Econometric Theory, **21**, 2005, 181–211. Printed in the United States of America.
DOI: 10.1017/S0266466605050127

ESTIMATING LINEAR DYNAMICAL SYSTEMS USING SUBSPACE METHODS

DIETMAR BAUER
TU Wien

This paper provides a survey on a class of so-called subspace methods whose main proponent is CCA proposed by Larimore (1983, *Proceedings of the 1983 American Control Conference 2*). Because they are based on regressions these methods for the estimation of ARMAX systems are attractive as a result of their conceptual simplicity and their numerical advantages as compared to traditional estimators based on criterion optimization. Under the assumption of correct specification the methods provide consistent and asymptotically normal estimates for stationary ARMAX processes where the innovations may be conditionally heteroskedastic and the exogenous inputs are strictly stationary of sufficiently short memory. For stationary autoregressive moving average (ARMA) processes with independent and identically distributed (i.i.d.) Gaussian innovations the estimates are even asymptotically efficient. For $I(1)$ ARMA processes the estimates of both the long-run and the short-run dynamics are consistent without using the knowledge that the data are integrated in the algorithm. Additionally the algorithms provide easily accessible information on the appropriateness of the chosen model complexity. The algorithms include a number of design parameters that have to be set by the user. These include the order of the estimated system. This paper collects up-to-date knowledge on the effects of these design parameters, leading to a number of suggested automated choices to obtain a fully automated estimation procedure.

1. INTRODUCTION

Autoregressive moving average (ARMA) and autoregressive moving average with exogenous inputs (ARMAX) systems are traditionally estimated using optimization of some criterion function such as the Gaussian likelihood or related criteria over a set of parametric models. Except in special cases such as ARX

The author is indebted to Manfred Deistler and Wolfgang Scherrer for guidance and support during the past decade. As visible from the references many results are due to their technical skills and insights. I am also grateful for the very useful comments provided by Martin Wagner, two anonymous referees, and Peter C.B. Phillips, which have led to a substantial improvement of a first version of this paper. This paper originated while the author was holding a postdoc position financed by the Max Kade Foundation at the Cowles Foundation for Research in Economics, Yale University, New Haven, whose hospitality is gratefully acknowledged. Address correspondence to Dietmar Bauer, Department of Mathematical Methods in Economics, TU Wien, E105, Argentinierstr. 8, A-1040 Wien, Austria; e-mail: Dietmar.Bauer@tuwien.ac.at.

© 2005 Cambridge University Press 0266-4666/05 $12.00

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

models the criterion function is nonlinear in the parameters and does not allow for closed form solutions. Therefore the optimization typically is performed using gradient methods starting from an initial guess. In particular for high-dimensional parameter spaces arising especially for multivariate data sets the optimization is numerically demanding, and local optima are a concern. In these cases the choice of the initial guess is critical. Using suboptimal starting values might lead to a substantial decrease in estimation accuracy. Additionally, for large data sets the gradient search becomes numerically exhaustive and sometimes prohibitive.

Subspace methods on the other hand are based on regressions and hence lead to easily implementable estimators carrying a relatively low computational cost. Moreover, in particular the subspace method called CCA (canonical correlation analysis) has been shown to possess remarkable accuracy properties, being asymptotically efficient in some cases. CCA is known to provide consistent estimators under a variety of conditions, including stationary ARMAX processes. Also for integrated ARMA processes consistency is achieved without using the knowledge of the cointegrating rank. Additionally, order estimation can be performed within the framework of subspace methods at extremely low computational cost. This paper presents the main ideas of subspace methods and reviews the literature on asymptotic results relating to consistency and the asymptotic distribution of the estimated parameters.

Subspace methods can be seen as evolutions of methods based on stochastic realization theory such as the algorithms proposed in Ho and Kalman (1966), Zeiger and McEwen (1974), and Aoki and Havenner (1991). The most popular subspace methods are CCA (canonical correlation analysis, introduced as CVA [canonical variate analysis] proposed in Larimore (1983), N4SID (numerical methods for subspace state space identification, Van Overschee and DeMoor, 1994), and MOESP (multivariate output error state space, Verhaegen, 1994). The main idea of subspace methods is to exploit the structure of the state space representation of ARMAX processes (see Section 2). Given the exogenous input the state summarizes in a certain sense all information of the past that is useful for mean square prediction. The various algorithms use this structure differently. As will be argued in Section 3 the CCA approach seems to be most suited for typical econometric analyses. Hence this survey focuses on this method.

The basic underlying idea of CCA is similar to the so-called Hannan–Rissanen method (sometimes also-called Durbin's method; Durbin 1960), which is discussed in great detail in Chapter 6 of Hannan and Deistler (1988). This method uses preliminary ARX modeling to obtain estimates of the innovations. Using these estimates in place of the lagged true innovations in the ARMAX system equations the ARMAX(p,q,r) system matrices can be estimated by regression for any integers (p,q,r) specifying the maximum lag lengths. Because the state space system equations are linear in the system matrices when the state is known (see Section 2 for details) CCA estimates the state in a first step. Then in a

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

second step the system matrices are estimated using regression in the system equations substituting the estimated state for the true state.

To use CCA a number of design parameters have to be set, resulting in a whole class of algorithms that have been called the *state approach* to subspace methods. One particular algorithm in this class has been shown to provide system estimators for ARMA processes that have the same asymptotic distribution as the preliminary estimators obtained using the Hannan–Rissanen procedure (see Deistler, Peternell, and Scherrer, 1995, Proposition 4.1). Note that these first step estimators provided by the Hannan–Rissanen procedure are not asymptotically efficient (cf. Hannan and Deistler, 1988, Remark, p. 264), which could be problematic, especially in high-dimensional data sets and in situations where a large number of local extrema are present, and it is hence important to choose the initial guess close to the global optimum. Therefore by optimizing the choices in the state approach the resulting algorithm is not worse than the Hannan–Rissanen procedure. A large portion of the literature on subspace methods in fact focuses on the effects of the various design parameters on the asymptotic distribution of the corresponding parameter estimators. Supplying rules for setting these design parameters then leads to fully automated procedures in the sense that the user only has to supply the data and the algorithm returns an estimated model. It should always be borne in mind that these rules are typically motivated by considerations about asymptotic properties in hypothetical situations that might not be reflected in finite samples in real-world data sets. Thus automated procedures are seen as a starting point for analysis rather than a substitute for research skills.

There are (at least) two shortcomings of this survey. First the paper is focused on a particular method (namely, the state approach) and does not take into account alternative algorithms such as the previously mentioned N4SID and MOESP and different adaptations of CCA such as the algorithms of Peternell, Scherrer, and Deistler (1996). A partial justification for this omission will be provided later. Second this survey neglects a wide range of topics that form a sizable part of the contributions to the engineering literature on subspace methods but do not seem to be relevant for economic data sets. For a related survey including some of these issues see Bauer (2004a).

The paper is organized as follows. In the next section the state space framework is discussed. Section 3 provides a detailed description of the state approach to subspace methods, and Section 4 presents some asymptotic theory for the stationary case under the assumption of correct specification, leading to a number of suggestions for the choice of the design parameters for given order of the system. Section 5 then discusses methods for specification of the order. Section 6 deals with integrated processes. A small simulation study compares subspace methods to pseudo-maximum-likelihood estimation in Section 7. Section 8 concludes the paper. Two Appendixes provide technical details on the relation between state space and ARMAX systems (Appendix A) and the parameterization of state space models based on overlapping echelon forms (Appendix B).

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

2. MODEL CLASS

In this paper we consider ARMAX processes that are defined as follows: Given a sequence of observed exogenous inputs $(z_t)_{t\in\mathbb{N}}, z_t \in \mathbb{R}^m$ and a martingale difference sequence $(\varepsilon_t)_{t\in\mathbb{N}}, \varepsilon_t \in \mathbb{R}^s$ (for detailed assumptions see Section 4) the process $(y_t)_{t\in\mathbb{N}}, y_t \in \mathbb{R}^s$ is an ARMAX process if there exist matrices $A_j \in \mathbb{R}^{s\times s}, j = 1,\dots,p, B_j \in \mathbb{R}^{s\times s}, j = 1,\dots,q$, and $D_j \in \mathbb{R}^{s\times m}, j = 0,\dots,r$ and (possibly random) initial conditions $y_t, t = 1-p,\dots,0, z_t, t = 1-r,\dots,0, \varepsilon_t, t = 1-q,\dots,0$ such that $(y_t)_{t\in\mathbb{N}}$ solves the vector difference equation (VDE)

$$y_t + A_1 y_{t-1} + \cdots + A_p y_{t-p} = D_o z_t + D_1 z_{t-1} + \cdots + D_r z_{t-r} + \varepsilon_t + B_1 \varepsilon_{t-1} + \cdots + B_q \varepsilon_{t-q} \tag{1}$$

for $t \in \mathbb{N}$. The integers p, q, r define the complexity of the model.

Equivalently every ARMAX process $(y_t)_{t\in\mathbb{N}}$ can be written as the solution to the state space equations

$$\begin{aligned} x_{t+1} &= Ax_t + Bz_t + K\varepsilon_t, \\ y_t &= Cx_t + Dz_t + \varepsilon_t \end{aligned} \tag{2}$$

for $t \in \mathbb{N}$. Here $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{s\times n}, D \in \mathbb{R}^{s\times m}, K \in \mathbb{R}^{n\times s}$ are real matrices. The recursions are started using the initial state $x_1 \in \mathbb{R}^n$, which is assumed to be a random variable with finite second moments uncorrelated with $(\varepsilon_t)_{t\in\mathbb{N}}$. The n-dimensional process $(x_t)_{t\in\mathbb{N}}$ is the so-called state process. The state x_t is not observed. For explicit expressions of the matrices (A, B, C, D, K) for given ARMAX system see Appendix A, where some relations between ARMAX and state space representations also are provided. A more detailed exposition can be found in Chapters 1 and 2 of Hannan and Deistler (1988). In the following discussion we will deal exclusively with the state space framework. To make the paper accessible for readers unfamiliar with these concepts, the main ideas needed in this paper are discussed in more detail in Appendix A, where the corresponding notions for ARMAX systems also are given for illustration purposes.

The solution to the state space equations is easily seen to be (cf. Hannan and Deistler, 1988, p. 14)

$$y_t = Dz_t + \sum_{j=1}^{t-1} CA^{j-1}Bz_{t-j} + \varepsilon_t + \sum_{j=1}^{t-1} CA^{j-1}K\varepsilon_{t-j} + CA^{t-1}x_1 = \tilde{y}_t + CA^{t-1}x_1, \tag{3}$$

where the last equation defines $\tilde{y}_t$. Therefore $\tilde{y}_t$ denotes the solution to the initial conditions $x_1 = 0$. Denoting $K_0 := I_s, L_0 := D, K_j := CA^{j-1}K, L_j := CA^{j-1}B, j > 0$ the transfer functions $k(z) := I_s + zC(I_n - zA)^{-1}K = \sum_{j=0}^{\infty} K_j z^j, l(z) := D + zC(I_n - zA)^{-1}B = \sum_{j=0}^{\infty} L_j z^j$ are both rational functions in $z \in \mathbb{C}$. These equations define the mapping $\pi : (A, B, C, D, K) \mapsto$

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

$(k(z), l(z)) = (I_s + zC(I_n - zA)^{-1}K, D + zC(I_n - zA)^{-1}B)$. The transfer functions converge at least for $z \in \mathbb{C}$ in a neighborhood of the origin. The state space system (A, B, C, D, K) is called *state space representation* of the pair $(k(z), l(z))$ of transfer functions. From (3) it follows that y_t can be additively decomposed into $(\tilde{y}_t)_{t\in\mathbb{N}}$, the solution corresponding to zero initial state, and the effects of a nonzero initial state, i.e., $CA^{t-1}x_1$. From the representation $\tilde{y}_t = \sum_{j=0}^{t-1} K_j \varepsilon_{t-j} + L_j z_{t-j}$ it is clear that $\tilde{y}_t$ depends only on the pair $(k(z), l(z))$ and not on the particular state space representation (A, B, C, D, K) thereof. Thus if the effects of the initial conditions are modeled separately it may be assumed without loss of generality that the system (A, B, C, D, K) is minimal (as defined in Hannan and Deistler, 1988, p. 47; see also Appendix A). For a given pair of transfer functions $(k(z), l(z))$ a minimal state space representation uses the smallest possible state dimension to represent $(k(z), l(z))$. The dimension of the state in a minimal state space representation is called the *order* of the system. The order measures the complexity of the model.

With the exception of Section 6 we will also impose the stability assumption $\lambda_{max}(A) < 1$ (where $\lambda_{max}(A)$ denotes the maximum modulus of the eigenvalues of A), implying that $k(z)$ and $l(z)$ converge absolutely for $|z| \leq 1$, $z \in \mathbb{C}$. This implies that for stationary input $(z_t)_{t\in\mathbb{N}}$ the initial state x_1 can be chosen such that $(y_t)_{t\in\mathbb{N}}$ is stationary. Furthermore the so-called strict minimum-phase assumption $\lambda_{max}(A - KC) < 1$ is imposed throughout, implying that $k(z)$ is stably invertible on $|z| \leq 1$. Let M_n denote the set of pairs of rational transfer functions $(k(z), l(z))$ having a state space representation of order n where $(k(0), l(0)) = (I_s, 0)$, which are analytic on the closed unit disc and where $k(z)^{-1}$ is analytic on the closed unit disc. The set of pairs M_n is equipped with the so-called pointwise topology (cf. Hannan and Deistler, 1988, p. 54).

Minimal state space systems are not identified. Changing coordinates in the state space (i.e., using $\tilde{x}_t = \mathcal{T}x_t$ for nonsingular $\mathcal{T} \in \mathbb{R}^{n\times n}$ as the state) leads to a different state space representation. It is possible to restrict sufficiently many entries of the system matrices (A, B, C, D, K) to zero or one such that the remaining unrestricted entries are identifiable parameters for pairs of transfer functions $(k(z), l(z))$ in a small enough neighborhood of any given pair $(k_0(z), l_0(z))$. In this paper we will use the parameterization based on the overlapping echelon forms described in Section 2.6 of Hannan and Deistler (1988) using such local identification conditions. In the following discussion we will only establish the relevant notation. For details on the construction of these forms see Appendix B. It should also be noted that the parameterization is only introduced as a means to conveniently compare the estimation accuracy of the estimators. Most other results and in particular the description of the algorithms will not use a particular coordinate system for M_n.

The parameterization consists of a finite collection of mappings φ_α indexed by a multi-index $\alpha = (n_1, n_2, \ldots, n_s), n_i \in \mathbb{Z}, 0 \leq n_i \leq n, i = 1, \ldots, s, \sum_{i=1}^{s} n_i = n$. For each given α the mapping φ_α is defined on a subset $M_{n,\alpha} \subset M_n$ such that

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

$\varphi_\alpha : (k(z), l(z)) \in M_{n,\alpha} \mapsto \theta \in \Theta_\alpha \subset \mathbb{R}^{2sn+sm+mn}$ attaches the parameter vector θ collecting all free entries to each pair of transfer functions in $M_{n,\alpha}$. Theorems 2.6.4 and 2.6.5 of Hannan and Deistler (1988) imply the following properties: The parameter space Θ_α is an open subset of $\mathbb{R}^{2sn+sm+nm}$ for each α. For each multi-index α it holds that $M_{n,\alpha}$ is open and dense in M_n and φ_α is continuous on $M_{n,\alpha}$. It follows that the specification of α once the order n has been specified is less critical. Finally denoting with $|\alpha| := \sum_{i=1}^{s} n_i$ the order of the parameterized pairs of transfer functions, we have $\bigcup_{\alpha,|\alpha|=n} M_{n,\alpha} = M_n$.

From this discussion and Appendix B it is seen that the question of finding a parameterization of ARMAX systems (or for that matter state space systems) for multivariate systems is not trivial. The underlying technical reason is that there does not exist a continuous parameterization for M_n (cf. the discussion in Hannan and Deistler, 1988, Sect. 2.6.). It might seem to be the case that using the maximal lag lengths (p, q, r) as the measure of complexity rather than the system order n simplifies the situation. However, as discussed on p. 77 of Hannan and Deistler (1988), such a parameterization has theoretical disadvantages. Therefore we will use the relatively complicated parameterization given previously. This carries the advantage that for echelon forms the relations between ARMAX and state space parameters are very simple (see Hannan and Deistler, 1988, Theorem 2.6.5, p. 72).

3. DESCRIPTION OF THE ALGORITHM

The basic idea underlying the state approach to subspace algorithms is the fact that if the state were known the state space system equations (2) could be used to obtain estimates of the system matrices using simple regressions. The main problem then is to obtain a good estimate of the state.

In this respect the properties of the state for prediction purposes are central. Consider the problem of predicting $y_{t+j}, j \geq 0$ based on the knowledge of $z_l, l \in \mathbb{N}, y_l, l < t$ and x_t. Let $y_{t+j|t}$ denote the best mean square prediction. Analogous to (3) for $j \geq 0$ we have the equation

$$y_{t+j} = CA^j x_t + \sum_{i=0}^{j} L_i z_{t+j-i} + \sum_{i=0}^{j} K_i \varepsilon_{t+j-i} = y_{t+j|t} + \sum_{i=0}^{j} K_i \varepsilon_{t+j-i} \tag{4}$$

because $(\varepsilon_t)_{t\in\mathbb{N}}$ is assumed to be a martingale difference sequence. Note that $y_{t+j|t}$ depends on $z_l, y_l, l < t$ only via x_t and that furthermore $y_{t+j|t}$ is a linear function of the state x_t and the future inputs $z_{t+i}, i \geq 0$ for all $j \geq 0$. On the other hand it follows also from the system equations that[1]

$$\begin{aligned} x_t &= Ax_{t-1} + Bz_{t-1} + K\varepsilon_{t-1} = Ax_{t-1} + Bz_{t-1} + K(y_{t-1} - Cx_{t-1} - Dz_{t-1}) \\ &= (A - KC)x_{t-1} + (B - KD)z_{t-1} + Ky_{t-1} = \cdots \\ &= (A - KC)^p x_{t-p} + \sum_{j=0}^{p-1} (A - KC)^j (Ky_{t-1-j} + (B - KD)z_{t-1-j}) \\ &= (A - KC)^p x_{t-p} + \mathcal{K}_p W_{t,p}^- = (A - KC)^{t-1} x_1 + \mathcal{K}_{t-1} W_{t,t-1}^-, \end{aligned}$$

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

where for $\bar{A} := A - KC, \bar{B} := B - KD$ we have

$$\mathcal{K}_p := [K, \bar{B}, \bar{A}[K, \bar{B}], \bar{A}^2[K, \bar{B}], \ldots, \bar{A}^{p-1}[K, \bar{B}]]$$

and $W_{t,p}^- := [w_{t-1}', \ldots, w_{t-p}']'$ denoting with w_t the joint process, i.e., $w_t := [y_t', z_t']'$. Therefore it follows that x_t is a linear function of the data obtained up to time t and the initial state x_1. Under the strict minimum-phase assumption, $(A - KC)^p$ will be small for large p, and hence a reasonable approximation of x_t is given by $\mathcal{K}_p W_{t,p}^-$. Therefore an estimate of the state can be obtained by replacing $\mathcal{K}_p$ with an estimate $\hat{\mathcal{K}}_p$ in this approximation.

These two facts build the basis for the estimation of the state. Consider the prediction equation (4) written jointly for $j = 0, \ldots, f - 1$ for some integer f (where $t > p$ is assumed):

$$Y_{t,f}^+ = \mathcal{O}_f x_t + \mathcal{U}_f Z_{t,f}^+ + N_{t,f}^+ = \mathcal{O}_f \mathcal{K}_p W_{t,p}^- + \mathcal{O}_f \bar{A}^p x_{t-p} + \mathcal{U}_f Z_{t,f}^+ + N_{t,f}^+, \tag{5}$$

where $Y_{t,f}^+ := [y_t', \ldots, y_{t+f-1}']'$ collects the predicted variables, $Z_{t,f}^+ := [z_t', \ldots, z_{t+f-1}']'$ is the future of the exogenous inputs, and $N_{t,f}^+$ denotes the prediction errors. Finally $\mathcal{O}_f := [C', A'C', \ldots, (A^{f-1})'C']' \in \mathbb{R}^{fs \times n}$ is sometimes called the truncated observability matrix. Expressions for $\mathcal{U}_f \in \mathbb{R}^{fs \times fm}$ and $N_{t,f}^+$ are easy to derive but not essential for our purposes and hence are omitted.

Equation (5) lies at the core of the subspace idea. Neglecting $\mathcal{O}_f \bar{A}^p x_{t-p}$ under the assumption $\bar{A}^p \approx 0$ the matrices $\mathcal{O}_f \mathcal{K}_p$ and $\mathcal{U}_f$ can be estimated using regression because $W_{t,p}^-$ and $Z_{t,f}^+$ are observed and the residuals $N_{t,f}^+$ are orthogonal to the regressors. Moreover for f sufficiently large the fact that $\operatorname{rank}(\mathcal{O}_f \mathcal{K}_p) = n$, where n denotes the order of the system, can be exploited. This motivates the following procedure.

1. Choose the integers f and p.
2. Estimate the equation $Y_{t,f}^+ = \beta_1 W_{t,p}^- + \beta_2 Z_{t,f}^+ + N_{t,\perp}, t = p + 1, \ldots, T - f + 1$ under the restriction $\operatorname{rank}(\beta_1) = n$. This leads to the estimates $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$ of β_1 (rank n). In this step also the order n of the estimated system has to be specified.
3. Estimate the state $\hat{x}_t := \hat{\mathcal{K}}_p W_{t,p}^-, t = p + 1, \ldots, T + 1$.
4. Regress y_t onto $\hat{x}_t$ and $z_t, t = p + 1, \ldots, T$ to obtain estimates $\hat{C}, \hat{D}$ and residuals $\hat{\varepsilon}_t := y_t - \hat{C}\hat{x}_t - \hat{D}z_t, t = p + 1, \ldots, T$.
5. Regress $\hat{x}_{t+1}$ onto $\hat{x}_t, z_t$ and $\hat{\varepsilon}_t$ for $t = p + 1, \ldots, T$ to obtain estimates $\hat{A}, \hat{B}$, and $\hat{K}$ to compute the estimates $\hat{k}(z) := I_s + z\hat{C}(I_n - z\hat{A})^{-1}\hat{K}, \hat{l}(z) := \hat{D} + z\hat{C}(I_n - z\hat{A})^{-1}\hat{B}$.
6. Check the minimum-phase assumption. If the estimated system is not minimum-phase the estimated spectrum $(\hat{k}(z)\hat{\Omega}\hat{k}(z)^*)/(2\pi)$ where $\hat{\Omega} := (T - p)^{-1} \sum_{t=p+1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t'$ (and $.^*$ denotes conjugate transposition) has to be refactored to ensure that the minimum-phase assumption holds (cf. Hannan and Deistler, 1988, Theorem 1.3.3, p. 26).

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

The central step in the algorithm is the reduced rank regression. Let $\hat{\beta}_1$ denote the least squares estimator ignoring the rank restriction on β_1. Then the rank n approximation $\hat{\mathcal{O}}_f\hat{\mathcal{K}}_p$ of $\hat{\beta}_1$ is obtained using the singular value decomposition (SVD)

$$\hat{\Xi}_f^+\,\hat{\beta}_1\,\hat{\Xi}_p^- = \hat{U}\hat{\Sigma}\hat{V}' = \hat{U}_n\hat{\Sigma}_n\hat{V}_n' + \hat{R}_n,$$

where $\hat{U}_n \in \mathbb{R}^{fs\times n}$ denotes the matrix whose columns are the left singular vectors of $\hat{\Xi}_f^+\,\hat{\beta}_1\,\hat{\Xi}_p^-$ corresponding to the n dominating singular values $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_n > 0$, $\hat{\Sigma}_n = \operatorname{diag}(\hat{\sigma}_1,\ldots,\hat{\sigma}_n)$ and $\hat{V}_n \in \mathbb{R}^{p(s+m)\times n}$ denotes the matrix of the corresponding right singular vectors. The approximation error arising from using a system order n is denoted as $\hat{R}_n$. Finally $\hat{\Xi}_f^+ \in \mathbb{R}^{fs\times fs}$ and $\hat{\Xi}_p^- \in \mathbb{R}^{p(s+m)\times p(s+m)}$ denote two weighting matrices that for the moment only are assumed to be nonsingular a.s. More specific conditions will be provided subsequently. This results in estimates $\hat{\mathcal{O}}_f := (\hat{\Xi}_f^+)^{-1}\hat{U}_n\hat{\Sigma}_n, \hat{\mathcal{K}}_p := \hat{V}_n'(\hat{\Xi}_p^-)^{-1}$. If a different factorization into $\hat{\mathcal{O}}_f = (\hat{\Xi}_f^+)^{-1}\hat{U}_n\hat{\Sigma}_n\mathcal{T}^{-1}, \hat{\mathcal{K}}_p = \mathcal{T}\hat{V}_n'(\hat{\Xi}_p^-)^{-1}$ for some nonsingular matrix $\mathcal{T} \in \mathbb{R}^{n\times n}$ is used the resulting estimates will be observationally equivalent; i.e., the estimated pair of transfer functions will be unchanged. The asymptotic properties of the state space representation $(\hat{A},\hat{B},\hat{C},\hat{D},\hat{K})$ on the contrary depend on the choice of the factorization.

The term canonical correlation analysis (CCA) has its origin in the approximation step. Let $\langle a_t, b_t\rangle := T^{-1}\sum_{t=p+1}^{T-f+1} a_t b_t'$ for two vector sequences $(a_t)_{t\in\mathbb{N}}, (b_t)_{t\in\mathbb{N}}$ somewhat sloppily using the same symbol for the processes and their realizations at time t. Introduce the notation

$$Y_{t,f}^{+,\Pi} := Y_{t,f}^+ - \langle Y_{t,f}^+, Z_{t,f}^+\rangle\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1} Z_{t,f}^+,$$

$$W_{t,p}^{-,\Pi} := W_{t,p}^- - \langle W_{t,p}^-, Z_{t,f}^+\rangle\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1} Z_{t,f}^+$$

such that $Y_{t,f}^{+,\Pi}$ denotes the residuals of a regression of $Y_{t,f}^+$ onto $Z_{t,f}^+$ and correspondingly $W_{t,p}^{-,\Pi}$ are the residuals of the regression of $W_{t,p}^-$ onto $Z_{t,f}^+$. Then the original version of CCA uses

$$\hat{\Xi}_f^+ := \langle Y_{t,f}^{+,\Pi}, Y_{t,f}^{+,\Pi}\rangle^{-1/2}, \qquad \hat{\Xi}_p^- := \langle W_{t,p}^{-,\Pi}, W_{t,p}^{-,\Pi}\rangle^{1/2}, \tag{6}$$

where $X^{1/2} = (X^{1/2})', (X^{1/2})^2 = X$ denotes the symmetric square root of the positive definite matrix X. Further the unrestricted estimate $\hat{\beta}_1 = \langle Y_{t,f}^{+,\Pi}, W_{t,p}^{-,\Pi}\rangle\langle W_{t,p}^{-,\Pi}, W_{t,p}^{-,\Pi}\rangle^{-1}$. Therefore in the case of no inputs the weightings used in CCA imply that the SVD amounts to the estimation of the canonical correlations and vectors of the two subspaces spanned by the components of $Y_{t,f}^+$ and $W_{t,p}^-$, respectively (because there is no input here $Y_{t,f}^{+,\Pi} = Y_{t,f}^+, W_{t,p}^{-,\Pi} = W_{t,p}^-$).

The choice of weightings according to (6) corresponds to the maximum-likelihood solution to the reduced rank regression $Y_{t,f}^+ = \beta_1 W_{t,p}^- + \beta_2 Z_{t,f}^+ + N_{t,\perp}$ such that $\operatorname{rank}(\beta_1) = n$ based on Gaussanity (and uncorrelatedness over time)

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

assumptions on $N_{t,\perp}$. The least squares solution is obtained by using $\hat{\Xi}_f^+ = I_{fs}$. With respect to $\hat{\Xi}_p^-$ only the choice given previously will be considered throughout. In the following discussion the algorithm using the weightings (6) will be referred to as CCA. Algorithms using a different weighting scheme but the same outline in the literature have been called *Larimore approach* or *state approach*. We will use the second term in the following discussion.

In the reduced rank regression both the outputs and the regressors are lags of y_t, z_t and w_t, respectively, and hence the problem has a similar structure as autoregressions. As in autoregressions there are two possible implementations that in the stationary case possess identical asymptotics. Either only time instants are used in the regression where all observations are available, i.e., the range of t is limited to $t = p + 1, \ldots, T - f + 1$, or alternatively, the regression can be based on the estimated second moments as in the Yule–Walker equations, essentially imposing zero initial and end conditions (cf. Hannan and Deistler, 1988, p. 212). The first approach might seem to be favorable for small sample sizes and relatively large values of f and p as the imposition of zero initial conditions introduces avoidable bias. On the other hand the usage of covariance estimates makes the procedure easily adaptable to data sets including missing values. Using robust estimators of the covariances leads to (statistically) robust versions of the state approach. A recursive (in sample size) formulation of the regression problem leads to recursive estimators.

Finally the term *subspace algorithm* needs explanation. It originates from the reduced rank regression in the second step as the most popular methods in the class of subspace algorithms, that is, MOESP (Verhaegen, 1994) and N4SID (Van Overschee and DeMoor, 1994), use the matrix $\hat{\mathcal{O}}_f$ in subsequent steps rather than $\hat{\mathcal{K}}_p$. Because the factorization into $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$ is of no importance for the estimated pair of transfer functions, the reduced rank regression problem essentially estimates the subspace spanned by the columns of the matrix $\hat{\mathcal{O}}_f$, hence the name. MOESP continues to obtain estimates for (A, B, C, D) but does not estimate K, which implies that only $l(z)$ is estimated. Frequently in econometric analysis one is primarily interested in the noise dynamics, and therefore MOESP does not seem to be a suitable algorithm for these purposes. Therefore it is not considered further in this survey. The same remark is true for one variant of N4SID. An alternative version of N4SID and a recently proposed algorithm by Chiuso and Picci (2003) use a more complicated method to estimate the shifted state $\hat{x}_{t+1}$. Because the analysis of these methods is still in its infancy we will not consider these algorithms further. We refer to Chiuso and Picci (2003) for details.

The description of the state approach makes it obvious that there are a couple of design parameters involved. First f and p have to be chosen; second the weighting matrices $\hat{\Xi}_f^+$ and $\hat{\Xi}_p^-$ and the order n have to be specified. A large part of the literature on subspace methods deals with the effects of these choices on the asymptotic properties of the corresponding estimators. Section 4 will survey the main results in this respect for the stationary case, and Section 6

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

will deal with the case of integrated processes, whereas Section 5 discusses procedures to estimate the order n.

4. ASYMPTOTIC RESULTS IN THE STATIONARY CASE

Most of the literature on the asymptotic properties of subspace methods deals with the stationary case. Throughout this section we will make the following assumptions on the data generating mechanism.

Assumption 1. The process $(y_t)_{t\in\mathbb{N}}$ is generated by a state space system (2) where the true pair of transfer functions $(k_0(z), l_0(z)) \in M_{n_0,\alpha}$, i.e., the true order of the system is denoted as n_0. Accordingly $\theta_{0,\alpha} = \varphi_\alpha(k_0(z), l_0(z)) \in \Theta_\alpha$ denotes the true parameter vector which is an interior point of Θ_α. Thus in particular it is assumed that the system is stable and strictly minimum-phase. The strictly stationary ergodic process $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_t = \sigma\{\varepsilon_t, \varepsilon_{t-1}, \ldots\}$ having the following properties:

$$\mathbb{E}\{\varepsilon_t|\mathcal{F}_{t-1}\} = 0, \qquad \lim_{k\to\infty} \mathbb{E}\{\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-k}\} = \mathbb{E}\{\varepsilon_t\varepsilon_t'\} > 0, \quad \text{a.s.}, \qquad \mathbb{E}\varepsilon_{t,a}^4 < \infty,$$

$$\lim_{k\to\infty} \mathbb{E}\{\varepsilon_{t,a}\varepsilon_{t,b}\varepsilon_{t-r,c}|\mathcal{F}_{t-k}\} = \mathbb{E}\{\varepsilon_{t,a}\varepsilon_{t,b}\varepsilon_{t-r,c}\} \quad \text{a.s.}, \qquad \forall r > 0.$$

Here $\varepsilon_{t,a}$ denotes the ath component of ε_t. If no exogenous inputs $(z_t)_{t\in\mathbb{N}}$ are present, the condition on the third moments is not needed.

The input process $(z_t)_{t\in\mathbb{Z}}$ has a representation as $z_t = \sum_{j=0}^{\infty} K_{z,j}\eta_{t-j}$ where for $k_z(z) := \sum_{j=0}^{\infty} K_{z,j} z^j$ it holds that $\sum_{j=0}^{\infty} \|K_{z,j}\| j^\delta < \infty$ for some $\delta > \frac{1}{2}, \sup_{j\in\mathbb{N}} j^\epsilon \|K_{z,j}\| < \infty$ for some $\epsilon > 1$ and $\det k_z(z) \neq 0, |z| \leq 1$. Here the strictly stationary and ergodic innovations process $(\eta_t)_{t\in\mathbb{Z}}$ is independent of $(\varepsilon_t)_{t\in\mathbb{Z}}$ where for the filtration $\mathcal{G}_t = \sigma\{\eta_t, \eta_{t-1}, \ldots\}$ it holds that

$$\mathbb{E}\{\eta_t|\mathcal{G}_{t-1}\} = 0, \qquad \lim_{k\to\infty} \mathbb{E}\{\eta_t\eta_t'|\mathcal{G}_{t-k}\} = \mathbb{E}\{\eta_t\eta_t'\} > 0, \quad \text{a.s.}, \qquad \mathbb{E}\eta_{t,a}^4 < \infty.$$

The initial state $x_1 = \sum_{j=0}^{\infty} A^j(Bz_{-j} + K\varepsilon_{-j})$.

A couple of remarks on these assumptions seem to be necessary.

- Note that here we use all processes except for $(y_t)_{t\in\mathbb{N}}$ defined on $\mathbb{Z}$ rather than on $\mathbb{N}$ as their time index. This is done for convenience because it allows us to directly use the results for stationary processes in Hannan and Deistler (1988). Because $A^j \to 0$ for $j \to \infty$ at an exponential rate as a result of the stability assumption different choices of the initial state x_1 (subject to mild conditions such as finite second moments) do not change any of the results. Similarly restricting the definition of the input to $t \in \mathbb{N}$ is not critical.

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

- The assumptions on the innovations sequence are identical to Assumptions AX on p. 131 of Hannan and Deistler (1988) used in proving asymptotic normality for pseudo-maximum-likelihood estimation. For some of the following results these seem to be overly restrictive. In particular consistency results might be expected to hold without the assumption of bounded fourth moments. In the literature no results for weaker assumptions have been reported.
- Note that the innovations sequence is not restricted to be conditionally homoskedastic. Certain autoregressive conditionally heteroskedastic (ARCH) and generalized autoregressive conditionally heteroskedastic (GARCH) processes satisfy the assumptions (cf. the discussion in Bauer, 2004c).
- The input sequence is assumed to be linearily regular excluding typical deterministic terms such as the constant or harmonic terms. This is done to avoid multicollinearity problems in the vectors $W_{t,p}^{-}$ and $Z_{t,f}^{+}$. Rather than including such components as inputs it is suggested that this information be extracted prior to using subspace methods. Analogously to the pseudo-maximum-likelihood case the asymptotic properties of the algorithms applied to the demeaned series are identical to the properties if the true mean is subtracted. Some discussion on this issue may be found in Bauer (2004c).
- Long memory processes such as fractionally integrated processes (see Baillie, 1996) are not admitted as inputs. As we assume that the spectrum of the inputs is nonsingular on the unit circle, differenced long memory processes are also disqualified. This does not imply that subspace methods are not applicable in this setting. To the best of my knowledge this area simply has not been investigated.

The state approach essentially consists of a reduced rank regression used to obtain an estimate of the state and regressions in the system equations (2) to estimate the system matrices. The analysis of the second step is standard, and hence most of the literature focuses on the asymptotic properties of the estimation of the state that are determined by the asymptotics for $\hat{\mathcal{K}}_p$. The methods used in the analysis are standard. $\hat{\mathcal{K}}_p$ is obtained as the right factor of the matrix $\hat{\mathcal{O}}_f\hat{\mathcal{K}}_p$ estimated in the reduced rank regression. Because regression estimates in a stationary framework essentially are nonlinear functions of the estimated covariances, the asymptotic distribution of $\hat{\mathcal{K}}_p$ follows using standard linearization techniques of the mapping attaching $\hat{\mathcal{K}}_p$ to the estimated covariance sequence from the well-documented properties of covariance sequence estimates (under the stated assumptions convergence follows from Theorem 7.4.3 on p. 326 of Hannan and Deistler, 1988; asymptotic normality follows using their Lemma 4.3.4, p. 146). The main technical complications in the analysis are due to the fact that $p \to \infty$ at some rate to make the bias introduced by neglecting the term corresponding to $\mathcal{O}_f(A - KC)^p x_{t-p}$ tend to zero.

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

Another complication is the nonuniqueness of the SVD. Typically $\hat{\Xi}_f^+ \hat{\beta}_1 \hat{\Xi}_p^- - U_{n_0}(f,p)\Sigma_{n_0}(f,p)V_{n_0}(f,p)' = o(1)$ is shown as a first step where $U_{n_0}(f,p) \times \Sigma_{n_0}(f,p)V_{n_0}(f,p)' \in \mathbb{R}^{fs \times p(s+m)}$ is a matrix of rank n_0. Here the dependence of the SVD on the integers f and p is made explicit. However, the SVD written as $U_{n_0}(f,p)\Sigma_{n_0}(f,p)V_{n_0}(f,p)'$ is not unique where $\Sigma_{n_0}(f,p) = \operatorname{diag}(\sigma_1(f,p),\dots,\sigma_{n_0}(f,p)) \in \mathbb{R}^{n_0 \times n_0}$ contains the singular values. Even in the case where the singular values $\sigma_1(f,p) > \sigma_2(f,p) > \cdots > \sigma_{n_0}(f,p) > 0$ are distinct the singular vectors (being the columns of $U_{n_0}(f,p)$ and $V_{n_0}(f,p)$) are only unique up to the orientation. If (u_i,v_i) denotes the pair of left and right singular vectors corresponding to $\sigma_i(f,p)$ then $(-u_i,-v_i)$ is another such pair. Therefore it is necessary to impose additional conditions on the SVD to achieve uniqueness and subsequently continuity and differentiability of the factorization of $\hat{\mathcal{O}}_f\hat{\mathcal{K}}_p$ into $\hat{\mathcal{O}}_f$ and $\hat{\mathcal{K}}_p$ (for a detailed discussion see Bauer, Deistler, and Scherrer, 1999, p. 1246, bottom right column). In cases of repeated singular values the situation is more complicated.

As noted previously the choice of the factorization of $\hat{\mathcal{O}}_f\hat{\mathcal{K}}_p$ is not essential for the estimated pair of transfer functions. Any factorization will lead to the same estimate $(\hat{k}(z),\hat{l}(z))$. The main differences lie in the asymptotic properties of the corresponding estimated system $(\hat{A},\hat{B},\hat{C},\hat{D},\hat{K})$. For consistency of $(\hat{k}(z),\hat{l}(z))$ the existence of one factorization such that the corresponding system estimate $(\hat{A},\hat{B},\hat{C},\hat{D},\hat{K})$ converges is sufficient. There are factorizations that simplify the mapping attaching $\hat{\mathcal{K}}_p$ to the estimated covariance sequence significantly (for details in the case of no inputs see Bauer, 2004c, Lemma 3, and the references given therein).

Using these techniques consistency for the transfer function estimators obtained via the state approach under various different assumptions has been established in a number of papers. Deistler et al. (1995, Proposition 3.2) prove consistency for the transfer function estimates obtained by CCA in the case of no inputs under Assumption 1 and the additional assumptions that $\mathbb{E}\{\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}\} = \mathbb{E}\varepsilon_t\varepsilon_t'$, and $f = p = \lfloor d\hat{p}_{BIC} \rfloor$ for some $d > 1$. Here $\lfloor x \rfloor$ denotes the integer part of the real x and $\hat{p}_{BIC}$ denotes the optimal lag length for an autoregressive approximation of y_t selected by minimizing the BIC over $0 \leq p \leq (\log T)^a$ for some constant $0 < a < \infty$. Peternell et al. (1966, Theorem 3.1) extend these results by including ARMA inputs such that the innovations of the joint process w_t form a conditionally homoskedastic martingale difference sequence with bounded unconditional fourth moments and $\hat{\Xi}_f^+ = \langle Y_{t,f}^+ - \hat{\beta}_2 Z_{t,f}^+, Y_{t,f}^+ - \hat{\beta}_2 Z_{t,f}^+ \rangle^{-1/2}$ (their algorithm uCCA) or $\hat{\Xi}_f^+ = I_{fs}$, respectively (their algorithm N4SID).[2] Bauer (2004c, Theorem 1) extends the result in the no inputs case to conditionally heteroskedastic innovations using Assumption 1. Because these results primarily are based upon the uniform convergence of the sample covariance sequence (cf. Deistler et al., 1995, equation (13); and Peternell et al., 1966, equation (26)) it is straightforward to generalize their results to the present setting with the use of Theorem 7.4.3 of Hannan and Deistler (1988). The extension to conditionally heteroskedastic innova-

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

tions $(\varepsilon_t)_{t\in\mathbb{Z}}$ and $(\eta_t)_{t\in\mathbb{Z}}$ follows along the lines of Theorem 1 in Bauer (2004c). A detailed proof of the following statement is contained in a preliminary version of this paper and can be obtained from the author on request.

Sufficient Conditions for Consistency. Let the process $(y_t)_{t\in\mathbb{N}}$ and the input process $(z_t)_{t\in\mathbb{Z}}$ fulfill Assumption 1. Suppose that n_0 is known and used in the state approach. Assume that either $\hat{\Xi}_f^+ = I_{fs}$ or $\hat{\Xi}_f^+ = \langle Y_{t,f}^{+,\Pi}, Y_{t,f}^{+,\Pi}\rangle^{-1/2}$ and $\hat{\Xi}_p^- = \langle W_{t,p}^{-,\Pi}, W_{t,p}^{-,\Pi}\rangle^{1/2}$ is used in the estimation and that either $f \geq n_0$ fixed independent of the sample size or $f = p$ is used where $p \to \infty$ a.s., $p = o((\log T)^a)$ for some $a < \infty$. Then $(\hat{k}(z), \hat{l}(z)) \to (k_0(z), l_0(z))$ a.s.

Some of these conditions clearly are not necessary. If f is kept fixed then the weighting matrix $\hat{\Xi}_f^+$ can be chosen arbitrarily such that $\hat{\Xi}_f^+ \to \Xi_f^+ > 0$ a.s. Also $f = p$ might seem to be too restrictive, and other conditions could be derived. Further the upper bound on the increase of p seems to be too conservative, and it seems likely that less restrictive upper bounds of the form $p = o(T^\delta)$ for some $\delta > 0$ might be sufficient. It should be noted that the assumptions given previously also include the case that f and p are chosen depending on the data such as the choice $\lfloor d\hat{p}_{BIC}\rfloor$. Almost surely convergence to infinity of p subject to the upper bound $(\log T)^a$ is sufficient in this respect. For example any choice of $\lfloor d\hat{p}_{BIC}\rfloor, d > 0$ satisfies these bounds for $0 < \rho_0 = \lambda_{max}(A - KC) < 1$ by Theorem 6.6.3 of Hannan and Deistler (1988, p. 261) stating that $-\hat{p}_{BIC}(2\log\rho_0)/\log T \to 1$ a.s. Alternatively $\hat{p}_{AIC} \geq \hat{p}_{BIC}$ could be used in the definition of f and p.

Note that $(\hat{k}(z), \hat{l}(z)) \to (k_0(z), l_0(z)) \in M_{n_0,\alpha}$ implies that $(\hat{k}(z), \hat{l}(z)) \in M_{n_0,\alpha}$ a.s. for large enough sample size T. Therefore $\hat{\theta}_\alpha = \varphi_\alpha(\hat{k}(z), \hat{l}(z))$ is defined a.s. for large enough T. In order to complete the definition of $\hat{\theta}_\alpha$ we assign an arbitrary parameter vector θ_c, say, to pairs $(\hat{k}(z), \hat{l}(z)) \notin M_{n,\alpha}$. Then $\hat{\theta}_\alpha \to \theta_{0,\alpha}$ a.s. holds as a result of the continuity of φ_α.

The sufficient conditions given here are too weak to obtain the usual square root sample size rate of convergence. Such a rate is required for the derivation of the asymptotic distribution. It has been noted in Theorem 1 of Bauer (2004c) in the case of no inputs that in addition to the sufficient conditions given previously $p \geq -d\log T/(2\log\rho_0), 0 < \rho_0 = \lambda_{max}(A - KC) < 1, d > 1, p = O((\log T)^a)$ is sufficient for $\hat{\theta}_\alpha - \theta_{0,\alpha} = O((\text{loglog } T/T)^{1/2})$ the rate of the law of the iterated logarithm (LIL). The case $\rho_0 = 0$ corresponds to the special case of ARX systems, which will not be dealt with here. In the following it will always be assumed that $\rho_0 > 0$. For exogenous inputs fulfilling Assumption 1 it can be shown that $p \geq -d\log T/(2\log\rho_0), p = O((\log T)^a), d > 1$ is sufficient for $\hat{\theta}_\alpha - \theta_{0,\alpha} = O((\log T/T)^{1/2})$.[3] Furthermore this rate of increase for p is also contained in sufficient conditions for asymptotic normality, as follows:

Sufficient Conditions for Asymptotic Normality. Let the process $(y_t)_{t\in\mathbb{N}}$ and the input process $(z_t)_{t\in\mathbb{Z}}$ fulfill Assumption 1. Suppose that n_0 and α are

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

known and used in the state approach. Assume that either $\hat{\Xi}_f^+ = I_{fs}$ or $\hat{\Xi}_f^+ = \langle Y_{t,f}^{+,\Pi}, Y_{t,f}^{+,\Pi} \rangle^{-1/2}$ and $\hat{\Xi}_p^- = \langle W_{t,p}^{-,\Pi}, W_{t,p}^{-,\Pi} \rangle^{1/2}$ is used in the estimation. Then if $f \geq n_0$ is fixed independent of the sample size and the data and $p \geq -d \log T / (2 \log \rho_0)$, $0 < \rho_0 = \lambda_{max}(A - KC) < 1$, $p = O((\log T)^a)$ for some $a < \infty$, $d > 1$ where p is chosen deterministically we have that $\sqrt{T}(\hat{\theta}_\alpha - \theta_{0,\alpha})$ converges in distribution to a normal random variable with mean zero. If no exogenous inputs are present the same holds true for $f = p$ under the assumptions on p given previously.

Asymptotic normality results under Assumption 1 in the case of no inputs and conditionally homoskedastic innovations have been obtained in Bauer et al. (1999, Theorem 1) and extended to ARMA inputs such that the joint process w_t has conditionally homoskedastic innovations in Bauer (1998). See Theorem 4.3.1 (f fixed) and Theorem 4.3.2 ($f \to \infty$). All three results consider only the case where the true system lies in the subset $M_{n_0}^+ \subset M_{n_0}$ characterized by the fact that $\lim_{T \to \infty} \min_{i=1,\ldots,n_0-1} [\sigma_i(f,p) - \sigma_{i+1}(f,p)] > 0$ (where the assumptions ensure that the limit exists). Hence in $M_{n_0}^+$ asymptotically the singular values are distinct. Bauer (2004c, Theorem 2) extends these results in the case of no inputs to the statement given in the sufficient conditions. The proof of the general result for $f \geq n_0$ as stated earlier is contained in a preliminary version of this paper and can be obtained on request. The proof for $f \to \infty$ in the no inputs case is analogous to the proof in Bauer et al. (1999) using the result of Bauer, Deistler, and Scherrer (2000).

Note that $p \geq -d \log T / (2 \log \rho_0)$ ensures that $(A - KC)^p = o(T^{-1/2})$. This condition is the analogue of conditions typically used in the analysis of autoregressive approximations (cf. Lewis and Reinsel, 1985, Theorem 2(iii)) and ensures that the asymptotic mean of $\sqrt{T}(\hat{\theta}_\alpha - \theta_{0,\alpha})$ is zero. It has been advocated in Deistler et al. (1995, p. 1866) to choose $p = \lfloor d\hat{p}_{BIC} \rfloor$ for some $d > 1$ where again $\hat{p}_{BIC}$ denotes the lag length of an autoregressive approximation of y_t of the ARX type selected by minimizing the BIC over $0 \leq p \leq (\log T)^a$. The choice of the constant d is an open question. It should also be noted that the choice $\lfloor d\hat{p}_{BIC} \rfloor$ is not covered by the sufficient conditions for asymptotic normality given previously because it depends on the data. Asymptotic normality results for stochastic choices of f and p have not been derived in the literature. In the autoregressive setting asymptotic normality of the estimated autoregressive coefficients for data dependent lag length has been obtained recently by Kuersteiner (2005). It seems likely that similar techniques also can be used to prove the asymptotic normality of $\sqrt{T}(\hat{\theta}_\alpha - \theta_{0,\alpha})$ for the suggested stochastic choices of f and p. This is an open point in the literature with serious consequences for the validity of the arguments to follow.

In this respect also note that the relation between CCA and autoregressive approximations is very close. Dahlen and Scherrer (2004) show that in the case of no inputs for $f = p$ the CCA estimates essentially are obtained via model reduction based on a preliminary autoregressive approximation.

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

In the general case no interpretable expression for the asymptotic variance, i.e., the variance of the limiting normal variable, is known. This implies that the effects of the various design parameters on the asymptotic distribution in the general case are unknown. The sufficient conditions only lead to upper and lower bounds on the increase of p as a function of the sample size. Bauer et al. (2000) show that in the situation of Bauer et al. (1999) (i.e., $m = 0$, $\mathbb{E}\{\varepsilon_t \varepsilon_t' | \mathcal{F}_{t-1}\} = \mathbb{E}\varepsilon_t \varepsilon_t', k_0(z) \in M_{n_0}^+$) the choice of the weighting $\hat{\Xi}_p^-$ does not influence the asymptotic distribution if there exists a sequence of deterministic matrices Ξ_p^- such that $\limsup_{p\in\mathbb{N}} \|(\Xi_p^-)^{-1}\|_2 < \infty$ and $\|\hat{\Xi}_p^- - \Xi_p^-\|_2 = o(T^{-\epsilon})$ a.s. for some $\epsilon > 0$. These conditions hold for all choices of $\hat{\Xi}_p^-$ proposed in the literature.

It follows from the proof of Lemma 3 in Bauer (2004c) that the results of Bauer et al. (2000) rely only on the uniform boundedness and convergence (both with respect to the two norm) of the weighting matrices and the convergence of $\hat{\beta}_1$ to $\mathcal{O}_f \mathcal{K}_p$ in the Frobenius norm. Hence the independence of the limiting distribution from the weighting $\hat{\Xi}_p^-$ (subject to the previously mentioned conditions) is easily extended to the general case. Therefore only one choice of $\hat{\Xi}_p^-$ is contained in this paper.

To obtain simple expressions for the asymptotic variance in addition to the sufficient conditions for asymptotic normality given earlier assume that the input is a martingale difference sequence; in particular it is assumed that $K_{z,j} = 0$, $j > 0$. The innovations ε_t are assumed to be conditionally homoskedastic, i.e., $\mathbb{E}\{\varepsilon_t \varepsilon_t' | \mathcal{F}_{t-1}\} = \mathbb{E}\varepsilon_t \varepsilon_t' > 0$. Further assume that $(k_0(z), l_0(z)) \in M_{n_0}$ is contained in a generic subset $M_{n_0}^i$ of M_{n_0} characterized by the fact that the matrix of the first n_0 columns of $\mathcal{K}_p$ is nonsingular (for discussion see Bauer, 2004c, Lemma 3). Using a parameterization of $M_{n_0}^i$ based on the unique state space representation for which the matrix of the first n_0 columns of $\mathcal{K}_p$ equals the identity matrix Bauer and Ljung (2002) provide a rather explicit expression for the variance of the normal limit of the (appropriately scaled) estimation error for the parameter vectors.[4] It is easy to see that the change of coordinates to obtain $\hat{\theta}_\alpha$ is differentiable on $M_{n_0}^i \cap M_{n_0,\alpha}$. Then the Delta method can be used to derive the corresponding expression for the asymptotic variance of $\sqrt{T}(\hat{\theta}_\alpha - \theta_{0,\alpha})$. Thus if $\hat{\theta}_\alpha$ denotes as previously the state approach estimate obtained using $\hat{\Xi}_f^+$ such that $\|\hat{\Xi}_f^+ - \Xi_f^+\|_2 = o(T^{-\epsilon})$ for some deterministic matrix $\Xi_f^+ > 0$ and $\epsilon > 0$ and $f \geq n_0$ fixed and finite we have that $\sqrt{T}(\hat{\theta}_\alpha - \theta_{0,\alpha})$ converges in distribution to a normal random variable with mean zero and variance V_f of the form

$$V_f = M_1 M_1' + M_2[\Gamma_{w,w}^{-1} \otimes \mathcal{O}_f^\dagger (\mathbb{E}N_{t,f}^+ (N_{t,f}^+)')(\mathcal{O}_f^\dagger)']M_2'.$$

Here $M_1 \in \mathbb{R}^{(2n_0 s+sm+n_0 m)\times(2n_0 s+sm+n_0 m)}$ and $M_2 \in \mathbb{R}^{(2n_0 s+sm+n_0 m)\times\infty}$ are matrices that depend on $(k_0(z), l_0(z))$ and on $\mathbb{E}\varepsilon_t \varepsilon_t', \mathbb{E}z_t z_t'$ but not on $f, \hat{\Xi}_f^+$ or $\hat{\Xi}_p^-$. Further $\Gamma_{w,w} = [\mathbb{E}w_j w_i']_{i,j\in\mathbb{N}}$. Explicit expressions for M_1 and M_2 can be derived from the proof of Theorem 1 in Bauer and Ljung (2002). The central term here is $\mathcal{O}_f^\dagger := (\mathcal{O}_f'(\Xi_f^+)'\Xi_f^+\mathcal{O}_f)^{-1}\mathcal{O}_f'(\Xi_f^+)'\Xi_f^+$ where $\hat{\Xi}_f^+ \to \Xi_f^+$. The design

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

parameters f and $\hat{\Xi}_f^+$ influence V_f only via $\mathcal{O}_f^\dagger N_{t,f}^+$. Corollary 2 of Bauer and Ljung (2002) shows that for each f the minimum of V_f is achieved using $\Xi_f^+ = \lim_{T\to\infty}\langle Y_{t,f}^{+,\Pi}, Y_{t,f}^{+,\Pi}\rangle^{-1/2}$. Given the identity of the asymptotic distribution of the estimators obtained in the first step in the Hannan–Rissanen procedure with subspace estimators using a particular choice of $\hat{\Xi}_f^+$ (Deistler et al., 1995, Proposition 4.1) it follows that the estimates obtained using CCA improve upon the Hannan–Rissanen estimates in the case of no inputs. Moreover, the minimum value is monotonically decreasing in f. Therefore in the case of no exogenous inputs or white exogenous inputs and conditional homoskedasticity f should be chosen large, and the CCA choice for the weighting matrices should be used under the assumption of specifying the true order in the estimation.

Further specializing, consider the situation of no inputs. Then using $f = p \geq -d\log T/(2\log\rho_0), p = o((\log T)^a)$ for some $d > 1$ the CCA estimate $\hat{\theta}_\alpha$ is asymptotically equivalent to the pseudo-maximum-likelihood estimate $\tilde{\theta}_\alpha$ maximizing the pseudo-likelihood based on Gaussianity and independence of $(\varepsilon_t)_{t\in\mathbb{Z}}$ over Θ_α. Here equivalence means that $\sqrt{T}(\hat{\theta}_\alpha - \tilde{\theta}_\alpha) = o_P(1)$. For conditionally homoskedastic innovations this is shown in Bauer (2000). Hence if no inputs are present and the innovations are independent and identically Gaussian distributed CCA for $f = p$ is not only optimal among the state approach subspace algorithms but in fact efficient. In other words, in this case one does not pay a price in terms of asymptotic accuracy for the numerical advantages. Bauer (2004c) extends the result that $\sqrt{T}(\hat{\theta}_\alpha - \tilde{\theta}_\alpha) = o_P(1)$ to the conditionally heteroskedastic case (i.e., under the assumption on ε_t stated in Assumption 1). This implies that for conditionally heteroskedastic or non-Gaussian innovation processes CCA does not achieve the optimal accuracy and its asymptotic properties are analogous to the asymptotic properties of the pseudo-maximum-likelihood estimator ignoring the conditional heteroskedasticity.

In the case of general colored exogenous input process $(z_t)_{t\in\mathbb{N}}$ examples have been found where CCA does not achieve optimal performance (for an example, see Bauer and Jansson, 2000). This lack of optimality has been the motivation for the proposal of slightly modified versions of CCA in Peternell et al. (1996). Although these algorithms decrease the asymptotic variance in some cases (see the calculations of the asymptotic variance in Bauer, 1998, p. 131) no analytical results are known comparing the asymptotic distributions of the basic CCA algorithms and these refined methods in general.

All results up to now relate to the case of correctly specified order n_0. In fact few results have been obtained up to now in misspecified cases. For fixed f under the assumptions given in the sufficient conditions for consistency $\|\hat{\Xi}_f^+\hat{\beta}_1\hat{\Xi}_p^- - U_n(f,p)\Sigma_n(f,p)V_n(f,p)'\|_2$ converges to zero. Let $\hat{U}_{\tilde{n}} \in \mathbb{R}^{fs\times\tilde{n}}$ denote the matrix of left singular vectors corresponding to the dominating $\tilde{n}$ eigenvalues of $\hat{\Xi}_f^+\hat{\beta}_1\hat{\Xi}_p^-$. Then the space spanned by the columns of $\hat{U}_{\tilde{n}}$ converges in the gap metric for $\liminf_{p\to\infty}[\sigma_{\tilde{n}}(f,p) - \sigma_{\tilde{n}+1}(f,p)] > 0$ (for a definition of the gap metric see Chatelin, 1983, p. 86). Consequently it can be shown that for $\liminf_{p\to\infty}[\sigma_{\tilde{n}}(f,p) - \sigma_{\tilde{n}+1}(f,p)] > 0$ the subspace estimators

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

$(\hat{k}_{\tilde{n}}(z), \hat{l}_{\tilde{n}}(z))$ using order $\tilde{n} < n_0$ (fixed and deterministic) converge a.s. to an approximating system $(k_{\tilde{n}}(z), l_{\tilde{n}}(z))$ that in some cases can be explicitly given (for details see Bauer, 1998, Sect. 5.3.2). However, nothing is known about the properties of the approximating system.

5. ORDER ESTIMATION

Up to now it was always assumed that the true order of the system is known. This section will discuss methods to specify the order.

The first step in the subspace procedure is a reduced rank regression in the equation

$$Y_{t,f}^{+} = \beta_1 W_{t,p}^{-} + \beta_2 Z_{t,f}^{+} + N_{t,\perp}$$

under the constraint $\operatorname{rank}(\beta_1) = n$. The first idea for order estimation hence would be to use the reduced rank regression framework. Under the assumption that $N_{t,\perp}$ is i.i.d. Gaussian the likelihood ratio between using $\operatorname{rank}(\beta_1) = n$ and $\operatorname{rank}(\beta_1) = n + 1$ is equal to $-\log(1 - \hat{\sigma}_{n+1}^2)$ where $\hat{\sigma}_i$ denotes the ith largest singular value of $\langle Y_{t,f}^{+,\Pi}, Y_{t,f}^{+,\Pi} \rangle^{-1/2} \langle Y_{t,f}^{+,\Pi}, W_{t,p}^{-,\Pi} \rangle \langle W_{t,p}^{-,\Pi}, W_{t,p}^{-,\Pi} \rangle^{-1/2}$ (Anderson, 1951). But $N_{t,\perp} = N_{t,f}^{+} + \mathcal{O}_f (A - KC)^p x_{t-p}^{\Pi}$ where

$$x_{t-p}^{\Pi} = x_{t-p} - \mathbb{E} x_{t-p} \begin{pmatrix} W_{t,p}^{-} \\ Z_{t,f}^{+} \end{pmatrix}' \left[\mathbb{E} \begin{pmatrix} W_{t,p}^{-} \\ Z_{t,f}^{+} \end{pmatrix} \begin{pmatrix} W_{t,p}^{-} \\ Z_{t,f}^{+} \end{pmatrix}' \right]^{-1} \begin{pmatrix} W_{t,p}^{-} \\ Z_{t,f}^{+} \end{pmatrix}.$$

Therefore $(N_{t,\perp})_{t=p+1,\ldots,T-f+1}$ is not uncorrelated over time and moreover depends on p. Consequently the distribution of the pseudo-likelihood ratio does not follow from standard theory. To the best of my knowledge the asymptotic distribution of the likelihood ratio is not known.

An alternative idea then is to formally use the information criteria defined as

$$SVC(n; C_T) := \|\hat{R}_n\|^2 + \frac{d(n) C_T}{T}, \qquad n = 0, 1, \ldots, M - 1,$$

where $\|\hat{R}_n\|^2 = -\sum_{j=n+1}^{M} \log(1 - \hat{\sigma}_j^2)$ and $M := \min(fs, p(s+m))$ denotes the number of nonzero singular values. Here $d(n) = 2sn + sm + nm$ equals the number of parameters included in a model of order n and $C_T > 0, C_T/T \to 0$ denotes the penalty term punishing complex models. Here *SVC* stands for *singular value criterion*. It should be stressed that this criterion is not an information criterion being defined in a misspecified case for the same reasons that the likelihood ratio does not follow standard asymptotics. This criterion has been discussed by Camba-Mendez and Kapetanios (2001). Peternell (1995) introduced a similar criterion using $\|\hat{R}_n\|^2 = \sum_{j=n+1}^{M} \hat{\sigma}_j^2$, the Frobenius norm of the approximation error. Bauer (2000) proposes $\|\hat{R}_n\|^2 = \hat{\sigma}_{n+1}^2$, i.e., the two norm of the approximation error, based on simulations. The advantage of the last

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

two criteria is that they can be used for general weighting matrices $\hat{\Xi}_f^+$ and $\hat{\Xi}_p^-$, whereas the first criterion uses the CCA choices to guarantee $\hat{\sigma}_i < 1$. In the following discussion we will only deal with these three choices.

The order is estimated as the minimizing integer of $SVC(n;C_T)$ resulting in an estimator $\hat{n}$. The rationale behind these estimates is that $\hat{\Xi}_f^+ \hat{\beta}_1 \hat{\Xi}_p^- - U_{n_0}(f,p)\Sigma_{n_0}(f,p)V_{n_0}(f,p)' = o(1)$ where the matrix $U_{n_0}(f,p)\Sigma_{n_0}(f,p) \times V_{n_0}(f,p)'$ contains at most n_0 nonzero singular values, the remaining ones being zero. Recall that it is the rank of β_1 that we want to infer from the noisy estimate $\hat{\beta}_1$. Letting the penalty factor $C_T > 0$ tend to infinity such that C_T/T converges more slowly to zero than the maximal estimation error in the singular values then ensures that asymptotically the penalty dominates and asymptotically the order is not overestimated, implying that[5] $\hat{n} \leq \lim_{T\to\infty} \operatorname{rank}(U_{n_0}(f,p)\Sigma_{n_0}(f,p) \times V_{n_0}(f,p)')$. In fact we have the following condition.

Convergence of Order Estimation Using SVC(n;C_T). Let the processes $(y_t)_{t\in\mathbb{N}}$ and $(z_t)_{t\in\mathbb{Z}}$ fulfill Assumption 1 with homoskedastic innovations $(\varepsilon_t)_{t\in\mathbb{Z}}$ and $(\eta_t)_{t\in\mathbb{Z}}$ where the true order is denoted as n_0. Additionally it is assumed that for some constants $0 < c < \infty$ and $0 < \rho < 1$ it holds that $\|K_{z,j}\| \leq c\rho^j, j \in \mathbb{N}$. Assume that f and p are either fixed and finite or tend to infinity, in which case $f \leq p, p = O((\log T)^a), a < \infty$ is assumed. Let $U_{n_0}(\infty)\Sigma_{n_0}(\infty)V_{n_0}(\infty)'$ denote the limit of $\hat{\Xi}_f^+ \hat{\beta}_1 \hat{\Xi}_p^-$, which for $f \to \infty$ or $p \to \infty$ (a.s.) has to be interpreted in an appropriate sense (for details see Bauer, 2000). Then $C_T > 0, C_T/T \to 0, C_T/(fp \log\log T) \to \infty$ is sufficient for $\hat{n} := \arg\min_{0\leq n\leq M} SVC(n;C_T)$ to converge to $\operatorname{rank}(U_{n_0}(\infty)\Sigma_{n_0}(\infty)V_{n_0}(\infty)')$.

Note that $U_{n_0}(\infty)\Sigma_{n_0}(\infty)V_{n_0}(\infty)'$ depends on the choice of f and p. Jansson and Wahlberg (1998) show by example that $f \geq n_0 + 1, p \geq n_0$ is not sufficient for $\sigma_{n_0}(f,p) > 0$. However, violation of this condition is rare because Bauer and Jansson (2000) show that for fixed pseudostationary input sequence $(z_t)_{t\in\mathbb{N}}$ that is persistently exciting of order $f + p - 1$ (i.e., fulfilling conditions (4.1.1.), (4.1.14), and (4.2.21) of Hannan and Deistler, 1988, the latter with $f + p - 1$ inserted for sp; these assumptions hold for a particular realization of a process $(z_t)_{t\in\mathbb{N}}$ fulfilling Assumption 1) for fixed f and p the subset of pairs of transfer functions in M_{n_0} with $\sigma_{n_0}(f,p) = 0$ is the complement of an open and dense subset. This implies that for any given f,p it is unlikely that the true system happens to have the property $\sigma_{n_0}(f,p) = 0$. Nevertheless the problem is relevant because for systems close (in an appropriate sense) to these exceptional systems the smallest singular value will be very small and hence a large number of observations will be needed to differentiate these singular values from noise. Chui (1997) gives sharp conditions on f and p for $n_0 = \operatorname{rank}(U_{n_0}(f,p) \times \Sigma_{n_0}(f,p)V_{n_0}(f,p)')$ that imply the sufficient conditions $\min(f,p) \geq 3n_0$. Another set of sufficient conditions is $f \geq n_0$ and $p \to \infty$.

Bauer (2001, Theorem 3) states consistency of order estimation using $SVC(n;C_T)$ for $p \to \infty$ under the assumptions on the input given previ-

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

ously. Only the Frobenius norm and the two norm of $\hat{R}_n$ in the definition of $SVC(n;C_T)$ are considered in that paper. There convergence of $\hat{n}$ for fixed and finite f and p also is discussed (allowing a harmonic process to be added to the process fulfilling the assumptions just stated). Because again mainly the uniform convergence of the sample covariances is used in the proof of Theorem 3 of Bauer (2001) the consistency proof for $SVC(n;C_T)$-based order estimation given there can be straightforwardly generalized to yield the result given earlier.[6] In particular note that the result also applies in the case of f and p depending on the sample size possibly tending to infinity a.s. This is the case for $f = p = \lfloor d\hat{p}_{BIC} \rfloor, d > 0$. In this case the condition $C_T/(fp \log\log T) \to \infty$ has to hold a.s. Again the sufficient conditions are by no means necessary.

Simulation evidence shows that for $\|\hat{R}_n\|^2 = \hat{\sigma}^2_{n+1}$ the sufficient condition on the penalty term seems to be too pessimistic, leading to a high risk of underestimating the order in finite samples (see the simulations in Bauer, 2001). The criterion based on the two norm seems to be less sensitive with respect to the choice of f and p than the method based on the Frobenius norm. No thorough motivation for any penalty term has been found.

A second approach to the problem of order specification is to note that the problem of selecting the order is equivalent to inferring the rank of the noisy matrix $\hat{\beta}_1$. Different procedures using sequential testing are developed in Sorelius (1999) in the case of no exogenous inputs included ($m = 0$) and Camba-Mendez and Kapetanios (2001). All these procedures center on the assessment of the rank of $\langle Y^+_{t,f}, W^-_{t,f} \rangle$ and proceed by sequentially testing for singularity of the matrix and increasing the size by one in case of a rejection until the test is not rejected for the first time.

The Hankel structure of the matrix $\langle Y^+_{t,f}, W^-_{t,f} \rangle$ complicates the derivation of the test statistic considerably. Additionally one runs into a sequential testing problem. Having said that, it should also be mentioned that this approach is statistically much sounder than the ad hoc definition of criteria such as $SVC(n;C_T)$. At present the development of this strand of literature is deemed to be premature, however. We will not go into detail in this respect, and we refer the reader to Sorelius (1999) and Camba-Mendez and Kapetanios (2001) for details.

A third approach to order estimation uses analogies to traditional information criteria. In the context of pseudo-maximum-likelihood estimation the order is estimated using information criteria such as the Akaike information criterion (AIC) or BIC (for details see Hannan and Deistler, 1988, Ch. 5). Assume that the Gaussian likelihood is maximized over M_n for $n = 0,1,2,\ldots,H_T$ for some upper bound H_T such as $(\log T)^a$, $0 < a < \infty$, resulting in an estimate $\tilde{\theta}_{n,\alpha_n}$ for each n. Then an estimate of the one step ahead prediction errors $\hat{\varepsilon}_t(\tilde{\theta}_{n,\alpha_n})$ is obtained from Kalman filtering (for details see Hannan and Deistler, 1988, Ch. 3), leading to an estimated one step ahead prediction variance $\hat{\Omega}(\tilde{\theta}_{n,\alpha_n}) := T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t(\tilde{\theta}_{n,\alpha_n}) \hat{\varepsilon}_t(\tilde{\theta}_{n,\alpha_n})', n = 0,\ldots,H_T$. Then the information criterion estimate is taken as a minimizing integer of the criterion

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

$$\widetilde{IC}(n;C_T) = \log\det\hat{\Omega}(\tilde{\theta}_{n,\alpha_n}) + \frac{d(n)C_T}{T},$$

where again $d(n) = 2sn + sm + nm$ denotes the number of parameters included in a model of order n and $C_T > 0$, $C_T/T \to 0$ denotes the penalty term punishing complex models. BIC uses $C_T = \log T$ and AIC uses $C_T = 2$.

For subspace methods there are two obvious generalizations of this approach. The pseudo-maximum-likelihood estimator $\tilde{\theta}_{n,\alpha_n}$ could simply be replaced by the subspace estimator $\hat{\theta}_{n,\alpha_n}$ in the definition of $\hat{\Omega}(\hat{\theta}_{n,\alpha_n})$. Alternatively the estimation of the equation $y_t = \hat{C}(n)\hat{x}_t(n) + \hat{D}(n)z_t + \hat{\varepsilon}_t(n)$ immediately leads to an estimate of the residual variance $\hat{\Omega}_n := (T-p)^{-1}\sum_{t=p+1}^{T}\hat{\varepsilon}_t(n)\hat{\varepsilon}_t(n)'$ that can be inserted into $\widetilde{IC}(n;C_T)$ for $\hat{\Omega}(\tilde{\theta}_{n,\alpha_n})$. Here the dependence of $\hat{\varepsilon}_t(n)$ on the usage of the n-dimensional state $\hat{x}_t(n) := \hat{V}_n'(\hat{\Xi}_p^-)^{-1}W_{t,p}^- \in \mathbb{R}^n$ is emphasized. These two alternatives are fundamentally different. Whereas $\hat{\Omega}(\hat{\theta}_{n,\alpha_n})$ takes the full system dynamics into account, $\hat{\Omega}_n$ is calculated on the basis of the n-dimensional estimated state, which in the limit need not correspond to an n-dimensional system.

To illustrate the point assume for the moment that we are selecting the model order based on the knowledge of the true state $x_t \in \mathbb{R}^{n_0}$ corresponding to a state space representation such that $C = [I_s,0]$ where $n_0 > s$ is assumed. Then let $x_t(n), 1 \le n \le n_0$ denote the vector of the first n components of x_t. Now order estimation based on $\hat{\Omega}_n$, i.e., the residual variance of the equation $y_t = \hat{C}(n)x_t(n) + \hat{D}(n)z_t + \hat{\varepsilon}_t(n)$, will result in an order estimator converging to s (for suitable choice of the penalty term C_T) because the remaining coordinates of x_t do not contribute to the prediction of y_t. The order estimate based on $\hat{\Omega}(\hat{\theta}_{n,\alpha_n})$, however, will lead to consistent order estimates because in this case for each order n a new estimate of the state also is obtained from the Kalman filter, which for $n < n_0$ will not coincide with $x_t(n)$.

For the subspace methods the estimated state $\hat{x}_t := \hat{V}_{n_0}'(\hat{\Xi}_p^-)^{-1}W_{t,p}^-$ may not correspond to a true system where the last columns of C are zero for any pair $(k_0(z), l_0(z))$ because the limiting system is determined implicitly in the SVD. But problems for the order estimation procedure are expected also in situations where the norm of the last column is small. An example in this respect is provided in Bauer (2001), where some asymptotic properties of the order estimate obtained by minimizing $IVC(n;C_T) := \log\det\hat{\Omega}_n + d(n)C_T/T$ also are derived. Here *IVC* stands for innovation variance criterion. Because of these problems the use of the criterion is strongly discouraged.

Finally in this section the asymptotic properties of the order selection based on

$$\widehat{IC}(n;C_T) := \log\det\hat{\Omega}(\hat{\theta}_{n,\alpha_n}) + \frac{d(n)C_T}{T}, \qquad 0 \le n \le M,$$

where $\hat{\theta}_{n,\alpha_n}$ denotes the CCA estimator using $f = p, p \ge -d\log T/(2\log\rho_0)$, $d > 1, p = O((\log T)^a), p$ chosen deterministically, are investigated. Addition-

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

ally assume that $\mathbb{E}\{\varepsilon_t \varepsilon_t' | \mathcal{F}_{t-1}\} = \mathbb{E}\varepsilon_t \varepsilon_t' > 0$ and $m = 0$ so that no inputs are included. In this case it can be shown that

$$\widehat{IC}(n; C_T) \begin{cases} = \widetilde{IC}(n_0; C_T) + o_P(\sqrt{\log\log T}/T), & n = n_0, \\ \geq \widetilde{IC}(n; C_T), & n \neq n_0 \end{cases}$$

and therefore $C_T > 0, C_T/T \to 0, C_T/\log\log T \to \infty$ is sufficient for weak consistency of the order estimate obtained by minimizing $\widehat{IC}(n; C_T)$ over $0 \leq n \leq ps$. The expression for $n \neq n_0$ is immediate because $\tilde{\theta}_{n,\alpha_n}$ minimizes $\log\det\hat{\Omega}(\theta)$. For $n = n_0$ the expression is a consequence of $\tilde{\theta}_{n,\alpha_n} - \hat{\theta}_{n,\alpha_n} = o_P(T^{-1/2})$. The details of the proof are omitted.

In the situation without inputs z_t the criterion $\widehat{IC}(n; C_T)$ has the tremendous advantage compared to $\widetilde{IC}(n; C_T)$ that once the state has been estimated for a maximal order the subspace estimators for a range of different orders can be obtained extremely numerically efficiently (in fact, only one regression has to be estimated; for details, see Bauer, 2001, p. 1564). Therefore, Kalman filtering to obtain $\hat{\varepsilon}_t(\hat{\theta}_{n,\alpha_n})$ constitutes the only main computational cost. This idea also has another interesting feature. As stated in Section 4, CCA produces asymptotically optimal estimators in the correctly specified case for Gaussian innovations. For misspecified cases no such result is known, and therefore the subspace estimator might result in worse values of the criterion function than the pseudo-maximum-likelihood estimator that minimizes $\log\det\hat{\Omega}_n(\theta)$. Hence order estimation based on $\widehat{IC}(n; \log T)$ could be expected to improve on BIC order estimation. It should be stressed, however, that the performance in cases where there exists a true model might not be the only criterion to assess the quality of a model selection procedure (cf. the discussion in Shibata, 1981).

6. RESULTS IN THE INTEGRATED CASE

Because the state approach is primarily based on regressions it seems to be tempting also to apply the algorithm to integrated processes. The assumptions in this section will be as follows.

Assumption 2. The process $(y_t)_{t\in\mathbb{N}}$ is generated by a minimal state space system (2) of order n_0 of the form

$$A = \begin{bmatrix} I_c & 0 \\ 0 & A_\bullet \end{bmatrix}, \qquad K = \begin{bmatrix} K_1 \\ K_\bullet \end{bmatrix}, \qquad C = [C_1 \quad C_\bullet], \tag{7}$$

where $C_1 \in \mathbb{R}^{s\times c}, K_1 \in \mathbb{R}^{c\times s}$, and $c \leq s$ and the remaining matrices are of conformable size. Here $m = 0$, and hence no exogenous input is present. The initial state x_1 is assumed to be chosen according to $x_1 = [0, (\sum_{j=0}^{\infty} A_\bullet^j K_\bullet \varepsilon_{-j})']'$. Here $\lambda_{max}(A_\bullet) < 1$ and $0 < \rho_0 = \lambda_{max}(A - KC) < 1$ is assumed. The strictly stationary ergodic noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_t = \sigma\{\varepsilon_t, \varepsilon_{t-1}, \ldots\}$ such that $\mathbb{E}\{\varepsilon_t | \mathcal{F}_{t-1}\} = 0$, $\mathbb{E}\{\varepsilon_t \varepsilon_t' | \mathcal{F}_{t-1}\} =$

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

$\mathbb{E}\varepsilon_t\varepsilon_t' = \Omega > 0, \mathbb{E}\varepsilon_{t,i}^4 < \infty, 1 \le i \le s$ where $\varepsilon_{t,i}$ denotes the ith component of the vector ε_t.

As a consequence of the specific structure of the state space system $CA^{j-1}K = C_1K_1 + C_\bullet A_\bullet^{j-1}K_\bullet, j > 0$ follows. This basically amounts to a partial fraction decomposition of $k_0(z) := I_s + zC(I_n - zA)^{-1}K$. Using (3) it follows that ($t \in \mathbb{N}$)

$$y_t = \sum_{j=1}^{t-1} C_1K_1\varepsilon_j + \varepsilon_t + \sum_{j=1}^{t-1} C_\bullet A_\bullet^{j-1}K_\bullet\varepsilon_{t-j} + CA^{t-1}x_1.$$

Choosing $x_1 := [0, (\sum_{j=0}^{\infty} A_\bullet^j K_\bullet \varepsilon_{-j})']'$ it follows that $CA^{t-1}x_1 = \sum_{j=t}^{\infty} C_\bullet A_\bullet^{j-1} \times K_\bullet\varepsilon_{t-j}$. Hence there exist initial conditions y_0 such that $(y_t - y_{t-1})_{t\in\mathbb{N}}$ is stationary and also $(C_\perp' y_t)_{t\in\mathbb{N}}$ is stationary for each $C_\perp \in \mathbb{R}^{s\times(s-c)}, C_\perp' C_1 = 0$. Therefore $(y_t)_{t\in\mathbb{N}}$ is $I(1)$ and admits $s - c$ cointegrating relations because minimality implies that C_1 is of full column rank and K_1 of full row rank (see Bauer and Wagner, 2002b, Theorem 1). The cointegrating space equals the orthogonal complement of the column space of C_1. Restricting x_1 to be as in Assumption 2 is no severe loss of generality. All results to follow continue to hold for random x_1 with finite second moment and independent of $\varepsilon_t, t \in \mathbb{N}$.

Assume that CCA using $f \ge n_0$ fixed and finite independent of the sample size is used to estimate the transfer function $\hat{k}(z) := I_s + z\hat{C}(I_{n_0} - z\hat{A})^{-1}\hat{K}$. Here n_0 denotes the true order, which is assumed to be known and used in the estimation. Let p be chosen deterministically such that $p = o((\log T)^a)$ for some arbitrary $a < \infty$ and $p \ge -d \log T/(2 \log \rho_0), d > 2$. Then for $(y_t)_{t\in\mathbb{N}}$ generated according to Assumption 2 the following statements hold.

1. $\hat{k}(z) \to k_0(z) = I_s + zC(I_n - zA)^{-1}K$ a.s. (Bauer, 2004b, Theorem 3.1). The transfer function—and hence the long-run and short-run dynamics—is estimated consistently without using the knowledge that $(y_t)_{t\in\mathbb{N}}$ is integrated in the algorithm.
2. $T^\gamma(\hat{\sigma}_i - 1) \to 0$ in probability for $i = 1, \dots, c$ for any $\gamma < 1$ where $1 \ge \hat{\sigma}_1 \ge \cdots \ge \hat{\sigma}_c$ denote the c dominant singular values of $\hat{\Xi}_f^+ \hat{\beta}_1 \hat{\Xi}_p^-$ (Bauer and Wagner, 2002a, Theorem 4).[7] A plot of the estimated singular values provides visual information on possible numbers of common trends (see Bauer and Wagner, 2002a, Figure 1).
3. If $\hat{n} := \arg\min_{0\le n\le fs} SVC(n; C_T)$ where SVC uses $\|\hat{R}_n\|^2 = \hat{\sigma}_{n+1}^2$ then $\hat{n} \to n_0$ in probability for $C_T > 0, C_T/T \to 0, C_T/(p \log\log T) \to \infty$ (see Bauer and Wagner, 2002a, Theorem 3).
4. If $\hat{\lambda}_i, i = 1, \dots, n_0, |\hat{\lambda}_1| \ge |\hat{\lambda}_2| \ge \cdots \ge |\hat{\lambda}_{n_0}|$, denote the eigenvalues of $\hat{A}$ ordered decreasing in modulus then the asymptotic distribution of $T(\hat{\lambda}_1 - 1, \dots, \hat{\lambda}_c - 1)$ is identical to the asymptotic distribution of the vector of the c eigenvalues (ordered decreasing in modulus) of $T(\sum_{t=2}^{T} \nu_t n_t') \times$

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

$(\sum_{t=2}^{T} n_t n_t')^{-1}$ where $(\nu_t)_{t\in\mathbb{N}}, \nu_t \in \mathbb{R}^c$ is i.i.d. standard normally distributed and $n_t = \sum_{j=1}^{t-1} \nu_j$ (Bauer, 2004b, Theorem 4.1). This implies that the largest (in modulus) eigenvalues of the estimated matrix $\hat{A}$ can be used to test for the number of common trends. The resulting test statistic is pivotal and the asymptotic distribution standard in unit root tests.

Although the transfer function estimator is consistent according to statement 1 the (untransformed) estimators for the system matrices will diverge if the state is estimated as $\hat{x}_t := \hat{V}_n'(\hat{\Xi}_p^-)^{-1} W_{t,p}^-$ because this implies $T^{-1}\sum_{t=p+1}^{T-f+1} \hat{x}_t \hat{x}_t' = I_{n_0}$. Therefore the nonstationary state components are downweighed and the corresponding columns in $\hat{C}$ are multiplied by $\sqrt{T}$. Of course, if the system matrix estimates are transformed into an echelon overlapping form the system matrix estimators also will be consistent.

Note that as a result of consistency the c largest eigenvalues of $\hat{A}$ will be close to one albeit different without investing the knowledge of c in the algorithm. If knowledge of c is used in the algorithm a reduced rank regression in the equation

$$\hat{x}_{t+1} - \hat{x}_t = (\tilde{A} - I_{n_0})\hat{x}_t + \tilde{K}\hat{\varepsilon}_t + r_t$$

subject to $\operatorname{rank}(\tilde{A} - I_{n_0}) = n_0 - c$ ensures that $(\tilde{A}, \hat{C}, \tilde{K})$ corresponds a.s. to an estimated system having a representation as in Assumption 2.

Consistency ensures that CCA can be used to obtain an initial guess for subsequent pseudo-maximum-likelihood estimation. Moreover CCA provides easily accessible information on the model order (via the estimated singular values) and the number of common trends (via the eigenvalues of the matrix $\hat{A}$). This information can be used to specify the model prior to the application of pseudo-maximum-likelihood estimation.

The topic of the estimation of unit root processes using subspace methods really is only in its infancy. Although Bauer (2004b) derives the asymptotic distribution of the subspace estimators, this does not reveal much information. In particular no comparison to pseudo-maximum-likelihood estimators has been achieved so far.

There is a subtle difference in the choice of p for the integrated case as opposed to the stationary case that seems worth mentioning: In the stationary case it has been argued that $\lfloor d\hat{p}_{BIC} \rfloor, d > 1$ is a reasonable choice. In the integrated case, however, $d > 2$ is imposed in the sufficient conditions given previously. The reason for this change is simple: $d > 1$ ensures that $(A - KC)^p = o(T^{-1/2})$, which is sufficient in the stationary case to ensure that the term $\mathcal{O}_f(A - KC)^p x_{t-p}$ does not influence the asymptotic distribution. In the integrated case on the contrary the convergence in some components of the system matrices is of order $O_P(T^{-1})$, and therefore $d > 2$ is imposed, leading to $(A - KC)^p = o(T^{-1})$.

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

7. SIMULATIONS

It has been argued that the main advantage of subspace identification lies in the numerical simplicity and reliability as an initial guess. These claims have been confirmed in a number of simulation studies. The statistical properties of CCA have been highlighted by Peternell (1995), Deistler et al. (1995), Peternell et al. (1996), and Bauer (1998). Bauer (2004c) also contains a comparison of the numerical load. In the integrated case Bauer and Wagner (2002a) demonstrate the usage of CCA for estimating the cointegrating rank, and Wagner (2004) examines the finite sample properties of the estimation of the cointegrating space and the predictions based on the CCA estimates. In this section a small simulation study will be used to demonstrate the claims of the paper. The simulation is not intended to mirror any particular application, and the simulated systems are totally arbitrary. The results reflect the findings in the literature and therefore can be seen to be typical. The main point to be made in the simulations, however, is that there are situations in which subspace methods appear to be a useful alternative to the classical estimation methods.

In the simulation experiment CCA is compared to the algorithm PEM of the MATLAB system identification toolbox. This routine uses an initial estimate obtained by a version of N4SID to afterward maximize the Gaussian likelihood in a Gauss–Newton type method. The method has been chosen because of its widespread use in the engineering community. The basic setup is the following: For each experiment three integers (s, n, T) are given where s denotes the output dimension, n the system order, and T the sample size. Based on these integers an ARMA system is drawn randomly and 100 time series of dimension s and sample size $T + 1{,}000$ are generated using this one system and Gaussian white noise of zero mean and variance equal to the identity matrix. The initial conditions are set to zero, which might be seen as unrealistic. Then CCA setting $f = p = 2\hat{p}_{AIC}$ and estimating the order via $SVC(n; \log T)$ based on the two norm $\|\hat{R}_n\|^2 = \hat{\sigma}^2_{n+1}$ is used to estimate the system on the first T data points. After the completion of this estimation another system is estimated on the same data set using the function PEM and the order estimated in CCA. For both estimated systems the one step ahead prediction error on the remaining 1,000 time instants is calculated. As a measure of accuracy $\hat{L}_{T,i} := \log\det\hat{\Omega}_i, i = 1,2$ is chosen where $\hat{\Omega}_i = 1{,}000^{-1}\sum_{t=T+1}^{T+1,000}\hat{\varepsilon}_t\hat{\varepsilon}_t'$ denotes the estimated innovation variance on the validation sample for the two procedures. For the true system $\sqrt{T}\hat{L}_{T,i}$ asymptotically is normally distributed with mean zero and variance $2s$. For each experiment three numbers are reported for each of the two algorithms.

1. The average time spent on the algorithm as estimated from the MATLAB function profile.
2. The average of the 100 accuracy measures $\sqrt{T}\hat{L}_{T,i}$ where $i = 1$ for CCA and $i = 2$ for PEM.
3. The average over the subsample where PEM is not obviously trapped in a local minimum: In some cases the estimation error for PEM is many orders

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

TABLE 1. Numerical and statistical properties of CCA and PEM

	Sample Size	$T = 100$			$T = 1{,}000$			$T = 10{,}000$		
Integers	Algo.	Sec.	Av.	Trim.	Sec.	Av.	Trim.	Sec.	Av.	Trim.
$s = 3, n = 3$	CCA	0.04	16.5	17.6	0.08	0.48	0.48	1.27	0.11	0.11
(1/1.96/3)	PEM	0.25	17.6	17.3	0.42	0.45	0.45	5.70	0.11	0.11
$s = 6, n = 6$	CCA	0.05	71.9	71.7	0.13	4.66	4.66	2.56	−0.02	−0.02
(1/3.04/4)	PEM	0.33	110.3	71.9	2.73	4.59	4.59	20.36	−0.02	−0.02
$s = 9, n = 9$	CCA	0.05	104.9	105.2	0.17	17.4	17.4	3.23	−0.28	−0.28
(1/3/8)	PEM	0.58	188.4	110.3	6.09	17.4	17.4	129.15	−0.27	−0.27

Note: Each entry is calculated on the basis of 100 replications. Seconds are measured by the MATLAB function profile. Av. is the average of $\sqrt{T}\hat{L}_{T,i}$ over all replications. Trim. omits the replications where PEM seems to be stuck in local optima (see text for an exact definition). The number of parameters equals 18, 72, and 162. The numbers below the model specification give the mean of the estimated orders for sample size (100/1,000/10,000).

of magnitude too high compared to the error attained by CCA. Formally these cases are defined somewhat arbitrarily by $|\sqrt{T}\hat{L}_{T,2}| > 1.5|\sqrt{T}\hat{L}_{T,1}|$ and $|\sqrt{T}\hat{L}_{T,2}| > 3$.

These three numbers are reported for each setting of integers (s, n, T) in Table 1. The main findings are shown.

- CCA is always numerically cheaper than PEM. The difference becomes remarkable for large sample sizes and high-dimensional data sets.
- The specification of the initial guess is an issue. If one does not account for local optima the performance of PEM is worse than the performance of CCA in most cases.
- In medium sample sizes (here $T = 1{,}000$) PEM achieves slightly better performance.
- For large sample size PEM does not seem to be trapped in local optima. At the same time the performance is not differentiable from the performance of CCA.
- For small sample sizes SVC picks overly small orders. This is especially the case for high-dimensional data sets. The consequence is a sizable bias.

These findings demonstrate the claims based on asymptotic theory raised in the literature.

8. CONCLUSIONS

The main advantage of CCA and subspace methods in general in comparison to conventional estimation methods based on criterion minimization is the com-

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

putational burden. Because it is based on regressions rather than nonlinear optimization, CCA can be applied in situations where pseudo-maximum-likelihood methods are infeasible. Three such situations are data sets with many inputs/outputs, very large data sets, and input selection problems that require a huge number of models to be estimated. In these cases the computational load is one of the prime issues, and hence subspace methods are a valuable alternative to classical estimation methods.

In other cases CCA may be used to obtain initial guesses that are not worse then traditional methods such as the Hannan–Rissanen method. Additionally, in a regression framework issues such as outliers, missing observations, and recursive estimation are well understood. This provides the basis for simple adaptations of the CCA idea.

When using CCA a number of design parameters have to be set by the user. Only in the case of no or white inputs does the literature provide guidance on how to choose these parameters optimally (with respect to the asymptotic variance). Based on these results (albeit not fully justified) CCA using $f = p = 2\hat{p}_{AIC}$ and the weightings (6) is proposed. Combining this choice with order estimation using $SVC(n;C_T)$ and testing for possible integration one obtains an automated procedure in the sense that the algorithm estimates a model given only the data without any additional interaction with the user. As indicated in the Introduction such automatic procedures should be used with caution. Bauer and de Waele (2003) show that such a procedure works well in some examples but fails to exhibit the expected properties based on the asymptotic theory in other situations. A weak point in the algorithm as stated earlier is the simplistic order estimation, which might cause the observed deficiency. Nevertheless at the very least the subspace estimators provide in many cases good starting points for consecutive analyses.

A very attractive feature of CCA is the demonstrated robustness with respect to the data characteristics inherited from the regression framework. It has been stated previously that key properties such as consistency of the transfer function estimates and system order estimates are robust with respect to conditional heteroskedasticity of the innovations or certain nonstationarities of the data. We therefore consider CCA an interesting addition to the toolbox of time series analysts.

NOTES

1. Here and in the following discussion the integer p does not coincide with the maximal lag length of the ARMAX process used previously.

2. There the restriction $D = 0$ also is imposed. It is obvious that this is not essential for the result.

3. A proof of this statement is contained in a preliminary version of this paper, which can be obtained from the author on request.

4. To be precise: There $(\varepsilon_t)_{t\in\mathbb{Z}}$ is assumed to be i.i.d., where it is remarked that the result also holds in the martingale difference framework considered in this paper.

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

5. Here f and p are functions of the sample size T, which is not emphasized in the notation.

6. A full proof of this fact is contained in a preliminary version of this paper and can be obtained on request.

7. Unfortunately there is an error in that theorem. The given asymptotic distribution is incorrect; the order of convergence is correct, however.

REFERENCES

Anderson, T.W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics* 22, 327–351. Correction, *Annals of Statistics* 8 (1980), 1400.

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

Baillie, R. (1996) Long memory processes and fractional integration in econometrics. *Journal of Econometrics* 73, 5–59.

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. (2000) Comparing the CCA Subspace Method to Pseudo Maximum Likelihood Methods in the Case of No Exogenous Inputs. Technical report, Institute of Econometrics, Operations Research and System Theory, TU Wien.

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

Bauer, D. (2004a) Asymptotic properties of subspace estimators. *Automatica* (forthcoming).

Bauer, D. (2004b) Estimation and Testing in Cointegrated Systems Using Subspace Methods. Technical report, Cowles Foundation for Research in Economics, Yale University.

Bauer, D. (2004c) Using Subspace Methods for Estimating ARMA Models for Multivariate Time Series with Conditionally Heteroskedastic Innovations. Technical report CFDP 1452. Cowles Foundation for Research in Economics, Yale University.

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., M. Deistler, & W. Scherrer (2000) On the impact of weighting matrices in subspace algorithms. In R. Smith (ed.), *Proceedings of the IFAC Conference SYSID'00*. Elsevier.

Bauer, D. & M. Jansson (2000) Analysis of the asymptotic properties of the MOESP type of subspace algorithms. *Automatica* 36, 497–509.

Bauer, D. & L. Ljung (2002) Some facts about the choice of the weighting matrices in Larimore type of subspace algorithms. *Automatica* 38, 763–773.

Bauer, D. & S. de Waele (2003) A finite sample comparison of automatic model selection methods. In P. van den Hof, B. Wahlberg, & S. Weiland (eds.), *Proceedings of the IFAC Symposium on System Identification, SYSID'03, Rotterdam, The Netherlands*. Elsevier.

Bauer, D. & M. Wagner (2002a) Estimating cointegrated systems using subspace algorithms. *Journal of Econometrics* 111, 47–84.

Bauer, D. & M. Wagner (2002b) A Canonical Form for Unit Root Processes in the State Space Framework, Technical report, TU Wien, Austria.

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.

Chatelin, F. (1983) *Spectral Approximation of Linear Operators*. Academic Press.

Chiuso, A. & G. Picci (2003) The asymptotic variance of subspace estimates. *Journal of Econometrics* 118, 257–291.

Chui, N. (1997) Subspace methods and informative experiments for system identification. Ph.D. thesis, Cambridge University.

Dahlen, A. & W. Scherrer (2004) The relation of the CCA subspace method to a balanced reduction of an autoregressive model. *Journal of Econometrics* 118, 293–312.

Deistler, M., K. Peternell, & W. Scherrer (1995) Consistency and relative efficiency of subspace methods. *Automatica* 31, 1865–1875.

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

Durbin, J. (1960) The fitting of time series models. *Review of International Institute of Statistics* 28, 233–244.
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.
Jansson, M. & B. Wahlberg (1998) On consistency of subspace methods for system identification. *Automatica* 34, 1507–1519.
Kuersteiner, G. (2005) Automatic inference for infinite order vector autoregressions. *Econometric Theory* (this issue).
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.
Lewis, R. & G. Reinsel (1985) Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis* 16, 393–411.
Peternell, K. (1995) Identification of Linear Dynamic Systems by Subspace and Realization-Based Algorithms. Ph.D. thesis, TU Wien.
Peternell, K., W. Scherrer, & M. Deistler (1996) Statistical analysis of novel subspace identification methods. *Signal Processing* 52, 161–177.
Shibata, R. (1981) An optimal autoregressive spectral estimate. *Annals of Statistics* 9, 300–306.
Sorelius, J. (1999) Subspace-Based Parameter Estimation Problems in Signal Processing. Ph.D. thesis, Uppsala University, Sweden.
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.
Wagner, M. (2004) A comparison of Johansen's, Bierens' and the subspace algorithm method for cointegration analysis. *Oxford Bulletin of Economics and Statistics* 66, 399–424.
Zeiger, H.P. & A.J. McEwen (1974) Approximate linear realizations of given dimension via Ho's algorithm. *IEEE Transaction on Automatic Control* 19, 153.

APPENDIX A: RELATIONS BETWEEN ARMAX AND STATE SPACE REPRESENTATIONS

The properties of the stochastic processes $(\bar{y}_t)_{t\in\mathbb{N}}$ being the solutions to VDEs (1) are well understood (for a detailed discussion see, e.g., Hannan and Deistler, 1988, Chs. 1 and 2). Typically the ARMAX system is described using three matrix polynomials (in $z \in \mathbb{C}$) $a(z) := I_s + A_1 z + \cdots + A_p z^p, b(z) := I_s + B_1 z + \cdots + B_q z^q, d(z) := D_0 + D_1 z + \cdots + D_r z^r$ and the initial conditions for $\bar{y}_t, z_t, \varepsilon_t$. Given the noise $(\varepsilon_t)_{t\in\mathbb{N}}$, the input $(z_t)_{t\in\mathbb{N}}$, and the system $(a(z), b(z), d(z))$ each set of initial conditions determines a unique solution process $(\bar{y}_t)_{t\in\mathbb{N}}$. Holding everything else fixed the difference $v_t := \bar{y}_t^1 - \bar{y}_t^2$ between two such solutions $(\bar{y}_t^1)_{t\in\mathbb{N}}$ and $(\bar{y}_t^2)_{t\in\mathbb{N}}$ for different initial conditions is a solution to the homogenous equation

$$v_t + A_1 v_{t-1} + \cdots + A_p v_{t-p} = 0, \qquad t \in \mathbb{N}. \tag{A.1}$$

The matrix polynomials $(a(z), b(z), d(z))$ are called *ARMAX representation* of the pair of rational transfer functions $k(z) := a(z)^{-1}b(z) = \sum_{j=0}^{\infty} K_j z^j, l(z) := a(z)^{-1}d(z) = \sum_{j=0}^{\infty} L_j z^j$ that converge at least for $z \in \mathbb{C}$ in a neighborhood of the origin. Let $\tilde{y}_t = \sum_{j=0}^{t-1}(K_j \varepsilon_{t-j} + L_j z_{t-j}), t \in \mathbb{N}$. It follows that $(\tilde{y}_t)_{t\in\mathbb{N}}$ is the solution to the VDE(1) with zero initial conditions and any solution $(\bar{y}_t)_{t\in\mathbb{N}}$ to the VDE(1) can be written as $\bar{y}_t =$

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

$\tilde{y}_t + v_t$ where $(v_t)_{t\in\mathbb{N}}$ solves the homogenous equation (8). Note that $\tilde{y}_t$ depends on the pair of transfer functions $(k(z), l(z))$ but not on the particular representation $(a(z), b(z), d(z))$. In other words except for initial effects the pair of transfer functions $(k(z), l(z))$ fully specifies the input/output behavior of ARMAX systems.

Following p. 15 of Hannan and Deistler (1988) for an ARMAX process $(\bar{y}_t)_{t\in\mathbb{N}}$ define the state vector $x_t := [\bar{y}'_{t-1}, \ldots, \bar{y}'_{t-p}, z'_{t-1}, \ldots, z'_{t-r}, \varepsilon'_{t-1}, \ldots, \varepsilon'_{t-q}]'$, $t \in \mathbb{N}$ and the matrices

$$A = \left[\begin{array}{cccc|ccc|ccc} -A_1 & \ldots & \ldots & -A_p & D_1 & \ldots & D_r & B_1 & \ldots & B_q \\ I_s & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & I_s & & & & & & & \\ \hline & & & & I_m & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & I_m & & & \\ \hline & & & & & & & I_s & & \\ & & & & & & & & \ddots & \\ & & & & & & & & & I_s \end{array}\right],$$

$$[B, K] = \left[\begin{array}{cc} D_0 & I_s \\ \hline I_m & \\ \hline & I_s \end{array}\right],$$

$$C = [-A_1 \quad \ldots \quad -A_p \mid D_1 \quad \ldots \quad D_r \mid B_1 \quad \ldots \quad B_q], \qquad D = D_0,$$

where empty positions contain zeros. Then it is straightforward to check that $(\bar{y}_t)_{t\in\mathbb{N}}$ solves the corresponding state space equations, i.e., $\bar{y}_t = Cx_t + Dz_t + \varepsilon_t$ where $x_{t+1} = Ax_t + Bz_t + K\varepsilon_t$. For AR processes the matrix A thus obtained is called the companion form of the polynomial $a(z)$.

On the other hand (3) decomposes a state space process into a component $\tilde{y}_t$ depending only on the pair of transfer functions $(k(z), l(z))$ and the effects of the initial conditions $CA^{t-1}x_1$. The use of the same symbol $\tilde{y}_t$ for the solutions to zero initial conditions for ARMAX and state space representations is no coincidence. Theorem 1.2.1 of Hannan and Deistler (1988) states that any pair of transfer functions $(k(z), l(z))$ is rational iff it has a state space and an ARMAX representation. Therefore for ARMAX and state space

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

representations of a process $(y_t)_{t\in\mathbb{N}}$ we obtain an additive decomposition into a part depending only on the pair of transfer functions $(k(z), l(z))$ represented via $(a(z), b(z), d(z))$ or (A, B, C, D, K), respectively, and effects of the initial conditions. Therefore if the effects of the initial conditions are modeled properly the representations of $(k(z), l(z))$ via ARMAX representations, or via state space representations are equivalent.

Next we will discuss some properties of the state space representation that are needed in the paper. The discussion will be short and will draw heavily from Chapters 1 and 2 of Hannan and Deistler (1988). The main purpose of this exposition is to make connections between equivalent concepts for ARMAX and state space representations.

For ARMAX representations redundant complexity is ruled out by assuming $(a(z), b(z), d(z))$ left-coprime (for a definition see Hannan and Deistler, 1988, p. 40). For state space systems redundancy is eliminated by the concept of minimality. Recall the definition $\pi(A, B, C, D, K) = (I_s + zC(I_n - zA)^{-1}K, D + zC(I_n - zA)^{-1}B) = (k(z), l(z))$. A state space system (A, B, C, D, K) is called minimal if $\pi(A, B, C, D, K) = \pi(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{K}) \Rightarrow \dim(A) \leq \dim(\tilde{A})$ where $\dim(A)$ for a square matrix A denotes the number of rows. Therefore minimality refers to the property of using as few state components as needed to describe the pair of transfer functions, irrespective of the effects of the initial values. The minimal state dimension, say, n_0, is called *order of the system*.

Two representations (A, B, C, D, K) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{K})$ of a pair of transfer functions $(k(z), l(z))$ are called *observationally equivalent*. Two minimal representations (A, B, C, D, K) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{K})$ of order n_0 are observationally equivalent if and only if there exists a nonsingular matrix $\mathcal{T} \in \mathbb{R}^{n_0\times n_0}$ such that $\tilde{A} = \mathcal{T}A\mathcal{T}^{-1}, [\tilde{B}, \tilde{K}] = \mathcal{T}[B, K], \tilde{C} = C\mathcal{T}^{-1}, \tilde{D} = D$, and $\tilde{x}_t = \mathcal{T}x_t$ (Hannan and Deistler, 1988, Theorem 2.3.4). Hence the two observationally equivalent minimal representations differ only in choosing a different basis for the state.

For ARMAX systems it is well known that the location of the roots of $\det a(z)$ for a left-coprime representation determine the (asymptotic) stationarity properties of the solution processes. The roots of $\det a(z)$ in a left-coprime ARMAX representation are the inverse of the eigenvalues other than zero of A in a minimal state space representation of the same pair of transfer functions (cf. Hannan and Deistler, 1988, Theorem 1.2.2). Therefore if $\lambda_{max}(A) < 1$ (where $\lambda_{max}(A)$ denotes the maximum modulus of all eigenvalues of A) then there exists an initial state x_1 leading to a stationary solution of the state space system equations (2) (conditional on stationarity of $(z_t)_{t\in\mathbb{N}}$). An analogous relation exists between the zeros of $\det b(z)$ and $\lambda_{max}(A - KC)$. Thus the so-called strict minimum-phase assumption $\lambda_{max}(A - KC) < 1$ corresponds to the assumption of stable invertibility of $k(z)$.

APPENDIX B: PARAMETERIZATION BASED ON ECHELON OVERLAPPING FORMS

These parameterizations are based on a factorization of the so-called Hankel matrix

$$\mathcal{H} := \begin{bmatrix} C[B,K] & CA[B,K] & CA^2[B,K] & \dots \\ CA[B,K] & CA^2[B,K] & & \\ CA^2[B,K] & & & \\ \vdots & & & \end{bmatrix} = \mathcal{O}\mathcal{C},$$

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press

where $\mathcal{O} := [C', A'C', (A^2)'C', \ldots]'$ denotes the *observability matrix* and $\mathcal{C} := [[B,K], A[B,K], A^2[B,K], \ldots]$. Different factorizations of $\mathcal{H}$ into $\mathcal{O}$ and $\mathcal{C}$ correspond to observationally equivalent representations. The rank of the Hankel matrix is equal to the system order (Hannan and Deistler, 1988, Theorem 2.4.1). Let α denote a multi-index indicating n (where n denotes the order of the system) rows $h_i, i = 1, \ldots, n$ of $\mathcal{H}$ such that every row of $\mathcal{H}$ is a linear combination of $h_i, i = 1, \ldots, n$. Let $h(i,j)$ denote the row of $\mathcal{H}$ with index $(i-1)s + j$. Because of the block Hankel structure it is no loss of generality to choose only rows in the first sn rows of $\mathcal{H}$ (see Hannan and Deistler, 1988, Theorem 2.4.1, p. 51). This follows because for $\tilde{a}(z) := \det(zI_n - A) = z^n - \sum_{j=0}^{n-1} a_j z^j$ it holds that $\tilde{a}(A) = 0$. Consequently $A^n = \sum_{j=0}^{n-1} a_j A^j$. Furthermore without loss of generality we can restrict attention to selections of rows such that if $h(i,j) \in \{h_r, r = 1, \ldots, n\}$ then also $h(l,j) \in \{h_r, r = 1, \ldots, n\}, l = 1, \ldots, i-1$ (see Hannan and Deistler, 1988, Theorem 2.4.1(i), p. 51). Selections of rows with this property are called *nice selections* and can be described using the multi-index $\alpha = (n_1, \ldots, n_s)$. Here n_i denotes the largest index such that $h(n_i, i)$ is contained in the nice selection if $n_i > 0$ and else no row for the ith coordinate is included. The order of α is denoted as $|\alpha| = \sum_{j=1}^{s} n_s = n$. Define $\mathcal{H}_\alpha := [h(1,1)', h(2,1)', \ldots, h(n_1,1)', h(1,2)', \ldots, h(n_s,s)']'$ (where no rows occur for coordinates i such that $n_i = 0$) and $\mathcal{H}_{\alpha+s} := [h(2,1)', h(3,1)', \ldots, h(n_1+1,1)', h(2,2)', \ldots, h(n_s+1,s)']'$ (assuming that $n_1 \geq 2, n_2 \geq 1, n_s \geq 1$ for ease of presentation). Then a unique state space representation of $(k(z), l(z))$ is obtained from the equations $A\mathcal{H}_\alpha = \mathcal{H}_{\alpha+s}$, $C\mathcal{H}_\alpha = [I_s, 0]\mathcal{H}$, and $[B,K] = \mathcal{H}_\alpha[I_{s+m}, 0]'$ if $\mathcal{H}_\alpha$ is of full row rank (cf. Hannan and Deistler, 1988, Theorem 2.6.2, p. 69). It follows that $[B,K]$ contains free parameters and A and C are of a particular structure given in Hannan and Deistler (1988, Theorem 2.6.2). The matrix A is partitioned according to $A = [A_{ij}]_{i,j=1,\ldots,s}$, $A_{ij} \in \mathbb{R}^{n_i \times n_j}$ where

$$A_{ii} = \begin{bmatrix} 0 & & \\ \vdots & I_{n_i-1} & \\ -a_{ii}(0) & \ldots & -a_{ii}(n_i-1) \end{bmatrix},$$

$$A_{ij} = \begin{bmatrix} 0 & & \\ \vdots & & \\ -a_{ij}(0) & \ldots & -a_{ij}(n_j-1) \end{bmatrix}, \qquad j \neq i$$

and for $n_i > 0, i = 1, \ldots, s$

$$C = \left[\begin{array}{cc|c|c} 1 & 0 & 0 & 0 \\ & \ldots & \ldots & \ldots \\ 0 & & 0 & 1 \quad 0 \end{array}\right].$$

If $n_i = 0$ for some i then the ith row of C also contains unrestricted entries (for details see Hannan and Deistler, 1988, Theorem 2.6.2, p. 69). Using this unique representation a parameterization for $M_{n,\alpha}$ can be constructed (Hannan and Deistler, 1988, Theorem 2.6.5). Let $M_{n,\alpha}$ denote the set of all pairs $(k(z), l(z)) \in M_n$ such that the corresponding matrix $\mathcal{H}_\alpha$ is of full rank. The parameter vector $\theta \in \mathbb{R}^{2sn+sm+mn}$, say, collects the entries $a_{ij}(l), 0 \leq l \leq n_j - 1$, the entries in $[B,K]$ and D, and entries in the rows of C for all i such that $n_i = 0$. Let φ_α denote the mapping attaching the thus constructed parameter vector θ to the pair of transfer functions $(k(z), l(z)) \in M_{n,\alpha}$ and let Θ_α denote the corresponding parameter space, i.e., $\Theta_\alpha = \varphi_\alpha(M_{n,\alpha})$. Theorems 2.6.4 and 2.6.5 of Hannan and Deistler (1988) imply that Θ_α is an open subset of $\mathbb{R}^{2sn+sm+nm}$ and that φ_α is continuous on $M_{n,\alpha}$.

https://doi.org/10.1017/S0266466605050127 Published online by Cambridge University Press