Estimating cointegrated systems using subspace algorithms

Dietmar Bauer *
Inst. for Econometrics,
Operations Research and System Theory
TU Wien
Argentinierstr. 8, A-1040 Wien

Martin Wagner †
Department of Economics
and Finance
Institute for Advanced Studies
Stumpergasse 56, A-1060 Wien

January 17, 2000

Abstract

In this paper the properties of so called subspace methods in the context of cointegrated processes of order one are investigated. It is shown that the algorithms deliver consistent estimates of the transfer function in the case of general VARMA models and mild conditions on the underlying noise process. A procedure for the estimation of the dimension of the cointegrating space is presented and consistency for this procedure is proven. Also the estimation of the order of the system is discussed. Simulation examples demonstrate the usefulness of the subspace algorithms for the estimation of cointegrated systems.

JEL Classification: C13, C32
Keywords: Cointegration, subspace algorithms, state space representation

*Support by the Austrian FWF under the project number P-11213-MAT is gratefully acknowledged. Part of this work has been done while this author was on leave at the University of Newcastle, Australia.

†Correspondence to (from February 2000 onwards): Department of Economics, University of Berne, Gesellschaftsstrasse 49, CH-3012 Berne, Switzerland. Tel.: ++41 +31 631477, Fax: ++41 +31 6313992, email: martin.wagner@vwi.unibe.ch.

1 Introduction

This paper presents a new method for estimation and testing in (co)integrated processes of order 1. Compared to the most widely used Johansen (1988, 1991, 1995) procedure our approach allows estimation and testing for cointegration for more general integrated processes of order 1 by including, in terms of an ARMA representation of the underlying system, an MA part. We however exclude unit roots at other points than 1, so e.g. seasonal unit roots and seasonal cointegration are not handled by our method so far.
Cointegration analysis has become one of the most widely used techniques in econometrics. The majority of analyses is carried out by using the methods and procedures that have been developed by Johansen and his co-authors. This method, despite its advantages like elaborate possibilities to test for a variety of hypotheses on the cointegrating space and also on the short-run dynamics, has one limitation. It is restricted to the case where the data generating process is a pure autoregression. Although this assumption may be a good approximation in many cases, the possibility of more general data generating processes deserves some attention. There are already a couple of results available in the literature dealing with this issue. Yap and Reinsel (1995) derive the maximum likelihood estimator for the cointegrating space of cointegrated Gaussian ARMA processes integrated of order 1. They also show that the asymptotic null distribution of the test statistic is the same as for pure autoregressions. Saikkonen (1992) and Saikkonen and Lütkepohl (1996) derive consistency of Johansen type estimators for the case, when the data generating process is given by an infinite order vector autoregression, but one approximates this by a finite order VAR. More precisely, it is shown that the Johansen procedure delivers consistent estimates of the cointegrating space, if the order of the VAR approximation is increasing with the sample size at a sufficient rate. This is a generalization of the Said and Dickey (1984) result to the multivariate case. Also in the Saikkonen and Lütkepohl (1996) case the asymptotic null distribution is the same as in the case of a pure (finite order) autoregression. Wagner (1999a) shows that the Johansen procedure delivers consistent estimates of the cointegrating space, when one estimates a VAR with a fixed lag order, but the data are generated by a vector ARMA system. In that case however the short-run dynamics are not estimated consistently anymore.
Another method that is related in some sense to our approach in terms of its applicability is the non-parametric cointegration analysis developed by

Bierens (1995, 1997). This procedure derives consistent estimates of the cointegrating space and a test for its dimension on the basis of a non-parametric approximation of ARMA systems integrated of order 1. Due to the non-parametric nature of this method one obtains estimates of the cointegrating space only and does not obtain estimates of e.g. short-run coefficients. In terms of results however our method is more comparable to the already above mentioned method of Yap and Reinsel (1995), because our method derives a consistent estimator of the transfer function of the system as well.[1] Given an estimate of the transfer function it is then possible to derive e.g. an ARMA representation of the system.

Subspace algorithms have up to now mainly been used in a stationary context, with the exception of the work of Aoki (1990). However, the Aoki approach has never been given a thorough statistical foundation including the issues of estimating the integer parameters like the order of the system and the dimension of the cointegrating space. Subspace methods have been developed in the engineering literature in the last couple of years, see e.g. Larimore (1983), Van Overschee and DeMoor (1994) or Verhaegen (1994). The asymptotic properties of the estimates obtained by these procedures in a stationary setting are established in a number of papers: Deistler et al. (1995) and Peternell (1995) treat the consistency of the methods, Viberg et al. (1993) derive asymptotic normality of the estimated poles of the system for one class of methods usually denoted by `MOESP` type of methods, Bauer (1998) and Bauer et al. (1998) establish a central limit theorem for the estimates of the system as well as consistent order estimation algorithms. For stationary stochastic processes the subspace estimates have the usual limiting behavior, i.e. consistency and asymptotic normality. Up to now no optimality or sub-optimality results for the asymptotic covariance matrices of the subspace estimators have been derived. Also a consistent estimate of the system order may be obtained in a simple fashion (see Peternell 1995, Bauer 1998). Approximation properties of the transfer function estimates are known (Bauer 1998). In this paper the consistency of the estimates of the transfer function for a special class of algorithms due to Larimore (1983) is derived also for processes integrated of order one. Furthermore estimation procedures for the number of unit roots and therefore for the dimension of the cointegrating space are

[1] E.g. for an ARMA system $a(L)y_t = b(L)\varepsilon_t$, where the matrices a and b are left co-prime, the transfer function is given by $k(z) = a^{-1}(z)b(z)$, where L denotes the lag operator and z a variable in the complex plane.

provided. As indicated above, the analysis is restricted to the case, where the unit roots are located at $z = 1$ and where the highest geometric multiplicity (in the state space representation, see below) of the unit roots is equal to one, thus excluding e.g. processes with seasonal unit roots or $I(2)$ processes. The analysis uses similar techniques as have been used in Huang and Guo (1990), Lütkepohl and Saikkonen (1997) and Saikkonen and Luukkonen (1997).

The organization of the paper is as follows: In the next section the procedure is described and the theoretical results are stated. In Section 3 the estimation of the cointegrating rank and of the system order are discussed and in Section 4 results of a simulation study to assess the performance of our method are presented. In this section we also compare the performance of our method to the performance of the Johansen method. Section 5 summarizes and concludes. In Appendix A all proofs are given and in Appendix B the simulated systems are described.

2 The subspace method

In this paper we consider finite dimensional, time invariant, discrete time, state space systems of the form

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

where y_t denotes the s-dimensional observed series, which is observed for $t = 0, 1, \ldots, T$. ε_t denotes an s-dimensional white noise sequence. Throughout the paper ε_t is assumed to be an ergodic strictly stationary martingale difference sequence for which the following equations hold:

$$\mathbb{E}\{\varepsilon_t | \mathcal{F}_{t-1}\} = 0 \tag{2}$$

$$\mathbb{E}\{\varepsilon_t \varepsilon_t' | \mathcal{F}_{t-1}\} = \mathbb{E}\{\varepsilon_t \varepsilon_t'\} \tag{3}$$

$$\mathbb{E}\{\varepsilon_{t,a}\varepsilon_{t,b}\varepsilon_{t,c} | \mathcal{F}_{t-1}\} = \omega_{a,b,c} \tag{4}$$

$$\mathbb{E}\varepsilon_{t,a}^4 < \infty \tag{5}$$

where $\varepsilon_{t,a}$ denotes the a-th component of the vector ε_t and $\mathcal{F}_{t-1}$ denotes the σ algebra spanned by the past $\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots, \varepsilon_0, x_0$. These conditions will be referred to as the *standard assumptions* throughout the paper. The matrix E is assumed to be nonsingular and lower triangular with positive entries on the diagonal. This restriction is necessary to ensure the identifiability of

E and ε_t. Furthermore it is assumed, that the system is strictly *minimum-phase*, i.e. that the eigenvalues of the matrix $(A - KE^{-1}C)$ have an absolute value smaller than 1. Corresponding to the eigenvalues of the matrix A, i.e. the *system poles*, we assume, that they are inside the open unit disc or at $z = 1$, where the geometric multiplicities of the eigenvalues at $z = 1$ are restricted to be equal to one. This corresponds to the assumption of the order of integration to be equal to one. A companion paper Bauer and Wagner (1999a) develops a canonical form for state space systems of the form (1) containing unit roots. The results derived in that paper are used extensively in the following discussion of the properties of the system estimates. Note that in the special case treated in the present paper the canonical form has the following structure:

$$A = \begin{bmatrix} I_r & 0 \\ 0 & A_{st} \end{bmatrix}, K = \begin{bmatrix} K_1 \\ K_{st} \end{bmatrix}, C = \begin{bmatrix} C_1 & C_{st} \end{bmatrix}$$

Here $0 \leq r \leq s$ denotes the number of common trends, $I_r \in \mathbb{R}^{r \times r}$ denotes the identity matrix and (A_{st}, K_{st}, C_{st}) denotes a state space realization of the stationary subsystem. From the structure of the state space representation it follows that

$$y_t = C_1 K_1 \sum_{j=1}^{t-1} \varepsilon_t + k_{st}(L)\varepsilon_t \tag{6}$$

where $k_{st}(L) = E + LC_{st}(I - LA_{st})^{-1}K_{st}$. This representation makes clear, that the system depends only on the product C_1K_1 and not on the two factors C_1 and K_1 directly, in the sense that (C_1, K_1) and $(C_1T, T^{-1}K_1)$ result in the same system for any nonsingular matrix T of compatible dimensions. The same is true for (A_{st}, K_{st}, C_{st}) and $(SA_{st}S^{-1}, SK_{st}, C_{st}S^{-1})$ for nonsingular S. In other words for a given system only certain products are identified, but not the system matrices themselves. Therefore additional restrictions are introduced in order to achieve identifiability, leading to a canonical form. In the canonical form presented in Bauer and Wagner (1999a) C_1 is chosen to be part of an orthonormal matrix, i.e. $C_1 \in \mathbb{R}^{s \times r}, C_1'C_1 = I_r$. Therefore there exists a matrix C_2 with $C_2'C_2 = I_{s-r}$ and $C_2'C_1 = 0$, i.e. C_2 is in the orthogonal complement of C_1. Let $\bar{C}' = [C_1, C_2]$. Since all the eigenvalues of A_{st} are by construction restricted to be inside the unit circle it is easily seen, that $k_{st}(z)$ is analytic in the closed unit disc. Note that the representation given in equation (6) coincides with Granger's. It is immediate that the first component in (6) corresponds to the common trends and that the columns of

C_2 span the space of the cointegrating relations. Therefore the cointegrating rank is equal to $s - r$ and the number of common trends is equal to the number of eigenvalues of A at one. In the case of processes of higher order of integration matters get more complicated, but also then it is the structure of the eigenvalues at 1 (i.e. their algebraic and geometric multiplicities) of the corresponding matrix A, that determines the order of integration and the number of components of the process with different orders of integration; i.e. the number of common trends with different orders of integration (for a detailed discussion see Bauer and Wagner, 1999a).

Subspace algorithms originated in the engineering literature in the 1980ies. They provide an alternative to classical maximum likelihood estimation of linear time invariant systems, like e.g. ARMA systems. In the meantime a variety of algorithms is available, see e.g. Larimore (1983), Van Overschee and DeMoor (1994), Verhaegen (1994). In this paper we restrict attention to the algorithm described in Larimore (1983), which is well suited for the analysis of multivariate time series, where no exogenous observed variables are present.
The main idea of this algorithm lies in the interpretation of the state: Consider the problem of predicting $y_{t+j}, j \geq 0$ from its finite past up to time $t-1$, i.e. from $y_{t-1}, y_{t-2}, \cdots, y_0$ and x_0.[2] From the system equations (1) it follows, that

$$y_{t+j} = CA^j x_t + \sum_{i=0}^{j-1} CA^i K \varepsilon_{t+j-i-1} + \varepsilon_{t+j}.$$

Now, since

$$\begin{aligned} x_t &= A^t x_0 + \textstyle\sum_{i=0}^{t-1} A^i K \varepsilon_{t-i-1} \\ &= A^t x_0 + \textstyle\sum_{i=0}^{t-1} A^i E^{-1} K (y_{t-i-1} - C x_{t-i-1}) \\ &= (A - KE^{-1}C)^t x_0 + \textstyle\sum_{i=0}^{t-1} (A - KE^{-1}C)^i K E^{-1} y_{t-i-1} \end{aligned}$$

one obtains $y(t+j|t) = CA^j x_t$, where $y(t+j|t)$ denotes the best linear predictor of y_{t+j} from the knowledge of $y_{t-1}, \cdots, y_0, x_0$. Thus the state x_t is a basis for the predictor space and is contained in the past of the time series.

Next define for given positive integers f and p

$$Y_{t,f}^+ = [y_t', y_{t+1}', \ldots, y_{t+f-1}']'$$

[2] In the case, that x_0 is not known, the prediction is performed using the Kalman filter. This, however, does not change the asymptotic properties.

$$Y_{t,p}^- = [y'_{t-1}, y'_{t-2}, \ldots, y'_{t-p}]'$$

and

$$E_{t,f}^+ = [\varepsilon'_t, \varepsilon'_{t+1}, \ldots, \varepsilon'_{t+f-1}]'$$

Furthermore let

$$\mathcal{O}_f = [C', A'C', \ldots, (A^{f-1})'C']'$$

$$\mathcal{K}_p = [K, (A - KE^{-1}C)K, \ldots, (A - KE^{-1}C)^{p-1}K]$$

and let $\mathcal{E}_f$ denote the matrix, whose i-th block row is equal to the matrix $[CA^{i-1}K, \cdots, CK, E, 0]$.[3] Then it follows from the system equations, that

$$Y_{t,f}^+ = \mathcal{O}_f \mathcal{K}_p Y_{t,p}^- + \mathcal{O}_f (A - KE^{-1}C)^p x_{t-p} + \mathcal{E}_f E_{t,f}^+ \tag{7}$$

Here for notational simplicity $y_t = 0, t < 0, x_t = 0, t \leq 0$.[4] Now the subspace algorithm can be described as follows:

1) In a first step regress $Y_{t,f}^+$ on $Y_{t,p}^-$ to obtain an estimate $\hat{\beta}_{f,p}$ of $\mathcal{O}_f \mathcal{K}_p$.

2) Typically $\hat{\beta}_{f,p}$ will be of full rank, whereas $\mathcal{O}_f \mathcal{K}_p$ is of rank n for $f, p \geq n$. Thus approximate $\hat{\beta}_{f,p}$ by a rank n matrix with decomposition $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$.

3) Use the estimate $\hat{\mathcal{K}}_p$ to estimate the state as $\hat{x}_t = \hat{\mathcal{K}}_p Y_{t,p}^-$. Once the state has been estimated, the system equations can be used to obtain estimates of the system matrices (A, K, C, E) by ordinary least squares: First regress y_t on $\hat{x}_t$ to obtain an estimate $\hat{C}_T$ and residuals $\tilde{\varepsilon}_t$. Then $\hat{\Omega} = \frac{1}{T} \sum_{t=1}^T \tilde{\varepsilon}_t \tilde{\varepsilon}'_t$ is an estimate for the innovation variance. Thus $\hat{E}_T$ can be calculated as the lower triangular Cholesky factor of $\hat{\Omega}$ and $\hat{\varepsilon}_t = \hat{E}_T^{-1} \tilde{\varepsilon}_t$. Finally regress $\hat{x}_{t+1}$ on $\hat{x}_t$ and $\hat{\varepsilon}_t$ to obtain estimates $\hat{A}_T$ and $\hat{K}_T$ respectively.

The approximation in step 2 of this procedure is performed by using the singular value decomposition of $\hat{W}_f^+ \hat{\beta}_{f,p} \hat{W}_p^-$. Here $\hat{W}_f^+$ and $\hat{W}_p^-$ are weighting matrices, which in this paper are restricted to be $\hat{W}_f^+ = (\hat{\Gamma}_f^+)^{-1/2}$ and $\hat{W}_p^- =$

[3] Note e.g. that the matrix $\mathcal{O}_f$ is a truncation of the *observability* matrix $\mathcal{O} = [C', A'C', (A^2)'C', \ldots]$.

[4] The asymptotic results also hold for a nonzero initial state vector.

$(\hat{\Gamma_p^-})^{1/2}$ respectively.[5] Here $\hat{\Gamma_f^+} = \sum_{t=1}^{T} Y_{t,f}^+ (Y_{t,f}^+)'$ and $\hat{\Gamma_p^-} = \sum_{t=1}^{T} Y_{t,p}^- (Y_{t,p}^-)'$, where unobserved values are replaced with zeros, such that $\hat{\Gamma_f^+}$ and $\hat{\Gamma_p^-}$ have a block Toeplitz structure.[6] This amounts to estimating the canonical correlations of $Y_{t,f}^+$ and $Y_{t,p}^-$. This explains the name *canonical correlation analysis (CCA)* for this algorithm.
In the literature several different weighting schemes have been proposed and analyzed in the stationary case. Up to now it has not been analyzed under which properties of the weighting matrices consistent estimates in the case of unit root processes are obtained. Therefore the weighting matrices are restricted as mentioned above. Thus let $\hat{W_f^+}\hat{\beta}_{f,p}\hat{W_p^-} = \hat{U}\hat{\Sigma}\hat{V}' = \hat{U}_n\hat{\Sigma}_n\hat{V}_n' + \hat{R}$, where $\hat{U}_n \in \mathbb{R}^{fs\times n}, \hat{V}_n \in \mathbb{R}^{ps\times n}, \hat{\Sigma}_n \in \mathbb{R}^{n\times n}$. $\hat{\Sigma}_n$ contains the n dominant singular values ordered decreasing in size. $\hat{U}_n$ contains the corresponding left singular vectors and $\hat{V}_n$ the respective right singular vectors. The remaining singular values contribute to $\hat{R}$ and are neglected. Now the rank n approximation to $\hat{\beta}_{f,p}$ is given by $\hat{\mathcal{O}}_f\hat{\mathcal{K}}_p = [(\hat{W_f^+})^{-1}\hat{U}_n\hat{\Sigma}_n][\hat{V}_n'(\hat{W_p^-})^{-1}]$ and thus $\hat{\mathcal{K}}_p = \hat{V}_n'(\hat{W_p^-})^{-1}$.

In this step of the algorithm usually the order of the system is estimated, see e.g. Bauer (1998). In the stationary case one possible order estimation procedure is obtained by considering the size of the first neglected singular value. Define the following criterion:

$$SVC(n) = \hat{\sigma}_{n+1}^2 + 2nsC_T/T \tag{8}$$

Here $C_T > 0, C_T/T \to 0$ denotes a penalty term, which determines the asymptotic properties of the estimated order. $2ns$ is the number of parameters in a model with state dimension n, excluding the parameters in E, see e.g. Hannan and Deistler (1988), Theorem 2.6.3. The estimated order is the minimizing argument of the criterion function $SVC(n)$.
Let $U(n)$ denote the set of all transfer functions $k \in M(n)$, such that the a.s. limit for $T \to \infty$ and $p = p(T) \to \infty$ (which exists under the assumptions of Theorem 1 below) $W_f^+\beta W^-$ of $\hat{W_f^+}\hat{\beta}_{f,p}\hat{W_p^-}$ has n distinct nonzero singular values. It can be shown, see e.g. Bauer et al. (1999) that $U(n)$ is a generic

[5]Here $X^{1/2}$ denotes any square root of the positive definite matrix X such that $X^{1/2}(X^{1/2})' = X$. Note that the choice of the square root is of no importance for the estimation. Different choices lead to numerically identical estimates of the system matrices. However we will use the Cholesky factor since it proves to be convenient in the derivations.

[6]This corresponds to a special choice of the initial values, which however does not influence the asymptotic properties under investigation.

subset of $M(n)$. Finally the estimate $\hat{x}_t = \hat{\mathcal{K}}_p Y_{t,p}^-$ is used to obtain estimates of the system matrices. Then the following result, which clarifies the asymptotic properties in the stationary case, has been shown in Bauer et al. (1999) and Bauer (1998)

Theorem 1 *Let y_t be generated by a system of the form (1), where the white noise ε_t fulfills the standard assumptions. If $f \geq n$ is a user supplied integer and $p(T) \geq -\frac{d}{2}\frac{\log T}{\log|\rho_0|}$, where ρ_0 is an eigenvalue of $A - KE^{-1}C$ of maximum modulus and $d > 1$ is some real value, and if $p(T) = o((\log T)^a)$ for some $a > 0$, then:*

- *for $k_0 \in M(n)$ the estimate of the transfer function is almost sure consistent, i.e. $\hat{k} \to k_0$ a.s.*
- *for $k_0 \in U(n)$ the estimate of the system matrices is a.s. consistent, i.e. there exists a realization (A_0, K_0, C_0, E_0) of the true transfer function $k_0 \in U(n)$, such that $\|vec[\hat{A}_T - A_0, \hat{K}_T - K_0, \hat{C}_T - C_0, \hat{E}_T - E_0]\| \to 0$ a.s. Here vec denotes the operator stacking the vectorizations of the various matrices.*
- *for $k_0 \in U(n)$ a central limit theorem for the system matrix estimates holds, i.e.*

$$\sqrt{T}[vec(\hat{A}_T - A_0, \hat{K}_T - K_0, \hat{C}_T - C_0, \hat{E}_T - E_0)] \xrightarrow{d} Z$$

  where $\xrightarrow{d}$ denotes convergence in distribution and Z is a Gaussian random variable with zero mean and variance V.

- *if $C_T/(p(T)\log\log T) \to \infty$ then the order estimated using SVC is a.s. consistent.*

This clarifies the asymptotic properties in the stationary case to a large extent. One question that still remains to be answered is whether the asymptotic variance covariance matrices of subspace estimators achieve the Cramer Rao lower bound. Up to now it is only known that for a couple of cases the asymptotic variance of the estimator described above is at least close to optimality.

To the best of the authors knowledge the nonstationary case has not been discussed in the literature so far. The description of the algorithm does not

include any assumption concerning stationarity, thus one might hope to apply the algorithm in a straightforward way also for the case of integrated processes. It is the aim of this paper to show that from a theoretical point of view this leads to a reasonable procedure. For the actual implementation however, some steps might have to be adapted in order to avoid numerical problems. This is mainly due to the different orders of magnitude of the stationary and the nonstationary part of the data.
The algorithm as it has been presented above does not necessitate any information concerning the dimension of the cointegrating space. If this information is known somehow (one way to estimate the cointegrating rank is discussed below), the procedure could be adapted in the following way: Note that in the final step, after $\hat{x}_t, \hat{\varepsilon}_t$ and $\hat{C}, \hat{E}$ have been estimated, $\hat{x}_{t+1}$ is regressed on $\hat{x}_t, \hat{\varepsilon}_t$ in order to obtain estimates $\hat{A}, \hat{K}$. This is based on the observation that $x_{t+1} = Ax_t + K\varepsilon_t$ according to the system structure (1). Since r eigenvalues of A are equal to one, the matrix $A - I$ is of rank $n - r$. Thus if the number of common trends r is known, a different way to estimate $\hat{A}, \hat{K}$ would be to consider a reduced rank regression of $x_{t+1} - x_t = \tilde{A}x_t + K\varepsilon_t$ under the constraint $\mathrm{rank}(\tilde{A}) = n - r$. This leads to an estimate $\hat{A}$, which corresponds to an exactly cointegrated system, whereas the unrestricted regression approach only leads to an estimated transfer function, which is close to a cointegrated system in the sense given by the next theorem, which constitutes the main result of this paper. The proof is given in Appendix A.

Theorem 2 *Let y_t be generated by a system of the form (1), where the ergodic noise ε_t fulfills the standard assumptions. Assume, that the order n of the true transfer function k_0 is known, and that $p = p(T) = o((\log T)^a)$ for some $0 < a < \infty$, $f \geq n$ fixed. Furthermore assume, that $diag((I - z), I)\bar{C}k_0(z)$ lies in the generic neighborhood of the echelon canonical form. Then the estimate $\hat{E}_T + \hat{C}_T(zI - \hat{A}_T)^{-1}\hat{K}_T$ converges in probability to the true transfer function, if the unrestricted regression approach is used.*

If in addition the multiplicity r of the unit root is known, then the same result holds, if the reduced rank regression is used to obtain estimates $\hat{A}_T$ and $\hat{K}_T$.

The cointegrating space (which is equal to the orthogonal complement of the column span of C_1, the first r columns of C) is estimated at rate T, i.e. $T^\alpha[\hat{C}_{T,1} - C_1] \to 0$ in probability for $0 < \alpha < 1$.

It is remarkable that the estimates obtained by using the subspace procedures, are consistent without using any prior knowledge on the cointegration

structure of the system, for any integer $0 \leq r \leq s$, i.e. independent of whether the true system is stationary, cointegrated or integrated without cointegrating relationships. This is a similarity to autoregression, which also results in consistent estimates of the transfer function regardless if the true system contains a unit root or not. Note also, that this is a difference to maximum likelihood estimation, where the structure of the unit roots usually is built in explicitly in the parametrisation and thus in order to achieve consistency for all unit root configurations many different parameter sets have to be considered. Also note, that when the cointegrating rank is known indeed, this knowledge can be used to obtain an estimate, which is in the desired model set, i.e. which has the corresponding cointegrating rank. The natural next question corresponds to the asymptotic distribution of the estimates. Note at this point that the result states consistency for the transfer function estimates, whereas the actual system matrix estimates $(\hat{A}_T, \hat{K}_T, \hat{C}_T, \hat{E}_T)$ need not converge. This is in particular true for an implementation based on the stationarity assumption, which estimates a state having a finite covariance matrix, by choosing the estimate $\hat{\mathcal{K}}_p$ such that $1/T \sum_{t=1}^{T} \hat{x}_t \hat{x}_t'$ is convergent. Thus the estimated state $\hat{x}_t$ is not consistent for the true state x_t, but rather the nonstationary directions are downweighted by a factor $T^{-1/2}$. The implication of this is that the estimate of the matrix C_1 has to compensate the downweighting of the state estimate and tends to infinity at the rate $T^{1/2}$ and thus does not converge. Proper rescaling of the estimates however leads to an implementation of the subspace methods, which lead to consistent estimates of the system descriptions in generic cases, as can be seen from an inspection of the proof.

3 Estimating the structure indices

By structure indices we denote the number of common trends r (or equivalently the dimension of the cointegrating space $s - r$) and the order of the system n. For the calculation of the estimates no knowledge of r is required. However, if r and n were known, then the reduced rank regression could be performed to obtain an estimate of A, which takes into account the specific cointegration structure. In this section it will be demonstrated, how the integer r can be estimated using subspace methods. The central fact in this respect is the observation, that the singular values, which have to be calculated in the algorithm, provide easily accessible information for the as-

sessment of the cointegrating rank. For the stationary case, Theorem 1 states consistency for the particular estimates of the order obtained by using SVC with a special penalty term C_T. In this section we also clarify the properties of the order estimation techniques in the unit root case.
Let the process y_t be generated by a system of the form (1). Then exactly r singular values are equal to 1, the remaining $n-r$ are smaller than one. It is important to note, that this fact is only true, if no zeros on the unit circle are admitted[7] and the unit roots are restricted to lie at $z=1$. Any of the other cases also introduces singular values equal to one. Thus all results in this section are not robust against the presence of e.g. seasonal unit roots.
It is shown in the proof of Theorem 2 that the singular values are estimated consistently. Moreover it has been derived, that the first r singular values converge to one at rate T, whereas the remaining $n-r$ nonzero singular values tend to their limits at rate $T^{1/2}$. Therefore a procedure for estimating the number of common trends can be obtained from the asymptotic distribution of the estimates of the first r singular values, which is derived in the following theorem:

Theorem 3 *Let the process y_t be generated by a system of the form (1), where the true noise satisfies the conditions of Theorem 1. Let $\hat{\sigma}_i$ denote the estimate of the i-th singular value and let r denote the true number of common trends. Then $T(1-1/r\sum_{j=1}^{r}\hat{\sigma}_j^2)$ is (asymptotically) distributed as*

$$\frac{1}{r}tr[C_1'\Omega C_1(\int_0^1 W(w)W(w)'dw)^{-1}] \tag{9}$$

Here $\int_0^1 W(w)W(w)'dw$ denotes a mixture of Brownian motions, where the covariance associated with $W(w)$ is equal to K_1K_1'. $\Omega = EE'$ denotes the innovation covariance matrix.

This theorem provides an asymptotically valid test on the number of common trends in the time series. One could use subspace algorithms to estimate the system in a first step. The resultant estimates could then be used to approximate the distribution given above. This in turn leads to a test for the number of common trends. However, the distribution of the test statistic depends on unknown quantities and also the finite sample approximation seems to be unsatisfactory. Therefore we propose to estimate the

[7] Thus we exclude, in terms of an ARMA representation unit roots in the MA polynomial.

dimension of the cointegrating space rather than to test for it. An estimation algorithm is easily derived from the above arguments. Since the singular values corresponding to the nonstationary part converge to one at rate T, a simple idea is to take the number of common trends to be equal to the largest integer, r say, such that the r-th singular value is the smallest one to differ from 1 by less than $h(T)/T$, where $h(T) \to \infty$ as $T \to \infty$. This leads to a consistent estimation of r. Note however, that the choice of a specific form of the penalty $h(T)$ includes an element of arbitrariness and different thresholds may be considered, as is done in the next section. Of course the choice of $h(T)$ influences the finite sample properties of the estimation procedure.

Finally also the estimation of the system order can be handled in this framework. The estimation algorithm builds on the estimation procedures proposed in Peternell (1995) and Bauer (1998) for the stationary case. In Theorem 1 the consistency for the order estimation procedure in the stationary case has been stated. The next theorem ensures, that consistency carries over to the nonstationary situation as well:

Theorem 4 *Under the conditions of Theorem 2 the estimate of the order obtained by SVC is weakly consistent, i.e.* $\hat{n} \to n$ *in probability.*

The proof of this theorem is also given in Appendix A.

4 A simulation study

In this section the theoretical results obtained in the last sections are tested on simulated data. The performance of our procedure is analyzed with regard to two aspects. The performance of the estimation of the system, and the performance of the estimation of the cointegrating rank. Concerning the first aspect we are especially interested in the quality of the approximation of the true cointegrating space by the estimated cointegrating space. One measure of quality employed in this paper is the Hausdorff distance, which is defined as follows:
Let Ξ and Ψ be two subspaces of $\mathbb{R}^m$. The intersection of a subspace Θ of $\mathbb{R}^m$ with the closed unit circle in $\mathbb{R}^m$ is denoted by $C(\Theta)$, i.e.

$$C(\Theta) = \{z \in \Theta \mid\mid \| z \| \leq 1\},$$

where $\| z \|$ is the Euclidean norm of z. Using this notation the distance d of Ξ and Ψ is given by the Hausdorff distance d_H of $C(\Xi)$ and $C(\Psi)$, i.e.

$$d(\Xi, \Psi) = d_H(C(\Xi), C(\Psi)) = max(\rho(C(\Xi), C(\Psi)), \rho(C(\Xi), C(\Psi)))$$

where $\rho(C_1, C_2)$is given by

$$\rho(C_1, C_2) = \sup_{x \in C_1} \inf_{y \in C_2} \| x - y \| .$$

Let us start the analysis with a set of systems, which has already been used in Saikkonen and Luukkonen (1997). A precise description of the systems can be found in Appendix B. All three systems generate 3-dimensional outputs. The three scenarios include the cases of a 2-dimensional cointegrating space (Scheme 1), of a 1-dimensional cointegrating space (Scheme 2) and of an integrated system without cointegration (Scheme 3). For each system 1000 time series of length $T = 150$ and $T = 1050$ respectively have been generated using Gaussian white noise with the covariance matrix as specified in Appendix B. The first 50 observations are discarded in order to simulate a nonzero initial state vector. For each time series the cointegrating rank and the order of the system are estimated. In the algorithm the integers $f = p = 2\hat{p}_{AIC}$ are used, where $\hat{p}_{AIC}$ denotes the order estimate obtained by using AIC. The order estimation criterion is $SVC(n)$ as described in the previous section. The number of common trends is estimated as the number of estimated singular values, which differ from 1 by more than $\log(T)^2/T$. The true order is equal to $n = 3$ in all three cases and the true cointegrating ranks are $2, 1$ and 0 respectively. Concerning the estimation of the number of cointegrating relationships and the system order we obtain the results shown in Table 1. For the case of a 2-dimensional cointegrating space the results are best: For sample size $T = 1000$ the correct configuration of the structure indices is estimated in almost each replication. For sample size $T = 100$ the number of cointegrating vectors is already estimated quite accurately, whereas the order of the system is estimated only with a low degree of accuracy. For Scheme 2 the results are also quite good, again for both the large sample size $T = 1000$ and also for the small sample size $T = 100$. For Scheme 3 the performance is not as satisfactory as for the other two systems. It has become clear from the theory presented above that the asymptotic distribution of the estimated singular values depends on the true system. Therefore it cannot be expected, that the simple estimation criterion used

Scheme	Sample Size	Dim. of coint. space				System order			
		0	1	2	3	0	1	2	3
1	$T = 100$	0	0.317	0.670	0.013	0.193	0.153	0.177	0.477
	$T = 1000$	0	0	0.990	0.010	0	0	0	1
2	$T = 100$	0	0.863	0.137	0	0.003	0.008	0.365	0.624
	$T = 1000$	0	0.985	0.015	0	0	0	0	1
3	$T = 100$	0.323	0.658	0.019	0	0	0.004	0.010	0.986
	$T = 1000$	0.747	0.253	0.001	0	0	0	0	1

Table 1: Distributions of the estimated dimension of the cointegrating space and the estimated system order for Schemes 1,2 and 3 and sample sizes $T = 100$ and $T = 1000$ respectively.

Scheme	1		2		3	
Sample size	$T = 100$	$T = 1000$	$T = 100$	$T = 1000$	$T = 100$	$T = 1000$
	0.767	0.772	0.314	0.494	0.006	0.029

Table 2: Percentage of a correct estimation of the dimension of the cointegrating space when the threshold $log(T)/T$ is used.

here shows good performance for all situations. For the first two examples it happens to be the case, that $\log(T)^2$ has the same magnitude as the 95% percentile of the asymptotic expression. By bootstrapping this percentile can be estimated to be approximately 34 for Scheme 1 with cointegrating rank $r = 2$ and approximately 35 for Scheme 2 having cointegrating rank $r = 1$. A comparison of these numbers with $\log(100)^2 = 21.20$ and $\log(1000)^2 = 47.7$ explains the performance in these cases. Table 2 shows the frequency of a correct estimate of the dimension of the cointegrating space using the threshold $log(T)/T$. It is still observed, that the estimation accuracy increases with increasing sample size, although $log(T)/T$ is by no means a good choice as a threshold for the examples at hand. The simulations indicate, that the estimation algorithms perform reasonably well for the case of a high dimensional cointegrating space, whereas in the presence of many common trends the performance deteriorates.

Corresponding to the order estimation procedure it is remarkable that for $T = 1000$ in all cases the correct system order is detected. This demonstrates, that the order estimation procedure works satisfactory for large sample sizes, which can also clearly be seen from a plot of the estimated singular values. Figure 1 shows two examples of such plots. It can clearly be seen, that for

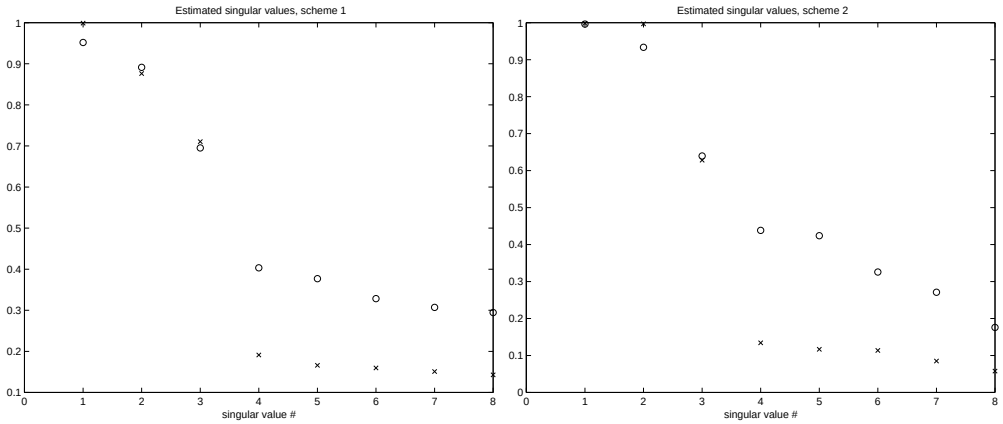

Figure 1: Estimated singular values for scheme 1 (left plot) and scheme 2 (right plot) for one example and sample size $T = 100$ (o) and $T = 1000$ (x) respectively.

sample size $T = 1000$ the gap between the third and the fourth singular value is very pronounced, which is reflected by the order estimates. The graphical information presented in Figure 1 gives, especially for $T = 1000$, a clear indication about the number of singular values equal to 1, and therefore about the dimension of the cointegrating spaces. Note however that the two plots for sample size $T = 100$ are quite similar, indicating the difficulty of estimating the cointegration rank for this sample size. Additional information can be gained from a plot of the eigenvalues of the estimated matrix A. For Scheme 1 this can be seen in Figure 2, which plots the eigenvalues of the estimated matrices A for sample size $T = 100$ (left plot) and $T = 1000$ (right plot). The three eigenvalues are ordered in size and the largest eigenvalue is indicated with '+', the second largest with 'o' and the smallest with 'x'. It can be clearly seen that the only possible unit root is located at $z = 1$. It can also be seen, that the information at sample size $T = 100$ is quite ambiguous, whereas for sample size $T = 1000$ the plot clearly shows the unit root at $z = 1$. Note that the true eigenvalues are at $z = 1, z = 0.8$ and $z = 0.7$ respectively.

Finally also the estimation of the cointegrating space is investigated. Table 3 summarizes the results: For both systems with cointegration the mean of the Hausdorff distance between the estimated and the true cointegrating spaces is decreasing with sample size, which reflects the consistency of the estimates. Again it can be seen that the subspace procedure performs satis-

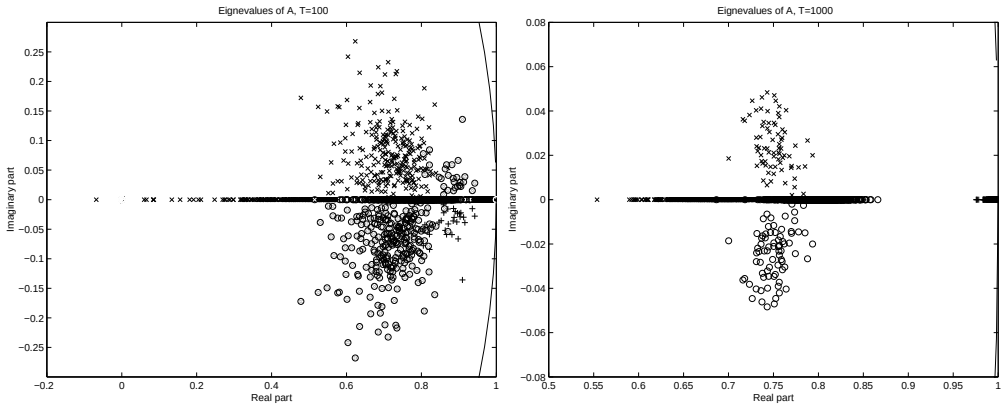

Figure 2: Estimated eigenvalues for Scheme 1 and sample sizes $T = 100$ (left plot) and sample size $T = 1000$ (right plot).

Sample size	$T = 100$	$T = 1000$
Scheme 1	0.070	0.006
Scheme 2	0.196	0.015

Table 3: Mean of the Hausdorff distances between the estimated and the true cointegrating space for the first two schemes and the two sample sizes $T = 100$ and $T = 1000$ respectively.

factorily especially in cases with few common trends. Problems seem to arise for small sample sizes and low dimensional cointegrating spaces.

As already noted in the introduction, in the cointegration literature the Johansen procedure is the by far most widely used method, therefore we also compare our method with this method on the simulated data. The results obtained by the Johansen method are documented in Table 4. The order of the AR model is chosen using AIC in each trial. The results show, that for the first two systems the subspace procedure results in more reliable estimates of the dimension of the cointegrating space. The dominance of the Johansen procedure is pronounced in the case of no cointegration.

Let us now analyze the results for eight 2-dimensional ARMA(2,1) systems given in equation (18) in Appendix B. These systems have already been analyzed in Wagner (1999a) where it is shown that the Johansen procedure derives consistent estimates of the cointegrating space also for ARMA sys-

Sample size	$T = 100$				$T = 1000$			
Coint. dim.	0	1	2	3	0	1	2	3
Sch. 1	0.187	0.484	0.294	0.035	0	0	0.93	0.07
Sch. 2	0.394	0.575	0.029	0.002	0	0.936	0.063	0.001
Sch. 3	0.962	0.035	0.003	0	0.942	0.035	0.003	0

Table 4: Frequencies of estimated dimensions of the cointegrating space for Schemes 1 to 3 and sample sizes $T = 100$ and $T = 1000$ using the Johansen method.

tems, given that the moving average polynomial is regular at $z = 1$. In that paper the systems have been used to investigate the finite sample implications of the theoretical robustness results. To assess the effect of the moving average polynomials, the simulations are performed over a set of ARMA systems where only the moving average polynomial is changed.

Here we use these systems to compare the performance of the subspace procedure with that of the Johansen procedure, to see whether our method produces possibly more reliable results in cases where the VAR based Johansen procedure operates on misspecified models. For the Johansen procedure it has been shown that only the estimates of the cointegrating space are consistent under the assumed form of misspecification, whereas of course our approach derives consistent estimates of the whole system. The simulations are performed for four different sample sizes $T = 100, 200, 300$ and 400.[8] Again the orders of the autoregressive approximation of the ARMA systems are selected according to AIC, the orders are given in Appendix B in Table 6 for these systems. The number of replications is 1000. The threshold $h(T)$ for the subspace procedure in these simulations is chosen to be $\log(T)/T$. In Figure 3 the acceptance frequencies for a 1-dimensional cointegrating space are shown for both procedures.[9]

The systems all have the same 1-dimensional cointegrating space. It can be seen that the nominal size is closer to the actual size for the Johansen procedure than for the subspace procedure only for systems MA1 and MA2. For systems MA3 to MA8 the performance of the two procedures is quite identical, with the subspace method showing better results already for the

[8]Time series with length 450 are generated, the first 50 observations are dropped, and then the first 100, 200 and so on observations are used to compute the estimates and test statistics.

[9]The labels MA1 to MA8 indicate the different systems.

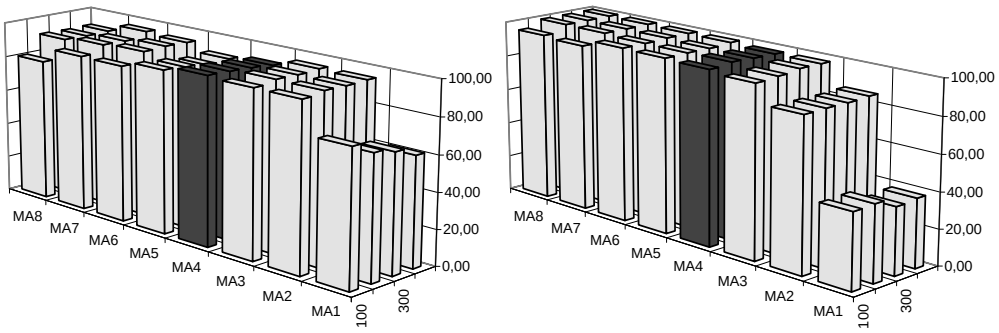

Figure 3: Acceptance frequencies for the correct number of cointegrating vectors for the 2-dimensional ARMA systems for all sample sizes. The left picture corresponds to the Johansen procedure (using the trace test), the right picture corresponds to the subspace procedure.

smaller sample sizes for these systems. The system MA4, the results for which are drawn in black in Figure 3, is a pure autoregressive system. Also for that system the performance of the subspace procedure is comparable to the results for the Johansen procedure. So, at least for these 2-dimensional systems the estimation of the dimension of the cointegrating space as described in the last section turns out to work reasonably well.
In Figure 4 the quality of the approximation of the true cointegrating space by the estimated cointegrating space is displayed. The measure of quality is the mean of the log of the Hausdorff distances between the estimated and true cointegrating space over all replications.[10]

The logarithm is taken to increase the sample variability of the observations, since prior to this transformation all observations are very close to 0.[11] Looking at the figure it can be seen that the performance is first of all very good for both methods, for all sample sizes and all moving average polynomials, and that the performance of both procedures is very similar.[12] In Figure 4 the two dotted lines around the solid line corresponding to the Johansen procedure are given by $\pm 2\sqrt{\frac{\hat{\sigma}^2_{Joh}}{1000} + \frac{\hat{\sigma}^2_{sub}}{1000}}$, with $\hat{\sigma}^2_{Joh}$ and $\hat{\sigma}^2_{sub}$ denoting the estimated variances of the distributions of the log Hausdorff distances for

[10] In Bauer and Wagner (1999b) more detailed results of simulation studies, including e.g. the empirical densities of the log Hausdorff distances, are presented.

[11] A Hausdorff distance of 0 means that the spaces are identical.

[12] Note the scale of the graphs, with a range between -7.3 and -6.6 for $T = 100$ and between -8.50 and -8.06 for $T = 400$.

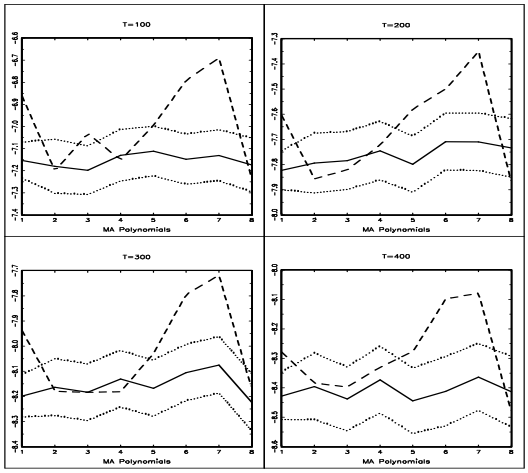

Figure 4: Means of the empirical densities of the log of the Hausdorff distances between the true and estimated cointegrating spaces for the 2-dimensional ARMA systems. The densities are computed over all replications where in each replication the correct number of cointegrating vectors is taken. The dashed line corresponds to the subspace procedure, the solid line to Johansen's procedure. The dotted lines are the $\pm 2\sqrt{\frac{\hat{\sigma}^2_{Joh}}{1000} + \frac{\hat{\sigma}^2_{sub}}{1000}}$ lines around the mean corresponding to the Johansen procedure.

the Johansen and the subspace procedure respectively. These lines are graphically indicating the acceptance and critical region for a test of equality of the means of the two distributions, for all eight systems. The null hypothesis of an equal mean is only rejected for systems MA1, MA6 and MA7. Thus for the 2-dimensional systems the subspace procedure is almost perfectly comparable to the Johansen procedure, with the advantage of obtaining consistent estimates of all system parameters.

The results of the simulation study indicate that the estimated cointegrating spaces obtained by application of the subspace method are of high quality, comparable to the ones obtained by the Johansen method. It is however required to gain further understanding of the properties of the methods for systems that are not so well described by low order autoregressions. At this point it may be worthwhile to mention that Wagner (1999b) compares the Johansen and the Bierens procedure amongst other systems on the same 2-dimensional systems. It is found that for these systems the performance of the Bierens procedure is much worse than for the Johansen or the subspace procedure. A more detailed account of the results of simulations of the various methods and also of applications to interest rate data is given in Bauer and Wagner (1999b). Especially the choice of the penalty term in the estimation of the cointegrating rank deserves some further investigations.

5 Summary and conclusions

This paper establishes consistency for subspace methods in the context of cointegrated time series. Also methods for estimating the dimension of the cointegrating space and the system order at the same time are developed and analyzed. The significance of these results lies in the fact, that the algorithm provides consistent estimates in the case of general ARMA systems and is thus not limited to AR processes such as the Johansen approach. The computationally cheap subspace estimates can e.g. also be used as consistent initial values to obtain efficient estimates of the parameters performing one Newton step for maximum likelihood estimation.
The simulation evidence shows results for the subspace procedure that are mostly comparable to the results for the Johansen method. However, further understanding concerning the choice of an optimal penalty in deciding about the number of singular values equal to one, which equals the number of common trends, has to be gathered. One advantage of the algorithm is

that it provides useful information on the structure of the cointegration and on the order of the system, which is easily accessible via the estimated singular values. Note that also the estimated eigenvalues of the matrix A can be used to decide on the order of cointegration. Thus the user obtains a variety of easily accessible information on the dimension of the cointegrating space. In Bauer and Wagner (1999b) it is demonstrated that the subspace procedure gives sensible results also on real world, in that particular case, interest rate data. Thus by application of subspace methods on cointegrated processes one may be able to gain additional insights in the properties of observed possibly cointegrated time series. Especially the applicability of our method for general integrated processes of order 1 allows for at least a "cross-validation" of the results obtained with more standard tools like e.g. the Johansen method.
Further research is concentrated on three important questions not dealt with in this contribution. One is the treatment of deterministic components, like constants and trends. The second is the derivation of test (statistics) of hypotheses on the cointegrating space, which is closely linked to the derivation of the asymptotic distribution of the estimates of the cointegrating space. The third research field finally lies in the exploration of the applicability of the subspace algorithms for processes arbitrary unit roots, i.e. processes with seasonal unit roots as well as processes integrated of higher orders.

A Proofs

Theorem 2 *Let y_t be generated by a system of the form (1), where the ergodic noise ε_t fulfills the standard assumptions. Assume, that the order n of the true transfer function k_0 is known, and that $p = p(T) = o((\log T)^a)$ for some $0 < a < \infty$, $f \geq n$ fixed. Furthermore assume, that $diag((I-z), I)\bar{C}k_0(z)$ lies in the generic neighborhood of the echelon canonical form. Then the estimate $\hat{E}_T + \hat{C}_T(zI - \hat{A}_T)^{-1}\hat{K}_T$ converges in probability to the true transfer function, if the unrestricted regression approach is used.*

If in addition the multiplicity r of the unit root is known, then the same result holds, if the reduced rank regression is used to obtain estimates $\hat{A}_T$ and $\hat{K}_T$.

The cointegrating space (which is equal to the orthogonal complement of the column span of C_1, the first r columns of C) is estimated at rate T, i.e. $T^\alpha[\hat{C}_{T,1} - C_1] \to 0$ in probability for $0 < \alpha < 1$.

Proof: The arguments developed below for integrated processes follow the lines of Shin and Lee (1997), Lütkepohl and Saikkonen (1997) and Saikkonen and

Luukkonen (1997). The key argument is the definition of transformations of $Y^+_{t,f}$ and $Y^-_{t,p}$ defined in the main part of the paper, which separate the stationary and nonstationary components of these random variables.
From the Granger representation theorem for cointegrated processes (of order 1) it follows that $y_t = C_1 K_1 \sum_{j=1}^{t} \varepsilon_{t-j} + k_{st}(L)\varepsilon_t$, where $k_{st}(z)$ denotes the stable part of transfer function and where $C_1 \in \mathbb{R}^{s\times r}, K_1 \in \mathbb{R}^{r\times s}, C_1'C_1 = I$. In this representation C_1 is not unique. Bauer and Wagner (1999a) show, how a unique choice for C_1 can be obtained. Note, however, that the cointegrating space does not depend on the choice of C_1. If $C_2 \in \mathbb{R}^{s\times(s-r)}$, where r is equal to the rank of C_1, is such that $C_2'C_2 = I, C_2'C_1 = 0$, then, with $\bar{C}$ as defined in Section 1, in $\bar{C}y_t$ the first r components are equal to $\sum_{j=1}^{t} K_1\varepsilon_{t-j} + z_t$, where $z_t = C_1'k_{st}(L)\varepsilon_t$ is stationary. The remaining $s-r$ components are stationary. Thus the dimension of the cointegrating space is equal to $s-r$. Then there exists a selector matrix $S_f \in \mathbb{R}^{fs\times fs}$ such that in $\tilde{Z}^+_{t,f} = S_f[I \otimes \bar{C}]Y^+_{t,f}$ the first fr rows correspond to the nonstationary part of the time series, whereas in the remaining $f(s-r)$ rows the stationary factors appear. Thus $\tilde{Z}^+_{t,f}$ is of the form

$$\tilde{Z}^+_{t,f} = \begin{bmatrix} I \\ \vdots \\ I \\ 0 \\ \vdots \\ 0 \end{bmatrix} (\sum_{j=0}^{t-2} K_1\varepsilon_j) + \begin{bmatrix} K_1\varepsilon_{t-1} \\ \vdots \\ K_1(\sum_{j=t-1}^{t+f-2}\varepsilon_j) \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} C_1'k_{st}(L)\varepsilon_t \\ \vdots \\ C_1'k_{st}(L)\varepsilon_{t+f-1} \\ C_2'k_{st}(L)\varepsilon_t \\ \vdots \\ C_2'k_{st}(L)\varepsilon_{t+f-1} \end{bmatrix}$$

Define $Q_f \in \mathbb{R}^{fr\times fr}$ to be the block Toeplitz matrix, whose (i,i) block is equal to I_r, the $r\times r$ identity matrix and the $(i+1,i)$ block equal to $-I_r$, the remaining blocks being zero. Then $Z^+_{t,f} = \mathrm{diag}(Q_f, I_{f(s-r)})\tilde{Z}^+_{t,f}$, so it can be represented as

$$Z^+_{t,f} = \begin{bmatrix} I \\ 0 \\ \vdots \\ \\ \\ \\ \vdots \\ 0 \end{bmatrix} (\sum_{j=0}^{t-2} K_1\varepsilon_j) + \begin{bmatrix} K_1\varepsilon_{t-1} \\ \\ \vdots \\ K_1\varepsilon_{t-2+f} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} C_1'k_{st}(L)\varepsilon_t \\ C_1'k_{st}(L)\Delta\varepsilon_{t+1} \\ \vdots \\ C_1'k_{st}(L)\Delta\varepsilon_{t+f-1} \\ C_2'k_{st}(L)\varepsilon_t \\ \vdots \\ C_2'k_{st}(L)\varepsilon_{t+f-1} \end{bmatrix}$$

Here $\Delta = 1 - L$ denotes the first difference operator. Analogously the vector $Y^+_{t,f}$ can be transformed to $Z^-_{t,p} = Q_p S_p Y^-_{t,p}$, which is given by

$$Z^-_{t,p} = \begin{bmatrix} I \\ 0 \\ \vdots \\ \\ \\ \vdots \\ 0 \end{bmatrix} (\sum_{j=0}^{t-2} K_1 \varepsilon_j) + \begin{bmatrix} 0 \\ K_1\varepsilon_{t-2} \\ \vdots \\ K_1\varepsilon_{t-p} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} C_1' k_{st}(L)\varepsilon_{t-1} \\ C_1' k_{st}(L)\Delta\varepsilon_{t-1} \\ \vdots \\ C_1' k_{st}(L)\Delta\varepsilon_{t-p+1} \\ C_2' k_{st}(L)\varepsilon_{t-1} \\ \vdots \\ C_2' k_{st}(L)\varepsilon_{t-p} \end{bmatrix}$$

Let $D_T = \mathrm{diag}(T^{-1}I_r, T^{-1/2}I_{fs-r})$, where T denotes sample size. From the construction of $Z^+_{t,f}$ and $Z^-_{t,p}$ it follows that only the first r components are nonstationary, whereas the remaining components are stationary. Hence the normalization. Let $\langle a_t, b_t \rangle = \sum_{t=1}^T a_t b_t'$. Then $D_T \langle Z^+_{t,f}, Z^+_{t,f} \rangle D_T'$ converges in distribution to Γ^+_f, where

$$\Gamma^+_f = \begin{bmatrix} K_1 \int_0^1 W(r)W(r)'dr K_1' & 0 \\ 0 & \tilde{\Gamma}^+_f \end{bmatrix}$$

Here $\int_0^1 W(r)W(r)'dr$ denotes the stochastic integral of the Brownian motion $W(r)$, which is the limit of $\frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} v_t$, where v_t is i.i.d. with unit variance.[13] $\tilde{\Gamma}^+_f$ denotes the covariance matrix of the stationary process composed of the last $fs-r$ components of $Z^+_{t,f}$. The convergence of these matrices is ensured by the conditions on the noise and the asymptotic results for nonstationary processes stated e.g. in Phillips and Solo (1992), Davidson (1994) or Johansen (1995). The off-diagonal entries converge to zero, since they consist of sums of products of nonstationary and stationary processes, normalized by $T^{-3/2}$. In order to simplify the analysis of the asymptotic distribution let $\Phi_T = \frac{1}{T^2} \sum_t (\sum_{j=0}^{t-2} K_1 \varepsilon_j)(\sum_{j=0}^{t-2} K_1 \varepsilon_j)'$. Clearly Φ_T is the (appropriately scaled) dominant term in the nonstationary component of both $Z^+_{t,f}$ and $Z^-_{t,p}$. Further let $\tilde{D}_T = \mathrm{diag}(\Phi_T^{-1/2} T^{-1}, T^{-1/2} I)$.[14] Then $\tilde{D}_T \langle Z^+_{t,f}, Z^+_{t,f} \rangle \tilde{D}_T$ converges to $\mathrm{diag}(I, \tilde{\Gamma}^+_f)$ in probability. Thus consider the difference

$$\tilde{D}_T \langle Z^+_{t,f}, Z^+_{t,f} \rangle \tilde{D}_T - \begin{bmatrix} I & 0 \\ 0 & \tilde{\Gamma}^+_f \end{bmatrix}$$

[13] $\lfloor x \rfloor$ denotes the smallest integer equal or larger than x.

[14] In order to simplify notation, the symbol $\tilde{D}_T$ will be used for any matrix of the form $\mathrm{diag}(\Phi_T^{-1/2} T^{-1}, T^{-1/2} I)$, irrespective of the dimension of the second block.

Denoting $n_t = \sum_{j=0}^{t-2} K_1 \varepsilon_j$ we obtain for the $(1,1)$ block of this expression

$$\Phi_T^{-1/2} T^{-2} \left[\sum_{t=0}^{T} (n_t + v_t)(n_t + v_t)' \right] (\Phi^{-1/2})' - \Phi_T^{-1/2} T^{-2} \sum_{t=0}^{T} n_t n_t' (\Phi^{-1/2})'$$

Here v_t stands for all stationary contributions. Thus we obtain $T^{-2} \sum_{t=0}^{T} \Phi_T^{-1/2} (n_t v_t' + v_t n_t' + v_t v_t')(\Phi_T^{-1/2})'$. This matrix converges, when multiplied by T, in distribution to a random variable, since $T^{-1} \sum_{t=0}^{T} n_t v_t'$ converges in distribution, see e.g. Theorem B.13 in Johansen (1995). The $(2,1)$ (and the $(1,2)$ block, which is the transpose thereof) are of the form $T^{-3/2} \sum_{t=0}^{T} \Phi_T^{-1/2} n_t v_t'$. Here v_t again stands for a stationary variable (not the same as before, though). It follows, that $T^{1/2}$ times this expression converges in distribution. Finally the $(2,2)$ term is the sample covariance of a stationary process and thus the error converges in distribution, when multiplied by $T^{1/2}$. Taking the Cholesky factor as the square root of a matrix, we obtain that $\tilde{D}_T \langle Z_{t,f}^+, Z_{t,f}^+ \rangle^{1/2}$ converges in probability to $\mathrm{diag}(I, (\tilde{\Gamma}_f^+)^{1/2})$, and again the blocks are of the same order of convergence.

The same arguments apply to $\tilde{D}_T \langle Z_{t,p}^-, Z_{t,p}^- \rangle \tilde{D}_T' \to \Gamma_p^-$, where Γ_p has a block structure analogous to Γ_f^+. Since Γ_p^- and Γ_f^+ are nonsingular a.s. the same result for the inverses of these matrices follows from the continuous mapping theorem. The remaining term, which has to be considered, consists of $\tilde{D}_T \langle Z_{t,f}^+, Z_{t,p}^- \rangle \tilde{D}_T'$. Completely the same arguments as for the other terms show, that $\tilde{D}_T \langle Z_{t,f}^+, Z_{t,p}^- \rangle \tilde{D}_T' \to \mathcal{H}_{f,p}$, which is given by

$$\mathcal{H}_{f,p} = \begin{bmatrix} I & 0 \\ 0 & \tilde{\mathcal{H}}_{f,p} \end{bmatrix}$$

Here again $\tilde{\mathcal{H}}_{f,p}$ corresponds to the stationary part. Note, that the matrix, on which the singular value decomposition is performed in the subspace algorithm is equal to

$$\langle Y_{t,f}^+, Y_{t,f}^+ \rangle^{-1/2} \langle Y_{t,f}^+, Y_{t,p}^- \rangle (\langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1/2})'$$

The left singular vectors of this matrix are equal to the eigenvectors of the matrix

$$\begin{aligned} \hat{X}_T &= \langle Y_{t,f}^+, Y_{t,f}^+ \rangle^{-1/2} \langle Y_{t,f}^+, Y_{t,p}^- \rangle \langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1} \langle Y_{t,p}^-, Y_{t,f}^+ \rangle (\langle Y_{t,f}^+, Y_{t,f}^+ \rangle^{-1/2})' \\ &= \langle Z_{t,f}^+, Z_{t,f}^+ \rangle^{-1/2} \langle Z_{t,f}^+, Z_{t,p}^- \rangle \langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1} \langle Z_{t,p}^-, Z_{t,f}^+ \rangle (\langle Z_{t,f}^+, Z_{t,f}^+ \rangle^{-1/2})' \end{aligned}$$

where the second expression can be analyzed more easily due to the fact, that in $Z_{t,p}^-$ the increase of $p(T)$ as a function of the sample size T only occurs in the rows corresponding to the stationary part and thus can be handled using the methods developed for the stationary case in e.g. Bauer et al. (1999). From the results

given above, it follows that this matrix converges to

$$(\Gamma_f^+)^{-1/2}\mathcal{H}_{f,\infty}(\Gamma_\infty^-)^{-1}\mathcal{H}'_{f,\infty}((\Gamma_f^+)^{-1/2})' = \begin{bmatrix} I & 0 \\ 0 & (\tilde{\Gamma}_f^+)^{-1/2}\tilde{\mathcal{H}}_{f,\infty}(\tilde{\Gamma}_\infty^-)^{-1}\tilde{\mathcal{H}}'_{f,\infty}((\tilde{\Gamma}_f^+)^{-1/2})' \end{bmatrix} \tag{10}$$

due to the continuous mapping theorem. Convergence is in distribution and since the matrix is deterministic also in probability. Also the rate of convergence can be investigated using the facts derived above. Let X_0 denote the limiting expression. Then $\hat{X}_T - X_0 =$

$$\begin{aligned}
&= (\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{1/2})^{-1}\langle Z_{t,f}^+, Z_{t,p}^-\rangle(\langle Z_{t,p}^-, Z_{t,p}^-\rangle)^{-1}\langle Z_{t,p}^-, Z_{t,f}^+\rangle(((\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{1/2})^{-1})' \\
&\quad - \begin{bmatrix} I & 0 \\ 0 & (\tilde{\Gamma}_f^+)^{-1/2}\tilde{\mathcal{H}}_{f,\infty}(\tilde{\Gamma}_\infty^-)^{-1}\tilde{\mathcal{H}}'_{f,\infty}((\tilde{\Gamma}_f^+)^{-1/2})' \end{bmatrix} \\
&\doteq \left\{(\tilde{D}_T\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{1/2})^{-1} - \begin{bmatrix} I & 0 \\ 0 & (\tilde{\Gamma}_f^+)^{-1/2} \end{bmatrix}\right\}\begin{bmatrix} I & 0 \\ 0 & \tilde{\mathcal{H}}_{f,\infty}(\tilde{\Gamma}_\infty^-)^{-1}\tilde{\mathcal{H}}'_{f,\infty}((\tilde{\Gamma}_f^+)^{-1/2})' \end{bmatrix} + \\
&\quad \begin{bmatrix} I & 0 \\ 0 & (\tilde{\Gamma}_f^+)^{-1/2} \end{bmatrix}\left\{\tilde{D}_T\langle Z_{t,f}^+, Z_{t,p}^-\rangle(\langle Z_{t,p}^-, Z_{t,p}^-\rangle)^{-1}\langle Z_{t,p}^-, Z_{t,f}^+\rangle\tilde{D}_T - \right. \\
&\quad \left.\begin{bmatrix} I & 0 \\ 0 & \tilde{\mathcal{H}}_{f,\infty}(\tilde{\Gamma}_\infty^-)^{-1}\tilde{\mathcal{H}}'_{f,\infty} \end{bmatrix}\right\}\begin{bmatrix} I & 0 \\ 0 & (\tilde{\Gamma}_f^+)^{-1/2} \end{bmatrix} + \\
&\quad \begin{bmatrix} I & 0 \\ 0 & (\tilde{\Gamma}_f^+)^{-1/2}\tilde{\mathcal{H}}_{f,\infty}(\tilde{\Gamma}_\infty^-)^{-1}\tilde{\mathcal{H}}'_{f,\infty} \end{bmatrix}\left\{(\tilde{D}_T\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{1/2})^{-1} - \begin{bmatrix} I & 0 \\ 0 & (\tilde{\Gamma}_f^+)^{-1/2} \end{bmatrix}\right\}
\end{aligned}$$

Here $\doteq$ has to be understood in the following sense: In the $(1,1)$ sub-block $\doteq$ denotes equality up to terms of order $O_P(T^{-1})$, in the remaining blocks $\doteq$ stands for equality up to terms of order $O_P(T^{-1/2})$. This follows from a repeated application of the same reasoning as has been given above for $\tilde{D}_T\langle Z_{t,f}^+, Z_{t,f}^+\rangle\tilde{D}_T$. One technical complication lies in the fact, that $p = p(T)$ has to tend to infinity at a certain rate of the sample size (see Theorem 1) in order to ensure consistency. However the increase of dimensions only concerns the stationary part of the process and thus can be treated with the same tools as used in the stationary case.

In subspace algorithms an eigenvalue decomposition is performed on $\hat{X}_T$. For the limit X_0 the first r eigenvalues are equal to 1, the corresponding eigenvectors span the space corresponding to the first r vectors of the canonical basis. With regard to the remaining eigenvalues and -vectors note that the term $(\tilde{\Gamma}_f^+)^{-1/2}\tilde{\mathcal{H}}_{f,p}((\tilde{\Gamma}_p^-)^{-1/2})'$ corresponds to the stationary transfer function

$$\tilde{k}(z) = \begin{bmatrix} I - zI & 0 \\ 0 & I \end{bmatrix}\bar{C}k(z)$$

as can be shown from the definition of $Z^+_{t,f}$ and $Z^-_{t,p}$. The transfer function $\tilde{k}(z)$ is of order n. This can be seen by considering the non-minimal representation

$$\tilde{A} = \begin{bmatrix} I & 0 & 0 \\ 0 & A_{st} & 0 \\ I & C_1' C_{st} & 0 \end{bmatrix}, \tilde{K} = \begin{bmatrix} K_1 \\ K_{st} \\ C_1' \end{bmatrix}, \tilde{C} = \left[\begin{pmatrix} I \\ 0 \end{pmatrix}, \bar{C} C_{st}, - \begin{pmatrix} I \\ 0 \end{pmatrix} \right], \tilde{E} = \bar{C}$$

of $\tilde{k}(z)$ [15]. Here the realization (A, K, C, E) of $k(z)$ is used, where

$$A = \begin{bmatrix} I & 0 \\ 0 & A_{st} \end{bmatrix}, K = \begin{bmatrix} K_1 \\ K_{st} \end{bmatrix}, C = [C_1, C_{st}], E = I$$

From the expressions for $Z^+_{t,f}$ and $Z^-_{t,p}$ and realization theory for the stationary case it follows, that $\tilde{\mathcal{H}}_{f,\infty}((\tilde{\Gamma}^-_\infty)^{-1/2})'$ is equal to a part of the Hankel matrix of the Markov parameters corresponding to $\tilde{k}(z)$ times an orthonormal matrix, which arises because of the specific choice for the square root of Γ^-_∞. If $\tilde{k}(z) \in M(n)$ is in the generic neighborhood of the echelon canonical forms, then it follows that $\tilde{\mathcal{H}}_{f,\infty}((\tilde{\Gamma}^-_\infty)^{-1/2})'$ is of rank $n - r$, since it is, up to the orthonormal transformation, essentially the Hankel matrix, where the first r rows have been omitted. Therefore the number of nonzero singular values of the limit of $\hat{X}_T = (\langle Y^+_{t,f}, Y^+_{t,f}\rangle)^{-1/2} \langle Y^+_{t,f}, Y^-_{t,p}\rangle (\langle Y^-_{t,p}, Y^-_{t,p}\rangle^{-1/2})'$ will (generically) be equal to n, the order of the system.

From equation (10) it thus follows, that the SVD leads to a factorization

$$(\Gamma^+_f)^{-1/2} \mathcal{H}_{f,p} ((\Gamma^-_p)^{-1/2})' = \begin{bmatrix} I & 0 \\ 0 & (\tilde{\Gamma}^+_f)^{-1/2} \mathcal{O}_f \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \mathcal{K}_p (\tilde{\Gamma}^-_p)^{1/2} \end{bmatrix}$$

Here $\mathcal{O}_f$ and $\mathcal{K}_p$ correspond to the decomposition of the stationary part. From the definition of $\tilde{A}$ and $\tilde{C}$ it follows, that $\tilde{C}\tilde{A} = [0, \bar{C} C_{st} A_{st} - \begin{bmatrix} I \\ 0 \end{bmatrix} C_1' C_{st}, 0]$ and thus only the $n - r$ columns in the middle of $\tilde{\mathcal{O}}_f$ contribute to $\tilde{\mathcal{H}}_{f,p}((\tilde{\Gamma}^-_p)^{-1/2})'$. It follows from the definition of $\tilde{A}$ and $\tilde{K}$, that the middle rows of $\tilde{\mathcal{C}}$ correspond to the controllability matrix corresponding to k_{st}. Therefore $\tilde{\mathcal{C}}_\infty((\tilde{\Gamma}^-)^{-1/2})' Z^-_{t,\infty} = \tilde{\mathcal{K}}_\infty (\tilde{\Gamma}^-)^{1/2} Z^-_{t,\infty} = x_{t,st}$, the stationary part of the state. Which particular realization A_{st}, K_{st} is used, is determined by the SVD. Furthermore the convergence of the matrix $\langle Y^+_{t,f}, Y^+_{t,f}\rangle^{-1/2} \langle Y^+_{t,f}, Y^-_{t,p}\rangle \langle Y^-_{t,p}, Y^-_{t,p}\rangle^{-1} \langle Y^-_{t,p}, Y^+_{t,f}\rangle (\langle Y^+_{t,f}, Y^+_{t,f}\rangle^{-1/2})'$ implies the convergence of the eigenvalues and also the eigenspaces. Thus let $\hat{U}_n$ denote the matrix, whose columns correspond to the eigenvectors to the n dominant eigenvalues. Then the following lemma (see e.g. Chatelin, 1983) provides tools to assess the estimation error.

[15] Non-minimal means that there exists a state space representation of lower state dimension, which corresponds to the same transfer function.

Lemma 1 *Let X_0 denote a symmetric, positive definite compact linear operator and let $\hat{X}_T$ denote a sequence of symmetric, positive definite compact operators converging to X_0. Let $\lambda_1 > \cdots > \lambda_k \geq 0$ denote the k, say, distinct non-zero eigenvalues of X_0 having geometric and algebraic multiplicities equal to k_i say. Further let P_i denote the (orthogonal) projection onto the eigenspace corresponding to the eigenvalue λ_i of X_0. Furthermore let $\hat{\lambda}_{i,j}$ and $\hat{P}_i$ denote the corresponding approximating quantities calculated from $\hat{X}_T$. Then:*

- $\hat{\lambda}_{i,j} \to \lambda_i$, *i.e. the eigenvalues converge to the true eigenvalues.*
- $\hat{P}_i \to P_i$, *where convergence is in the gap metric*

Furthermore the following first order approximations hold:

$$\frac{1}{k_i}\sum_{j=1}^{k_i}\hat{\lambda}_{i,j} = \lambda_i + \frac{1}{k_i}tr[(\hat{X}_T - X_0)P_i] \tag{11}$$

$$\hat{P}_i = P_i + \sum_{\lambda_j \neq \lambda_i}\frac{1}{\lambda_i - \lambda_j}P_j[\hat{X}_T - X_0]P_i \tag{12}$$

From the lemma it follows that (for T large enough with probability one) there exists a nonsingular matrix $\hat{S}_T$, such that $\tilde{U}_n = \hat{U}_n\hat{S}_T = \begin{bmatrix} I & 0 \\ \tilde{U}_{n,1} & \tilde{U}_{n,2} \end{bmatrix}$, which converges in probability to $U_0 = \begin{bmatrix} I & 0 \\ 0 & \tilde{U}_0 \end{bmatrix}$. Here again $\tilde{U}_0$ corresponds to the stationary part. The results in (Chatelin 1983, Proposition 3.25) further show, that the entries of the matrix $\tilde{U}_n$ are analytic functions of the entries in $\hat{X}_T$. Consider the estimate

$$\begin{aligned}
\hat{x}_t &= \tilde{U}_n^T\langle Y_{t,f}^+, Y_{t,f}^+\rangle^{-1/2}\langle Y_{t,f}^+, Y_{t,p}^-\rangle\langle Y_{t,p}^-, Y_{t,p}^-\rangle^{-1}Y_{t,p}^- \\
&= \tilde{U}_n^T\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\langle Z_{t,f}^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- \\
&= \tilde{U}_n^T\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\langle (W_f^+)^{-1}U_0x_t + \mathcal{E}E_{t,f}^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^-
\end{aligned}$$

Here the limit of $\tilde{D}_T^{-1}\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}$ is denoted with W_f^+. Recall that $x_t = \tilde{\mathcal{K}}Z_{t,\infty}^- = \tilde{\mathcal{K}}_pZ_{t,p}^- + (A - KE^{-1}C)^px_{t-p}$. Therefore $\tilde{D}_T^{-1}\hat{x}_t - x_t =$

$$\begin{aligned}
&= \tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^-\rangle^{-1/2}\langle (W_f^+)^{-1}U_0x_t + \mathcal{E}E_{t,f}^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- - x_t \\
&= (\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^-\rangle^{-1/2}(W_f^+)^{-1}U_0 - I)\langle\tilde{\mathcal{K}}_pZ_{t,p}^-, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- \\
&\quad + \tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\langle\mathcal{E}_fE_{t,f}^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- - (A - KE^{-1}C)^px_{t-p} +
\end{aligned}$$

$$
\begin{aligned}
&+\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^-\rangle^{-1/2}\langle (W_f^+)^{-1}U_0(A-KE^{-1}C)^p\langle x_{t-p}, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- \\
= \;& (\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}(W_f^+)^{-1}U_0 - I)\tilde{\mathcal{K}}_p Z_{t,p}^- - (A-KE^{-1}C)^p x_{t-p} + \\
&+\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^-\rangle^{-1/2}\langle (W_f^+)^{-1}U_0(A-KE^{-1}C)^p\langle x_{t-p}, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- \\
&+\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\langle \mathcal{E}_f E_{t,f}^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^-
\end{aligned}
$$

This fact is exploited to show that a regression of the system equations can be used to obtain a consistent estimate of the transfer function. Consider therefore the regression in the observation equation:

$$
\begin{aligned}
\hat{C}_T\tilde{D}_T - C_0 &= (\sum_{t=1}^{T}(y_t - C_0\tilde{D}_T^{-1}\hat{x}_t)\hat{x}_t')(\sum_{t=1}^{T}\hat{x}_t\hat{x}_t')^{-1}\tilde{D}_T \\
&= (\sum_{t}[C_0(x_t - \tilde{D}_T^{-1}\hat{x}_t) + \varepsilon_t]\hat{x}_t')(\sum_{t}\hat{x}_t\hat{x}_t')^{-1}\tilde{D}_T
\end{aligned}
$$

It follows from the definition of $\hat{x}_t$, that $\langle \hat{x}_t, \hat{x}_t\rangle$ converges to a deterministic limit, say P, which is nonsingular. It follows from standard arguments, that $\langle \varepsilon_t, \hat{x}_t\rangle$ converges in distribution. Considering the expression given above one can show, that $\langle \tilde{D}_T^{-1}\hat{x}_t - x_t, \hat{x}_t\rangle$ converges in distribution, if $p = p(T) \geq -d\frac{\log T}{\log|\rho_0|}$. Thus we impose a stronger requirement on the increase of the integer p in order to ensure, that $\|(A-KE^{-1}C)^p\|$ tends to zero faster than T^{-1}. Therefore $\hat{C}_T\tilde{D}_T$ converges in probability to C_0 and furthermore $(\hat{C}_T\tilde{D}_T - C_0)\tilde{D}_T^{-1}$ converges in distribution, establishing the familiar convergence of order T for the complement of the cointegrating space (and thus also for the cointegrating space).

Note that $y_t - \hat{C}_T\hat{x}_t = Cx_t + \varepsilon_t - \hat{C}_T\tilde{D}_T\tilde{D}_T^{-1}\hat{x}_t = (C - \hat{C}_T\tilde{D}_T)x_t + \hat{C}_T\tilde{D}_T(x_t - \tilde{D}_T^{-1}\hat{x}_t) + \varepsilon_t$. Since $1/T\langle \varepsilon_t, \varepsilon_t\rangle \to \Omega$, where convergence is in probability, the consistency of $1/T\langle \hat{\varepsilon}_t, \hat{\varepsilon}_t\rangle$ follows from application of the arguments given above, the consistency for $\hat{C}_T\tilde{D}_T$ and the expression obtained for $\tilde{D}_T^{-1}\hat{x}_t - x_t$. Therefore also the estimates $\hat{E}_T$ are consistent.

It remains to consider the estimation of A and K. Concerning $\hat{A}_T$ note, that the normalization of $\hat{x}_t$ implies, that $\tilde{D}_T^{-1}\hat{A}_T\tilde{D}_T$ is the relevant quantity. Note that $\tilde{D}_T^{-1}A_0\tilde{D}_T = A_0$ due to the block diagonal structure of A_0. Thus consider

$$
\begin{aligned}
\tilde{D}_T^{-1}\hat{A}_T\tilde{D}_T - A_0 &= \langle \tilde{D}_T^{-1}\hat{x}_{t+1} - A_0\tilde{D}_T^{-1}\hat{x}_t, \hat{x}_t\rangle\langle \hat{x}_t, \hat{x}_t\rangle^{-1}\tilde{D}_T \\
&= \langle \tilde{D}_T^{-1}\hat{x}_{t+1} - x_{t+1}, \hat{x}_t\rangle\langle \hat{x}_t, \hat{x}_t\rangle^{-1}\tilde{D}_T + \\
&\quad +\langle A_0(x_t - \tilde{D}_T^{-1})\hat{x}_t, \hat{x}_t\rangle\langle \hat{x}_t, \hat{x}_t\rangle^{-1}\tilde{D}_T + \langle K_0\varepsilon_t, \hat{x}_t\rangle\langle \hat{x}_t, \hat{x}_t\rangle^{-1}\tilde{D}_T
\end{aligned}
$$

It follows from standard arguments, that all these terms converge to zero in probability (using the expression for $\tilde{D}_T^{-1}\hat{x}_t - x_t$ and the analogous expression for $\tilde{D}_T^{-1}\hat{x}_{t+1} - x_{t+1}$). Note that the estimate $\hat{A}_T$ will have roots strictly inside the

unit circle, as it is derived from an autoregression. However, usually the main emphasis is put on the estimation of the cointegrating vectors, such that the fact, that the eigenvalues of the estimate $\hat{A}_T$ are smaller than one might be thought of as a minor problem. Alternatively the moduli of the eigenvalues of $\hat{A}_T$ contain information about the cointegrating rank.

Finally also the consistency of $\hat{K}_T$ is shown: Note, that $\hat{\varepsilon}_t = \hat{E}^{-1}(y_t - \hat{C}_T\hat{x}_t)$ is uncorrelated with $\hat{x}_t$, since $\hat{\varepsilon}_t$ denotes the residuals of the first regression, where $\hat{x}_t$ were used as regressors. The relevant quantity in accordance with the results for $\hat{A}_T$ and $\hat{C}_T$ is equal to $\tilde{D}_T^{-1}\hat{K}_T$. Therefore consider

$$\begin{aligned}
\tilde{D}_T^{-1}\hat{K}_T &= (\sum_t \tilde{D}_T^{-1}\hat{x}_{t+1}\hat{\varepsilon}_t')(1/T\sum_t \hat{\varepsilon}_t\hat{\varepsilon}_t')^{-1} \\
&= (T^{-1}\sum_t(\tilde{D}_T^{-1}\hat{x}_{t+1} - (A - KE^{-1}C)\tilde{D}_T^{-1}\hat{x}_t\hat{\varepsilon}_t')(T^{-1}\sum_t \hat{\varepsilon}_t\hat{\varepsilon}_t')^{-1} \\
&= T^{-1}\sum_t(\tilde{D}_T^{-1}\hat{x}_{t+1} - x_{t+1})\hat{\varepsilon}_t'(T^{-1}\sum_t \hat{\varepsilon}_t\hat{\varepsilon}_t')^{-1} \\
&\quad + T^{-1}\sum_t[(A - KE^{-1}C)(x_t - \tilde{D}_T^{-1}\hat{x}_t) + K\varepsilon_t]\hat{\varepsilon}_t'(T^{-1}\sum_t \hat{\varepsilon}_t\hat{\varepsilon}_t')^{-1}
\end{aligned}$$

Tedious but straightforward calculations show, that also this expression converges to K_0 in probability. It remains to proof the result for the constrained procedure, i.e. for the case, where the true number of common trends is known. The proof of consistency follows from straightforward arguments using the consistency of the state estimation as apparent from the equation for $\tilde{D}_T^{-1}\hat{x}_t - x_t$ and the consistency of e.g. the Johansen procedure, which is in fact a reduced rank regression problem. This completes the proof.

REMARK: Note however, that the proof only shows the consistency for the transfer function estimates. The system description $(\hat{A}_T, \hat{K}_T, \hat{C}_T, \hat{E}_T)$ on the contrary will be divergent. One way to obtain also consistent estimates of the system description is to transform the estimates to a canonical form, e.g. echelon canonical form (see e.g. Hannan and Deistler, 1988), then the proof given above shows the consistency for the estimated system matrices on a generic subset. Note that the echelon canonical form can easily be transformed to an ARMA representation, if this is the preferred system representation.

Theorem 3 *Let the process y_t be generated by a system of the form (1), where the true noise satisfies the standard assumptions. Let $\hat{\sigma}_i$ denote the estimate of the i-th singular value and let r denote the true number of common trends. Then $T(1 - \frac{1}{r}\sum_{j=1}^{r}\hat{\sigma}_j^2)$ is (asymptotically) distributed as*

$$\frac{1}{r}tr[C_1^T\Omega C_1(\int_0^1 W(w)W(w)^T dw)^{-1}] \tag{13}$$

Here $\int_0^1 W(w)W(w)^T dw$ denotes a mixture of Brownian motions, where the covariance associated with $W(w)$ is equal to $K_1 K_1'$. $\Omega = EE'$ denotes the innovation covariance matrix.

Proof: The asymptotic properties of the eigenvalues (or equivalently singular values) have already been investigated in equation (11) in the proof of Theorem 2 in this appendix. Thus we have to evaluate $\mathrm{tr}[P_1(\hat{X}_T - X_0)]$, which can easily be seen to be equal to $\mathrm{tr}[\hat{X}_T^{1,1} - X_0^{1,1}]$, where the superscript ^{1,1} denotes the $(1,1)$ block of the respective quantities. Let $z_t^+ = \sum_{j=0}^{t-1} K_1 \varepsilon_j + C_1' k_{st}(L)\varepsilon_t$ denote the vector of the first r components of $Z_{t,f}^+$. Then it is straightforward to see, that the relevant quantity is equal to

$$\begin{gathered}\mathrm{tr}[I - \langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}\langle Z_{t,p}^-, z_t^+\rangle\langle z_t^+, z_t^+\rangle^{-1}\langle z_t^+, Z_{t,p}^-\rangle] = \\ -\mathrm{tr}[\langle z_t^+, z_t^+\rangle^{-1}\{\langle z_t^+, z_t^+\rangle - \langle z_t^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}\langle Z_{t,p}^-, z_t^+\rangle\}]\end{gathered}$$

Let the first r rows of $Z_{t,p}^-$ be denoted by $z_t^- = \sum_{j=0}^{t-2} K_1 \varepsilon_j + C_1' k_S(L)\varepsilon_{t-1}$. Then it follows that $z_t^+ = z_t^- + K_1\varepsilon_{t-1} + C_1' k_{st}(L)(1-L)\varepsilon_t = z_t^- + C_1^T \Delta y_t$. Denote $a_t = C_1'\Delta y_t$, then $\langle z_t^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1} Z_{t,p}^- = z_t^+ - a_t + \langle a_t, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1} Z_{t,p}^-$, which shows that we have to consider

$$\mathrm{tr}[\langle z_t^+, z_t^+\rangle^{-1}\{-\langle a_t, z_t^+\rangle + \langle a_t, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}\langle Z_{t,p}^-, z_t^+\rangle\}]$$

The essential term in the second summand is seen to be equal to $\langle a_t, z_t^+\rangle - \langle a_t, a_t\rangle + \langle a_t, Z_{t,p}^{-,st}\rangle\langle Z_{t,p}^{-,st}, Z_{t,p}^{-,st}\rangle^{-1}\langle Z_{t,p}^{-,st}, a_t\rangle$, where $Z_{t,p}^{-,st}$ denotes the stationary part of $Z_{t,p}^-$. Therefore up to first order approximation we obtain

$$T(1 - \frac{1}{r}\sum_{j=1}^{r}\hat{\sigma}_j^2) \doteq \frac{T}{r}\mathrm{tr}[\langle z_t^+, z_t^+\rangle^{-1}\{\langle a_t, a_t\rangle - \langle a_t, Z_{t,p}^{-,st}\rangle\langle Z_{t,p}^{-,st}, Z_{t,p}^{-,st}\rangle^{-1}\langle Z_{t,p}^{-,st}, a_t\rangle\}]$$

Now the result follows from the facts, that $1/T^2\langle z_t^+, z_t^+\rangle \xrightarrow{d} \int_0^1 W(w)W(w)'dw$, a_t and $Z_{t,p}^{-,st}$ are stationary and ε_t are the innovations of the process, whose components form $Z_{t,p}^{-,st}$ and on which a_t is a linear transformation. The claim then follows from the continuous mapping theorem.

Theorem 4 *Under the conditions of Theorem 2 the estimate of the order obtained by SVC is weakly consistent, i.e. $\hat{n} \to n$ in probability.*

Proof: Consider the matrices $\hat{X}_T$ and the corresponding limit X_0 defined in the proof of Theorem 2. Since the rank of $\hat{X}_0$ is equal to n and $\hat{X}_T \to X_0$ in probability, the probability that $\hat{n} < n$ tends to zero. It remains to proof

that the probability of over-estimation of the order tends to zero. Overestimation occurs, if the decrease of the criterion function $SVC(n)$ for $\hat{n} > n$ is higher than the increase due to the inclusion of more parameters, which is penalized by the amount C_T for each parameter. Thus consistency is established by showing that the maximal estimation error in the elements of the matrix, which is decomposed in the SVD, tends to zero quicker than $\sqrt{C_T/p(T)T}$ in probability. It follows from the proof of Theorem 2 and the proof in the stationary case that for each element of $\hat{X}_T - X_0$ the probability that the error is bigger than $\epsilon\sqrt{C_T/p(T)T}$ tends to zero for all $\epsilon > 0$ uniformly in the elements. Here we have used the fact that if a_t and b_t are stationary processes with rational spectral density and innovations having finite fourth moments, then there exists a constant M such that the probability that $\max_{|j|\leq H_T} \sqrt{T/\log\log T}\|\frac{1}{T}\sum_{t=1}^{T} a_t b'_{t-j} - \mathbb{E}a_t b'_{t-j}\| > M$ tends to zero for $H_T = o(\log T^a)$ for some constant $a < \infty$. Therefore the probability that the square of the Frobenius norm of this matrix is larger than $\epsilon T/\log\log T p(T)$ tends to zero for $C_T/p\log\log T \to \infty$ as it is the sum of fps^2 terms of the specified order, which proves the conjecture.

B Simulated systems

In this appendix the simulated systems are described.
The systems taken from Saikkonen and Luukkonen (1997) are the following 3-dimensional VARMA(1,1) processes:

$$\Delta y_t = \Psi y_{t-1} + \varepsilon_t - \Gamma_1 \varepsilon_{t-1} \tag{14}$$

with $y_0 = y_{-1} = 0$ and ε_t normally independently distributed $N(0, \Sigma)$. The parameter matrices are defined as follows, $\Gamma_1 = C_\gamma \mathrm{diag}(0.297, -0.202, 0) C_\gamma^{-1}$ where

$$C_\gamma = \begin{pmatrix} -0.816 & -0.657 & -0.822 \\ -0.624 & -0.785 & 0.566 \\ -0.488 & 0.475 & 0.174 \end{pmatrix} \tag{15}$$

$$\Sigma = \begin{pmatrix} 0.47 & 0.20 & 0.18 \\ 0.20 & 0.32 & 0.27 \\ 0.18 & 0.27 & 0.30 \end{pmatrix} \tag{16}$$

and $\Psi = N diag(\phi_1, \phi_2, \phi_3) N^{-1} - I_3$ with

$$N^{-1} = \begin{pmatrix} -0.29 & -0.47 & -0.57 \\ -0.01 & -0.85 & 1.00 \\ -0.75 & 1.39 & -0.55 \end{pmatrix} \tag{17}$$

Scheme	ϕ_1	ϕ_2	ϕ_3
1	1.0	0.8	0.7
2	1.0	1.0	0.7
3	1.0	1.0	1.0

Table 5: Parameter values ϕ_i for Schemes 1 to 3.

	MA1	MA2	MA3	MA4	MA5	MA6	MA7	MA8
100	2	2	2	2	2	3	4	5
200	3	2	2	2	2	4	4	4
300	3	2	2	2	2	4	4	7
400	3	2	2	2	2	4	4	7

Table 6: Selected autoregressive order of an autoregressive approximation of systems (18) for different sample sizes using AIC.

The 3 sets of parameters ϕ_i are given in Table 5.

The number of parameters ϕ_i less than unity corresponds to the number of cointegrating relationships.

The 2-dimensional ARMA(2,1) systems that have been simulated are taken from Wagner (1999a) and are given in equation (18)

$$\begin{aligned}\begin{bmatrix} 1 & -2 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} = \begin{bmatrix} 1.5 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} u_{1t-1} \\ u_{2t-1} \end{bmatrix} + \\ + \begin{bmatrix} -0.5 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_{1t-2} \\ u_{2t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} + \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{bmatrix} \begin{bmatrix} \varepsilon_{1t-1} \\ \varepsilon_{2t-1} \end{bmatrix} \end{aligned} \tag{18}$$

The parameter values γ_1, γ_2 in the MA polynomials are given by $\gamma_1 = \gamma_2 = -0.8, -0.5. - 0.2, 0, 0.2, 0.5, 0.8,$ and $\gamma_1 = 1$ and $\gamma_2 = 0.8$. The initial values are set to 0 and the ε_t are standard normally independently distributed.

In the figures and tables the systems are referred to as MA1 to MA8. The true cointegrating vector of the systems is, suitably normalized, given by $(1, -3)$. For the Johansen procedure the order of an autoregressive approximation of the systems has to be chosen. The orders are selected according to AIC[16], the chosen orders (for the different sample sizes) are given in Table 6. In Table 6 it can be

[16]The results are essentially unchanged if one selects the order according to e.g. BIC.

seen that only for the systems with large positive autocorrelation of the ε_t's large lag lengths tend to be chosen.

References

Aoki, M. (1990). *State space modeling of time series.* Springer, New York.

Bauer, D. (1998). Some Asymptotic Theory for the Estimation of Linear Systems Using Maximum Likelihood Methods or Subspace Algorithms. PhD thesis. TU Wien.

Bauer, D., M. Deistler and W. Scherrer (1998). Asymptotic distributions of subspace estimates under misspecification of the order. In: *Proceedings of the 1998 MTNS.* Padua, Italy.

Bauer, D., M. Deistler and W. Scherrer (1999). Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs. *Automatica* **35**, 1243–1254

Bauer, D. and M. Wagner (1999a). Unit root analysis in a state space framework: Canonical form and maximum likelihood estimation. Mimeo.

Bauer, D. and M. Wagner (1999b). Estimating cointegrated systems using subspace algorithms: Simulation performance and applications. Mimeo.

Bierens, H.J. (1995) Nonparametric cointegration analysis. *CentER Discussion Paper* No. 95123, Tilburg.

Bierens, H.J. (1997). Nonparametric cointegration analysis. *Journal of Econometrics* **77**, 379 – 404.

Chatelin, F. (1983). *Spectral approximation of linear operators.* Academic Press.

Davidson, J. (1994). *Stochastic limit theory.* Oxford University Press, Oxford.

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

Hannan, E. and M. Deistler (1988). *The statistical theory of linear systems.* Wiley, New York.

Huang, D. and L. Guo (1990). Estimation of nonstationary armax models based on the Hannan-Rissanen method. *The Annals of Statistics* **18**, 1729–1756.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* **80**, 359 – 386.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59**, 1551 – 1580.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models*. Oxford University Press, Oxford.

Larimore, W. E. (1983). System identification, reduced order filters and modeling via canonical variate analysis. In: *Proc. 1983 Amer. Control Conference 2.* (H. S. Rao and P. Dorato, Eds.). Piscataway, NJ. pp. 445–451. IEEE Service Center.

Lütkepohl, H. and P. Saikkonen (1997). Impulse response analysis in infinite order cointegrated vector autoregressive processes. *Journal of Econometrics* **81**, 127–157.

Peternell, K. (1995). Identification of Linear Dynamic Systems by Subspace and Realization-Based Algorithms. PhD thesis. TU Wien.

Phillips, P.C.B and V. Solo (1992). Asymptotics for linear processes. *The Annals of Statistics* **20**, 971 – 1001.

Said, S.E. and D.A. Dickey (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* **71**, 599 – 607.

Saikkonen, P. (1992). Estimation and Testing of Cointegrated Systems by an Autoregressive Approximation. *Econometric Theory* **8**, 1 – 27.

Saikkonen, P. and H. Lütkepohl (1996). Infinite order cointegrated vector autoregressive processes: estimation and inference. *Econometric Theory* **12**, 814 – 844.

Saikkonen, P. and R. Luukkonen (1997). Testing cointegration in infinite order vector autoregressive processes. *Journal of Econometrics* **81**, 93–126.

Shin, D. W. and Y. D. Lee (1997). A study on misspecified nonstationary autoregressive time series with a unit root. *Journal of Time Series Analysis* **18**, 475–484.

Van Overschee, P. and 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**(1), 61–74.

Viberg, M., B. Ottersten, B. Wahlberg and L. Ljung (1993). Performance of subspace based state space system identification methods. In: *Proc. of the 12th IFAC World Congress*. Vol. 7. Sydney, Australia. pp. 369–372.

Wagner, M. (1999a). VAR Cointegration in VARMA Models. *Economics Series* **No. 65**, Institute for Advanced Studies, Vienna.

Wagner, M. (1999b). Bierens' and Johansen's methods - complements or substitutes? *Economics Series* **No. 74**, Institute for Advanced Studies, Vienna.

Yap, S.F. and G.C. Reinsel (1995). Estimating and Testing for Unit Roots in a Partially Nonstationary Vector Autoregressive Moving Average Model. *Journal of the American Statistical Association* **90**, 253 – 267.