Optimal Instrumental Variables Estimation for ARMA Models

By Guido M. Kuersteiner[1]

Massachusetts Institute of Technology

In this paper a new class of Instrumental Variables estimators for linear processes and in particular ARMA models is developed. Previously, IV estimators based on lagged observations as instruments have been used to account for unmodelled MA(q) errors in the estimation of the AR parameters. Here it is shown that these IV methods can be used to improve efficiency of linear time series estimators in the presence of unmodelled conditional heteroskedasticity. Moreover an IV estimator for both the AR and MA parts is developed. Estimators based on a Gaussian likelihood are inefficient members of the class of IV estimators analyzed here when the innovations are conditionally heteroskedastic.

Keywords: ARMA, conditional heteroskedasticity, instrumental variables, efficiency lowerbound, frequency domain.

JEL Classification: C13, C22

[1] Address: MIT Dept. of Economics, 50 Memorial Drive, Cambridge, MA 02142, USA. Phone: (617) 253 2118, Fax: (617) 253 1330, Email: gkuerste@mit.edu.

1. Introduction[2]

This paper considers instrumental variables (IV) estimators for linear time series models. Efficient estimation in this framework has been studied by Hayashi and Sims (1983), Stoica, Söderström and Friedlander (1985; 1987a,b) and Hansen and Singleton (1991, 1996). In these papers efficient estimation of autoregressive roots under the presence of moving average errors has been analyzed. The moving average part of the model is not estimated but rather treated as a nuisance parameter. The class of instruments is restricted to linear functions of past observations. It is also assumed in this literature that the innovations are conditionally homoskedastic. Instrumental variables methods using overidentifying restrictions have also been applied to autoregressive (AR) models in the context of missing observations by Chen and Zadrozny (1998).

Here it is shown that the same type of IV estimators based on linear functions of past observations can be used to improve efficiency of estimators for linear time series models in the presence of unmodelled conditional heteroskedasticity. A consequence of the results of this paper is that standard estimators of linear process models based on Gaussian pseudo maximum likelihood (PML) functions are inefficient generalized method

[1]This paper is partly based on results in my Ph.D. dissertation at Yale University, 1997. I wish to thank Peter C.B. Phillips for continued encouragement and support. I have also benefited from comments by Donald Andrews, Jinyong Hahn, Jerry Hausman, Whitney Newey, Oliver Linton, Chris Sims and participants of the econometrics workshop at Yale and the University of Pennsylvania. I have also received very helpful comments from an associate editor and two referees that lead to a substantial improvement of the paper. All remaining errors are my own. Financial support from an Alfred P. Sloan Doctoral Dissertation Fellowship is gratefully acknowledged.

of moments (GMM) estimators if the innovations are conditionally heteroskedastic. This means in particular that ordinary least squares (OLS) estimators for autoregressive models of order p (AR(p)) are inefficient GMM estimators if the innovations are heteroskedastic.

In Kuersteiner (1999) a feasible efficient IV estimator for the AR(p) model with martingale innovations satisfying some additional moment restrictions is developed. In this paper we extend the previous results in two directions. First, the form of the optimal instrument is analyzed under more general assumptions about the innovation process, relaxing some of the restrictions on the fourth moments of the innovations that were previously maintained. Second, the class of models is extended to include general ARMA(p,q) processes that require minimization of a nonlinear criterion function. While Kuersteiner (1999) is mainly concerned with the semiparametric implementation of the IV procedure we focus on identification issues which are the most important aspect in extending the previous results to nonlinear models.

In addition, the paper extends the current literature in two directions. First, IV estimation is extended to general autoregressive-moving average (ARMA) models when the innovations are conditionally heteroskedastic. Second, for the class of IV estimators with linear instruments the paper derives exact functional forms of optimal filters of the type developed in Hansen and Singleton (1991) for a simpler estimation problem. It is shown how the filters depend on fourth order cumulants of the innovation distribution and the impulse response function of the underlying process. This formulation allows to give exact conditions on the distribution of the error process under which optimal instrumental variables estimators are feasible. A detailed analysis of the properties of the optimal weight matrix is provided.

The results in this paper are presented for the case of martingale difference innovations driving the linear process. Alternatively similar formulas with the same efficiency implications could be obtained under the weaker assumption of white noise innovations. In this case the space of permissible instruments is generated by all linear combinations of past observations and the efficiency bounds would be identical to the bounds of Hansen (1985) and Hansen, Heaton and Ogaki (1988). In the case of martingale difference innovations Hansen's bounds are based on a larger class of instruments and are therefore tighter than the bounds obtained here.

The main technical difficulty in extending previous procedures to the estimation of the moving average case lies in the consistency proof. We give a general characterization of instrument processes that lead to consistent estimators. We then establish that the optimal instrument satisfies these criteria.

In this paper we do not focus on implementation issues. For most parts of the analysis it is assumed that the optimal instrument is known a priori. It is clear that in practice a procedure for estimating the weight matrix is needed. In Kuersteiner (1997) such a feasible procedure is developed under stronger assumptions on the martingale difference innovations. If these assumptions are satisfied then the procedures developed in Kuersteiner (1997) can be directly applied to the present context. Explicit formulas are provided for this case. We also discuss feasible versions of the optimal procedure under the more general conditions analyzed in this paper without giving proofs of feasibility. In this case the feasible estimator depends on a bandwidth parameter. Monte Carlo simulations for this case are reported to give some guidance in the choice of the bandwidth parameter.

The paper is organized as follows. Section 2 introduces the assumptions about the innovation sequence and specifies the inference problem. Section 3 develops an instrumental variables estimator for estimation of linear process models and proves consistency and asymptotic normality of estimators for the ARMA class. In Section 4 it is shown how to factorize the asymptotic covariance matrix of this class of instrumental variables estimators in a way to obtain a lower bound. Section 5 uses the lowerbound to derive an explicit formulation of the optimal IV estimator depending on the data periodogram and an optimal frequency domain filter. Numerical Examples for the ARMA(1,1) model and some Monte Carlo simulations are reported in Section 6. Proofs of some important lemmas are contained in Appendix A while the proofs of the results in the paper are contained in Appendix B.

2. Model Specification

The econometrician observes a finite stretch of data $\{y_t\}_{t=1}^n$ from a univariate process which is generated by the following mechanism

$$y_t = \sum_{j=0}^{\infty} c(\beta, j)\varepsilon_{t-j} \tag{1}$$

for a given $\beta = \beta_0 \in \mathbb{R}^d$ and $c(\beta, j) : \mathbb{R}^d \times \mathbb{N} \to \mathbb{R}$. The parameter β_0 is unknown but the functions $c(., j)$ are known. We define the lag polynomial $C(\beta, L) = \sum_{j=0}^{\infty} c(\beta, j)L^j$ where L is the lag operator and impose the identifying restriction $c(\beta, 0) = 1$.

The innovations ε_t are assumed to be a univariate martingale difference sequence. The martingale difference property imposes restrictions on the fourth order cumulants. These restrictions can be conveniently summarized by defining the following function

$$\sigma(s,r) = \begin{cases} E\left(\varepsilon_t^2 \varepsilon_{t-|s|} \varepsilon_{t-|r|}\right) & \text{if } r \neq s \\ E\left(\varepsilon_t^2 \varepsilon_{t-s}^2\right) - \sigma^4 & \text{if } r = s \end{cases} \quad \text{for } r, s \in \{0, \pm 1, \pm 2, ...\} \tag{2}$$

where $\sigma^4 = \left(E\varepsilon_t^2\right)^2$. It should be emphasized that $\sigma(s,r)$ is equal to the fourth order cumulant for $s, r > 0$. We assume that we have a probability space $(\Omega, \mathcal{F}, P)$ with a filtration $\mathcal{F}_t$ of increasing σ-fields such that $\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}\ \forall t$. The doubly infinite sequence of random variables $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ generates the filtration $\mathcal{F}_t$ such that $\mathcal{F}_t = \sigma(\varepsilon_t, \varepsilon_{t-1}, ...)$. The assumptions on $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ are summarized as follows:

Assumption A1. *(i) ε_t is strictly stationary and ergodic, (ii) $E(\varepsilon_t \mid \mathcal{F}_{t-1}) = 0$ almost surely, (iii) $E\left(\varepsilon_t^2\right) = \sigma^2 > 0$, (iv) $\sum_{r=1}^{\infty} \sum_{s=1}^{\infty} |\sigma(s,r)| = B < \infty$, (v) $E\left(\varepsilon_t^2 \varepsilon_{t-s}^2\right) > \underline{\alpha}$ some $\underline{\alpha} > 0$ for all s.*

Remark 1. *Assumption A1(ii) could be relaxed to $E\varepsilon_t\varepsilon_s = 0$ for $t \neq s$ at the cost of slightly more complicated expressions for the optimal instruments. Assumption (iii) guarantees that ε_t has a nondegenerate distribution. Assumption (iv) limits the dependence in higher moments by imposing a summability condition on the fourth cumulants. The assumption is needed to prove invertibility of the infinite dimensional weight matrix of the optimal GMM estimator. Assumption (v) together with (iii) rules out degenerate joint distributions of ε_t and ε_{t-s}.*

Remark 2. *It can be checked that processes in the autoregressive conditionally het-*

eroskedastic family such as ARCH, GARCH, EGARCH as well as stochastic volatility models satisfy the assumptions, provided that $E\varepsilon_t^4 < \infty$. It is well known from Milhoj (1985) or Nelson (1990) that this condition is satisfied only if additional restrictions limiting the temporal dependence of conditional variances and/or the innovation distribution are imposed.

Assumption (A1) implies that ε_t^2 is strictly stationary and ergodic and therefore covariance stationary. It should be emphasized that no assumptions about third moments are made. In particular this allows for skewness in the error process.

For the special case of an ARMA(p,q) process, the lag polynomial has the familiar rational form

$$C(\beta, z) = \frac{\theta(z)}{\phi(z)} \tag{3}$$

with $\theta(z) = 1-\theta_1 z - \ldots - \theta_q z^q$ and $\phi(z) = 1-\phi_1 z - \ldots - \phi_p z^p$ and $\beta' = (\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)$. Let $g_{yy}(\beta, \lambda) = \left|C(\beta, e^{i\lambda})\right|^2$ where $|z| = (zz^*)^{1/2}$ for $z \in \mathbb{C}$ and z^* is the complex conjugate of z. Under Assumption (A1), the spectrum of y_t is given by $f_{yy}(\beta, \lambda) = \frac{\sigma^2}{2\pi} g_{yy}(\beta, \lambda)$.

Further restrictions on $C(\beta, e^{i\lambda})$ are needed to insure identification of the model and for consistency and asymptotic normality of the estimators. The necessary assumptions are discussed in Hannan (1973), Dunsmuir and Hannan (1976), and Deistler, Dunsmuir and Hannan (1978). As shown in these articles, a careful distinction between convergence of the parameters in $c(\beta, j)$ and the structural form parameters is needed. Consistency proofs typically establish convergence in the pointwise topology. An identification condition is then needed to obtain convergence in the quotient topology.

Some of the results of this paper are presented for the general formulation $C(\beta, z)$.

At some points however a specialization to the ARMA case is made in order to obtain sharper results. This is especially the case for the consistency proof. In that case abstract high level assumptions can be made precise for the specific functional form of the ARMA model.

In the general case the functions $c(\beta, j) \in C(\left[\mathbb{R}^d \times \mathbb{N}\right], \mathbb{R})$ are restricted to satisfy the following additional constraints.

Assumption B1. *Let $C(\beta, z) = \sum_{j=0}^{\infty} c(\beta, j) z^j$. The parameter space Θ is a closed subset of Θ' defined by $\Theta' = \{\beta \in \mathbb{R}^d \left| |C(\beta, z)|^{-2} \neq 0 \text{ for } |z| \leq 1, \; |C(\beta, z)|^2 \neq 0 \text{ for } |z| \leq 1\right.\}$ where we assume that Θ' is open in $\mathbb{R}^d$ and has a compact closure denoted by $\overline{\Theta}$. Assume $\beta_0 \in \Theta$. The coefficients $c(\beta, j)$ are twice continuously differentiable in $\beta \in \Theta$ for all j and $c(\beta, 0) = 1$. We require for $\beta \in \Theta$ that $\sum_{j=0}^{\infty} |j| \, |c(\beta, j)| < \infty$ and $\sum_{j=0}^{\infty} |j| \left| \frac{\partial}{\partial \beta} c(\beta, j) \right| < \infty$.*

Assumption B2. *For all $\beta, \beta_0 \in \Theta$, $g_{yy}(\beta_0, \lambda) \neq g_{yy}(\beta, \lambda)$ whenever $\beta \neq \beta_0$ for some subsets $L \subset [-\pi, \pi]$ with nonzero Lebesgue measure.*

Assumption B3. *For a neighborhood U of β_0, $U \subset \Theta_0$, $\partial^2 g_{yy}(\beta, \lambda)/\partial\beta\partial\beta'$ is continuous in $\lambda \in [-\pi, \pi]$ and $\beta \in U$.*

Remark 3. *Assumption (B1) implies that the functions $g_{yy}(\beta, \lambda)$ and $\partial g_{yy}(\beta, \lambda)/\partial\beta$ are Lipschitz continuous. The Lipschitz condition also implies that $g_{yy}^{-1}(\beta, \lambda)$ is Lipschitz continuous on Θ and therefore that $\frac{\partial}{\partial \beta} \ln g_{yy}(\beta, \lambda)$ is Lipschitz continuous on Θ.*

Remark 4. *Assumption (B1) is stronger than C2.2 in Dunsmuir (1979) where on the other hand conditional homoskedasticity is assumed. The stronger summability restrictions are needed to justify approximations of the instruments based on the innovation sequence.*

Remark 5. *Assuming Θ to be compact is of little practical importance and is commonly done in the time series literature. See for example Hosoya and Taniguchi (1982), Kabaila (1980), Taniguchi (1983).*

The assumptions specified here are sufficient to identify the parameters β in $C(\beta, e^{i\lambda})$. For specific functional forms of $C(\beta, e^{i\lambda})$ the assumptions can be made more explicit. A leading example is the ARMA model where the identifiable subset of $\mathbb{R}^d$ can be described more accurately. The following Assumption is equivalent to the previous assumptions for the case of an ARMA model.

Assumption B4. *Let $C(\beta, z) = \theta(z)/\phi(z)$. The parameter space Θ is a closed subset of Θ' defined by $\Theta' = \{\beta \in \mathbb{R}^d \,|\phi(z) \neq 0$ for $|z| \leq 1,\ \theta(z) \neq 0$ for $|z| \leq 1, \theta(z), \phi(z)$ have no common zeros, $\theta_q \neq 0$ or $\phi_p \neq 0\}$. Assume $\beta_0 \in \Theta$.*

Remark 6. *Deistler, Dunsmuir and Hannan (1978) show that Θ defined in Assumption (B4) satisfies the conditions of Assumption (B1). It is easy to show that all ARMA models in Θ satisfy the summability and differentiability requirements of (B1).*

In the following analysis of the IV estimator results will first be obtained for the general linear process case. It will then be shown that high level assumptions needed for these results are satisfied for the case when Assumptions (B1-B3) are specialized to (B4).

3. Instrumental Variables Estimators

In this section a class of instrumental variables estimators is introduced. The instruments are constructed from linear filters of lagged innovations ε_t. An alternative, equivalent formulation would be to allow for linear filters of the observable process y_t. Estimators of this form have been proposed by Hayashi and Sims (1983), Stoica, Söderström and Friedlander (1985, 1987a,b) and Hansen and Singleton (1991). Efficiency of these procedures is achieved by exploiting all the moment conditions of the form $E\varepsilon_t\varepsilon_s = 0$ for $t \neq s$. The innovations ε_t are functions of the observable data $g(y_t, y_{t-1}, ..., \beta_0) = g_t(\beta_0) = \varepsilon_t$. If we stack a finite number m of innovations in $\varepsilon_t^m = [\varepsilon_{t-1}, ..., \varepsilon_{t-m}]'$ then a standard GMM estimator minimizes the population equivalent of $E\left(g(\beta)\varepsilon_t^m\right)' \Omega_m^{-1} E\left(g(\beta)\varepsilon_t^m\right)$ where Ω_m is a suitable weight matrix. The first order conditions of this problem are given by $E\left(\frac{\partial g(\beta)}{\partial \beta'}\varepsilon_t^m\right)' \Omega_m^{-1} E\left(g(\beta)\varepsilon_t^m\right) = 0$. We will later use the notation $P_m = E\left(\frac{\partial g(\beta)}{\partial \beta'}\varepsilon_t^m\right)$ where P_m is a $m \times d$ dimensional matrix. In the context of linear instrumental variables estimators P_m corresponds to the matrix of covariances between regressors and instruments. Using this notation we can now set up an equivalent problem which is to solve the set of equations $E\left(g(\beta)P_m'\Omega_m^{-1}\varepsilon_t^m\right) = 0$. Letting $z_t = P_m'\Omega_m^{-1}\varepsilon_t^m$ be the d dimensional vector of instruments then leads to the formulation of an equivalent, exactly identified IV estimator that solves the population equivalent of $E\left(g(\beta)z_t\right) = 0$. Clearly, the two formulations of the problem are equivalent as far as their first order asymptotic efficiency properties are concerned.

The advantage of this transformation lies in the fact that in our context of ARMA models the matrix P_m can be estimated $\sqrt{n}$ consistently irrespective of the size of m. It is shown in Kuersteiner (1997, 1999) that under additional restrictions on the weight matrix Ω_m it is possible to set $m = n$ or in other words to let the number of moment conditions grow at the same rate as the sample size. This is not the case for the usual implementation of GMM where $E\left(g(\beta)\varepsilon_t^m\right)$ is replaced by a sample average.

We now discuss the estimation problem for a general class of linear instruments. Introduce the space of absolutely summable sequences l^1 such that $x \in l^1$ if $\sum |x_j| < \infty$ for $x = \{x_j\}_{j=1}^{\infty}$. Define the set $\mathcal{A}$ of sequences of vectors $a_j \in \mathbb{R}^d$ such that

$$\mathcal{A} = \left\{ a = \{a_j\}_{j=1}^{\infty} : a_j \in \mathbb{R}^d, \left\{[a_j]_k\right\}_{j=1}^{\infty} \in l^1 \text{ for all } 1 \leq k \leq d \right\}$$

where $[.]_k$ denotes the k-th element of a vector. We define $z_t(\omega) : \Omega \to \mathbb{R}^d$ for all t as

$$z_t = \sum_{k=1}^{\infty} a_k \varepsilon_{t-k} \text{ a.s.}$$

for $a \in \mathcal{A}$, a fixed. The instruments satisfy the orthogonality condition

$$E\left[\left(C^{-1}(\beta_0, L)y_t\right) z_t\right] = 0 \tag{4}$$

since $C^{-1}(\beta_0, L)y_t = \varepsilon_t$ from (1). The estimator based on this condition is constructed in the time domain. If $C^{-1}(\beta_0, L)$ is of infinite order, as is the case for $MA(q)$ models, a sample analog to (4) needs to be based on an approximation. Such an approximation can be conveniently analyzed in the frequency domain. It should be stressed however that the estimator is set up in time domain. Let the expansion of the polynomial $C^{-1}(\beta, L)$ be $C^{-1}(\beta, L) = \sum_{j=0}^{\infty} \tilde{c}_j^{\beta} L^j$. The sample analog of the moment restriction is then given by

$$G_n(\beta, a) = \frac{1}{n} \sum_{t=1}^{n} z_t \sum_{j=0}^{t-1} \tilde{c}_j^{\beta} y_{t-j} = 0 \tag{5}$$

for $a \in \mathcal{A}$. The criterion function is indexed by a to emphasize the fact that each choice of an instrument results in a different estimator. From (4) we see that z_t has to be approximated as well. Discussion of this issue will be delayed until Section 5 where an optimal instrument is considered. For the time being it is therefore assumed that z_t is known.

In the frequency domain the analog of (4) is

$$\int_{-\pi}^{\pi} C^{-1}(\beta_0, e^{-i\lambda}) f_{yz}(\lambda) d\lambda = 0$$

where $f_{yz}(\lambda) = \sum_{j=-\infty}^{\infty} \gamma_{yz}(j) e^{-i\lambda j}$ and $\gamma_{yz}(j) = E y_t z_{t-j}$. We set

$$G(\beta, a) = (2\pi)^{-1} \int_{-\pi}^{\pi} C^{-1}(\beta, e^{-i\lambda}) f_{yz}(\lambda) d\lambda.$$

Note that $f_{yz}(\lambda)$ typically is a complex vector valued function $f_{yz}(\lambda) : [-\pi, \pi] \to \mathbb{C}^d$. Also note that $\int_{-\pi}^{\pi} C^{-1}(\beta, e^{-i\lambda}) f_{yz}(\lambda) d\lambda$ is real valued.

We introduce discrete Fourier transforms of the data defined as $\omega_{n,y}(\lambda) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} y_t e^{-it\lambda}$ and for the instrument as $\omega_{n,z}(\lambda) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} z_t e^{-it\lambda}$. The cross periodogram is $I_{n,yz}(\lambda) = \omega_{n,y}(\lambda)\omega_{n,z}(-\lambda)$. It is easy to check that $G_n(\beta, a)$ defined in (5) is identical to

$$G_n(\beta, a) = (2\pi)^{-1} \int_{-\pi}^{\pi} C^{-1}(\beta, e^{-i\lambda}) I_{n,yz}(\lambda) d\lambda.$$

We follow Hansen (1982) in defining the estimator β_n. Unless otherwise stated all conditions are for $a \in \mathcal{A}$, a fixed

Assumption C1. *The sequence of estimators* $\beta_n \in \mathbb{R}^d$ *is defined by*

$$\beta_n = \arg\min_{\beta\in\Theta} \|G_n(\beta, a)\|^2. \tag{6}$$

Assumption C2. *Let the sets* $B_k(\beta_0)$ *for* $k = 1, 2, ...$ *form a countable local base*[3] *around* β_0*. The sets* $B_k(\beta_0)$ *can be taken as the set of balls with rational radius centered at* β_0*. Let* $z_t = \sum_{k=1}^{\infty} a_k \varepsilon_{t-k}$ *a.s. where* ε_{t-k} *satisfies Assumption (A1). Let* $\mathcal{A}^* \subseteq \mathcal{A}$ *be the set of all sequences* $\{a_k\}_{k=1}^{\infty}$ *such that*

$$\mathcal{A}^* = \left\{ a \in \mathcal{A} \,\middle|\, \inf_{\beta \in B_k(\beta_0)^C \cap \Theta} \|G(\beta_n, a)\| > 0 \text{ for } k = 1, 2, ... \right\}$$

where $B_k(\beta_0)^C$ *are the complements of* $B_k(\beta_0)$*. Assume that* $\mathcal{A}^* \neq \emptyset$*.*

Remark 7. *Assumption (C1) is the definition of the estimator. We show in the consistency proof that* $\|G_n(\beta_n, a)\|^2 = 0$ *almost surely is implied by the assumptions on* G_n*. (C2) is a familiar identification condition for global identification. Assumption (C2) specializes (B2) to the case of IV estimators. The difference is that identification of an entire class of estimators indexed by* a *needs to be guaranteed. Identification depends on the choice of the instrument or* $a \in \mathcal{A}$ *and does not hold for all* $a \in \mathcal{A}$ *(see Remark 10 for an example). We therefore define the subset* $\mathcal{A}^*$ *of instruments that satisfy the identification condition.*

[3] *A collection of open subsets* $\mathcal{B}$ *of a space* X *is called a base if for each open set* $O \subset X$ *and each* $x \in O$ *there is a set* $B \in \mathcal{B}$ *such that* $x \in B \subset O$*. A collection* $\mathcal{B}_x$ *of open sets containing a point* x *is called a local base at* x *if for each open set* O *containing x there is a* $B \in \mathcal{B}_x$ *such that* $x \in B \subset O$*. Every metric space has a countable base at each point (see Royden (1988), p. 175).*

This imposes restrictions on z_t or a. A complete description of the set $\mathcal{A}^*$ is possible for a given parametric class $C(\beta, z)$. A characterization will be given for the ARMA case.

Lemma 3.1. *Assume (A1), (B1-B3), (C1-C2). Let $z_t = \lim_{m\to\infty} A_m' \varepsilon_t^m$ a.s. with $A_m' = [a_1, ..., a_m]$, $\{a_k\}_{k=1}^{\infty} \in \mathcal{A}^*$ and $\varepsilon_t^m = [\varepsilon_{t-1}, \ldots, \varepsilon_{t-m}]'$. Then $\beta_n \to \beta_0$ almost surely.*

Consistency of the IV estimator depends both on restrictions on the parameter space and the instruments z_t. Assumption (C2) restricts the class of allowable instruments. The conditions given are necessarily high level without further knowledge regarding the function $C(\beta, L)$. For practical purposes it is however important to characterize the set of instruments $\mathcal{A}^*$ leading to consistent estimators. In the case of an ARMA(p,q) model it is possible to give conditions on the sequences $a \in \mathcal{A}^*$. This is done in the next proposition.

Proposition 3.2. *Assume $C(\beta, L) = \theta_0(L)/\phi_0(L)$ is an ARMA(p,q) lag operator and the parameter space Θ satisfies Assumption (B4). Let*

$$S = \left\{ x = [x_1, ...] \in l^2 : \phi_0(L)x = 0 \text{ for } x_j, j > d, [x_1, ..., x_d]' = \kappa, \kappa \in \mathbb{R}^d \right\}$$

be the set of solutions to the difference equation $\phi_0(L)x = 0$ with d initial conditions κ. Define $\ker A' = \left\{x \in l^2 : A'x = 0\right\}$ for $A = [a_1, ...]'$ and $a \in \mathcal{A}$. Let $a \in \mathcal{A}$ with $A_d = [a_1, ..., a_d]'$ where $d = p + q$. If $p = 0$ and A_d is nonsingular then $a \in \mathcal{A}^*$. If $0 < p$, $A = [a_1, ...]'$ is of full column rank and $\ker A' \cap S = 0$ then $a \in \mathcal{A}^*$.

Remark 8. *Proposition (3.2) shows that ARMA models can be consistently estimated by instrumental variables techniques provided that the instruments satisfy the specified restrictions.*

Remark 9. *The usual conditions for consistency in IV estimation are* $E\varepsilon_t z_t = 0$ *and* $E\frac{\partial g_t(\beta)}{\partial\beta}z_t$ *is of full rank. For linear models these two conditions are equivalent to Assumption (C2). In our context C2 may hold even if* $E\frac{\partial g_t(\beta)}{\partial\beta}z_t$ *is of reduced rank. An example is an ARMA(1,1) model with instruments* z_t *defined by* $a_1 = [1,0]'$, $a_2 = \left[-\theta_0^{-1}, 1\right]'$ *and* $a_3 = \left[0, -\theta_0^{-1}\right]'$ *with* $a_j = 0$ *for* $j > 3$*. Then* A *is of full row rank,* $\ker A' \cap S = 0$ *but* $E\frac{\partial g_t(\beta)}{\partial\beta}z_t$ *is of reduced rank. On the other hand* $E\frac{\partial g_t(\beta)}{\partial\beta}z_t$ *being full rank implies* A' *to be of full row rank and* $A^{\perp} \cap S = 0$*. In other words the identification conditions given here are weaker than standard conditions would imply.*

Remark 10. *If the instruments are replaced by* $a_1 = [1,0]'$, $a_2 = \left[-\phi_0^{-1}, 1\right]$ *and* $a_3 = \left[0, -\phi_0^{-1}\right]$ *in the ARMA(1,1) example then* $\ker A' \cap S \neq 0$ *which shows that* A' *being full row rank is not sufficient for identification and clearly* $\mathcal{A}^* \neq \mathcal{A}$*.*

We now state additional assumptions that are sufficient to establish a result for the limiting distribution of $\sqrt{n}(\beta_n - \beta_0)$. Introduce the notation $\dot{\eta}(\beta,\lambda) = \partial \ln C(\beta, e^{-i\lambda})/\partial\beta$ and $b_k = (2\pi)^{-1}\int \dot{\eta}(\beta_0,\lambda)e^{ik\lambda}d\lambda$. It follows immediately that $b_{-k} = 0$ and $b_0 = 0$. Let $l_a(\lambda) = \sum_{k=1}^{\infty} a_k e^{-i\lambda k}$ and define the matrices $P_m^{'} = [b_1, ..., b_m]$, $A_m' = [a_1, ..., a_m]$ and

$$\Omega_m = \begin{bmatrix} \alpha_{1,1} & \cdots & \alpha_{1,m} \\ \vdots & \ddots & \vdots \\ \alpha_{m,1} & \cdots & \alpha_{m,m} \end{bmatrix}. \tag{7}$$

where

$$\alpha_{s,r} = \begin{cases} \sigma(s,r) & \text{if } s \neq r \\ \sigma(r,r) + \sigma^4 & \text{if } s = r \end{cases}. \tag{8}$$

It is easy to check that $\lim_m P_m' A_m = (2\pi)^{-1} \int \dot{\eta}(\beta_0, \lambda) l_a(-\lambda)' d\lambda$. Also note that using our earlier notation $\lim_m \sigma^2 P_m' A_m = E \frac{\partial g_t(\beta)}{\partial \beta} z_t$. The following additional conditions are needed to prove the existence of a limiting distribution of β_n.

Assumption D1. *Define* $\mathcal{A}^{**} \subseteq \mathcal{A}$ *as* $\mathcal{A}^{**} = \left\{ a \in \mathcal{A} \,\middle|\, \det \int \dot{\eta}(\beta_0, \lambda) l_a(-\lambda)' d\lambda \neq 0 \right\}$. *Assume that* $\mathcal{A}^* \cap \mathcal{A}^{**} \neq \emptyset$.

Remark 11. *Assumption (D1) guarantees that there is an instrument* $a \subseteq \mathcal{A}$ *that satisfies both the identification condition C2 and* $\det \int \dot{\eta}(\beta_0, \lambda) l_a(-\lambda)' d\lambda \neq 0$. *Assumption (D1) corresponds to Assumption 3.4 in Hansen (1982). Note that for linear models there is no difference between the identification assumptions and Assumption (D1) while this is not true for nonlinear models. In general C2 does not imply D1.*

The next lemma shows that for the ARMA model D1 does indeed imply C2 while we have seen in Remark 9 that the reverse is not true.

Lemma 3.3. *Assume* $C(\beta, L) = \theta_0(L)/\phi_0(L)$ *is an ARMA(p,q) lag operator and the parameter space* Θ *satisfies Assumption (B4) and that* $a \in \mathcal{A}$. *Then* $a \in \mathcal{A}^{**} \Rightarrow a \in \mathcal{A}^*$.

The limiting distribution of the instrumental variables estimator is stated in the next theorem. For notational efficiency define $\lim_{m\to\infty} \sigma^{-4}(P_m' A_m)^{-1} A_m' \Omega_m A_m (A_m' P_m)^{-1} = \sigma^{-4}(P' A)^{-1} A' \Omega A (A' P)^{-1}$. This notation will be justified in the next section in terms of operators on infinite dimensional spaces.

Theorem 3.4. *Assume (A1), (B1-B3), (C1, C2) and (D1). Let* $z_t = \lim_{m\to\infty} A_m' \varepsilon_t^m$ *with* $A_m' = [a_1, ..., a_m]$, $\{a_k\}_{k=1}^{\infty} \in \mathcal{A}^* \cap \mathcal{A}^{**}$ *and* $\varepsilon_t^m = [\varepsilon_{t-1}, \ldots, \varepsilon_{t-m}]'$. *Then the estimator* *defined by* $\beta_n = \arg\min \|G_n(\beta_n, a)\|^2$ *has a limiting distribution given by*

$$\sqrt{n}(\beta_n - \beta_0) \xrightarrow{d} N(0, \sigma^{-4}(P^{'}A)^{-1}A^{'}\Omega A(A^{'}P)^{-1})$$

Remark 12. *If* β_n *is obtained from minimizing a Gaussian PML criterion function then the asymptotic covariance matrix is* $\sigma^{-4}(P^{'}P)^{-1}P^{'}\Omega P(P^{'}P)^{-1}$. *Such an estimator therefore corresponds to an IV estimator where* $A = P$. *This shows that Gaussian estimators have the interpretation of inefficient IV or GMM estimators when the innovations are conditionally heteroskedastic.*

The main result of the paper will now be developed in two steps. We first obtain a lower bound for the covariance matrix

$$\sigma^{-4}(P^{'}A)^{-1}A^{'}\Omega A(A^{'}P)^{-1} \tag{9}$$

in the next section. This lower bound is then used to construct an optimal instrumental variables estimator.

4. Covariance Matrix Lowerbound

Finding a lower bound for (9) poses certain technical difficulties having to do with the infinite dimensional nature of the instrument space. We investigate the properties of the fourth order cumulant matrix Ω_m, first by holding m fixed and then by looking at a related infinite dimensional problem. In particular we establish that the infinite dimensional operator Ω, associated with Ω_m in a way to be defined, has a well behaved inverse.

Invertibility of Ω_m for all m is not enough to show that Ω is invertible. We briefly review the theory of invertible operators (see Gohberg and Goldberg, 1980), p.65. For two Banach spaces B_1 and B_2 denote the set of bounded linear operators mapping B_1 into B_2 by $L(B_1, B_2)$. Then $A \in L(B_1, B_2)$ is invertible if there exists an operator $A^{-1} \in L(B_2, B_1)$ such that $A^{-1}Ax = x$ for all $x \in B_1$ and $AA^{-1}y = y$ for all $y \in B_2$. Let $\ker A = \{x \in B_1 : Ax = 0\}$ and $\operatorname{Im} A = \{Ax : x \in B_1\}$. Then A is invertible if $\ker A = \{0\}$ and $\operatorname{Im} A = B_2$.

Following Hanani, Netanyahu and Reichaw (1968) we now choose B_1, B_2 as linear spaces whose points are sequences of real numbers denoted by $x = \{x_1, x_2, ...\}$ and $y = \{y_1, y_2, ...\}$. Define the norm $\|x\|_2 = (\sum_i |x_i|^2)^{1/2}$. Then B is the space of all sequences that are bounded under the $\|.\|_2$ norm and is denoted by l^2. An operator $A : l^2 \mapsto l^2$ is defined by the infinite dimensional matrix $A = (a_{i,j}), i, j = 1, 2, ...$ such that $y = Ax \in l^2$ for all $x \in l^2$. The operator A can be defined element by element as $y_i = \sum_j^\infty a_{i,j}x_j$ for all i. The operator A is invertible if the only solution to $Ax = 0$ is $x = \{0, 0, ...\}$ and $\operatorname{Im} A = l^2$. Note that l^2 is a Hilbert space with inner product $\langle x, y\rangle = \sum_j^\infty x_j y_j$. From Theorem 11.4 in Gohberg and Goldberg (1980) it follows $KerA^\perp = \operatorname{Im} A$ for a self adjoint operator A. It is thus enough to show $KerA = 0$ for $A : l^2 \to l^2$, A selfadjoint.

Consider now the following infinite dimensional operator associated with Ω_m. Define the operator Ω component-wise by its image for all $x \in l^2$ by $b_i = \lim_{m\to\infty} \sum_j^m \alpha_{i,j}x_j$ where $\alpha_{i,j}$ is defined in (8). In other words Ω is the infinite dimensional matrix such that any left upper corner sub matrix of dimension $m \times m$ has the same elements as Ω_m.

Lemma 4.1. *Let Ω_m be defined as in (7). Then Ω_m^{-1} exists for all m, $\Omega \in L(l^2, l^2)$ and*

Ω^{-1} *exists.*

Proof. See Appendix B ■

Remark 13. *The fact that the image of* Ω *is square summable, i.e.* $\Omega x \in l^2$*, depends on the summability properties of* $\sigma(k,l)$*. The interpretation of the summability condition is that the instruments* ε_t *become unrelated in their fourth cumulants as the time spread between them increases.*

By the Closed Graph Theorem (Gohberg and Goldberg (1980), Theorem X.4.2) it also follows that Ω^{-1} is bounded, i.e., $\left\|\Omega^{-1}\right\| = \sup_{\|x\|_2 \leq 1} \left\|\Omega^{-1}x\right\|_2 < \infty$. Thus $\sup_{i,j} |\omega_{i,j}| < \infty$ where $[\Omega^{-1}]_{i,j} = \omega_{i,j}$.

Next, we need to establish properties of the matrix Ω_m^{-1} as m tends to infinity. In particular we want to establish that the inverse Ω_m^{-1} approximates Ω^{-1} as $m \to \infty$.

Lemma 4.2. *Let* Ω_m *be as defined in (7). Define* Ω_m^{-1} *such that* $\Omega_m^{-1}\Omega_m = I_m$ *and* $\Omega_m \Omega_m^{-1} = I_m$ $\forall m$*. Let*

$$\Omega_m^* = \begin{bmatrix} \Omega_m & 0 \\ 0 & \sigma^4 I \end{bmatrix} \tag{10}$$

where I *stands for an infinite dimensional identity matrix. Then* Ω_m^{*-1} *exists and* $\left\|\Omega_m^{*-1} - \Omega^{-1}\right\| \to 0$ *as* $m \to \infty$.

Proof. See Appendix B ■

Remark 14. *A consequence of the convergence of* Ω_m^{*-1} *to* Ω^{-1} *in the operator norm is that* $\left\|P'(\Omega_m^{*-1} - \Omega^{-1})\right\|_2 \to 0$ *in the* l^2 *norm. In other words* $\operatorname{Var}(z_t^m - z_t) \to 0$ *as* $m \to \infty$

where $z_t^m = P'\Omega_m^{*-1}\varepsilon_t^\infty$ and $z_t = P'\Omega^{-1}\varepsilon_t^\infty$. In this sense Lemma (4.2) provides an algorithm to approximate the infinite dimensional inverse Ω^{-1}. For m fixed, the finite dimensional inverse of Ω_m is computed and used to construct Ω_m^{*-1}. The resulting instrument then converges in a mean squared sense.

We define the d dimensional product of sequence spaces $l_d^2 = l^2 \times ... \times l^2$. Define the infinite dimensional matrix $P = [b_1, ...]'$ by stacking elements of the sequence $\{b_k\}_{k=1}^\infty \in l_d^2$. Introduce notation for the reverse operation of extracting a sequence from the rows of a matrix by defining $b(P) := \{b_k\}_{k=1}^\infty$. Define the matrix $\Xi = (P'\Omega^{-1}P)^{-1}$.

Using this notation we can state our next theorem which establishes a lower bound for the covariance matrix (9).

Theorem 4.3. *For any $a \in \mathcal{A}$ let $A' = [a_1, ...]$ and P and Ω as previously defined. If $a(A) \in \mathcal{A}^{**}$ then the matrix $(P'A)^{-1}A'\Omega A(A'P)^{-1}$ satisfies*

$$(P'A)^{-1}A'\Omega A(A'P)^{-1} - (P'\Omega^{-1}P)^{-1} \geq 0$$

where ≥ 0 stands for positive semi-definite.

Proof. See Appendix B ■

Remark 15. *If $a \in \mathcal{A}^* \cap \mathcal{A}^{**}$ then $(P'A)^{-1}A'\Omega A(A'P)^{-1}$ is the asymptotic covariance matrix of an estimator based on a. However, it is important to point out that the lower-bound is for IV estimators in the class of all instruments which are linear functions of the innovation process and have an innovation filter in $\mathcal{A}^{**}$. The construction of the lower* *bound does not involve consistency restrictions for the instruments. In order to construct an efficient estimator in practice it has to be established that the optimal instrument does in fact satisfy consistency restrictions.*

5. Optimal Instrumental Variables Estimators

Theorem (4.3) immediately leads to the construction of an efficient IV estimator. The optimal instrument is determined by the linear filter $A' = P'\Omega^{-1}$. It is not a priori true that the optimal filter also results in a consistent estimator. However, for important parametric examples such as the ARMA class this is indeed the case.

Theorem 5.1. *Assume $C(\beta, L) = \theta(L)/\phi(L)$ and the parameter space Θ satisfies Assumption (B4). If $A^{'} = P'\Omega^{-1}$ then the sequence $a = a(P'\Omega^{-1})$ defined by the rows of A satisfies $a \in \mathcal{A}^* \cap \mathcal{A}^{**}$.*

Theorem (5.1) together with Theorem (3.4) and Theorem (4.3) establish that the IV estimator for the ARMA model constructed with instruments based on $A' = P'\Omega^{-1}$ achieves a lowerbound of the same type as in Hansen and Singleton (1991) but under the weaker martingale difference sequence assumptions on ε_t detailed in Assumption (A1). This result is summarized in the following Corollary.

Corollary 5.2. *Assume (A1), $C(\beta, L) = \theta_0(L)/\phi_0(L)$ is an ARMA(p,q) lag operator, the parameter space Θ satisfies Assumption (B4) and $z_t = \lim_{m\to\infty} P_m^{'}\Omega_m^{-1}\varepsilon_t^m$. Then*

$$\sqrt{n}(\beta_n - \beta_0) \xrightarrow{d} N\left(0, \sigma^{-4}\left(P^{'}\Omega^{-1}P\right)^{-1}\right)$$

Feasible versions of the optimal IV procedure have to be based on approximations of the optimal instrument z_t. Such approximations proceed in two steps. First, the infinite number of unobserved innovations is replaced by a finite number of proxies based on the observed sample and constructed from $\hat{\varepsilon}_t = y_t - \sum_{j=1}^{t-1} c(\beta_0, j) y_{t-j}$ for $t = 1, ..., n$ where $\hat{\varepsilon}_1 = y_1$. This leads to a pseudo feasible estimator $\tilde{\beta}$ based on the instruments $\hat{z}_t = \sum_{j=1}^{t-1} a_j \hat{\varepsilon}_{t-j}$. A fully feasible estimator denoted by $\tilde{\beta}(\hat{a})$ is obtained by substituting β_0 for a first stage consistent estimator $\hat{\beta}$ in the construction of $\hat{\varepsilon}_t$ and by replacing the weights a_j by estimated quantities $\hat{a}_j$. Gaussian PMLE procedures which are consistent but inefficient in our context can be used to generate first stage estimators $\hat{\beta}$.

The empirical analog of the moment restriction for the pseudo feasible estimator becomes

$$\tilde{G}_n(\tilde{\beta}, a) = \frac{1}{n} \sum_{t=1}^{n} \hat{z}_t \sum_{j=0}^{t-1} \tilde{c}_j^{\beta} y_{t-j} = 0. \tag{11}$$

It is shown in the proof of Corollary (5.4) that $\sup_{\beta \in \Theta} \left| \tilde{G}_n^F(\beta, a) - \tilde{G}_n(\tilde{\beta}, a) \right| = O_p(n^{-1})$ where

$$\tilde{G}_n^F(\beta, a) = (2\pi)^{-1} \int_{-\pi}^{\pi} C^{-1}(\beta, e^{-i\lambda}) h(\beta_0, \lambda) I_{n,yy}(\lambda) d\lambda. \tag{12}$$

Here $I_{n,yy}(\lambda)$ is the data periodogram and the filter $h(\lambda) : [-\pi, \pi] \to \mathbb{C}^d$ is defined as $h(\beta_0, \lambda) = l_\psi(-\lambda) C^{-1}(\beta_0, e^{i\lambda})$ with

$$l_\psi(\lambda) = \sum_{j=1}^{\infty} a_j e^{-i\lambda j}.$$

The coefficients of the optimal instrument are given by

$$a_j = \sum_{k=1}^{\infty} b_k \omega_{kj}$$

where b_k is the Fourier coefficient of the derivative of the log spectral density of y_t and ω_{kj} is the kj-th entry of the inverse Ω^{-1}. The b_k coefficients have simple interpretations in special parametric models. In the case of an $AR(p)$ model for example they are equivalent to the impulse response function and can therefore be computed easily. It can also be noted that the Gaussian estimators are obtained by setting $a_j = b_j$.

It is shown in Kuersteiner (1997) that a sufficient condition for the validity of the approximation $\hat{z}_t$ for z_t is that the coefficients of the instruments satisfy

$$\sum_{j=1}^{\infty} j \left| [a_j]_k \right| < \infty \text{ for } k = 1, ..., d. \tag{13}$$

The following lemma shows that under strengthened summability restrictions on the fourth order cumulants Condition (13) is satisfied for the optimal instrumental variables estimator of the $ARMA(p, q)$ model. The Corollary summarizes the result that based on Theorem (5.3) the instruments can be approximated without affecting the first order asymptotic distribution.

Theorem 5.3. *Assume $C(\beta, L) = \theta(L)/\phi(L)$ and the parameter space Θ satisfies Assumption (B4). Strengthen Assumption (A1v) to $\sum_{r=1}^{\infty} \sum_{s=1}^{\infty} s \left| \sigma(s, r) \right| = B < \infty$. By symmetry this implies $\sum_{r=1}^{\infty} \sum_{s=1}^{\infty} r \left| \sigma(s, r) \right| = B < \infty$. If $A' = P' \Omega^{-1}$ then $a = a(P' \Omega^{-1})$ satisfies (13).*

Proof. *See Appendix B* ■

Corollary 5.4. *Assume that the conditions of Theorem 5.3 hold together with Assumption A1. Let* $\tilde{\beta} \in \Theta$ *satisfy* $\tilde{G}_n(\tilde{\beta}, a) = o_p(n^{-1/2})$. *Then* $\sqrt{n}(\tilde{\beta} - \beta_n) = o_p(1)$.

Proof. See Appendix B ■

Feasible versions of the optimal estimator are then obtained by replacing $\tilde{G}_n(\beta, a)$ by $\tilde{G}_n(\beta, \hat{a})$ where in $\tilde{G}_n(\beta, \hat{a})$ we replace $h(\beta_0, \lambda)$ by $\hat{l}_\psi(\lambda)C^{-1}(\hat{\beta}, e^{i\lambda})$ and $\hat{\beta}$ is a consistent first stage estimate. The challenging part is to estimate $\hat{l}_\psi(\lambda)$ consistently. Here we only discuss a special case where Ω_m is diagonal. This case with the additional restrictions $\alpha_{j,k} = 0$ for $j \neq k$ on the moments of ε_t has been analyzed in Kuersteiner (1997) in the context of estimating an AR(p) model. The restriction $\alpha_{j,k} = 0$ is satisfied for GARCH processes with symmetric innovation distributions and stochastic volatility models. Under these circumstances it is possible to estimate $\hat{l}_\psi(\lambda)$ consistently without the need to introduce bandwidth parameters controlling the number of instruments. The simplification comes from the fact that in that particular case Ω^{-1} is diagonal such that $a_j = b_j/\alpha_{j,j}$. An estimate of the optimal instrument is obtained from

$$\hat{z}_t = \sum_{j=1}^{t-1} \hat{b}_j / \hat{\alpha}_{j,j} \hat{\varepsilon}_{t-j} \tag{14}$$

where $\hat{b}_j = (2\pi)^{-1} \int_{-\pi}^{\pi} \dot{\eta}(\hat{\beta}, \lambda) e^{i\lambda j} d\lambda$. To define $\hat{\alpha}_{j,j}$ introduce a truncation sequence $d_n = cn^{-1/2+\nu}$ for some $0 < \nu < 1/2$ and some constant $c > 0$. Then define $\hat{\alpha}_{j,j} = \hat{\alpha}^*_{j,j}$ if $\hat{\alpha}^*_{j,j} > d_n$ and $\hat{\alpha}_{j,j} = d_n$ otherwise where $\hat{\alpha}^*_{j,j} = \frac{1}{n} \sum_{t=1+j}^{n} \hat{\varepsilon}_t^2 \hat{\varepsilon}_{t-j}^2$. In simulation experiments this truncation parameter does not affect the outcome and in practice $d_n = 1$ works well. In order to prove the result formally we need to require that additional higher order moments exists.

Assumption E1. *Let* $c_{\varepsilon\ldots\varepsilon}(t_1,\ldots,t_{k-1})$ *be the k-th order cumulant of the error process* ε_t *and assume that*

$$\sum_{t_1}\cdots\sum_{t_{k-1}}(1+|t_j|)\left|c_{\varepsilon\ldots\varepsilon}(t_1,\ldots,t_{k-1})\right|<\infty,\ \textit{for all } j=1,...,k-1 \textit{ and } k=2,3,..,8.$$

Theorem 5.5. *Assume (A1),* $\alpha_{j,k}=0$ *for* $j\neq k$ *and (E1),* $C(\beta,L)=\theta_0(L)/\phi_0(L)$ *is an ARMA(p,q) lag operator, the parameter space* Θ *satisfies Assumption (B4). Let* $\tilde{\beta}(\hat{a})\in\Theta$ *satisfy* $\tilde{G}_n(\tilde{\beta},\hat{a})=o_p(n^{-1/2})$ *with* $\hat{z}_t$ *defined in (14). Then* $\sqrt{n}(\tilde{\beta}(\hat{a})-\beta_n)=o_p(1)$.

Proof. See Appendix B ■

In the more general case where $\alpha_{j,k}\neq 0$ for $j\neq k$ the elements ω_{kj} can be estimated from a sample analog of the approximation matrix Ω_m^* defined in (10). Denote by ω_{kj}^* the k,j-th element of the inverse Ω_m^{*-1} and let $z_t^*=\sum_{j=1}^{t-1}\sum_{k=1}^{t-1}b_k\omega_{kj}^*\hat{\varepsilon}_{t-j}$ be the instrument based on the approximation and denote by $\tilde{G}_n(\beta,a^*(m))$ the corresponding criterion function where $a^*(m)$ indicates that instruments depend on m. Let $\tilde{\beta}^*=\arg\min\left\|\tilde{G}_n(\beta,a^*(m))\right\|^2$. From $\hat{\Omega}_m=n^{-1}\sum_{t=m+1}^{n}\hat{\varepsilon}_t^2\hat{\varepsilon}_t^m\hat{\varepsilon}_t^{m\prime}$ we form the $n\times n$ matrix

$$\hat{\Omega}_m^{*-1}=\begin{bmatrix}\hat{\Omega}_m^{-1} & 0\\ 0 & I_{n-m}\left(\hat{\sigma}^4\right)^{-1}\end{bmatrix}$$

and obtain estimates $\hat{\omega}_{kj}^*$ of ω_{kj} where $\hat{\omega}_{kj}^*$ is the k,j-th element of $\hat{\Omega}_m^{*-1}$. Here I_{n-m} is an $n-m$ dimensional identity matrix and $\hat{\sigma}^4=\left(\frac{1}{n}\sum_{t=1}^{n}\hat{\varepsilon}_t^2\right)^2$. We then form the truncated estimate of the a_j coefficients by setting $\hat{a}_j(m) = \sum_{k=1}^{n} \hat{b}_k \hat{\omega}_{kj}^*$. The feasible estimator resulting from minimizing $\left\| \tilde{G}_n(\beta, \hat{a}(m)) \right\|^2$ is denoted by $\tilde{\beta}^*(\hat{a}(m))$. In the next section we use Monte Carlo experiments to evaluate some selection rules for m.

6. Numerical Examples

In this section we take the case of an ARMA(1,1) model and analyze its asymptotic efficiency properties for the case when ε_t is an ARCH(1) process. We also show plots of the likelihood contour. We finally conduct a small Monte Carlo experiment to explore the performance of the IV procedures in finite samples.

The model we investigate is a univariate process y_t defined as the stationary solution to

$$y_t = \phi y_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}. \tag{15}$$

The innovations ε_t are an ARCH(1) process generated by $\varepsilon_t = u_t h_t^{-1/2}$ and $h_{t+1} = \gamma_0 + \gamma_1 \varepsilon_t^2$ where $u_t \tilde{} N(0,1)$ is an iid sequence of random variables. We assume that $\gamma_0 > 0$ and set it to .1 for all numerical calculations. In this case $\beta = [\phi, \theta]'$ and $\Theta' = \{\beta | \, |\phi| < 1, |\theta| < 1\}$. The parameter α_1 satisfies $0 \leq \gamma_1 < \sqrt{1/3}$ which guarantees that $E\varepsilon_t^4 < \infty$. It is shown in Milhoj (1985) that $\sigma^4 = (\gamma_0/1 - \gamma_1)^2$, $\text{Cov}(\varepsilon_t^2, \varepsilon_{t-i}^2) = 2\gamma_0^2 \gamma_1^i / [(1-\gamma_1)^2(1-3\gamma_1^2)]$ for $i > 0$ and $E(\varepsilon_t^2 \varepsilon_{t-k} \varepsilon_{t-j}) = 0$ for $j \neq k$. It follows that $\Omega_m = \text{diag}(\alpha_{1,1}, ..., \alpha_{m,m})$ where $\alpha_{k,k} = \text{Cov}(\varepsilon_t^2, \varepsilon_{t-k}^2) + \sigma^4$.

The polynomial $C(\beta, L) = (1 - \theta L) / (1 - \phi L)$ has log derivative $\dot{\eta}(\beta, \lambda) = [e^{-i\lambda}/(1 - \phi e^{-i\lambda}), -e^{-i\lambda}/(1 - \theta e^{-i\lambda})]'$. It follows that $b_j = \left[\phi^{j-1}, -\theta^{j-1}\right]'$.

For $\gamma_1 = 0$ it follows that $\text{Cov}(\varepsilon_t^2, \varepsilon_{t-k}^2) = 0$ and $\alpha_{k,k} = \sigma^4$ for $k > 0$ since $\varepsilon_t = \gamma_0 u_t$ is iid in this case. The optimal instrument then is $z_t = \lim_{m\to\infty} P'_m \varepsilon_t^m$ a.s. and the limiting distribution of β_n is $\sqrt{n}(\beta_n - \beta_0) \to N(0,(P'P)^{-1})$ which is the same as the limiting distribution of the maximum likelihood estimator. Figure 1 shows a contour plot of the function $-\arctan \|G(\beta,a)\|^2$ which has the shape of a long flat valley with a unique global minimum at β_0 where $-\arctan \|G(\beta,a)\|^2 = 0$.

Figure 1

When $\gamma_1 > 0$ the optimal instrument is given by

$$z_t = \left[\sum_{k=1}^{\infty} \phi_0^{k-1}/\alpha_{k,k}\varepsilon_{t-k}, -\sum_{k=1}^{\infty} \theta_0^{k-1}/\alpha_{k,k}\varepsilon_{t-k}\right]'.$$

In Figure 2 we plot the asymptotic relative efficiency $\sigma^2(\phi_{PML})/\sigma^2(\phi_{IV})$ of the PML estimator which reaches .7 around $\phi = -.9$ and improves as ϕ gets closer to zero. The overall shape and magnitude of the efficiency improvement of IV over PML is very similar to the pure AR(1) model analyzed in Kuersteiner (1999). In Figure 3 we plot the same efficiency gains for the parameter θ when ϕ is varied over $[-1,1]$. Note that here the shape of the efficiency curve is much less symmetric.

Figure 2,3

We now report a Monte Carlo experiment to investigate how well the asymptotic efficiency properties of the IV procedure hold in finite samples. We generate samples of size

$n = 2^k$ for $k = 7, 8, ..., 10$ from Model (15) with ARCH(1) innovations.

Starting values are $y_0 = 0$, $h_0 = 0$ and $\varepsilon_0 = 0$. In each sample the first 500 observations are discarded to eliminate dependence on initial conditions. Small sample properties are evaluated for different values of $\beta,\ \gamma_1 \in [0, 1)$. For GARCH processes it was shown in Milhoj (1985) and Bollerslev (1986) that asymptotic normality established in previous chapters only obtains for a subset of values for γ_1. Nevertheless, simulation results are reported for parametrizations outside this range in order to analyze the robustness of the proposed IV procedure to departures from the assumptions.

The parameter β is estimated in several different ways. The PML estimator based on a Gaussian likelihood is denoted by $\hat{\beta}_{OLS}$ and was computed using the NAG Fortran routine G13AFF, Mark 18. This routine forces the parameters to be in Θ. The optimal instrumental variables estimator is obtained from the consistent first stage estimator $\hat{\beta}_{OLS}$ as $\tilde{\beta}(\hat{a}) = \arg\min \left\| \tilde{G}_n(\beta, \hat{a}) \right\|^2$. We also compute the more general IV estimator $\tilde{\beta}(\hat{a}(m)) = \arg\min \left\| \tilde{G}_n(\beta, \hat{a}(m)) \right\|^2$ where m is set to $5, 10, \sqrt{n}, n^{2/5}$. Both problems are solved numerically with the constrained nonlinear optimization routine *nag_nlp_sol* of the NAG fl90 (release 3) library of Fortran 90 subroutines using $\hat{\beta}_{OLS}$ as a starting value.

We can compare the asymptotic gains reported in Figure 1 with the empirical efficiency of the estimators $\tilde{\beta}(\hat{a})$ and $\tilde{\beta}(\hat{a}(m))$ based on $1,000$ replications for sample sizes ranging from 128 to 1024. The results are summarized in Tables 1-16. The results for $\gamma_1 = .5$ in Tables 1-8 can be directly compared to the theoretical efficiency gain calculations. For the estimator $\tilde{\beta}(\hat{a})$ based on the diagonal weight matrix, reported in Tables 1-4, the sample sizes needed to achieve efficiency gains are relatively large with 1024 observations. As expected, the strongest gains are achieved for $|\phi| > .5$. The moving average parameter θ is generally less well estimated by this method than the autoregressive parameter ϕ. For parameter values near the point of non-identification where $\theta = \phi$ the IV procedure performs relatively poorly compared to Gaussian pseudo likelihood estimators. Overall the performance of the IV estimator is superior for large sample sizes and at well chosen points in the parameter space. For smaller sample sizes and at less favorable points in the parameter space the Gaussian estimators tend to dominate but IV performs reasonable even under these circumstances.

In Tables 5-8 we report the performance of the unrestricted IV procedures $\tilde{\beta}(\hat{a}(m))$. These procedures perform better than $\tilde{\beta}(\hat{a})$ and achieve gains over the Gaussian estimator at favorable points in the parameter space even in small samples of 128 observations. In larger samples they almost uniformly dominate Gaussian estimators, with the exception of points very close to $\theta = \phi$. As has to be expected from the theoretical calculations the estimators for θ again fare slightly worse than for ϕ. The overall preferred choice for m is $m = n^{2/5}$.

In Tables 9-16 we report results for $\gamma_1 = .9$. Strictly speaking these experiments are not covered by the theoretical results because $\gamma_1 > \sqrt{1/3}$ implies that $E\varepsilon_t^4$ does not exist. The IV estimators nevertheless perform well and in fact achieve even larger efficiency gains. This is especially true for $\tilde{\beta}(\hat{a}(m))$ which achieves variance reductions of up to 60% in large samples. There is however an outlier problem for this estimator. This is particularly evident from Table 14 where the mean absolute error for $\tilde{\beta}(\hat{a}(m))$ is improved over β_{OLS} while the same is not always true for the variance measure. Looking at the inter quantile range however shows that the distribution of $\tilde{\beta}(\hat{a}(m))$ is more concentrated than for β_{OLS} at least as long as $|\phi| > .5$.

The simulations show that even if the true weight matrix Ω is diagonal it is better not to impose this restriction and to estimate the full non-diagonal weight matrix. The recommendation that can be given based on the simulation results is to use $\tilde{\beta}(\hat{a}(m))$ with $m = n^{2/5}$. The choice of m reported here is clearly limited by the small scale of the Monte Carlo experiment and no claim is made that this choice is optimal from a theoretical point of view. The simulations however also seem to indicate that the performance of $\tilde{\beta}(\hat{a}(m))$ is not very sensitive to the choice of m at least in the cases considered.

7. Conclusions

In this paper we have analyzed the instrumental variables estimator for stationary linear process models and ARMA models in particular. It was shown that a GMM estimator based on the infinite number of moment conditions $E\varepsilon_t\varepsilon_s$ can be constructed. The maintained hypothesis in this paper is that the innovations ε_t are martingale difference sequences. The overidentified GMM estimator can be conveniently represented in the form of an exactly identified IV estimator. It is shown that under the additional restrictions used in Kuersteiner (1999) a feasible version of the GMM estimator is available even when the number of moment conditions used in estimation is the same as the number of observations.

The procedures proposed in this paper are optimal in the class of GMM estimators based on instruments that are linear in the observed data. While inclusion of nonlinear instruments is possible in principle and would improve asymptotic efficiency it creates difficult issues of implementation. Linear instruments have the advantage of being approximately ordered by their time lag in terms of their importance. This ordering is lost when nonlinear instruments are included and a more sophisticated selection procedure has to be implemented to make any such procedure feasible.

The proposed procedures are developed for univariate time series models. Extensions to the class of vector ARMA models could be proved along similar lines. If $h_0(L)y_t = g_0(L)\varepsilon_t$ in the notation of Dunsmuir and Hannan (1976) then IV procedures would be constructed based on the moment conditions $E\varepsilon_t \otimes \varepsilon_s$. Consistent estimates then have to be based on instruments that are not orthogonal to any process $g^{-1}(L)h(L)y_t$ in the VARMA class.

Simulation evidence for a univariate ARMA(1,1) model shows that the procedures do achieve, sometimes significant, efficiency gains in finite samples. This is especially true when the model is well identified and the autoregressive parameter is larger than .5 in absolute value.

A. Appendix - Lemmas

Lemma A.1. *Under Assumption (A1) for each $m \in \{1, 2...\}$, m fixed, the vector*

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} [\varepsilon_t \varepsilon_{t-1}, ..., \varepsilon_t \varepsilon_{t-m}] \Rightarrow N(0, \Omega_m)$$

with Ω_m defined in (7).

Proof. We note that individually all the terms $\varepsilon_t \varepsilon_{t-k}$ with $k \geq 1$ are martingale difference sequences (mds). Now define $Y_t' = [\varepsilon_t \varepsilon_{t-1}, ..., \varepsilon_t \varepsilon_{t-m}]$. Then it is enough to show that for all $\ell \in \mathbb{R}^m$ such that $\ell' \ell = 1$ we have $\frac{1}{\sqrt{n}} \sum \ell' \tilde{Y}_t \Rightarrow N(0,1)$ where now $\tilde{Y}_t = \Omega_m^{-1/2} Y_t$ and $\Omega_m = E Y_t Y_t'$. Note that $\ell' \tilde{Y}_t$ is a mds and a martingale CLT (see Hall and Heyde , 1980, Theorem 3.2, p.52) can be applied to the sum $\sum_t Y_{nt} = \frac{1}{\sqrt{n}} \sum_t \tilde{Y}_t$. ■

Lemma A.2. *Let ε_t satisfy Assumption (A1). Then $\nexists \alpha \in l^2$, $\alpha \neq 0$ such that $\sum_{i=0}^{\infty} \alpha_i \varepsilon_{t-i} = 0$ a.s.*

Proof. If $\exists \alpha \in l^2$, $\alpha \neq 0$ such that $\sum \alpha_i \varepsilon_{t-i} = 0$ *a.s* assume without loss of generality $\alpha_1 \neq 0$. If $\alpha_i = 0$ for all $i = 2, 3, ...$ then $\sum \alpha_i \varepsilon_{t-i} = 0$ *a.s.* is trivially contradicted. Now assume $\alpha_i \neq 0$ for at least one $i = 2, 3, ...$ such that $\varepsilon_{t-1} = -\alpha_1^{-1} \sum_{i=2}^{\infty} \alpha_i \varepsilon_{t-i}$ *a.s.* But then $E(\varepsilon_{t-1} | \mathcal{F}_{t-2}) = -\alpha_1^{-1} \sum_{i=2}^{\infty} \alpha_i \varepsilon_{t-i}$ *a.s.* so that $E(\varepsilon_{t-1} | \mathcal{F}_{t-2}) \neq 0$ with positive probability. This contradicts the martingale difference assumption. ■

Lemma A.3. *Let $I_{n,yz}(\lambda) = \omega_{n,y}(\lambda) \omega_{n,z}(-\lambda)$. $I_{n,\varepsilon\varepsilon}(\lambda)$ is the periodogram of $\{\varepsilon_1, \ldots, \varepsilon_n\}$. Assume ε_t satisfy Assumption (A1) and that $y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ with $\psi(\lambda) = \sum_{j=0}^{\infty} \psi_j e^{-i\lambda j}$*

such that $\sum_{j=0}^{\infty} |j| \left|\psi_j\right| < \infty$. *Also let* $z_t = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j}$ *with* $a \in \mathcal{A}$. *Let* $\varsigma(.)$ *be a function on* $[-\pi, \pi] \to \mathbb{C}$ *with absolutely summable Fourier coefficients* $\{c_k, -\infty < k < \infty\}$ *such that* $\varsigma(\lambda) = \sum_{j=-\infty}^{\infty} c_j e^{-i\lambda j}$. *Then for any* $\eta, \epsilon > 0$

$$P\left(\sqrt{n}(2\pi)^{-1}\left|\int_{-\pi}^{\pi} I_{n,yz}(\lambda)\varsigma(\lambda)\,d\lambda - \int_{-\pi}^{\pi} I_{n,\varepsilon z}(\lambda)\,C(\beta_0,\lambda)\varsigma(\lambda)\,d\lambda\right| > \eta\right) < \epsilon$$

as $n \to \infty$.

Proof. First an expression for $R_n(\lambda) = I_{n,yz}(\lambda) - I_{n,\varepsilon z}(\lambda)\,\psi(\lambda)$ is obtained. Using

$$\omega_{n,y}(\lambda) = \psi(\lambda)\omega_{n,\varepsilon}(\lambda) + n^{-1/2}\sum_{j=0}^{\infty}\psi_j e^{-i\lambda j} U_{nj}(\lambda) \tag{16}$$

where $U_{nj}(\lambda) = \sum_{t=1-j}^{n-j} \varepsilon_t e^{-i\lambda t} - \sum_{t=1}^{n} \varepsilon_t e^{-i\lambda t}$ leads to

$$R_n(\lambda) := I_{n,yz}(\lambda) - \psi(\lambda) I_{n,\varepsilon z}(\lambda) = \omega_z(-\lambda)\, n^{-1/2}\sum_{j=0}^{\infty}\psi_j e^{-i\lambda j} U_{nj}(\lambda)$$

Note that $(2\pi)^{-1}\int R_n(\lambda)\varsigma(\lambda)\,d\lambda = n^{-1}\sum_{k=1}^{\infty}\sum_{l=0}^{\infty}\sum_{r=1}^{l}\sum_{m=-\infty}^{\infty} a_k\psi_l c_m \varepsilon_{r+j+m-k}\left(\varepsilon_{r-l} - \varepsilon_{n-l+r}\right)$.

Then using the Markov inequality it is enough to consider

$$E\sqrt{n}\left|(2\pi)^{-1}\int_{-\pi}^{\pi} R_n(\lambda)\varsigma(\lambda)\,d\lambda\right| \le 2\sup_k \alpha_k^{1/2} n^{-1/2}\sum_{k=1}^{\infty}\sum_{l=0}^{\infty}\sum_{m=-\infty}^{\infty}\left|a_k\psi_l c_m\right| |l| \to 0$$

since the last term is bounded from $\sum_{k=1}^{\infty}|a_k| < \infty$ and $\sum_{l=0}^{\infty}|l|\,|\psi_l| < \infty$. ■

Lemma A.4. *Let* $I_{n,\varepsilon z}(\lambda) = \omega_{n,\varepsilon}(\lambda)\omega_{n,z}(-\lambda)$. *Assume* ε_t *satisfy Assumption* (*A*1) *and* *let* $z_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ *with* $a \in \mathcal{A}$*. Then for any* $\ell \in \mathbb{R}^d$ *such that* $\ell'\ell = 1$,

$$n^{1/2}(2\pi)^{-1} \int_{-\pi}^{\pi} \ell' I_{n,\varepsilon z}(\lambda)\, d\lambda \xrightarrow{d} N\left(0, \sum_{l=1}^{\infty}\sum_{k=1}^{\infty} \alpha_{k,l} \ell' a_k a_l' \ell\right).$$

Proof. First note that $(2\pi)^{-1} \int_{-\pi}^{\pi} I_{n,\varepsilon z}(\lambda)\, d\lambda = n^{-1} \sum_{t=1}^{n} \varepsilon_t z_t$ such that $E n^{1/2}(2\pi)^{-1} \int_{-\pi}^{\pi} I_{n,\varepsilon z}(\lambda)\, d\lambda = 0$. It also follows that $\varepsilon_t z_t$ is a martingale difference sequence. However $z_t = \sum_{k=1}^{\infty} a_k \varepsilon_{t-k}$ such that a direct application of Lemma (A.1) is not possible.

For a fixed m we introduce $z_t^m = \sum_{k=1}^{m} a_k \varepsilon_{t-k}$ such that $\omega_{n,z}^m(\lambda) = n^{-1/2} \sum_{t=1}^{k} z_t^m e^{-i\lambda k}$ and $I_{n,\varepsilon z}^m(\lambda) = \omega_{n,\varepsilon}(\lambda) \omega_{n,z}^m(-\lambda)$. From Billingsley (1968, Theorem 4.2) it is enough to show that for all $\epsilon > 0$,

$$\lim_{m\to\infty} \limsup_{n\to\infty} P\left\{\left| n^{1/2} \int_{-\pi}^{\pi} \ell'(I_{n,\varepsilon z}^m(\lambda) - I_{n,\varepsilon z}(\lambda)) d\lambda \right| \geq \epsilon \right\} = 0$$

where

$$n^{1/2}(2\pi)^{-1} \int_{-\pi}^{\pi} \ell'(I_{n,\varepsilon z}^m(\lambda) - I_{n,\varepsilon z}(\lambda)) d\lambda = n^{-1/2} \sum_{t=1}^{n} \sum_{k>m}^{\infty} \ell' a_k \varepsilon_t \varepsilon_{t-k}$$

Since $E a_k \varepsilon_t \varepsilon_{t-k} = 0$ it is enough to consider

$$n^{-1} E\left(\sum_{t=1}^{n} \sum_{k>m}^{\infty} \ell' a_k \varepsilon_t \varepsilon_{t-k}\right)^2 = n^{-1} \sum_{t=1}^{n} \sum_{j>m}^{\infty} \sum_{k>m}^{\infty} \ell' a_k a_l' \ell \alpha_{k,l} \to 0 \text{ as } m \to \infty.$$

Applying Lemma (A.1) then gives the result. ■

Lemma A.5. *Assumption (B1) implies that* $\tilde{c}(\beta, j) = (2\pi)^{-1} \int C^{-1}(\beta, \lambda) e^{i\lambda j} d\lambda$ *satisfies* $\sum j \left|\tilde{c}(\beta, j)\right| < \infty$ *for all* $\beta \in \Theta$.

Proof. Since $C^{-1}(\beta, \pi) = C^{-1}(\beta, -\pi)$ it follows from integration by parts that

$$|\tilde{c}(\beta, j)| = j^{-1} \left| (2\pi)^{-1} \int \partial C^{-1}(\beta, \lambda)/\partial\lambda e^{i\lambda j} d\lambda \right|. \tag{17}$$

From $\partial C^{-1}(\beta, \lambda)/\partial\lambda = C^{-2}(\beta, \lambda)\partial C(\beta, \lambda)/\partial\lambda$ and the fact that $C(\beta, \lambda)$ satisfies $\sum j |c(\beta, j)| < \infty$ it follows that $\partial C^{-1}(\beta, \lambda)/\partial\lambda$ has absolutely summable Fourier coefficients. Rearranging (17) and summing over j then gives the result.

Lemma A.6. *Assume (A1), (B1-B3), (C1-C2). Let* $z_t = \lim_{m\to\infty} A_m^{'} \varepsilon_t^m$ *with* $A_m^{'} = [a_1, ..., a_m]$, $\{a_k\}_{k=1}^{\infty} \in \mathcal{A}^*$ *and* $\varepsilon_t^m = [\varepsilon_{t-1}, \ldots, \varepsilon_{t-m}]^{'}$. *Then for any convergent sequence* $\beta_n \in \Theta$ *with* $\beta_n \to \beta \in \Theta$ *there exists an event* E *with probability one such that for all outcomes in* E, $G_n(\beta_n, a) \to G(\beta, a)$.

Proof. Without loss of generality assume that z_t takes values in $\mathbb{R}$. Let $Ey_t z_s = \gamma_{yz}(t-s)$, and $cum(y_t, z_s, y_q, z_r) = c_{yyzz}(t-s, t-q, t-r)$. Then, from Assumption (A1) and the proof of Theorem 2.8.1 in Brillinger (1981) it follows that $\sum_j \left|\gamma_{yz}(j)\right| < \infty$ and $\sum_{s,q,r} |c_{yyzz}(s, q, r)| < \infty$. For each $\epsilon > 0$ there exists an $n_0 < \infty$ and $\delta > 0$ such that $\|\beta_n - \beta\| < \delta$ and

$$\sup_{\left\|\beta' - \beta\right\| < \delta} \sup_{\lambda} \left|C^{-1}(\beta', \lambda) - C^{-1}(\beta, \lambda)\right| < \epsilon$$

for $n > n_0$ by continuity of $C^{-1}(\beta, \lambda)$ at $\beta \in \Theta$. For $\beta^{'}$ such that $\left\|\beta' - \beta\right\| < \delta$ the lag polynomial $C^{-1}(\beta', z)$ has an expansion with coefficients $\tilde{c}(\beta', j)$ such that $\sum_{j=1}^{\infty} j \left|\tilde{c}(\beta', j)\right| < \infty$. We will use the short hand notation $\tilde{c}' = \tilde{c}(\beta', j)$. Let $X_n(\beta) = G_n(\beta, a) - EG_n(\beta, a)$ and define $X_n = \sup_{\left\|\beta' - \beta\right\| < \delta} \left|X_n(\beta')\right|$. Since $EG_n(\beta', a) \to G(\beta', a)$ and $\left|G(\beta', a) - G(\beta, a)\right| \leq \epsilon \int |f_{yz}(\lambda)| d\lambda$ uniformly for all β' such that $\left\|\beta' - \beta\right\| < \delta$ it is enough to show that $X_n \to 0$

almost surely. Thus letting $X_n(j) = \sum_{t=1}^{n-j} y_t z_{t+j} - \gamma_{yz}(-j)$

$$X_n \le \sup_{\left\|\beta'-\beta\right\|<\delta} n^{-1} \sum_{j=0}^{n} \left|\tilde{c}_j'\right| |X_n(j)| \le K_0 n^{-1} \left(\sum_{j=0}^{n} j^{-2} \left| X_n(j) \right|^2 \right)^{1/2}$$

where $K_0 = \sup_{\left\|\beta'-\beta\right\|<\delta} \left(\sum_{j=0}^{\infty} \left|\tilde{c}_j'\right| j \right)$. We consider

$$EX_n^2 \le K_0^2 n^{-2} \sum_{j=0}^{n} j^{-2} \left(EX_n(j)^2 \right).$$

Since

$$EX_n(j)^2 \le n \sum_{k=-\infty}^{\infty} \left| \gamma_{yy}(k)\gamma_{zz}(k) + \gamma_{yz}(k)\gamma_{yz}(k) \right| + n \sum_{j,k,l=-\infty}^{\infty} \left| c_{yyzz}(j,k,l) \right|$$

for all j there is a K_1such that $EX_n^2 \le K_2 n^{-1}$ where $K_2 = \frac{\pi^2}{6} K_1 K_0^2$. For $n/2 \le n_1 < n$ consider $X_{n,n_1} = \sup_{\left\|\beta'-\beta\right\|<\delta} \left| X_n(\beta') - X_{n_1}(\beta') \right|$ such that

$$\begin{aligned} X_{n,n_1} \quad &\le \quad K_0 (n - n_1) (n n_1)^{-1} \left(\sum_{j=0}^{n_1} j^{-2} \left| X_{n_1}(j) \right|^2 \right)^{1/2} \\ &\quad + K_0 n^{-1} \left(\sum_{j=0}^{n} j^{-2} \left(\sum_{t=\max(n_1-j,1)}^{n-j} y_t z_{t+j} - \gamma_{yz}(-j) \right)^2 \right)^{1/2}. \end{aligned}$$

Now

$$K_0^2 (n - n_1)^2 (n n_1)^{-2} E \sum_{j=0}^{n_1} j^{-2} \left| X_{n_1}(j) \right|^2 \le K_2 (n - n_1) n^{-2}$$

and

$$K_0^2 n^{-2} \sum_{j=0}^{n} j^{-2} E \left(\sum_{t=\max(n_1-j,1)}^{n-j} y_t z_{t+j} - \gamma_{yz}(-j) \right)^2 \leq K_2 n^{-2} (n - n_1)$$

together with $E\left(Y+Z\right)^2 \leq EY^2 + 2\left(EY^2 EZ^2\right)^{1/2} + EZ^2$ implies that $EX_{n,n_1}^2 \leq K_2 n^{-2}(n-n_1)$. It now follows from Lemma 3 in Gaposhkin (1980) that $X_n \to 0$ almost surely. Let this event be E. From $|G_n(\beta_n, a) - G_n(\beta, a)| \leq X_n$ for all $n > n_0$ the result follows. ■

B. Appendix - Proofs

Proof of Lemma 3.1 From the definition of β_n it follows that

$$0 \leq \liminf_n \|G_n(\beta_n, a)\|^2 \leq \limsup_n \|G_n(\beta_n, a)\|^2 \leq \limsup_n \|G_n(\beta_0, a)\|^2 . \qquad (18)$$

From Lemma (A.6) it follows that $G_n(\beta_0, a) \to G(\beta_0, a) = 0$ almost surely. Thus

$$\limsup_n \|G_n(\beta_n, a)\|^2 = \lim_n \|G_n(\beta_n, a)\|^2 = 0 \text{ almost surely.} \qquad (19)$$

Let E be the probability one event in Lemma (A.6). Now consider the sequence $\beta_n \in \Theta$. If β_n does not converge to β_0 then by compactness of Θ there exists a subsequence β_{n_k} such that $\beta_{n_k} \to \beta \in \Theta$. By Lemma (A.6) and Assumption (C2) $\liminf_k \left\|G_{n_k}(\beta_{n_k}, a)\right\|^2 > 0$ $a.s.$ contradicting (19). Therefore $\beta_n \to \beta_0$. ■

Proof of Proposition 3.2 We only prove that Assumption (C2) holds. We first note that $f_{yz}(\lambda) = C(\beta_0, e^{-i\lambda})l_a(-\lambda)$ where $l_a(\lambda) = \sum_{k=1}^{\infty} a_k e^{-i\lambda k}$ such that

$$G(\beta, a) = (2\pi)^{-1} \int_{-\pi}^{\pi} \psi(\beta, e^{-i\lambda}) l_a(-\lambda) d\lambda$$

with $\psi(\beta, e^{-i\lambda}) = C^{-1}(\beta, e^{-i\lambda})C(\beta_0, e^{-i\lambda})$. It is clear that $\psi(\beta_0, e^{-i\lambda}) = 1$ so that $G(\beta_0, a) = 0$.

We need to show that for $C(\beta, e^{-i\lambda}) = \theta(e^{-i\lambda})/\phi(e^{-i\lambda})$ there is no other $\beta \in \Theta$ such that $G(\beta, a) = 0$. The orthogonality conditions can be written as

$$(2\pi)^{-1} \int_{-\pi}^{\pi} (\psi(\beta, e^{-i\lambda}) - 1) l_a(-\lambda) d\lambda = 0. \tag{20}$$

We want to show that the only function $\psi(\beta, e^{-i\lambda}) - 1 : [-\pi, \pi] \to \mathbb{C}$ satisfying this condition is

$$\psi(\beta, e^{-i\lambda}) - 1 \equiv 0. \tag{21}$$

If the assumptions of Proposition (3.2) hold then the only value β for which $\psi(\beta, e^{-i\lambda}) - 1 \equiv 0$ is β_0.

Now showing that $\psi(\beta, e^{-i\lambda}) - 1 \equiv 0$ is equivalent to showing $\phi(e^{-i\lambda})\theta_0(e^{-i\lambda})/\phi_0(e^{-i\lambda}) - \theta(e^{-i\lambda}) \equiv 0$ since the polynomial $\theta(e^{-i\lambda})$ is not zero for any $\lambda \in [-\pi, \pi]$ for $\beta \in \Theta$. It is more convenient to rewrite this equation as

$$\left(\phi(e^{-i\lambda}) - \phi_0(e^{-i\lambda})\right) C(\beta_0, e^{-i\lambda}) - \left(\theta(e^{-i\lambda}) - \theta_0(e^{-i\lambda})\right) \equiv 0.$$

Here $C(\beta_0, e^{-i\lambda}) = \theta_0(e^{-i\lambda})/\phi_0(e^{-i\lambda})$ is the lag polynomial of an $ARMA(p, q)$ with a one sided Fourier expansion $\sum_{j=0}^{\infty} c_j e^{-i\lambda j}$.

For $j \geq \max(p, q+1) - p$ the coefficients c_j satisfy the well known restriction

$$c_j - \phi_{0,1} c_{j-1} - ... - \phi_{0,p} c_{j-p} = 0. \tag{22}$$

We define the infinite dimensional matrix C with p rows as $C = \left[c, [0, c']', [0_2, c']', ..., [0_{p-1}, c']'\right]$ with $c' = [c_0, c_1, ...]$ and 0_k is the k-dimensional column vector of zeros then Condition (20) has a matrix representation

$$A'C(\phi - \phi_0) - A_q'(\theta - \theta_0) = 0. \tag{23}$$

which can be written as $R\delta = 0$ where R is the $d \times d$ matrix $R = A'D$ where

$$D = \begin{bmatrix} C, & \begin{matrix} -I \\ 0 \end{matrix} \end{bmatrix}$$

with 0 an $\infty \times q$ dimensional matrix of zeros and $\delta = \beta - \beta_0$. We need to show that $\ker R = 0$ which follows if R is of full rank. We can distinguish two cases. If $p = 0$ then $R = A_q'$ and $\delta = (\theta - \theta_0)$ such that $\delta = 0$ if A_q is of full rank. If $p > 0$ then C contains p linearly independent vectors in l^2 which are also linearly independent of $[-I, 0]'$. So D has full column rank. It is a finite rank operator mapping $\mathbb{R}^d$ into the d-dimensional subspace $\operatorname{Im} D$ of l^2. Since l^2 is a Hilbert space this subspace is closed and has an orthogonal complement $(\operatorname{Im} D)^{\perp}$ (see Gohberg and Goldberg, p.205). The finite rank operator A' maps l^2 into $\mathbb{R}^d$. If $\operatorname{Im} D \cap \ker A' = 0$ then $\operatorname{Im} D = (\ker A')^{\perp}$ since $l^2 = \ker A' \oplus (\ker A')^{\perp}$

where $\oplus$ is the direct sum. Then, by theorem II.11.4 in Gohberg and Goldberg (1980), $\operatorname{Im} A' = (\ker A)^{\perp} = \mathbb{R}^d$ where the last equality is due to $\ker A = 0$ since A is of full column rank. It follows that $\{A'x | x \in \operatorname{Im} D\} = \mathbb{R}^d$ and $A'x = 0$ for $x \in \operatorname{Im} D$ if and only if $x = 0$. But this means that $\operatorname{Im} R = \mathbb{R}^d$ showing that R is of full rank.

Finally, we show that $\operatorname{Im} D$ is the space of all the solutions $x = [x_1, ...]$ to $\phi_0(L)x = 0$ for $x_j,\ j > d$ with $d = p + q$ initial conditions determining $x_1, ..., x_d$. To see this note that $c = \{c_j\}_{j=0}^{\infty}$ is the solution to $\phi_0(L)c = 0$ for $c_j,\ j > \max(p, q + 1) - p$ which has general form $c_j = \sum_{i=1}^{k} \sum_{n=0}^{r_i} \kappa_{in} j^n \xi_i^{-j}$ where $\xi_i, i = 1, ..., k$ are the distinct zeros of $\phi_0(L)$ with multiplicity r_i. The first $\max(p, q+1) - p$ coefficients c_j are determined by the $\max(p, q+1)$ boundary conditions implied by $C(\beta_0, L)$. The l^2 sequence $D\delta$ then has j-th element $\tilde{c}_j = \sum_{i=1}^{p} \delta_i c_{j-i}$ with $c_{j-m} = \sum_{i=1}^{k} \sum_{n=0}^{r_i} \sum_{l=0}^{n} \binom{n}{l} \tilde{\kappa}_{in} j^l m^{n-l} \xi_i^{-j-m}$ such that the p coefficients $\tilde{\kappa}_{in}$ of $\tilde{c}_j$ can be set arbitrarily to satisfy p initial conditions. The remaining q initial conditions can be set by appropriately choosing $\delta_{p+1}, ..., \delta_d$. ■

Remark 16. *For finite dimensional matrices it is known from Corollary 6.2 in Marsaglia and Styan (1974) that* $rank(A'D) = rank(D) - \dim(\ker A' \cap \operatorname{Im} D)$*. Our proof extends this result to finite rank operators on Hilbert spaces when* A *and* D *are of identical and full column rank.*

Proof of Lemma 3.3: Remember that $\int_{-\pi}^{\pi} \dot{\eta}(\beta_0, \lambda) l_a(-\lambda) d\lambda = A'P$ with $P = [b_1, b_2, ...]'$ where $b_k = (2\pi)^{-1} \int_{-\pi}^{\pi} \partial \ln C(\beta_0, e^{-i\lambda}) / \partial\beta e^{i\lambda k} d\lambda$. For $C(\beta_0, e^{-i\lambda}) = \theta_0(\lambda)/\phi_0(\lambda)$ we have

$$\partial \ln C(\beta_0, e^{-i\lambda}) / \partial\beta = \left[\frac{e^{-i\lambda}}{\phi_0(\lambda)}, \cdots, \frac{e^{-i\lambda p}}{\phi_0(\lambda)}, \frac{e^{-i\lambda}}{\theta_0(\lambda)}, \cdots, \frac{e^{-i\lambda q}}{\theta_0(\lambda)} \right]'.$$

Define the expansions of $\phi_0^{-1}(z) = \sum_{j=0}^{\infty} \psi_{\phi,j} z^j$ and $\theta_0^{-1}(z) = \sum_{j=0}^{\infty} \psi_{\theta,j} z^j$. The coefficients in the expansion satisfy the difference equation $\psi_{\phi,j} - \phi_{0,1}\psi_{\phi,j-1} - ... - \phi_{0,p}\psi_{\phi,j-p} = 0$ which has p linearly independent solutions. A similar expression holds for $\psi_{\theta,j}$. Set $\psi_{\phi,j} = \psi_{\theta,j} = 0$ for $j < 0$. Then $b_k = \left[\psi_{\phi,k-1}, \cdots, \psi_{\phi,k-p}, \psi_{\theta,k-1}, \cdots, \psi_{\theta,k-q}\right]'$. Any set of $d = p + q$ vectors $b_{k_1}, b_{k_2}, ..., b_{k_d}$ is linearly independent because of the linear independence of the solutions to $\phi_0(L)x = 0$ and $\theta_0(L)x = 0$ together with the requirement that $\phi(L)$ and $\theta(L)$ have no common zeros and that $\phi_p \neq 0$ or $\theta_q \neq 0$. Thus P has full column rank.

When $p > 0$ we distinguish two cases. For the case where $q = 0$, $\operatorname{Im} P = \operatorname{Im} D$ where D was defined in the proof of Proposition (3.2). This implies that $\ker P' \cap \operatorname{Im} D = 0$ since $\operatorname{Im} P = (\ker P')^{\perp}$.

When $q > 0$ then $\operatorname{Im} P = S_1$ where

$$S_1 = \left\{ x = [x_1, ...] \in l^2 : \phi_0(L)\theta_0(L)x = 0 \text{ for } x_j, j > d, [x_1, ..., x_d]' = \kappa, \kappa \in \mathbb{R}^d \right\}$$

while $\operatorname{Im} D = S$. Since $\phi_0(L)x = 0 \Rightarrow \phi_0(L)\theta_0(L)x = 0$ it follows that $\operatorname{Im} P \supset \operatorname{Im} D$. But $\operatorname{Im} P = (\ker P')^{\perp}$ and $\ker P' \cap (\ker P')^{\perp} = 0$ which implies that $\operatorname{Im} D \cap \ker P' = 0$. To see that $\operatorname{Im} P = S_1$ note that the j-th element c_j in $\operatorname{Im} P$ is $c_j = \sum_{i=1}^{p} \delta_i \psi_{\phi,j-i} + \sum_{i=1}^{q} \delta_{p+i} \psi_{\theta,j-1}$ for $\delta \in \mathbb{R}^d$. Since $\psi_{\phi,j-m} = \sum_{i=1}^{k} \sum_{n=0}^{r_i} \sum_{l=0}^{n} \binom{n}{l} \kappa_{in} j^l m^{n-l} \xi_{\phi,i}^{-j-m}$ from the general solution with a similar expression for $\psi_{\theta,j-m}$ it follows that $\delta_m \psi_{\phi,j-m} + \delta_{p+w} \psi_{\theta,j-w}$ is the general solution of a difference equation with roots $\xi_{\phi,i}$ and $\xi_{\theta,i}$ which is the same as $\phi_0(L)\theta_0(L)x = 0$. Since there are d free parameters δ_i, c_j can be made to satisfy d initial conditions.

We now show that $\mathcal{A}^{**} \subset \mathcal{A}^{*}$. First let $p = 0$. Then $a \in \mathcal{A}^{**}$ implies $rank(A'P) = q$.

Assume that $rank A < q$. Then $\dim(\operatorname{Im} A') < q$ because $\ker A \neq 0$. This contradicts $A'P$ to be of full rank. Thus A_q is of full rank and $a \in \mathcal{A}^*$.

For $p > 0$ it follows by the same argument that the row rank of A' has to be full. To show that $\ker A' \cap S = 0$ assume that $\ker A' \cap S \neq 0$. We have shown that $S = \operatorname{Im} D$ and $\operatorname{Im} D \subseteq \operatorname{Im} P$. This implies $\ker A' \cap \operatorname{Im} P \neq 0$. But then $\exists x \in \mathbb{R}^d$, $x \neq 0$ such that $Px \in \ker A'$ thus $A'Px = 0$. This contradicts $A'P$ being full rank. ■

Proof of Theorem 3.4: Let $M_n(\beta, a) = G_n(\beta, a) - G(\beta, a)$. We use a mean value expansion for $M_n(\beta, a) - M_n(\beta_0, a) = \frac{\partial}{\partial \beta'} M_n(\beta^+, a)(\beta - \beta_0)$ with $\left\|\beta^+ - \beta_0\right\| \leq \|\beta - \beta_0\|$. Then

$$\sup_{\|\beta-\beta_0\|<\delta} \|M_n(\beta, a) - M(\beta_0, a)\| \leq \delta n^{-1} \sum_{j=0}^{n-1} \left\| \frac{\partial}{\partial \beta} \tilde{c}_j^{\beta^+} - \frac{\partial}{\partial \beta} \tilde{c}_j \right\| \|X_n(j)\| + \delta \sum_{j=n}^{\infty} \left\| \frac{\partial}{\partial \beta} \tilde{c}_j^{\beta^+} - \frac{\partial}{\partial \beta} \tilde{c}_j \right\| \|\gamma_{yz}(-j)\|$$

where the second term is $O(n^{-1})$. Note that $\frac{\partial}{\partial \beta} C^{-1}(\beta, \lambda) = C^{-2}(\beta, \lambda) \frac{\partial}{\partial \beta} C(\beta, \lambda)$ such that $j \left\| \frac{\partial}{\partial \beta} \tilde{c}_j^{\beta} \right\|$ is summable by Assumption (B1). It then follows from arguments similar to the proof of Lemma (A.6) that $E \sup_{\|\beta-\beta_0\|<\delta} \|M_n(\beta, a) - M(\beta_0, a)\| \leq K\delta/\sqrt{n}$ for some constant K. For any sequence $\delta_n \to 0$ the Markov inequality then implies that $\sup_{\|\beta-\beta_0\|<\delta_n} \sqrt{n} \|M_n(\beta, a) - M(\beta_0, a)\| = o_p(1)$. From Theorem 3.2.5 in van der Vaart and Wellner (1996) it follows that $\|\beta - \beta_0\| = O_p(n^{-1/2})$. Following the proof of Theorem 3.3 in Pakes and Pollard (1989) we now consider

$$\left\| G_n(\beta_n) - G_n(\beta_0) - \left[\frac{\partial}{\partial \beta_i} G_n(\beta_n)' \right] (\beta_n - \beta_0) \right\|$$

$$\leq \|M_n(\beta_n, a) - M(\beta_0, a)\| + \left\| G(\beta_n) - \left[\frac{\partial}{\partial \beta_i} G_n(\beta_n^i)'\right](\beta_n - \beta_0) \right\|.$$

Pick a sequence δ_n such that $P(\|\beta_n - \beta_0\| \geq \delta_n) \to 0$. Then for any $\epsilon, \eta > 0$ there exists an n such that

$$\begin{aligned} P(\sqrt{n}\,\|M_n(\beta_n, a) - M(\beta_0, a)\| &> \epsilon) \leq P(\sup_{\|\beta-\beta_0\|<\delta_n} \sqrt{n}\,\|M_n(\beta_n, a) - M(\beta_0, a)\| > \epsilon) \\ +P(\|\beta_n - \beta_0\| &\geq \delta_n) \leq \eta \end{aligned}$$

The term $\left\| G(\beta_n) - \left[\frac{\partial}{\partial \beta_i} G_n(\beta_n^i)'\right](\beta_n - \beta_0) \right\| = o_p(n^{-1/2})$ by a mean value expansion argument of $G(\beta_n)$ around β_0 and the fact that convergence of $\left[\frac{\partial}{\partial \beta_i} G_n(\beta_n^i)'\right]$ can be shown by the same arguments as convergence of $G_n(\beta, a)$ noting that both y_t and z_t are strictly stationary and $\partial \ln C(\beta, e^{-i\lambda})/\partial\beta$ is uniformly continuous on $[-\pi, \pi] \times U$ for $U \subset \Theta$, U compact, $\beta_0 \in \Theta$. A set U with these properties exists by local compactness of the parameter space. The details are omitted.

We have thus established that

$$[\frac{\partial}{\partial \beta} G_n(\beta_n)]\sqrt{n} G_n(\beta_n) = (M + o_p(1))[\sqrt{n} G_n(\beta_0) + \left[\frac{\partial}{\partial \beta_i} G_n(\beta_n^i)'\right]\sqrt{n}(\beta_n - \beta_0)] + o_p(1)$$

where $M = \sigma^2(2\pi)^{-1}\int_{-\pi}^{\pi} \partial \ln C(\beta_0, e^{-i\lambda})/\partial\beta l_a(\lambda) d\lambda$. Next, turn to

$$\sqrt{n} G_n(\beta_0) = \sqrt{n}(2\pi)^{-1}\int_{-\pi}^{\pi} C^{-1}(\beta_0, e^{-i\lambda}) I_{n,yz}(\lambda) d\lambda.$$

From Lemma (A.3) it follows that $\sqrt{n}\int_{-\pi}^{\pi} C^{-1}(\beta_0, e^{-i\lambda}) I_{n,yz}(\lambda) d\lambda - \sqrt{n}\int_{-\pi}^{\pi} I_{n,\varepsilon z}(\lambda) d\lambda =$

$o_p(1)$. Using Lemma (A.4) then shows that $\sqrt{n}G_n(\beta_0) \xrightarrow{d} N(0, \sum_{l=1}^{\infty}\sum_{k=1}^{\infty}\alpha_{k,l}a_k a_l')$ where it should be noted that $\sum_{l=1}^{\infty}\sum_{k=1}^{\infty}\alpha_{k,l}a_k a_l' = \lim A_m'\Omega_m A_m$. The result now follows from $\partial \ln C(\beta_0, e^{-i\lambda})/\partial\beta = \sum_{k=1}^{\infty} b_k e^{-i\lambda k}$ such that $(2\pi)^{-1}\int_{-\pi}^{\pi}\partial \ln C(\beta_0, e^{-i\lambda})/\partial\beta l_a(\lambda)' d\lambda = \sum_{k=1}^{\infty} b_k a_k'$. ■

Proof of Lemma 4.1 From Assumption (A1) it is clear that $\Omega x \in l^2$ for all $x \in l^2$. It remains to show that $\ker \Omega = 0$. Assume there is $x \in l^2$ such that $x \neq \{0, 0, ...\}$ and $\Omega x = 0$. Then also $x'\Omega x = 0$ which can be written as $E(\sum_{i=1}^{\infty} x_i \varepsilon_t \varepsilon_{t-i})^2 = 0$. But this is only possible if $\sum x_i \varepsilon_t \varepsilon_{t-i} = 0$ with probability one. Now $\sum x_i \varepsilon_t \varepsilon_{t-i} = 0$ $a.s.$ if $\varepsilon_t \varepsilon_{t-i} = 0$ $a.s.$ or the functions ε_{t-i} are linearly dependent $a.s.$ which is ruled out by Lemma (A.2).

On the other hand if $\varepsilon_t \varepsilon_{t-i} = 0$ $a.s.$ for all i then $\varepsilon_t^2 \varepsilon_{t-i}^2 = 0$ $a.s.$ But then $E(\varepsilon_t^2 \varepsilon_{t-i}^2) = 0$ for all i which contradicts Assumption (A1). Therefore $\Omega x = 0$ can only hold if $x = 0$. Thus $\ker \Omega = 0$. Symmetry of Ω now implies that $\operatorname{Im} \Omega = l^2$ therefore Ω^{-1} exists and is bounded on l^2.

Finally, it follows at once from before that $x_m'\Omega_m x_m = E(\sum x_{i,m}\varepsilon_t \varepsilon_{t-i})^2 > 0$ where the inequality is strict by Assumption (A1). So Ω_m is positive definite such that $\lambda_j^m > 0$ $\forall j, m$. This shows that Ω_m has full rank. ■

Proof of Lemma 4.2 By Assumption (A1) we know that $\sum\sum |\sigma(k,l)| < B$ thus $\sum_k |\sigma(k,l)| < B$ for any l. Therefore for any fixed l, $\sigma(k,l) \to 0$ as $k \to \infty$. This holds also if the roles of k and l are reversed. Also $\sum_k |\sigma(k,k)| < B$ such that $\sigma(k,k) \to 0$ as $k \to \infty$. Define the infinite dimensional matrices S_{12}^m, S_{21}^m and S_{22}^m according to the following partition

$$\Omega = \begin{bmatrix} \Omega_m & S_{12}^m \\ S_{21}^m & S_{22}^m \end{bmatrix}.$$

Then $tr(S_{12}^m S_{12}^{m'}) = \sum_{l=m+1}^{\infty} \sum_{k=1}^{m} |\sigma(k,l)|^2 \to 0$, $tr(S_{21}^m S_{21}^{m'}) \to 0$ and $tr(S_{22}^m - \sigma^4 I)(S_{22}^m - \sigma^4 I)' \to 0$ as $m \to \infty$. Define the infinite dimensional approximation matrix

$$\Omega_m^* = \begin{bmatrix} \Omega_m & 0 \\ 0 & \sigma^4 I \end{bmatrix}.$$

Clearly Ω_m^{*-1} exists $\forall m$ by Lemma (**4.1**) and the partitioned inverse formula. We now have

$$(\Omega^{-1} - \Omega_m^{*-1}) = \Omega_m^{*-1}(\Omega - \Omega_m^*)\Omega^{-1}$$

such that

$$\left\|\Omega^{-1} - \Omega_m^{*-1}\right\| \leq \left\|\Omega_m^{*-1}\right\| \left\|\Omega - \Omega_m^*\right\| \left\|\Omega^{-1}\right\|.$$

where $\|.\|$ is the matrix norm defined by $\|A\| = \sup_{\|x\|_2 \leq 1} \|Ax\|_2$. First show that $\left\|\Omega_m^{*-1}\right\|$ is bounded. By the partitioned inverse formula

$$\Omega_m^{*-1} = \begin{bmatrix} \Omega_m^{-1} & 0 \\ 0 & \sigma^{-4} I \end{bmatrix}$$

such that $\left\|\Omega_m^{*-1}\right\| \leq \left\|\Omega_m^{-1}\right\| + \sigma^{-4}$. We have shown in Lemma (4.1) that $0 < \min_{x'x=1} x'\Omega x$ which together with $\min_{x'x=1} x'\Omega x \leq \min_{x_m' x_m=1} x_m' \Omega_m x_m$ $\forall m$ implies that the smallest eigenvalue λ_1^m of Ω_m is bounded away from zero uniformly in m. Then by a familiar inequality for all $x \in \mathbb{R}^m$ $x^{'}\Omega_m^{-1}x/x^{'}x \leq 1/\lambda_1^m < \infty$ $\forall m$ such that $\left\|\Omega_m^{-1}\right\| < \infty$ since for finite m all norms are equivalent. Also $\left\|\Omega^{-1}\right\| < \infty$ by Lemma (4.1) and

$$\begin{aligned}
\|\Omega - \Omega_m^*\| &= \sup_{\|x\|\leq 1}\left(\sum_{k=1}^{m}\left|\sum_{l=m+1}^{\infty}\sigma\left(k,l\right)x_l\right|^2 + \sum_{k=m+1}^{\infty}\left|\sum_{l=1}^{\infty}\sigma\left(k,l\right)x_l\right|^2\right)^{1/2} \\
&\leq \sup_{\|x\|\leq 1}\sum_{k=1}^{m}\sum_{l=m+1}^{\infty}\left|\sigma\left(k,l\right)\right|\left|x_l\right| + \sup_{\|x\|\leq 1}\sum_{k=m+1}^{\infty}\sum_{l=1}^{\infty}\left|\sigma\left(k,l\right)\right|\left|x_l\right| \\
&\leq 2\sum_{l=m+1}^{\infty}\sum_{k=1}^{\infty}\left|\sigma\left(k,l\right)\right| \to 0 \text{ as } m\to\infty.
\end{aligned}$$

Thus $\left\|\Omega^{-1} - \Omega_m^{*-1}\right\| \to 0$ as $m \to \infty$ ■

Proof of Theorem 4.3 For all m fixed it follows from standard results that

$$(P_m'A_m)^{-1}(A_m'\Omega_m A_m)(A_m'P_m)^{-1} - (P_m'\Omega_m^{-1}P_m)^{-1} \geq 0.$$

But since for any sequence $\{x_m\}$ such that $x_m \geq 0$ for all m it follows that $\liminf_m x_m \geq 0$ the above inequality also holds in the limit. Since both $b(P) \in l^1$ and $a(A) \in l^1$ it follows from a bounded convergence argument that $\lim_m P_m'A_m$ exists and is finite. If $a \in \mathcal{A}^{**}$ then the inverse exists as well. The same arguments can be used to show that $\lim_m A_m'\Omega_m A_m$ exists and is finite.

Finally note that

$$\lim_m P_m^{'}\Omega_m^{-1}P_m = P'\Omega^{-1}P$$

by Lemma (4.2) since $P_m \in l^2$. ■

Proof of Theorem 5.1 We first show that $a(A) \in \mathcal{A}$ for $A' = P'\Omega^{-1}$. From Assumption

(A1) it follows that Ω maps l^1 into l^1. To see this write $\Omega = \Sigma + \sigma^4 I$ where the matrix Σ consists of elements $\sigma(k,l)$. For $x \in l^1$ we have $\Omega x = \Sigma x + \sigma^4 x$ with $\Sigma x \in l^1$ because of the summability restrictions on $\sigma(k,l)$. From Lemma 4.1 we know that for $x \in l^1 \subset l^2$ we have $\Omega^{-1} x \in l^2$. Assume $\Omega^{-1} x \notin l^1$. Then $x = \Omega\Omega^{-1}x = \Sigma\Omega^{-1}x + \sigma^4\Omega^{-1}x$. But $\Sigma\Omega^{-1}x \in l^1$. Thus $\left\|\sigma^4\Omega^{-1}x\right\|_1 = \left\|x - \Sigma\Omega^{-1}x\right\|_1 \leq \|x\|_1 + \left\|\Sigma\Omega^{-1}x\right\|_1$ and $\|x\|_1$ becomes unbounded because of $\left\|\sigma^4\Omega^{-1}x\right\|_1$. But this contradicts the assumption that $x \in l^1$. It follows that the image of l^1 under Ω^{-1} is also in l^1 which in turn implies that $\sum_{k=1}^{\infty} |\omega_{lk}| < \infty$ for all l. This can be seen by considering the image under Ω^{-1} of the l-th unit vector. Since $P \in \mathcal{A}$ it now follows that $P'\Omega^{-1} \in \mathcal{A}$.

In light of Lemma (3.3) we only need to show that $a(A) \in \mathcal{A}^{**}$ which implies $a(A) \in \mathcal{A}^{*}$. The optimal instrument is defined by $A' = P'\Omega^{-1}$ or $A'\Omega = P^{'}$. The row rank of $A^{'}$ is therefore the same as the column rank of P which has full column rank, thus establishing that $A_d = [a_1, ..., a_d]$ is nonsingular.

Next, $\int \dot{\eta}(\beta,\lambda) l_a(-\lambda)' d\lambda = P'\Omega^{-1}P$ and $P'\Omega^{-1}P = E(\varepsilon_t^2 z_t z_t')$. Now, $\det P'\Omega^{-1}P = 0 \Rightarrow \exists \ell \in \mathbb{R}^d$, $\ell \neq 0$ such that $\ell' E(\varepsilon_t^2 z_t z_t') = 0 \Rightarrow \ell' E(\varepsilon_t^2 z_t z_t')\ell = 0$. Then for $x_t := \ell' z_t$, $0 = E(\varepsilon_t^2 x_t^2) = E x_t^2 E\left[\varepsilon_t^2 \left| \mathcal{F}_{t-1}\right.\right] \Rightarrow E\left[\varepsilon_t^2 \left| \mathcal{F}_{t-1}\right.\right] = 0$ a.s. or $x_t^2 = 0$ a.s. Now, clearly $E\left[\varepsilon_t^2 \left| \mathcal{F}_{t-1}\right.\right] = 0$ a.s. is ruled out by Assumption $(A1)$. Then $x_t^2 = 0$ a.s. implies $x_t = \ell' z_t = 0$ a.s. From Lemma (A.2) it follows that $z_t = 0$ a.s. is impossible and we have shown before that the column rank of A is full so that $\ell' z_t = 0$ a.s. is also impossible. ■

Proof of Theorem 5.3: We need to show that $\sum_{k=1}^{\infty} |k| \left| [a_k]_j \right|$ for $j = 1, ..., d$ is bounded.

Since $P \in \mathcal{A}$ we can write $P^{'} = P^{'}\Omega^{-1}\Omega = P^{'}\Omega^{-1}(\sigma^4 I + \Sigma)$. Define the vector $\ell_k = k e_k$

where e_k is the k-th unit vector. Then

$$P^{'}\ell_k = P^{'}\Omega^{-1}(\sigma^4 I + \Sigma)\ell_k. \tag{24}$$

Now, the sequence $\left\{P^{'}\ell_k\right\}_{k=1}^{\infty} \in \mathcal{A}$ and $\Sigma\ell_k \in l^1$ for all k. Therefore, by the fact that $a(P'\Omega^{-1}) \in \mathcal{A}$ and by the summability assumption of Lemma (5.3), $\left\{P^{'}\Omega^{-1}\Sigma\ell_k\right\}_{k=1}^{\infty} \in \mathcal{A}$. From (24) we have

$$\begin{aligned} \left|P^{'}\Omega^{-1}\ell_k\sigma^4\right| &= \left|P^{'}\ell_k - P^{'}\Omega^{-1}\Sigma\ell_k\right| \\ &\leq \left|P^{'}\ell_k\right| + \left|P^{'}\Omega^{-1}\Sigma\ell_k\right|. \end{aligned}$$

where $|.|$ is a vector norm on $\mathbb{R}^d$. Without loss of generality we use $|x| = \sup_i |x_i|$ for $x \in \mathbb{R}^d$. Summing over k gives $\sigma^4 \sum_{k=1}^{\infty} \left|P^{'}\Omega^{-1}\ell_k\right| \leq \sum_{k=1}^{\infty} \left(\left|P^{'}\ell_k\right| + \left|P'\Omega^{-1}\Sigma\ell_k\right|\right) < \infty$. Note that $\left|P^{'}\Omega^{-1}\ell_k\right| = k\left|\sum_{l=1}^{\infty} b_l\omega_{lk}\right|$. This establishes the result ■.

Proof of Corollary 5.4: We first need to establish consistency. For this we show that uniformly in Θ, $\left\|\tilde{G}_n(\beta,a) - G_n(\beta,a)\right\| \xrightarrow{p} 0$. Note that $z_t - \hat{z}_t = \sum_{j=1}^{t-1} a_j(\varepsilon_{t-j} - \hat{\varepsilon}_{t-j}) + \sum_{j=t}^{\infty} a_j\varepsilon_{t-j}$ and $\varepsilon_{t-j} - \hat{\varepsilon}_{t-j} = \sum_{l=t-j}^{\infty} c(\beta_0,l)y_{t-j-l}$. Without loss of generality assume $a_j \in \mathbb{R}$. Then

$$\begin{aligned} \left|\tilde{G}_n(\beta,a) - G_n(\beta,a)\right| \leq\ & n^{-1}\left|\sum_{t=1}^{n}\sum_{j=1}^{t-1} a_j \sum_{l=t-j}^{\infty} c(\beta_0,l)y_{t-j-l}\sum_{r=0}^{t-1}\tilde{c}_r^{\beta}y_{t-r}\right| \\ &+n^{-1}\left|\sum_{t=1}^{n}\sum_{j=t}^{\infty} a_j\varepsilon_{t-j}\sum_{r=0}^{t-1}\tilde{c}_r^{\beta}y_{t-r}\right| \end{aligned}$$

where, using $t \leq 2|j||l|$ on the relevant range of summation, the first term can be bounded by

$$2n^{-1}\sum_{t=1}^{n}\sum_{j=1}^{t-1} j\left|a_j\right| \sum_{l=t-j}^{\infty} t^{-1}l\left|c(\beta_0,l)\right| \sum_{r=0}^{t-1}\left|\tilde{c}_r^{\beta}\right|\left|y_{t-j-l}y_{t-r}-\gamma_{yy}(j+l-r)\right|$$
$$+2n^{-1}\sum_{t=1}^{n}\sum_{j=1}^{t-1} j\left|a_j\right| \sum_{l=t-j}^{\infty} t^{-1}l\left|c(\beta_0,l)\right| \sum_{r=0}^{t-1}\left|\tilde{c}_r^{\beta}\right|\left|\gamma_{yy}(j+l-r)\right|.$$

Note that $E\left|y_{t-j-l}y_{t-r}-\gamma_{yy}(j+l-r)\right|$ is uniformly bounded in t, j, l, r and the remaining terms in the expression are summable in Θ. We can therefore bound the sum by

$$Kn^{-1}\sum_{t=1}^{n} t^{-1} < Kn^{-1+\nu}\sum_{t=1}^{n} t^{-(1+\nu)} = O(n^{-1+\nu})$$

for $\nu \in (0,1/2)$ and some constant K. Thus by the Markov inequality the first term is $O_p(n^{-1})$. The second term can be bounded by

$$n^{-1}\sum_{t=1}^{n} t^{-1}\sum_{j=t}^{\infty} j\left|a_j\right|\sum_{r=0}^{t-1}\left|\tilde{c}_r^{\beta}\right|\left|y_{t-r}\varepsilon_{t-j}-\gamma_{y\varepsilon}(r-j)\right|+n^{-1}\sum_{t=1}^{n} t^{-1}\sum_{j=t}^{\infty} j\left|a_j\right|\sum_{r=0}^{t-1}\left|\tilde{c}_r^{\beta}\right|\left|\gamma_{y\varepsilon}(r-j)\right|$$

where again $E\left|y_{t-r}\varepsilon_{t-j}-\gamma_{y\varepsilon}(r-j)\right|$ is uniformly bounded and $n^{-1}\sum_{t=1}^{n} t^{-1} = O(n^{-1+\nu})$ for $\nu \in (0,1/2)$. This establishes $\sup_{\beta\in\Theta}\left|\tilde{G}_n(\beta,a)-G_n(\beta,a)\right| \xrightarrow{p} 0$ and thus $\tilde{\beta} \xrightarrow{p} \beta_0$. Next we show that $\sup_{\beta\in\Theta}\left|\tilde{G}_n(\beta,a)-\tilde{G}_n^F(\beta,a)\right| = O_p(n^{-1})$. Note that

$$\begin{aligned}
\tilde{G}_n^F(\beta,a) &= \frac{1}{n}\sum_{j=0}^{\infty}\sum_{k=1}^{\infty}\sum_{l=0}^{\infty} c_l\tilde{c}_j^{\beta}a_k \sum_{t=\max(k+l,j)}^{\min(n+k+l,n+j)} y_{t-j}y_{t-k-l} \\
&= G_n(\beta,a)+\frac{1}{n}\sum_{j=0}^{\infty}\sum_{k=1}^{\infty}\sum_{l=0}^{\infty} c_l\tilde{c}_j^{\beta}a_k\left\{\max(j,k+l)<n\right\}\left(\sum_{t=n}^{\min(n+k+l,n+j)} y_{t-j}y_{t-k-l}\right)
\end{aligned}$$

$$+\frac{1}{n}\sum_{j=0}^{\infty}\sum_{k=1}^{\infty}\sum_{l=0}^{\infty} c_l \tilde{c}_j^{\beta} a_k \{\max(j,k+l)\geq n\}\left(\sum_{t=\max(k+l,j)+1}^{\min(n+k+l,n+j)} y_{t-j}y_{t-k-l}\right)$$

where the second term is uniformly bounded in expectation by

$$\frac{1}{n}\sup_{t,s} E\left|y_t y_s\right| \sup_{\beta}\sum_{j=0}^{n}\sum_{k=1}^{n}\sum_{l=0}^{n} jlk\left|c_l\right|\left|\tilde{c}_j^{\beta}\right|\left|a_k\right| = O(n^{-1})$$

where we can exchange sup and E by a dominated convergence argument since $j\left|\tilde{c}_j^{\beta}\right|$ is uniformly summable on Θ. The third term can be uniformly bounded in expectation by

$$\begin{aligned}&\frac{1}{n}\sup_{t,s} E\left|y_t y_s\right| \sup_{\beta}\sum_{k=1}^{\infty}\sum_{l=0}^{\infty}\sum_{j=0}^{k+l}(k+l)\left|c_l\right|\left|\tilde{c}_j^{\beta}\right|\left|a_k\right|\{k+l\geq n\}\\ &+\frac{1}{n}\sup_{t,s} E\left|y_t y_s\right| \sup_{\beta}\sum_{k=1}^{\infty}\sum_{l=0}^{\infty}\sum_{j=n}^{\infty} j\left|c_l\right|\left|\tilde{c}_j^{\beta}\right|\left|a_k\right|\end{aligned}$$

which is also $O(n^{-1})$. It then follows that $E\sup_{\beta}\left|\tilde{G}_n(\beta,a)-\tilde{G}_n^F(\beta,a)\right| = O(n^{-1})$. The result follows by the Markov inequality. For the limiting distribution expand (12) as

$$\sqrt{n}(\tilde{\beta}-\beta_0) = \left[\partial\tilde{G}_n^F(\beta^+,a)/\partial\beta\right]^{-1}\sqrt{n}\tilde{G}_n^F(\beta_0,a)+o_p(1)$$

where $\left\|\beta^+-\beta_0\right\|\leq\left\|\tilde{\beta}-\beta_0\right\|$ where β^+ varies across different rows in $\partial\tilde{G}_n^F(\beta^+,a)$ by the mean value theorem. From Lemma A.2 and A.3 in Kuersteiner (1999) it follows that $\sqrt{n}\tilde{G}_n^F(\beta_0,a)\xrightarrow{d} N(0,P'\Omega^{-1}P)$ and by standard arguments

$$\partial\tilde{G}_n^F(\beta^+,a)/\partial\beta \xrightarrow{p}\int\dot{\eta}(\beta_0,\lambda)l_{\psi}(-\lambda)'d\lambda = P'\Omega^{-1}P. \blacksquare$$

Proof of Theorem 5.5: For consistency we establish that $\sup_{\beta\in\Theta}\left|\tilde{G}_n^F(\beta,a)-\tilde{G}_n^F(\beta,\hat{a})\right|\xrightarrow{p}$ 0. Here

$$\sup_{\beta\in\Theta}\left|\tilde{G}_n^F(\beta,a)-\tilde{G}_n^F(\beta,\hat{a})\right|\leq\frac{1}{2\pi}\sup_{\substack{\beta\in\Theta,\\ \lambda\in[-\pi,\pi]}}\left|C^{-1}(\beta,e^{-i\lambda})\right|\sup_{\lambda}\left|h(\beta_0,\lambda)-\hat{h}(\hat{\beta},\lambda)\right|\int_{-\pi}^{\pi}I_{n,yy}(\lambda)d\lambda$$

where $\sup_\lambda\left|h(\beta_0,\lambda)-h(\hat{\beta},\lambda)\right| = o_p(1)$ by Theorem 5.1 in Kuersteiner (1999) such that consistency follows by standard arguments. The same arguments also lead to

$$\sup_{\beta\in\Theta}\left|\frac{\partial}{\partial\beta}\tilde{G}_n^F(\beta,a)-\frac{\partial}{\partial\beta}\tilde{G}_n^F(\beta,\hat{a})\right|\xrightarrow{p}0$$

The limit theory is established by showing that $\sqrt{n}\left(\tilde{G}_n^F(\beta_0,a)-\tilde{G}_n^F(\beta_0,\hat{a})\right)=o_p(1)$. The proof is essentially identical to the proof of Theorem 5.2 in Kuersteiner (1999) and is omitted. ■

Brillinger, D. R., 1981. Time Series, Data Analysis and Theory, Expanded Edition. Holden-Day, Inc.

Brockwell, P. J. and R. A. Davis, 1987. Time Series: Theory and Methods. Springer Verlag-New York, Inc.

Chen, B. and P.A. Zadrozny, 1998. An Extended Yule-Walker Method for Estimating a Vector Autoregressive Model with Mixed Frequency Data.In: Advances in Econometrics, vol 13, T. Fomby and R.C. Hill, eds., JAI Press, 47-73.

Deistler, M., W. Dunsmuir, and E.J. Hannan, 1978. Vector Linear Time Series Models: Corrections and Extensions. Adv. Appl. Prob., 10:360–372.

Dunsmuir, W.,1979. A Central Limit Theorem for Parameter Estimation in Stationary Vector Time Series and its Application to Models for a Signal Observed with Noise. The Annals of Statistics, 490–506.

Dunsmuir, W. and E.J. Hannan, 1976. Vector Linear Time Series Models. Adv. Appl. Prob., 8:339–364.

Gaposhkin, V.F., 1980. Almost Sure Convergence of Estimates for the Spectral Density of a Stationary Process. Theory of Probability and its Applications, 169–176.

Gohberg, I. and S. Goldberg, 1991. Basic Operator Theory. Birkhauser.

Hanani, H., E. Netanyahu, and M. Reichaw, 1968. Eigenvalues of Infinite Matrices. Colloquium Mathematicum 19, 89–101.

Hannan, E.J., 1973. The Asymptotic Theory of Linear Time Series Models. Journal of Applied Probability 10, 130–145.

Hansen, L.P., 1982. Large Sample Properties of Generalized Method of Moments Estimators. Econometrica 50, 1029–1053.

Hansen, L.P., 1985. A Method for Calculating Bounds on the Asymptotic Covariance Matrices of Generalized Method of Moments Estimators. Journal of Econometrics 30, 203–238.

Hansen, L. P., J. C. Heaton, and M. Ogaki, 1988. Efficiency Bounds Implied by Multiperiod Conditional Moment Restrictions. Journal of the American Statistical Association 83, 863–871.

Hansen, L.P. and K. J. Singleton, 1991. Computing Semiparametric Efficiency Bounds for Linear Time Series Models. In William A. Barnett, James Powell, and George Tauchen, editors, Nonparametric and Semiparametric Methods in Econometrics and Statistics (Cambridge University Press), 387–412.

Hansen, L.P. and K. J. Singleton, 1996. Efficient Estimation of Linear Asset-Pricing Models with Moving Average Errors. Journal of Business and Economic Statistics, 53–68.

Hayashi, F. and C. Sims, 1983. Nearly Efficient Estimation of Time Series Models with Predetermined, but not Exogenous, Instruments. Econometrica, 51, 783–798.

Hosoya, Y. and M. Taniguchi, 1982. A Central Limit Theorem for Stationary Processes and the Parameter Estimation of Linear Processes. The Annals of Statistics 10, 132–153.

Kabaila, P.V., 1980. An Optimality Property of the Least Squares Estimate of the Parameter of the Spectrum of a Purely Nondeterministic Time Series. The Annals of Statistics 8, 1082–1092.

Kuersteiner, G.M., 1997. Efficient Inference in Time Series Models with Conditional Heterogeneity. Ph.D. Dissertation, Yale University.

Kuersteiner, G.M., 1999. Efficient IV Estimation for Autoregressive Models with Conditional Heterogeneity. MIT Department of Economics Working Paper No. 99-08.

Magnus, J.R. and H. Neudecker, 1988. Matrix Differential Calculus with Applications in Statistics and Econometrics. John Wiley and Sons.

Marsaglia, G. and G. P. H. Styan, 1974. Equalities and Inequalities for Ranks of Matrices. Linear and Multilinear Algebra, 2, 269-292.

Milhoj, A., 1985. The Moment Structure of ARCH Processes. Scandinavian Journal of Statistics 12, 281–292.

Nelson, D.B. (1990): Stationarity and Persistence in the GARCH(1,1) model. Econometric Theory 6, 318–334.

Nelson, D.B. (1991): Conditional Heteroskedasticity in Asset Returns: A New Approach. Econometrica 59, 347–370.

Pakes, A. and D. Pollard (1989): Simulation and the Asymptotics of Optimization Estimators. Econometrica 57, 1027–1057.

Royden, H.L., 1988. Real Analysis. Macmillan Publishing Company.

Stoica, P., T. Söderström, and B. Friedlander, 1985. Optimal Instrumental Variable Estimates of the AR Parameters of an ARMA process. IEEE Transactions on Automatic Control 30, 1066–1074.

Stoica, P., T. Söderström, and B. Friedlander, 1987a. Optimal Instrumental Variable Multistep Algorithms for Estimation of the AR parameters of an ARMA Process. International Journal of Control, vol 45, 2083-2107.

Stoica, P., T. Söderström, and B. Friedlander, 1987b. Instrumental Variable Methods for ARMA models. In: Control and Dynamic Systems-Advances in Theory and Applications, vol. 25: System Identification and Adaptive Control (C.T. Leondes, Ed.), 79-150, New York, Academic Press.

Taniguchi, M., 1983. On the Second Order Asymptotic Efficiency of Estimators of Gaussian ARMA Processes. The Annals of Statistics 11, 157–169.