Econometric Theory, **24**, 2008, 1063–1092. Printed in the United States of America.
doi:10.1017/S0266466608080419

USING SUBSPACE METHODS FOR ESTIMATING ARMA MODELS FOR MULTIVARIATE TIME SERIES WITH CONDITIONALLY HETEROSKEDASTIC INNOVATIONS

DIETMAR BAUER
arsenal research

This paper deals with the estimation of linear dynamic models of the autoregressive moving average type for the conditional mean for stationary time series with conditionally heteroskedastic innovation process. Estimation is performed using a particular class of subspace methods that are known to have computational advantages as compared to estimation based on criterion minimization. These advantages are especially strong for high-dimensional time series. Conditions to ensure consistency and asymptotic normality of the subspace estimators are derived in this paper. Moreover asymptotic equivalence to quasi maximum likelihood estimators based on the Gaussian likelihood in terms of the asymptotic distribution is proved under mild assumptions on the innovations. Furthermore order estimation techniques are proposed and analyzed.

1. INTRODUCTION

This paper considers the estimation of autoregressive moving average (ARMA) type models for stationary time series using so-called subspace methods. Traditionally ARMA models have been estimated using criterion optimization, the quasi maximum likelihood (quasi ML) method being the most prominent example. In this paper the term *quasi likelihood* is used for the Gaussian likelihood neglecting possible conditional heteroskedasticity of the innovations. The asymptotic properties of the quasi ML estimators are well documented, and conditions ensuring consistency and asymptotic normality of the estimators are known (see, e.g., Hannan and Deistler, 1988 [henceforth HD], Ch. 4). These results

Financial support via the Austrian FWF under project P-17065-N12 and the Max Kade foundation is gratefully acknowledged. Part of this work was done while the author held a postdoctoral position at the Cowles Foundation for Research in Economics, Yale University, New Haven, Connecticut, USA, whose hospitality is appreciated. The very helpful comments of the anonymous referees led to a substantial improvement of the paper that is gratefully acknowledged. Finally I thank Wolfgang Scherrer, Martin Wagner, and Michael McAleer for fruitful discussions on the topics of this paper. Address correspondence to Dietmar Bauer, arsenal research, Giefingg. 2, A-1210 Vienna, Austria; e-mail: Dietmar.Bauer@arsenal.ac.at.

© 2008 Cambridge University Press 0266-4666/08 $15.00

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

cover both conditionally homoskedastic and conditionally heteroskedastic innovations. In the Gaussian homoskedastic case it is well known that the quasi ML estimator attains the smallest possible asymptotic variance matrix.[1]

In the conditionally heteroskedastic case the quasi ML estimator in general loses its efficiency (for a specific example, see, e.g., Engle, 1982, Sect. 7) but still provides an initial guess for subsequent efficient estimation including either a parametric model for the conditional variance (such as, e.g., the generalized autoregressive conditional heteroskedasticity model) or a nonparametric model such as is used, e.g., in the efficient instrumental variables estimators of Kürsteiner (2001, 2002).

Typically the maximization of the quasi likelihood leads to a nonlinear optimization problem that is solved using standard numerical methods such as the Gauss–Newton method or other gradient-based search algorithms. In particular for high-dimensional data sets computational speed and numerical robustness become relevant issues.[2] This is a fortiori an issue if the estimate only serves as an initial guess for subsequent estimation.

Subspace methods have been proposed in the engineering literature as a possible solution to the numerical problems of quasi ML methods. They are based primarily on regressions and hence can be implemented numerically robust and fast. The most popular algorithms are CCA (canonical correlation analysis)[3] proposed in Larimore (1983), MOESP (multivariate output error state space; Verhaegen, 1994) and N4SID (numerical methods for subspace state space identification; Van Overschee and DeMoor, 1994). They are largely based on realization algorithms such as the method of Ho and Kalman (1966). In the econometrics literature the work of Aoki (1990) and Aoki and Havenner (1991) introduced algorithms based on realization theory similar to the Ho–Kalman algorithm (Ho and Kalman, 1966). For a survey of subspace methods including the realization-based approaches see Chapter 3 of Bauer (1998).

The asymptotic properties for the particular method considered in this paper have been analyzed thoroughly in the stationary conditionally homoskedastic case (for a survey, see Bauer, 2005b). The bottom line of these results is that in the case of known order of the true system generating the data, sufficient conditions are known such that the estimators obtained by using CCA are consistent, asymptotically normal, and asymptotically equivalent to estimates obtained by maximizing the Gaussian likelihood in the sense that the asymptotic distributions of the suitably normalized estimation errors coincide (see Sect. 4 for precise statements).

The aim of this paper is to extend the asymptotic theory for a particular subspace algorithm, namely, CCA, from the conditionally homoskedastic to the conditionally heteroskedastic case. In particular conditions ensuring consistency, asymptotic normality, and also the asymptotic equivalence to quasi ML estimation will be derived. Furthermore order estimation is discussed. It will be established that analogously to quasi ML estimation also estimation using CCA is robust to the occurrence of conditional heteroskedasticity of unknown

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

form subject to mild conditions on the innovations. Hence CCA will be demonstrated to be an attractive alternative to quasi ML estimation to obtain initial guesses for subsequent efficient estimation taking the conditional variance of the innovations into account.

The paper is organized as follows. In the next section the model set and the assumptions are discussed. Section 3 presents the CCA subspace method. Section 4 contains the main results of this paper, which are proved in the Appendix. Section 5 discusses the inclusion of deterministic terms into the model. Finally Section 6 concludes the paper. The technical details are contained in the Appendix.

Throughout the paper we will use the notation $F_T = o(g_T)$ for random matrix sequences $F_T \in \mathbb{R}^{a_T \times b_T}$ of dimension $a_T \times b_T$ possibly depending on the sample size T and scalar sequences g_T meaning that $\max_{1\le i\le a_T, 1\le j\le b_T} |F_{T,i,j}/g_T| \to 0$ almost surely (a.s.). Here $F_{T,i,j}$ denotes the (i,j) entry of the matrix F_T. Further $F_T = O(g_T)$ means that there exists a constant $M < \infty$ such that $\limsup_{T\to\infty} \max_{1\le i\le a_T, 1\le j\le b_T} |F_{T,i,j}/g_T| \le M$ a.s. Let $o_P(g_T)$ and $O_P(g_T)$ denote the corresponding in probability versions. Note that these concepts differ from the usual definition in situations where the dimension of the matrix F_T depends on T. We assume uniform boundedness of the entries rather than boundedness of the norm of the matrix. The term $\lambda_{max}(A)$ denotes an eigenvalue of maximum modulus of the matrix A. The Kronecker product is denoted by $\otimes$. The euclidean norm of a matrix or a vector will be denoted by $\|.\|$; furthermore the Frobenius norm ($\|X\|_{Fr} := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} X_{i,j}^2}$), the one norm ($\|X\|_1 := \max_{1\le j\le n}\sum_{i=1}^{m}|X_{i,j}|$), and the infinity norm ($\|X\|_\infty := \max_{1\le i\le n}\sum_{j=1}^{m}|X_{i,j}|$) will be used. Here $X = [X_{i,j}]_{i=1,\dots,n,j=1,\dots,m} \in \mathbb{R}^{n\times m}$.

2. MODEL SET AND ASSUMPTIONS

In this paper we consider processes $(y_t)_{t\in\mathbb{Z}}$ generated by a linear, time invariant, finite-dimensional, discrete time, state space system of the form

$$\begin{aligned} x_{t+1} &= Ax_t + K\varepsilon_t, \\ y_t &= Cx_t + \varepsilon_t \end{aligned} \tag{1}$$

for $t \in \mathbb{Z}$ where $y_t \in \mathbb{R}^s$ denotes the s-dimensional observed output, $(x_t)_{t\in\mathbb{Z}}$ the unobserved n-dimensional state process, and $(\varepsilon_t)_{t\in\mathbb{Z}}$ the s-dimensional innovation sequence. Here $A \in \mathbb{R}^{n\times n}, K \in \mathbb{R}^{n\times s}, C \in \mathbb{R}^{s\times n}$ are real matrices. Throughout the paper it is assumed that the system is stable, i.e., $|\lambda_{max}(A)| < 1$, and strictly minimum-phase, i.e., $|\lambda_{max}(A - KC)| < 1$. Under these assumptions for each stationary innovation sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$ there exists a unique stationary solution $(y_t)_{t\in\mathbb{Z}}$ to the state space equation (see, e.g., HD, p. 11):

$$y_t = \varepsilon_t + \sum_{j=1}^{\infty} K_j \varepsilon_{t-j},$$

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

where $K_j = CA^{j-1}K, j > 0$. The transfer function $k(z) := I_s + zC(I_n - zA)^{-1}K = I_s + \sum_{j=1}^{\infty} K_j z^j$ for $z \in \mathbb{C}, |z| \leq 1$ then describes the input/output mapping. The transfer function $k(z)$ is a matrix-valued function that is rational in $z \in \mathbb{C}$ and therefore has a representation as an ARMA system according to a left matrix fraction representation $k(z) = a^{-1}(z)b(z)$ where $a(z) = I_s + a_1 z + \cdots + a_p z^p, b(z) = I_s + b_1 z + \cdots + b_q z^q$ for some integers $p, q \in \mathbb{N} \cup \{0\}$. Furthermore $(y_t)_{t\in\mathbb{Z}}$ satisfies the corresponding ARMA vector difference equation, i.e., $a(L)(y_t)_{t\in\mathbb{Z}} = b(L)(\varepsilon_t)_{t\in\mathbb{Z}}$ where L as usual denotes the backward shift operator. Conversely each ARMA process has a representation as a solution to state space equations. See HD (p. 15) for an explicit construction of a state space system based on the ARMA representation. In this sense ARMA representations and state space representations are equivalent. In what follows we will use the state space framework. For a more detailed discussion on the relation between ARMA and state space systems see HD, Section 1.2.

In what follows we will provide a very brief discussion of the main concepts in the state space framework that are needed for the estimation theory in this paper. The presentation of the concepts is intended to serve as a reference list of technical terms in the state space framework and is hence far from being self-contained. For a detailed discussion we refer the interested reader to Chapters 1 and 2 of HD.

Let π denote the mapping attaching the transfer function $k(z) := I_s + zC(I_n - zA)^{-1}K = \pi(A, K, C)$ to the state space system (A, K, C). A state space system (A, K, C) is called *minimal* if no other state space system $(\tilde{A}, \tilde{K}, \tilde{C})$ exists where $\pi(A, K, C) = \pi(\tilde{A}, \tilde{K}, \tilde{C})$ and where $\tilde{A} \in \mathbb{R}^{\tilde{n}\times\tilde{n}}, A \in \mathbb{R}^{n\times n}$ such that $\tilde{n} < n$. In other words *minimality* refers to minimal state dimension. When modeling the transfer function $k(z)$ there is no loss of generality when restricting attention to minimal systems, which will be done henceforth. The set of transfer functions corresponding to minimal, stable, strictly minimum-phase state space systems with state dimension equal to n will be denoted as $M(n)$. For $k(z) \in M(n)$ the integer n will be called the *order* of the system or of the transfer function, respectively.

Two systems (A, K, C) and $(\tilde{A}, \tilde{K}, \tilde{C})$ are called *observationally equivalent* if $\pi(A, K, C) = \pi(\tilde{A}, \tilde{K}, \tilde{C})$. Two minimal systems (A, K, C) and $(\tilde{A}, \tilde{K}, \tilde{C})$ are observationally equivalent if and only if there exists a nonsingular matrix $\mathcal{T} \in \mathbb{R}^{n\times n}$ such that $A = \mathcal{T}\tilde{A}\mathcal{T}^{-1}, K = \mathcal{T}\tilde{K}, C = \tilde{C}\mathcal{T}^{-1}$ corresponding to $x_t = \mathcal{T}\tilde{x}_t$ (cf. HD, Thm. 2.3.4, p. 48). Therefore minimality is not restrictive enough to uniquely identify one state space system corresponding to a particular transfer function $k(z)$, and hence more restrictions are needed to define a parameterization of $M(n)$. In this paper we will use the overlapping forms discussed in Section 2.6 of HD that achieve identification by restricting certain entries of A and C to be equal to either zero or one. All unrestricted entries are free parameters. The overlapping forms are defined as a collection of homeomorphisms $\varphi_\alpha : \Theta_\alpha \subset \mathbb{R}^{2ns} \to M(n; \alpha), \alpha \in \Lambda$ such that each set $M(n; \alpha)$ is open in $M(n)$ and each $k(z) \in M(n)$ lies in the interior of some $M(n; \alpha)$. Here Λ denotes a finite set of

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

multi-indices α indexing the sets $M(n;\alpha)$. The mappings φ_α are continuous where the topology in $M(n)$ is the so-called pointwise topology (for a definition see HD, p. 55) and Θ_α is equipped with the euclidean topology. For details on the structure of the various pieces $M(n;\alpha)$ of $M(n)$ and the geometrical properties of the decomposition see Section 2.6 in HD.

For this paper it is sufficient to state that to each $\theta \in \Theta_\alpha$ the parameterization attaches a system $(A(\theta), K(\theta), C(\theta))$ such that $k(z,\theta) := \pi(A(\theta), K(\theta), C(\theta)) \in M(n;\alpha)$. Conversely for each $k(z) \in M(n)$ there exists for each index α such that $k(z) \in M(n;\alpha)$ a unique parameter vector $\theta_\alpha = \varphi_\alpha^{-1}(k(z)) \in \Theta_\alpha$ where typically $k(z) \in M(n;\alpha)$ for more than one multi-index α. Changing the coordinate system is a smooth operation; i.e., for $\alpha_1 \neq \alpha_2$ such that $k(z,\theta_{\alpha_1}) \in M(n;\alpha_2)$ the mapping $\theta_{\alpha_2} = \varphi_{\alpha_2}^{-1}(\varphi_{\alpha_1}(\theta_{\alpha_1}))$ is differentiable for each $\theta_{\alpha_1} \in \Theta_{\alpha_1}$ (see HD, Thm. 2.6.3). Furthermore the mapping $\rho_\alpha : \theta \mapsto (A(\theta), K(\theta), C(\theta))$ trivially is differentiable for $\theta \in \Theta_\alpha$ because the parameters contained in θ_α are the free entries of $(A(\theta), K(\theta), C(\theta))$. Finally Θ_α is an open subset of $\mathbb{R}^{2ns}$.

Throughout this paper we will always use the following assumptions on the noise.

Assumption 1. The process $(\varepsilon_t)_{t\in\mathbb{Z}}$ is assumed to be an ergodic, strictly stationary, martingale difference sequence with respect to the sequence of increasing sigma fields $\mathcal{F}_t = \sigma\{\varepsilon_t, \varepsilon_{t-1}, \ldots\}$ having the following properties:

$$\mathbb{E}\{\varepsilon_t | \mathcal{F}_{t-1}\} = 0, \qquad \lim_{k\to\infty} \mathbb{E}\{\varepsilon_t \varepsilon_t' | \mathcal{F}_{t-k}\} = \Omega_0 = \mathbb{E}\varepsilon_t \varepsilon_t' > 0, \quad \text{a.s.}$$

$$\mathbb{E}\varepsilon_{t,j}^4 < \infty, \qquad j = 1, \ldots, s,$$

where $\varepsilon_{t,j}$ denotes the jth component of the vector ε_t.

These assumptions on the innovation sequence hold for processes generated by a number of commonly used models. Among these are in the univariate case independent and identically distributed (i.i.d.) innovations with finite fourth moments, autoregressive conditionally heteroskedastic (ARCH) processes (Engle, 1982), generalized autoregressive conditionally heteroskedastic (GARCH) processes (Bollerslev, 1986), and exponential GARCH (E-GARCH) processes (Nelson, 1991). For ARCH and GARCH processes this can be verified using Theorem 1.3 of Bougerol and Picard (1992) where the assumption of finite fourth moment introduces restrictions on the parameters (for the ARCH case, see Gourieroux, 1997, Exer. 3.4; for the GARCH case, see Ling and McAleer, 2002). For E-GARCH this follows from Nelson (1991, Thm. 2.2). In the multivariate case verification of the assumption is (besides the trivial i.i.d. case) only known for a small number of models such as specific Baba, Engle, Kraft, and Kroner–ARCH (BEKK-ARCH) (Engle and Kroner, 1995) models (see Rahbek, Hansen, and Dennis, 2003). It should be noted, however, that the results in this paper do not rely on any assumption on the specific

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

structure of the conditional heteroskedasticity but instead only use the high level assumption given by Assumption 1.

3. ESTIMATION METHODS

The main problem dealt with in this paper is the estimation of the state space model for the conditional mean based on observations $y_1, \dots, y_T$ of the output. The traditional method of estimation is quasi ML, which is described subsequently. Typically, prior to estimation of the model for the mean, deterministic terms will be dealt with. Throughout the paper we will not deal with these issues to facilitate discussion. Section 5 at the end of the paper provides the main facts in this respect.

Assume that a constant conditional variance $\Omega = \Omega(\omega)$ is parameterized using the parameter vector $\omega \in S \subset \mathbb{R}^{s(s+1)/2}$ and the true innovation $\varepsilon_t = \Omega^{1/2}(\omega_0)\eta_t$ where η_t is i.i.d. standard normally distributed. Here $X^{1/2}$ denotes the uniquely defined symmetric square root of a matrix X. Further assume that $(y_t)_{t\in\mathbb{Z}}$ is generated by a state space system of the form (1) for some system of order n_0. For some prespecified order n and multi-index α then $-2/T$ times the logarithm of the Gaussian likelihood is equal to

$$L(y_t, \tau) := (\log\det\Gamma_T^+(\tau) + (Y_{1,T}^+)'\Gamma_T^+(\tau)^{-1}Y_{1,T}^+)/T, \tag{2}$$

where $\tau := [\theta', \omega']' \in \Psi_\alpha := \overline{\Theta_\alpha} \times S$ ($\overline{\Theta_\alpha}$ denoting the closure of Θ_α) and where $\Gamma_T^+(\tau)$ denotes the variance of the vector $Y_{1,T}^+ := [y_1', y_2', \dots, y_T']'$ according to the parameter θ and the unconditional innovation variance $\Omega(\omega)$ assuming $(y_t)_{t\in\mathbb{Z}}$ to be the stationary solution to the state space equations. The quasi ML estimator then is obtained as the minimizing argument[4] of this function, i.e.,

$$\hat{\tau}_\alpha := \arg\min_{\tau\in\Psi_\alpha} L(y_t, \tau).$$

In the misspecified case that $(\eta_t)_{t\in\mathbb{Z}}$ is not i.i.d. standard normally distributed, but the innovations fulfill Assumption 1 and the order $n = n_0$ and the multi-index α are correctly specified, then the theory of HD, Chapter 4, still applies, and the corresponding estimator $\hat{\tau}_\alpha$ is consistent (HD, Thm. 4.2.1) and asymptotically normal (HD, Thm. 4.3.1). However, asymptotic efficiency cannot be claimed in general because of misspecification of the conditional distribution and/or the temporal dependencies.

In the preceding discussion it is assumed that the order n_0 and the multi-index α are correctly specified. In practice the order will be specified using order selection techniques. Once the order is specified the specification of α is less critical as each set $M(n_0;\alpha)$ is open and dense in $M(n_0)$ and hence in a loose sense contains almost all transfer functions. Also the consistency proof of HD, Section 4.2, is coordinate independent, showing that the quasi ML esti-

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

mator of $k_0(z)$ obtained by maximizing the likelihood over $M(n_0)$ will lie in $M(n_0;\alpha)$ a.s. for large enough T for $k_0(z) \in M(n_0;\alpha)$. Therefore for correctly specified n_0 the restriction of only considering estimators in $M(n_0;\alpha)$ is not severe from an asymptotic standpoint. In what follows we will always assume that a correct multi-index is chosen.

Contrary to the quasi ML estimators described previously the CCA estimator is not defined implicitly as the optimizer of a criterion function but instead is given explicitly from the solutions of a sequence of least squares problems. The main idea of CCA builds on the properties of the state. For a detailed motivation of CCA see Bauer (2005b); here we will only provide the details of the algorithm. Defining $\mathcal{K}_p := [K,(A - KC)K,\dots,(A - KC)^{p-1}K], Y^-_{t,p} := [y'_{t-1}, y'_{t-2},\dots, y'_{t-p}]'$ it holds for $t \in \mathbb{Z}$ that

$$Y^+_{t,f} = \mathcal{O}_f x_t + \mathcal{E}_f E^+_{t,f} = \mathcal{O}_f \mathcal{K}_p Y^-_{t,p} + \mathcal{O}_f (A - KC)^p x_{t-p} + \mathcal{E}_f E^+_{t,f},$$

where $Y^+_{t,f} := [y'_t, y'_{t+1},\dots, y'_{t+f-1}]'$, and $E^+_{t,f} := [\varepsilon'_t, \varepsilon'_{t+1},\dots,\varepsilon'_{t+f-1}]'$. Further $\mathcal{O}_f := [C', A'C',\dots,(A')^{f-1}C']'$, and $\mathcal{E}_f$ denotes the matrix whose ith block row for $i > 1$ is equal to $[CA^{i-2}K,\dots,CK, I_s, 0^{s\times(f-i)s}]$, whereas the first block row is equal to $[I_s, 0^{s\times(f-1)s}]$. Here f and p are two integers to be chosen by the user (see the discussion that follows for details). In what follows we use the notation $\langle a_t, b_t\rangle_i^{T-j} := T^{-1}\sum_{t=i}^{T-j} a_t b'_t$ somewhat sloppily, using the same symbol for the processes $(a_t)_{t\in\mathbb{Z}}, (b_t)_{t\in\mathbb{Z}}$ and the random variables a_t and b_t, respectively, which should not cause any confusion.

Then CCA obtains an estimate of the system using the following steps.

1. Obtain a rank n matrix $\hat{\mathcal{O}}_f\hat{\mathcal{K}}_p$ estimating $\mathcal{O}_f\mathcal{K}_p$ by using rank restricted regression techniques in the equation $Y^+_{t,f} = \beta Y^-_{t,p} + N^+_{t,f}, t = p+1,\dots, T-f+1$ where $\beta \in \mathbb{R}^{fs\times ps}$, i.e.,

   $$\hat{\mathcal{O}}_f\hat{\mathcal{K}}_p := \arg\min_{\beta:\operatorname{rank}(\beta)=n} \operatorname{tr}[\hat{W}\langle Y^+_{t,f} - \beta Y^-_{t,p}, Y^+_{t,f} - \beta Y^-_{t,p}\rangle_{p+1}^{T-f+1}],$$

   where $\hat{W} = \hat{W}' > 0$ is a design variable to be chosen by the user. Typically $\hat{W} = (\hat{W}^+_f)'\hat{W}^+_f$ for a nonsingular matrix $\hat{W}^+_f$ is used. The estimate can be obtained using the singular value decomposition (SVD)

   $$\hat{W}^+_f\langle Y^+_{t,f}, Y^-_{t,p}\rangle_{p+1}^{T-f+1}(\langle Y^-_{t,p}, Y^-_{t,p}\rangle_{p+1}^{T-f+1})^{-1/2} = \hat{U}\hat{\Sigma}\hat{V}' = \hat{U}_n\hat{\Sigma}_n\hat{V}'_n + \hat{R}_n,$$

   where $\hat{U}_n \in \mathbb{R}^{sf\times n}$ denotes the matrix of left singular vectors corresponding to the n dominating singular values and $\hat{\Sigma}_n = \operatorname{diag}(\hat{\sigma}_1,\dots,\hat{\sigma}_n)$, $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_n > 0$, denotes the diagonal matrix containing the dominant n singular values ordered decreasing in size. Here $\hat{V}_n \in \mathbb{R}^{sp\times n}$ conformably contains the right singular vectors. Further $\hat{R}_n$ is a residual term collecting the contribution of the remaining singular values. This leads to

   $$\hat{\mathcal{O}}_f := (\hat{W}^+_f)^{-1}\hat{U}_n\hat{\Sigma}_n, \quad \hat{\mathcal{K}}_p := \hat{V}'_n(\langle Y^-_{t,p}, Y^-_{t,p}\rangle_{p+1}^{T-f+1})^{-1/2}.$$

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

2. Estimate the state as $\hat{x}_t := \hat{\mathcal{K}}_p Y^-_{t,p}, t = p+1,\dots,T+1$. The system matrix estimates are obtained using least squares fitting in the system equations where the estimate $\hat{x}_t$ of the state is used in place of the true state x_t. Then

$$\hat{C} := \langle y_t, \hat{x}_t\rangle^T_{p+1}(\langle \hat{x}_t, \hat{x}_t\rangle^T_{p+1})^{-1}, \qquad \hat{A} := \langle \hat{x}_{t+1}, \hat{x}_t\rangle^T_{p+1}(\langle \hat{x}_t, \hat{x}_t\rangle^T_{p+1})^{-1},$$
$$\hat{\varepsilon}_t := y_t - \hat{C}\hat{x}_t, t = p+1,\dots,T, \qquad \hat{K} := \langle \hat{x}_{t+1}, \hat{\varepsilon}_t\rangle^T_{p+1}(\langle \hat{\varepsilon}_t, \hat{\varepsilon}_t\rangle^T_{p+1})^{-1}.$$

In step 1 note that $\hat{W} := (\langle Y^+_{t,f}, Y^+_{t,f}\rangle^{T-f+1}_{p+1})^{-1}$ corresponds to the maximum likelihood estimate assuming Gaussian i.i.d. errors $(N^+_{t,f})_{t\in\mathbb{Z}}$. The algorithm obtained using this choice has been called CVA by Larimore (1983) and CCA by Deistler, Peternell, and Scherrer (1995).

Different decompositions of $\hat{\mathcal{O}}_f\hat{\mathcal{K}}_p$ into $\hat{\mathcal{O}}_f$ and $\hat{\mathcal{K}}_p$ correspond to different choices of the basis for $\hat{x}_t$ and result in observationally equivalent estimated systems. Therefore the estimated transfer function is not affected by the choice of the decomposition.

One possibility for estimating the order n of the system is to use the estimated singular values. Define the criterion

$$SVC(n) := F(\hat{R}_n) + \frac{d(n)C_T}{T}, \tag{3}$$

which can be minimized over $n = 0,1,\dots,H_T$ for some upper bound $H_T \le \min(fs, ps)$ to obtain the order estimate. Here $d(n) := 2ns$ denotes the number of parameters in a model of order n and $C_T > 0, C_T \to \infty, C_T/T \to 0$ denotes a penalty factor. Although the form is similar to the definition of information criteria no relation has been established. Typical choices for $F(X): \mathbb{R}^{fs\times ps} \to [0,\infty)$ are $F(X) = \|X\|^2_{Fr}$ (Peternell, 1995), $F(X) = \|X\|^2_2$ (Bauer, 2001), and restricted to the case of CCA weights $F(\hat{R}_n) := -\sum_{j=n+1}^{\min(fs,ps)} \log(1-\hat{\sigma}_j^2)$ (Camba-Mendez and Kapetanios, 2001).

An open point in the description of the algorithm is the choice of the integers f and p. In the conditionally homoskedastic stationary case use of $f = p = 2\hat{p}_{AIC}$ has been suggested, where $\hat{p}_{AIC}$ denotes the Akaike information criterion (AIC) order estimate in an autoregressive approximation of y_t (see Deistler et al., 1995, p. 1866). Also in the case of conditionally heteroskedastic innovations this choice seems to be useful because the motivation for this choice is Theorem 6.6.3 of HD, which remains valid under our assumptions. Throughout the text we will use two different scenarios for the choice of the user parameters as shown in Assumption 2.

Assumption 2. The integer f and the weighting matrix $\hat{W}^+_f \in \mathbb{R}^{fs\times fs}$ are chosen corresponding to one of the following two possibilities:

(a) $f \ge n_0$ fixed and independent of the sample size where n_0 denotes the true order, and $\hat{W}^+_f$ is such that there exists a positive definite deterministic matrix W^+_f where $\hat{W}^+_f - W^+_f = O(Q_T), Q_T := \sqrt{\log\log T/T}$.

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

(b) $f = p$ depending on the sample size, and $\hat{W}_f^+ = (\langle Y_{t,f}^+, Y_{t,f}^+ \rangle_{p+1}^{T-f+1})^{-1/2}$. This choice will be labeled CCA in what follows.

In any case $p \to \infty, p \leq (\log T)^a$ a.s. for some $a < \infty$.

In this paper we will somewhat loosely use the term *subspace method* to denote any algorithm as described earlier where f, p, and $\hat{W}_f^+$ are chosen as in Assumption 2. Note that this does not conform with the literature, where the term *subspace method* is used for a larger class of algorithms including, e.g., the MOESP algorithm.

Note that we do not specify how p is chosen. In particular choosing p based on the autoregressive order estimator $\hat{p}_{AIC}$ as described previously is included in Assumption 2.

4. ASYMPTOTIC PROPERTIES

In this section we state the results of this paper. The true system will throughout be denoted as (A_0, K_0, C_0), and $k_0(z) := \pi(A_0, K_0, C_0)$ denotes the true transfer function where $k_0(z) \in M(n_0; \alpha)$ and hence n_0 denotes the true order. The multi-index α here is not necessarily unique. We will always assume that in the estimation a suitable index has been chosen.

The key to the results in the conditionally homoskedastic case was the representation of the subspace estimators approximately as a nonlinear function of the estimated covariance sequence. Then the asymptotic properties follow from a verification that the nonlinear mapping is differentiable using the well-known properties of the estimators of the covariance sequence. The extension to the conditionally heteroskedastic case follows the same lines: With respect to consistency the uniform convergence for covariance sequences as stated in HD (Thm. 5.3.2; see also (5.3.7)) builds the cornerstone. Let $\hat{\gamma}_j := \langle y_t, y_{t-j} \rangle_{j+1}^T$ for $0 \leq j < T$ and $\dot{\gamma}_j := \sum_{i=0}^{\infty} K_{i+j} \dot{\Omega} K_i'$, where $\dot{\Omega} := \langle \varepsilon_t, \varepsilon_t \rangle_1^T$. Then (5.3.7) of HD states that for $H_T = O((\log T)^a)$ for $a < \infty$ it holds that

$$\max_{0 \leq j \leq H_T} \|\hat{\gamma}_j - \dot{\gamma}_j\| = O(Q_T). \tag{4}$$

Note that $\dot{\gamma}_j - \mathbb{E} y_t y_{t-j}' = \sum_{i=0}^{\infty} K_{i+j} (\dot{\Omega} - \Omega_0) K_i'$ also converges to zero according to the ergodicity assumption. However, to the best of my knowledge no result exists ensuring convergence of order $O(Q_T)$ for $\dot{\Omega} - \Omega_0$ under the Assumptions 1 on $(\varepsilon_t)_{t \in \mathbb{Z}}$; cf. the discussion in HD, p. 327. The lack of such a result, however, does not conflict with the rate of convergence of the subspace estimators as follows from the next result, which is proved in the Appendix.

THEOREM 1. *Let $(y_t)_{t \in \mathbb{Z}}$ be generated by a stable and strictly minimum-phase state space system (A_0, K_0, C_0) of order n_0, where the innovations process $(\varepsilon_t)_{t \in \mathbb{Z}}$ fulfills Assumption 1. Let $\theta_{0,\alpha} := \varphi_\alpha^{-1}(\pi(A_0, K_0, C_0))$ denote*

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

the true parameter vector where the index α is chosen such that $k_0(z) := \pi(A_0, K_0, C_0) \in M(n_0;\alpha)$. Further let $\hat{\theta}_{sub,\alpha} := \varphi_\alpha^{-1}(\pi(\hat{A}, \hat{K}, \hat{C}))$ for $\pi(\hat{A}, \hat{K}, \hat{C}) \in M(n_0;\alpha)$ and $\hat{\theta}_{sub,\alpha} := \theta_c$ else for some arbitrary parameter vector $\theta_c \in \Theta_\alpha$. Here $(\hat{A}, \hat{K}, \hat{C})$ denotes the subspace estimator based on $y_t, t = 1, \ldots, T$ using the true order n_0. The subspace estimator depends on the choice of $\hat{W}_f^+$ and f, p.

(i) Then under Assumption 2 it holds that $\hat{k}(z) := \pi(\hat{A}, \hat{K}, \hat{C}) \in M(n_0;\alpha)$ a.s. for large enough T and further $\hat{\theta}_{sub,\alpha} \to \theta_{0,\alpha}$ a.s.

(ii) If $\liminf_{T\to\infty} - 2p \log \rho / \log T > 1$ a.s. holds where $0 \le \rho_0 = |\lambda_{max}(A_0 - K_0 C_0)| < \rho < 1$, then $\|\hat{\theta}_{sub,\alpha} - \theta_{0,\alpha}\| = O(Q_T)$.

Part (i) of the theorem states consistency for $k_0(z) \in M(n_0)$ because for any α such that $k_0(z) \in M(n_0;\alpha)$ we obtain $\hat{\theta}_{sub,\alpha} \to \theta_{0,\alpha}$ implying $\varphi_\alpha(\hat{\theta}_{sub,\alpha}) \to k_0(z)$. The estimators $\varphi_\alpha(\hat{\theta}_{sub,\alpha})$ differ for different values of α, but for $k_0(z) \in M(n_0;\alpha_1) \cap M(n_0;\alpha_2)$ for $\alpha_1 \neq \alpha_2$ the difference $\varphi_{\alpha_1}(\hat{\theta}_{sub,\alpha_1}) - \varphi_{\alpha_2}(\hat{\theta}_{sub,\alpha_2})$ is nonzero only for $\hat{k}(z) \in M(n_0) - M(n_0;\alpha_i)$ for one $i = 1, 2$. This does not happen a.s. for large enough T because $M(n_0;\alpha)$ is open in $M(n_0)$ for each α. Thus the consistency result holds coordinate free as in the quasi ML case. To quantify the rate of convergence, however, the choice of a particular coordinate system is convenient.

Note that the consistency result given here also holds for stochastic choice of p, e.g., for the suggestion $p := 2\hat{p}_{AIC}$. Moreover according to HD, Theorem 6.6.3, in this case the lower bound for the increase of p holds a.s., and hence $\|\hat{\theta}_{sub,\alpha} - \theta_{0,\alpha}\| = O(Q_T)$.

Part (ii) of the theorem gives a close to maximal result in terms of the speed of convergence of the estimator being of the form of a law of the iterated logarithm except for the fact that the constant in the $O(\cdot)$ statement is not evaluated. The same rate also applies in the conditionally homoskedastic case; hence in this sense the algorithms are robust with respect to the conditional heteroskedasticity of the innovations subject to Assumption 1.

Beside consistency the asymptotic distribution also is of interest. As in the conditionally homoskedastic case the estimators can be shown to fulfill a central limit theorem (CLT).

THEOREM 2. *Let the assumptions of Theorem 1 hold. Assume that* $\liminf_{T\to\infty} - 2p \log \rho / \log T > 1$ *holds a.s. where ρ is as in Theorem 1. Then under Assumption 2, $\sqrt{T}(\hat{\theta}_{sub,\alpha} - \theta_{0,\alpha}) \xrightarrow{d} Z$ where Z is multivariate normally distributed with mean zero and some variance V that depends on the choice of f and $\hat{W}_f^+$.*

Again the choice $p := 2\hat{p}_{AIC}$ fulfills the assumptions of the result. The proof of the theorem is contained in the Appendix. The asymptotic variance V depends on both f and $\hat{W}_f^+$. Because the parameters are the free entries in the system matrices an expression for V can easily be obtained from (using $r := f + p - 1$)

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

$$\operatorname{vec}(\hat{A} - A_0, \hat{K} - K_0, \hat{C} - C_0)$$
$$= M_r \operatorname{vec}[\langle \varepsilon_t, y_{t-1}\rangle_{r+1}^T, \langle \varepsilon_t, y_{t-2}\rangle_{r+1}^T, \ldots, \langle \varepsilon_t, y_{t-r}\rangle_{r+1}^T] + o_P(T^{-1/2}),$$

where equations for M_r can be obtained using the following set of equations:

$$\hat{C} - C_0 = \langle \varepsilon_t, Y_{t,p}^-\rangle_{p+1}^T \mathcal{K}_p'(\mathcal{K}_p \dot{\Gamma}_p^- \mathcal{K}_p')^{-1} - C_0(\hat{\mathcal{K}}_p - \mathcal{K}_p)\dot{\Gamma}_p^- \mathcal{K}_p'(\mathcal{K}_p \dot{\Gamma}_p^- \mathcal{K}_p')^{-1} + o_P(T^{-1/2}),$$
$$\hat{\mathcal{K}}_p - \mathcal{K}_p = \mathcal{O}_f^\dagger(\hat{\beta} - \mathcal{O}_f \mathcal{K}_p)P_{\mathcal{K}} + o_P(T^{-1/2}),$$
$$\hat{\beta} - \mathcal{O}_f \mathcal{K}_p = \mathcal{E}_f \langle E_{t,f}^+, Y_{t,p}^-\rangle_{p+1}^{T-f+1}(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1})^{-1} + o(T^{-1/2})$$

and similar equations for $\hat{A} - A_0$ and $\hat{K} - K_0$. Here $\dot{\Gamma}_p^- := [\dot{\gamma}_{j-i}]_{i,j=1,\ldots,p}$. These equations hold under the assumptions of Theorem 2 as is shown in the Appendix. Here in all equations $(\hat{A}, \hat{K}, \hat{C}) = (A(\hat{\theta}_{sub,\alpha}), K(\hat{\theta}_{sub,\alpha}), C(\hat{\theta}_{sub,\alpha}))$; i.e., the estimators for the system matrices have been transformed into the coordinate system corresponding to the overlapping form. All other matrices such as $\hat{\mathcal{K}}_p, \mathcal{O}_f, \mathcal{K}_p$ are transformed accordingly. Finally $P_{\mathcal{K}} := (I - S_p \mathcal{K}_p), \mathcal{O}_f^\dagger := (\mathcal{O}_f' W \mathcal{O}_f)^{-1}\mathcal{O}_f' W$ where $W = (\hat{W}_f^+)'\hat{W}_f^+$ or $W = (\dot{\Gamma}_f^+)^{-1}$ in the CCA case. Here $S_p = [S_{n_0}', 0]', S_{n_0} \in \mathbb{R}^{n_0 s \times n_0}$ denotes a matrix composed of columns of the $n_0 s$-dimensional identity matrix such that $\det \mathcal{K}_p S_p \neq 0$.

Note that the matrix M_r depends on the true system. Using the estimated system in place of the true system and denoting the corresponding matrix with $M_r(\hat{\tau})$ it is straightforward to show the consistency for any fixed r. The exponential decrease of the entries of $M_r(\hat{\tau})$, which can be shown to be uniform in a sufficiently small compact neighborhood of the true value (see the end of the proof of Theorem 2), then shows convergence also for $r \to \infty$. In the univariate case expressions for the asymptotic variance, say, B_k, of $\sqrt{T}\operatorname{vec}[\langle \varepsilon_t, y_{t-1}\rangle_{k+1}^T, \ldots, \langle \varepsilon_t, y_{t-k}\rangle_{k+1}^T]$ can be found in Goncalves and Kilian (2003).[5] Under stronger assumptions on the innovations including conditions on sums of cumulants up to the eighth order (for details, see Goncalves and Kilian, 2003, Assumps. 1(i), (ii), and (iii′), pp. 4, 7), a consistent (uniformly in $1 \leq k = r \leq (\log T)^a$) estimate $\hat{B}_k$ such that $\|\hat{B}_k - B_k\| = o_P(1)$ also is provided. Consequently in the univariate case a consistent estimate for V is obtained as $\hat{V}_r := M_r(\hat{\tau})\hat{B}_r M_r(\hat{\tau})'$. We will not attempt to extend this result to the multivariate case or try to weaken the assumptions on the existence of moments but instead take a different approach.

For the conditionally homoskedastic case the equivalence to quasi ML estimation is shown in Bauer (2005a). It turns out that the equivalence also prevails for the conditionally heteroskedastic situation.

THEOREM 3. *Let the assumptions of Theorem 1 hold. Assume that the CCA algorithm using $f = p, p \geq -\log T/(2\log\rho)$, where ρ is as in Theorem 1, $p \leq (\log T)^a$ for $a < \infty$ is used based on $y_t, t = 1, \ldots, T$, to obtain the estimator*

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

$\hat{\theta}_{CCA,\alpha}$. Let $(\hat{\theta}_{ML,\alpha}, \hat{\omega})$ denote the estimate obtained by minimizing $L(y_t, [\theta', \omega']')$. Then $\sqrt{T}\|\hat{\theta}_{CCA,\alpha} - \hat{\theta}_{ML,\alpha}\| = o_P(1)$.

For the proof see the Appendix. Therefore for estimators fulfilling the restrictions of Theorem 3 the asymptotic equivalence admits a different method for estimating the asymptotic variance of the CCA estimators via the asymptotic variance of the quasi ML estimator without the need to strengthen the assumptions on the innovations: Let $\hat{\varepsilon}_t(\theta)$ denote the estimates of the innovations obtained using the steady state Kalman filter under the assumption of zero initial state calculated as (for $t \geq 1$)

$$\begin{aligned}\hat{x}_{t+1}(\theta) &= (A(\theta) - K(\theta)C(\theta))\hat{x}_t(\theta) + K(\theta)y_t,\\ \hat{\varepsilon}_t(\theta) &= y_t - C(\theta)\hat{x}_t(\theta),\end{aligned} \tag{5}$$

where $\hat{x}_1(\theta) = 0$. Note that $\hat{\varepsilon}_t(\theta)$ does not denote the innovations corresponding to the quasi ML estimation problem because in this case the nonstationary Kalman filter is used.

Central in the expressions for the asymptotic variance is the derivative $\psi_t(\theta)$ of the estimated innovations $\hat{\varepsilon}_t(\theta)$. Using the Kalman filter recursions (5) it is simple to check that $\psi_{t,j}(\theta)$, say, the jth column of $\psi_t(\theta)$ (i.e., the derivative of $\hat{\varepsilon}_t(\theta)$ with respect to the jth coordinate of θ) can be calculated as (∂_j denoting partial derivative with respect to the jth coordinate):

$$\begin{aligned}\partial_j \hat{x}_{t+1}(\theta) &= (A(\theta) - K(\theta)C(\theta))\partial_j \hat{x}_t(\theta)\\ &\quad + (\partial_j A(\theta) - [\partial_j K(\theta)]C(\theta) - K(\theta)[\partial_j C(\theta)])\hat{x}_t(\theta) + \partial_j K(\theta)y_t,\\ \psi_{t,j}(\theta) &= -\partial_j C(\theta)\hat{x}_t(\theta) - C(\theta)\partial_j \hat{x}_t(\theta).\end{aligned} \tag{6}$$

Using this notation the next theorem provides expressions for the asymptotic variance.

THEOREM 4. *Let the assumptions of Theorem 3 hold. Let $\hat{\varepsilon}_t(\theta)$ denote the estimate of the innovation ε_t corresponding to the parameters θ obtained using the steady state Kalman filter according to (5) and let $\psi_t(\theta) := [\partial_1 \hat{\varepsilon}_t(\theta), \ldots, \partial_{2sn_0} \hat{\varepsilon}_t(\theta)] \in \mathbb{R}^{s \times 2sn_0}$ denote the matrix of the partial derivatives of $\hat{\varepsilon}_t(\theta)$ with respect to the coordinates of θ evaluated at θ calculated according to (6).*

(i) Then $\sqrt{T}(\hat{\theta}_{ML,\alpha} - \theta_{0,\alpha}) \xrightarrow{d} Z$ where Z is multivariate normally distributed with mean zero and variance $V_{ML} := \lim_{T\to\infty} \mathcal{Q}_T^{-1}\mathcal{P}_T\mathcal{Q}_T^{-1}$ where

$$\mathcal{Q}_T := \mathbb{E}\psi_T(\theta_{0,\alpha})'\Omega_0^{-1}\psi_T(\theta_{0,\alpha}),$$

$$\mathcal{P}_T := \mathbb{E}\psi_T(\theta_{0,\alpha})'\Omega_0^{-1}\varepsilon_T\varepsilon_T'\Omega_0^{-1}\psi_T(\theta_{0,\alpha}).$$

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

(ii) A consistent estimator of V_{ML} is given by $\hat{V}_{ML} := \hat{\mathcal{Q}}_T^{-1}\hat{\mathcal{P}}_T\hat{\mathcal{Q}}_T^{-1}$ where

$$\hat{\mathcal{Q}}_T := \frac{1}{T}\sum_{t=1}^{T}\psi_t(\hat{\theta}_{CCA,\alpha})'\hat{\Omega}^{-1}\psi_t(\hat{\theta}_{CCA,\alpha}),$$

$$\hat{\mathcal{P}}_T := \frac{1}{T}\sum_{t=1}^{T}\psi_t(\hat{\theta}_{CCA,\alpha})'\hat{\Omega}^{-1}\hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha})\hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha})'\hat{\Omega}^{-1}\psi_t(\hat{\theta}_{CCA,\alpha}),$$

where $\hat{\Omega} := \langle \hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha}), \hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha})\rangle_1^T$.

A proof of this theorem can be found in the Appendix. Note that no assumptions exceeding those contained in Theorem 3 are needed. From the form of the asymptotic variance V_{ML} it can be verified using standard arguments that in general the quasi ML estimators—and consequently also the CCA estimators—are not efficient: In the scalar case the asymptotic variance is identical to the asymptotic variance obtained from suitably scaling the estimation error in an ordinary least squares (OLS) regression of $\Omega_0^{-1/2}\varepsilon_t$ onto $\Omega_0^{-1/2}\psi_t(D_0,\theta_{0,\alpha})$, which in general is a generalized least squares (GLS) problem with block diagonal covariance matrix having $\mathbb{E}\{\Omega_0^{-1/2}\varepsilon_t\varepsilon_t'\Omega_0^{-1/2}|\mathcal{F}_{t-1}\}$ as its tth diagonal block. For the simple case of a regression with ARCH(1) errors this has been exemplified in detail in Section 4.2 of Gourieroux (1997).

All results stated previously deal with the case where the true order of the system is known. The last result of this paper provides sufficient conditions for the order estimation obtained by minimizing $SVC(n)$ to be consistent.

THEOREM 5. *Let the assumptions of Theorem 1 hold. Assume that the order is estimated using $SVC(n)$ for $F(\hat{R}_n) = \hat{\sigma}_{n+1}^2$. Let $\hat{n}$ be obtained as the minimal minimizing argument of $SVC(n)$ over $0 \le n \le \min(fs, ps)$. Assume further that Assumptions 2 hold. Then $C_T > 0, C_T/T \to 0, C_T/(fp\log\log T) \to \infty$ a.s. is a sufficient condition for consistency of $\hat{n}$, i.e., $\hat{n} \to n_0$ a.s. where n_0 denotes the true order of the data generating process.*

Again the proof of this theorem is given in the Appendix.

5. DETERMINISTIC TERMS

In the previous sections of the paper it was assumed that the process $(y_t)_{t\in\mathbb{Z}}$ is a stationary process of the form $y_t = \sum_{j=0}^{\infty}K_j\varepsilon_{t-j}$. In practice data preprocessing contains elements such as mean extraction and compensating for seasonal factors. Hence let $T_t := [1, \sin(\omega_1 t), \cos(\omega_1 t), \ldots, \sin(\omega_k t), \cos(\omega_k t)]'$ for some integer k such that $0 < \omega_1 < \cdots < \omega_k < \pi$. Then, if the subspace methods are applied using $\hat{y}_t := y_t - \langle y_t, T_t\rangle_1^T(\langle T_t, T_t\rangle_1^T)^{-1}T_t, t = 1, \ldots, T$ as the data, all results of the paper continue to hold if the observed process equals $(y_t + D_0T_t)_{t\in\mathbb{Z}}$ for some matrix $D_0 \in \mathbb{R}^{s\times 2k+1}$. In particular the corresponding estimator $\hat{\theta}_{sub,\alpha}$ will be consistent for $(y_t)_{t\in\mathbb{Z}}, f, p$, and $\hat{W}_f^+$ fulfilling the

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

assumptions of Theorem 1; a CLT will hold for the assumptions of Theorem 2. The estimators are asymptotically equivalent to quasi ML estimation under the assumptions of Theorem 3 where the asymptotic variance can be estimated using Theorem 4. The asymptotic variance is identical to the one obtained using $y_t, t = 1, \ldots, T$ as the data. This implies that eliminating the deterministic terms by regression prior to estimation does not carry a penalty in terms of asymptotic variance as is to be expected. Finally the order can be estimated consistently using $SVC(n)$ according to Theorem 5.[6]

6. CONCLUSIONS

In this paper the knowledge of the asymptotic properties of estimators obtained using subspace algorithms has been extended to the conditionally heteroskedastic case by a more careful study of the assumptions sufficient for the previously published results. The main message is that—except for the expressions for the asymptotic variance—all results achieved for the conditionally homoskedastic case also carry over to certain cases of conditionally heteroskedastic innovations. In particular the equivalence of CCA estimators to quasi ML estimators ignoring the conditional heteroskedasticity prevails. This implies that the CCA estimators are not asymptotically efficient in this case in general. Hence they should not be used as final estimators. Nevertheless they are seen as valuable tools, as they provide estimates for multivariable systems in situations where quasi ML estimation is numerically not feasible.[7] In this situation CCA can be used to obtain good initial estimates for subsequently constructing efficient estimators.

NOTES

1. Here asymptotic variance denotes the variance of the limiting Gaussian variable.
2. An earlier version of this paper—which can be obtained from the author on request—contained simulations demonstrating this point.
3. Originally the algorithm was proposed under the name CVA (canonical variate analysis).
4. Here we as usual implicitly assume that a minimum exists. This is justified from asymptotic arguments; see below.
5. This reference has been pointed out by an anonymous referee, who is gratefully acknowledged.
6. A proof of these facts can be found in an earlier version of this paper to be obtained from the author on request.
7. Simulations underpinning this behavior can be obtained from the author on request.

REFERENCES

Aoki, M. (1990) *State Space Modeling of Time Series*. Springer-Verlag.

Aoki, M. & A. Havenner (1991) State space modeling of multiple time series. *Econometric Reviews* 10, 1–59.

Bauer, D. (1998) Some asymptotic theory for the estimation of linear systems using maximum likelihood methods or subspace algorithms. Ph.D. Thesis, TU Wien, Austria.

Bauer, D. (2001) Order estimation for subspace methods. *Automatica* 37, 1561–1573.

Bauer, D. (2005a) Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs. *Journal of Time Series Analysis* 26, 631–668.

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

Bauer, D. (2005b) Estimating linear dynamical systems using subspace methods. *Econometric Theory* 21, 181–211.
Bauer, D., M. Deistler, & W. Scherrer (1999) Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs. *Automatica* 35, 1243–1254.
Bauer, D. & L. Ljung (2002) Some facts about the choice of the weighting matrices in Larimore type of subspace algorithms. *Automatica* 38, 763–773.
Billingsley, P. (1968) *Convergence of Probability Measures*. Wiley.
Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–327.
Bougerol, P. & N. Picard (1992) Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics* 52, 115–127.
Camba-Mendez, G. & G. Kapetanios (2001) Testing the rank of the Hankel covariance matrix: A statistical approach. *IEEE Transactions on Automatic Control* 46, 331–336.
Deistler, M., K. Peternell, & W. Scherrer (1995) Consistency and relative efficiency of subspace methods. *Automatica* 31, 1865–1875.
Engle, R. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1007.
Engle, R. & K. Kroner (1995) Multivariate simultaneous generalized ARCH. *Econometric Theory* 11, 122–150.
Golub, G. & C. VanLoan (1989) *Matrix Computations*, 2nd ed. Johns Hopkins University Press.
Goncalves, S. & L. Kilian (2003) Asymptotic and Bootstrap Inference for AR(∞) Processes with Conditional Heteroskedasticity. Technical report 2003s-28, CIRANO.
Gourieroux, C. (1997) *ARCH Models and Financial Applications*. Springer Series in Statistics. Springer-Verlag.
Hannan, E.J. & M. Deistler (1988) *The Statistical Theory of Linear Systems*. Wiley.
Ho, B. & R.E. Kalman (1966) Efficient construction of linear state variable models from input/output functions. *Regelungstechnik* 14, 545–548.
Kürsteiner, G. (2001) Optimal instrumental variables estimation for ARMA models. *Journal of Econometrics* 104, 359–405.
Kürsteiner, G. (2002) Efficient IV estimation for autoregressive models with conditional heteroskedasticity. *Econometric Theory* 18, 547–583.
Larimore, W.E. (1983) System identification, reduced order filters and modeling via canonical variate analysis. In H.S. Rao & P. Dorato (eds.), *Proceedings of the 1983 American Control Conference 2*, pp. 445–451. IEEE Service Center.
Ling, S. & M. McAleer (2002) Necessary and sufficient moment conditions for the GARCH(r,s) and asymmetric power GARCH(r,s) models. *Econometric Theory* 18, 722–729.
Nelson, D. (1991) Conditional heteroskedasticity in asset returns. A new approach. *Econometrica* 59, 347–370.
Peternell, K. (1995) Identification of linear dynamic systems by subspace and realization-based algorithms. Ph.D. Thesis, TU Wien, Austria.
Rahbek, A., E. Hansen, & J. Dennis (2003) ARCH Innovations and Their Impact on Cointegration Rank Testing. Technical report, Department of Theoretical Statistics, University of Copenhagen.
Van Overschee, P. & B. DeMoor (1994) N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica* 30, 75–93.
Verhaegen, M. (1994) Identification of the deterministic part of mimo state space models given in innovations form from input-output data. *Automatica* 30, 61–74.

APPENDIX: Proofs

LEMMA 1. *Let* $\hat{\Phi} \in \mathbb{R}^{a\times b}$ *and* $\hat{\Psi} \in \mathbb{R}^{b\times c}$ *where the dimensions* a,b,c *possibly depend on the sample size. Assume that there exist matrices* $\Phi \in \mathbb{R}^{a\times b}$ *and* $\Psi \in \mathbb{R}^{b\times c}$ *such*

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

that $\hat{\Phi} - \Phi = O(a_T)$ *and* $\hat{\Psi} - \Psi = O(b_T)$. *Moreover assume that* $\max(\limsup_{a,b} \|\Phi\|_\infty, \limsup_{b,c} \|\Psi'\|_\infty) < \infty$. *Then* $\hat{\Phi}\hat{\Psi} - \Phi\Psi = O(\max(a_T, b_T, a_T b_T b))$. *If* Φ *is equal to zero, then* $\hat{\Phi}\hat{\Psi} = O(\max(a_T, a_T b_T b))$.

Proof. Note that $\hat{\Phi}\hat{\Psi} - \Phi\Psi = (\hat{\Phi} - \Phi)\hat{\Psi} + \Phi(\hat{\Psi} - \Psi) = (\hat{\Phi} - \Phi)(\hat{\Psi} - \Psi) + (\hat{\Phi} - \Phi)\Psi + \Phi(\hat{\Psi} - \Psi)$. Because of the uniform bound on the infinity norms the latter two terms are of the order $O(a_T)$ and $O(b_T)$, respectively. The entries in the first term are of order $O(a_T b_T b)$, being the sum of b products of elements in $\hat{\Phi} - \Phi$ and $\hat{\Psi} - \Psi$, respectively. If Φ is zero then the last term is zero, and hence b_T does not appear. ■

Note that extensions to products containing more than two matrices are straightforward: $\hat{\Phi}\hat{\Psi}\hat{\Theta} - \Phi\Psi\Theta = (\hat{\Phi}\hat{\Psi} - \Phi\Psi)\hat{\Theta} + \Phi\Psi(\hat{\Theta} - \Theta) = (\hat{\Phi}\hat{\Psi} - \Phi\Psi)\Theta + (\hat{\Phi}\hat{\Psi} - \Phi\Psi)(\hat{\Theta} - \Theta) + \Phi\Psi(\hat{\Theta} - \Theta) = O(\max(a_T, b_T, c_T))$ for $b_T b \to 0, c_T c \to 0, \hat{\Theta} - \Theta = O(c_T)$, $\limsup \|\Phi\Psi\|_\infty < \infty$, and $\limsup \|\Theta'\|_\infty < \infty$ in addition to the assumptions of the lemma.

LEMMA 2.

(i) Let $a_t(\theta) := \sum_{j=0}^{\infty} A_j(\theta)\varepsilon_{t-j}, b_t(\theta) := \sum_{j=0}^{\infty} B_j(\theta)\varepsilon_{t-j}$, *where* $(\varepsilon_t)_{t\in\mathbb{Z}}$ *fulfills Assumption 1. Assume that the matrices* $A_j(\theta)$ *and* $B_j(\theta)$ *are bounded uniformly for* $\theta \in \Theta$*:*

$$\|A_j(\theta)\| \le \mu\zeta^j, \qquad \|B_j(\theta)\| \le \mu\zeta^j, \qquad 0 < \mu < \infty, \quad \forall j \ge 0 \quad \text{for some } 0 < \zeta < 1.$$

Then

$$\max_{1\le f,p\le(\log T)^a} \max_{-f<j\le p} \sup_{\theta\in\Theta} \|\langle a_t(\theta), b_t(\theta)\rangle_1^T - \langle a_{t-j}(\theta), b_{t-j}(\theta)\rangle_{p+1}^{T-f+1}\| = o_P(T^{-1/2})$$

and

$$\max_{-f<j\le p} \sup_{\theta\in\Theta} \|\langle a_t(\theta), b_t(\theta)\rangle_1^T - \langle a_{t-j}(\theta), b_{t-j}(\theta)\rangle_{p+1}^{T-f+1}\| = O(Q_T) + o((f+p)/\sqrt{T})$$

uniformly in $1 \le f, p \le (\log T)^a$ *for arbitrary* $a < \infty$. *For* $a_t(\theta) := \sum_{j=0}^{t-1} A_j(\theta) y_{t-j}, b_t(\theta) := \sum_{j=0}^{t-1} B_j(\theta) y_{t-j}$ *the same results hold.*

(ii) Uniformly in $1 \le f, p \le (\log T)^a$ *we have* $\max_{0\le i\le f, 1\le j\le p} \|\langle \varepsilon_{t+i}, y_{t-j}\rangle_{p+1}^{T-f+1}\| = O(Q_T)$.

Proof.

(i) The proof of the first bound is identical to the proof of Lemma B.3 of Bauer (2005a), wherein additional to Assumption 1 on the noise sequence imposed in this paper the condition $\mathbb{E}\{\varepsilon_t \varepsilon_t' | \mathcal{F}_{t-1}\} = \mathbb{E}\{\varepsilon_t \varepsilon_t'\}$ and a condition on the conditional third moment also are imposed but are not used in the proof of the first statement. Hence the result remains valid under the current assumptions. For the second claim note that in the proof of Lemma B.3 of Bauer (2005a) two types of error appear: Neglecting observations at the start of the sample and at the end of the sample. For the start of the sample the bounds provided in Bauer (2005a) do not use the conditional homoskedasticity and are based on $\langle a_t(\theta), b_t(\theta)\rangle \to \mathbb{E}a_t(\theta)b_t(\theta)'$ a.s., which continues to hold. Hence these arguments are seen to remain valid. The derivation of the almost sure bounds for the error due to omitting observations at the end of the sample, however, makes use of conditional

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

homoskedasticity by including the rate of convergence $\langle a_t(\theta), b_t(\theta)\rangle - \mathbb{E}a_t(\theta)b_t(\theta)' = O(Q_T)$, which cannot be assumed any more. For the in probability statements strict stationarity implies that omitting observations at the end of the sample introduces an error that is $o_P(T^{-1/2})$ because omitting observations at the start of the sample results in an error of order $o(T^{-1/2})$. Further at most $(f+p)$ observations are omitted, where each term is of the form $a_t(\theta)b_t(\theta)'$. As in the proof of Lemma B.3 of Bauer (2005a) it follows that $a_t(\theta) = o(t^{1/4})$, and the same holds for $b_t(\theta)$. This shows the second claim. Then it follows as in the proof of Lemma B.3 that replacing ε_t with y_t does not change the results. Details are omitted.

(ii) Because $\langle \varepsilon_t, y_{t-j}\rangle_{j+1}^{T} = O(Q_T), j > 0$ according to Theorem 5.3.1 on p. 166 of HD where $\mathbb{E}\varepsilon_t y_{t-j}' = 0$, it follows that $\langle \varepsilon_t, y_{t-j}\rangle_{j+1}^{T} - \langle \varepsilon_t, y_{t-j}\rangle_{j+1}^{T-f-p+1} = O(Q_{T-f-p+1}) = O(Q_T)$ uniformly in the given range for f, p. The claim then follows analogously to (i). Again details are omitted. ■

LEMMA 3. *Let the assumptions of Theorem 1 hold and let* $\dot{\Gamma}_p^- := [\dot{\gamma}(j-i)]_{i,j=1,\dots,p}$ *where* $\dot{\gamma}(i) = \sum_{j=0}^{\infty} K_{i+j}\dot{\Omega}K_j', i \geq 0$ *and* $\dot{\gamma}(i) = \dot{\gamma}(-i)', i < 0$. *Then* $\|\dot{\Gamma}_\infty^-\|_\infty < \infty$ *a.s.,* $\sup_{1\leq p\leq((\log T)^a)}\|\dot{\Gamma}_p^-\| = O(1)$, $\sup_{1\leq p\leq((\log T)^a)}\|(\dot{\Gamma}_p^-)^{-1}\| = O(1)$, *and* $\sup_{1\leq p\leq((\log T)^a)}\|(\dot{\Gamma}_p^-)^{-1}\|_\infty = O(1)$ *for* $a < \infty$. *The same bounds hold for* $\dot{\Gamma}_p^-$ *replaced with* $\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1}$.

Proof. The norm bounds for $\Gamma_p^- := \mathbb{E}Y_{t,p}^-(Y_{t,p}^-)'$ can be found in Theorems 6.6.10 and 6.6.11 of HD. In these theorems the only assumption on Ω_0 is nonsingularity. Then using a continuity argument noting that Γ_p^- depends linearly on Ω_0 the proofs hold uniformly in a compact neighborhood of every nonsingular Ω. Because $\hat{\Omega} \to \Omega_0$ a.s. as a result of ergodicity the norm bounds hold for $\dot{\Gamma}_p^-$ a.s. for large enough T. Now according to (5.3.7) on p. 169 of HD, $\hat{\gamma}_j - \dot{\gamma}_j = O(Q_T)$ uniformly for $0 \leq j \leq (\log T)^a$. In $\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1}$ entries of the form $\langle y_{t-r}, y_{t-s}\rangle_{p+1}^{T-f+1}$ appear for $1 \leq r, s \leq p$. For $r \geq s$ we have

$$\hat{\gamma}_{r-s} - \langle y_{t-s}, y_{t-r}\rangle_{p+1}^{T-f+1} = O(Q_T) + o((f+p)/\sqrt{T})$$

according to Lemma 2(i). Therefore $\|\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1} - \dot{\Gamma}_p^-\|_{Fr} = O(Q_T p + p(f+p)/\sqrt{T}) = o(1)$ uniformly in the range given for f, p. Analogously $\|\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1} - \dot{\Gamma}_p^-\|_\infty = O(Q_T p + p(f+p)/\sqrt{T}) = o(1)$ follows. Consequently the norm bounds also hold for $\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1}$. ■

LEMMA 4. *Let the assumptions of Theorem 1 hold and let* $\tilde{\mathcal{K}}_p - \mathcal{K}_p^\circ = O(\max(Q_T, \rho^p))$ *where* $\mathcal{K}_p^\circ := [K_\circ, (A_\circ - K_\circ C_\circ)K_\circ, \dots, (A_\circ - K_\circ C_\circ)^{p-1}K_\circ]$ *and* $\tilde{\mathcal{K}}_p = \tilde{T}\hat{\mathcal{K}}_p$ *for nonsingular (a.s. for large enough T) matrix* $\tilde{T} \in \mathbb{R}^{n_0\times n_0}$. *Here* $(A_\circ, K_\circ, C_\circ)$ *denotes a minimal representation of the true system, and* $0 \leq |\lambda_{max}(A_\circ - K_\circ C_\circ)| < \rho < 1$. *Then*

$$(A_\circ - K_\circ C_\circ)^p\langle x_{t-p}, z_t\rangle_{p+1}^{T} = O(\rho^p),$$

$$\tilde{\mathcal{K}}_p\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T}\tilde{\mathcal{K}}_p' - \mathcal{K}_p^\circ\dot{\Gamma}_p^-(\mathcal{K}_p^\circ)' = O(\max(Q_T, \rho^p)),$$

where $(z_t)_{t\in\mathbb{Z}}$ *is any process such that* $\langle z_t, z_t\rangle_{p+1}^{T} = O(1)$.

Proof. The first claim is immediate from $(A_\circ - K_\circ C_\circ)^p = O(\rho^p)$ and $\langle x_{t-p}, z_t\rangle_{p+1}^{T} = O(1)$, which follows from a componentwise application of the Cauchy–Schwarz inequal-

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

ity noting that $\langle x_{t-p}, x_{t-p}\rangle_{p+1}^{T} = O(1)$ as is straightforward to show and $\langle z_t, z_t\rangle_{p+1}^{T} = O(1)$ by assumption.

The second claim follows from repeated application of $\tilde{\mathcal{K}}_p - \mathcal{K}_p^\circ = O(\max(Q_T, \rho^p))$ and Lemma 1 as

$$\tilde{\mathcal{K}}_p\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T}\tilde{\mathcal{K}}_p' - \mathcal{K}_p^\circ\dot{\Gamma}_p^-(\mathcal{K}_p^\circ)' = (\tilde{\mathcal{K}}_p\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T} - \mathcal{K}_p^\circ\dot{\Gamma}_p^-)\tilde{\mathcal{K}}_p' + \mathcal{K}_p^\circ\dot{\Gamma}_p^-(\tilde{\mathcal{K}}_p - \mathcal{K}_p^\circ)'$$

$$= \mathcal{K}_p^\circ(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T} - \dot{\Gamma}_p^-)(\mathcal{K}_p^\circ)' + O(\max(Q_T, \rho^p)).$$

Here the infinity norm of $\mathcal{K}_p^\circ\dot{\Gamma}_p^-, \dot{\Gamma}_p^-$, and $\mathcal{K}_p^\circ\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T}$ is uniformly bounded (see Lemma 3 for $\dot{\Gamma}_p^-$ and $\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T}$; for $\mathcal{K}_p^\circ$ this follows from $|\lambda_{max}(A_\circ - K_\circ C_\circ)| < 1$). Then $x_t = \mathcal{K}_p^\circ Y_{t,p}^- + (A_\circ - K_\circ C_\circ)^p x_{t-p}$ where $\|(A_\circ - K_\circ C_\circ)^p\| = O(\rho^p)$ shows that

$$\mathcal{K}_p^\circ(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T} - \dot{\Gamma}_p^-)(\mathcal{K}_p^\circ)' = \langle x_t, x_t\rangle_{p+1}^{T} - \dot{\Sigma}_x + O(\rho^p) = O(Q_T + \rho^p),$$

where $\dot{\Sigma}_x = \sum_{j=0}^{\infty} A_\circ^j K_\circ \dot{\Omega} K_\circ'(A_\circ^j)'$ and the first claim is used to show that, e.g., $(A_\circ - K_\circ C_\circ)^p\langle x_{t-p}, Y_{t,p}^-\rangle_{p+1}^{T}(\mathcal{K}_p^\circ)' = O(\rho^p)$. Then the second claim follows from $\dot{\Sigma}_x - \langle x_t, x_t\rangle_{p+1}^{T} = O(Q_T)$ using arguments analogous to those contained in the proof of Lemma 2(i). ■

A.1. Proof of Theorem 1. The proof parallels the proof of Lemma A.1 in Bauer (2005a) with two differences: First it is necessary to replace the true covariance sequence with the sequence $\dot{\gamma}_j$ (see the beginning of Sect. 4 for a definition) to maintain the order of convergence of the covariance sequence estimates. Second in Theorem 1 the assumptions on f, p, and $\hat{W}_f^+$ are more general than in Lemma A.1. In particular no lower bound on the increase of p as a function of the sample size is imposed. This will lead to a number of changes in the proof. We will not provide a self-contained proof but rather highlight only the differences in the argumentation.

The central estimate in subspace procedures is the estimate $\hat{\mathcal{O}}_f\hat{\mathcal{K}}_p$ of $\mathcal{O}_f\mathcal{K}_p$. Let the unrestricted estimator be defined as $\hat{\beta} := \langle Y_{t,f}^+, Y_{t,p}^-\rangle_{p+1}^{T-f+1}(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1})^{-1}$ where $Y_{t,f}^+ := [y_t', \ldots, y_{t+f-1}']'$ and $Y_{t,p}^- := [y_{t-1}', \ldots, y_{t-p}']'$. Because $Y_{t,f}^+ = \mathcal{E}_f E_{t,f}^+ + \mathcal{O}_f\mathcal{K}_p Y_{t,p}^- + \mathcal{O}_f\bar{A}_0^p x_{t-p}$ for $\bar{A}_0 := A_0 - K_0 C_0$ we obtain

$$\hat{\beta} - \mathcal{O}_f\mathcal{K}_p = \mathcal{E}_f\langle E_{t,f}^+, Y_{t,p}^-\rangle_{p+1}^{T-f+1}(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1})^{-1}$$

$$+ \mathcal{O}_f\bar{A}_0^p\langle x_{t-p}, Y_{t,p}^-\rangle_{p+1}^{T-f+1}(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1})^{-1}.$$

Here (A_0, K_0, C_0) denotes a representation of the true system. From Lemma 2(ii) we obtain $\langle E_{t,f}^+, Y_{t,p}^-\rangle_{p+1}^{T-f+1} = O(Q_T)$. Then according to the uniform bound on $\|(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1})^{-1}\|_\infty$ (Lemma 3) and also $\sup_{f\in\mathbb{N}}\|\mathcal{E}_f\|_\infty < \infty$ as is easy to verify from the structure of $\mathcal{E}_f$ the first term is of order $O(Q_T)$ by Lemma 1. In the second term it is easy to show that $\langle x_{t-p}, Y_{t,p}^-\rangle_{p+1}^{T-f+1} = O(1)$ (cf. Lem. B.2 of Bauer, 2005a, which only uses ergodicity but not conditional homoskedasticity of $(\varepsilon_t)_{t\in\mathbb{Z}}$). Therefore $\mathcal{O}_f\bar{A}_0^p\langle x_{t-p}, Y_{t,p}^-\rangle_{p+1}^{T-f+1}(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1})^{-1} = O(\|\bar{A}_0^p\|)$. It is straightforward to show that $\|\bar{A}_0^p\| = O(\rho^p)$ for ρ as in the statement of the theorem. This shows

$$\hat{\beta} - \mathcal{O}_f\mathcal{K}_p = \mathcal{E}_f\langle E_{t,f}^+, Y_{t,p}^-\rangle_{p+1}^{T-f+1}(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1})^{-1} + O(\rho^p)$$

$$= \mathcal{E}_f\langle E_{t,f}^+, Y_{t,p}^-\rangle_{p+1}^{T-f+1}(\dot{\Gamma}_p^-)^{-1} + O(\rho^p) + o(T^{-1/2}). \qquad \textbf{(A.1)}$$

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

Consequently $\hat{\beta} - \mathcal{O}_f \mathcal{K}_p = O(\max(Q_T, \rho^p))$ for arbitrary $\rho_0 = |\lambda_{max}(A_0 - K_0 C_0)| < \rho < 1$. Hence for $\liminf_{T\to\infty} -2p \log \rho / \log T > 1$ a.s. it holds that $\hat{\beta} - \mathcal{O}_f \mathcal{K}_p = O(Q_T)$. Because all error bounds given previously hold uniform for p in the range given in the theorem they hold in particular for any possibly random sequence within the bounds.

The next step in the proof of Lemma A.1 is to derive a bound on the error in $\hat{\mathcal{K}}_p$ in a specific basis: Let $S_{n_0} \in \mathbb{R}^{n_0 s \times n_0}$ be defined as a matrix composed of n_0 columns of the $n_0 s$-dimensional identity matrix and let $S_p = [S'_{n_0}, 0]' \in \mathbb{R}^{ps \times n_0}$ where $p \geq n_0$. Then define $\check{\mathcal{K}}_p := [\hat{\mathcal{K}}_p S_p]^{-1} \hat{\mathcal{K}}_p$ where the nonsingularity of $\hat{\mathcal{K}}_p S_p$ for appropriate choice of S_p follows as in the proof of Lemma A.1 of Bauer (2005a). Accordingly the true matrix is denoted as $\check{\mathcal{K}}^\circ_p := [\mathcal{K}_p S_p]^{-1} \mathcal{K}_p$. Then the order of convergence $\hat{\beta} - \mathcal{O}_f \mathcal{K}_p = O(\max\{Q_T, \rho^p\})$ can be used analogously to the derivations in the proof of Lemma A.1 of Bauer (2005a) to show that

$$\check{\mathcal{K}}_p - \check{\mathcal{K}}^\circ_p = \mathcal{O}_f^\dagger (\hat{\beta} - \mathcal{O}_f \mathcal{K}_p) P_{\mathcal{K}} + O(\max\{f\rho^{2p}, Q_T^2 f^2 p\}), \tag{A.2}$$

where $P_{\mathcal{K}} := (I - S_p \check{\mathcal{K}}^\circ_p), \mathcal{O}_f^\dagger := ((\check{\mathcal{O}}^\circ_f)' W \check{\mathcal{O}}^\circ_f)^{-1} (\check{\mathcal{O}}^\circ_f)' W$ where $W = (\hat{W}_f^+)' \hat{W}_f^+$ or $W = (\dot{\Gamma}_f^+)^{-1}$ in the CCA case. Here $\check{\mathcal{O}}^\circ_f \check{\mathcal{K}}^\circ_p = \mathcal{O}_f \mathcal{K}_p$ defines $\check{\mathcal{O}}^\circ_f$. Thus for $p \geq -\log T/(2 \log \rho)$ it follows that $\check{\mathcal{K}}_p - \check{\mathcal{K}}^\circ_p = O(Q_T)$. Essential here is the bound

$$\|\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p - \mathcal{O}_f \mathcal{K}_p\| \leq \|\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p - \mathcal{O}_f \mathcal{K}_p\|_{Fr} \leq \delta \|\hat{\beta} - \mathcal{O}_f \mathcal{K}_p\|_{Fr} = O(Q_T \sqrt{fp} + \rho^p) = o(1)$$

for some constant $\delta < \infty$ (possibly redefining $\rho < 1$) where $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$ denotes the best rank n_0 approximation to $\hat{\beta}$ in a weighted Frobenius norm sense being defined via the SVD.

The next step in the proof of Lemma A.1 of Bauer (2005a) is the estimation of the matrix C in the coordinate system corresponding to $\check{\mathcal{K}}^\circ_p$. Let as in Lemma A.1 the system $(\check{A}_\circ, \check{K}_\circ, \check{C}_\circ)$ denote the representation of the true transfer function corresponding to $\check{\mathcal{K}}^\circ_p$. Further let $(\check{A}, \check{K}, \check{C})$ denote the system obtained from using the state $\check{x}_t := \check{\mathcal{K}}_p Y^-_{t,p}, t = p+1, \ldots, T+1$. Clearly $\pi(\check{A}, \check{K}, \check{C}) = \pi(\hat{A}, \hat{K}, \hat{C}) = \hat{k}(z)$. Then

$$\begin{aligned}
\check{C} &= \langle y_t, \check{x}_t \rangle_{p+1}^T (\langle \check{x}_t, \check{x}_t \rangle_{p+1}^T)^{-1} \\
&= \langle \varepsilon_t + \check{C}_\circ \check{\mathcal{K}}^\circ_p Y^-_{t,p} + \check{C}_\circ (\check{A}_\circ - \check{K}_\circ \check{C}_\circ)^p x_{t-p}, Y^-_{t,p} \rangle_{p+1}^T \check{\mathcal{K}}'_p (\check{\mathcal{K}}_p \hat{\Gamma}_p^- \check{\mathcal{K}}'_p)^{-1} \\
&= \check{C}_\circ + \langle \varepsilon_t, Y^-_{t,p} \rangle_{p+1}^T \check{\mathcal{K}}'_p (\check{\mathcal{K}}_p \hat{\Gamma}_p^- \check{\mathcal{K}}'_p)^{-1} - \check{C}_\circ (\check{\mathcal{K}}_p - \check{\mathcal{K}}^\circ_p) \hat{\Gamma}_p^- \check{\mathcal{K}}'_p (\check{\mathcal{K}}_p \hat{\Gamma}_p^- \check{\mathcal{K}}'_p)^{-1} + O(\rho^p) \\
&= \check{C}_\circ + O(Q_T) + O(\max\{Q_T, \rho^p\}) + O(\rho^p) = \check{C}_\circ + O(\max\{Q_T, \rho^p\})
\end{aligned} \tag{A.3}$$

using $\hat{\Gamma}_p^- := \langle Y^-_{t,p}, Y^-_{t,p} \rangle_{p+1}^T$. Here in the third equality the results of Lemma 4 are used; the fourth equality follows from $\langle \varepsilon_t, Y^-_{t,p} \rangle_{p+1}^T = O(Q_T)$ (Lemma 2(ii)) and the uniform bounds on $\|\check{\mathcal{K}}_p\|_\infty$ and $\|(\check{\mathcal{K}}_p \hat{\Gamma}_p^- (\check{\mathcal{K}}_p)')^{-1}\|$, which are easily established from the uniform norm bound on $\check{\mathcal{K}}^\circ_p$ and the convergence $\check{\mathcal{K}}_p - \check{\mathcal{K}}^\circ_p = O(\max\{Q_T, \rho^p\})$ and also the facts that $\check{\mathcal{K}}^\circ_p$ contains the identity matrix as a submatrix and the eigenvalues of $\hat{\Gamma}_p^-$ are uniformly bounded away from zero a.s. (see Lemma 3). Similar arguments show $\check{A} - \check{A}_\circ = O(\max(Q_T, \rho^p))$ and $\check{K} - \check{K}_\circ = O(\max(Q_T, \rho^p))$. Because $p \to \infty$ a.s. we obtain $\rho^p = o(1)$. Therefore consistency of $(\check{A}, \check{K}, \check{C})$ follows, and consequently $\hat{k}(z) := \pi(\check{A}, \check{K}, \check{C}) \to k_0(z)$ by continuity of π. For p fulfilling the lower bound it follows that $\rho^p = o(T^{-1/2})$ for $\rho_0 < \rho$ small enough, and hence $\check{A} - \check{A}_\circ = O(Q_T), \check{K} - \check{K}_\circ = O(Q_T), \check{C} - \check{C}_\circ = O(Q_T)$. Here we again use the uniformity of the achieved bounds in p for the given range to conclude the convergence also in the case of stochastic choice of p. The rest of

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

the proof consists in suitably transforming the basis of the state space conforming to the multi-index α and is identical to the remainder of the proof of Lemma A.1 of Bauer (2005a) and hence omitted. This proves the theorem. ■

A.2. Proof of Theorem 2. Again the proof of the theorem follows similar ideas as the proof for the CLT in the case of conditionally homoskedastic innovations as contained in Bauer, Deistler, and Scherrer (1999) also using the arguments contained in Bauer and Ljung (2002). Besides the more general assumptions on the innovations the main difference with respect to the results presented in Bauer et al. (1999) lies in the fact that a different decomposition of $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$ into $\hat{\mathcal{O}}_f$ and $\hat{\mathcal{K}}_p$ is used, which allows a bigger set of transfer functions to be included in the results.

To simplify the notation let $(\hat{A}, \hat{K}, \hat{C}) = (A(\hat{\theta}_{sub,\alpha}), K(\hat{\theta}_{sub,\alpha}), C(\hat{\theta}_{sub,\alpha}))$ and $(A_0, K_0, C_0) = (A(\theta_{0,\alpha}), K(\theta_{0,\alpha}), C(\theta_{0,\alpha}))$ such that both systems correspond to a particular coordinate neighborhood of the overlapping forms, i.e., to a particular multi-index α. Because in the overlapping forms used in this paper the entries of the system matrices are either free parameters or restricted to zero or one it is sufficient to show asymptotic normality of the estimation errors in the system matrices, which will be done subsequently.

Lemma 3 of Bauer and Ljung (2002) shows that

$$\operatorname{vec}(\hat{A} - A_0, \hat{K} - K_0, \hat{C} - C_0) = \overline{M}_1 \operatorname{vec}[\langle \varepsilon_t, x_t \rangle_{p+1}^T] + \overline{M}_{2,p} \operatorname{vec}(\check{\mathcal{K}}_p - \check{\mathcal{K}}_p^\circ) + o_P(T^{-1/2}). \tag{A.4}$$

(To be precise the lemma shows the result for a different coordinate system. The change of the coordinate system, however, does not change the result as follows easily from using the Delta method.) For $\hat{C} - C_0$ this is almost immediate from (A.3). Here $\overline{M}_1 \in \mathbb{R}^{(n_0^2+2n_0 s)\times n_0 s}$ and $\overline{M}_{2,p} \in \mathbb{R}^{(n_0^2+2n_0 s)\times n_0 ps}$ are matrices depending on $(A_\circ, K_\circ, C_\circ), \dot{\Omega}$, and $\dot{\Gamma}_p^-$. Hence both are random. Expressions for $\overline{M}_1$ and $\overline{M}_{2,p}$ can be found using the equations on p. 766 of Bauer and Ljung (2002). Because the main argument used in the proof of Lemma 3 of Bauer and Ljung (2002) is the uniform convergence of the estimated covariance sequence the proof of the lemma is unaltered (after exchanging all expressions involving γ_j by the identical expressions using $\dot{\gamma}_j$) under the assumptions on the noise of this paper with one exception: In Bauer and Ljung (2002) it is argued that $\sqrt{T}\langle \varepsilon_t - C_0(\hat{x}_t - x_t), \hat{x}_t \rangle_{p+1}^T$ converges in distribution, which has not been shown under the current assumptions. However, this is not needed because the proof that

$$\langle \varepsilon_t - C_0(\hat{x}_t - x_t), \hat{x}_t \rangle_{p+1}^T = (\hat{C} - C_0)\langle \hat{x}_t, \hat{x}_t \rangle_{p+1}^T = O(Q_T)$$

is sufficient for the argument used in Bauer and Ljung (2002). Here $\hat{C} = \langle y_t, \hat{x}_t \rangle_{p+1}^T (\langle \hat{x}_t, \hat{x}_t \rangle_{p+1}^T)^{-1}$ is used. But $\hat{C} - C_0 = O(Q_T)$ has been shown in the proof of Theorem 1 given earlier and $\langle \hat{x}_t, \hat{x}_t \rangle_{p+1}^T = O(1)$ as is easy to verify using Lemma 4. Therefore the conclusions of the lemma also hold in the setting of the current paper. This lemma holds for $-\log T/(2 \log \rho) \leq p \leq (\log T)^a$, which holds a.s. for T large enough under the present assumptions. Changes on zero measure sets do not alter the distribution.

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

From equation (A.2) we obtain $\check{\mathcal{K}}_p - \check{\mathcal{K}}_p^\circ = \mathcal{O}_f^\dagger(\hat{\beta} - \mathcal{O}_f\mathcal{K}_p)P_{\mathcal{K}} + o(T^{-1/2})$, whose central component is according to (A.1) (note that in Theorem 2 the lower bound on p implies $\rho^p = o(T^{-1/2})$)

$$\hat{\beta} - \mathcal{O}_f\mathcal{K}_p = \mathcal{E}_f\langle E_{t,f}^+, Y_{t,p}^-\rangle_{p+1}^{T-f+1}(\dot{\Gamma}_p^-)^{-1} + o(T^{-1/2}).$$

Therefore the main ingredients in the asymptotic expression are terms of the form $\langle \varepsilon_{t+i}, y_{t-j}\rangle_{p+1}^{T-f+1} = \langle \varepsilon_t, y_{t-j-i}\rangle_{p+f}^{T} + o_P(T^{-1/2}), 0 \le i \le f-1, 1 \le j \le p$ using arguments similar to those applied in Lemma 2. The same terms also appear in the proof of $\langle \varepsilon_t, x_t\rangle_{p+1}^T = \langle \varepsilon_t, Y_{t,p}^-\rangle_{p+1}^T\mathcal{K}_p' + o(T^{-1/2})$ using Lemma 4.

The remainder of the proof is based on the arguments provided in HD, p. 145ff., in particular on Bernstein's lemma (see, e.g., HD, Lem. 4.3.3; see also Billingsley, 1968, Thm. 4.2). The proof provided here deviates much from the proof given in Bauer et al. (1999), and hence a self-contained proof will be given. Applying

$$\check{\mathcal{K}}_p - \check{\mathcal{K}}_p^\circ = \mathcal{O}_f^\dagger\mathcal{E}_f\langle E_{t,f}^+, Y_{t,p}^-\rangle_{p+1}^{T-f+1}(\dot{\Gamma}_p^-)^{-1}P_{\mathcal{K}} + o(T^{-1/2})$$

to (A.4) and expressing $\langle E_{t,f}^+, Y_{t,p}^-\rangle_{p+1}^{T-f+1}$ as a linear function of $\langle \varepsilon_t, y_{t-j}\rangle_{r+1}^T, j = 1, \ldots, r$ we obtain (using $r := f + p - 1$)

$$\begin{aligned}\operatorname{vec}(\hat{A} - A_0, \hat{K} - K_0, \hat{C} - C_0)\\ = M_r \operatorname{vec}[\langle \varepsilon_t, y_{t-1}\rangle_{r+1}^T, \langle \varepsilon_t, y_{t-2}\rangle_{r+1}^T, \ldots, \langle \varepsilon_t, y_{t-r}\rangle_{r+1}^T] + o_P(T^{-1/2}),\end{aligned} \tag{A.5}$$

where the matrix $M_r \in \mathbb{R}^{(2n_0+s+n_0^2)\times rs^2}$ depends on $\overline{M}_1, \overline{M}_{2,p}, (A_0, K_0, C_0)$, and $\dot{\Gamma}_p^-$. This approximation holds uniformly in the range of r implied by the range of p given in the theorem.

In the notation of Lemma 4.3.3 of HD we show that we can use $X_T = Y_T(m) + Z_T(m)$ where

$$X_T := \sqrt{T}M_r \operatorname{vec}[\langle \varepsilon_t, y_{t-1}\rangle_{r+1}^T, \langle \varepsilon_t, y_{t-2}\rangle_{r+1}^T, \ldots, \langle \varepsilon_t, y_{t-r}\rangle_{r+1}^T],$$

$$Y_T(m) := \sqrt{T}M_{r,1:m} \operatorname{vec}[\langle \varepsilon_t, y_{t-1}\rangle_{r+1}^T, \langle \varepsilon_t, y_{t-2}\rangle_{r+1}^T, \ldots, \langle \varepsilon_t, y_{t-m}\rangle_{r+1}^T],$$

$$Z_T(m) := \sqrt{T}M_{r,m+1:r} \operatorname{vec}[\langle \varepsilon_t, y_{t-m-1}\rangle_{r+1}^T, \ldots, \langle \varepsilon_t, y_{t-r}\rangle_{r+1}^T]$$

to deduce asymptotic normality of X_T and therefore of the system matrix estimates. Here m is a fixed integer, i.e., not depending on T. Further the notation $M_{r,a:b}$ is used to denote the matrix composed of the block columns (where the blocks correspond to the matrices given in the argument of the vectorization) indexed a up to (and including) b of M_r. Then for asymptotic normality of $\sqrt{T}\operatorname{vec}(\hat{A} - A_0, \hat{K} - K_0, \hat{C} - C_0)$ it is sufficient to show that for each $\zeta > 0, \eta > 0$

1. there exists m such that $Y_T(m)$ is asymptotically normal,
2. the asymptotic variance matrix $\Sigma(m)$ of $Y_T(m)$ converges to Σ for $m \to \infty$,
3. and $\mathbb{P}\{\|Z_T(m)\|_2 > \zeta\} < \eta$.

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

Using Bernstein's lemma again to write

$$X_T^* = \sqrt{T}\,\mathrm{vec}[\langle \varepsilon_t, y_{t-j}\rangle_{r+1}^T]$$

$$= \sum_{i=0}^{m^*} \sqrt{T}\,\mathrm{vec}[\langle \varepsilon_t, \varepsilon_{t-j-i}\rangle_{r+1}^T K_i'] + \sum_{i=m^*+1}^{\infty} \sqrt{T}\,\mathrm{vec}[\langle \varepsilon_t, \varepsilon_{t-j-i}\rangle_{r+1}^T K_i']$$

$$= Y_T^*(m^*) + Z_T^*(m^*)$$

Lemma 4.3.4 of HD can be used in combination with arguments analogous to those presented later to show that

$$\sqrt{T}\,\mathrm{vec}[\langle \varepsilon_t, y_{t-1}\rangle_{r+1}^T, \langle \varepsilon_t, y_{t-2}\rangle_{r+1}^T, \ldots, \langle \varepsilon_t, y_{t-m}\rangle_{r+1}^T]$$

converges to a multivariate normal distribution with zero mean and variance $\mathbb{E}[Y_{t,m}^-(Y_{t,m}^-)' \otimes \varepsilon_t \varepsilon_t']$ under Assumption 1. Therefore $M_{r,1:m} \to M_{1:m}$ for some deterministic matrix $M_{1:m}$ is a sufficient condition for $Y_T(m)$ to be asymptotically normal for every fixed m. For (ii) $\|M_{1:\infty}\|_\infty < \infty$ is sufficient because the variance matrix $\mathbb{E}[Y_{t,m}^-(Y_{t,m}^-)' \otimes \varepsilon_t \varepsilon_t']$ has elements decreasing exponentially with the distance to the diagonal.

It remains to verify (3), i.e., that by choosing m and T sufficiently large $\mathbb{P}\{\|Z_T(m)\|_2 > \zeta\} < \eta$ for arbitrary $\zeta > 0, \eta > 0$. This will be done subsequently. Assume for the moment that it holds that $\|[M_r, 0] - M_\infty\|_\infty = o(1)$ for some deterministic matrix $M_\infty \in \mathbb{R}^{(n_0^2+2n_0 s)\times\infty}$ with $\|M_\infty\|_\infty < \infty$ and for $R_r := \sup_{1\le j\le r}\|M_{r,j:j}\|_1/(\mu\rho_Z^j)$ for some $0 < \mu < \infty$ and $0 < \rho_Z < 1$ it holds that $\limsup_{r\to\infty} R_r = O(1)$. Then note that

$$\|Z_T(m)\|_1 = \sqrt{T}\|M_{r,m+1:r}\,\mathrm{vec}[\langle \varepsilon_t, y_{t-m-1}\rangle_{r+1}^T, \ldots, \langle \varepsilon_t, y_{t-r}\rangle_{r+1}^T]\|_1$$

$$\le \sum_{j=m+1}^{r} \|M_{r,j:j}\|_1 \sqrt{T}\|\mathrm{vec}\langle \varepsilon_t, y_{t-j}\rangle_{r+1}^T\|_1. \tag{A.6}$$

Further for a scalar sequence (i.e., $y_t \in \mathbb{R}$)

$$\mathbb{E}T\langle \varepsilon_t, y_{t-j}\rangle_{r+1}^T \langle \varepsilon_t, y_{t-j}\rangle_{r+1}^T = \frac{1}{T}\sum_{t,s=r+1}^{T} \mathbb{E}\varepsilon_t \varepsilon_s y_{t-j} y_{s-j}$$

$$= \frac{1}{T}\sum_{t=r+1}^{T} \mathbb{E}\varepsilon_t^2 y_{t-j}^2 \le (\mathbb{E}\varepsilon_t^4)^{1/2}(\mathbb{E}y_t^4)^{1/2} < \infty$$

because of the martingale difference assumption, strict stationarity, and assumed finite fourth moments of ε_t. Consequently $\sqrt{T}\mathbb{E}\|\mathrm{vec}\langle \varepsilon_t, y_{t-j}\rangle_{r+1}^T\|_1 \le (\mathbb{E}T\|\mathrm{vec}\langle \varepsilon_t, y_{t-j}\rangle_{r+1}^T\|_2^2)^{1/2} < \Lambda$ for some $\Lambda < \infty$.

For vector-valued ε_t the bound can be established for each coordinate separately. Note that the bound is uniform in $j > 0$ and $1 \le r \le T$. It follows that

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

$$\mathbb{P}\{\|Z_T(m)\|_2 > \zeta\} \le \mathbb{P}\{\|Z_T(m)\|_1 > \zeta\}$$
$$\le \mathbb{P}\{(\|Z_T(m)\|_1 > \zeta) \wedge (R_r > c)\} + \mathbb{P}\{(\|Z_T(m)\|_1 > \zeta) \wedge (R_r \le c)\}$$
$$\le \mathbb{P}\{R_r > c\} + \mathbb{P}\left\{ \sum_{j=m+1}^{r} \mu \rho_Z^j \sqrt{T} \|\mathrm{vec}\langle \varepsilon_t, y_{t-j}\rangle_{r+1}^T\|_1 > \zeta/c \right\}$$

using inequality (A.6) and the definition of R_r. The first probability can be made arbitrarily small by choosing c large enough because $R_r = O(1)$. With respect to the second term note that

$$\mathbb{P}\left\{ \sum_{j=m+1}^{r} \mu \rho_Z^j \sqrt{T} \|\mathrm{vec}\langle \varepsilon_t, y_{t-j}\rangle_{r+1}^T\|_1 > \zeta/c \right\}$$
$$\le \mathbb{E}\left(\sum_{j=m+1}^{r} \mu \rho_Z^j \sqrt{T} \|\mathrm{vec}\langle \varepsilon_t, y_{t-j}\rangle_{r+1}^T\|_1 \right) c/\zeta < \frac{c\Lambda\mu\rho_Z^{m+1}}{\zeta(1-\rho_Z)},$$

which can be made arbitrarily small by choosing m large enough because $\rho_Z < 1$.

Note that all of the preceding evaluations hold uniformly in r and thus in f, p. This implies that the particular choice of these integers has no influence on the result as long as the lower and upper bounds included in the assumptions hold (a.s. for large enough T).

Therefore to finish the proof it is sufficient to show that $\|[M_r, 0] - M_\infty\|_\infty = o(1)$ for some deterministic matrix $M_\infty \in \mathbb{R}^{(n_0^2+2n_0 s)\times\infty}$ with $\|M_\infty\|_\infty < \infty$ and $\limsup_{r\to\infty} R_r = O(1)$ for some $0 < \mu < \infty$ and $0 < \rho_Z < 1$. Note that M_r is the sum of a number of matrices. Clearly it is sufficient to verify the convergence for each summand separately. For the term corresponding to $\sqrt{T}\langle \varepsilon_t, x_t\rangle_{r+1}^T = \sqrt{T}\langle \varepsilon_t, Y_{t,p}^-\rangle_{r+1}^T \mathcal{K}_p' + o_P(1)$ the convergence is obvious. Corresponding to Lemma 3 of Bauer and Ljung (2002) $\overline{M}_{2,p}$ consists of a finite sum of terms of the form $N_j(L_{p,j} \otimes I_{n_0})$ where N_j are finite-dimensional matrices independent of the sample size and where $L_{p,j}$ is equal to either $[I_s, 0], \mathcal{K}_p\dot{\Gamma}_p^-$ or $\mathcal{K}_p[\dot{\mathcal{H}}_{1,p}', \dot{\Gamma}_p^-][I_{ps}, 0]'$ where $\dot{\mathcal{H}}_{1,p} := [\dot{\gamma}_1, \ldots, \dot{\gamma}_p]$. Note that

$$\overline{M}_{2,p}\,\mathrm{vec}(\check{\mathcal{K}}_p - \check{\mathcal{K}}_p^\circ) = \overline{M}_{2,p}((I - (\check{\mathcal{K}}_p^\circ)' S_p')(\dot{\Gamma}_p^-)^{-1} \otimes \mathcal{O}_f^\dagger \mathcal{E}_f)\mathrm{vec}[\langle E_{t,f}^+, Y_{t,p}^-\rangle_{p+1}^{T-f+1}]$$
$$+ o_P(T^{-1/2}).$$

Thus consider $L_{p,j}(\dot{\Gamma}_p^-)^{-1}$ for the terms given: For $L_{p,j} = [I_s, 0]$ it follows that $\|[L_{p,j}(\dot{\Gamma}_p^-)^{-1}, 0] - \dot{\Omega}_b^{-1}[I_s, -C_b\dot{\mathcal{K}}_\infty^b]\|_\infty \to 0$ a.s. Here $\dot{\mathcal{K}}_\infty^b$ has elements decreasing exponentially along the block columns (for definitions and details, see Bauer, 2005a, proof of Lem. 5.5). For $L_{p,j} = \mathcal{K}_p\dot{\Gamma}_p^-$ we obtain $L_{p,j}(\dot{\Gamma}_p^-)^{-1} = \mathcal{K}_p$ and finally $\|[\mathcal{K}_p[\dot{\mathcal{H}}_{1,p}', \dot{\Gamma}_p^-][I_{ps}, 0]'(\dot{\Gamma}_p^-)^{-1}, 0] - [\Sigma_x C_0', \mathcal{K}_\infty\Gamma_\infty^-](\Gamma_\infty^-)^{-1}\|_\infty \to 0$. Also the fact that $[S_p'(\dot{\Gamma}_p^-)^{-1}, 0]$ converges to a matrix whose elements decrease exponentially can be verified using the arguments in the proof of Lemma 5.5 of Bauer (2005a). In all cases the speed of convergence is elementwise of order $O(\rho^p)$ where $|\lambda_{max}(A_0 - K_0C_0)| < \rho < 1$. Clearly for each term, $L_{p,j}(\check{\mathcal{K}}_p^\circ)'$ converges a.s. for $p \to \infty$. This shows that $[\overline{M}_{2,p}((I - (\check{\mathcal{K}}_p^\circ)' S_p')(\dot{\Gamma}_p^-)^{-1} \otimes \mathcal{O}_f^\dagger \mathcal{E}_f), 0]$ converges for every fixed and finite f in infinity norm to a deterministic matrix with bounded infinity norm. Consequently for f fixed and finite convergence of $[M_r, 0]$ follows because for every $1 \le j \le r$ terms equivalent to $\langle \varepsilon_t, y_{t-j}\rangle_{r+1}^T$ up to initial and end effects in this case occur at most f times

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

in $\langle E_{t,f}^{+}, Y_{t,p}^{-}\rangle_{p+1}^{T-f+1}$. The uniform speed of convergence of order $O(\rho^p)$ in combination with the exponential decrease of the entries of M_∞ implies that $\limsup_{r\to\infty} R_r = O(1)$. In the case that $f = p \to \infty$ a more careful analysis of M_r leads to the same result. Because these evaluations are cumbersome we refrain from presenting them. Instead it is noted that asymptotic normality in this case is implied by Theorem 3. ■

A.3. Proof of Theorem 3. Again the proof consists in paralleling the proof for the conditionally homoskedastic situation with analyzing the arguments that involve the assumption $\mathbb{E}\{\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}\} = \Omega_0$ that has been used in Bauer (2005a) but is not imposed in the present paper. In what follows we will reconsider the steps of the proof of Theorem 3.1 of Bauer (2005a) by noting the points where the arguments have to be changed rather than presenting a self-contained proof. Such a proof would essentially replicate the arguments given in Bauer (2005a) as will be clear from the following analysis. Instead we will only discuss the changes in the proof to hold also under the assumptions of this paper. This implies that the following proof can only be understood once Bauer (2005a) has been studied. We also refrain from introducing concepts that are used in this proof but nowhere else in this paper. For details see Bauer (2005a). All cited lemmas and theorems refer either to the current paper or to Bauer (2005a).

The first three lemmas in Appendix B of Bauer (2005a) are not specific to subspace methods. Lemma B.1 obviously continues to hold. For Lemma B.2 γ_j has to be replaced by $\hat{\gamma}_j$ for the orders of convergence to hold. Note that $\max_{j\in\mathbb{Z}}\|\hat{\gamma}_j - \mathbb{E}y_t y_{t-j}'\| = o(1)$ continues to hold by setting $\hat{\gamma}_j := 0, |j| \geq T$ as a result of Theorem 7.4.1 of HD. Lemma B.2(ii) holds for in probability bounds replacing almost sure bounds as can be seen from the proof of Lemma 2 in the current paper. For the almost sure versions the slightly larger bounds $O(Q_T) + o((f+p)/\sqrt{T})$ follow from Lemma 2. Lemma B.2(iii) only uses the convergence $\langle y_t, y_t\rangle_1^T$ and thus continues to hold. Finally Lemma B.2(iv) uses the convergence of $\mathcal{K}_p(\theta_T) - \mathcal{K}_p$ (which clearly does not depend on the assumptions on the noise), the results of Lemma 2(ii) and boundedness of a number of estimated second moments, which continues to hold because of ergodicity. Lemma 2(i) provides the relevant results of Lemma B.3. In Lemma B.3 $\dot{\Omega} - \Omega_0 = O(Q_T)$ is used to show that neglecting observations at the end of the sample introduces an error of order $O(Q_T)$. This cannot be assumed any more in our setting. Throughout the proof, however, only the weaker in probability norm bound is needed, and this continues to hold according to Lemma 2(i). Lemma B.4 discusses the links between quasi ML estimation and prediction error methods. The proof of this lemma does not use the assumption on conditional homoskedasticity and hence can be shown to remain true under the current set of assumptions.

Appendix A of Bauer (2005a) describes the properties of the CCA estimator. Lemma A.1 is implied by Theorem 1 as has been noted earlier. The proof of Lemma A.2 holds unchanged because in no place is $\hat{\gamma}_j - \gamma_j$ used, but only $\hat{\gamma}_j = O(1), \langle \varepsilon_t, y_{t-j}\rangle_{p+1}^{T-f+1} = O(Q_T)$ (see Lemma 2) and Lemmas B.1, B.2, and B.3. The error bounds contained in the second line of Lemma A.3 hold for the calculated backward system using Ω_0, the true innovation variance, rather than an estimate thereof. The bound on the error between a fixed backward system corresponding to (A_0, K_0, C_0) and Ω_0 and the backward system calculated on the basis of the estimated forward system and $\dot{\Omega}$ cannot be shown with the tools provided previously to be of order $O(Q_T)$ because no result showing $\dot{\Omega} - \Omega_0 = O(Q_T)$ is known, to the best of my knowledge. However, if the true backward system $(A(\theta_{0,\alpha})_b, K(\theta_{0,\alpha})_b, C(\theta_{0,\alpha})_b)$ is calculated using (A_0, K_0, C_0) and $\dot{\Omega}$, it corresponds to a random choice of the basis because $\dot{\Omega}$ is random. Then the error between the backward

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

system calculated using $(\hat{A}, \hat{K}, \hat{C})$ and $\dot{\Omega}$ and $(A(\theta_{0,\alpha})_b, K(\theta_{0,\alpha})_b, C(\theta_{0,\alpha})_b)$ can be shown to be of order $O(Q_T)$ using the identical arguments as used in the proof of Lemma A.3. With respect to the estimation error for the estimated backward system note that all error bounds of Lemma A.2 and Lemma A.4 can be written (up to higher order terms) in terms of in probability bounds on the estimation error in the covariance sequence. It follows from Bauer et al. (1999) that the CCA system estimators (neglecting higher order terms) can be viewed as a nonlinear function of the estimated covariance sequence.

Analogously the estimated backward system $(\tilde{A}_b, \tilde{K}_b, \tilde{C}_b)$ can be seen to be obtained from the same mapping applied to the transpose of the estimated covariance sequence. Clearly the uniform almost sure error bounds on the covariance sequence estimates imply identical bounds for the transposed covariance sequence estimates. Note that also here the true system $(A(\theta_{0,\alpha})_b, K(\theta_{0,\alpha})_b, C(\theta_{0,\alpha})_b)$ does not correspond to a fixed basis because it corresponds to $\dot{\mathcal{K}}_\infty^b := \mathcal{O}_\infty'(\dot{\Gamma}_\infty^+)^{-1}$ generating the backward state. Hence the first part of Lemma A.3 and Lemma A.4 hold true with replacing almost sure statements by in probability versions. Therefore it is concluded that all results contained in the preliminary lemmas continue to hold under the less restrictive assumption given by Assumption 1 subject to the replacement of γ_j by $\dot{\gamma}_j$ for the uniform convergence argument and replacing almost sure results by in probability results.

In the main text of Bauer (2005a) it is straightforward to verify that all arguments prior to Lemma 5.1 hold unchanged given the fact that all preliminary lemmas hold with the appropriate changes indicated previously. In Lemma 5.1 the first claims only use the properties of the parameterization. To show that $\hat{\mathcal{O}}_f - \mathcal{O}_f(\theta_0) = O(Q_T)$ note that $\hat{\mathcal{O}}_f - \mathcal{O}_f(\theta_0) = \hat{\beta}\hat{\mathcal{K}}_p^\dagger - \mathcal{O}_f(\theta_0)\mathcal{K}_p(\theta_0)\mathcal{K}_p(\theta_0)^\dagger$ where $\hat{\beta} - \mathcal{O}_f(\theta_0)\mathcal{K}_p(\theta_0) = O(Q_T)$ (see the proof of Theorem 1) and $\hat{\mathcal{K}}_p^\dagger - \mathcal{K}_p(\theta_0)^\dagger = O(Q_T)$ for $\hat{\mathcal{K}}_p^\dagger := \langle Y_{t,p}^-, Y_{t,p}^- \rangle_{p+1}^{T-f+1} \hat{\mathcal{K}}_p'(\hat{\mathcal{K}}_p \langle Y_{t,p}^-, Y_{t,p}^- \rangle_{p+1}^{T-f+1} \hat{\mathcal{K}}_p')^{-1}$ and $\mathcal{K}_p(\theta_0)^\dagger := \langle Y_{t,p}^-, Y_{t,p}^- \rangle_{p+1}^{T-f+1} \mathcal{K}_p(\theta_0)'(\mathcal{K}_p(\theta_0) \langle Y_{t,p}^-, Y_{t,p}^- \rangle_{p+1}^{T-f+1} \mathcal{K}_p(\theta_0)')^{-1}$ as a result of $\hat{\mathcal{K}}_p - \mathcal{K}_p(\theta_0) = O(Q_T)$ (as a result of $\check{\mathcal{K}}_p - \check{\mathcal{K}}_p^\circ = O(Q_T)$ and differentiability of the transformation matrix relating $(\check{A}, \check{K}, \check{C})$ to $(A(\hat{\theta}_{CCA,\alpha}), K(\hat{\theta}_{CCA,\alpha}), C(\hat{\theta}_{CCA,\alpha}))$; see the end of the proof of Theorem 1). Further $\hat{\mathcal{O}}_f'(\hat{\Gamma}_f^+)^{-1} - \mathcal{O}_f(\theta_0)'(\dot{\Gamma}_f^+)^{-1} = O_P(Q_T)$ follows from $\hat{\Gamma}_f^+ - \dot{\Gamma}_f^+ = O_P(Q_T)$ (using (4) and Lemma 2). Moreover $\mathcal{O}_f(\theta_0)'(\dot{\Gamma}_f^+)^{-1} - \dot{\mathcal{K}}_f^b(\theta_0) = O(\rho^f)$ for $\rho_0 < \rho < 1$ holds where $\dot{\mathcal{K}}_f^b$ corresponds to the backward system corresponding to (A_0, K_0, C_0) and $\dot{\Omega}$ (again without changing the arguments of the proof of Lemma 5.1 of Bauer, 2005a). Further a matrix Z_f appears in the statement of Lemma 5.1, which uses the true covariance sequence in its definition. Replacing Ω_0 by $\dot{\Omega}$ we can introduce a matrix $\dot{Z}_f$ converging a.s. to $\dot{Z}_\infty$ such that the result given in Lemma 5.1 still holds (a.s.):

$$\dot{\Xi}_f^{-1}\mathcal{O}_f = (\dot{\Gamma}_f^+)^{-1}\mathcal{O}_f + (\dot{\Gamma}_f^+)^{-1}\mathcal{O}_f(\dot{\Sigma}_x^{-1} - \mathcal{O}_f'(\dot{\Gamma}_f^+)^{-1}\mathcal{O}_f)^{-1}\mathcal{O}_f'(\dot{\Gamma}_f^+)^{-1}\mathcal{O}_f = (\dot{\Gamma}_f^+)^{-1}\mathcal{O}_f\dot{Z}_f',$$

where all quantities correspond to the estimated system $(\hat{A}, \hat{K}, \hat{C})$ and $\dot{\Omega} = \langle \varepsilon_t, \varepsilon_t \rangle_1^T$ as the innovation variance. Here $\dot{\Xi}_f := \mathcal{E}_f(\hat{\theta}_{CCA,\alpha})(I_f \otimes \dot{\Omega})\mathcal{E}_f(\hat{\theta}_{CCA,\alpha})'$. This follows because in the derivation of the results for Z_f only the nonsingularity of the true covariance Ω_0 is used. Now $\dot{\Omega}$ is nonsingular a.s. for T large enough as a result of consistency. It is straightforward to see from the proof of Lemma 5.1 that all arguments for the convergence of $\dot{Z}_f$ continue to hold a.s. if the true covariances γ_j are replaced by $\dot{\gamma}_j$. The remaining statements of Lemma 5.1 obviously hold unchanged (a.s.). Therefore Lemma 5.1 also continues to hold.

Next note that $\hat{\Omega}(\hat{\theta}_{CCA,\alpha})$ in Bauer (2005a) is equal to $\langle \hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha}), \hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha}) \rangle_1^T = \langle \varepsilon_t, \varepsilon_t \rangle_1^T + o(T^{-1/2})$ as follows from

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

$$\hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha}) = [I_s, -C(\hat{\theta}_{CCA,\alpha})\mathcal{K}_p(\hat{\theta}_{CCA,\alpha})]Y^-_{t+1,p+1} - C(\hat{\theta}_{CCA,\alpha})\bar{A}(\hat{\theta}_{CCA,\alpha})^p x_{t-p}(\hat{\theta}_{CCA,\alpha})$$

$$= \varepsilon_t + [C(\theta_{0,\alpha})\mathcal{K}_p(\theta_{0,\alpha}) - C(\hat{\theta}_{CCA,\alpha})\mathcal{K}_p(\hat{\theta}_{CCA,\alpha})]Y^-_{t,p}$$

$$+ C(\theta_{0,\alpha})\bar{A}(\theta_{0,\alpha})^p x_{t-p}(\theta_{0,\alpha}) - C(\hat{\theta}_{CCA,\alpha})\bar{A}(\hat{\theta}_{CCA,\alpha})^p x_{t-p}(\hat{\theta}_{CCA,\alpha})$$

for $t > p$ using the consistency of $\hat{\theta}_{CCA,\alpha}$ and the results of Lemmas 5.1, 2(ii), and 4 stated previously. Therefore $\hat{\Omega}(\hat{\theta}_{CCA,\alpha})$ can be replaced with $\dot{\Omega}$ in (13) of Bauer (2005a) without changing the equality. This result is used in the proof of Lemma 5.2, where $Z_\infty \mathcal{K}^b_f(\hat{\theta}_{CCA,\alpha})$ now has to be replaced by $\dot{Z}_\infty \dot{\mathcal{K}}^b_f(\hat{\theta}_{CCA,\alpha})$ using

$$\dot{\mathcal{K}}^b_\infty(\hat{\theta}_{CCA,\alpha}) = \mathcal{O}_\infty(\hat{\theta}_{CCA,\alpha})'\Gamma^+_\infty(\hat{\theta}_{CCA,\alpha}, \dot{\Omega})^{-1}.$$

The evaluations to show that the lemma holds with these replacements are identical to those given in Bauer (2005a) and hence are omitted.

The paragraph following Lemma 5.2 discusses the replacement of $\dot{\mathcal{K}}^b_f(\hat{\theta}_{CCA,\alpha})$ by $\hat{\mathcal{O}}'_f(\hat{\Gamma}^+_f)^{-1}$. The discussion needs to be changed slightly because now $\hat{\Gamma}^+_f - \dot{\Gamma}^+_f = O_P(Q_T)$ rather than $O(Q_T)$ as a result of end effects resulting in a larger almost sure error bound in this case, which does not affect the conclusion, however.

Lemma 5.3 does not use conditional homoskedasticity, and hence its conclusions remain true under the current set of assumptions. In the lines just above Lemma 5.4 again the bound on the order of convergence for $\dot{\mathcal{K}}^b_f(\hat{\theta}_{CCA,\alpha}) - \hat{\mathcal{O}}'_f(\hat{\Gamma}^+_f)^{-1}$ is only in probability rather than almost sure. In Lemma 5.4 the error bound $\langle \hat{x}_t, y_{t-p-1}\rangle^T_{p+1} = o(T^{-\epsilon})$ is derived using the uniform convergence of sample covariances in combination with the fact $\mathbb{E}x_t y'_{t-p-1} = o(T^{-\epsilon})$ for suitably small $\epsilon > 0$. An alternative argument is

$$\langle \hat{x}_t, y_{t-p-1}\rangle^T_{p+1} = \langle x_t, y_{t-p-1}\rangle^T_{p+1} + O(Q_T)$$

$$= \langle A_0^p x_{t-p}, y_{t-p-1}\rangle^T_{p+1} + \sum_{j=1}^{p} A_0^{j-1} K_0 \langle \varepsilon_{t-j}, y_{t-p-1}\rangle^T_{p+1} + O(Q_T)$$

$$= \langle A_0^p x_{t-p}, y_{t-p-1}\rangle^T_{p+1} + O(Q_T) = o(T^{-\epsilon}).$$

This bound is derived from $x_t = A_0^p x_{t-p} + \sum_{j=1}^p A_0^{j-1} K_0 \varepsilon_{t-j}$ where $A_0^p = o(T^{-\epsilon})$ for p obeying the lower bound. Therefore replacing the true covariances with the dotted quantities leaves the error bound unchanged. Note that

$$\dot{\mathcal{K}}^b_f(\hat{\theta}_{CCA,\alpha}) = \mathcal{O}_f(\hat{\theta}_{CCA,\alpha})'\Gamma^+_f(\hat{\theta}_{CCA,\alpha}, \dot{\Omega})^{-1} + O(Q_T) = \hat{\mathcal{O}}'_f(\hat{\Gamma}^+_f)^{-1} + O_P(Q_T)$$

$$= \mathcal{K}^b_f(\hat{A}_b, \hat{K}_b, \hat{C}_b) + O_P(Q_T),$$

where the first equality has been stated previously. The second follows from Lemma 5.1 (discussed earlier), and the third equality is the backward analogue to $\hat{\mathcal{K}}_p - \mathcal{K}_p(\hat{\theta}_{CCA,\alpha}) = O(Q_T)$ according to Lemma 5.1 and the proof of Lemma A.1. Here $(\hat{A}_b, \hat{K}_b, \hat{C}_b)$ denotes the estimated backward system (for a definition, see Bauer, 2005a), and $\mathcal{K}^b_f(\hat{A}_b, \hat{K}_b, \hat{C}_b) := [\hat{K}_b, (\hat{A}_b - \hat{K}_b\hat{C}_b)\hat{K}_b, \ldots, (\hat{A}_b - \hat{K}_b\hat{C}_b)^{f-1}\hat{K}_b]$. Then the proof of Lemma 5.4 follows analogously to the conditionally homoskedastic case from

$$(\dot{\mathcal{K}}^b_f(\hat{\theta}_{CCA,\alpha}) - \mathcal{K}^b_f(\hat{A}_b, \hat{K}_b, \hat{C}_b))(\hat{\mathcal{O}}_f - \mathcal{O}_f(\hat{\theta}_{CCA,\alpha})) = o_P(T^{-1/2})$$

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

using the error bound on $\dot{\mathcal{K}}_f^b(\hat{\theta}_{CCA,\alpha}) - \mathcal{K}_f^b(\hat{A}_b, \hat{K}_b, \hat{C}_b)$ derived previously and $\hat{\mathcal{O}}_f - \mathcal{O}_f(\hat{\theta}_{CCA,\alpha}) = O(Q_T)$ by Lemma 5.1. This finishes the proof for derivatives with respect to entries in K.

Corresponding to derivatives with respect to entries in A the same reasoning holds as the previously established lemmas are used with the indicated changes (i.e., exchanging Ω_0 for $\dot{\Omega}$ in the formulations where needed and weakening almost sure bounds to in probability bounds). For derivatives with respect to entries in C the proof of Lemma 5.5 follows along the same lines as before. This concludes the proof of the theorem. ■

A.4. Proof of Theorem 4. Essential in this proof are the quantities $\hat{\varepsilon}_t(\theta)$ and $\psi_t(\theta)$ defined in (5) and (6), respectively. In what follows quantities built on the basis of $\hat{\theta}_{CCA,\alpha}$ will be denoted with a "hat." Quantities without a hat always refer to the true system.

Define the *prediction error* criterion function

$$\tilde{L}(y_t, \tau) = \log\det\Omega(\omega) + \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t(\theta)'\Omega(\omega)^{-1}\hat{\varepsilon}_t(\theta),$$

which differs from the criterion function used in the quasi ML formulation by the assumptions on y_0. Whereas in the quasi ML method it is assumed that y_0 corresponds to the stationary solution of the state space equations the prediction error methods assume $y_0 = 0$. Minimization of this criterion function defines the so-called prediction error estimator $\hat{\tau}_{PE,\alpha}$, which under the present set of assumptions can be shown to be asymptotically equivalent to the quasi ML estimator in the sense that $\sqrt{T}(\hat{\tau}_{ML,\alpha} - \hat{\tau}_{PE,\alpha}) = o_p(1)$. (A detailed proof of this fact is contained in a preliminary version of this paper, available from the author.)

Because the asymptotic distribution of $\hat{\theta}_{ML,\alpha}$ equals the asymptotic distribution of the prediction error estimator $\hat{\theta}_{PE,\alpha}$ we will use $\tilde{L}$ as the criterion function rather than L. Then Theorem 4.3.1 on p. 132 of HD states asymptotic normality of $\hat{\theta}_{PE,\alpha}$. The arguments for the derivation are standard and hence omitted. (A detailed derivation can also be found in a previous version of this paper, which can be obtained from the author on request.) Also the expressions for the asymptotic variance can be verified using the material contained in Section 4.3 of HD; see in particular equation (4.3.15) and the discussion following that equation. This shows part (i).

Note that

$$\hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha}) = \sum_{j=0}^{t-1}\hat{K}_j^- y_{t-j} = \sum_{j=0}^{t-1}\hat{K}_j^-\sum_{i=0}^{\infty}K_i\varepsilon_{t-j-i} = \sum_{l=0}^{\infty}\left(\sum_{j=0}^{\min(l,t-1)}\hat{K}_j^- K_{l-j}\right)\varepsilon_{t-l}.$$

Here $k^{-1}(z, \hat{\theta}_{CCA,\alpha}) = \sum_{i=0}^{\infty}\hat{K}_i^- z^i$ and $k_0(z) = \sum_{i=0}^{\infty}K_i z^i$. Given that $\|\hat{K}_i^-\| \le M_- \rho_-^i$ for some $0 < \rho_- < 1$ (which holds a.s. for large enough T because of the strict minimum-phase assumption and consistency) and $\|K_i\| \le M_k \rho_k^i$ where $\rho_- \ne \rho_k$ can be assumed without loss of generality we obtain with $\rho_* := \max(\rho_-, \rho_k)$

$$\left\|\sum_{j=0}^{\min(l,t-1)}\hat{K}_j^- K_{l-j}\right\| \le \mu\rho_*^l$$

for some constant μ depending on M_-, M_k, ρ_-, and ρ_k. Let $\hat{K}_{t,l} := \sum_{i=0}^{\min(l,t-1)}\hat{K}_i^- K_{l-i}$. Then

$$\sum_{l=0}^{\infty}\|\hat{K}_{t,l} - K_{t,l}\| \le \sum_{l=0}^{\infty}\sum_{i=0}^{l}\|\hat{K}_i^- - K_i^-\|\|K_{l-i}\| = \sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\|\hat{K}_i^- - K_i^-\|\|K_j\| = o(1).$$

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

Here $\sum_{j=0}^{\infty}\|K_j\| = O(1)$ follows from $\|K_i\| \le M_k \rho_k^i$ and $\sum_{i=0}^{\infty}\|\hat{K}_i^- - K_i^-\| = o(1)$ follows from the arguments given in the proof of Lemma 5.1 of Bauer (2005a).

The analogous representation holds for the columns $\psi_{t,r}(\hat{\theta}_{CCA,\alpha})$, say, of $\psi_t(\hat{\theta}_{CCA,\alpha})$:

$$\psi_{t,r}(\hat{\theta}_{CCA,\alpha}) = \sum_{j=0}^{t-1}\hat{L}_{j,r}\,y_{t-j} = \sum_{l=0}^{\infty}\left(\sum_{j=0}^{\min(l,t-1)}\hat{L}_{j,r}\,K_{l-j}\right)\varepsilon_{t-l},$$

where $\partial_r k^{-1}(z,\hat{\theta}_{CCA,\alpha}) = \sum_{j=1}^{\infty}\hat{L}_{j,r}\,z^j$ denotes the partial derivative with respect to the rth component of θ. Further let $\hat{L}_{t,l,r} := \sum_{i=0}^{\min(l,t-1)}\hat{L}_{i,r}\,K_{l-i}$. Analogously to $\hat{K}_{t,l}$ also $\|\hat{L}_{t,l,r}\| \le \mu\rho_*^l$ a.s. and $\sum_{l=0}^{\infty}\|\hat{L}_{t,l,r} - L_{t,l,r}\| = o(1)$ for large enough T.

With respect to (ii) note that a consistent estimator of $\lim_{T\to\infty}\mathcal{Q}_T$ is $\hat{\mathcal{Q}}_T := T^{-1}\sum_{t=1}^{T}\psi_t(\hat{\theta}_{CCA,\alpha})'\hat{\Omega}^{-1}\psi_t(\hat{\theta}_{CCA,\alpha})$, due to HD, (4.3.15), and the approximation properties of the prediction error criterion function $\partial^2_{\theta,\theta'}\tilde{L}(\hat{\tau}_{CCA,\alpha}) - 2\mathcal{Q}_T \to 0$ a.s. for $T\to\infty$ because of consistency of $\hat{\theta}_{CCA,\alpha}$. Here $\hat{\tau}_{CCA,\alpha} := [\hat{\theta}'_{CCA,\alpha}, \hat{\omega}']'$ for $\Omega(\hat{\omega}) = \hat{\Omega} = \langle\hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha}), \hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha})\rangle_1^T$. Then

$$\partial^2_{\theta,\theta'}\tilde{L}(\hat{\tau}_{CCA,\alpha}) - 2\hat{\mathcal{Q}}_T = 2\left[T^{-1}\sum_{t=1}^{T}\partial_r\psi_t(\hat{\theta}_{CCA,\alpha})'\hat{\Omega}^{-1}\hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha})\right]_{r=1,\dots,2sn_0} = o(1)$$

from standard arguments using the convergence of $\hat{K}_{t,l}$ and $\hat{L}_{t,l,r}$ given before because $\mathbb{E}\partial_r\psi_t(\tau_{0,\alpha})'\Omega_0^{-1}\varepsilon_t = 0$ (the notation $[.]_{r=1,\dots,2sn_0}$ indicates the columns of the matrix with typical element given).

Thus for (ii) it remains to verify that

$$\hat{\mathcal{P}}_T := T^{-1}\sum_{t=1}^{T}\psi_t(\hat{\theta}_{CCA,\alpha})'\hat{\Omega}^{-1}\hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha})\hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha})'\hat{\Omega}^{-1}\psi_t(\hat{\theta}_{CCA,\alpha}) \to \lim_{T\to\infty}\mathcal{P}_T,$$

where convergence is in probability. Because $\hat{\Omega}\to\Omega_0$ and $\hat{\theta}_{CCA,\alpha}\to\theta_{0,\alpha}$ the probability that the estimators $\hat{\tau}_{CCA,\alpha}$ are outside an arbitrarily small compact neighborhood Ψ_c of the true system tends to zero. Therefore it is sufficient to show convergence conditional on $\hat{\tau}_{CCA,\alpha}\in\Psi_c$. Consequently $\|\hat{\Omega}^{-1}\| < \lambda$, and the bounds $\|\hat{K}_j^-\| \le M_-\rho_-^j$ and $\|\hat{L}_{j,r}\| \le \mu M_-\rho_-^j, j\in\mathbb{N}, r = 1,\dots,2sn_0$, can be assumed to hold. From these bounds we obtain (using the usual notation for conditional expectation)

$$\mathbb{E}(\|(\hat{L}_{t,a,r}\varepsilon_{t-a})'\hat{\Omega}^{-1}\hat{K}_{t,b}\varepsilon_{t-b}\varepsilon'_{t-c}\hat{K}'_{t,c}\hat{\Omega}^{-1}\hat{L}_{t,d,s}\varepsilon_{t-d}\| \,|\, \hat{\tau}_{CCA,\alpha}\in\Psi_c) \le \mu\rho_*^{a+b+c+d}$$

for some constant μ not depending on a, b, c, or d.

Therefore consider $\hat{\mathcal{P}}_{T,r,s}$, say, the (r,s) element of $\hat{\mathcal{P}}_T$:

$$\begin{aligned}\hat{\mathcal{P}}_{T,r,s} &= T^{-1}\sum_{t=1}^{T}\psi_{t,r}(\hat{\theta}_{CCA,\alpha})'\hat{\Omega}^{-1}\hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha})\hat{\varepsilon}_t(\hat{\theta}_{CCA,\alpha})'\hat{\Omega}^{-1}\psi_{t,s}(\hat{\theta}_{CCA,\alpha})\\ &= T^{-1}\sum_{t=1}^{T}\sum_{a,b,c,d=0}^{\infty}\varepsilon'_{t-a}\hat{L}'_{t,a,r}\hat{\Omega}^{-1}\hat{K}_{t,b}\varepsilon_{t-b}\varepsilon'_{t-c}\hat{K}'_{t,c}\hat{\Omega}^{-1}\hat{L}_{t,d,s}\varepsilon_{t-d}\\ &= T^{-1}\sum_{t=1}^{T}\sum_{a,b,c,d=0}^{m}\varepsilon'_{t-a}\hat{L}'_{t,a,r}\hat{\Omega}^{-1}\hat{K}_{t,b}\varepsilon_{t-b}\varepsilon'_{t-c}\hat{K}'_{t,c}\hat{\Omega}^{-1}\hat{L}_{t,d,s}\varepsilon_{t-d} + R_{m,r,s}. \qquad \textbf{(A.7)}\end{aligned}$$

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

Here m is an arbitrary fixed integer. Thus $R_{m,r,s}$ collects all terms where at least one of a,b,c,d is bigger than m. It follows from the preceding discussion that $\mathbb{E}(|R_{m,r,s}|\,|\hat{\tau}_{CCA,\alpha} \in \Psi_c) \le \rho_*^m \delta$ for some constant $\delta < \infty$. On the other hand for every fixed m it follows from ergodicity and consistency of $\hat{L}_{t,a,r}$ and $\hat{K}_{t,l}$ that for any $0 \le a,b,c,d \le m$

$$T^{-1}\sum_{t=1}^{T} \varepsilon'_{t-a}\hat{L}'_{t,a,r}\hat{\Omega}^{-1}\hat{K}_{t,b}\varepsilon_{t-b}\varepsilon'_{t-c}\hat{K}'_{t,c}\hat{\Omega}^{-1}\hat{L}_{t,d,s}\varepsilon_{t-d}$$

$$\to \mathbb{E}\varepsilon'_{t-a}L'_{\infty,a,r}\Omega_0^{-1}K_{\infty,b}\varepsilon_{t-b}\varepsilon'_{t-c}K'_{\infty,c}\Omega_0^{-1}L_{\infty,d,s}\varepsilon_{t-d},$$

where $K_{\infty,\alpha}$ and $L_{\infty,j,r}$ denote the limits of $\hat{K}_{t,j}$ and $\hat{L}_{t,j,r}$, respectively. Consequently for arbitrary $\epsilon > 0$ for large enough T

$$\begin{aligned}\mathbb{P}\{\|\hat{\mathcal{P}}_T - \mathcal{P}_T\| > \eta\} &= \mathbb{P}\{\|\hat{\mathcal{P}}_T - \mathcal{P}_T\| > \eta \wedge \hat{\tau}_{CCA,\alpha} \in \Psi_c\}\\ &\quad + \mathbb{P}\{\|\hat{\mathcal{P}}_T - \mathcal{P}_T\| > \eta \wedge \hat{\tau}_{CCA,\alpha} \notin \Psi_c\}\\ &\le \mathbb{P}\{\|\hat{\mathcal{P}}_T - \mathcal{P}_T\| > \eta \wedge \hat{\tau}_{CCA,\alpha} \in \Psi_c\} + \mathbb{P}\{\hat{\tau}_{CCA,\alpha} \notin \Psi_c\} < \epsilon.\end{aligned}$$

This holds because (using $T_{m,r,s} = \hat{\mathcal{P}}_T - R_{m,r,s}$)

$$\begin{aligned}\mathbb{P}\{\|\hat{\mathcal{P}}_T - \mathcal{P}_T\| > \eta \wedge \hat{\tau}_{CCA,\alpha} \in \Psi_c\} &= \mathbb{P}\{\|\hat{\mathcal{P}}_T - \mathcal{P}_T\| > \eta | \hat{\tau}_{CCA,\alpha} \in \Psi_c\}\mathbb{P}\{\hat{\tau}_{CCA,\alpha} \in \Psi_c\}\\ &\le \mathbb{P}\{\|\hat{\mathcal{P}}_T - \mathcal{P}_T\| > \eta | \hat{\tau}_{CCA,\alpha} \in \Psi_c\}\\ &= \mathbb{P}\{\|T_{m,r,s} - \mathcal{P}_T + R_{m,r,s}\| > \eta | \hat{\tau}_{CCA,\alpha} \in \Psi_c\}\\ &\le \mathbb{P}\{\|T_{m,r,s} - \mathcal{P}_T\| > \eta/2 | \hat{\tau}_{CCA,\alpha} \in \Psi_c\}\\ &\quad + \mathbb{P}\{\|R_{m,r,s}\| > \eta/2 | \hat{\tau}_{CCA,\alpha} \in \Psi_c\}.\end{aligned}$$

Both terms can be made arbitrarily small by choosing m and T large enough. This concludes the proof of the theorem. ■

A.5. Proof of Theorem 5. Following the main idea of the proof of Theorem 3 in Bauer (2001) let $\hat{X}_{f,p} := \hat{W}_f^+ \hat{\beta}(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1})^{1/2}$ and define $\dot{X}_{f,p} := \hat{W}_f^+ \mathcal{O}_f\langle x_t, Y_{t,p}^-\rangle_{p+1}^{T-f+1}$ $(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1})^{-1/2}$. Note that the rank of $\dot{X}_{f,p}$ a.s. for large enough T equals n_0 and hence asymptotically does not depend on the sample size, whereas the dimensions of $\dot{X}_{f,p}$ do. Now using the characterization of singular values (cf., e.g., Golub and VanLoan, 1989, Cor. 2.3-3), we obtain

$$\hat{\sigma}_{n_0+1} = \min_{\mathrm{rank}(X)=n_0} \|\hat{X}_{f,p} - X\| \le \|\hat{X}_{f,p} - \dot{X}_{f,p}\| = O(Q_T\sqrt{fp})$$

because

$$\hat{X}_{f,p} - \dot{X}_{f,p} = \hat{W}_f^+(\hat{\beta} - \mathcal{O}_f\langle x_t, Y_{t,p}^-\rangle_{p+1}^{T-f+1}(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1})^{-1})(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^{T-f+1})^{1/2},$$

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press

where

$$\|\hat{\beta} - \mathcal{O}_f \langle x_t, Y_{t,p}^- \rangle_{p+1}^{T-f+1} (\langle Y_{t,p}^-, Y_{t,p}^- \rangle_{p+1}^{T-f+1})^{-1} \|_{Fr} = O(Q_T \sqrt{fp})$$

(see the proof of Theorem 1) and $\|\hat{W}_f^+\| = O(1), \|(\langle Y_{t,p}^-, Y_{t,p}^- \rangle_{p+1}^{T-f+1})^{1/2}\| = O(1)$, the former following from Assumption 2 and the latter by Lemma 3. The bound then follows by bounding the two norm via the Frobenius norm. Therefore $\hat{\sigma}_{n_0+1}^2 T/(fp \log\log T) = O(1)$. Then the rest of the proof of Theorem 3 of Bauer (2001) (in particular see p. 1566) holds unchanged. Therefore we refer to the original paper for details. ■

https://doi.org/10.1017/S0266466608080419 Published online by Cambridge University Press