lambda}_\psi = Z_\psi := \mathcal{J}_\psi^{-1}G_\psi, \tag{12.12}$$

where $\mathcal{J}_\psi$ and G_ψ are defined in (12.7) and (12.10) respectively.

By Corollary 1(b),

$$T(\widehat{\beta} - \beta_0) \xrightarrow{d} \widehat{\lambda}_\beta, \tag{12.13}$$

where $\widehat{\lambda}_\beta$ solves $q_\beta(\widehat{\lambda}_\beta) = \inf_{\lambda_\beta \in \Lambda_\beta} q_\beta(\lambda_\beta)$ with Λ_β as in (12.10) and $q_\beta(\lambda_\beta)$ is defined using (12.7) and (12.10).

By Corollary 1(b),

$$\begin{aligned} & T(\widehat{\delta} - \delta_0) \xrightarrow{d} \widehat{\lambda}_\delta = \mathcal{J}_\delta^{-1} G_\delta - \mathcal{J}_\delta^{-1} \mathcal{J}_{\delta\beta} \lambda, \text{ where} \\ & G_* := \begin{pmatrix} G_\beta \\ G_\delta \end{pmatrix}, \ G_\delta := \int_0^1 B_\delta(r) dB_1(r) + \Pi_\delta, \ B_*(r) := \begin{pmatrix} B_\beta(r) \\ B_\delta(r) \end{pmatrix}, \\ & \Pi_* := \begin{pmatrix} \Pi_\beta \\ \Pi_\delta \end{pmatrix}, \ \mathcal{J}_* := \begin{bmatrix} \mathcal{J}_\beta & \mathcal{J}_{\beta\delta} \\ \mathcal{J}_{\delta\beta} & \mathcal{J}_\delta \end{bmatrix} = \begin{bmatrix} \int_0^1 B_\beta(r) B_\beta(r)' dr & \int_0^1 B_\beta(r) B_\delta(r)' dr \\ \int_0^1 B_\delta(r) B_\beta(r)' dr & \int_0^1 B_\delta(r) B_\delta(r)' dr \end{bmatrix}, \end{aligned} \tag{12.14}$$

G_δ, $B_\delta(r)$, $\Pi_\delta \in R^q$, and $\mathcal{J}_\delta \in R^{q \times q}$.

13. Appendix of Proofs for the GARCH(1,q^*) Model

13.1. Stationarity and Ergodicity

First, we establish strict stationarity and ergodicity (S&E) of $\{h_{0t} : t = ..., 0, 1, ...\}$. From this and the assumed S&E of $\{X_t : t = ..., 0, 1, ...\}$, we obtain directly the S&E of numerous random variables including ε_t, Y_t, $h_t(\theta)$, $e_t(\theta)$, $(:= Y_t - \theta_5 - X_t'\theta_6)$, $\ell_{tt}(\theta)$, $\frac{\partial}{\partial\theta} h_t(\theta)$, $\frac{\partial}{\partial\theta} e_t(\theta)$, $\frac{\partial}{\partial\theta} \ell_{tt}(\theta)$, $\frac{\partial^2}{\partial\theta\partial\theta'} h_t(\theta)$, $\frac{\partial^2}{\partial\theta\partial\theta'} e_t(\theta)$, and $\frac{\partial^2}{\partial\theta\partial\theta'} \ell_{tt}(\theta)$. We use S&E to obtain laws of large numbers (LLNs) for various quantities.

To establish the S&E of h_{0t}, we write $h_{0t} = c_0 \sum_{k=0}^{\infty} M(t,k)$, where $c_0 := \theta_{40}(1 - \theta_{30})$ and $M(t,k)$ is a linear combination of products of the squared innovations z_s^2 for $s = t-1, t-2, ...,$ as in Bollerslev (1986, Proof of Theorem 1). Because $Eh_{0t}^{p^*} < \infty$ for some t by Assumption (c) of (10.9), $c_0 \sum_{k=0}^{\infty} M(t,k) < \infty$ a.s. By stationarity of z_t, this implies that $c_0 \sum_{k=0}^{\infty} M(t,k) < \infty$ a.s. $\forall t = ..., 0, 1, ...$ That is, $c_0 \sum_{k=0}^{\infty} M(t,k)$ is R-valued a.s. $\forall t$. Furthermore, this sum is clearly a *measurable* R-valued function of the S&E sequence $\{z_t : t = ..., 0, 1, ...\}$ and, hence, is S&E itself (e.g., by Theorem 3.5.8 of Stout (1974)).

13.2. Verification of Assumption 1

Our approach to verifying Assumptions 1 and 4 for this example is to extend results in the literature (mainly moment bounds) for a GARCH(1,1) model with intercept (including an IGARCH(1,1) model) to a GARCH$(1,q^*)$ model with regression function. This approach dramatically reduces the length of the proof because moment bounds in GARCH models are complicated to derive, but very similar methods can be used for GARCH$(1,q^*)$ models as for GARCH(1,1) models.

Specifically, we extend results of Lee and Hansen (1994), hereafter denoted LH. To make the extension of LH's results as simple as possible to understand, we adopt the same symbols as LH for the parameters that are common to both models. We let β denote the GARCH–AR parameter θ_3 (thus, β differs here from in Section), ω denotes the variance intercept parameter θ_4, and $(\alpha_1, ..., \alpha_{q^*})$ denote the GARCH–MA parameters on the variables $e^2_{t-1}(\theta), ..., e^2_{t-q^*}(\theta)$, respectively, i.e., the elements of θ_1 and θ_2 arranged in the order specified. LH's single GARCH–MA parameter is denoted α. (Our model does not require that all of the lags $e^2_{t-1}(\theta), ..., e^2_{t-q^*}(\theta)$ enter the model. If $e^2_{t-j}(\theta)$ does not enter the model, the corresponding α_j is set equal to zero.) The mean in LH's model is γ. In our model, it is $\theta_5 + X_t'\theta_6$. For present purposes, we let θ denote $(\alpha_1, ..., \alpha_{q^*}, \beta, \omega, \theta_5, \theta_6')'$.

We need to extend LH's results to allow for $q^* > 1$ and the mean function $\theta_5 + X_t'\theta_6$, rather than $q^* = 1$ and the mean function γ. In other respects, our models are the same. In particular, the assumptions on the innovations $\{z_t : t = ..., 0, 1, ...\}$ are the same. The restrictions on the parameter space are the same if one views our α_1 as their α — both are bounded away from zero. Our restrictions on $\alpha_2, ..., \alpha_{q^*}$ (that $\alpha_j \in [0, \alpha_{ju}]$ for some $0 < \alpha_{ju} < \infty$ $\forall j = 2, ..., q^*$) are novel of course, because these parameters do not appear in LH.

Define

$$\mathcal{L}_T(\theta) := \sum_{t=1}^{T} \ell_{tt}(\theta) := -\tfrac{T}{2}\ln(2\pi) - \tfrac{1}{2}\sum_{t=1}^{T}\ln h_t(\theta) - \tfrac{1}{2}\sum_{t=1}^{T} e_t^2(\theta)/h_t(\theta). \tag{13.1}$$

Note that $\mathcal{L}_T(\theta)$ is comprised of sums of strictly stationary and ergodic random variables. In consequence, it is more amenable to analysis than is the quasi-likelihood $\ell_T(\theta)$.

We verify Assumption 1 for $\ell_T(\theta)$ by showing that it holds for $\mathcal{L}_T(\theta)$ and that $\ell_T(\theta)$ is closely approximated by $\mathcal{L}_T(\theta)$ in the sense that (3.6) holds.

To establish (3.6), we write the left-hand side of (3.6) as

$$\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T}\left|\sum_{t=1}^{T} a_t(\theta)\right|, \text{ where}$$

$$a_t(\theta) := \ln\left(\frac{h_t(\theta)}{h_t^*(\theta)}\right) - \ln\left(\frac{h_t(\theta_0)}{h_t^*(\theta_0)}\right) - \tfrac{1}{2}\left(\frac{e_t^2(\theta)}{h_t^*(\theta)} - \frac{e_t^2(\theta)}{h_t(\theta)} - \frac{\varepsilon_t^2}{h_t^*(\theta)} + \frac{\varepsilon_t^2}{h_t(\theta_0)}\right). \tag{13.2}$$

The functions $h_t(\theta)$, $h_t^*(\theta)$, and $e_t^2(\theta)$ are continuous in θ a.s. Hence, $\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T} |a_t(\theta)| \to 0$ a.s. as $T \to \infty$ $\forall t \geq 1$. The argument of the proof of Lemma LH3 applies to our model with minor adjustments and yields $\sum_{t=1}^{\infty}\sup_{\theta\in\Theta}|a_t(\theta)| < \infty$ a.s. (Note that the proof uses Lemma LH2(3) which is extended below to cover our model.) Hence, given $\varepsilon > 0$, $\exists T_1 < \infty$ such that $\sum_{t=T_1+1}^{\infty}\sup_{\theta\in\Theta}|a_t(\theta)| < \varepsilon/2$. And, given $\varepsilon > 0$ and $T_1 < \infty$, $\exists T_2 < \infty$ such that $\forall T \geq T_2$ we have $\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T}|a_t(\theta)| < \varepsilon/(2T_1)$ $\forall t \leq T_1$. Combining these results gives

$$\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T}\left|\sum_{t=1}^{T} a_t(\theta)\right| \leq \sum_{t=1}^{T_1}\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\gamma_T}|a_t(\theta)| + \sum_{t=T_1+1}^{\infty}\sup_{\theta\in\Theta}|a_t(\theta)| < \varepsilon. \tag{13.3}$$

Hence, (3.6) holds.

We verify Assumption 1 for $\mathcal{L}_T(\theta)$ using Assumption 1* and Lemma 6(b). (Note that the latter applies because it is shown below that $E\left\|\frac{\partial^2}{\partial\theta\partial\theta'}\ell_{tt}(\theta_0)\right\| < \infty$ and, hence, $-T^{-1}\frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{L}_T(\theta_0) \xrightarrow{p} \mathcal{J}$ by the ergodic theorem.

We verify Assumption 1*(c) for $\mathcal{L}_T(\theta)$ by showing that $\sup_{\theta\in\Theta_0}\left\|\frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{L}_T(\theta) - E\frac{\partial^2}{\partial\theta\partial\theta'}\ell_{tt}(\theta)\right\| \xrightarrow{p} 0$, for some set $\Theta_0 \subset \Theta$ that contains $\Theta \cap S(\theta_0, \varepsilon)$ for some $\varepsilon > 0$, and $E\frac{\partial^2}{\partial\theta\partial\theta'}\ell_{tt}(\theta)$ is continuous at θ_0. Both of these results follow from the uniform LLN given in Theorem 6 of Andrews (1992) using Assumption TSE–1D provided

$$E\sup_{\theta\in\Theta_0}\left\|\frac{\partial^2}{\partial\theta\partial\theta'}\ell_{tt}(\theta)\right\| < \infty, \tag{13.4}$$

because $\frac{\partial^2}{\partial\theta\partial\theta'}\ell_{tt}(\theta)$ is stationary and ergodic and continuous in θ a.s. (Note that the above uniform LLN has simpler conditions to verify than the one used by LH. It avoids the need to consider third derivatives of $\ell_{tt}(\theta)$ as is done in LH. It is applicable because $\left\{\frac{\partial^2}{\partial\theta\partial\theta'}\ell_{tt}(\theta) : t = 1, 2, ...\right\}$ are identically distributed.)

We verify (13.4) by strengthening Lemma LH11(1) from "for all $\theta \in \Theta_4$, $E\left\|\frac{\partial^2}{\partial\theta\partial\theta'}\ell_{tt}(\theta)\right\| < \infty$," where Θ_4 is a neighborhood of θ_0, to the result of (13.4) (which takes the supremum before the expectation) and extending Lemma LH11(1) to our more general model. Lemma LH11(1) relies on Lemmas LH1, LH2, LH4–LH7, LH8(1), and LH10(1). Hence, we need to strengthen and extend each of these Lemmas in the same direction. The strengthening of these results in terms of taking the supremum over θ before the expectation turns out to be simple because the bounds used by LH are always in terms of upper and lower bounds on the parameter space and these bounds hold uniformly for all points in the parameter space.

To extend LH's Lemmas to our model, we define the function $h_t^\varepsilon(\theta)$ to be

$$h_t^\varepsilon(\theta) := \omega + \sum_{k=0}^{\infty}\beta^k\sum_{j=1}^{q^*}\alpha_j\varepsilon_{t-k-j}^2. \tag{13.5}$$

Most of LH's moment bounds for functions of $h_t(\theta)$ are derived using $h_t^\varepsilon(\theta)$ (with $\alpha_j = 0$ for $j \neq 1$) in place of $h_t(\theta)$, which is justified by their Lemma LH1, which states that $B^{-1}h_t^\varepsilon(\theta) \leq h_t(\theta) \leq Bh_t^\varepsilon(\theta)$ a.s. for a constant $0 < B < \infty$. Lemma LH1 extends to our model with minor adjustments, such as the replacement of terms like $\alpha e_{t-1-k}^2(\theta)$ by terms like $\sum_{j=1}^{q^*}\alpha_j e_{t-j-k}^2(\theta)$, and with the adjustment that B is of the form

$$B = C(1 + ||X_t|| + ||X_t||^2) \text{ for some constant } 0 < C < \infty. \tag{13.6}$$

The latter adjustment occurs because LH's upper and lower bounds on the intercept, γ_ℓ and γ_u, must be replaced by upper and lower bounds on the regression function, which depend on $||X_t||$.

We relate $h_t^\varepsilon(\theta)$ for our model to $h_t^\varepsilon(\theta)$ for a GARCH(1,1) model using the following simple, but key, bounds:

$$\omega + \alpha_1 \sum_{k=0}^{\infty} \beta^k \varepsilon_{t-k-1}^2 \leq \omega + \sum_{k=0}^{\infty} \beta^k \sum_{j=1}^{q^*} \alpha_j \varepsilon_{t-k-j}^2 := h_t^\varepsilon(\theta)$$

$$= \omega + \left(\sum_{j=1}^{q^*} \alpha_j \beta^{1-j} \right) \sum_{k=0}^{\infty} \beta^k \varepsilon_{t-k-1}^2 - \sum_{k=0}^{q^*-2} \beta^k \sum_{j=k+2}^{q^*} \alpha_j \beta^{1-j} \varepsilon_{t-k-1}^2$$

$$\leq \omega + \left(\sum_{j=1}^{q^*} \alpha_j \beta^{1-j} \right) \sum_{k=0}^{\infty} \beta^k \varepsilon_{t-k-1}^2, \tag{13.7}$$

where both inequalities use the fact that $\alpha_2, ..., \alpha_{q^*}$ are non-negative and the equality holds by summing the coefficients on all of the lags of ε_t^2 of the same lag order.

Note that the lower bound on $h_t^\varepsilon(\theta)$ for the GARCH$(1, q^*)$ model is exactly the same function of the parameters and lagged ε_t^2's as for a GARCH(1,1) model with a GARCH–MA parameter of α_1 and the upper bound is exactly the same as for a GARCH(1,1) model with a GARCH–MA parameter of $\sum_{j=1}^{q^*} \alpha_j \beta^{1-j}$. (Of course, the random variables ε_{t-k-1}^2 for $k = 0, 1, ...$ have different properties in the GARCH$(1, q^*)$ case than in the GARCH (1,1) case, but this turns out not to cause any difficulties.)

Given our assumption that the first element of θ_2 is bounded away from zero, see (10.4), and that this element is the GARCH–MA parameter on the first lag of ε_t^2, see (10.1), we have that α_1 is bounded away from zero for $\theta \in \Theta$. Thus, the lower bound in (13.7) is not "degenerate." That is, it does not reduce to just ω for any $\theta \in \Theta$.

Let α_{j0}, $\alpha_{j\ell}$, and α_{ju} denote the true value of α_j, the lower bound on α_j given in Θ, and the upper bound on α_j given in Θ, respectively, for $j = 1, ..., q^*$. By assumption, $\alpha_{1\ell} > 0$, $\alpha_{j\ell} = 0 \; \forall j = 2, ..., q^*$, and $\alpha_{ju} < \infty \; \forall j = 1, ..., q^*$. Define ω_0, ω_ℓ, and ω_u analogously. Define η_ℓ and η_u as in LH. (They are values that satisfy $0 < \eta_\ell < \beta_0 < \eta_u < 1$.)

Now, the conclusions of Lemma LH2(1) and LH2(2) regarding stationarity hold in our case by the argument given at the beginning of this section. Lemma LH2(3) holds by the proof in LH using Assumption (c) of (10.9) in place of Theorem 4 of Nelson (1990), with α_0 replaced by α_{10}, and using (13.7).

Parts (1) and (2) of Lemma LH4 hold without change in our GARCH$(1, q^*)$ model because these parts only involve the innovations z_t, which are the same here as in LH. In the proof of part (3) of Lemma LH4, the first equation on p. 44 holds for the GARCH$(1, q^*)$ case with the equality replaced by the inequality "$\leq$" and α_0 replaced by α_{10}, because $\sum_{j=2}^{q^*} \alpha_{j0} h_{0t-j} z_{t-j}^2 \geq 0$. The rest of the proof holds unchanged. Parts (4) and (5) hold by the argument of LH using equation (13.7). The bounding constants in this case are $K_\ell := \omega_u/\omega_0 + \sum_{j=1}^{q^*} \alpha_{ju}/\alpha_{10}$ and $H_u := \omega_0/\omega_\ell + \sum_{j=1}^{q^*} \alpha_{j0}/\alpha_{1\ell}$.

Lemma LH5(1) holds for our GARCH$(1, q^*)$ model by LH's proof except that the equality in the third last equation on p. 44 is an inequality ($\leq$), α is replaced by α_{10}, α_ℓ by $\alpha_{1\ell}$, etc. Also, the proof of LH contains a small error that needs to be corrected.

The first equality in the last equation on p. 44 is not correct. The expectation of the product does not equal the product of the conditional expectations because the terms in the product contain overlapping random variables. One needs to replace the $h^{\varepsilon}_{t-i}(\theta)/h^{\varepsilon}_{t-i+1}(\theta)$ terms by the bound given in the third last equation on p. 44 before switching the order of the expectation and the product. Lemma LH5(2) holds with $H_c := \omega_0/\omega_\ell + \sum_{j=1}^{q^*} \alpha_{j0}/(\alpha_{1\ell}\eta_\ell)$. The proof is analogous to that of LH using (13.7).

Lemma LH6(1) holds for our GARCH$(1, q^*)$ model with $K_c := \omega_u/\omega_0 + \sum_{j=1}^{q^*} \alpha_{j0}/(\alpha_{10}\eta_u)$. The proof is analogous to that of LH using (13.7). Lemma LH6(2) holds with H_u and K_u defined as above by the same argument as in LH. We note that, in the last equation on p. 45 of LH, h_{0t-k+1} should be h_{0t-k} and $\mathcal{F}_{t-k}$ should be $\mathcal{F}_{t-k-1}$.

Lemma LH7(1) holds by the proof in LH with B replaced by EB in the definition of H_1, with $(\gamma_u - \gamma_\ell)^2$ replaced by $E||X_t^2|| \cdot ||\theta_{6u} - \theta_{6\ell}||^2$ in the definition of H_1, and with B and H_c defined as above. Thus, H_1 equals a constant times $(1 + E||X_t|| + E||X_t||^2)$. Lemma LH7(2) holds as in LH.

Lemma LH8(1) holds with B^2 replaced by $||B^2||_r$ in the definition of H_β, where B is as in (13.6). Thus, we need $E||X_t||^{4r} < \infty$. By optimizing the split in the application of Holder's inequality in the proof of Lemma LH11(1), we find that we need Lemma LH8(1) to hold for $r = 5/2$. Hence, we require $E||X_t||^{10} < \infty$, as is assumed in Assumption (b) of (10.9).

Lemma LH10(1) holds as in LH, but with H_β defined with B^2 replaced by $||B^2||_{2r}$. Thus, we need $E||X_t||^{8r} < \infty$. By optimizing the split in Holder's inequality in the proof of Lemma LH11(1), we find that we need Lemma LH10(1) to hold for $r = 5/4$. Hence, we require $E||X_t||^{10} < \infty$, as is assumed in (10.9).

Lemma LH11(1) holds for our model by the proof in LH with $g\ (= \gamma - \gamma_0)$ replaced by $\theta_5 - \theta_{50} + X_t'(\theta_6 - \theta_{60})$, with the first H_β in the fourth last line of p. 50 defined as described above in the discussion of Lemma LH8(1) with $r = 2$, with the second H_β in the same line defined as described above in the discussion of Lemma LH10(1) with $r = 1$, and with the application of Holder's inequality between the third-last and second-last lines on p. 50 of LH altered to give a split of $||W_1 W_2|| \leq ||W_1||_5 \cdot ||W_2||_{5/4}$, where W_1 is the first term in parentheses on the third-last line and W_2 is the term in brackets on that line. This completes the proof for the second derivative of $\ell_{tt}(\theta)$ with respect to β. For the second partial and cross-partial derivatives with respect to the other parameters in θ, bounds can be obtained by similar methods using our extensions of Lemmas LH5–LH7, LH8(1), and LH10(1). The moment conditions on $||X_t||$ required for these cases are weaker than those required for the second partial derivative with respect to β.

This completes the proof of (13.4), the proof that Assumption 1 holds for $\mathcal{L}_T(\theta)$, and the proof that Assumption 1 holds for $\ell_T(\theta)$.

13.3. Verification of Assumptions 2 and 7

Next, we establish that $E\left\|\frac{\partial}{\partial\theta}\ell_{tt}(\theta_0)\right\|^2 < \infty$. This condition is required to apply the martingale difference CLT when verifying Assumptions 2 and 7 of Sections 3 and

5 respectively. This result follows by an extension of Lemma LH9(1) to our model. Note that we only require the result of Lemma LH9(1) for the case of $\theta = \theta_0$, which simplifies the proofs somewhat. Lemma LH9(1) relies on Lemma LH8(2). We only need the latter to hold for $\theta = \theta_0$.

Lemma LH8(2) extends to our model with the proof unchanged for the derivatives with respect to (wrt) ω and α. The proof for the derivative wrt γ (i.e., wrt θ_5 and θ_6 in our case) differs from LH because an X_t multiplicand appears when considering θ_6. This requires that Lemma LH7(1) is extended to include an $||X_t||$ multiplicand. It can be, given that $\{X_t\}$ is independent of $\{z_t\}$ and $E||X_t||^3 < \infty$. The proof for the derivative wrt β is the same as in LH except that $g = 0$ because we only need to consider the case where $\theta = \theta_0$.

Lemma LH9(1) now holds for our model by the proof given by LH, but with some simplifications due to the fact that we only require it to hold for $\theta = \theta_0$. In particular, we only require Lemma LH8(1) to hold with $r = 2$ rather than $r = 4$. Thus, the moment conditions on X_t given in (10.9) are sufficient.

13.4. Verification of Consistency

Here we establish consistency of the QML estimator in the non-integrated GARCH$(1, q^*)$ case (i.e., when $\beta_0 + \sum_{j=1}^{q^*} \alpha_{0j} < 1$) and "local consistency" of the QML estimator in the integrated GARCH$(1, q^*)$ case (i.e., when $\beta_0 + \sum_{j=1}^{q^*} \alpha_{0j} = 1$) by verifying Assumptions 4*(a) and 4*(b*). By "local consistency" we mean that, for some set Θ_1 that is a subset of Θ that contains $\Theta \cap S(\theta_0, \varepsilon)$ for some $\varepsilon > 0$, the maximization of $\ell_T(\theta)$ over $\theta \in \Theta_1$ leads to a consistent estimator of θ_0. The set Θ_1 that we consider is $\Theta_1 := \{\theta \in \Theta : \beta_{1\ell} \leq \beta \leq \beta_{1u}\}$, where $\beta_{1\ell}$ and β_{1u} are defined on p. 36 of LH such that $\beta_{1\ell} \leq \beta_0 < \beta_{1u}$. The difficulty in treating β values that are distant from β_0 is that they may yield a value of the likelihood function whose expectation is negative infinity in the integrated model.

To verify Assumption 4*(a) for $T^{-1}\ell_T(\theta)$, it suffices to verify Assumption 4*(a) for $T^{-1}\mathcal{L}_T(\theta)$, where $\mathcal{L}_T(\theta)$ is defined in (13.1), and to show that

$$\sup_{\theta \in \Theta} |T^{-1}\ell_T(\theta) - T^{-1}\mathcal{L}_T(\theta)| \xrightarrow{p} 0. \tag{13.8}$$

Equation (13.8) holds by the extension of Lemma LH3 to our model, which requires only minor adjustments to the proof given in LH.

Now, to verify Assumption 4*(a) for $\mathcal{L}_T(\theta)$, we use the same uniform LLN as in Section 13.2 just above equation (13.4). Because $\ell_{tt}(\theta)$ is stationary and ergodic and continuous in θ, it suffices to show that (i) $E \sup_{\theta \in \Theta} |\ell_{tt}(\theta)| < \infty$ in the non-integrated GARCH case and (ii) $E \sup_{\theta \in \Theta_1} |\ell_{tt}(\theta)| < \infty$ in the integrated GARCH case. Result (i) holds by the first part of the proof (that does not involve derivatives) of Theorem LH2, using (13.7), and adjusted (easily) to take the supremum over θ before the expectation. Result (ii) holds by the proof of Lemma LH7(2), which has been extended to our model in Section 13.2.

Next, we verify Assumption 4*(b*). The uniform LLN used above delivers continuity of the limit function

$$\ell(\theta) := E\ell_{tt}(\theta) \tag{13.9}$$

on Θ_1 in the integrated case and on Θ in the non-integrated case. Both Θ_1 and Θ are compact. Hence, it remains to show that $\ell(\theta)$ is uniquely minimized at θ_0 over Θ_1 in the integrated case and over Θ in the non-integrated case. These two cases can be treated simultaneously by considering an arbitrary element θ for which $\ell(\theta)$ is well-defined, i.e., $\theta \in \Theta_1$ in the integrated case and $\theta \in \Theta$ in the non-integrated case.

We have

$$\ell(\theta) := E\ell_{tt}(\theta) = -\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}E\ln(h_t(\theta)) - \tfrac{1}{2}Ee_t^2(\theta)/h_t(\theta) \text{ and}$$

$$\begin{aligned} Ee_t^2(\theta)/h_t(\theta) &= E\varepsilon_t^2/h_t(\theta) + \begin{pmatrix}\theta_5-\theta_{50}\\ \theta_6-\theta_{60}\end{pmatrix}' E\begin{pmatrix}1\\ X_t\end{pmatrix}\begin{pmatrix}1\\ X_t\end{pmatrix}'\begin{pmatrix}\theta_5-\theta_{50}\\ \theta_6-\theta_{60}\end{pmatrix} \\ &\geq E\varepsilon_t^2/h_t(\theta) = Eh_{0t}/h_t(\theta) \end{aligned} \tag{13.10}$$

with strict inequality unless $\theta_5 = \theta_{50}$ and $\theta_6 = \theta_{60}$ because $E\begin{pmatrix}1\\ X_t\end{pmatrix}\begin{pmatrix}1\\ X_t\end{pmatrix}'$ is positive definite (pd). The function $\ln(x)+y/x$ is uniquely minimized over x at $x = y$. Hence,

$$\ell(\theta) \leq -\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}E\ln(h_{t0}) - \tfrac{1}{2} = \ell(\theta_0) \tag{13.11}$$

with equality iff $\theta_4 = \theta_{40}$, $\theta_5 = \theta_{50}$, and $P(h_t(\theta) = h_{0t}) = 1$.

It suffices to show that for any θ with $\theta_5 = \theta_{50}$ and $\theta_6 = \theta_{60}$, $P(h_t(\theta) = h_{0t}) = 1$ iff $\theta = \theta_0$. For θ with $\theta_5 = \theta_{50}$ and $\theta_6 = \theta_{60}$, some algebraic manipulations yield

$$h_t(\theta) - h_{0t} = \theta_3(h_{t-1}(\theta) - h_{0t-1}) + (\varepsilon_{1t}^{2\prime}, \varepsilon_{2t}^{2\prime}, h_{0t-1}, 1)'(\overline{\theta} - \overline{\theta}_0), \tag{13.12}$$

where $\overline{\theta} = (\theta_1', \theta_2', \theta_3, \theta_4(1-\theta_3))'$ and $\overline{\theta}_0 = (\theta_{10}', \theta_{20}', \theta_{30}, \theta_{40}(1-\theta_{30}))'$. By stationarity of $\{h_t(\theta) - h_{0t} : t = ..., 0, 1...\}$, $h_t(\theta) - h_{0t} = 0$ a.s. iff $h_{t-1}(\theta) - h_{0t-1} = 0$ a.s. Combining this with (13.12), we find that $h_t(\theta) - h_{0t} = 0$ a.s. iff

$$W_t'(\overline{\theta} - \overline{\theta}_0) = 0 \text{ a.s., where } W_t := (\varepsilon_{1t}^{2\prime}, \varepsilon_{2t}^{2\prime}, h_{0t-1}, 1)'. \tag{13.13}$$

We now show that $W_t'\lambda = 0$ a.s. iff $\lambda = \mathbf{0}$. This implies that $\overline{\theta} = \overline{\theta}_0$, which completes the proof that θ_0 uniquely minimizes $\ell(\theta)$. By repeated substitution into the formula in (10.1) for h_{0t}, we can write

$$W_t'\lambda = c_0 + \sum_{j=1}^{q^*} c_j z_{t-j}^2 + \lambda_3 K h_{0t-q^*} \tag{13.14}$$

for some constants $c_0, ..., c_{q^*}$, and K, where $\lambda = (\lambda_1', \lambda_2', \lambda_3, \lambda_4)'$, λ_3 is the coefficient on h_{0t-1} in $W_t'\lambda$, and $0 < \beta_0^{q^*-1} \leq K < \infty$.

Suppose $W_t'\lambda = 0$ a.s. By Assumption (a) of (10.9), $P(z_t^2 = 1) \neq 1$. This implies that $P(z_{t-1}^2 = 1|\mathcal{F}_{t-2}) \neq 1$ with positive probability. Thus, conditional on $\mathcal{F}_{t-2}$, all the terms on the right-hand side (rhs) of (13.14) are constants except z_{t-1}^2. Because the rhs equals zero by assumption, we must have $c_1 = 0$. Repeating this argument for z_{t-j}^2 with $j = 2, ..., q^*$ yields $0 = c_2 = \cdots = c_{q^*}$. Next, h_{0t-q^*} is not a constant because $P(z_t^2 = 1) \neq 1$ $\forall t$, so $\lambda_3 = 0$. Given that $\lambda_3 = 0$, c_0 must equal λ_4. Hence, $\lambda_4 = 0$. Given $\lambda_3 = \lambda_4 = 0$, c_1 equals the first element of λ_2 multiplied by $\omega_0(1-\beta_0) > 0$ and, hence, the first element of λ_2 equals zero. Continuing for $c_2, ..., c_{q^*}$, we obtain $\lambda_1 = \mathbf{0}$ and $\lambda_2 = \mathbf{0}$, as desired. This completes the verification of Assumption 4*(b*).

13.5. Verification of Assumption 3

We now show that $\mathcal{J}$, defined in (10.7), is pd, as required by Assumption 3. By (10.13), $\mathcal{J}$ is block diagonal between θ_6 and the other parameters. So, it suffices to show positive definiteness of each block. The block corresponding to θ_6 equals $EX_t^* X_t^{*\prime}/h_{0t}$, which is pd because $EX_t^* X_t^{*\prime}$ is pd and $0 < \omega_0(1 - \beta_0) \leq h_{0t} < \infty$ a.s. Note that $EX_t^* X_t^{*\prime}$ is pd iff $E \begin{pmatrix} 1 \\ X_t \end{pmatrix} \begin{pmatrix} 1 \\ X_t \end{pmatrix}'$ is pd and the latter is assumed in (10.9).

Next, consider the block that corresponds to the other parameters. For notational convenience, let θ denote θ with θ_6 deleted. Write the upper blocks of the two summands of $\mathcal{J}$ in (10.7) as $M_1 + M_2$. Suppose $M_1 + M_2$ is not pd. (We will derive a contradiction.) Then, there exists a $\lambda_* \neq \mathbf{0}$ such that $\lambda_*'(M_1 + M_2)\lambda_* = 0$. Because M_1 is positive semi-definite, this implies that $\lambda_*' M_2 \lambda_* = 0$. The matrix M_2 equals $E(\mathbf{0}' \;\; 1)(\mathbf{0}' \;\; 1)'/h_{0t}$. Hence, the last element of λ_*, which corresponds to θ_5 must be zero.

Let λ denote λ_* with its last element deleted. Let M_{11} equal M_{11} with its last row and column deleted. By the results above, $\lambda_*'(M_1 + M_2)\lambda_* = \lambda' M_{11} \lambda$. Note that $M_{11} = \frac{1}{2} E \frac{\partial}{\partial \overline{\theta}} h_t(\theta_0) \frac{\partial}{\partial \overline{\theta}'} h_t(\theta_0)/h_{0t}^2$, where $\overline{\theta}$ denotes θ with θ_5 and θ_6 deleted. Thus, $\lambda' M_{11} \lambda = 0$ implies that $\frac{\partial}{\partial \overline{\theta}} h_t(\theta_0)'\lambda = 0$ a.s. We have

$$\frac{\partial}{\partial \overline{\theta}} h_t(\theta_0) = W_t + \beta_0 \frac{\partial}{\partial \overline{\theta}} h_{t-1}(\theta_0) \tag{13.15}$$

for W_t as in (13.13). Hence, by stationarity of $\frac{\partial}{\partial \overline{\theta}} h_t(\theta_0)$, $\frac{\partial}{\partial \overline{\theta}} h_t(\theta_0)'\lambda = 0$ a.s. implies that $W_t'\lambda = 0$ a.s. The latter implies that $\lambda = 0$ by the argument given in the paragraphs containing and following (13.14), which is a contradiction. We conclude that $M_1 + M_2$ and $\mathcal{J}$ are pd.

14. Footnotes

[1]The author thanks Arthur Lewbel, David Pollard, and Chris Sims for helpful comments, Glena Ames for typing the manuscript, and Rosemarie Lewis for proof-reading the manuscript. The author gratefully acknowledges the research support of the National Science Foundation via grant number SBR–9410675.

[2]The sufficiency of Assumption 4* for Assumption 4 holds by the following argument: Given $\varepsilon > 0$, there exists a $\delta > 0$ such that $\theta \notin S(\theta_0, \varepsilon)$ implies that $\ell(\theta_0) - \ell(\theta) \geq \delta > 0$. Thus,

$$\begin{aligned} P(\widehat{\theta} \notin S(\theta_0, \varepsilon)) &\leq P(\ell(\theta_0) - T^{-1}\ell_T(\widehat{\theta}) + T^{-1}\ell_T(\widehat{\theta}) - \ell(\widehat{\theta}) \geq \delta) \\ &\leq P(\ell(\theta_0) - T^{-1}\ell_T(\theta_0) + o_p(1) + T^{-1}\ell_T(\widehat{\theta}) - \ell(\widehat{\theta}) \geq \delta) \\ &\leq P(2 \sup_{\theta \in \Theta} |T^{-1}\ell_T(\theta) - \ell(\theta)| + o_p(1) \geq \delta) \xrightarrow{p} 0. \end{aligned}$$

15. 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," *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): "A Simple Counterexample to the Bootstrap," unpublished manuscript, Cowles Foundation for Research in Economics, Yale University.

_______ (1997b): "Hypothesis Testing with a Restricted Parameter Space," *Journal of Econometrics,* forthcoming.

_______ (1997c): "Testing When a Parameter Is on the Boundary of the Maintained Hypothesis: Theory and Applications," manuscript in preparation, Cowles Foundation, Yale University.

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.

Arcones, M. A. and B. Yu (1994): "Central Limit Theorems for Empirical and *U*-processes of Stationary Mixing Sequences," *Journal of Theoretical Probability,* 7, 47–71.

Berry, S. (1994): "Estimating Discrete Choice Models of Product Differentiation," *Rand Journal of Economics*, 25, 242–262.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Billingsley, P. (1968): *Convergence of Probability Measures.* New York: Wiley.

Bollerslev, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics,* 31, 307–327.

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.

Doukhan, P., P. Massart, and E. Rio (1995): "Invariance Principles for Absolutely Regular Empirical Processes," *Annals of the Institute of Henri Poincaré*, 31, 393–427.

Gallant, A. R. and G. H. Golub (1984): "Imposing Curvature Restrictions on Flexible Functional Forms," *Journal of Econometrics,* 26, 295-321.

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.

Hausman, J. A. and D. Wise (1978): "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46, 403–426.

Horowitz, J. L. (1993): "Semiparametric Estimation of a Work-trip Mode Choice Model," *Journal of Econometrics*, 58, 49–70.

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, 221–233. Berkeley: University of California.

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.

Lee, S. W. and B. E. Hansen (1994): "Asymptotic Theory for the GARCH(1,1) Quasi-maximum Likelihood Estimator," *Econometric Theory,* 10, 29–52.

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.

Lumsdaine, R. L. (1996): "Consistency and Asymptotic Normality of the Quasi-Maximum Likelihood Estimator in IGARCH(1,1) and Covariance Stationary GARCH(1,1) Models," *Econometrica*, 64, 575–596.

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57, 995–1026.

Moon, H. (1997): "Minimum Distance Estimator with Random Weight Matrix," unpublished manuscript, Cowles Foundation for Research in Economics, Yale University.

Moran, P. A. P. (1971): "Maximum-likelihood Estimation in Non-standard Conditions," *Proceedings of the Cambridge Philosophical Society*, 70, 441–450.

Nelson, D. B. (1990): "Stationarity and Persistence in the GARCH(1,1) Model," *Econometric Theory,* 6, 318–334.

Nelson, D. B. and C. Q. Cao (1992): "Inequality Constraints in the Univariate GARCH Model," *Journal of Business and Economic Statistics,* 10, 229–235.

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

Park, J. Y. and P. C. B. Phillips (1988): "Statistical Inference in Regressions with Integrated Processes: Part 1," *Econometric Theory,* 4, 468–497.

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.

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.

Rockafellar, R. T. (1970): *Convex Analysis.* Princeton, NJ: Princeton University Press.

Rothenberg, T. J. (1973): *Efficient Estimation with A Priori Information.* Cowles Foundation Monograph No. 23. New Haven: Yale University Press.

Royden, H. L. (1968): *Real Analysis.* New York: MacMillan.

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.

Stout, W. F. (1974): *Almost Sure Convergence.* New York: Academic Press.

Wolak, F. (1989): "Local and Global Testing of Linear and Nonlinear Inequality Constraints in Nonlinear Econometric Models" *Econometric Theory*, 5, 1–35.

Wu, C. F. J. (1990): "On the Asymptotic Properties of the Jackknife Histogram," *Annals of Statistics,* 18, 1438-1452.