Minimum Distance Estimation of Nonstationary Time Series Models

Hyungsik Roger Moon and Frank Schorfheide

USC Center for Law, Economics & Organization
Research Paper No. C02-2

CENTER FOR LAW, ECONOMICS AND ORGANIZATION RESEARCH PAPER SERIES

Sponsored by the John M. Olin Foundation

University of Southern California Law School
Los Angeles, CA 90089-0071

Minimum Distance Estimation of Nonstationary Time Series Models*

Hyungsik Roger Moon
Department of Economics
University of Southern California

Frank Schorfheide†
Department of Economics
University of Pennsylvania

January 2002

JEL CLASSIFICATION: C32

KEY WORDS: Minimum Distance Estimator, Nonlinear Parameter Restriction, Random Limit Weight Matrix.

*The first version of the paper was written in 1997 by the first author and circulated under the title "A Minimum Distance Estimator with a Random Limit Weight Matrix".

†We thank Bruce Hansen and three anonymous referees for helpful comments. We are also grateful to Peter Phillips, Donald Andrews, Oliver Linton, Peter Robinson, Doug Steigerwald, and seminar participants at Arizona State University, Boston College, UC Santa Barbara, University of Illinois at Urbana-Champaign, USC, the Econometrics Society Summer Meetings in Maryland, and the European Econometric Society Meetings in Lausanne for helpful comments on earlier versions of the paper. The first author acknowledges the financial support from the Academic Research Committee of the University of California at Santa Barbara. The second author acknowledges financial support from the University of Pennsylvania Research Foundation.

Abstract

This paper establishes the consistency and limit distribution of minimum distance (MD) estimators for time series models with deterministic or stochastic trends. We consider models that are linear in the variables, but involve nonlinear restrictions across parameters. Two complications arise. First, the unrestricted and restricted parameter space have to be rotated to separate fast converging components of the MD estimator from slowly converging components. Second, if the model includes stochastic trends it is desirable to use a random matrix to weigh the discrepancy between the unrestricted and restricted parameter estimates. In this case, the objective function of the MD estimator has a stochastic limit. We provide regularity conditions for the non-linear restriction function that are easier to verify than the stochastic equicontinuity conditions that typically arise from direct estimation of the restricted parameters. We derive the optimal weight matrix when the limit distribution of the unrestricted estimator is mixed normal and propose a goodness-of-fit test based on over-identifying restrictions. To illustrate the MD estimation we analyze a permanent-income model based on a linear-quadratic dynamic programming problem and a present-value model.

1 Introduction

This paper considers the limit distribution of minimum distance (MD) estimators for time series models that involve deterministic or stochastic trends. The models are indexed by a $q \times 1$ vector of parameters a. At the pseudo-true value a_0, the parameter vector satisfies the nonlinear restriction $a_0 = g(b_0)$, where b_0 is a lower dimensional $p \times 1$ vector. Such a restriction can, for instance, stem from a optimization-based economic model. We assume that an estimator $\hat{a}_T$ for the unrestricted parameter vector is available. Our main interest is to analyze the limit distribution of MD estimators $\hat{b}_T$ of b_0 and to test whether the restriction $a_0 = g(b_0)$ is satisfied.

In principle, the time series model could be re-parameterized in terms of the parameter vector b, to estimate b_0 directly. However, an attractive alternative is to estimate the unrestricted parameter vector a_0 first and then to minimize a measure of discrepancy between $\hat{a}_T$ and $g(b)$. This procedure is known as minimum distance (MD) estimation. The distance measure used in our paper is $\|W_T(\hat{a_T} - g(b))\|$, where $\{W_T\}$ is a sequence of weight matrices and $\| \cdot \|$ denotes the Euclidean norm. The properly standardized discrepancy between the unrestricted estimate $\hat{a}_T$ and the restriction function evaluated at the MD estimate $g(\hat{b}_T)$ provides a natural goodness-of-fit measure for the restricted specification. We show that the regularity conditions for consistency and weak convergence of the MD estimator can be stated in terms of equicontinuity conditions for the derivatives of the restriction function and provide useful sufficient conditions. Our conditions are easier to verify than the stochastic equicontinuity conditions, e.g. Saikkonen (1995), that commonly arise in non-stationary time series models if b is directly estimated.

In the context of linear regression models without trends, the asymptotic properties of MD estimators of nonlinear restricted parameters are well known, e.g. Chamberlain (1984). The unrestricted estimator $\hat{a}_T$ is $\sqrt{T}$-consistent and has a multivariate normal limit distribution. If the restriction function is smooth, a first-order Taylor expansion of $g(b)$ immediately yields the limit distribution of the MD estimator. The optimal weight matrix is the inverse of the covariance matrix of the unrestricted estimator $\hat{a}_T$.

Two complications arise in the presence of time trends. First, some linear combinations of $\hat{a}_T$ will converge at a faster rate than $\sqrt{T}$. The limit theory of $\hat{a}_T$ is usually based on a rotation that separates the fast converging components from the slow components. The analysis of the minimum distance estimator requires an additional rotation of the restricted parameter vector b. We describe the appropriate rotation and how to determine the order of consistency of the rotated MD estimator.

Second, if the model includes stochastic trends, the limit distribution of $\hat{a}_T$ is non-standard. In the examples studied in this paper the limit distribution of $\hat{a}_T$ is mixed normal with a random covariance matrix. Optimality considerations suggest to use a sequence of weight matrices $\{W_T\}$ that converges in distribution to the inverse of this random covariance matrix. In this case the objective function of the MD estimator does not converge to a non-stochastic limit and the standard consistency argument for extremum estimators, e.g. Amemiya (1985), cannot be employed. Following the methods used in the empirical process literature, e.g. Kim and Pollard (1990) and van der Waart and Wellner (1996), we present an argument based on an almost-sure representation of the objective function.

The paper is organized as follows. Section 2 reviews the existing literature and presents two examples to motivate the MD estimation problem. Nonlinear restriction functions are derived from a permanent-income model that is based on a linear-quadratic dynamic programming problem and a present-value model. The two examples highlight the complications in MD estimation that are addressed in this paper. A general definition of the MD estimator is provided in Section 3 and some fundamental assumptions are stated. Section 4 establishes the consistency of the MD estimator and Section 5 characterizes its limit distribution under various assumptions on the rates of convergence of $\hat{a}_T$ and the smoothness of $g(b)$. In Section 6 we consider the case in which the limit distribution of the unrestricted parameter estimates $\hat{a}_T$ is mixed normal. We define an optimality criterion for the MD estimator and derive the optimal weight matrix. Moreover, a J-type test for the hypothesis $a_0 = g(b_0)$ is provided. Section 7 concludes and the appendix contains mathematical derivations and proofs.

The notation "$\equiv$" is used to signify distributional equivalence, "$\Longrightarrow$" denotes convergence in distribution, "$\xrightarrow{p}$" denotes convergence in probability, "$\xrightarrow{a.s.}$" is almost-sure convergence, and "$\otimes$" is the Kronecker-product. We will use $\int B$ and $\int BB'$ to abbreviate integrals of vector Brownian motion $B(r)$ sample paths $\int B(r)dr$ and $\int B(r)B(r)'dr$, respectively.

2 Examples

Economic theory often implies nonlinear restrictions across parameters of econometric models. Many structural economic models involve solutions to stochastic dynamic programming problems, which arise from the intertemporal maximization of, for instance, households' utility, firms' profits, or social welfare. We consider an optimization problem with a quadratic objective function and linear state transition equations. The optimal decision rule generates non-linear restrictions across the parameters that govern the joint evolution of state and control variables. In macroeconomics, linear-quadratic programming problems are frequently used to approximate smooth non-linear model economies (see Ljungqvist and Sargent (2000)).

Example 1: A consumer chooses consumption $\{C_t\}_{t=0}^{\infty}$ to maximize the expected utility

$$-\frac{1}{2}I\!E_0\left[\sum_{t=0}^{\infty}\left(\frac{1}{1+r}\right)^t (C_t-\epsilon_{1,t}-\alpha)^2\right] \tag{1}$$

subject to the constraints

$$\begin{aligned} I_{t+1} &= \mu(1-\phi)+\phi I_t+\epsilon_{2,t+1} \\ W_{t+1} &= (1+r)(W_t-C_t)+I_{t+1}+\epsilon_{3,t+1}, \end{aligned} \tag{2}$$

where W_t is wealth, I_t is a stochastic income process, and r is the real interest rate. The exogenous shocks are collected in the vector $\epsilon_t=[\epsilon_{1,t},\epsilon_{2,t},\epsilon_{3,t}]'$. We will assume that $\epsilon_t \sim iid(0,\Sigma_{\epsilon\epsilon})$. The shock $\epsilon_{1,t}$ can be interpreted as taste shock that shifts the utility of time t consumption, and $\epsilon_{2,t}$ is the innovation of the observed income process I_t. It is assumed that there exists a stochastic income component $\epsilon_{3,t}$ that is unobserved by the econometrician. In other words, we do not require measured wealth, consumption, and income to satisfy the wealth accumulation constraint in Equation (2) exactly.

Under the optimal consumption rule the law of motion for consumption, wealth, and income is

$$\begin{bmatrix} C_t \\ \Delta W_t \\ I_{1,t} \end{bmatrix} = \begin{bmatrix} \frac{r}{1+r} & \frac{\mu}{1+r} & \frac{\phi r(2-\phi)}{(1+r-\phi)(1+r)} \\ 0 & 0 & \frac{\phi(1-\phi)}{1+r-\phi} \\ 0 & \mu_1 & \phi \end{bmatrix} \begin{bmatrix} W_{t-1} \\ 1 \\ (I_{1,t-1}-\mu_1) \end{bmatrix} + \begin{bmatrix} \frac{r}{1+r} & \frac{r}{1+r} & \frac{r}{1+r} \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \epsilon_{3,t} \end{bmatrix} + \begin{bmatrix} -\frac{r^2}{1+r} & 0 & 0 \\ -r & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \epsilon_{1,t-1} \\ \epsilon_{2,t-1} \\ \epsilon_{3,t-1} \end{bmatrix}. \tag{3}$$

The first row of (3) corresponds to the decision rule for C_t as a function of the state variables. We will explore the estimation of this system for a stationary ($0 \leq \phi < 1$) and non-stationary income ($\phi = 1$) process. While our version of the permanent-income model is more stylized than specifications that are estimated in practice, it can be solved analytically and allows us to illustrate the MD estimation problem.[1]

□

2.1 Restricted Cointegration Relationships

Suppose $\phi = 1$ and $\mu = 0$ in Example 1. Define $y_{1,t} = C_t$ and $y_{2,t} = [W_t, I_t]$. According to the permanent-income model all three variables are integrated of order one ($I(1)$). In this case the system (3) has the form of a cointegration regression model

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} A' \\ \mathcal{I}_2 \end{bmatrix} y_{2,t-1} + u_t, \tag{4}$$

where $\mathcal{I}_j$ denotes the $j \times j$ identity matrix and u_t is a moving-average (MA) process of ϵ_t and ϵ_{t-1}. Let $a = vec(A)$ and $b = r$. The restriction imposed by the optimal

[1] Detailed calculations for Example 1 are available from the authors upon requests.

decision rule on the cointegration vector A is

$$g(b) = \left[\frac{b}{1+b}, \frac{1}{1+b}\right]'. \tag{5}$$

The distribution theory for estimators of the unrestricted cointegration vector A in Equation (4) is well developed. The estimators typically converge a rate T^{-1}. Both the maximum likelihood estimator of a_0 (Phillips (1991)) and the fully modified least squares estimator (Phillips and Hansen (1990)) have a mixed normal (MN) distribution with random covariance matrix.[2]

Several results concerning the estimation of the restricted cointegration vector have been published. Phillips (1991) developed a theory of optimal inference for cointegration regressions based on the likelihood function for $[y'_{1,t}, y'_{2,t}]'$. He derives the limit distribution of $\hat{b}_T$ for linear restriction functions. Moreover, Saikkonen (1993) studied a general approach for the estimation of cointegration vectors with linear restrictions.

Saikkonen (1995) extended the analysis of the maximum likelihood estimator to the case in which the restriction function is nonlinear and twice differentiable. He provided stochastic equicontinuity conditions to make the conventional Taylor approximation approach valid. Unfortunately, it is in general difficult to verify these conditions (Saikkonen, 1995, page 893). One advantage of the MD approach is that it leads to conditions that only involve the (deterministic) restriction function $g(b)$. Phillips (1993) also investigated the MLE estimation of a cointegration model in which nonlinear restrictions are imposed on the cointegration parameters.

Nagaraj and Fuller (1991) studied a univariate nonstationary autoregressive time series regression model with restrictions across parameters that are estimated at different rates. The restrictions are given by an implicit function. Nagaraj and Fuller showed that the constrained nonlinear least squares estimator is consistent and derived its limit distribution under a stochastic equicontinuity condition for the restriction function.

[2]See also Park (1992), Saikkonen (1991), and Stock and Watson (1993).

2.2 Restrictions between Short-run and Long-run Dynamics

Even if the income process is stationary ($0 \leq \phi < 1,\ \mu > 0$) both consumption and wealth are $I(1)$ processes under the optimal consumption choice (see, for instance, Hall (1978)). In this case the optimal decision rule creates restrictions between parameters that are associated with long-run relationships and parameters that control the short-run dynamics. Define $y_{1,t} = C_t$, $y_{2,t} = [\Delta W_t, I_t]'$, $x_{1,t} = W_{t-1}$, $x_{2,t} = [1, I_{t-1}]'$, $y_t = [y_{1,t}, y'_{2,t}]'$, and $x_t = [x_{1,t}, x'_{2,t}]'$. The regressor $x_{1,t}$ is $I(1)$, whereas $x_{2,t}$ is stationary. The consumption model is nested in the following general specification

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} A'_{11} & A'_{21} \\ 0 & A'_{22} \end{bmatrix} \begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}, \tag{6}$$

where u_t is an MA(1) process. Define $a_{ij} = vec(A_{ij})$. The unrestricted parameter vector is $a = [a'_{11}, a'_{21}, a'_{22}]'$ and $b = [r, \mu, \phi]'$ is composed of the structural parameters. The restriction function $g(b)$ is

$$\begin{aligned} g(b) \quad = \quad & \Bigg[\frac{r}{1+r}, \frac{\mu}{1+r} - \frac{\mu\phi r(2-\phi)}{(1+r)(1+r-\phi)}, \frac{\phi r(2-\phi)}{(1+r-\phi)(1+r)}, \\ & -\frac{\mu\phi(1-\phi)}{1+r-\phi}, \frac{\phi(1-\phi)}{1+r-\phi}, \mu(1-\phi), \phi\Bigg]'. \end{aligned} \tag{7}$$

Since the errors $u_{1,t}$ and $u_{2,t}$ are uncorrelated with the stationary regressors $x_{2,t}$, the system (6) can be estimated by quasi-maximum likelihood ignoring the MA(1) structure of u_t. The limit distribution of the estimator $\hat{a}_T$ is of the form

$$D_T R(\hat{a}_T - a_0) \Longrightarrow \eta^{1/2} Z, \tag{8}$$

where

$$D_T = \begin{bmatrix} T & 0 & 0 \\ 0 & T^{-1/2}\mathcal{I}_2 & 0 \\ 0 & 0 & T^{-1/2}\mathcal{I}_4 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & R_* & 0 \\ 0 & 0 & I_2 \otimes R_* \end{bmatrix}, \quad R_* = \begin{bmatrix} 1 & \mu \\ 0 & 1 \end{bmatrix},$$

and $Z \equiv \mathcal{N}(0, \mathcal{I}_7)$. The matrix R rotates the unrestricted parameters and its inverse removes the mean μ from the regressor I_{t-1}. The diagonal matrix D_T contains the rates of convergence of the rotated parameters. Assume that the partial sum process of ΔW_t converges to a vector Brownian motion: $T^{-1/2}\sum_{t=1}^{[Tr]} \Delta W_t \Longrightarrow B(r) \equiv$

$BM(\Omega)$. Ω is the long-run variance of ΔW_t defined as $\lim_{T\to\infty} \frac{1}{T}\mathbb{E}[(\sum_{t=1}^{T} \Delta W_t)(\sum_{t=1}^{T} \Delta W_t)']$. The random covariance matrix η is given by

$$\eta = \begin{bmatrix} (\Sigma_{11.2} \otimes Q_{11.2}^{-1}) & -(\Sigma_{11.2} \otimes Q_{11.2}^{-1} Q_{12} Q_{22}^{-1}) & 0 \\ -(\Sigma_{11.2} \otimes Q_{22}^{-1} Q_{21} Q_{11.2}^{-1}) & (\Sigma_{11.2} \otimes Q_{22.1}^{-1}) + (\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \otimes Q_{22}^{-1}) & (\Sigma_{12} \otimes Q_{22}^{-1}) \\ 0 & (\Sigma_{21} \otimes Q_{22}^{-1}) & (\Sigma_{22} \otimes Q_{22}^{-1}) \end{bmatrix}. \tag{9}$$

Let Σ be the long-run variance of u_t. The $\Sigma'_{ij}s$ are the partitions of Σ that conform with the partition of u_t. The various elements of η are defined as follows: $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, $Q_{11} = \int_0^1 B^2$, $Q_{12} = [\int_0^1 B, 0]$, $Q_{21} = Q'_{12}$, $Q_{22} = \mathbb{E}[R_*^{-1'} x_{2,t} x'_{2,t} R_*^{-1}]$, $Q_{11.2} = Q_{11} - Q_{12}Q_{22}^{-1}Q_{12}$, and $Q_{22.1} = Q_{22} - Q_{21}Q_{11}^{-1}Q_{12}$.

Define the rotated unrestricted parameters $\tilde{a} = Ra$ and let $\tilde{a} = [\tilde{a}_1, \tilde{a}_2]'$, where $\tilde{a}_1$ corresponds to the cointegration parameter a_{11}. The permanent-income model has the following features. First, the function $\tilde{g}(b) = Rg(b)$ creates restrictions across $\tilde{a}_1$ and $\tilde{a}_2$, which are estimated at the rates T^{-1} and $T^{-1/2}$, respectively. Second, the fast parameter estimate $\hat{\tilde{a}}_{1,T}$ is asymptotically correlated with the slow parameter estimates $\hat{\tilde{a}}_{2,T}$ due to the intercept $\mu/(1+r)$ in the cointegration relationship of C_t and W_t. Third, the rotated restriction function is block-diagonal. The vector b can be expressed as $b = [b_1, b'_2]'$ and the restriction function is of the form $\tilde{g}(b) = [\tilde{g}_1(b_1)', \tilde{g}_2(b_1, b_2)']'$.

There are many other economic models that generate restrictions between parameters that control short-run dynamics and long-run relationships. A well-known example is the present-value model, which has been widely studied in the empirical finance literature, for instance by Campbell and Shiller (1987).

Example 2: Let $y_{1,t}$ be a stock price and $y_{2,t}$ a dividend payment. A risk-neutral investor is indifferent between the stock and a bond that guarantees to pay the interest rate r if

$$y_{1,t} = \frac{1}{1+r}\mathbb{E}_t[y_{1,t+1} + y_{2,t+1}]. \tag{10}$$

Campbell and Shiller essentially modelled the joint behavior of $y_{1,t}$ and $y_{2,t}$ as

$$\begin{bmatrix} y_{1,t} - a_1 y_{2,t-1} \\ \Delta y_{2,t} \end{bmatrix} = (\mathcal{I}_2 - A'_2 L)^{-1}\epsilon_t, \quad \epsilon_t \sim iid(0, \Sigma_{\epsilon\epsilon}), \tag{11}$$

where L denotes the lag operator. If the cointegration parameter $a_1 = 1/r$ then the linear combination $y_{1,t} - \frac{1}{r}y_{2,t}$ is called the spread and reflects the difference between the stock price and the present discounted value of future dividends under the assumption that $y_{2,t+j} = y_{2,t}$ for all j.

The model assumes that spread and dividend growth follow a stationary VAR(1) process. In addition to $a_1 = 1/r$ Equation (10) imposes restrictions on the matrix of short-run dynamics A_2. Let $a_2 = vec(A_2)$, $a = [a_1, a_2']'$ and define the parameter vector $b = [r, b_1, b_2]'$. The restriction function for this model is

$$g(b) = \left[\frac{1}{r},\, b_2,\, b_3,\, 1 + r - b_2,\, -b_3 - \frac{1+r}{r}\right]. \tag{12}$$

The unrestricted parameter vector a can be estimated, for instance, by maximum likelihood. The ML estimator has the property that a_1 is estimated at rate T^{-1} and a_2 at rate $T^{-1/2}$. The restriction function is block-diagonal. However, unlike in Example 1 the estimators $\hat{a}_{1,T}$ and $\hat{a}_{2,T}$ are asymptotically uncorrelated since there is no intercept in the cointegration relationship between stock prices and dividends. $\square$

After the first draft of this paper was written we learnt that Elliot (2000) analyzed an MD estimator for b in the context of a cointegration regression model (4). He applied the MD estimator to the six variable cointegration model of King *et al.* (1991) in which the cointegration coefficients have to satisfy linear exclusion restrictions. He does not provide examples of nonlinear restrictions across cointegration parameters. While his asymptotic theory suffices to analyze the permanent-income model with an I(1) income process, it is not general enough to be applied to the important class of problems in which $g(b)$ imposes restrictions across parameters that are estimated at different rates, such as in the permanent-income model with stationary income process and the present-value model. Thus, our paper provides a much more complete theory of MD estimators for non-stationary time series models, which encompasses Elliot's results as a special case.

3 Minimum Distance Estimation

This section provides a rigorous definition of the MD estimator and introduces regularity conditions that are used throughout the paper. Let $Y_T(\omega)$ be a data matrix of sample size T defined on the probability space $(\Omega, \mathcal{F}, I\!P)$. Moreover, let $\{\hat{a}_T\}$ be a sequence of estimators of a q-dimensional parameter vector $a \in \mathcal{A} \subset I\!R^q$. It is assumed that there exists a unique pseudo-true value $a_0 \in \mathcal{A}$ which is consistently estimated by $\{\hat{a}_T\}$, that is, $\hat{a}_T \xrightarrow{p} a_0$ under $I\!P$. Let $b \in \mathcal{B} \subset I\!R^p$ be a second parameter vector and $g(\cdot)$ be a mapping from $\mathcal{B}$ to $\mathcal{A}$. We assume that $p \leq q$ and $\{a \in \mathcal{A} : a = g(b), b \in \mathcal{B}\} \subset \mathcal{A}$. Thus, $g(b)$ can be interpreted as a restriction on the parameter space $\mathcal{A}$. We will make the following assumptions with respect to $g(b)$ and $\mathcal{B}$.

Assumption 1 *(Parameter Restriction (I))*

(i) The parameter space $\mathcal{B}$ is compact.

(ii) The restriction function $g(b)$ is continuous.

(iii) There is a unique b_0 in the interior of $\mathcal{B}$ such that $g(b_0) = a_0$.

Suppose $\{W_T\}$ is a sequence of $q \times q$ weight matrices. Any vector $\hat{b}_T$ that minimizes the criterion function

$$Q_T(b) = \frac{1}{2}\|\tilde{W}_T(\hat{a}_T - g(b))\|, \tag{13}$$

will be called a minimum distance (MD) estimator of b. $\|\cdot\|$ denotes the Euclidean norm in $I\!R^q$. The estimator $\hat{b}_T$ exists because $Q_T(b)$ is continuous in b on the compact set $\mathcal{B}$ by Assumption 1. Moreover, $\hat{b}_T$ is measurable due to Lemma 2 of Jennrich (1969). We will now impose conditions on the joint asymptotic behavior of the unrestricted estimator $\hat{\alpha}_T$ and the weight matrix $\tilde{W}_T$.[3]

[3]Our main interest is the limit distribution of the MD estimator. To establish its consistency, it is not necessary to make such detailed assumptions about the limit distribution of $\hat{a}_T$.

Assumption 2 *(Unrestricted Parameter Estimation and Weight Matrix)*

(i) $\tilde{W}_T$ *and* a_T *are defined on the same probability space.*

(ii) There exists a non-stochastic and invertible matrix R *and a non-stochastic diagonal matrix* D_T *whose elements tend to infinity as* $T \longrightarrow \infty$ *such that*

$$\begin{bmatrix} D_T R(\hat{a}_T - a_0) \\ vec(\tilde{W}_T R^{-1} D_T^{-1}) \end{bmatrix} \Longrightarrow \begin{bmatrix} \alpha \\ vec(W) \end{bmatrix}$$

where α *is a* $q \times 1$ *random vector and* W *a* $q \times q$ *random matrix that is non-singular with probability 1.*

Assumption 2 implies that the limit distribution of the unrestricted estimator $\hat{a}_T$ is equivalent to the distribution of the random variable α. The matrix R rotates the unrestricted parameter space and separates directions in which convergence of $\hat{a}_T$ is fast from directions in which convergence is slow. The diagonal elements of D_T correspond to the convergence rates for these different directions. More weight should be assigned to the rotated elements of $\hat{a}_T$ that provide the most precise measurements of the corresponding a_0 elements. Therefore, it is assumed that the weight matrix $\tilde{W}_T$ is designed to grow more rapidly along the directions in which the convergence of $\hat{a}_T$ to a_0 is fast. For the remainder of the paper we define the standardized weight matrix $W_T = \tilde{W}_T R^{-1} D_T^{-1}$. Assumption 2 allows the limit of the sequence $\{W_T\}$ to be random.

4 Consistency

The minimum distance estimator is consistent provided that the unrestricted estimator $\hat{a}_T$ is consistent and b_0 is uniquely identifiable based on the restriction $b_0 = g(a_0)$. The result is formally stated in the following theorem and proved in the Appendix.

Theorem 1 *(Consistency of MD Estimator)*

If Assumptions 1 and 2 are satisfied, then $\hat{b}_T \xrightarrow{p} b_0$ *as* $T \longrightarrow \infty$.

The traditional proof of the consistency of extremum estimators is based on the uniform convergence of the random sample objective function to a non-random limit function coupled with some identification condition for the "true" parameter values, e.g. Amemiya (1985). However, this traditional method is not applicable to our MD estimator for two reasons. First, since the weight matrix converges in distribution to a random matrix, the objective function $Q_T(b)$ converges for each value of b to a random variable. Second, in time series models with trends, the convergence rate of $Q_T(b)$ will generally depend on b.

The key idea to overcome the first difficulty is to work with an almost-sure representation of the probability distributions of $\hat{a}_T$ and $\tilde{W}_T$. This idea has been used in the empirical process literature to establish limit distributions for extremum estimators. Kim and Pollard (1990), for instance, employ Dudley's almost-sure representation. We are using the Skorohod representation in this paper, see for instance Billingsley (1986). To cope with the different convergence rates, Lemma 1 of Wu (1981) is employed (see Appendix).

5 Limit Distribution

Without loss of generality it is assumed that the diagonal elements of the matrix D_T are equal to T^{ν_j}, $j = 1, \ldots, q$, where $\nu_j \geq \nu_{j+1} > 0$. This section will develop the limit distribution of $\hat{b}_T$ under various assumptions on the restriction function $g(b)$. Define $\alpha_T = D_T R(\hat{a}_T - a_0)$ and $W_T = \tilde{W}_T R^{-1} D_T^{-1}$. The objective function of the MD estimator can be rewritten as

$$\begin{aligned} Q_T(b) &= Q_T(b_0) - \alpha_T' W_T' W_T D_T R(g(b) - a_0) \\ &\quad + \frac{1}{2}(g(b) - a_0)' R' D_T' W_T' W_T D_T R(g(b) - a_0) \end{aligned} \tag{14}$$

We will begin with a linear restriction function and then consider Taylor approximations to non-linear restriction functions subsequently. Throughout this section we will state additional assumptions on the restriction function $g(b)$ and the domain $\mathcal{B}$ of b. We use $g^{(1)}(b_*)$ to denote the $q \times p$ matrix $\frac{\partial g}{\partial b'}|_{b=b_*}$ of first derivatives.

5.1 Linear Restriction Functions

Suppose that $g(b) = Gb$, where G is a $q \times p$ matrix. In this case the objective function is quadratic in b. Thus, $Q_T(b) = Q_{q,T}(b)$, where

$$\begin{aligned} Q_{q,T}(b) &= Q_{q,T}(b_0) - \alpha_T' W_T' W_T D_T R G(b - b_0) \\ &+ \frac{1}{2}(b - b_0)' G' R' D_T' W_T' W_T D_T R G(b - b_0). \end{aligned} \tag{15}$$

To analyze the limit distribution of the MD estimator, the restricted parameter space has to be rotated. Define $\tilde{g}(b) = Rg(b)$, $\tilde{G} = RG$, and the $p \times p$ matrix G^* consisting of the first p linearly independent rows of $\tilde{G}$. Moreover define the function $\iota(j)$ such that the j'th row of G^* equals to the $\iota(j)$'s row of $\tilde{G}$. Decompose $G^* = L_* U_*$, where L_* is lower triangular and U_* is upper triangular. Let

$$\Lambda_T = \text{diag}[T^{\nu_{\iota(1)}}, \ldots, T^{\nu_{\iota(p)}}] U_*, \quad \text{and} \quad \Gamma_T = D_T \tilde{G} \Lambda_T^{-1},$$

where $T^{\nu_{\iota(j)}}$ is the convergence rate that corresponds to the j'th row of G_*. Define the local parameter vector $s = \Lambda_T(b - b_0)$ with domain $\mathcal{S} = \Lambda_T(\mathcal{B} - b_0)$. The role upper triangle matrix U_* in Λ_T is to rotate the restricted parameter b. The diagonal matrix $\text{diag}[T^{\nu_{\iota(1)}}, \ldots, T^{\nu_{\iota(p)}}]$ in Λ_T controls the convergence rates of the rotated parameter. The sample objective function of the MD estimator in terms of s is

$$Q_{q,T}(b_0 + \Lambda_T^{-1} s) = Q_{q,T}(b_0) - \alpha_T' W_T' W_T \Gamma_T s + \frac{1}{2} s' \Gamma_T' W_T' W_T \Gamma_T s. \tag{16}$$

Example 3: Let $R = \mathcal{I}$. Thus, $\tilde{G} = G$. A_{ij} denotes element ij of a matrix A. $A_{i.}$ and $A_{.j}$ denote its i'th row and j'th column, respectively. Suppose $D_T = diag[T^{3/2}, T, T, T^{1/2}, T^{1/2}]$ and $G_{3.} = \lambda_1 G_{1.} + \lambda_2 G_{2.}$ for some scalars λ_1, λ_2. In this case $G^* = [G_{1.}', G_{2.}', G_{4.}']'$. The function $\iota(j)$ takes the values $\iota(1) = 1$, $\iota(2) = 2$, $\iota(3) = 4$. Therefore, $\nu_{\iota(1)} = 3/2$, $\nu_{\iota(2)} = 1$, $\nu_{\iota(3)} = 1/2$. Since L_* and U_* are defined through and LU-decomposition of G^*, it follows that $G_{i.}^*[U_*^{-1}]_{.j} = 0$ for $j > i$. Since $G_{3.}$ is a linear combination of $G_{1.}^*$ and $G_{2.}^*$ we can deduce

$$G_{3.}[U_*^{-1}]_{.3} = \lambda_1 G_{1.}^*[U_*^{-1}]_{.3} + \lambda_2 G_{2.}^*[U_*^{-1}]_{.3} = 0$$

Therefore,

$$\Gamma_T = \begin{bmatrix} G^*_{1.}[U_*^{-1}]_{.1}T^{3/2-3/2} & 0 & 0 \\ G^*_{2.}[U_*^{-1}]_{.1}T^{1-3/2} & G^*_{2.}[U_*^{-1}]_{.2}T^{1-1} & 0 \\ G^*_{3.}[U_*^{-1}]_{.1}T^{1-3/2} & G^*_{3.}[U_*^{-1}]_{.2}T^{1-1} & 0 \\ G^*_{4.}[U_*^{-1}]_{.1}T^{1/2-3/2} & G^*_{4.}[U_*^{-1}]_{.2}T^{1/2-1} & G^*_{4.}[U_*^{-1}]_{.3}T^{1/2-1/2} \\ G^*_{5.}[U_*^{-1}]_{.1}T^{1/2-3/2} & G^*_{5.}[U_*^{-1}]_{.2}T^{1/2-1} & G^*_{5.}[U_*^{-1}]_{.3}T^{1/2-1/2} \end{bmatrix}$$

$$\longrightarrow \begin{bmatrix} G^*_{1.}[U_*^{-1}]_{.1} & 0 & 0 \\ 0 & G^*_{2.}[U_*^{-1}]_{.2} & 0 \\ 0 & G^*_{3.}[U_*^{-1}]_{.2} & 0 \\ 0 & 0 & G^*_{4.}[U_*^{-1}]_{.3} \\ 0 & 0 & G^*_{5.}[U_*^{-1}]_{.3} \end{bmatrix} \quad \square$$

The asymptotic behavior of Γ_T is summarized in the following Lemma.

Lemma 1 *Suppose that $g(b) = Gb$. $\lim_{T\to\infty} \Gamma_T = \Gamma$, where Γ has full row rank.*

Since it is assumed that the "true" parameter is in the interior of $\mathcal{B}$, the parameter space $\mathcal{S}$ of the local parameter s expands to $I\!R^p$ as $T \longrightarrow \infty$. Define

$$\hat{s}_{q,T} = (\Gamma_T' W_T' W_T \Gamma_T)^{-1} \Gamma_T' W_T' W_T \alpha_T \tag{17}$$

as the minimizer of the quadratic sample objective function $Q_{q,T}(b_0 + \Lambda_T^{-1} s)$ over $I\!R^p$.

Theorem 2 *Suppose Assumptions 1 and 2 are satisfied and the restriction is of the form $g(b) = Gb$. Then $\hat{s}_{q,T} \Longrightarrow (\Gamma' W' W \Gamma)^{-1} \Gamma' W' W \alpha$ and $\Lambda_T(\hat{b}_T - b_0) = \hat{s}_{q,T} + o_p(1)$.*

Lemma 1 can be used to derive the limit distribution of $\hat{s}_{q,T}$. The $o_p(1)$ term arises because for small sample sizes the objective function might attain its minimum on the boundary of the parameter space and not satisfy the first-order conditions.

Remark: If the convergence rates in Λ_T are different and Λ_T is not a diagonal matrix, the convergence rate of the unrotated restricted parameter estimator is determined by a slower convergence rate and its limit may have a degenerated asymptotic covariance matrix. For example, suppose $p = 2$ and $\Lambda_T = diag\left(T, \sqrt{T}\right) U_*$. If $U_{*12} \neq 0$, we deduce from Theorem 2 that

$$\left[\begin{array}{c} \sqrt{T}\left(\hat{b}_{1,T} - b_{1,0}\right) \\ \sqrt{T}\left(\hat{b}_{2,T} - b_{2,0}\right) \end{array}\right] = \left[\begin{array}{c} -\frac{U_{*12}}{U_{*11}U_{*22}}\hat{s}_{2,qT} \\ \frac{1}{U_{*22}}\hat{s}_{2,qT} \end{array}\right] + o_p(1).$$

However, if $U_{*12} = 0$, then

$$\left[\begin{array}{c} T\left(\hat{b}_{1,T} - b_{1,0}\right) \\ \sqrt{T}\left(\hat{b}_{2,T} - b_{2,0}\right) \end{array}\right] = \left[\begin{array}{c} \frac{1}{U_{*11}}\hat{s}_{1,qT} \\ \frac{1}{U_{*22}}\hat{s}_{2,qT} \end{array}\right] + o_p(1)$$

and the limit distribution of the unrotated parameter vector is non-degenerated. □

5.2 Block-triangular Restriction Matrices

The restriction functions in the examples of Section 2 are block-triangular. In this subsection we study the linear case. Suppose that $R = \mathcal{I}$ and the unrestricted parameter vector a can be partitioned as follows: $a = [a_1', a_2']'$. The subvector a_1 consists of long-run parameters that can be estimated at a fast rate, say $\nu_1 = 1$. Assume that a_2 consists of short-run parameters that are estimated at a slower rate T^{ν_2}, e.g. $\nu_2 = 1/2$. We refer to the restrictions as block-triangular if the matrix G and the restricted parameter b can be partitioned as follows:

$$\left[\begin{array}{c} a_1 \\ a_2 \end{array}\right] = \left[\begin{array}{cc} G_{11} & 0 \\ G_{21} & G_{22} \end{array}\right] \left[\begin{array}{c} b_1 \\ b_2 \end{array}\right],$$

where G_{11} and G_{22} have full row rank. The b_i's are $p_i \times 1$ subvectors of b. The setup implies that the subvector b_2 does not restrict the long-run parameters a_1. The rank condition guarantees, that it is possible to solve for b_1 based on a_1, and for b_2 based on a_2 conditional on a_1. This case is also discussed in Phillips (1991, Remark (m)). Before examining the consequences of the block-triangular structure we provide a general definition.

Definition 1 *(Block-triangular Restrictions) Let $\tilde{a} = Ra$. Partition the rotated unrestricted parameter vector $\tilde{a} = [\tilde{a}_1', \dots, \tilde{a}_L']'$ such that the estimators of the elements of $\tilde{a}_l$ converge all at the same rate T^{ν_l}, $l = 1, \dots, L$. Let $\tilde{g}_l(b)$ be the rotated restriction function corresponding to $\tilde{a}_l$ and G_{lk} be the submatrix of G that conforms with the partitions of $\tilde{a}$ and b. The restriction function is block-triangular if it is possible to rearrange the elements of b such that the restricted parameter vector can be partitioned into $k = 1, \dots, K \leq L$ subvectors b_k and (i) $\tilde{g}_l(b) = f_l([b_1', \dots, b_l']')$, $l = 1, \dots, K$, and (ii) G_{ll} has full row rank, $l = 1, \dots, K$.*

Due to the block-triangular structure, the matrices G^* and U_*, defined above, are of the form

$$G^* = \begin{bmatrix} G_{11}^* & 0 \\ G_{21}^* & G_{22}^* \end{bmatrix}, \quad U_* = \begin{bmatrix} U_{11}^* & 0 \\ 0 & U_{22}^* \end{bmatrix},$$

where G_{11}^* consists of the first p_1 linearly independent rows of G_{11} and U_* corresponds to the upper-triangular matrix of the LU-decomposition of G_*. The matrix Γ_T converges to

$$\Gamma_T = \begin{bmatrix} T^{\nu_1} G_{11} & 0 \\ T^{\nu_2} G_{21} & T^{\nu_2} G_{22} \end{bmatrix} \begin{bmatrix} T^{-\nu_1} U_{11}^{*-1} & 0 \\ 0 & T^{-\nu_2} U_{22}^{*-1} \end{bmatrix} \longrightarrow \begin{bmatrix} \Gamma_{11} & 0 \\ 0 & \Gamma_{22} \end{bmatrix},$$

where $\Gamma_{ii} = G_{ii} U_{ii}^{*-1}$. Define the selection matrices $M_1' = [I_{p_1}, 0_{p_1 \times p_2}]$, and $M_2' = [0_{p_2 \times p_1}, I_{p_2}]$. The limit distribution of the subvectors $\hat{b}_{i,T}$ is given by

$$T^{\nu_i}(\hat{b}_{i,T} - b_i) \Longrightarrow U_{ii}^{*-1} M_i' (\Gamma' W' W \Gamma)^{-1} \Gamma' W' W \alpha, \quad i = 1, 2. \tag{18}$$

Thus, the parameters b_1 that enter the long-run parameters a_1 can be estimated at the fast rate T^{ν_1}, whereas b_2 is estimated at the slower rate T^{ν_2}. In general, the limit distribution for both subvectors depends on the entire vector α. Thus, even the restrictions embodied in the short-run parameters a_2 are informative with respect to b_1.

If the limit weight matrix W is with probability one of the form

$$W = \begin{bmatrix} W_{11} & 0 \\ 0 & W_{22} \end{bmatrix},$$

where the partitions of W correspond to the partitions of a, then the limit distribution simplifies considerably

$$T^{\nu_i}(\hat{b}_{i,T} - b_i) \Longrightarrow (G_{ii}' W_{ii}' W_{ii} G_{ii})^{-1} G_{ii} W_{ii}' W_{ii} \alpha_i, \quad i = 1, 2. \tag{19}$$

The simplification arises because the system is now block-diagonal and $U_{ii}^* \Gamma_{ii} = G_{ii}$. Thus, if the weight matrix cross-terms between short-run and long-run parameters are zero, the distribution of the long-run parameter estimates $\hat{a}_{1,T}$ does not affect the limit of $\hat{b}_{2,T}$, and vice versa, the limit distribution of $\hat{b}_{1,T}$ is solely determined through the distribution of $\hat{\alpha}_{1,T}$. Now consider the following two step estimator $\tilde{b}_T$ of the restricted parameter vector b.

Two-step Estimation Procedure:

(i) Estimate b_1 according to

$$\tilde{b}_{1,T} = \mathrm{argmin}_{b_1 \in \mathcal{B}_1} \frac{1}{2} \|\tilde{W}_{11,T}(\hat{a}_{1,T} - G_{11} b_1)\|,$$

where $T^{-\nu_1} \tilde{W}_{11,T} \Longrightarrow W_{11}$.

(ii) Estimate b_2 according to

$$\tilde{b}_{2,T} = \mathrm{argmin}_{b_2 \in \mathcal{B}_2} \frac{1}{2} \|\tilde{W}_{22,T}(\hat{a}_{1,T} - G_{21} \hat{b}_{1,T} - G_{22} b_2)\|,$$

where $T^{-\nu_2} \tilde{W}_{22,T} \Longrightarrow W_{22}$. $\square$

In the second step, b_1 is replaced by its first-step estimate. Conventional arguments imply that the estimation uncertainty of $\hat{b}_{1,T}$ does not affect the limit distribution of $\hat{b}_{2,T}$ because $\hat{b}_{1,T}$ is estimated at a faster rate. Thus, the limit distribution of the two-step estimator is equivalent to the limit of the minimum distance estimator with block-diagonal weight matrix, given in Equation (19). The arguments provided in this subsection easily extend to the case of more than two different convergence rates ($L > 2$).

5.3 Nonlinear Restriction Functions

To derive the limit distribution of the MD estimator for nonlinear restriction functions we will impose smoothness conditions and use a Taylor series approximation of the form

$$g(b) = g(b_0) + G(b - b_0) + \Phi[b_+, b_0](b - b_0), \tag{20}$$

where $G = g^{(1)}(b_0)$, $\Phi[b_+, b_0] = g^{(1)}(b_+) - g^{(1)}(b_0)$, and b_+ is located between b_0 and b. Define $\tilde{G}$, G_*, L_*, U_*, Λ_T, and Γ_T as in Section 5.1. The term $D_T R(g(b) - a_0)$ that appears in the objective function of the MD estimator (Equation 14) is approximated by

$$D_T R(g(b) - a_0) = \Gamma_T s + D_T R \Phi[b_+, b_0] \Lambda_T^{-1} s, \tag{21}$$

where s is the local parameter $s = \Lambda_T(b - b_0)$. In order to be able to bound the remainder term $\Phi[\cdot, \cdot]$, we impose some smoothness conditions on $g(b)$.

Assumption 3 *(Parameter Restriction (II))*

(i) The parameter restriction function $g(b)$ is differentiable in a neighborhood of b_0.

(ii) For any sequence $\delta_T \to 0$, $\sup_{\|b_1 - b_0\| \le \delta_T} \|D_T R \Phi[b_1, b_0] \Lambda_T^{-1}\| = o(1)$.

Assumption 3(ii) is an equicontinuity condition for the first derivative of the restriction function that allows us to use the conventional quadratic approximation of the objective function. An advantage of the MD approach is that one only has to verify a deterministic condition. Maximum likelihood estimators, such as the one discussed by Saikkonen (1995) for the restricted cointegrated regression model, and the constrained least squares estimator proposed by Nagaraj and Fuller (1991, see Assumption 4) require the verification of stochastic equicontinuity conditions.

The following assumption provides a sufficient condition for Assumption 3. A proof can be found in the Appendix.

Assumption 3* *(Parameter Restriction (Sufficient Condition)) Suppose that $\frac{\partial g(b)}{\partial b}$ is continuous in a neighborhood of the true parameter b_0. Moreover, at least one of the following conditions is satisfied:*

(i) For each pair i, j such that $\nu_i > \nu_{\iota(j)}$, there exists an $\epsilon > 0$ such that the ij'th element of the matrix $R\Phi[b_1, b_0]U_*^{-1}$ is equal to zero for $\|b_1 - b_0\| < \epsilon$.

(ii) For each i such that $\nu_i > \nu_{\iota(p)}$ the i'th component $\tilde{g}_i(b) = \sum_{j=1}^{q} R_{ij} g_j(b)$ of the rotated restriction function is linear in b.

(iii) The rotated restriction function $Rg(b)$ is block-triangular, as in Definition 1.

Remark: Conditions (ii) and (iii) are special cases of condition (i), that are easy to verify and cover many economic models. For instance, both the restrictions imposed by the permanent-income model (stationary income process) and the present-value model have a block-triangular structure and satisfy Assumption 3*(iii). Moreover, from Equations (7) and (12) it is easy to see that the first derivative of $g(b)$ is continuous on $\mathcal{B}$ if $\mathcal{B}$ is compact. $\square$

The limit distribution of the MD estimator is obtained by showing that the sample objective function $Q_T(b_0 + \Lambda_T^{-1} s)$ can be approximated by the quadratic function $Q_{q,T}(b_0 + \Lambda_T^{-1} s)$ given in Equation (16).[4] Assumption 3 guarantees that the contributions of the remainder term from the Taylor series approximation are asymptotically negligible. The following Theorem characterizes the large sample behavior of $\hat{b}_T$.

Theorem 3 *Suppose that Assumptions 1 to 3 are satisfied. Then (i) $\|\Lambda_T(\hat{b}_T - b_0)\| = O_p(1)$, and (ii) $\Lambda_T(\hat{b}_T - b_0) = \hat{s}_{q,T} + o_p(1)$.*

The first part characterizes the order of consistency. The second part states that the limit distribution of the nonlinear MD estimator $\hat{b}_T$ is equivalent to the distribution of the local estimator $\hat{s}_{q,T}$, which minimizes the quadratic approximation of the sample objective function $Q_T(b_0 + \Lambda_T^{-1} s)$. The limit distribution of $\hat{s}_{q,T}$ is given in Theorem 2.

[4] An advantage of deriving limit distributions of extremum estimators through quadratic approximations of sample objective functions is that the derivation can be extended to the case in which b_0 is on the boundary of $\mathcal{B}$. Based on the very general results in Andrews (1999) this extension is straightforward. However, the boundary case is not pursued in this paper.

6 Mixed Normality of the Unrestricted Estimator

The limit distribution of the MD estimator generally depends on the choice of the sequence of weight matrices $\{\tilde{W}_T\}$. If the asymptotic distribution of the unrestricted estimator $\hat{a}_T$ is mixed normal, it is possible to develop an optimality theory for the MD estimation. Moreover, we can construct a test for the null hypothesis that $a_0 = g(b_0)$ for a $b_0 \in \mathcal{B}$. The mixed normal case is of great practical importance. It arises, for instance, in the examples considered in Section 2. Park and Phillips (2000) showed that the maximum likelihood estimator of the regression coefficients in a non-stationary binary choice model also has a mixed-normal limit distribution.

6.1 Optimal Weight Matrix

Consider the following class $\mathcal{M}$ of minimum distance estimators:

> $\mathcal{M}$: $\hat{b}_T$, $\hat{a}_T$, and $\tilde{W}_T$ satisfy Assumption 2, $\alpha \equiv \eta^{1/2} Z$, where η is a $q \times q$ random matrix that is positive definite with probability one and Z is a $q \times 1$ standard normal random vector that is independent of η and W.

A common criterion for asymptotic efficiency of an estimator is the concentration of the limit distribution, e.g., Basawa and Scott (1983). This criterion does not require the competing estimates to be asymptotically normal and has been widely used in the non-stationary time series literature, such as Saikkonen (1991), Phillips (1991), and Jeganathan (1995).

Definition 2 *(Efficiency) Let $\hat{b}_T$ and b_T^* be two estimator for b_0 belonging to class $\mathcal{M}$. b_T^* is asymptotically more efficient than $\hat{b}$ if*

$$\lim_{T\to\infty} I\!P\{\Lambda_T(\hat{b}_T - b_0) \in C\} \leq \lim_{T\to\infty} I\!P\{\Lambda_T(b_T^* - b_0) \in C\}$$

for any convex set $C \subset \mathbb{R}^p$, symmetric about the origin. If the inequality holds for all $\hat{b}_T \in \mathcal{M}$, then b_T^* is asymptotically efficient within the class $\mathcal{M}$.

We obtain the following result with respect to the optimal choice of weight matrix.

Theorem 4 *(Optimal weight) Suppose that the limit distribution α of the unrestricted estimator $\hat{a}_T$ is mixed normal $\mathcal{MN}(0,\eta)$. Then, an optimal sequence of weight matrices $\{\tilde{W}_T^o\}$ is a sequence of random matrices such that*

$$\begin{pmatrix} D_T^{-1'} R^{-1'} \tilde{W}_T^{o'} \tilde{W}_T^o R^{-1} D_T^{-1} \\ D_T R\left(\hat{a}_T - a_0\right) \end{pmatrix} \Rightarrow \begin{pmatrix} \eta^{-1} \\ \alpha \end{pmatrix}, \tag{22}$$

where η is the random variance in the mixed normal distribution.

If the b_0 is in the interior of the parameter space, Theorems 2 and 3 imply that the limit distribution of the MD estimator is mixed normal of the form

$$\Lambda_T(\hat{b}_T - b) \Longrightarrow \mathcal{MN}\Big(0, (\Gamma'\eta^{-1}\Gamma)^{-1}\Big).$$

Now reconsider the special case discussed in Section 5.1. We argued that the two-step estimation $\tilde{b}_{1,T}$ and $\tilde{b}_{2,T}$ can be interpreted as a minimum-distance estimator with block-diagonal weight matrix. Let η_{ij} denote the partitions of η that correspond to the partitions of the unrestricted parameter vector a. To implement the two-step estimation procedure it is reasonable to choose the weight matrices such that $T^{-2\nu_i}\tilde{W}'_{ii,T}\tilde{W}_{ii,T} \Longrightarrow \eta_{ii}^{-1}$. This yields the following limit distribution

$$T^{\nu_i}(\tilde{b}_{i,T} - b_i) \Longrightarrow \mathcal{MN}\Big(0, (G'_{ii}\eta_{ii}^{-1}G_{ii})^{-1}\Big).$$

However, Theorem 4 implies that the two-step estimator $\tilde{b}_T$ is dominated by minimum-distance estimators for which $D_T^{-1'} R^{-1'} \tilde{W}_T^{o'} \tilde{W}_T^o R^{-1} D_T^{-1}$ converges to η^{-1}. Thus, whenever η is not block-diagonal, it is inefficient to disregard the information in the short-run parameter estimates $\hat{a}_{2,T}$ about the parameters b_1. It is also inefficient to estimate b_2 by treating b_1 as known.

Remark: The restriction function for the present-value model in Section 2 is block-diagonal and the asymptotic covariance matrix of the unrestricted parameter estimates turns out to be block-diagonal. Hence, the two-step MD estimator is asymptotically efficient in Example 2. In the permanent-income model (Example 1) the asymptotic covariance matrix of the rotated parameter estimates, however, is not block-diagonal (see Equation (9)) and the two-step procedure is inefficient. □

6.2 Testing the Validity of the Restriction

To test the validity of the imposed restrictions we consider the following J-test statistic

$$J_T = [\hat{a}_T - g(\hat{b}_T)]'\tilde{W}_T'\tilde{W}_T[\hat{a}_T - g(\hat{b}_T)].$$

Under the assumption that b_0 is in the interior of $\mathcal{B}$ and $\hat{a}_T$ has a mixed normal limit distribution, we obtain the following result.

Theorem 5 *Suppose that Assumptions 1 to 3 are satisfied. Assume that the normalized unrestricted estimator* $D_T R(\hat{a}_T - a) \Longrightarrow \mathcal{MN}\left(0, \eta\right)$.

(i) Then,

$$J_T \Longrightarrow \sum_{i=1}^{q-p} d_i \chi_1^2(i),$$

where $\chi_1^2(i)$ *denote iid* χ^2 *random variables with one degree of freedom, that are independent of* d_i. *The* d_i*'s are non-zero random variables that correspond to the eigenvalues of*

$$\eta'^{1/2}W'\left(I_q - W\Gamma\left(\Gamma'W'W\Gamma\right)^{-1}\Gamma'W'\right)W\eta^{1/2}.$$

(ii) Under a sequence of optimal weight matrices $\{\tilde{W}_T^o\}$ *(defined in Theorem 4)*

$$J_T \Longrightarrow \chi^2_{q-p},$$

where χ^2_{p-q} *is a* χ^2 *random variable with* $q - p$ *degrees of freedom.*

7 Conclusions

In this paper we studied the asymptotic properties of the MD estimator in nonstationary time series models that are linear in the variables but involve nonlinear parameter restrictions. We analyze two applications of the proposed MD estimator: a permanent-income model based on a linear-quadratic dynamic programming problem and a present-value model.

To construct optimal MD estimators we allow the criterion function of the estimator to depend on a sequence of weight matrices that converges to a stochastic limit. We showed the consistency of the estimator using a Skorohod representation of the weakly converging objective function and derived the limit distribution of the MD estimator for smooth restriction functions. Our analysis relies on an equicontinuity condition for the parameter restriction function that allows a conventional first-order Taylor series approximation. If the equicontinuity condition is violated and some of the remainder terms are being amplified through convergence rate differentials of the unrestricted estimators, higher order expansions of the restriction function may become necessary. However, the results will be very model specific and difficult to generalize.

References

Amemiya, T (1985): *Advanced Econometrics*, Cambridge: Harvard University Press.

Andrews, D (1999): Estimation When a Parameter is on a Boundary, *Econometrica*, 67, 1341–1384.

Basawa, I. and D. Scott (1983): *Asymptotic Optimal Inference for Nonergodic Models,* New York: Springer-Verlag.

Billingsley, P. (1986): *Probability and Measure,* New York: John Wiley.

Chamberlain, G. (1984): Panel Data, pp 1465–1514 in Z. Griliches and M. Intriligator, eds. Handbook of Econometrics, Vol. 3, New York.

Campbell, J. and R. Shiller (1987): Cointegration and Tests of Present Value Models, *Journal of Political Economy*, 95, 1062–1088.

Elliot, G. (2000): Estimating Restricted Cointegrating Vectors, *Journal of Business & Economic Statistics*, 18, 91–99.

Hall, R.E. (1978): Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence, *Journal of Political Economy*, Vol. 86(6), 971–988.

Jeganathan, P. (1995): Some Aspects of Asymptotic Theory with Applications to Time Series Models, *Econometric Theory,* 11, 818–887.

Jennrich, R.I. (1969): Asymptotic Properties of Non-linear Least Squares Estimators, *Annals of Mathematical Statistics*, 40, 633–643.

Kim, J. and D. Pollard (1990): Cube Root Asymptotics, *Annals of Statistics*, 18, 191–219.

King, R., C. Plosser, J. Stock, and M. Watson (1991): Stochastic Trends and Economic Fluctuations, *American Economic Review*, 81, 819–840.

Ljungqvist, L. and T.J. Sargent (2000), *Recursive Macroeconomic Theory*, Cambridge: MIT Press.

Nagaraj, N. and W. Fuller (1991): Estimation of the Parameters of Linear Time Series Models subject to Nonlinear Restrictions, *Annals of Statistics*, 19, 1143–1154.

Park, J. (1992): Canonical Cointegrating Regression, *Econometrica*, 60, 119–144.

Park, J. and P.C.B. Phillips (2000): Nonstationary Binary Choice, *Econometrica*, 68, 1249–1280.

Phillips, P.C.B. (1991): Optimal Inference in Cointegrated Systems, *Econometrica*, 59, 238–306.

Phillips, P.C.B. (1993): Nonlinear Restriction and Cointegration, *mimeo.*

Phillips, P.C.B. and B. Hansen (1990): Statistical Inference in Instrumental Variables Regression with I(1) Processes, *Review of Economic Studies,* 57, 99–125.

Saikkonen, P. (1991): Asymptotically Efficient Estimation of Cointegration Regressions, *Econometric Theory*, 7, 1–21.

Saikkonen, P. (1993): Estimation of Cointegration Vectors with Linear Restrictions, *Econometric Theory*, 9, 19–35.

Saikkonen, P. (1995): Problems with the Asymptotic Theory of Maximum Likelihood Estimation in Integrated and Cointegrated systems, *Econometric Theory*, 11, 888–911.

Stock, J. and M. Watson (1993): A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems, *Econometrica*, 61, 783–820.

van der Waart, A. and Wellner (1996): *Weak Convergence and Empirical Processes*, New York, Springer-Verlag.

Wu, C. (1981): Asymptotic Theory of Nonlinear Least Squares Estimation, *Annals of Statistics*, 9, 501–513.

Appendix: Proofs

Lemma 2 *(Lemma 1 in Wu (1981))*

Suppose that for any $\delta > 0$

$$\liminf_{T\to\infty} \inf_{\|b-b_0\|\geq\delta} \left(Q_T(b) - Q_T(b_0)\right) > 0 \ \text{a.s. (or in prob.)}.$$

Then, $\hat{b}_T \longrightarrow b_0$ a.s. (or in prob.) as $T \longrightarrow \infty$.

Proof of Theorem 1

Define $\alpha_T = D_T R(\hat{a}_T - a_0)$ and $W_T = \tilde{W}_T R^{-1} D_T^{-1}$. By Assumption 2

$$[\alpha_T', vec(W_T)']' \Longrightarrow [\alpha', vec(W)']'.$$

Using the Skorohod construction, e.g. Billingsley (1986), one can find a probability space $(\Omega^*, \mathcal{F}^*, I\!P^*)$ with random variables $\alpha^*, \alpha_T^*, W^*, W_T^*$ that are distributionally equivalent to α, α_T, W, W_T, respectively, and $[\alpha_T^{*\prime}, vec(W_T^*)'] \xrightarrow{a.s.} [\alpha^{*\prime}, vec(W^*)']$ in $[I\!P^*]$.

Define $q_T(b) = D_T R(g(b) - a_0)$. From the uniqueness assumption that $a_0 = g(b)$ only if $b = b_0$, it follows that $q_T(b) = 0$ if and only if $b = b_0$. Furthermore, under Assumptions 1 and 2, we have

$$\inf_{\|b-b_0\|\geq\delta} \|q_T(b)\|^2 \geq \lambda_{\min}\left(D_T\right)\lambda_{\min}\left(R'R\right) \inf_{\|b-b_0\|\geq\delta} \|q_T(b)\|^2 \to \infty, \tag{23}$$

where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of matrix A.

The objective function can be rewritten as

$$Q_T(b) = \frac{1}{2}\alpha_T' W_T' W_T \alpha_T - q_T(b)' W_T' W_T \alpha_T + \frac{1}{2} q_T(b)' W_T' W_T q_T(b).$$

Moreover, we define

$$Q_T^*(b) = \frac{1}{2}\alpha_T^{*\prime} W_T^{*\prime} W_T^* \alpha_T^* - q_T' W_T^{*\prime} W_T^* \alpha_T^* + \frac{1}{2} q_T' W_T^{*\prime} W_T^* q_T(b),$$

which is distributionally equivalent to $Q_T(b)$. Let $\hat{b}_T^*$ be the MD estimator based on the objective function $Q_T^*(b)$. We will use Lemma 2 to show that $\hat{b}_T^* \xrightarrow{a.s.} b_0$. Since $\hat{b}_T^* \equiv \hat{b}_T$ on the original probability space, it can be deduced that $\hat{b}_T \xrightarrow{p} b_0$.

We show that the sufficient condition in Lemma 2 is satisfied. For a given $\delta > 0$

$$\begin{aligned}
&\liminf_{T\to\infty} \inf_{\|b-b_0\|\geq\delta} \left(Q_T^*(b) - Q_T^*(b_0)\right) \\
&= \liminf_{T\to\infty} \inf_{\|b-b_0\|\geq\delta} \left\{ q_T(b)'W_T^{*\prime}W_T^*q_T(b) \left[\frac{1}{2} - \frac{q_T(b)'W_T^{*\prime}W_T^*\alpha_T^*}{q_T(b)'W_T^{*\prime}W_T^*q_T(b)}\right]\right\} \\
&\geq \liminf_{T\to\infty} \left(\inf_{\|b-b_0\|\geq\delta} \|W_T^*q_T(b)\|^2\right)\left(\frac{1}{2} - \sup_{\|b-b_0\|\geq\delta} \frac{|q_T(b)'W_T^{*\prime}W_T^*\alpha_T^*|}{\|W_T^*q_T(b)\|^2}\right).
\end{aligned}$$

The equality and the inequality hold because W_T^* converges almost surely to a non-singular matrix by Assumption 2 and in consequence, together with (23) we have

$$\liminf_{T\to\infty} \left(\inf_{\|b-b_0\|\geq\delta} \|W_T^*q_T(b)\|^2\right) > 0.$$

The Cauchy-Schwarz inequality implies that

$$\sup_{\|b-b_0\|\geq\delta} \frac{|q_T(b)'W_T^{*\prime}W_T^*\alpha_T^*|}{\|W_T^*q_T(b)\|^2} \leq \frac{\|W_T^*\alpha_T^*\|}{\inf_{\|b-b_0\|\geq\delta}\|W_T^*q_T(b)\|} \longrightarrow 0$$

almost surely, because $\|W_T^*\alpha_T^*\| \longrightarrow \|W\alpha\|$ and $\|W_T^*q_T(b)\| \longrightarrow \infty$ almost surely for $\|b-b_0\| > \delta$ as in (23). Thus, it can be deduced that

$$\liminf_{T\to\infty} \inf_{\|b-b_0\|\geq\delta} \left(Q_T^*(b) - Q_T^*(b_0)\right) > 0. \quad \square$$

Proof of Lemma 1

$\Gamma_T = D_T RG\Lambda_T^{-1}$. Let $\Gamma_{T,ij}$ and Γ_{ij} denote the elements ij of the matrices Γ_T and Γ, respectively. Suppose the first p rows of $\tilde{G}$ are linearly independent such that $\nu_{(j)} = \nu_j$. Then $\Gamma_{T,ij} = T^{\nu_i-\nu_j}\Gamma_{ij}$. If $i \geq j$ then $\nu_i \leq \nu_j$ and $T^{\nu_i-\nu_j}$ is $O(1)$. If $j > i$, then $\Gamma_{ij} = 0$ because $G_*U_*^{-1}$ is lower triangular. Moreover, $\Gamma_{T,ii} = \Gamma_{ii} \neq 0$ because the diagonal elements of L_* are non-zero since G_* has full rank. Therefore, Γ has full row rank. The argument can be easily extended to the case in which there is linear dependence among the first p rows of $\tilde{G}$, by noting that $\nu_{(j)} \leq \nu_j$ and $G_{i.}[U_*^{-1}]_{.j} = 0$ for $i < \iota(j)$. $\square$

Proof of Sufficiency of Assumption 3*

Part (i): Since $\partial g(b)/\partial b$ is continuous and the parameter space $\mathcal{B}$ is compact, we can deduce that $\partial g(b)/\partial b$ is uniformly continuous around b_0. Suppose $\nu_i \leq \nu_{\iota(j)}$.

Then

$$\sup_{\|b_1-b_0\|\leq\delta_T}\left|[D_TR\Phi[b_1,b_0]\Lambda_T^{-1}]_{(ij)}\right| = T^{\nu_i-\nu_{\iota(j)}}\sup_{\|b_1-b_0\|\leq\delta_T}\left|[D_TR\Phi[b_1,b_0]U_*^{-1}]_{(ij)}\right| \longrightarrow 0$$

because $T^{\nu_i-\nu_{\iota(j)}}$ is $O(1)$ and $\partial g(b)/\partial b'$ is uniformly continuous around b_0.

If $\nu_i > \nu_{\iota(j)}$,

$$\sup_{\|b_1-b_0\|\leq\delta_T}\left|[D_TR\Phi[b_1,b_0]\Lambda_T^{-1}]_{(ij)}\right| = T^{\nu_i-\nu_{\iota(j)}}[D_TR\Phi[b_1,b_0]U_*^{-1}]_{(ij)} = 0$$

by Assumption 3(ii).

To prove (ii) and (iii) we will verify that condition (i) is satisfied.

Part (ii): Define $\tilde{\Phi}[b_1,b_0] = R\Phi[b_1,b_0]$. Now consider

$$\left[D_T\tilde{\Phi}[b_1,b_0]\Lambda_T^{-1}\right]_{ij} = T^{\nu_i-\nu_{\iota(j)}}\left(\sum_{l=1}^{p}\tilde{\Phi}_{il}[b_1,b_0][U_*^{-1}]_{lj}\right).$$

For $\nu_i > \nu_{\iota(j)}$, $[D_T\tilde{\Phi}[b_1,b_0]\Lambda_T^{-1}]_{ij} = 0$ because of the linearity assumption of $\tilde{g}_i(b)$ in b so that $\tilde{\Phi}_{ij}[b_1,b_0] = 0$ for $j = 1,\ldots,p$.

Part (iii): Let $\tilde{\Phi}_{lk}[b_1,b_0]$ and $[U_*^{-1}]_{lk}$ denote the submatrices of $\tilde{\Phi}[b_1,b_0]$ and U_*^{-1} that conform with the partitions of $\tilde{a}$ and b in Definition 1. Due to the block-triangular structure, $\tilde{\Phi}[b_1,b_0] = 0$ for $k > l$. Moreover, $[U_*^{-1}]_{lk} = 0$ for $l \neq k$. Thus,

$$\begin{aligned}\left[D_T\tilde{\Phi}[b_1,b_0]\Lambda_T^{-1}\right]_{lk} &= T^{\nu_l-\nu_k}\sum_{j=1}^{K}\tilde{\Phi}_{lj}[b_1,b_0][U_*^{-1}]_{jk} \\ &= T^{\nu_l-\nu_k}\sum_{j=1}^{l}\tilde{\Phi}_{lj}[b_1,b_0][U_*^{-1}]_{jk},\end{aligned}$$

which is zero whenever $k > l$. Hence condition (i) is satisfied. □

Proof of Theorem 3

Part (i): The proof is similar to the proof of Theorem 1 in Andrews (1999). Write $s = \Lambda_T(b - b_0)$. The objective function is of the form

$$\begin{aligned}Q_T(b) &= Q_T(b_0) - \alpha_T'W_T'W_T\Gamma_Ts + \frac{1}{2}s'\Gamma_T'W_T'W_T\Gamma_Ts \\ &\quad -\alpha_T'W_T'W_TD_TR\Phi(b_+,b_0)\Lambda_T^{-1}(\Gamma_T'W_T'W_T\Gamma_T)^{-1}\Gamma_T'W_T'W_T\Gamma_Ts \\ &\quad +s'\Gamma_T'W_T'W_TD_TR\Phi(b_+,b_0)\Lambda_T^{-1}(\Gamma_T'W_T'W_T\Gamma_T)^{-1}\Gamma_T'W_T'W_T\Gamma_Ts \\ &\quad +\frac{1}{2}s'\Gamma_T'W_T'W_T\Gamma_T(\Gamma_T'W_T'W_T\Gamma_T)^{-1}\Lambda_T^{-1'}\Phi'(b_+,b_0)R'D_T'W_T' \\ &\quad \times W_TR\Phi(b_+,b_0)\Lambda_T^{-1}(\Gamma_T'W_T'W_T\Gamma_T)^{-1}\Gamma_T'W_T'W_T\Gamma_Ts.\end{aligned}$$

Let $\hat{b}_T$ be the MD estimator and $\hat{s}_T = \Lambda_T(\hat{b}_T - b_0)$. Then,

$$\begin{aligned} 0 &\leq Q_T(b_0) - Q_T(\hat{b}_T) \\ &= O_p(1)\|W_T\Gamma_T\hat{s}_T\| - \frac{1}{2}\|W_T\Gamma_T\hat{s}_T\|^2 + o_p(1)\|W_T\Gamma_T\hat{s}_T\| \\ &\quad -o_p(1)\|W_T\Gamma_T\hat{s}_T\| - \frac{1}{2}o_p(1)\|W_T\Gamma_T\hat{s}_T\| \end{aligned}$$

where the last equality holds because $\alpha_T = O_p(1)$ and by Assumption 3. Denote the $O_p(1)$ term by ξ_T and rewrite the inequality as

$$\frac{1}{2}\|W_T\Gamma_T\hat{s}_T\|^2 \leq \xi_T\|W_T\Gamma_T\hat{s}_T\| + o_p(1)$$

Thus,

$$\left(\|W_T\Gamma_T\hat{s}_T\| - \xi_T\right)^2 \leq O_p(1)$$

and therefore

$$\|\hat{s}_T\| \leq \left\|(\Gamma_T'W_T'W_T\Gamma_T)^{-1}\right\|\|W_T\Gamma_T\hat{s}_T\| = O_p(1),$$

which implies the desired result.

Part (ii): Follows from Theorem 3(a) in Andrews (1999). □

Proof of Theorem 4

We follow the arguments in Theorem 3.1 in Saikkonen (1991). Recall that $\alpha \equiv \eta^{1/2}Z$, where $Z \equiv \mathcal{N}(0, I_q)$ and Z is independent of the random covariance matrix η and the limit weight matrix W. Let

$$\theta = (\Gamma'\eta^{-1}\Gamma)^{-1}\Gamma'\eta^{-1}\alpha, \quad \psi = (\Gamma'W'W\Gamma)^{-1}(\Gamma'W'W\alpha)$$

and $\phi = \psi - \theta$. It can be easily verified that $I\!E[\theta\phi'|\eta, W] = 0$, which implies that ϕ and θ are independent conditional on η and W. Let C be any convex set, symmetric about the origin, and $\hat{b}_T^o$ an MD estimator with an optimal sequence of weight matrices, then for any MD estimator in $\mathcal{M}$:

$$\begin{aligned} \lim_{T\to\infty} I\!P\{\Lambda_T(\hat{b}_T - b_0) \in C\} &= I\!E\Big[I\!P[\{\theta + \phi \in C\}|\eta, W]\Big] \\ &\leq I\!E\Big[I\!P[\{\theta \in C\}|\eta, W]\Big] \\ &= \lim_{T\to\infty} I\!P\{\Lambda_T(\hat{b}_T^o - b_0) \in C\} \end{aligned}$$

The inequality follows from Lemma 2.3.1 in Basawa and Scott (1983). $\square$

Proof of Theorem 5

Write $\hat{s}_T = \Lambda_T \left(\hat{b}_T - b_0\right)$. Under the assumptions of the theorem, we may write

$$J_T = \alpha_T' W_T' W_T \alpha_T - 2\alpha_T' W_T' W_T \Gamma_T \hat{s}_T + \hat{s}_T' \Gamma_T' W_T' W_T \Gamma_T \hat{s}_T + o_p(1). \tag{24}$$

Since Γ_T is asymptotically full rank and b_0 is in the interior of $\mathcal{B}$, we may write

$$\hat{s}_T = \left(\Gamma_T W_T' W_T \Gamma_T\right)^{-1} \Gamma_T' W_T' W_T \alpha_T + o_p(1)$$

Replacing $\hat{s}_T$ in (24), we have

$$J_T = \alpha_T' W_T' \left(I_q - W_T \Gamma_T \left(\Gamma_T' W_T' W_T \Gamma_T\right)^{-1} \Gamma_T' W_T'\right) W_T \alpha_T + o_p(1).$$

Under the assumptions, it follows that

$$J_T \Rightarrow Z_q' \eta'^{1/2} W' \left(I_q - W\Gamma \left(\Gamma' W' W \Gamma\right)^{-1} \Gamma' W'\right) W \eta^{1/2} Z_q$$

as $T \to \infty$. Notice that $I_q - W\Gamma\left(\Gamma' W' W\Gamma\right)^{-1} \Gamma' W'$ is an idempotent (random) matrix of rank $q-p$ with probability one and recall that $W\eta^{1/2}$ is of full rank with probability one. From the spectral decomposition of $\eta'^{1/2} W' \left(I_q - W\Gamma \left(\Gamma' W' W\Gamma\right)^{-1} \Gamma' W'\right) W\eta^{1/2}$ and Z_q being independent of $\eta'^{1/2} W' \left(I_q - W\Gamma\left(\Gamma' W' W\Gamma\right)^{-1} \Gamma' W'\right) W\eta^{1/2}$, the result in Part (i) follows.

Part (ii) is straightforward because $W\eta^{1/2} = I_q$ and all the non-zero eigenvalues of $I_q - W\Gamma\left(\Gamma' W' W\Gamma\right)^{-1} \Gamma' W'$ are 1's. $\square$