Penn Institute For Economic Research

Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19103-6297

pier@ssc.upenn.edu
http://www.econ.upenn.edu/pier

PIER Working Paper 01-016

"Minimum Distance Estimation of Nonstationary Time Series Models"

by

Hyungsik Roger Moon and Frank Schorfheide

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

April 2001

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. As applications, we investigate cointegration regression models, present-value models, and a permanent-income model based on a linear-quadratic dynamic programming problem.

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 "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. 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. 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. An advantage of the MD estimator over the direct estimator is that in the presence of stochastic trends and nonlinear restrictions it is easier to verify the regularity conditions that guarantee consistency and weak convergence of the estimator.

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. Equicontinuity conditions for the derivatives of the restriction function together with some useful sufficient conditions are provided. Our conditions are easier to verify than the stochastic equicontinuity conditions, e.g. Saikkonen (1995), that commonly arise if b is directly estimated.

Second, if the model includes stochastic trends, the limit distribution of $\hat{a}_T$ is non-standard. Among the examples studied in this paper is the cointegration regression model

$$
\begin{aligned}
y_{1,t} &= A_0' y_{2,t-1} + u_{1,t} \\
y_{2,t} &= y_{2,t-1} + u_{2,t}
\end{aligned}
\tag{1}
$$

and $a_0 = vec(A_0)$. The vec-operator stacks the columns of a matrix in a vector. 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 limit distribution with a random covariance matrix. Optimality considerations suggest to use a sequence of weight matrices 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 introduces several specific examples that highlight the complications in MD estimation addressed in this paper and reviews the existing literature. 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 revisits the examples laid out in Section 2 and Section 8 concludes. The appendix contains mathematical derivations and proofs.

The notation "$\equiv$" is used to signify distributional equivalence, "$\Longrightarrow$" denotes convergence in distribution, "$\overset{p}{\longrightarrow}$" denotes convergence in probability, and "$\overset{a.s.}{\longrightarrow}$" is almost-sure convergence. 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

2.1 Restricted Cointegration Relationships

Consider the cointegration model in Equation (1). Suppose that the short-run dynamics are not explicitly modeled and that $g(b)$ restricts only the matrix of long-run parameters A. 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.

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. 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. His theoretical results are reproduced as a special case of our more general analysis of MD estimators for time series models.

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.

The distribution theory for estimators of the unrestricted parameter matrix a_0 in Model (1) is well developed. Assume that the partial sum process of u_t converges to a vector Brownian motion: $T^{-1/2} \sum_{t=1}^{[Tr]} u_t \Longrightarrow B(r) \equiv BM(\Omega)$. Ω is the long-run variance of u_t given by $\lim_{T \to \infty} \frac{1}{T} I\!E[(\sum_{t=1}^{T} u_t)(\sum_{t=1}^{T} u_t)']$. We will use $B_i(r)$ and Ω_{ij} to denote the partitions of $B(r)$ and Ω that conform to the partitions of u_t. The limit distribution of the OLS estimator $\hat{a}_{T,OLS}$ is skewed and miscentered due to possible correlation between $u_{1,t}$ and $u_{2,t}$ and serial correlation of the u_t's. To overcome these problems, several methods have been proposed. For example[1], Phillips and Hansen (1990) proposed the fully modified (FM) least squares estimator $\hat{a}_{T,FM}$ and found that its limit distribution is mixed normal (MN):

$$T(\hat{a}_{T,FM} - a_0) \Longrightarrow \eta^{1/2} Z, \tag{2}$$

where

$$\eta = \Omega_{11.2} \otimes \left(\int_0^1 B_2 B_2' \right)^{-1},$$

---

[1]Other examples include Park (1992), Saikkonen (1991), and Stock and Watson (1993).

and $\Omega_{11.2} = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$. $Z \equiv \mathcal{N}(0, I_q)$ is a $q \times 1$ standard normal random vector that is independent of η and $\otimes$ is the Kronecker product.

Example 1: A special case of a restricted cointegration regression model is the seemingly unrelated cointegration (SURC) model, for instance considered by Moon (1999):

$$s_{m,t} = b_m' x_{m,t-1} + v_{m,t}, \quad m = 1, \ldots, M \qquad (3)$$

$$x_{m,t} = x_{m,t-1} + w_{m,t},$$

where $x_{m,t}$ is a $L \times 1$ vector of integrated regressors, and $[v_{m,t}', w_{m,t}']'$ is a stationary process. The SURC model can be interpreted as multivariate cointegration model with parameter restrictions. To cast it into the form of Equation (1) define $y_{1,t} = [s_{1,t}, \ldots, s_{M,t}]'$, $y_{2,t} = [x_{1,t}', \ldots, x_{M,t}']'$, $u_{1,t} = [v_{1,t}, \ldots, v_{M,t}]'$, $u_{2,t} = [w_{1,t}', \ldots, w_{M,t}']'$, and $b = [b_1', \ldots, b_M']'$. The restriction is linear and can be expressed as

$$a = vec(A) = Gb, \qquad (4)$$

where $G = diag(i_1, \ldots, i_M) \otimes I_L$. i_m denotes the m'th column of the $M \times M$ identity matrix I_M and $diag(A_1, \ldots, A_k)$ is a block-diagonal matrix with blocks $A_1, \ldots, A_k$ on its diagonal. The limit distribution of the unrestricted estimator is given by Equation (2).

Park and Ogaki (1991), Moon (1999), and Moon and Perron (2000) proposed efficient estimation of the cointegration parameters b using canonical cointegration regression (CCR), fully modified (FM) estimation, and dynamic GLS estimation, respectively. Interestingly, the three estimators are asymptotically equivalent. We will demonstrate below that the MD estimator has the same limit distribution as the above mentioned estimators, provided the sequence of weight matrices $\{W_T\}$ is chosen optimally. $\square$

2.2 Restrictions between Short-run and Long-run Dynamics

Many economic models imply restrictions across parameters that control short-run adjustments and long-run relationships. The "long-run" parameters are usually

estimated at a faster rate than the "short-run" parameters. This has important consequences for minimum distance estimation.

Example 2: A present value model for two scalar variables $y_{1,t}$ and $y_{2,t}$ states that $y_{1,t}$ is the present discounted value of expected future $y_{2,t}$:

$$y_{1,t} = \sum_{j=1}^{\infty} \left(\frac{1}{1+r} \right)^j \mathbb{E}_t[y_{2,t+j}]. \tag{5}$$

This relationship is a solution to the following difference equation

$$y_{1,t} = \frac{1}{1+r} \mathbb{E}_t[y_{1,t+1} + y_{2,t+1}]. \tag{6}$$

If $y_{1,t}$ is a stock price and $y_{2,t}$ a dividend payment, then a risk neutral investor is indifferent between the stock and bond that guarantees to pay the interest rate $1/\beta - 1$ if Equation (6) is satisfied. This model has been widely studied in the empirical finance literature, for instance by Campbell and Shiller (1987). They assumed that

$$
\begin{aligned}
y_{1,t} &= \frac{1}{r} y_{2,t-1} + u_{1,t}, \tag{7}\\
y_{2,t} &= y_{2,t-1} + u_{2,t}.
\end{aligned}
$$

Thus, the "dividend" process $y_{2,t}$ is a random walk with innovations $u_{2,t}$. The linear combination $y_{1,t} - \frac{1}{r} y_{2,t}$ is called the spread and reflects the difference between $y_{1,t}$ and the present discounted value of future $y_{2,t}$ under the assumption that $y_{2,t+j} = y_{2,t}$ for all j.

Define $u_t = [u_{1,t}, u_{2,t}]'$. Suppose the short-run dynamics are modeled as a vector autoregressive process, as in Campbell and Shiller (1987). To simplify the exposition we consider a VAR(1):

$$u_t = A' u_{t-1} + \epsilon_t. \tag{8}$$

The difference equation (6) imposes the following restriction on the short-run dynamics of u_t:

$$A_{11} + A_{12} = (1+r), \quad r(A_{21} + A_{22}) = -(1+r).$$

Let a_{lr} be the unrestricted coefficient of $y_{2,t-1}$ in the first equation of (7), define $a_{sr} = vec(A)$, and let b be a 3×1 vector of parameters. The present value model restrictions can be expressed as follows:

$$
\begin{aligned}
a_{lr} &= 1/b_1 \\
A_{11} &= b_2 & A_{12} &= 1 + b_1 - b_2 \\
A_{21} &= b_3 & A_{22} &= -[b_3 + (1 + b_1)/b_1]
\end{aligned}
$$

The restriction function in this example is nonlinear and has the block-triangular structure

$$a_{lr} = g_{lr}(b_1), \quad a_{sr} = g_{sr}(b_1, b_2, b_3). \tag{9}$$

Assume that $T^{-1/2} \sum_{t=1}^{[Tr]} u_t \Longrightarrow B(r) \equiv BM(\Omega)$, where Ω is the long-run covariance matrix of u_t. Use $B_i(r)$ and Ω_{ij} to denote the partitions that conform with the partitions of u_t. The unrestricted parameters a can be estimated, for instance, by maximum likelihood or quasi-maximum likelihood. Suppose $\epsilon_t \sim iid(0, \Sigma_{\epsilon\epsilon})$. Let $a = [a_{lr}, a'_{sr}]'$. The limit distribution of $\hat{a}_T$ is of the mixed normal form

$$
D_T \left[\begin{array}{c} \hat{a}_{T,lr} - a_{lr} \\ \hat{a}_{T,sr} - a_{sr} \end{array} \right] \Longrightarrow \eta^{1/2} Z, \tag{10}
$$

where $D_T = diag(T, T^{1/2} I_4)$ and

$$
\eta = \left[\begin{array}{cc} \Omega_{11.2} \left(\int_0^1 B_2^2 \right)^{-1} & 0 \\ 0 & \Sigma_{\epsilon\epsilon} \otimes Q_{uu}^{-1} \end{array} \right].
$$

The random covariance matrix η is block-diagonal. $Z \equiv \mathcal{N}(0, I_5)$ is independent of η and $Q_{uu} = \mathbb{E}[u_t u'_t]$.

Informal inspection suggests that the elements of b can be estimated in a two-step procedure. First, estimate b_1 based on $\hat{a}_{T,lr}$, which leads a T-consistent estimator. Second, estimate b_2 and b_3 conditional on $\hat{b}_{1,T}$ by minimizing the discrepancy between $\hat{a}_{T,sr}$ and $g_{sr}(\hat{b}_{1,T}, b_2, b_3)$. We will show in Section 5, that this two-step procedure is asymptotically equivalent to the optimal MD estimation, whenever $\hat{a}_{lr,T}$ and $\hat{a}_{sr,T}$ are asymptotically uncorrelated. $\square$

Example 3: 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. It is well known that the optimal decision rule is a linear function of the state variables. However, it generates non-linear restrictions across the parameters that govern the joint evolution of states and controls. In macroeconomics, linear-quadratic programming problems are frequently used to approximate smooth non-linear model economies, e.g., Kydland and Prescott (1982).

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

$$-\frac{1}{2}\mathbb{E}_0\left[\sum_{t=0}^{\infty}\left(\frac{1}{1+r}\right)^t(C_t - Z_t - \alpha)^2\right] \qquad (11)$$

subject to the constraints

$$
\begin{aligned}
Z_{t+1} &= \psi C_t + \epsilon_{1,t+1} \\
W_{t+1} &= (1+r)(W_t - C_t) + I_{1,t+1} + I_{2,t+1} \\
I_{1,t+1} &= \mu(1-\phi) + \phi I_{1,t} + \epsilon_{2,t+1} \\
I_{2,t+1} &= \epsilon_{3,t+1},
\end{aligned}
\qquad (12)
$$

where C_t is consumption, Z_t is a habit stock, r is the real interest rate, and $I_{1,t}$ and $I_{2,t}$ are income processes. Habit is a fraction of past consumption and is shifted by a taste shock $\epsilon_{1,t}$. Habit formation is often introduced to improve the asset pricing implication of representative agent models. Define the vector of exogenous shocks $\epsilon_t = [\epsilon_{1,t}, \epsilon_{2,t}, \epsilon_{3,t}]'$. We will assume that $\epsilon_t \sim iid(0, \Sigma_{\epsilon\epsilon})$.

To simplify the subsequent analysis of the optimization model, we impose $\psi = 0$. Moreover, we will assume that the income component $I_{2,t}$ 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 (12) exactly.

The optimal decision rule for consumption is

$$C_t = \frac{1}{1+r}\mu + \frac{r}{1+r}W_t + \frac{\phi r}{(1+r-\phi)(1+r)}(I_{1,t} - \mu_1) + \frac{r}{1+r}\epsilon_{1,t} \qquad (13)$$

Suppose the model is fitted to data on C_t, W_t, and $I_{1,t}$. The law of motion for these three variables is

$$
\begin{bmatrix} C_t \\ \Delta W_t \\ I_{1,t} \end{bmatrix} = \begin{bmatrix} \frac{r}{1+r} & \frac{\mu}{1+r} & \frac{\phi r(2-\phi)}{(1+r-\phi)(1+r)} \\ 0 & 0 & \frac{\phi(1-\phi)}{1+r-\phi} \\ 0 & \mu_1 & \phi \end{bmatrix} \begin{bmatrix} W_{t-1} \\ 1 \\ (I_{1,t-1} - \mu_1) \end{bmatrix}
$$
$$
+ \begin{bmatrix} \frac{r}{1+r} & \frac{r}{1+r} & \frac{r}{1+r} \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \epsilon_{3,t} \end{bmatrix} + \begin{bmatrix} -\frac{r^2}{1+r} & 0 & 0 \\ -r & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \epsilon_{1,t-1} \\ \epsilon_{2,t-1} \\ \epsilon_{3,t-1} \end{bmatrix} . \tag{14}
$$

According to the model, both consumption and wealth are $I(1)$ processes. As emphasized by Hall (1978), consumption is a Martingale. If $\phi = 1$ and $\mu_1 = 0$ the observable income process becomes a random walk and the model simplifies to a cointegration regression model with stochastic trends W_t and $I_{1,t}$ and a non-linear parameter restriction. We will focus on the case $0 \leq \phi < 1$, in which the income process is stationary.

Define $y_{1,t} = C_t$, $y_{2,t} = [\Delta W_t, I_{1,t}]'$, $x_{1,t} = W_{t-1}$, $x_{2,t} = [1, I_{1,t-1}]'$, $y_t = [y_{1,t}, y_{2,t}']'$, and $x_t = [x_{1,t}, x_{2,t}']'$. The regressor $x_{1,t}$ is $I(1)$, whereas $x_{2,t}$ is $I(0)$. The consumption model is nested in the following general model

$$
\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} A_{11}' & A_{21}' \\ 0 & A_{22}' \end{bmatrix} \begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}, \tag{15}
$$

where $u_t = [u_{1,t}, u_{2,t}']' = C_0 \epsilon_t + C_1 \epsilon_{t-1}$. According to the structural model, the errors $u_{1,t}$ and $u_{2,t}$ are uncorrelated with the stationary regressor $x_{1,t}$. This is a necessary condition for the consistent estimation of A_{21} and A_{22}. Define $a_{ij} = vec(A_{ij})$. The unrestricted parameter vector is $a = [a_{11}', a_{21}', a_{22}']'$ and the vector b is composed of the structural parameters r, μ, and ϕ. The restriction function $g(b)$ can be obtained from Equation (14).

Let Y, Y_i, X, X_i, U, and U_i be matrices with rows y_t', $y_{i,t}'$, x_t', $x_{i,t}'$, u_t', $u_{i,t}'$. Define the rotation matrix

$$
R = \left[\begin{array}{c|cc} 1 & 0 & 0 \\ \hline 0 & 1 & \mu \\ 0 & 0 & 1 \end{array} \right] .
$$

The unrestricted model can be expressed in matrix form as

$$Y = XR^{-1}RA + U = \tilde{X}\tilde{A} + U, \tag{16}$$

where $\tilde{X}$ and $\tilde{A}$ denote the rotated regressors and the rotated unrestricted parameters. The rotation demeans the stationary income process and preserves the block-triangular structure of the coefficient matrix: $\tilde{A}_{12} = 0$. Define $\eta_t = (I_{t-1} - \mu) + \epsilon_{2,t} + \epsilon_{3,t} - r\epsilon_{1,t-1}$. Assume that $\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]}\eta_t \Longrightarrow B(r) \equiv BM(\Omega)$, where Ω is the long-run variance of η_t.

Following the approach in Phillips and Solo (1992), we decompose u_t as

$$u_t = (C_0 + C_1)\epsilon_t + C_1(L-1)\epsilon_t = \nu_t + C_1(L-1)\epsilon_t \tag{17}$$

and define $\Sigma = (C_0 + C_1)\Sigma_{\epsilon\epsilon}(C_0 + C_1)'$. The unrestricted model can be estimated quasi-maximum likelihood, ignoring the $C_1(1-L)\epsilon_t$ component of u_t. Let V be the matrix with columns ν_t'. Suppose Σ were known. In this case the maximum likelihood estimator of A_{22} corresponds to the OLS estimator of

$$Y_2 = \tilde{X}_2\tilde{A}_{22} + V_2. \tag{18}$$

The estimator of A_{11} and A_{21} is given by the OLS estimation of

$$Y_1 = \tilde{X}_1\tilde{A}_{11} + \tilde{X}_2\tilde{A}_{21} + \hat{V}_2\Sigma_{22}^{-1}\Sigma_{21} + V_1. \tag{19}$$

The Σ_{ij}'s are partitions of Σ that conform with the partitions of u_t. The correction term $\hat{V}_2\Sigma_{22}^{-1}\Sigma_{21}$ reflects the conditional expectation of V_1 given $\hat{V}_2$ and is comparable to the correction term that arises in the maximum likelihood estimation of the cointegration regression model (Phillips (1991)). It can be shown that the limit distribution of $\hat{a}_T$ is of the form:[2]

$$D_T R \begin{bmatrix} \hat{a}_{T,11} - a_{11} \\ \hat{a}_{T,21} - a_{21} \\ \hat{a}_{T,22} - a_{22} \end{bmatrix} \Longrightarrow \eta^{1/2}Z, \tag{20}$$

[2]Detailed derivations are available from the authors on request.

where $Z \equiv \mathcal{N}(0, I_7)$, $D_T = diag(T, T^{1/2}I_2, T^{1/2}I_4)$, and

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

The various elements of the random covariance matrix η are defined as follows: $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, $Q_{11} = \int_0^1 B^2$, $Q_{12} = [\int_0^1 B, 0]$, $Q_{21} = Q_{12}'$, $Q_{22} = \mathbb{E}[\tilde{x}_{2,t}\tilde{x}_{2,t}']$, $Q_{11.2} = Q_{11} - Q_{12}Q_{22}^{-1}Q_{12}$, and $Q_{22.1} = Q_{22} - Q_{21}Q_{11}^{-1}Q_{12}$.

Unlike in the cointegration regression model examined by Phillips (1991), the quasi-maximum likelihood estimate of model (15) is not fully efficient. Ignoring the MA structure of u_t in the estimation of the coefficients on the stationary regressors introduces an inefficiency. So far, we assumed that Σ was known. In practice, one can estimate Σ by a consistent estimator as in a fully modified least squares procedure.

In this example the (rotated) restriction function is again block-triangular. However, due to the intercept in the cointegration equation, there is now a correlation between the T-consistent parameter estimates $\hat{a}_{T,11}$ and the $T^{1/2}$-consistent parameter estimates $\hat{a}_{T,21}$. In this case the optimal MD estimator will asymptotically dominate the two-step estimator described in Example 2. $\square$

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 be a data matrix of sample size T defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P}_a)$. It is assumed that the probability measures are indexed by a $q \times 1$ dimensional parameter vector $a \in \mathcal{A} \subset \mathbb{R}^q$. The data are drawn from the distribution $\mathbb{P}_{a_0}$. Let $b \in \mathcal{B} \subset \mathbb{R}^p$ be a second parameter vector and $g(\cdot)$ be a mapping from $\mathcal{B}$ to $\mathcal{A}$. We assume that $p \le q$ 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 $g(b_0) = a_0$.*

Let $\hat{a}_T$ be an estimator of the unrestricted parameter a_0 and $\tilde{W}_T$ be a weight matrix. Any vector $\hat{b}_T$ that minimizes the criterion function

$$Q_T(b) = \frac{1}{2}\|\tilde{W}_T(\hat{a}_T - g(b))\|, \tag{21}$$

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]

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 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 unrestricted estimator $\hat{a}_T$ is weakly consistent and that its limit distribution is distributionally equivalent to the random variable α. In non-stationary time series models the parameter estimates potentially converge at different rates. The matrix R rotates the unrestricted parameter space and separates directions in which convergence is fast from directions in which convergence is slow.

---

[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 $\hat{a}_T$.

The diagonal elements of D_T reflect the rates that correspond to 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 2 and 1 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

$$Q_T(b) = Q_T(b_0) - \alpha_T' W_T' W_T D_T R(g(b) - a_0) + \frac{1}{2}(g(b) - a_0)' R' D_T' W_T' W_T D_T R(g(b) - a_0)$$
(22)

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

$$\begin{aligned} Q_{q,T}(b) &= Q_{q,T}(b_0) - \alpha_T' W_T' W_T D_T RG(b - b_0) \\ &\quad + \frac{1}{2}(b - b_0)' G' R' D_T' W_T' W_T D_T RG(b - b_0). \end{aligned}$$
(23)

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

$$Q_{q,T}(b_0 + \Lambda_T^{-1}s) = Q_{q,T}(b_0) - \alpha_T' W_T' W_T \Gamma_T s + \frac{1}{2} s' \Gamma_T' W_T' W_T \Gamma_T s. \qquad (24)$$

Example 4: Let $R = I$. 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,

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

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

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

$$\hat{s}_{q,T} = (\Gamma_T' W_T' W_T \Gamma_T)^{-1} \Gamma_T' W_T' W_T \alpha_T \tag{25}$$

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

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

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

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

and the limit distribution of the unrotated parameter vector is non-degenerated. $\square$

5.2 Block-triangular Restriction Matrices

The restriction functions in Examples 2 and 3 are block-triangular. In this subsection we study the linear case. Suppose that $R = I$ 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:

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} G_{11} & 0 \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

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 $\tilde{a} = [\tilde{a}_1', \ldots, \tilde{a}_L']'$ such that the estimators of the elements of $\tilde{a}_l$ converge all at the same rate T^{ν_l}, $l = 1, \ldots, L$. 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 $k = 1, \ldots, K \leq L$ subvectors b_k and (i) $\tilde{g}_l(b) = f_l([b_1', \ldots, b_l']')$, $l = 1, \ldots, K$, and (ii) G_{ll} has full row rank, $l = 1, \ldots, K$.*

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

$$
T^{\nu_i}(\hat{b}_{i,T} - b_i) \Longrightarrow U_{ii}^{*-1} M_i' (\Gamma' W' W' \Gamma)^{-1} \Gamma' W' W \alpha, \ i = 1, 2. \tag{26}
$$

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

$$
T^{\nu_i}(\hat{b}_{i,T} - b_i) \Longrightarrow (G_{ii}' W_{ii}' W_{ii} G_{ii})^{-1} G_{ii} W_{ii}' W_{ii} \alpha_i, \ i = 1, 2. \tag{27}
$$

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

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

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

(ii) Estimate b_2 according to

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

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 (27). 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

$$g(b) = g(b_0) + G(b - b_0) + \Phi[b_+, b_0](b - b_0), \tag{28}$$

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 22) is approximated by

$$D_T R(g(b) - a_0) = \Gamma_T s + D_T R \Phi[b_+, b_0]\Lambda_T^{-1}s, \tag{29}$$

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 $\delta_T \to 0$, $\sup_{\|b_1 - b_0\| \leq \delta_T} \|D_T R\Phi[b_1, b_0]\Lambda_T^{-1}\| = o(1)$.*

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.*

Conditions (ii) and (iii) are special cases of condition (i), that are easy to verify and cover many economic models, such as the ones discussed in the examples. 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 (24).[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.

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:

---

[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.

$\mathcal{M}$: $\hat{b}_T$, $\hat{a}_T$, 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*

$$\begin{pmatrix} D_T^{-1'} R^{-1'} \tilde{W}_T^{o'} \tilde{W}_T^o R^{-1} D_T^{-1} \\ D_T R \left(\hat{a}_T - a_0 \right) \end{pmatrix} \Rightarrow \begin{pmatrix} \eta^{-1} \\ \alpha \end{pmatrix}, \tag{30}$$

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

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

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

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

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.

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 $D_T R(\hat{a}_T - a) \Longrightarrow \mathcal{MN}(0, \eta)$.*

(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 Examples Revisited

In the case of the cointegration regression model the limit distribution of $\hat{a}_T$ is given by Equation (2). Theorem 4 suggests to use a covariance estimate for $\hat{a}_T$ as weight matrix

$$\tilde{W}_T^o = \hat{\Omega}_{11.2}^{-1/2} \otimes \left(\sum_{t=1}^{T} y_{2,t} y_{2,t}'\right)^{1/2}, \tag{31}$$

where $\hat{\Omega}_{11.2}$ is a consistent estimate of $\Omega_{11.2}$, e.g. Phillips (1995). Since all elements of the vector $\hat{a}_T$ converge at the same rate T, it is not necessary to rotate the unrestricted parameter space: $R = I_p$ and $D_T = TI_p$. Since $\Lambda_T = TU_*$ and $\Gamma = GU_*^{-1}$,

$$T(\hat{b}_T - b_0) \Longrightarrow \mathcal{MN}\left(0, \left[G'\left(\Omega_{11.2}^{-1} \otimes \int_0^1 B_2 B_2'\right)G\right]^{-1}\right). \tag{32}$$

Example 1 (Continued): If $G = (diag(i_1, \ldots, i_M) \otimes I_L)$ then it can be verified that

$$G'\left(\Omega_{11.2}^{-1} \otimes \int_0^1 B_2 B_2'\right)G = \int_0^1 \tilde{B}_2 \Omega_{11.2}^{-1} \tilde{B}_2', \tag{33}$$

where $\tilde{B}_2 = diag(B_{w_1}, \ldots, B_{w_M})$ and B_{w_m} is the Brownian motion limit of the partial sum of $w_{m,t}$. Thus, the limit distribution of the optimal MD estimator is equivalent to the limit distribution of the efficient SURC estimators proposed by Park and Ogaki (1991), Moon (1999), and Moon and Perron (2000). $\square$

Example 2 (Continued): The rotation matrix is simply $R = I_5$. Since the restriction function is non-linear we have to check the equicontinuity condition stated in Assumption 3*. Provided that the parameter space $\mathcal{B}$ is compact, it is easy to see that the first derivative of $g(b)$ is continuous on $\mathcal{B}$. The restriction function is of the form $a_{lr} = g_{lr}(b_1)$ and $a_{sr} = g_{sr}(b_1, b_2, b_3)$. Hence, Assumption 3*(iii) is satisfied.

The covariance matrix of $\hat{a}_T = [\hat{a}_{lr,T}, \hat{a}'_{sr,T}]'$ is block-diagonal. Therefore the limit distribution simplifies to

$$T(\hat{b}_{lr,T} - b_{lr}) \implies \mathcal{MN}\left(0, \left[G'_{lr}\left(\Omega_{11.2}^{-1} \otimes \int_0^1 B_2^2\right)G_{lr}\right]^{-1}\right) \qquad (34)$$

$$T^{1/2}(\hat{b}_{sr,T} - b_{sr}) \implies \mathcal{N}\left(0, \left[G'_{sr}\left(\Sigma_{\epsilon\epsilon}^{-1} \otimes Q_{uu}\right)G_{sr}\right]^{-1}\right),$$

where $G_{lr} = \partial g_{lr}/\partial b_{lr}$ and $G_{sr} = \partial g_{sr}/\partial b'_{sr}$. In this case the optimal minimum distance estimator is asymptotically equivalent to the two-step estimation procedure described in Section 5.1. The two-step procedure has been used in applied work, for instance by Campbell and Shiller (1987). $\square$

Example 3 (Continued): Define $a_{lr} = a_{11}$, $a_{sr} = [a'_{21}, a'_{22}]'$. The rotated restriction function is block-triangular: $\tilde{a}_{lr} = \tilde{g}_{lr}(r)$, and $\tilde{a}_{sr} = \tilde{g}_{sr}(r, \mu, \phi)$. It satisfies Assumption 3*. Since the random covariance matrix η is not block-diagonal, it is optimal to estimate the structural parameters r, μ, and ϕ jointly, rather than with the two-step procedure. $\square$

8 Conclusions

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

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 trough 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.

Kydland, F. and E. Prescott (1982): Time to Build and Aggregate Fluctuations, *Econometrica*, 50, 1345–1370.

Moon, H.R. (1999): A Note on Fully-Modified Estimation of Seemingly Unrelated Regression Models with Integrated Regressors, *Economics Letters*, 65, 25–31.

Moon, H.R. and B. Perron (2000): The Seemingly Unrelated Dynamic Cointegration Regression Model and Testing for Purchasing Power Parity, *mimeo*.

Nagaraj, N. & 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 M. Ogaki (1991): Seemingly Unrelated Canonical Cointegrating Regressions, *mimeo*.

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. (1995): Fully Modified Least Squares and Vector Autoregression, *Econometrica*, 63, 1023–1078.

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

Phillips, P.C.B. and V. Solo (1992): Asymptotics for Linear Processes, *Annals of Statistics*, 20, 971–1001.

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 $\delta > 0$

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

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 $(\Omega^*, \mathcal{F}^*, I\!\!P_a^*)$ with random variables $\alpha^*, \alpha_T^*, W^*, W_T^*$ that are distributionally equivalent to α, α_T, W, W_T, respectively, and $[\alpha_T^{*'}, vec(W_T^*)'] \xrightarrow{a.s.} [\alpha^{*'}, vec(W^*)']$ in $[I\!\!P_a^*]$.

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 $b = b_0$. 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

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

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 $\delta > 0$

$$\liminf_{T \to \infty} \inf_{\|b - b_0\| \geq \delta} (Q_T^*(b) - Q_T^*(b_0))$$

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

The second equality and the inequality hold because for $b's$ outside of the δ-neighborhood of b_0, $q_T(b) \neq 0$ for all T. Since W_T^* converges almost surely to a non-singular matrix by Assumption 2, $\|W_T^*q_T(b)\|^2 > 0$ if T is large.

The Cauchy-Schwarz inequality implies that

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

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

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

Proof of Lemma 1

$\Gamma_T = D_T R G \Lambda_T^{-1}$. 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

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

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

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

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

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 $l \ne k$. 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. $\square$

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

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

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

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

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

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

Thus,

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

and therefore

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

which implies the desired result.

Part (ii): Follows from Theorem 3(a) in Andrews (1999). $\square$

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 $\mathbb{E}[\theta\phi'|\eta, W] = 0$, 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 $\mathcal{M}$:

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

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

$$J_T = \alpha_T' W_T' W_T \alpha_T - 2\alpha_T' W_T' W_T \Gamma_T \hat{s}_T + \hat{s}_T' \Gamma_T' W_T' W_T \Gamma_T \hat{s}_T + o_p(1). \qquad (35)$$

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 (35), 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. $\square$