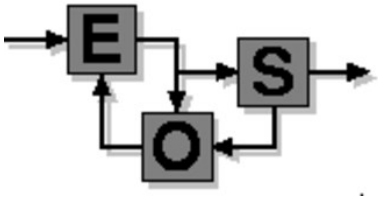


H A B I L I T A T I O N S S C H R I F T

Analyse der asymptotischen Eigenschaften von Subspace Algorithmen zur Schätzung von linearen dynamischen Systemen stationärer und integrierter Prozesse

ausgeführt zum Zwecke der Erlangung der venia docendi im Fach 'Ökonometrie'

eingereicht an der Technischen Universität Wien

von
Dietmar Bauer

Wien, am 24. Januar 2003

Einleitung

Diese Habilitationsschrift besteht aus den folgenden sechs Artikeln:

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

2. D. Bauer, M. Jansson: "Analysis of the asymptotic properties of the MOESP type of subspace algorithms"; *Automatica*, **36** (2000), p. 497 - 509.

3. D. Bauer, L Ljung: "Some facts about the choice of the weighting matrices in Larimore type of subspace algorithms"; *Automatica*, **38** (2002), p. 763 - 773.

4. D. Bauer: "Order estimation for subspace methods"; *Automatica*, **37** (2001), p. 1561 - 1573.

5. D. Bauer, M. Wagner: "Estimating Cointegrated Systems Using Subspace Algorithms"; *Journal of Econometrics*, **111** (2002), p. 47 - 84.

6. D. Bauer: "Identification of state space systems with conditionally heteroskedastic innovations"; in: "Proceedings of the 15th IFAC world congress", Bareclona, Spain, July 2002.

Alle sechs Beiträge beschäftigen sich mit den asymptotischen Eigenschaften von sogenannten 'Subspace'-Algorithmen, einer Klasse von Algorithmen zur Schätzung von linearen dynamischen Modellen für multivariate Input-Output Daten. Die Arbeiten beschäftigen sich sowohl mit der Analyse der Eigenschaften der Algorithmen für bereits spezifizierte Modellstruktur, sowie mit der Entwicklung und Analyse von Verfahren zur Modellspezifikation.
Hierbei beschäftigen sich die ersten vier Arbeiten mit dem klassischen stationären Fall mit homoskedastischen Fehlern. Arbeit Nummer fünf behandelt den Fall von integrierten Prozessen, welcher in der ökonometrischen Literatur des letzten Jahrzehnts einen prominenten Platz einnimmt. Die letzte Arbeit untersucht den heteroskedastischen Fall, der vor allem in der Modellierung von Finanzmarktdaten eine Rolle spielt.
Die ersten zwei Arbeiten beweisen die asymptotische Normalität für zwei verschiedene Klassen von 'subspace algorithms'. In der dritten Arbeit konnte für den Spezialfall keiner beobachteten Inputs (beziehungsweise weißen Rauschens als beobachteten Input) gezeigt werden, daß eine spezielle Klasse untersuchter Algorithmen, welche manchmal als `CCA` bezeichnet wird und die von (Larimore, 1983) vorgeschlagen wurde, optimal bezüglich der asymptotischen Varianz der Transferfunktionsschätzer ist. Weiters konnten für diese speziellen Fälle relativ einfache Ausdrücke für die asymptotische Varianz einer ganzen Klasse von Algorithmen gefunden werden. Die noch nicht veröffentlichte Arbeit (Bauer, 2000) beweist, daß für den Fall keiner beobachteten Inputs, `CCA` sogar asymptotisch äquivalent zu Pseudo-Maximum-Likelihood Schätzung ist und daher im Gauß-schen Fall asymptotisch effizient.

Die vierte Arbeit beschäftigt sich mit der Schätzung der Ordnung der Zustandsraumsysteme und untersucht drei Verfahren zur Ordnungsschätzung, die am Institut entwickelt wurden. Konsistenzaussagen bilden den Kern dieser Arbeit.
Die fünfte Arbeit beschäftigt sich mit der Schätzung von integrierten Prozessen, d.h. Prozessen, die mittels Bildung der ersten zeitlichen Differenz stationär gemacht werden können. Eine Erweiterung von `CCA` auf diese Klasse von Systemen wird entwickelt und dessen Konsistenz gezeigt. Weiters werden Fragen der Ordnungsschätzung und der Schätzung des kointegrierenden Ranges behandelt. Hinsichtlich der Spezifikation des kointegrierenden Ranges wird ein Test vorgeschlagen und dessen asymptotische Verteilung hergeleitet.
Die letzte Arbeit beschäftigt sich mit der Thematik von heteroskedastischen Innovationen und zeigt im wesentlichen, daß `CCA` einige Robustheitseigenschaften hat hinsichtlich der Konsistenz. Im Speziellen wird gezeigt, daß die Ordnung der fast sicheren Konvergenz der Schätzer der Systemmatrizen auch im Falle von bestimmten heteroskedastischen Innovationen erhalten bleibt (im stationären Fall). Die Klasse von Innovationen, für die dieses Robustheitsresultat gilt, enthält die häufig verwendeten univariaten GARCH-Prozesse (unter Einschränkung an die Parameterwerte) sowie univariate E-GARCH Prozesse.
Die ersten fünf Veröffentlichungen werden in der erschienenen Originalversion wiedergegeben. Im letzten Artikel wurde ein Druckfehler korrigiert, die Arbeit sonst aber nicht verändert.

Literaturverzeichnis

Bauer, D. (2000). Asymptotic efficiency of the CCA subspace method in the case of no exogenous inputs. Technical report. Department of Automatic Control, Linköping Universitetet. eingereicht bei Journal of Time Series Analysis.

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.

Introduction

This thesis contains the following six papers:

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

2. D. Bauer, M. Jansson: "Analysis of the asymptotic properties of the MOESP type of subspace algorithms"; *Automatica*, **36** (2000), p. 497 - 509.

3. D. Bauer, L Ljung: "Some facts about the choice of the weighting matrices in Larimore type of subspace algorithms"; *Automatica*, **38** (2002), p. 763 - 773.

4. D. Bauer: "Order estimation for subspace methods"; *Automatica*, **37** (2001), p. 1561 - 1573.

5. D. Bauer, M. Wagner: "Estimating Cointegrated Systems Using Subspace Algorithms"; *Journal of Econometrics*, **111** (2002), p. 47 - 84.

6. D. Bauer: "Identification of state space systems with conditionally heteroskedastic innovations"; in: "Proceedings of the 15th IFAC world congress", Bareclona, Spain, July 2002.

All six papers deal with the investigation of the asymptotic properties of so called subspace algorithms, a class of algorithms for the estimation of linear, dynamical models for multivariate input-output data. The papers analyze the properties of the estimators obtained for fixed model structure, as well as the problem of specifying the model structure based on the data at hand.
The first four papers deal with the stationary case, where the innovations are assumed to be homoskedastic. Paper number five deals with the case of integrated processes, which feature prominently in the econometrics literature over the last decade. The last paper analyzes the case of heteroskedastic innovations, which is a common assumption in financial econometrics.
The first two papers provide asymptotic normality results for two classes of subspace algorithms. The asymptotic variance for one class of estimators is analyzed in the third paper for the special case of no observed inputs (or white noise observed inputs respectively). The main result therein is the proof, that a certain subspace algorithm, sometimes termed as `CCA`, proposed by (Larimore, 1983) leads to optimal - within the class of subspace algorithms considered - estimates in the sense of lowest asymptotic variance matrix of the transfer function estimates. This result is based on the derivation of relatively simple expressions for the asymptotic variance of the estimated system matrices. This is interesting in connection with the results of the unpublished article (Bauer, 2000), which shows, that `CCA` in the case of Gaussian innovations is asymptotically efficient.
The fourth paper deals with questions of estimating the order of the state space systems. Three different algorithms for the estimation of the order are analyzed,

all of which were developed at the institute. The main result in this respect is the proof of consistency for all three algorithms.
The fifth paper deals with the extensions of `CCA` to the case of integrated processes, i.e. processes, which can be transformed into stationary processes by taking first differences. The properties of the original `CCA` algorithm in this case are analyzed and an adaptation is suggested, which also in the case of integrated processes leads to consistent estimation of the system matrices. This paper also contains a discussion on order estimation issues and develops and analyzes a testing procedure for finding the cointegrating rank.
The final paper investigates the robustness properties of `CCA` with respect to heteroskedastic innovations. The main result in this respect is that the estimates obtained using `CCA` are robust to heteroskedasty in the sense, that consistency is preserved and also the order of almost sure convergence is unchanged for a certain class of heteroskedastic innovations. This class contains i.a. univariate GARCH processes (subject to restrictions on the parameters) as well as E-GARCH processes.
The first five papers are reprinted in the original published version. In the sixth paper a misprint has been corrected, the remaining paper has been unaltered.

PERGAMON

Automatica 35 (1999) 1243–1254

Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs[1]

D. Bauer, M. Deistler, W. Scherrer*,[2]

Institut für Ökonometrie, Operations Research und Systemtheorie, Technische Universität Wien, Argentinierstr. 8/119, A-1040 Vienna, Austria

Received 7 January 1998; revised 28 July 1998; received in final form 26 January 1999

Asymptotic normality for a class of subspace algorithms, which estimate the state in a first step, is derived. Expressions for the asymptotic variance are given.

Abstract

Linear systems with unobserved white noise inputs are considered. A class of subspace estimates for the system matrices obtained by estimating the state in the first step is analyzed. The main result presented here states asymptotic normality of subspace estimates. In addition, a consistency result for the system matrix estimates is given. An algorithm to compute the asymptotic variances of the estimates is presented. In a final section the implications of the result are discussed. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Subspace methods; Identification; Asymptotic analysis; Discrete time systems; Linear multivariable systems; State-space systems

1. Introduction

Subspace algorithms for the identification of linear dynamic systems recently have attained great attention (Larimore, 1983; Van Overschee and DeMoor, 1994; Verhaegen, 1994; Peternell, 1995; Deistler et al., 1995). The advantage of subspace methods compared to methods based on optimization of a criterion function such as the likelihood or the prediction error lies in their numerical properties. They can be implemented numerically efficiently and use only standard reliable numerical tools such as the singular-value decomposition. Subspace algorithms make use of the structure of the realization problem (see e.g. Akaike, 1975; Lindquist and Picci, 1985). In addition to classical realization of course in identification model reduction has to be performed (see e.g. Glover, 1984; Desai et al., 1985).

On the other hand the statistical properties of these algorithms are not fully understood yet. In Deistler et al. (1995) consistency has been proved. Simulation studies e.g. in Peternell (1995) and Bauer et al. (1997) indicate, that the relative efficiency of some subspace methods, compared to the maximum likelihood estimates, is close to one. Up to now no general analytical results concerning the asymptotic efficiency of subspace algorithms have been obtained. In Viberg et al. (1993) the asymptotic distribution of the estimates of the poles of the system has been derived. These lines have been further developed in Wahlberg and Jansson (1994) and Jansson (1995). In a frequency domain setting results are given in McKelvey (1995). Our contribution (for more details see Bauer (1998)) is a further step towards an analytical understanding of this problem. In this paper, asymptotic normality for the estimates of the system matrices (A, B, C, D) described below is derived.

The paper is organized as follows: In the next section, the class of subspace algorithms under consideration is presented. In the third section some definitions and notations are introduced as well as some preliminary facts, which are proved in the appendix. In the fourth section

* Corresponding author. Tel.: + 43 1 58801 11944; fax.: + 43 1 58801 11999; e-mail: dietmar.bauer@tuwien.ac.at.

[1] Most of the material contained in this paper originally was presented at 'SYSID'97, Fukuoka, Japan'. This paper was recommended for publication in revised form by Associate Editor B. Ninness under the direction of Editor T. Söderström.

[2] Support by the Austrian 'Fonds zur Förderung der wissenschaftlichen Forschung' Projekt P11213-MAT is gratefully acknowledged.

0005-1098/99/$ - see front matter
PII: S0005-1098(99)00031-X

asymptotic normality of the parameter estimates is stated. The proof is given in Section 5. In Section 6 the implications of the result are discussed.

2. Model set and algorithms

In this paper, linear, time invariant, finite-dimensional state-space systems are considered. Only the case is considered, where the inputs are unobserved white noise. The system considered is of the form

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

where $(y_t)_{t\in\mathbb{Z}}$ denotes the s-dimensional measured output, $(\varepsilon_t)_{t\in\mathbb{Z}}$ denotes s-dimensional zero mean white noise with variance equal to the identity. x_t denotes the n-dimensional state. $A\in\mathbb{R}^{n\times n}$, $B\in\mathbb{R}^{n\times s}$, $C\in\mathbb{R}^{s\times n}$ and $D\in\mathbb{R}^{s\times s}$ are parameter matrices. The transfer function of the system is given by: $k(z)=D+C(zI-A)^{-1}B=\sum_{j=0}^{\infty}K(j)z^{-j}$, where $K(0)=D$ and $K(j)=CA^{j-1}B, j>0$. It will always be assumed that D is nonsingular. Throughout the paper the system will be assumed to be minimal, stable (i.e. $|\lambda_{\max}(A)|<1$ holds) and strictly minimum-phase (i.e. $|\lambda_{\max}(A-BD^{-1}C)|<1$ holds). Here $\lambda_{\max}(.)$ denotes an eigenvalue of maximum modulus. Thus the system is assumed to be in innovation form. The spectrum f of the stationary process $(y_t)_{t\in Z}$ is equal to $f(\lambda)=(1/2\pi)k(e^{i\lambda})k^*(e^{i\lambda})$. Here* denotes conjugate transpose. Since it is assumed, that the noise (ε_t) is not observed, the stability and the strict minimum-phase assumption exclude only spectra, which have zeros on the unit circle. The factorization of the spectrum is only unique, if a unique representative for the matrix D from the class $\{DQ\colon Q^TQ=I\}$ is chosen. Therefore, D is restricted to be lower triangular with positive entries on the diagonal throughout the paper.

The subspace algorithms considered here, use the fact that the state represents, in a certain sense, the interface between the past and the future of the process $(y_t)_{t\in Z}$. Let $Y_t^+=[y_t^T, y_{t+1}^T, \dots]^T$, $Y_t^-=[y_{t-1}^T, y_{t-2}^T, \dots]^T$ and let E_t^+ be the analogously defined vector of the future of the noise. Using system equations (1) it is easy to show that $Y_t^+=\mathcal{O}x_t+\mathcal{E}E_t^+$ and $x_t=\mathcal{K}Y_t^-$ holds, where $\mathcal{O}=[C^T, A^TC^T, (A^2)^TC^T, \dots]^T$ denotes the observability matrix, $\mathcal{K}=[BD^{-1}, (A-BD^{-1}C)\,BD^{-1}, (A-BD^{-1}C)^2BD^{-1}, \dots]$ and finally

$$\mathcal{E}=\begin{bmatrix} D & & & \\ CB & D & & 0 \\ CAB & CB & D & \\ \vdots & & \ddots & \ddots \end{bmatrix}.$$

Both equations together give

$$Y_t^+=\mathcal{O}\mathcal{K}Y_t^-+\mathcal{E}E_t^+. \tag{2}$$

Thus, since the future of the noise and the past of the process $(y_t)_{t\in Z}$ are uncorrelated, $\mathcal{O}\mathcal{K}Y_t^-$ is the orthogonal projection of Y_t^+ onto the space spanned by the elements of Y_t^-. (Here projection has to be understood in the context of the Hilbert space $\overline{\operatorname{span}}\{y_{t,i}\colon t\in\mathbb{Z}, i=1,\dots,s\}$ endowed with the inner product $\langle a,b\rangle=\mathbb{E}ab$ where $\mathbb{E}$ denotes expectation.)

Now commence from a process $(y_t)_{t\in\mathbb{Z}}$ rather than from the system representation (1). Every decomposition of the linear operator attaching to the past (Y_t^-) the projection $\mathcal{O}\mathcal{K}Y_t^-$ into two rank n operators, $\mathcal{O}$ and $\mathcal{K}$ then fixes a basis in the state space. Using such a decomposition, $x_t=\mathcal{K}Y_t^-, \forall t\in\mathbb{Z}$ is a state sequence, which then defines the system matrices via projecting y_t on x_t and x_{t+1} on x_t and ε_t i.e. $C=\mathbb{E}\{y_tx_t^T\}(\mathbb{E}\{x_tx_t^T\})^{-1}$, $A=\mathbb{E}\{x_{t+1}x_t^T\}(\mathbb{E}\{x_tx_t^T\})^{-1}$, $B=\mathbb{E}\{x_{t+1}\varepsilon_t^T\}(\mathbb{E}\{\varepsilon_t\varepsilon_t^T\})^{-1}$. D can be calculated as the lower triangular Cholesky factor of $\gamma(0)-CPC^T>0$, where $\gamma(0)$ denotes the variance of y_t and P the variance of x_t.

For given sample size T the (infinite dimensional) eq. (2) cannot be used and thus a decision on the number of block rows, f say, and the number of block columns, p say, which are included, has to be made. In the following these integers are called *truncation indices*. Throughout the paper it is assumed, that $f\geq n$ holds. Let $Y_{t,f}^+=[y_t^T, y_{t+1}^T, \dots, y_{t+f-1}^T]^T$ and $Y_{t,p}^-=[y_{t-1}^T, y_{t-2}^T, \dots, y_{t-p}^T]^T$ be finite-dimensional vectors of stacked outputs. Define $\mathcal{O}_f=[C^T, A^TC^T, \dots, (A^{f-1})^TC^T]^T$, $\mathcal{K}_p=[BD^{-1}, (A-BD^{-1}C)\,BD^{-1}, \dots, (A-BD^{-1}C)^{p-1}BD^{-1}]$. $\mathcal{E}_f$ denotes the first f block rows of $\mathcal{E}$. Then, of course, eq. (2) gives the following equation:

$$Y_{t,f}^+=\mathcal{O}_f\mathcal{K}_pY_{t,p}^-+\mathcal{O}_f(A-BD^{-1}C)^p\underbrace{\mathcal{K}Y_{t-p}^-}_{=x_{t-p}}+\mathcal{E}_fE_t^+.$$

Using this equation, the subspace methods considered here can be decomposed into three main steps (comp. Peternell, 1995):

(1) Regress $Y_{t,f}^+$ on $Y_{t,p}^-$ to get an estimate $\hat{\beta}_{f,p}$ of $\mathcal{O}_f\mathcal{K}_p$. If $\hat{\Gamma}_p^-$ and $\hat{\mathcal{H}}_{f,p}$ denote the sample variance of $Y_{t,p}^-$ and the sample covariance between $Y_{t,f}^+$ and $Y_{t,p}^-$, respectively, then this estimate is given by $\hat{\beta}_{f,p}=\hat{\mathcal{H}}_{f,p}(\hat{\Gamma}_p^-)^{-1}$.
(2) Approximate $\hat{\beta}_{f,p}$ by a rank n matrix and decompose this approximation into the product $\hat{\mathcal{O}}_f\hat{\mathcal{K}}_p$ of two rank n matrices to get an estimate $\hat{\mathcal{K}}_p$ of $\mathcal{K}_p$.
(3) Use the estimate of the state $\hat{x}_t=\hat{\mathcal{K}}_pY_{t,p}^-$ to estimate C by regressing y_t on $\hat{x}_t$. In the next step estimate $[A, BD^{-1}]$ by regressing $\hat{x}_{t+1}$ on $\hat{x}_t$ and $\hat{\varepsilon}_t$, the residuals of the first regression. Finally, the estimate of D is calculated as the lower triangular Cholesky factor of the sample covariance of $\hat{\varepsilon}_t$.

The approximation step (2) is performed by a singular-value decomposition of the matrix $\hat{W}_f^+\hat{\beta}_{f,p}\hat{W}_p^-=\hat{U}\hat{\Sigma}\hat{V}^T=\hat{U}_n\hat{\Sigma}_n\hat{V}_n^T+\hat{R}$, where $\hat{W}_f^+$ and $\hat{W}_p^-$ are weighting

matrices, $\hat{U}_n \in \mathbb{R}^{fs \times n}$, $\hat{V}_n \in \mathbb{R}^{ps \times n}$ and $\hat{\Sigma}_n$ is a diagonal matrix containing the n largest singular values as its diagonal entries (in decreasing order). The remaining singular values contribute to $\hat{R}$ and thus are neglected. This corresponds to approximating $\hat{\beta}_{f,p}$, which will typically be of full rank due to finite data length, by a rank n approximation $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p = [(\hat{W}_f^+)^{-1} \hat{U}_n (\hat{\Sigma}_n)^{1/2}][(\hat{\Sigma}_n)^{1/2} \hat{V}_n^{\mathrm{T}} (\hat{W}_p^-)^{-1}]$. Note, that the singular vectors, i.e. the columns of the matrix $\hat{U}_n$, are unique up to sign changes, if the singular values are distinct. This corresponds to a basis transformation of the form diag($\pm 1, \ldots, \pm 1$). This will be further discussed in the next section.

Throughout the paper it is assumed that the state dimension n is known. In practice the SVD is also used to determine the model order, either by inspection of the singular values and user's choice or by means of a criterion function (see Fuchs, 1990; Peternell, 1995).

The choice of the weighting matrices is essential. Different choices lead to algorithms, which have different asymptotic properties. Larimore's CCA procedure (Larimore, 1983) is obtained by choosing $\hat{W}_f^+ = (\hat{\Gamma}_f^+)^{-1/2}$ and $\hat{W}_p^- = (\hat{\Gamma}_p^-)^{1/2}$, where $\hat{\Gamma}_f^+$ denotes the sample variance of $Y_{t,f}^+$. A variant of N4SID (Van Overschee and DeMoor, 1994) corresponds to $\hat{W}_f^+ = I_f$ and $\hat{W}_p^- = (\hat{\Gamma}_p^-)^{1/2}$. In the following the discussion will be restricted to these two particular choices of the weighting matrices. Note that the estimates of (A, B, C, D) do not depend on the particular choice of the square root $(\hat{\Gamma}_p^-)^{1/2}$, $(\hat{\Gamma}_p^-)^{1/2}(\hat{\Gamma}_p^-)^{\mathrm{T}/2} = \hat{\Gamma}_p^-$ or $(\hat{\Gamma}_f^+)^{1/2}, (\hat{\Gamma}_f^+)^{1/2}(\hat{\Gamma}_f^+)^{\mathrm{T}/2} = \hat{\Gamma}_f^+$.

3. The set $M^+(n)$

In order to state a central limit theorem for the estimates $(\hat{A}_{\mathrm{T}}, \hat{B}_{\mathrm{T}}, \hat{C}_{\mathrm{T}}, \hat{D}_{\mathrm{T}})$ of the system matrices, first the particular representation of the true transfer function k_0, which is the limit of the estimates for $T \to \infty$, has to be determined. As is well known, for minimal state-space representations, the class of all observationally equivalent system matrices corresponding to the transfer function k is given by different choices of the basis in the state space. In the algorithms considered here, the choice of the basis is done implicitly by decomposing the rank n approximation to $\hat{\beta}_{f,p}$ into $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$. This decomposition is performed using the SVD of $\hat{W}_f^+ \hat{\beta}_{f,p} (\hat{\Gamma}_p^-)^{1/2}$ and is unique, if the first n singular values of this matrix have multiplicity one and if the orientation of the singular vectors is fixed. Let Γ^- denote the population variance of Y_t^- and let $\mathcal{H}_f$ denote the population covariance between $Y_{t,f}^+$ and Y_t^-. As will be shown in the proof in Section 5, the matrix $\hat{X}_p = \hat{W}_f^+ \hat{\beta}_{f,p} (\hat{\Gamma}_p^-) \hat{\beta}_{f,p}^{\mathrm{T}} (\hat{W}_f^+)^{\mathrm{T}} = \hat{W}_f^+ \hat{\mathcal{H}}_{f,p} (\hat{\Gamma}_p^-)^{-1} \hat{\mathcal{H}}_{f,p}^{\mathrm{T}} (\hat{W}_f^+)^{\mathrm{T}}$ converges to $\bar{X} = W_f^+ \mathcal{H}_f (\Gamma^-)^{-1} \mathcal{H}_f^{\mathrm{T}} (W_f^+)^{\mathrm{T}} = W_f^+ \mathcal{O}_f \mathcal{K} \Gamma^- \mathcal{K}^{\mathrm{T}} \mathcal{O}_f^{\mathrm{T}} (W_f^+)^{\mathrm{T}}$ a.s., if the index p is a function of the sample size T, which tends to infinity at a certain rate (see Theorem 1). Here $W_f^+ = (\Gamma_f^+)^{-1/2}$ for the CCA algorithm and $W_f^+ = I$ for the N4SID algorithm and Γ_f^+ denotes the population variance of $Y_{t,f}^+$. Thus if the n nonzero eigenvalues of $\bar{X}$ are distinct, then for T large enough, the n largest singular values of $(\hat{W}_f^+) \hat{\beta}_{f,p} (\hat{\Gamma}_p^-)^{1/2}$ will be distinct too, by the continuity of the singular values (see the forthcoming Lemma 7).

Let $M(n)$ denote the set of all rational, stable, strictly minimum-phase transfer functions k of McMillan degree n with a constant term, which is lower diagonal and has strictly positive diagonal entries. Furthermore, let $M^+(n) \subset M(n)$ denote the subset of all transfer functions $k \in M(n)$, for which $\bar{X}$ has n distinct (non zero) eigenvalues. This subset $M^+(n)$ is generic in the sense that it is an open and dense subset of $M(n)$, where $M(n)$ is endowed with the so-called pointwise topology. Since the proof of the genericity of $M^+(n)$ is lengthy and not essential for the understanding of the rest of the paper, it is shifted to the appendix. Note, that the set $M^+(n)$ depends on the choice of the index f as well as on the choice of the weighting matrix W_f^+, thus in particular on the choice of either CCA or N4SID, however, for the sake of notational simplicity, this will not be explicitly indicated in the notation.

Now for each transfer function $k_0 \in M^+(n)$ a particular representation may be obtained as follows. Note that the eigenvalue decomposition of $\bar{X} = U_n \Sigma_n^2 U_n^{\mathrm{T}}$ fixes a basis in the state space by the choice $\mathcal{O}_f = (W_f^+)^{-1} U_n \Sigma_n^{1/2}$, $\mathcal{K} = \Sigma_n^{-1/2} U_n^{\mathrm{T}} W_f^+ \mathcal{H}_f (\Gamma^-)^{-1}$, and $x_t = \mathcal{K} Y_t^-$. Since the eigenvalues of $\bar{X}$ are distinct, the eigenvectors U_n are unique up to sign changes. Now the choice of the basis in the state space is uniquely defined if in each column of U_n a nonzero entry is chosen to be strictly positive. In this way a unique realization of $k_0 \in M^+(n)$ is constructed. By fixing the elements in the same positions in $\hat{U}_n$ to be positive, a unique algorithm is obtained. This can be done from a certain T_0 onwards, since $\hat{X}_p$ converges to $\bar{X}$ a.s. under our assumptions.

4. A central limit theorem

In this section the main result of this paper, i.e. asymptotic normality will be stated. In Deistler et al. (1995) consistency of the algorithms was shown in the sense of convergence of the estimated transfer function to the true transfer function rather than convergence of the system matrix estimates. Here it is proved, that the estimates of the system matrices are consistent. To achieve consistency in our framework, the truncation index p has to tend to infinity. This is essentially due to the fact, that in the first step a regression is performed, neglecting information from the far past (contained in x_{t-p}). For a central limit theorem convergence of the estimates of order $\sqrt{T}$ is needed, thus a lower bound on the increase of the index p has to be imposed in order to ensure that

the effect of neglecting the far past does not show up in the limiting distribution. On the other hand, the limited amount of data imposes upper bounds for the increase of p in order to ensure a uniform convergence of the estimates of the covariances.

The following theorem contains the main result of the paper:

Theorem 1. *Let $(y_t)_{t\in\mathbb{Z}}$ be generated by a true transfer function $k_0 \in M^+(n)$, where the ergodic white noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ fulfills the following conditions*:

$$\mathbb{E}\{\varepsilon_t|\mathscr{F}_{t-1}\} = 0,$$

$$\mathbb{E}\{\varepsilon_t\varepsilon_t^{\mathrm{T}}|\mathscr{F}_{t-1}\} = \mathbb{E}\{\varepsilon_t\varepsilon_t^{\mathrm{T}}\} = I,$$

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

$$\mathbb{E}\{\varepsilon_{t,i}^4\} < \infty,$$

where $\mathbb{E}$ denotes expectation, $\mathscr{F}_t$ the σ-algebra spanned by the past and present of the noise and additional subscripts here indicate components of the vector ε_t. $\omega_{a,b,c}$ is a constant not depending on t and *$f \geq n$ is a fixed integer.*

If p fulfills the following conditions:

(1) $p \geq -(d\log T)/(2\log|\rho_0|), \forall T > T_0$ *for some* $d > 1$, *where ρ_0 is a zero of k_0 of maximum modulus,*
(2) $p/(\log T)^a \to 0$ *for some* $a < \infty$.

then

$$\sqrt{T}\,\mathrm{vec}[\hat{A}_T - A_0, \hat{B}_T - B_0, \hat{C}_T - C_0, \hat{D}_T - D_0] \xrightarrow{\mathrm{d}} Z,$$

where Z is a multivariate normal random variable with zero mean and variance V_f^θ, and $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution. Here (A_0, B_0, C_0, D_0) denotes the particular realization of k_0 described in Section 3.

The asymptotic variance V_f^θ depends on f, k_0 and on the choice of the weighting matrices. However, this is not emphasized in the expressions for notational convenience. Note, that the assumptions on the noise process are exactly the same as those given in Hannan and Deistler (1988), where the asymptotic normality of maximum likelihood estimates is derived.

Note that the lower bound of the increase of the truncation index p depends on the true system. However, it is possible to estimate this bound consistently as follows: Fit a (long) autoregression to the sample data and estimate the order of this AR model by the BIC criterion. If the true process is ARMA, then the estimated order $\hat{p}_{\mathrm{BIC}}$ will fulfill $\lim_{T\to\infty} -(2\hat{p}_{\mathrm{BIC}}\log|\rho_0|)/(\log T) = 1$ *a.s.* (see e.g. Hannan and Deistler, 1988). Therefore $d\hat{p}_{\mathrm{BIC}}$ (or $d\hat{p}_{\mathrm{AIC}}$), for some $d > 1$ seems to be a reasonable choice for the truncation index p.

The central limit theorem for the system matrix estimates also implies a central limit theorem for other quantities, which are derived from the system matrix estimates:

Corollary 2. *Let $g: \mathbb{R}^{n^2+2ns+(s^2+s)/2} \to \mathbb{R}^m$ be a mapping attaching the vector $x \in \mathbb{R}^m$ to the matrices (A, B, C, D). If, under the assumptions of Theorem* 1, *(A_0, B_0, C_0, D_0) denotes the realization of $k_0 \in M^+(n)$ described in Section* 3 *and if g is differentiable at (A_0, B_0, C_0, D_0), then the following holds:*

$$\sqrt{T}(g(\hat{A}_T, \hat{B}_T, \hat{C}_T, \hat{D}_T) - g(A_0, B_0, C_0, D_0)) \xrightarrow{\mathrm{d}} Z, \qquad (3)$$

where Z is a multivariate normally distributed random variable with mean zero and variance $V = J_g V_f^\theta J_g^{\mathrm{T}} \in \mathbb{R}^{m\times m}$, where J_g denotes the matrix of partial derivatives of g evaluated at (A_0, B_0, C_0, D_0).

In particular, three applications of this corollary are of interest:

- The poles of the system depend differentiably on the entries in the matrix A, if the eigenvalues are distinct (see Lemma 7). Thus a CLT for the estimates of the system poles is obtained on some generic subset of $M(n)$ (comp. Wahlberg and Jansson, 1994). The same statement is true for the estimates of the system zeros.
- For fixed frequency ω the transfer function evaluated at ω is equal to $k_0(\mathrm{e}^{\mathrm{i}\omega}) = D_0 + C_0(\mathrm{e}^{\mathrm{i}\omega}I - A_0)^{-1}B_0$. Thus a central limit theorem for the estimates of the transfer function at fixed frequency points is obtained. This can be used, to compare different choices of procedures (i.e. different choices of f and of W_f^+) for a given system.
- For given system matrices the transformation to Echelon coordinates is differentiable for system matrices corresponding to a transfer function in the generic neighborhood corresponding to the Echelon parametrization. Thus a CLT for the Echelon parameter estimates on a generic neighborhood holds (since the intersection of two generic sets is still a generic set). This can also be used to compare different procedures correponding to their asymptotic behaviour.

Note that the algorithm has been made unique by restricting certain elements in each column of $\hat{U}_n$ to be positive. However, the actual implementation of the SVD algorithm may use a different selection of the signs of the singular vectors. Thus, the system obtained by the algorithm may be related to $(\hat{A}_T, \hat{B}_T, \hat{C}_T, \hat{D}_T)$ as defined above by a basis transformation corresponding to $\mathrm{diag}(\pm 1, \ldots, \pm 1)$. Consistency and asymptotic normality for the estimates of the actually implemented SVD will hold, if this SVD of $\bar{X}$ is continuous at the true system. The results of Corollary 2 will hold, if g depends only on the transfer function k, even if the actual SVD is

not continuous at the true transfer function k_0. (In this case the estimates of the system matrices may not converge to a single point, but to the equivalence class $\{(TA_0T^{-1}, TB_0, C_0T^{-1}, D_0)\colon T = \mathrm{diag}(\pm 1, \ldots, \pm 1)\}$).

5. Proof of the theorem

To simplify the notation, in the proof only the case of Larimore's procedure will be considered, i.e. the case where $\hat{W}_f^+ = (\hat{\Gamma}_f^+)^{-1/2}$. The N4SID algorithm can be treated in a completely analogous manner. In fact for the N4SID algorithm some steps simplify. In the proof extensive use of the following notation will be made: $g(T) = o(f(T))$ means $\lim_{T\to\infty} g(T)/f(T) = 0$ a.s., $g(T) = O(f(T))$ means $\sup_{T\in\mathbb{N}}|g(T)/f(T)| < M$ a.s. for some constant $M < \infty$. $g(T) = o_{\mathrm{P}}(f(T))$ means that for every $\varepsilon > 0$, $\mathbb{P}\{|g(T)/f(T)| > \varepsilon\} \to 0$ for $T \to \infty$. The Frobenius norm of a matrix X is denoted by $\|X\| = \sqrt{\mathrm{tr}\, X^{\mathrm{T}}X}$ and the ℓ^1-norm of a (possibly infinite dimensional) vector is denoted with $\|X\|_1 = \sum_i |X_i|$.

5.1. Preliminaries

For the class of algorithms considered in this paper it is straightforward to see, that the estimates $(\hat{A}_T, \hat{B}_T, \hat{C}_T, \hat{D}_T)$ are a nonlinear function of the sample autocovariances. For given indices f and p the estimates depend only on the sample covariances up to lag $f + p - 1$ i.e. on $\hat{\gamma}(0), \hat{\gamma}(1), \ldots, \hat{\gamma}(f + p - 1)$, where $\hat{\gamma}(j)$ denotes the estimate $1/T\sum_{t=j+1}^{T} y_t y_{t-j}^{\mathrm{T}}$ of $\gamma(j) = \mathbb{E}y_t y_{t-j}^{\mathrm{T}}, j \geq 0$. For consistency the column truncation index p has to tend to infinity at a certain rate (Deistler et al., 1995). Thus also the number of included covariance estimates tends to infinity, which causes the main technical complication in the proof.

Introduce the shorthand notation $\langle a_t, b_t\rangle = 1/T\sum_{t=1}^{T} a_t b_t^{\mathrm{T}}$. Then the regressions in step 3 can be written as follows:

$$\hat{C}_T = \langle y_t, \hat{x}_t\rangle\langle \hat{x}_t, \hat{x}_t\rangle^{-1}, \tag{4}$$

$$\hat{D}_T = (\hat{\gamma}(0) - \hat{C}_T\langle \hat{x}_t, \hat{x}_t\rangle\hat{C}_T^{\mathrm{T}})^{1/2}, \tag{5}$$

$$\hat{A}_T = \langle \hat{x}_{t+1}, \hat{x}_t\rangle\langle \hat{x}_t, \hat{x}_t\rangle^{-1}, \tag{6}$$

$$\hat{B}_T = \langle \hat{x}_{t+1}, \hat{\varepsilon}_t\rangle = (\langle \hat{x}_{t+1}, y_t\rangle - \langle \hat{x}_{t+1}, \hat{x}_t\rangle\hat{C}_T^{\mathrm{T}})\hat{D}_T^{-\mathrm{T}}, \tag{7}$$

where the innovations ε_t are estimated by $\hat{\varepsilon}_t = \hat{D}_T^{-1}(y_t - \hat{C}_T\hat{x}_t)$. Note that the estimated residuals are orthogonal to the estimates of the state, i.e. $\langle \hat{x}_t, \hat{\varepsilon}_t\rangle = 0$, and thus $\hat{A}_T$ and $\hat{B}_T$ may be obtained by the two separate regressions (6) and (7).

In the following, the above expressions will be further analyzed. The estimates of the states are defined by $\hat{x}_t = \hat{\mathscr{K}}_p Y_{t,p}^-$ and $\hat{x}_{t+1} = \hat{\mathscr{K}}_p Y_{t+1,p}^-$. The matrix $\hat{\mathscr{K}}_p$ is obtained from the matrix $\hat{X}_p = \hat{\mathscr{H}}_{f,p}(\hat{\Gamma}_p^-)^{-1}\hat{\mathscr{H}}_{f,p}^{\mathrm{T}}$ in the following way: Recall from Section 3, that $\hat{\Sigma}_n$ contains the square roots of the largest n eigenvalues of the matrix $(\hat{\Gamma}_f^+)^{-1/2}\hat{X}_p(\hat{\Gamma}_f^+)^{-T/2}$ and $\hat{U}_n$ contains the corresponding eigenvectors, i.e.

$$((\hat{\Gamma}_f^+)^{-1/2}\hat{X}_p(\hat{\Gamma}_f^+)^{-\mathrm{T}/2})\,\hat{U}_n = \hat{U}_n\hat{\Sigma}_n^2 \quad \text{and} \quad \hat{U}_n^{\mathrm{T}}\hat{U}_n = I_n.$$

The matrices $\hat{\mathscr{O}}_f$ and $\hat{\mathscr{K}}_p$ are defined as

$$\hat{\mathscr{O}}_f = (\hat{\Gamma}_f^+)^{1/2}\hat{U}_n\hat{\Sigma}_n^{1/2},$$

$$\hat{\mathscr{K}}_p = \hat{\Sigma}_n^{1/2}\hat{V}_n^{\mathrm{T}}(\hat{\Gamma}_p^-)^{-1/2} = \hat{\Sigma}_n^{-1}\hat{\mathscr{O}}_f^{\mathrm{T}}(\hat{\Gamma}_f^+)^{-1}\hat{H}_{f,p}(\hat{\Gamma}_p^-)^{-1}.$$

Furthermore, let

$$E_k = \begin{bmatrix} I_s \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{ks\times s} \quad \text{and} \quad S_p = \begin{bmatrix} 0 & \cdots & & & \\ I_s & 0 & & & \\ 0 & I_s & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & & I_s & 0 \end{bmatrix}$$

$\in \mathbb{R}^{ps\times ps}$.

Using the identities $y_t = E_f^{\mathrm{T}}Y_{t,f}^+ = E_p^{\mathrm{T}}Y_{t+1,p}^-$ and $Y_{t+1,p}^- = E_pE_f^{\mathrm{T}}Y_{t,f}^+ + S_pY_{t,p}^-$ it is straightforward to derive the following expressions:

$$\langle \hat{x}_t, \hat{x}_t\rangle = \hat{\Sigma}_n, \tag{8}$$

$$\langle y_t, \hat{x}_t\rangle = E_f^{\mathrm{T}}\hat{\mathscr{O}}_f\hat{\Sigma}_n, \tag{9}$$

$$\langle y_t, \hat{x}_{t+1}\rangle = E_p^{\mathrm{T}}\hat{\mathscr{H}}_{f,p}^{\mathrm{T}}(\hat{\Gamma}_f^+)^{-1}\hat{\mathscr{O}}_f\hat{\Sigma}_n^{-1}, \tag{10}$$

$$\begin{aligned}\langle \hat{x}_{t+1}, \hat{x}_t\rangle = {} & \hat{\Sigma}_n^{-1}\hat{\mathscr{O}}_f^{\mathrm{T}}(\hat{\Gamma}_f^+)^{-1}(\hat{\mathscr{H}}_{f,p}(\hat{\Gamma}_p^-)^{-1}E_p)E_f^{\mathrm{T}}\hat{\mathscr{O}}_f\hat{\Sigma}_n \\ & + \hat{\Sigma}_n^{-1}\hat{\mathscr{O}}_f^{\mathrm{T}}(\hat{\Gamma}_f^+)^{-1}(\hat{\mathscr{H}}_{f,p}(\hat{\Gamma}_p^-)^{-1}S_p\hat{\mathscr{H}}_{f,p}^{\mathrm{T}}) \\ & \times(\hat{\Gamma}_f^+)^{-1}\hat{\mathscr{O}}_f\hat{\Sigma}_n^{-1}.\end{aligned} \tag{11}$$

From the expressions given above it can be seen that the estimates of the true parameter matrices (A_0, B_0, C_0, D_0) are obtained via a nonlinear map attaching to $\hat{\gamma}(0), \ldots, \hat{\gamma}(f)$ and the finite-dimensional matrices $\hat{X}_p = \hat{\mathscr{H}}_{f,p}(\hat{\Gamma}_p^-)^{-1}\hat{\mathscr{H}}_{f,p}^{\mathrm{T}}$, $\hat{Y}_p = \hat{\mathscr{H}}_{f,p}(\hat{\Gamma}_p^-)^{-1}S_p\hat{\mathscr{H}}_{f,p}^{\mathrm{T}}$ and $\hat{Z}_p = \hat{\mathscr{H}}_{f,p}(\hat{\Gamma}_p^-)^{-1}E_p$ the corresponding matrices $(\hat{A}_T, \hat{B}_T, \hat{C}_T, \hat{D}_T)$.

In order to outline the proof of the CLT some further notation is introduced:

$$\hat{\theta}_T = \mathrm{vec}(\hat{A}_T, \hat{B}_T, \hat{C}_T, \hat{D}_T),$$

$$\theta_0 = \mathrm{vec}(A_0, B_0, C_0, D_0),$$

$$\hat{m}_p = \mathrm{vec}(\hat{\gamma}(0), \ldots, \hat{\gamma}(f), \hat{X}_p, \hat{Y}_p, \hat{Z}_p),$$

$$m_p = \mathrm{vec}(\gamma(0), \ldots, \gamma(f), X_p, Y_p, Z_p),$$

$$m_0 = \lim_{p\to\infty} m_p = \mathrm{vec}(\gamma(0), \ldots, \gamma(f), X_0, Y_0, Z_0),$$

$$\hat{g}_{T,h} = \mathrm{vec}(\hat{\gamma}(0), \ldots, \hat{\gamma}(h-1)),$$

$$g_h = \mathrm{vec}(\gamma(0), \ldots, \gamma(h-1)).$$

Here e.g. X_p is defined as $X_p = \mathscr{H}_{f,p}(\Gamma_p^-)^{-1}\mathscr{H}_{f,p}^{\mathrm{T}}$ and $X_0 = \lim_{p\to\infty} X_p = \mathscr{H}_f(\Gamma^-)^{-1}\mathscr{H}_f^{\mathrm{T}}$. Y_p, Y_0, Z_p, Z_0 are defined analogously, vec will be used to denote the vector of stacked vectorizations of several matrices with slight abuse of notation. It will be part of the proof to show the existence of all the required limits. As has been stated already $\hat{m}_p$ is a function of the sample autocovariances $\hat{g}_{T,f+p}$, $\hat{m}_p = \phi(\hat{g}_{T,f+p})$ say. Their population counterparts are given by $m_p = \phi(g_{f+p})$. The estimates $\hat{\theta}_T$ of the system matrices are obtained as a function $\hat{\theta}_T = \psi(\hat{m}_p)$. Then the proof of the CLT may be decomposed into the following four steps:

(1) A central limit theorem for the covariances $\hat{g}_{T,h}$, with $h = (f+p) \to \infty$ at a suitable rate for $T \to \infty$ (see Section 5.2).
(2) The proof, that $\sqrt{T}(m_p - m_0) \to 0$, and that $\psi(m_0) = \theta_0$ where θ_0 corresponds to the particular realization of the true system described in Section 3 (see Section 5.3).
(3) A central limit theorem for $\hat{m}_p$ i.e. $\sqrt{T}(\hat{m}_p - m_0) \xrightarrow{\mathrm{d}} Z$, where Z is multivariate normally distributed with mean zero and variance V^m, for $p \to \infty$ at a suitable rate for $T \to \infty$ (see section 5.4).
(4) The proof of the differentiability of the mapping ψ at the point m_0 (see Section 5.5). In other words it is proved that $\psi(m_0 + \delta m) - \psi(m_0) = J_\psi \delta m + o(\|\delta m\|)$ and thus one obtains

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = \sqrt{T}(\psi(\hat{m}_p) - \psi(m_0)) \xrightarrow{\mathrm{d}} Z,$$

where Z is multiviariate normally distributed with zero mean and variance $V_f^\theta = (J_\psi)\ V^m(J_\psi)^{\mathrm{T}}$. Here J_ψ denotes the Jacobian of ψ evaluated at m_0.

Note that the 'intermediate' variable $\hat{m}_p$ has been introduced since $\hat{m}_p$ opposed to $\hat{g}_{T,h}$ is of fixed finite dimension as p tends to infinity.

5.2. A CLT for the covariance estimates

As has already been stated, the main technical complication lies in the fact, that the index p and thus also the dimension of the stacked vector of autocovariances $\hat{g}_{T,f+p}$ has to tend to infinity. Tools to handle the growth of dimensions are provided in Lewis and Reinsel (1985). Their techniques are used to prove the following lemma:

Lemma 3. *For fixed h let $V_h^g \in \mathbb{R}^{hs^2 \times hs^2}$ denote the covariance matrix of $\sqrt{T}(\hat{g}_{T,h} - g_h)$ defined above. Let h depend on T such that $h = o((\log T)^\alpha)$ for some $\alpha < \infty$. Then under the assumptions on the process (y_t) and on the noise sequence (ε_t) given in Theorem 1, for every sequence of vectors $l(T) \in \mathbb{R}^{hs^2}$ satisfying $0 < c_1 \le l(T)^{\mathrm{T}} V_h^g l(T)$ and $\|l(T)\|_1 \le c_2 < \infty$ it follows that*

$$\sqrt{T}\frac{l(T)^{\mathrm{T}}(\hat{g}_{T,h} - g_h)}{(l(T)^{\mathrm{T}} V_h^g l(T))^{1/2}} \xrightarrow{\mathrm{d}} Z, \tag{12}$$

where Z is scalar normally distributed with zero mean and unit variance.

Proof. In a first step the lemma is proved for the case $y_t = \varepsilon_t$. Note that in this case

$$\sqrt{T}(\hat{g}_{T,h} - g_h) = 1/\sqrt{T}\sum_{t=1}^{\mathrm{T}} \mathrm{vec}[\varepsilon_t\varepsilon_t^{\mathrm{T}} - I, \varepsilon_t\varepsilon_{t-1}^{\mathrm{T}}, \ldots, \varepsilon_t\varepsilon_{t-h+1}^{\mathrm{T}}]$$
$$= \mathrm{vec}[e_0, e_1, \ldots, e_{h-1}],$$

where

$$e_i = 1/\sqrt{T}\sum_{t=1}^{\mathrm{T}} \varepsilon_t\varepsilon_{t-i}^{\mathrm{T}},\ i = 1, \ldots, h-1$$

and

$$e_0 = 1/\sqrt{T}\left[\sum_{t=1}^{\mathrm{T}} \varepsilon_t\varepsilon_t^{\mathrm{T}} - I\right].$$

Now let $l(T) \in \mathbb{R}^{hs^2}$ be a sequence of vectors satisfying $0 < c_1 \le l(T)^{\mathrm{T}} V_h^\varepsilon l(T)$, where the notation indicates, that V_h^ε corresponds to $y_t = \varepsilon_t$. Furthermore, let $v_{T,h}^2 = l(T)^{\mathrm{T}} V_h^\varepsilon l(T)$ denote the variance of $\sqrt{T}l(T)^{\mathrm{T}}(\hat{g}_{T,h} - g_h)$, which is bounded due to $\|l(T)\|_1 \le c_2 < \infty$ (see also Remark 4). Then following the proof of Theorem 3 in Lewis and Reinsel (1985) $\sqrt{T}l(T)^{\mathrm{T}}(\hat{g}_{T,h} - g_h)/v_{T,h} = \sum_{t=1}^{T} X_t(T) = \sum_{t=1}^{T} 1/\sqrt{T}l(T)^{\mathrm{T}}\mathrm{vec}[\varepsilon_t\varepsilon_t^{\mathrm{T}} - I, \ldots, \varepsilon_t\varepsilon_{t-h+1}^{\mathrm{T}}]/v_{T,h}$. The assumptions on ε_t imply that $X_t(T)$ and $X_s(T)$ are uncorrelated for $t \ne s$ with expectation equal to zero and that the variance of $X_t(T)$ is equal to $1/T$. $\sum_{t=1}^{n} X_t(T)$, $n = 1, \ldots, T$ is a martingale sequence for each T. In order to prove the convergence of $\sqrt{T}l(T)^{\mathrm{T}}(\hat{g}_{T,h} - g_h)/v_{T,h}$ to a normal distribution, it is sufficient to show, that

(a) $\sup_{t\le T} X_t^2(T) \xrightarrow{\mathrm{p}} 0$ for $T \to \infty$,
(b) $\sum_{t=1}^{\lfloor \tau T \rfloor} X_t^2(T) \xrightarrow{\mathrm{p}} \tau,\ 0 < \tau \le 1$ for $T \to \infty$,

where $\xrightarrow{\mathrm{p}}$ denotes convergence in probability and $\lfloor \tau T \rfloor$ denotes the integer part of τT. Write

$$X_t(T) = X_{t,0}(T) + X_{t,1}(T),$$

where

$$X_{t,0}(T) = (1/\sqrt{T v_{T,h}^2})l_0(T)^{\mathrm{T}}\,\mathrm{vec}(\varepsilon_t\varepsilon_t^{\mathrm{T}} - I),$$

where $l_0(T)$ denotes the vector of the first s^2 elements of $l(T)$. Then $\sup_{t\le T} X_{t,0}(T)^2 \xrightarrow{\mathrm{p}} 0$, since $\sup_{t\le T} X_{t,0}(T)^2 \le \sup_{t\le T} X_{t,0}(t)^2$ and $\mathbb{E}X_{t,0}(t)^2 \to 0$ for $t \to \infty$ (compare Hannan and Deistler, 1988, p. 149). $\sup_{t\le T} X_{t,1}(T)^2 \xrightarrow{\mathrm{p}} 0$ follows from the arguments in Lewis and Reinsel (1985). Thus the convergence of $\sup_{t\le T} X_t(T)^2$ to zero follows from $X_t(T)^2 \le 2(X_{t,0}(T)^2 + X_{t,1}(T)^2)$.

For condition (b) note, that $\sum X_{t,0}(T)^2$ converges due to ergodicity of ε_t and thus of $(\varepsilon_t\varepsilon_t^{\mathrm{T}} - I)^2$ and the assumption of finite fourth moments. The convergence of $\sum X_{t,1}(T)^2$ can be seen, using the arguments of Lewis and Reinsel (1985). Finally, the contribution of the mixed terms $\sum X_{t,0}(T)X_{t,1}(T) \xrightarrow{\mathrm{p}} 0$, which can be seen as follows: $|\sum X_{t,0}(T)X_{t,1}(T)| \le \sup_{t\le T}|X_{t,0}(T)|\sum|X_{t,1}(T)|$. Now $\sup_{t\le T}|X_{t,0}(T)|$ converges in probability to zero, since $\sup_{t\le T} X_{t,0}(T)^2$ does, and $\mathbb{P}\{\sum|X_{t,1}(T)| > \delta\} \le \sum\mathbb{P}\{|X_{t,1}(T)| > \delta\} \le \sum\mathbb{P}\{X_{t,1}(T)^2 > \delta^2\} \to 0$ (see the proof in Lewis and Reinsel, 1985). This shows the convergence in distribution to a random variable, which is normally distributed with mean zero and of unit variance. Note, that the normalization by $v_{T,h}$ is necessary here, if no conditions on the limiting behaviour of $l(T)^{\mathrm{T}}V_h^{\varepsilon}l(T)$ except for its boundedness are imposed.

In order to extend this result to the autocovariances of a process $(y_t)_{t\in Z}$, generated by model (1), note that $y_t = \sum_{i=0}^{\infty}K(i)\varepsilon_{t-i}$ holds, and that the Markov parameters $K(i)$ converge exponentially to zero i.e. $\|K(i)\| \le c(\rho_p')^i$ for some constants $c > 0$ and $1 > \rho_p' > |\lambda_{\max}(A)|$. Now substituting this expression for y_t in $\hat{\gamma}(j)$ and in $\gamma(j)$ one obtains

$$\sqrt{T}(\hat{\gamma}(j) - \gamma(j)) = \sum_{i,l=0}^{a} K(i)\bar{e}_{j+l-i}K(l)^{\mathrm{T}} + r(j), \tag{13}$$

where $\bar{e}_j = e_j$ for $j \ge 0$ and $\bar{e}_j = e^{\mathrm{T}}_{-j}$ for $j < 0$. The term $r(j)$ may be decomposed into four components: The first one is due to the replacement of $\mathbb{E}\hat{\gamma}(j) = ((T-j)/T)\gamma(j)$ by $\gamma(j)$. This term may be bounded by ch/T.

To obtain a bound for the other contributions the following assessment will be heavily used:

$$\mathbb{E}\left\|\sum_{t\in\mathscr{I}}(\varepsilon_t\varepsilon_{t-i}^{\mathrm{T}} - \delta_{0,i}I)\right\|^2 \le \sum_{t\in\mathscr{I}}\mathbb{E}\|(\varepsilon_t\varepsilon_{t-i}^{\mathrm{T}} - \delta_{0,i}I)\|^2 \le c|\mathscr{I}|. \tag{14}$$

Here $|\mathscr{I}|$ denotes the number of elements of the indexset $\mathscr{I}$. The above inequlatities follow from the fact that the terms $\mathrm{vec}(\varepsilon_t\varepsilon_{t-i}^{\mathrm{T}} - \delta_{0,i}I)$ and $\mathrm{vec}(\varepsilon_s\varepsilon_{s-i}^{\mathrm{T}} - \delta_{0,i}I)$ are uncorrelated for $t \ne s$.

The second contribution is due to the approximation of the process y_t by the finite sums $\sum_{i=0}^{a}K(i)\varepsilon_{t-i}$. Because of Eq. (14) and the exponential decrease of the Markov parameters the expectation of the Frobenius norm of this term is bounded by $c(\rho_p')^a$. Therefore, this term converges to zero if a converges to infinity faster than $-\log T/(2\log|\rho_p'|)$.

The third term is due to the fact, that sums of the form $\sum_{t=j+1-i}^{T-i}(\varepsilon_t\varepsilon_{t-j-l+i}^{\mathrm{T}} - \delta_{0,j+l-i}I)$ are replaced by e_{j+l-i}, i.e. by sums where the summation index runs from $1, \dots, T$. The difference is the sum of at most $(2a+1)$ summands. Therefore by eq. (14) and by the exponential decrease of the Markov parameters the expectation of the Frobenius norm of this term may be bounded by $c\sqrt{a/T}$.

The last contribution stems from the replacement of e_i with e^{T}_{-i} for $i < 0$. Now e_i and e^{T}_{-i} differ only in the first i and last i sumands. Therefore, by the same reasoning as above this term may be bounded by $c\sqrt{(a+h)/T}$.

Putting together these considerations imply that

$$\mathbb{E}\|r(j)\| \le c_1\sqrt{\frac{a+h}{T}} + c_2(\rho_p')^a, \tag{15}$$

where the constants c_1, c_2 do not depend on j.

Now the first term in eq. (13) will be analyzed in more detail. Define $M_l^a = \sum_{i=\max(0,-l)}^{\min(a,a-l)}K(l+i)\otimes K(i)$ and $M_l = \lim_{a\to\infty}M_l^a = \sum_{i=\max(0,-l)}^{\infty}K(l+i)\otimes K(i)$, then

$$\mathrm{vec}\left(\sum_{i,l=0}^{a}K(i)\bar{e}_{j+l-i}K(l)^{\mathrm{T}}\right) = \sum_{l=-a}^{a}M_l^a\,\mathrm{vec}(\bar{e}_{j+l}).$$

Note that $M = (\dots, M_{-1}, M_0, M_1, \dots)$ has rows which are elements of ℓ_2 and $M^a = (\dots, M^a_{-1}, M^a_0, M^a_1, \dots)$ converges to M, since it can be shown that $\|M^a - M\| \le c(\rho_p')^a$. Here the sequence M_l^a is extended on both sides with zeros, i.e. $M_l^a = 0 \in \mathbb{R}^{s^2\times s^2}$ for $|l| > a$. Now let the permutation matrix $P \in \mathbb{R}^{s^2\times s^2}$ be defined such that $P\,\mathrm{vec}(H) = \mathrm{vec}(H^{\mathrm{T}})$ for every matrix $H \in R^{s\times s}$. Then it follows that the linear term in eq. (13) can be written as

$$\mathrm{vec}\left(\sum_{i,l=0}^{a}K(i)\bar{e}_{j+l-i}K(l)^{\mathrm{T}}\right) = \sum_{l=0}^{a+h}L_{j,l}^a\,\mathrm{vec}(e_l) = L_j^{a,h}\,\mathrm{vec}(e_0, e_1, \dots, e_{a+h}),$$

where $L_{j,l}^a = M^a_{-j}$ for $l = 0$ and $L_{j,l}^a = M^a_{l-j} + M^a_{-l-j}P$ for $l > 0$. Clearly, the rows of the matrix $L_j^{a,h}$, when extended with zeros, converge in the ℓ_2 sense to the rows of the $(s^2\times\infty)$ matrix $L_j = (L_{j,0}, L_{j,1}, \dots)$, where $L_{j,l} = \lim_{a\to\infty}L_{j,l}^a$. In fact by the convergence of $M^a \to M$, it follows that $\|L_j^a - L_j\| \le c(\rho_p')^a$ holds, where the constant does not depend on j.

Assembling the expressions for the covariances $\sqrt{T}(\hat{\gamma}(j) - \gamma(j))$ in the vector $\sqrt{T}(\hat{g}_{T,h} - g_h)$ then leads to

$$\sqrt{T}(\hat{g}_{T,h} - g_h) = L^{a,h}\,\mathrm{vec}(e_0, e_1, \dots, e_{a+h}) + \mathrm{vec}(r(0), r(1), \dots, r(h)),$$

where the matrix $L^{a,h} \in \mathbb{R}^{hs^2\times(a+h)s^2}$ has as its jth block row the matrix $L_j^{a,h}$. By the convergence of the rows of $L^{a,h}$ and by the block diagonal structure of V_{a+h}^{ε} it follows that, for fixed h, the variance of $\sqrt{T}(\hat{g}_{T,h} - g_h)$ exists and is given by $V_h^g = \lim_{a\to\infty}L^{a,h}V_{a+h}^{\varepsilon}(L^{a,h})^{\mathrm{T}}$, see Remark 4 below.

Finally, let $l(T) \in \mathbb{R}^{hs^2}$ be a sequence of vectors fullfilling $0 < c_1 \le l(T)^{\mathrm{T}}V_h^g l(T)^{\mathrm{T}}$ and $\|l(T)\|_1 \le c_2 < \infty$. Then

$$\sqrt{T}l(T)^{\mathrm{T}}(\hat{g}_{T,h} - g_h) = l(T)^{\mathrm{T}}L^{a,h}\mathrm{vec}(e_0, e_1, \dots, e_{a+h}) + o_{\mathrm{P}}(1),$$

where the second term on the right hand side is equal to $l(T)^{\mathrm{T}}\,\mathrm{vec}(r(0),\ \dots,\ r(h))$ and converges to zero in probability, as follows from eq. (15). The convergence of L_j^a to L_j and the bounded ℓ^1-norm of $l(T)$ imply (*for* $h \le a$)

$$\lim_{a,h\to\infty} (l(T)^{\mathrm{T}} L^{a,h} V^{\varepsilon}_{a+h} (L^{a,h})^{\mathrm{T}} l(T) - l(T)^{\mathrm{T}} V_h^g l(T)) = 0.$$

Finally, it follows that $l(T)^{\mathrm{T}} L^{a,h} V^{\varepsilon}_{a+h} (L^{a,h})^{\mathrm{T}} l(T)$ is a sequence bounded from below and above and thus the first part of the proof gives the desired result. □

Remark 4. Recall, that the variance matrix V_h^g can be calculated as follows: $V_h^g = \lim_{a\to\infty} L^{a,h} V^{\varepsilon}_{h+a} (L^{a,h})^{\mathrm{T}}$. This limit exists due to the structure of V_x^{ε} and $L^{a,h}$ and due to the convergence of $L^{a,h}$ for fixed h. Now the variance of $\mathrm{vec}[e_0, \dots, e_{x-1}]$, which has been denoted as $V_x^{\varepsilon} \in \mathbb{R}^{xs^2 \times xs^2}$, where x is an arbitrary integer, has as its $[js^2 + (b-1)s + a, is^2 + (d-1)\ s + c]$ entry the following expression: $1/T \sum_{t,s=1}^{\mathrm{T}} [\mathbb{E}(\varepsilon_{t,a}\varepsilon_{t-j,b} - \delta_{j0}\delta_{ab})(\varepsilon_{s,c}\varepsilon_{s-i,d} - \delta_{i0}\delta_{cd})]$, where δ denotes the Kronecker delta function. For $s < t$ ($t < s$) conditional expectation on $\mathscr{F}_{t-1}$ ($\mathscr{F}_{s-1}$) shows, that the contribution is zero. Thus only the contribution for $t = s$ has to be examined. If $j \neq i$, then again taking the expectation conditional on $\mathscr{F}_{t-\min(i,j)}$ shows, that the expectation is zero (here the assumption on the third order moments is needed to simplify the expressions). Thus V_x^{ε} is blockdiagonal. In order to calculate the blockdiagonal entries, we distinguish the two cases $i = j > 0$ and $i = j = 0$. As can easily be seen, for $j > 0$, the expectation $\mathbb{E}(\varepsilon_{t,a}\varepsilon_{t-j,b})(\varepsilon_{t,c}\varepsilon_{t-j,d}) = \delta_{ac}\delta_{bd}$, which is equal to 1, if $a = c$ and $b = d$ and zero else. Thus the variance matrix in this case is equal to the identity. For $j = 0$, the expectation $\mathbb{E}(\varepsilon_{t,a}\varepsilon_{t,b} - \delta_{ab})\ (\varepsilon_{t,c}\varepsilon_{t,d} - \delta_{cd}) = \mathbb{E}\varepsilon_{t,a}\varepsilon_{t,b}\varepsilon_{t,c}\varepsilon_{t,d} - \delta_{ab}\delta_{cd}$. Now for Gaussian ε_t, the fourth moment $\mathbb{E}\varepsilon_{t,a}\varepsilon_{t,b}\varepsilon_{t,c}\varepsilon_{t,d}$ is equal to $\delta_{ab}\delta_{cd} + \delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}$. Thus the expectation $\delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}$ is equal to 2, if $a = b = c = d$, equal to 1, if $a \neq b$ and $a = c \wedge b = d$ or $a \neq b$ and $a = d \wedge b = c$, and equal to zero else.

Remark 5. The lemma also shows, that the condition $L_h V_h^g L_h^{\mathrm{T}} \to V$, $L_h \in R^{m\times hs^2}$ (m fixed and finite) is a sufficient condition for $\sqrt{T} L_h(\hat{g}_{T,h} - g_h) \xrightarrow{\mathrm{d}} Z$, where here Z is multivariate normally distributed with mean zero and variance equal to V (see e.g. Anderson, 1971, Theorem 7.7.7). Since V_h^g is a matrix with elements of bounded infinity norm for a stable system, where the bound is independent of h, it is straightforward to see, that a sufficient condition for $L_h V_h^g L_h^{\mathrm{T}} \to V$ to hold for some V is that the rows of L_h embedded in ℓ^2 converge in the ℓ^2 norm to an infinite-dimensional vector, having elements decreasing exponentially. This will be the condition used in the sequel. Note, that the requirement $l(T)^{\mathrm{T}} V_h^g l(T) > 0$ is only needed for the normalization of the variance and thus can be dropped for our purposes.

5.3. *Convergence of* m_p *to* m_0

First, it will be proved that $\|m_p - m_0\| = O(|\rho_0'|^p)$, where $1 > \rho_0' > |\rho_0|$ and ρ_0 denotes an eigenvalue of $(A - BD^{-1}C)$, i.e. a zero of the transfer function, of maximum modulus. For this purpose the following lemma is proved:

Lemma 6. $\|\mathscr{H}_{f,p}(\Gamma_p^-)^{-1} - \mathscr{O}_f \mathscr{K}_p\| = O(|\rho_0'|^p)$.

Proof. The equality $\mathscr{H}_f(\Gamma^-)^{-1} = \mathscr{O}_f \mathscr{K}$ implies that $[\mathscr{H}_{f,p}(\Gamma_p^-)^{-1}, 0](\Gamma^-)_p - \mathscr{O}_f \mathscr{K}(\Gamma^-)_p = 0$, where $(\Gamma^-)_p$ denotes the first p block columns of the infinite-dimensional matrix Γ^-. From this it follows, that $\mathscr{H}_{f,p}(\Gamma_p^-)^{-1} - \mathscr{O}_f \mathscr{K}_p = \mathscr{O}_f (A - BD^{-1}C)^p \mathscr{K} \tilde{\mathscr{H}}_p (\Gamma_p^-)^{-1}$. Here $\tilde{\mathscr{H}}_p$ denotes the matrix obtained by omitting the first p block rows in $(\Gamma^-)_p$, which is a (reordered) part of the covariance Hankelmatrix $\mathscr{H} = \mathbb{E} Y_t^+ (Y_t^-)^{\mathrm{T}}$ and thus has finite Frobenius norm, independently of p. The Frobenius norm of $\mathscr{O}_f$ can also be bounded (independently of f), Γ_p^- has bounded eigenvalues independently of p and finally $\mathscr{K}$ is of finite Frobenius norm. Thus, the Frobenius norm of the error can be bounded by the Frobenius norm of $(A - BD^{-1}C)^p$ times a constant, which depends only on the underlying system and not on the choice of the truncation indices. Now $\|(A - BD^{-1}C)^p\|$ can be bounded by $|\rho_0'|^p$, for all $1 > \rho_0' > |\rho_0|$. □

This lemma immediately implies $\|Z_p - Z_0\| = \|\mathscr{H}_{f,p}(\Gamma_p^-)^{-1} E_p - \mathscr{H}_f(\Gamma^-)^{-1} E_\infty\| = O(|\rho_0'|^p)$. Furthermore, it is easy to see that

$$\begin{aligned}\|X_p - X_0\| &= \|\mathscr{H}_{f,p}(\Gamma_p^-)^{-1} \mathscr{H}_{f,p}^{\mathrm{T}} - \mathscr{H}_f \Gamma^{-1} \mathscr{H}_f^{\mathrm{T}}\| \\ &\le \|(\mathscr{H}_{f,p}(\Gamma_p^-)^{-1} - \mathscr{O}_f \mathscr{K}_p)\ \mathscr{H}_{f,p}^{\mathrm{T}}\| \\ &\quad + \|\mathscr{O}_f (A - BD^{-1}C)^p \mathscr{K}\| \|\mathscr{H}_f^{\mathrm{T}}\|\end{aligned}$$

is of the desired order $O(|\rho_0'|^p)$. Applying the techniques of Lemma 6 it follows that $\|Y_p - Y_0\| = \|\mathscr{H}_{f,p}(\Gamma_p^-)^{-1} S_p \mathscr{H}_{f,p}^{\mathrm{T}} - \mathscr{H}_f(\Gamma^-)^{-1} S_\infty \mathscr{H}_f^{\mathrm{T}}\| = O(|\rho_0'|^p)$.

Now in order to prove $\sqrt{T}\|m_p - m_0\| \to 0$ it is sufficient that $\sqrt{T}|\rho_0'|^p \to 0$ holds, which is guaranteed by the condition $p \ge -(d \log T)/(2 \log|\rho_0|)$ for some $d > 1$. These considerations show, that one has to impose a lower bound on the convergence rate of $p \to \infty$ to ensure $\sqrt{T}$-consistency of the estimate.

Next, it is proved that $\psi(m_0) = \theta_0$. In other words the subspace algorithms considered give a realization algorithm, if the true autocovariances are used and $p = \infty$. First note that by construction the eigenvalue decomposition $(\Gamma_f^+)^{-1/2} X_0 (\Gamma_f^+)^{-T/2} = U_n \Sigma_n^2 U_n^{\mathrm{T}}$ gives a factorization $\mathscr{H} = \mathscr{O}_f \bar{\mathscr{C}}$ and $\mathscr{H}(\Gamma^-)^{-1} = \mathscr{O}_f \mathscr{K}$, where

$$\mathscr{O}_f = (C_0^{\mathrm{T}}, A_0^{\mathrm{T}} C_0^{\mathrm{T}}, (A_0^2)^{\mathrm{T}} C_0^{\mathrm{T}}, \dots)^{\mathrm{T}},$$

$$\bar{\mathscr{C}} = (M_0, A_0 M_0, A_0^2 M_0, \dots),$$

$$\mathscr{K} = (B_0 D_0^{-1}, (A_0 - B_0 D_0^{-1} C_0)\, B_0 D_0^{-1}, \ \dots),$$

$$D_0 D_0^{\mathrm{T}} = \gamma(0) - C_0 \Sigma_n C_0^{\mathrm{T}},$$

$$M_0 = A_0 \Sigma_n C_0^{\mathrm{T}} + B_0 D_0^{\mathrm{T}},$$

and the matrices (A_0, B_0, C_0, D_0) are the realization of the transfer function k_0, explained in Section 3. Now the mapping ψ evaluated at the point m_0 is considered. From eqs. (4) and (9) one obtains that $C = E_f^{\mathrm{T}} \mathscr{O}_f = C_0$ and thus from eq. (5) that $D = D_0$. Term (11) simplifies to

$$\begin{aligned}
&\Sigma_n^{-1} \mathscr{O}_f^{\mathrm{T}} (\Gamma_f^+)^{-1} [\mathscr{H}_f (\Gamma^-)^{-1} E_\infty E_f^{\mathrm{T}} \mathscr{O}_f \Sigma_n \\
&\quad + (\mathscr{H}_f (\Gamma^-)^{-1} S_\infty \mathscr{H}_f^{\mathrm{T}})\, (\Gamma_f^+)^{-1} \mathscr{O}_f \Sigma_n^{-1}] \\
&= \Sigma_n^{-1} \mathscr{O}_f^{\mathrm{T}} (\Gamma_f^+)^{-1} [\mathscr{O}_f B_0 D_0^{-1} C_0 \Sigma_n \\
&\quad + \mathscr{O}_f (A_0 - B_0 D_0^{-1} C_0) \Sigma_n] = A_0 \Sigma_n
\end{aligned}$$

since $\mathscr{H}_f(\Gamma^{-1}) E_\infty = \mathscr{O}_f \mathscr{K}_1 = \mathscr{O}_f B_0 D_0^{-1}$, $E_f^{\mathrm{T}} \mathscr{O}_f = C_0$, $\mathscr{H}_f (\Gamma^-)^{-1} S_\infty = \mathscr{O}_f (A_0 - B_0 D_0^{-1} C_0)\, \mathscr{K}$ and finally $\mathscr{K} \mathscr{H}^{\mathrm{T}} (\Gamma_f^+)^{-1} \mathscr{O}_f = \Sigma_n^2$ and $\mathscr{O}_f^{\mathrm{T}} (\Gamma_f^+)^{-1} \mathscr{O}_f = \Sigma_n$. Thus it follows that $A = A_0$. Now eq. (10) gives $\langle y_t, x_{t+1} \rangle \xrightarrow{\mathrm{d}} M_0$, which implies that $B = B_0$.

5.4. A CLT for $\hat{m}_p$

In this subsection the results of the previous subsection will be used to prove a central limit theorem for the vector $\hat{m}_p = \mathrm{vec}(\hat{\gamma}(0), \dots, \hat{\gamma}(f), \hat{X}_p, \hat{Y}_p, \hat{Z}_p)$. This will be done by linearizing the map ϕ attaching $\hat{m}_p$ to the sample covariances $\hat{g}_{T, f+p}$. It will be shown that $\hat{m}_p = m_p + L_p(\hat{g}_{T, p+f} - g_{T, p+f}) + o_{\mathrm{P}}(T^{-1/2})$. In order to apply Lemma 3 it then remains to show, that the rows of L_p converge in the ℓ^2 norm to vectors with elements decreasing exponentially.

First the term $\hat{X}_p = \hat{\mathscr{H}}_{f,p} (\hat{\Gamma}_p^-)^{-1} \hat{\mathscr{H}}_{f,p}^{\mathrm{T}}$ is considered. By linearizing this expression one obtains that

$$\begin{aligned}
&\sqrt{T}(\hat{\mathscr{H}}_{f,p} (\hat{\Gamma}_p^-)^{-1} \hat{\mathscr{H}}_{f,p}^{\mathrm{T}} - \mathscr{H}_{f,p} (\Gamma_p^-)^{-1} \mathscr{H}_{f,p}^{\mathrm{T}}) \\
&\quad = (\sqrt{T}(\hat{\mathscr{H}}_{f,p} - \mathscr{H}_{f,p}))\, (\Gamma_p^-)^{-1} \mathscr{H}_{f,p}^{\mathrm{T}} \\
&\qquad - \mathscr{H}_{f,p} (\Gamma_p^-)^{-1} (\sqrt{T}(\hat{\Gamma}_p^- - \Gamma_p^-))\, (\Gamma_p^-)^{-1} \mathscr{H}_{f,p}^{\mathrm{T}} \\
&\qquad + \mathscr{H}_{f,p} (\Gamma_p^-)^{-1} (\sqrt{T}(\hat{\mathscr{H}}_{f,p} - \mathscr{H}_{f,p})^{\mathrm{T}}) \\
&\qquad + \text{higher-order terms.} \qquad (16)
\end{aligned}$$

In order to prove that the higher-order terms are of order $o_{\mathrm{P}}(1)$, the uniform convergence of the sample autocovariances has to be used: Under the upper bound on the increase of p as a function of T, it follows that $\max_{|j| \le f+p-1} \|\hat{\gamma}(j) - \gamma(j)\| = O(Q_{\mathrm{T}})$, where $Q_{\mathrm{T}} = \sqrt{\log \log T / T}$ (see e.g. Hannan and Deistler, 1988). This result implies that $\|\hat{\mathscr{H}}_{f,p} - \mathscr{H}_{f,p}\| = O(pQ_{\mathrm{T}})$, $\|\hat{\Gamma}_p^- - \Gamma_p^-\| = O(p^2 Q_{\mathrm{T}})$. In addition, it can be shown that $\|(\Gamma_p^-)^{-1}\|$ and $\|(\hat{\Gamma}_p^-)^{-1}\|$ are of order $O(p)$. Using these bounds some simple but tedious calculations show that the higher-order terms are of order $\sqrt{T} p^k Q_{\mathrm{T}}^j$, $j \ge 2$, $k \le 6$ and thus converge to zero in probability under the assumptions on the increase of p.

It remains to show that the rows of L_p corresponding to the term $\hat{X}_p$ converge in the ℓ^2 norm to vectors with elements decreasing exponentially. This follows immediately from Lemma 6.

The terms $\hat{Y}_p = \hat{\mathscr{H}}_{f,p} (\hat{\Gamma}_p^-)^{-1} S_p \mathscr{H}_{f,p}^{\mathrm{T}}$ and $\hat{Z}_p = \hat{\mathscr{H}}_{f,p} (\hat{\Gamma}_p^-)^{-1} E_p$ can be analyzed in a completely analogous manner, by showing that $\|\mathscr{H}_{f,p} S_p^{\mathrm{T}} (\Gamma_p^-)^{-1} - (\mathscr{H}_f S_\infty^{\mathrm{T}} (\Gamma^-)^{-1})_p\|$ converges to zero and that $\mathscr{H}_f S_\infty^{\mathrm{T}} (\Gamma^-)^{-1}$ has rows with exponentially decreasing entries.

5.5. Differentiability of ψ

The last part of the proof consists of the proof of the differentiability of ψ at the value $m_0 = \mathrm{vec}[\gamma(0), \dots, \gamma(f), X_0, Y_0, Z_0]$. Thus it remains to show, that in the neighborhood of m_0, the approximation $\psi(m_0 + \delta m) = \psi(m_0) + J_\psi \delta m + o(\|\delta m\|)$ holds. For this purpose the essential steps for the computation of ψ are repeated.

First, a Cholesky factorization $(\Gamma_f^+)^{1/2}$ of Γ_f^+ and the inverse of Γ_f^+ and of $(\Gamma_f^+)^{1/2}$ have to be computed. Since Γ_f^+ is positive definite, these computations are differentiable. Corresponding to the Cholesky decomposition this can be seen from the recursions defining the Cholesky factor (see e.g. Golub and Van Loan, 1989). The entries of the inverse of a matrix X depend differentiably on the entries of X, if X is nonsingular, which is straightforward to see.

Next the n largest eigenvalues and the corresponding eigenvectors of the matrix $(\Gamma_f^+)^{-1/2} X (\Gamma_f^+)^{-T/2}$ have to be computed. The differentiability of the mapping attaching the eigenvectors and the eigenvalues to the matrix $(\Gamma_f^+)^{-1/2} X (\Gamma_f^+)^{-T/2}$ holds for $k_0 \in M^+(n)$, due to the following result, which can be found e.g. in Chatelin (1983):

Lemma 7. *If λ_i is an eigenvalue of multiplicity one of a symmetric matrix $X \in \mathbb{R}^{m \times m}$, which has a basis of eigenvectors, with corresponding eigenvector u_i, then the eigenvalue $\tilde{\lambda}_i$ and the corresponding eigenvector $\tilde{u}_i$ of the perturbed matrix $X + \varepsilon X_1$ are given for first-order approximation by*

$$\tilde{\lambda}_i \doteq \lambda_i + u_i^{\mathrm{T}} \varepsilon X_1 u_i, \qquad (17)$$

$$\tilde{u}_i \doteq u_i + \sum_{j=1, j \ne i}^{m} \frac{u_j^{\mathrm{T}} \varepsilon X_1 u_i}{\lambda_i - \lambda_j} u_j. \qquad (18)$$

Here $u_1, \dots, u_m$ are the eigenvectors of X and $\doteq$ means, that the error is $o(\varepsilon)$.

Although the theorem only states the existence of the directional derivatives in direction X_1, the differentiability follows from the fact, that the directional derivatives are continuous in X_1. This result can be found in standard textbooks on analysis (see e.g. Kowalsky, 1974).

Thus the singular values σ_i and the corresponding singular vectors u_i depend differentiably on the entries in the matrix $(\Gamma_f^+)^{-1/2}X(\Gamma_f^+)^{-T/2}$. This shows, that $\hat{\mathcal{O}}_f = (\hat{W}_f^+)^{-1}\hat{U}_n\hat{\Sigma}_n^{1/2}$ depends differentiably on $\hat{m}_p$. The remaining steps are easily analyzed using eqs. (4)–(11). One only has to keep in mind that Σ_n and $(\gamma(0) - C\Sigma_n C^{\mathrm{T}})$ are positive definite and thus the Cholesky factorizations and the inverses, which have to be computed, are differentiable with respect to the entries of m_p.

5.6. Calculation of the asymptotic variance

The main result of this paper states asymptotic normality of the estimates. The asymptotic variance can be written formally as $V_f^\theta = \lim_{p\to\infty} L_{f,p}V_{f,p}^g L_{f,p}^{\mathrm{T}}$ as follows from Section 5.4, where $L_{f,p} = J_\psi L_p$ (L_p is defined in the first paragraph of Section 5.4 and J_ψ denotes the matrix of partial derivatives of the function ψ). The resulting formulas are too complicated to be investigated analytically. However, actual computations provide approximations to V_f^θ, where the approximation error can be made arbitrarily small by choosing a and p suitably large. This approximation can be calculated along the following lines:

Recall, that the asymptotic variance V_h^g of $\sqrt{T}[\hat{g}_{T,h} - g_h]$ for fixed h (see Lemma 3) is the limit $\lim_{a\to\infty} L^{a,h}V_{a+h}^\varepsilon(L^{a,h})^{\mathrm{T}}$, where the evaluation of V_{h+a}^ε is documented in Remark 4. Expressions for $L^{a,h}$ can be found in the proof of Lemma 3. The convergence of this expression is related to the magnitude of ρ_p^a (see the proof of Lemma 3). Corresponding to the asymptotic variance, V^m say, of $\sqrt{T}[\hat{m}_p - m_0]$ it has been shown, that $\sqrt{T}[\hat{m}_p - m_0] = \sqrt{T}L_p[\hat{g}_{T,p+f} - g_{p+f}] + o_{\mathrm{P}}(1)$. Thus $V^m = \lim_{p\to\infty} L_p V_{f+p}^g (L_p)^{\mathrm{T}}$, where the existence of the limit has been shown in Section 5.4. L_p can be found from eq. (16). Lemma 6 shows, that the convergence of L_p and thus of V^m depends heavily on $|\rho_0^p|$. Finally, ψ is a mapping between two finite-dimensional vector spaces and thus the derivative of ψ can be calculated without any approximation using the results of Lemma 7 and eqs. (4)–(11). Thus the approximation of V_f^θ can be found as $J_\psi L_p L^{a,f+p}V_{a+h}^\varepsilon(L^{a,f+p})^{\mathrm{T}}L_p^{\mathrm{T}}J_\psi^{\mathrm{T}}$ by taking a and p large, where the meaning of large depends on the location of the systems zeros (for p) and poles (for a).

6. Conclusions

In this paper the asymptotic properties of the estimates of system matrices are discussed, when the estimation is performed using a particular class of subspace algorithms. Here only the case, where no observed inputs are included, is treated. The discussion centers on the asymptotic distribution of the estimates. The paper contains a new consistency result, which states consistency for the system matrices. The main result states a central limit theorem for the estimates of the system matrices, if the true system is contained in a generic set (see Section 3). The estimates are found to be asymptotically normal and the variance may be calculated, since the proof of the CLT hinges on the linearization of the mapping attaching system matrix estimates to covariance estimates. This makes it possible, to compare for a given system the effects of different weighting matrices W_f^+, which is done in Bauer et al. (1997). It is also possible to compare for a given system the asymptotic variance to the optimal asymptotic variance, as obtained by the maximum likelihood approach. This is also done for some examples in Bauer et al. (1997).

An important condition for the central limit theorem is that the truncation index p has to tend to infinity at a certain rate, which depends on the true system. However, as has been stated already, this rate can be consistently estimated. The truncation index f however is fixed and has to fulfill $f \geq n$ in order for our result to hold. Simulation evidence suggests, that in some situations the choice of this index is rather important for the N4SID-type procedure, whereas it seems to be less critical for CCA.

Appendix: Genericity of the set $M^+(n)$

In order to simplify the notation, here only the case of the CCA weighting $W_f^+ = (\Gamma_f^+)^{-1/2}$ is considered. The case of N4SID can be treated in a completely analogous way.

Let $\Theta \subseteq \mathbb{R}^{(n+s)^2 - s(s-1)/2}$ denote the set of quadruples (A, B, C, D), where A and $A - BD^{-1}C$ are stable and D is lower triangular with strictly positive diagonal elements. Furthermore, let $\Theta_n \subset \Theta$ denote the set of all minimal realizations. It is easy to see, that the set of all realizations Θ is an open and nonvoid subset of $\mathbb{R}^{(n+s)^2 - s(s-1)/2}$ and that the set of all minimal realizations Θ_n is an open and dense subset of Θ.

Recall the definition of the central matrix $\bar{X} = (\Gamma_f^+)^{-1/2}\mathscr{H}_f(\Gamma^-)^{-1}\mathscr{H}_f^{\mathrm{T}}(\Gamma_f^+)^{-T/2}$ (see Section 3) and consider the set $\Theta_n^+ \subset \Theta$, where the corresponding matrix $\bar{X}$ has n distinct (nonzero) eigenvalues. Clearly Θ_n^+ is a subset of Θ_n and the next step is to prove that Θ_n^+ is open and dense in Θ.

First note that $\mathscr{H}_f(\Gamma^-)^{-1}\mathscr{H}_f^{\mathrm{T}} = \mathcal{O}_f\mathscr{K}\Gamma^-\mathscr{K}^{\mathrm{T}}\mathcal{O}_f^{\mathrm{T}} = \mathcal{O}_f\mathscr{C}\mathscr{C}^{\mathrm{T}}\mathcal{O}_f^{\mathrm{T}}$, where $\mathscr{C} = (B, AB, A^2B, \ldots)$ is the controllability matrix. Therefore, the n nonzero eigenvalues of $\bar{X}$ are equal to the eigenvalues of $\bar{Z} = [\mathcal{O}_f^{\mathrm{T}}(\Gamma_f^+)^{-1}\mathcal{O}_f][\mathscr{C}\mathscr{C}^{\mathrm{T}}]$. Since A is stable $P = \mathscr{C}\mathscr{C}^{\mathrm{T}} = \sum A^j BB^{\mathrm{T}}(A^j)^{\mathrm{T}}$ is an analytic function of (A, B). This implies that the autocovariances $\gamma(0) = CPC^{\mathrm{T}} + DD^{\mathrm{T}}$, $\gamma(j) = CA^{j-1}(APC^{\mathrm{T}} + BD^{\mathrm{T}})$, $j > 0$ are analytic functions of (A, B, C, D). Finally, since D has full rank and by the strict minimum phase assumption it follows that the entries of $\bar{Z}$ are analytic functions of (A, B, C, D) on Θ.

The characteristic polynomial $a(\lambda) = \det(\bar{Z} - \lambda I)$ and its derivative $b(\lambda) = (\mathrm{d}/\mathrm{d}\lambda)a(\lambda)$ have a common root iff $\bar{Z}$ has some eigenvalues of multiplicity larger than one. Thus the determinant of the corresponding Sylvester matrix

$$R = \begin{bmatrix} a_0 & a_1 & \cdots & \cdots & a_n & 0 & \cdots & 0 \\ 0 & a_0 & a_1 & \cdots & \cdots & a_n & & \\ & & \ddots & & & & \ddots & 0 \\ 0 & \cdots & 0 & a_0 & a_1 & \cdots & \cdots & a_n \\ b_0 & \cdots & \cdots & b_{n-1} & 0 & & \cdots & 0 \\ 0 & \ddots & & & & \ddots & & \\ & & \ddots & & & & & 0 \\ 0 & \cdots & 0 & b_0 & \cdots & \cdots & \cdots & b_{n-1} \end{bmatrix} \in \mathbb{R}^{(2n-1)\times(2n-1)}$$

is zero for all $(A, B, C, D) \in \Theta_n \backslash \Theta_n^+$ and nonzero for all $(A, B, C, D) \in \Theta_n^+$. Since $\det R$ is an analytic function of the parameters (A, B, C, D) it follows that Θ_n^+ is open in Θ_n and thus in Θ. Now suppose that Θ_n^+ were not dense in Θ_n. Then $\det R$ is zero on an open subset $\mathscr{V} \subset \Theta_n$. Now from the analycity of $\det R$ it follows that $\det R$ is zero on the largest pathwise connected subset of Θ, which contains $\mathscr{V}$. Since Θ is pathwise connected, one may conclude that $\det R$ is zero on Θ.

To find the desired contradiction now it is sufficient to construct an element in Θ_n^+, i.e. to prove that Θ_n^+ is not empty. This is done in a recursive manner. Clearly, for $n = 1$, $\Theta_1^+ = \Theta_1$ is nonvoid. Now suppose that the conjecture is true for $n - 1$. Therefore, there exists a transfer-function $k_0 \in M(n-1)$ for which $\bar{Z}$ has $n - 1$ nonzero distinct eigenvalues. Now let $(A_0, B_0, C_0, D_0) \in \Theta$ be a (nonminimal) realization of k_0 with n states. Then $\bar{Z}_0$ corresponding to (A_0, B_0, C_0, D_0) has one zero eigenvalue and $n - 1$ distinct eigenvalues larger than zero. Since Θ_n is dense in Θ, there exists a sequence $(A_k, B_k, C_k, D_k) \in \Theta_n$ which converges to (A_0, B_0, C_0, D_0). Furthermore, by the continuity of the eigenvalues of $\bar{Z}$, it follows that the eigenvalues of $\bar{Z}_k$ converge to the eigenvalues of $\bar{Z}_0$. Thus, there must exist an index k such that $\bar{Z}_k$ has n distinct nonzero eigenvalues, which proves that $(A_k, B_k, C_k, D_k) \in \Theta_n^+$.

Now let $M^+(n) = \pi(\Theta_n^+) \subset M(n)$, where π: $\Theta \to M(n)$ denotes the mapping attaching the transfer function $k(z) = D + C(zI - A)^{-1}B$ to matrix quadrupels (A, B, C, D). In the following, it is proved that $M^+(n)$ is an open and dense subset of $M(n)$. Here $M(n)$ is endowed with the so-called pointwise topology, which is defined as follows: The transfer function is identified with the sequence of its Markov parameters $(K(j)|j \in \mathbb{Z}^+)$. The set $(\mathbb{R}^{s\times s})^{\mathbb{Z}^+}$ is endowed with the product topology and the pointwise topology is the corresponding relative topology for transfer functions (comp. Hannan and Deistler, 1988). Consider a $k_0 \in M(n)$ and let $\mathscr{V} \subseteq M(n)$ be a neighbourhood of k_0 with a continuous parametrization, i.e. with a mapping φ: $\mathscr{V} \mapsto \Theta_n$, $k \mapsto (A, B, C, D)$, which is continuous. (One possibility is to use e.g. the overlapping parametrizations presented in Hannan and Deistler (1988, Chap. 2). Now consider the concatenated mapping $k \to \varphi(k) \to \det R$. Since φ is continuous this mapping is continuous. Thus if $k_0 \in M^+(n)$ then $\det R \neq 0$ holds in a neighbourhood of k_0, which gives the openess of $M^+(n)$. If $k_0 \in M(n)\backslash M^+(n)$, then by the denseness of Θ_n^+ and by the continuity of π, one can construct a sequence of transfer functions in $M^+(n)$, which converges to k_0. Thus $M^+(n)$ is dense in $M(n)$.

References

Akaike, H. (1975). Markovian representation of stochastic processes by canonical variables. *SIAM Journal of Control*, *13*(1), 162–172.

Anderson, T. W. (1971). *The statistical analysis of time series*. New York: Wiley.

Bauer, D., Deistler, M., & Scherrer, W. (1997). The analysis of the asymptotic variance of subspace algorithms. *Proceedings of the 11th IFAC Symposium on System Identification* (pp. 1087–1091). Fukuoka, Japan.

Bauer, D. (1998). Some asymptotic theory for the estimation of linear systems using maximum likelihood methods or subspace algorithms. Ph.D. thesis, Tu Wein.

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

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

Desai, U. B., Pal, D., & Kirkpatrick, R. D. (1985). A realization approach to stochastic model reduction. *International Journal of Control*, *42*(4), 821–838.

Fuchs, J. J. (1990). Structure and order estimation of multivariable stochastic processes. *IEEE Transactions on Automatic Control*, *35*, 1338–1341.

Glover, K. (1984). All optimal Hankel norm approximations of linear multivariable systems and their l_∞-error bound. *International Journal of Control*, *39*, 1115–1193.

Golub, G., & Van Loan, C. (1989). *Matrix computations* (2nd ed.), USA: John Hopkins University Press.

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

Jansson, M. (1995). On the performance of subspace methods in system identification and array processing. *Licentiate Thesis*, KTH, Stockholm.

Kowalsky, H. J. (1974). *Vektoranalysis I*. De Gruyter.

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,* vol. 2 (pp. 445–451). Piscataway, NJ: IEEE Service Center.

Lewis, R., & Reinsel, G. (1985). Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis*, *16*, 393–411.

Lindquist, A., & Picci, G. (1985). Realization theory for multivariable stationary gaussian processes. *SIAM Journal on Control and optimization*, *23*, 809–857.

McKelvey, T. (1995). Identification of State-Space Models from Time and Frequency Data. Ph.D. Thesis. Dept. of Electr. Eng., Linköping.

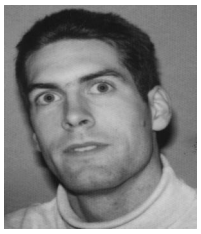
Peternell, K. (1995). Identification of Linear Dynamic Systems by Subspace and Realization-Based Algorithms. Ph.D. thesis, TU Wien.

Van Overschee, P., & DeMoor, B. (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., Ottersten, B., Wahlberg, B., & Ljung, L. (1993). Performance of subspace based state space system identification methods. *Proceedings of the 12th IFAC World Congress*, Vol. 7 (pp. 369–372). Sydney, Australia.

Wahlberg, B., & Jansson, M. (1994). 4sid linear regression. In: *Proceedings of the 33rd Conference on Decision and Control* (pp. 2858–2863). Orlando, USA.



Dietmar Bauer was born in St. Pölten, Austria, on June 21st 1972. He received the masters degree in applied mathematics from the TU Wien in 1995 and his Ph.D. degree in applied mathematics at the TU Wien in 1998. He is currently an assistant at the Institut für Ökonometrie, Operations Research und Systemtheorie, TU Wien. Research interests include parametrization and estimation of linear systems, in particular subspace algorithms.

Manfred Deistler was born in St. Pölten, Austria. He received the Dipl. Ing. (corresponding to M. Sc.) degree in electrical engineering in 1964 and the Ph.D. degree in applied mathematics in 1971, both from the 'Technische Universität' (University of Technology) Vienna. From 1964 to 1966 he worked in industry on control problems. From 1966 to 1968 he had a Ford Foundation scholarship to study econometrics. From 1968 to 1978 he was an assistant and an associated professor at the universities of Regensburg and Bonn, respectively, in econometrics and statistics. Since 1978 he has been full professor of Econometrics at the Technische Universität, Vienna. His main research interests are systems indentification, time series analysis, econometrics and environmental modeling. He is coauthor (with E. J. Hannan) of the book 'The Statistical Theory of Linear Systems' (New York, Wiley 1988). He is on the editorial board of (among others) Journal of Econometrics, Journal of Time Series Analysis, SIAM Journal on Control and Optimization and SIAM Journal on Matrix Analysis and Applications.

Wolfgang Scherrer was born in Feldkirch, Austria, on 20 September 1958. He has received his Mag.rer.nat. (corresponding to a B.Sc. degree), teaching profession for secondary schools in mathematics and physics, (1981) from the Universität Innsbruck, his Dipl.Ing. (corresponding to an M.Sc. degree) in applied mathematics (1986) and his Ph.D. degree in applied Mathematics (1991) from the Technische Universität Wien. From 1982 to 1983 he was working as a teacher at a secondary school in Bregenz. From 1986 to 1996 he was research assistant at the Institut für Ökonometrie, Operations Research und Systemtheorie, Technische Universität Wien. Since 1997, after his qualification as university lecturer, he has been Associate Professor at the above department. His main research areas are: system identification, multivariate linear dynamic systems, errors-in-variables models, econometrics.

PERGAMON

Automatica 36 (2000) 497–509

www.elsevier.com/locate/automatica

Analysis of the asymptotic properties of the MOESP type of subspace algorithms☆

D. Bauer[a], M. Jansson[b,*,1]

[a]Institut für Ökonometrie, Operations Research und Systemtheorie, Technische Universität Wien, Argentinierstr. 8/119, A-1040 Vienna, Austria
[b]Department of Electrical & Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA

Received 8 September 1998; revised 10 May 1999; received in final form 2 August 1999

In this paper consistency and asymptotic normality of the estimates of MOESP type of subspace algorithms are established under fairly general assumptions on the input process.

Abstract

The MOESP type of subspace algorithms are used for the identification of linear, discrete time, finite-dimensional state-space systems. They are based on the geometric structure of covariance matrices and exploit the properties of the state vector extensively. In this paper the asymptotic properties of the algorithms are examined. The main results include consistency and asymptotic normality for the estimates of the system matrices, under suitable assumptions on the noise sequence, the input process and the underlying true system. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Subspace methods; Linear systems; Asymptotic analysis; Identification

1. Introduction

Subspace algorithms are used for the estimation of linear, time-invariant, finite-dimensional, discrete time, state-space systems. They are an alternative to the more classical maximum likelihood and prediction error methods. The main advantages of subspace algorithms are their conceptual simplicity and their numerical properties. The main idea of these algorithms lies in the observation that the predictions of a time series from the whole past of the outputs and possibly the whole series of observed exogenous inputs for different time horizons are a function of the state vector and the future of the exogenous inputs: Every optimal (in the least-squares sense) predictor of the future of the process based on the entire past of the output process and the whole input process is a linear function of the state and the future of the exogenous inputs under appropriate assumptions on the noise and the data generating process. This fact can be used for estimation of the state (cf. Larimore, 1983; Peternell, Scherrer & Deistler, 1996) or the estimation of the linear mapping attaching the predictions to the state vectors and the future of the exogenous inputs (cf. Van Overschee & De Moor, 1994, 1996; Verhaegen, 1994). The statistical properties of the first type of algorithms are clarified to a large extent by Deistler, Peternell and Scherrer (1995), Peternell et al. (1996), Bauer, Deistler and Scherrer (1999) and Bauer (1998). Within the second type of algorithms, the `MOESP` class of algorithms is very popular. `MOESP` has been developed by Verhaegen and coworkers in a series of papers (Verhaegen & Dewilde, 1992a,b; Verhaegen & Dewilde, 1993; Verhaegen, 1994). The numerical properties of the latter algorithms have been investigated thoroughly in these papers. The consistency of this approach has been investigated in Jansson and Wahlberg (1997, 1998). The main conclusion from these papers is that, in general, it is not enough to impose persistence of excitation type of conditions on the exogenous inputs in order to guarantee consistency. However, there are some special cases (see

☆This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor B. Ninness under the direction of Editor T. Söderström.

* Corresponding author.
E-mail addresses: dietmar.bauer@tuwien.ac.at (D. Bauer), magnus.jansson@s3.kth.se (M. Jansson)
[1] On leave from S3-Automatic Control, Royal Institute of Technology (KTH), Stockholm, Sweden.

PII: S0005-1098(99)00174-0

Jansson & Wahlberg, 1998). Asymptotic normality of the estimates of the poles of the transfer function has been established in Viberg, Ottersten, Wahlberg and Ljung (1993). In the current paper the asymptotic properties of the subspace estimates using various conditions on the exogenous inputs are considered. The analysis will center on conditions ensuring consistency of the approach in generic situations, and on asymptotic normality of the system matrix estimates.

The paper is organized as follows: Section 2 introduces the model class used for identification and presents some standard assumptions. Section 3 presents the class of algorithms considered. Section 4 then contains the main results of this paper, namely consistency and asymptotic normality of the system matrix estimates. Section 5 presents some numerical examples and finally Section 6 concludes the paper.

Throughout the paper the following notation will be used: Bold face symbols are used for matrices and vectors, lower case latin and greek symbols are used for scalars. As usual $\rightarrow$ will denote convergence for deterministic quantities and $\rightarrow$ a.s. stands for almost sure convergence of stochastic quantities. $\xrightarrow{\mathrm{d}}$ will denote convergence in distribution. Also the notation $\langle \mathbf{a}_t, \mathbf{b}_t \rangle = (1/T)\sum_{t=1}^{T} \mathbf{a}_t \mathbf{b}_t^{\mathrm{T}}$, where T denotes the sample size, is introduced. Here the initial conditions are such that $\langle \mathbf{a}_t, \mathbf{b}_t \rangle = \langle \mathbf{a}_{t+j}, \mathbf{b}_{t+j} \rangle$ holds for $|j| \leq \alpha + \beta$, where α and β are integers to be specified in the following section. Finally $f_n = o(g_n)$ means $\lim_{n\to\infty} f_n/g_n = 0$.

2. Model set

In this paper the model class is restricted to linear, finite-dimensional, discrete time, time-invariant, state-space systems of the form

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{K}\boldsymbol{\varepsilon}_t, \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{D}\mathbf{u}_t + \mathbf{E}\boldsymbol{\varepsilon}_t, \end{aligned} \tag{1}$$

where $t \in \mathbb{Z}$, $\mathbf{y}_t \in \mathbb{R}^s$ is the s-dimensional observed output, $\boldsymbol{\varepsilon}_t \in \mathbb{R}^s$, denotes the s-dimensional white noise with zero mean and covariance matrix equal to unity. $\mathbf{u}_t \in \mathbb{R}^m$ denotes the m-dimensional exogenous input series, which is assumed to be independent of the noise $\boldsymbol{\varepsilon}_t$ in an appropriate sense to be defined below. Finally, $\mathbf{x}_t \in \mathbb{R}^n$ denotes the n-dimensional state, and $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathbf{C} \in \mathbb{R}^{s\times n}$, $\mathbf{D} \in \mathbb{R}^{s\times m}$, $\mathbf{E} \in \mathbb{R}^{s\times s}$ and $\mathbf{K} \in \mathbb{R}^{n\times s}$ are parameter matrices. The matrix $\mathbf{E}$ is assumed to be lower triangular with strictly positive entries on the main diagonal. In particular, it is thus assumed that $\mathbf{E}$ is nonsingular. Throughout the paper it will also be assumed that the matrix $\mathbf{A}$ is stable, i.e. that $|\lambda_{\max}(\mathbf{A})| < 1$, where $\lambda_{\max}(\mathbf{A})$ denotes an eigenvalue of $\mathbf{A}$ of maximum modulus, and that $|\lambda_{\max}(\mathbf{A} - \mathbf{K}\mathbf{E}^{-1}\mathbf{C})| < 1$. Using the forward-shift operator z, the output $\mathbf{y}_t$ can be written as

$$\begin{aligned} \mathbf{y}_t &= \mathbf{D}\mathbf{u}_t + \sum_{j=1}^{\infty} \mathbf{C}\mathbf{A}^{j-1}\mathbf{B}\mathbf{u}_{t-j} + \mathbf{E}\boldsymbol{\varepsilon}_t + \sum_{j=1}^{\infty} \mathbf{C}\mathbf{A}^{j-1}\mathbf{K}\boldsymbol{\varepsilon}_{t-j} \\ &= \sum_{j=0}^{\infty} \mathbf{L}(j) z^{-j} \mathbf{u}_t + \sum_{j=0}^{\infty} \mathbf{K}(j) z^{-j} \boldsymbol{\varepsilon}_t. \end{aligned}$$

Here $z\mathbf{u}_t = \mathbf{u}_{t+1}$, $z\boldsymbol{\varepsilon}_t = \boldsymbol{\varepsilon}_{t+1}$ and the Markov parameters $\mathbf{K}(j)$ and $\mathbf{L}(j)$ are defined by the above equality. Using this notation the transfer functions $\mathbf{k}(z) = \sum_{j=0}^{\infty} \mathbf{K}(j) z^{-j} = \mathbf{E} + \mathbf{C}(zI - \mathbf{A})^{-1}\mathbf{K}$, and $\mathbf{l}(z) = \sum_{j=0}^{\infty} \mathbf{L}(j) z^{-j} = \mathbf{D} + \mathbf{C}(zI - \mathbf{A})^{-1}\mathbf{B}$ can be defined. Note that the bound on the eigenvalues of $\mathbf{A}$ implies the convergence of the series defining $\mathbf{k}(z)$ on the complement of the open unit disc. The assumption on the eigenvalues of $\mathbf{A} - \mathbf{K}\mathbf{E}^{-1}\mathbf{C}$ implies that the inverse of $\mathbf{k}$ exists and is analytic on the complement of the open unit disc.

Definition 1 (*Standard assumptions*). The process $\mathbf{y}_t, t \in \mathbb{Z}$ is generated by a system of the form (1), where $(\mathbf{A},\mathbf{C})$ is observable and $(\mathbf{A},[\mathbf{B},\mathbf{K}])$ is reachable. The white noise $\boldsymbol{\varepsilon}_t$ is independently identical distributed (i.i.d.) with mean zero and covariance equal to unity. Furthermore, the third- and the fourth-order moments of the noise exist and thus are finite. The input process $\mathbf{u}_t$ is assumed to be independent of the noise.

Note that it is not assumed that the system is reachable from the exogenous inputs only, i.e. that the pair $(\mathbf{A}, \mathbf{B})$ is reachable. Also, note that it is assumed that the matrix $\mathbf{A}$ describes the dynamics of $\mathbf{k}$ and of $\mathbf{l}$, i.e. the matrix $\mathbf{A}$ contains the dynamics due to the exogenous inputs as well as the dynamics due to the noise. Therefore $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ may be a nonminimal realization of $\mathbf{l}$. The assumptions on the white noise are overly strong. However, the authors decided to keep the assumptions on the noise simple, since it will be clear from the exposition, which properties of the noise indeed are needed. The results will obviously hold also for much weaker requirements on the noise. Concerning the inputs, there will be different sets of assumptions for the results on consistency and on the asymptotic normality.

3. The algorithms

In this section a brief presentation of the algorithms considered in this paper will be given. The main fact that is used by subspace algorithms can be formulated as follows: Let $\mathbf{Y}_{t,\beta} = [\mathbf{y}_{t-1}^{\mathrm{T}}, \mathbf{y}_{t-2}^{\mathrm{T}}, \ldots, \mathbf{y}_{t-\beta}^{\mathrm{T}}]^{\mathrm{T}}$ be the vector of the stacked (finite) past of the process and let $\mathbf{Y}_{t,\alpha} = [\mathbf{y}_t^{\mathrm{T}}, \mathbf{y}_{t+1}^{\mathrm{T}}, \ldots, \mathbf{y}_{t+\alpha-1}^{\mathrm{T}}]^{\mathrm{T}}$ be the vector of the stacked (finite) future of the output process. Define $\mathbf{U}_{t,\beta}$ and $\mathbf{U}_{t,\alpha}$ analogously from $\mathbf{u}_t$, and let $\mathbf{P}_{t,\beta} = [\mathbf{Y}_{t,\beta}^{\mathrm{T}}, \mathbf{U}_{t,\beta}^{\mathrm{T}}]^{\mathrm{T}}$. In what follows, it is assumed that $\alpha > n$ and $\beta \geq n$. Furthermore, let $\boldsymbol{\Gamma}_\alpha = [\mathbf{C}^{\mathrm{T}}, \mathbf{A}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}, \ldots, (\mathbf{A}^{\mathrm{T}})^{\alpha-1}\mathbf{C}^{\mathrm{T}}]^{\mathrm{T}}$ denote

the extended observability matrix. Then the following equation can easily be shown to hold:

$$\mathbf{Y}_{t,\alpha} = \mathbf{\Gamma}_\alpha \mathbf{x}_t + \mathbf{\Phi}_\alpha \mathbf{U}_{t,\alpha} + \mathbf{N}_{t,\alpha}. \tag{2}$$

Here $\mathbf{N}_{t,\alpha}$ is equal to the contribution due to the future of the noise and $\mathbf{\Phi}_\alpha$ is defined as

$$\mathbf{\Phi}_\alpha = \begin{bmatrix} \mathbf{D} & 0 & \cdots & 0 \\ \mathbf{CB} & \mathbf{D} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \mathbf{CA}^{\alpha-2}\mathbf{B} & \cdots & \mathbf{CB} & \mathbf{D} \end{bmatrix}.$$

Now the MOESP type of algorithms can be described as follows. The discussion will be restricted to PO-MOESP first (Verhaegen, 1994). At the end of Section 4 also the algorithm denoted by PI-MOESP (Verhaegen and Dewilde, 1993) will be dealt with.

Remark 2. The notation 'MOESP type' is introduced in order to emphasize that the considered class of algorithms is obtained from direct modifications of the original procedure proposed in Verhaegen (1994). From the discussion it will be clear that there are several possibilities to compute the intermediate steps in the estimation algorithm. Different choices lead to variations of the algorithm, which also change the asymptotic properties of the corresponding estimates. In order to avoid dealing with all variants of the original algorithm MOESP, one particular version (which is chosen somewhat arbitrarily) is analyzed. However, the tools used in the analysis below are the basis of the analysis for some of the variants proposed in the literature.

In a first step, define $[\hat{\mathbf{H}}_{\alpha,\beta}, \hat{\mathbf{\Phi}}_\alpha]$ from the regression of $\mathbf{Y}_{t,\alpha}$ onto $\mathbf{P}_{t,\beta}$ and $\mathbf{U}_{t,\alpha}$:

$$[\hat{\mathbf{H}}_{\alpha,\beta}, \hat{\mathbf{\Phi}}_\alpha] = \left\langle \mathbf{Y}_{t,\alpha}, \begin{pmatrix} \mathbf{P}_{t,\beta} \\ \mathbf{U}_{t,\alpha} \end{pmatrix} \right\rangle \left\langle \begin{pmatrix} \mathbf{P}_{t,\beta} \\ \mathbf{U}_{t,\alpha} \end{pmatrix}, \begin{pmatrix} \mathbf{P}_{t,\beta} \\ \mathbf{U}_{t,\alpha} \end{pmatrix} \right\rangle^{-1}.$$

Note that the column space of $\mathbf{H}_{\alpha,\beta}$, the population analog to $\hat{\mathbf{H}}_{\alpha,\beta}$, is contained in the column space of $\mathbf{\Gamma}_\alpha$ (cf. Verhaegen, 1994, see also below).

Remark 3. A more complex method can be obtained easily using a similar approach as proposed in Peternell et al. (1996): Note that $\mathbf{\Phi}_\alpha$ is a lower triangular block Toeplitz matrix, i.e. a matrix of the form

$$\mathbf{L}_\alpha = \begin{bmatrix} \mathbf{L}_0 & 0 & \cdots & 0 \\ \mathbf{L}_1 & \mathbf{L}_0 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ \mathbf{L}_{\alpha-1} & \cdots & \mathbf{L}_1 & \mathbf{L}_0 \end{bmatrix},$$

where $\mathbf{L}_i \in \mathbb{R}^{s \times m}$. These restrictions can be imposed in the regression given above. Introducing more of the structure of the problem, it is hoped to obtain better estimates (see Peternell et al., 1996, for some simulation studies).

In a second step, the singular value decomposition (SVD) of a weighted version of the matrix $\hat{\mathbf{H}}_{\alpha,\beta}$ is used: Let $\hat{\mathbf{W}}_\alpha^+ \hat{\mathbf{H}}_{\alpha,\beta} \hat{\mathbf{W}}_\beta^- = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^{\mathrm{T}} = \hat{\mathbf{U}}_n \hat{\mathbf{\Sigma}}_n \hat{\mathbf{V}}_n^{\mathrm{T}} + \hat{\mathbf{R}}$. The diagonal matrix $\hat{\mathbf{\Sigma}}_n \in \mathbb{R}^{n \times n}$ contains the largest n singular values contained in the diagonal matrix $\hat{\mathbf{\Sigma}}$, and $\hat{\mathbf{U}}_n \in \mathbb{R}^{\alpha s \times n}$ and $\hat{\mathbf{V}}_n \in \mathbb{R}^{\beta(m+s) \times n}$ denote the matrices containing the corresponding singular vectors as columns. Here $\hat{\mathbf{W}}_\alpha^+$ and $\hat{\mathbf{W}}_\beta^-$ are weighting matrices, which are assumed to be nonsingular (for some comments on this, see Jansson & Wahlberg, 1998). Common choices for $\hat{\mathbf{W}}_\alpha^+$ are $\mathbf{I}$, $(\langle \mathbf{Y}_{t,\alpha} - \hat{\mathbf{\Phi}}_\alpha \mathbf{U}_{t,\alpha}, \mathbf{Y}_{t,\alpha} - \hat{\mathbf{\Phi}}_\alpha \mathbf{U}_{t,\alpha} \rangle)^{-1/2}$, $(\langle \mathbf{Y}_{t,\alpha}, \mathbf{Y}_{t,\alpha} \rangle - \langle \mathbf{Y}_{t,\alpha}, \mathbf{U}_{t,\alpha} \rangle \langle \mathbf{U}_{t,\alpha}, \mathbf{U}_{t,\alpha} \rangle^{-1} \langle \mathbf{U}_{t,\alpha}, \mathbf{Y}_{t,\alpha} \rangle)^{-1/2}$ or $\hat{\mathbf{W}}_\alpha^+ = \mathbf{W}_\alpha^+$ lower triangular block Toeplitz and independent of the data. For $\hat{\mathbf{W}}_\beta^-$ the restriction that either $\hat{\mathbf{W}}_\beta^- = (\langle \mathbf{P}_{t,\beta}, \mathbf{P}_{t,\beta} \rangle)^{1/2}$ or

$$\hat{\mathbf{W}}_\beta^- = (\langle \mathbf{P}_{t,\beta}, \mathbf{P}_{t,\beta} \rangle - \langle \mathbf{P}_{t,\beta}, \mathbf{U}_{t,\alpha} \rangle \langle \mathbf{U}_{t,\alpha}, \mathbf{U}_{t,\alpha} \rangle^{-1} \times \langle \mathbf{U}_{t,\alpha}, \mathbf{P}_{t,\beta} \rangle)^{1/2}$$

is imposed. Here $\mathbf{X} = (\mathbf{Y})^{1/2}$ denotes any square root of a positive-definite matrix $\mathbf{Y}$ such that $\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{Y}$. It is straightforward to see, that the particular choice of the square root does not affect the estimates.

In this second step also the order of the system has to be determined. For the construction of order estimation procedures in the context of subspace identification methods, see Bauer (1998). Given the true order n an estimate of $\mathbf{\Gamma}_\alpha$ is defined by $\hat{\mathbf{\Gamma}}_\alpha = (\hat{\mathbf{W}}_\alpha^+)^{-1} \hat{\mathbf{U}}_n$. Then an estimate of $(\mathbf{A},\mathbf{C})$ is obtained by using the shift-invariance property of $\mathbf{\Gamma}_\alpha$: Note that $\mathbf{\Gamma}_{\alpha-1}\mathbf{A} = \mathbf{\Gamma}_\alpha^\uparrow$, where $\mathbf{\Gamma}_\alpha^\uparrow$ is obtained from $\mathbf{\Gamma}_\alpha$ by omitting the first block row, i.e. $\mathbf{\Gamma}_\alpha^\uparrow = [\mathbf{A}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}, (\mathbf{A}^{\mathrm{T}})^2\mathbf{C}^{\mathrm{T}}, \ldots, (\mathbf{A}^{\mathrm{T}})^{\alpha-1}\mathbf{C}^{\mathrm{T}}]^{\mathrm{T}}$. Replacing true quantities with estimates, the least-squares estimate $\hat{\mathbf{A}}_T = \hat{\mathbf{\Gamma}}_{\alpha-1}^\dagger \hat{\mathbf{\Gamma}}_\alpha^\uparrow$ is obtained. Here $\hat{\mathbf{\Gamma}}_{\alpha-1}^\dagger$ denotes the Moore–Penrose pseudoinverse of $\hat{\mathbf{\Gamma}}_{\alpha-1}$. $\hat{\mathbf{C}}_T$ is estimated as the first block row of $\hat{\mathbf{\Gamma}}_\alpha$. There have been several different proposals on how to estimate the pair $(\mathbf{A},\mathbf{C})$ from an estimate of $\mathbf{\Gamma}_\alpha$ (see e.g. Viberg, Wahlberg & Ottersten, 1997; Lovera, Falcetti & Bittanti, 1998). All these methods basically lead to a (explicit or implicit) definition of a mapping attaching estimates $(\hat{\mathbf{A}}_T, \hat{\mathbf{C}}_T)$ to the estimate $\hat{\mathbf{\Gamma}}_\alpha$. It will be clear from the discussion in Section 4, what the 'key properties' of these mappings are, in order to ensure consistency and asymptotic normality.

In the remaining step, the estimate $\hat{\mathbf{\Gamma}}_\alpha$ is used to obtain estimates of $\mathbf{B}$ and $\mathbf{D}$ from Eq. (2). Note that $\mathbf{\Phi}_\alpha$ is a linear function of $\mathrm{vec}[\mathbf{B},\mathbf{D}]$, i.e. $\mathrm{vec}\,\mathbf{\Phi}_\alpha = \mathbf{L}_{\mathbf{B},\mathbf{D}}\,\mathrm{vec}[\mathbf{B},\mathbf{D}]$, where $\mathbf{L}_{\mathbf{B},\mathbf{D}}$ depends only on $\mathbf{\Gamma}_\alpha$. Let $\mathbf{\Gamma}_\alpha^\perp \in \mathbb{R}^{\alpha s \times (\alpha s - n)}$ be a full-rank matrix, such that $\mathbf{\Gamma}_\alpha^T \mathbf{\Gamma}_\alpha^\perp = 0$, i.e. the columns of $\mathbf{\Gamma}_\alpha^\perp$ span the orthogonal complement of the space spanned by the columns of $\mathbf{\Gamma}_\alpha$. Then from Eq. (2) it follows that $(\mathbf{\Gamma}_\alpha^\perp)^{\mathrm{T}}\mathbf{Y}_{t,\alpha} = (\mathbf{\Gamma}_\alpha^\perp)^{\mathrm{T}}\mathbf{\Phi}_\alpha \mathbf{U}_{t,\alpha} + (\mathbf{\Gamma}_\alpha^\perp)^{\mathrm{T}}\mathbf{N}_{t,\alpha}$. For the

estimation it is tempting to replace the true quantity $\mathbf{\Gamma}_\alpha^\perp$ with a corresponding estimate $\hat{\mathbf{\Gamma}}_\alpha^\perp \in \mathbb{R}^{\alpha s \times (\alpha s - n)}$, such that $\hat{\mathbf{\Gamma}}_\alpha^{\mathrm{T}} \hat{\mathbf{\Gamma}}_\alpha^\perp = 0$. In this paper the choice $\hat{\mathbf{\Gamma}}_\alpha^\perp = (\hat{\mathbf{W}}_\alpha^+)^{\mathrm{T}} \hat{\mathbf{U}}_2$ is used, where $\hat{\mathbf{U}}_2 \in \mathbb{R}^{\alpha \times (\alpha s - n)}$ is an orthonormal matrix spanning the orthogonal complement of the space spanned by the columns of $\hat{\mathbf{U}}_n$. (Again, the choice for $\hat{\mathbf{\Gamma}}_\alpha^\perp$ made here is somewhat arbitrary.) Therefore, $[\mathbf{B}, \mathbf{D}]$ can be estimated as follows.[2]

$$\begin{aligned}\mathrm{vec}[\hat{\mathbf{B}}_T, \hat{\mathbf{D}}_T] &= \arg\min_{\mathbf{B},\mathbf{D}} \|\mathrm{vec}[(\hat{\mathbf{\Gamma}}_\alpha^\perp)^{\mathrm{T}} \langle \mathbf{Y}_{t,\alpha}, \mathbf{U}_{t,\alpha} \rangle \langle \mathbf{U}_{t,\alpha}, \mathbf{U}_{t,\alpha} \rangle^{-1}] \\ &\quad - (\mathbf{I} \otimes [\hat{\mathbf{\Gamma}}_\alpha^\perp]^{\mathrm{T}}) \hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}} \mathrm{vec}[\mathbf{B},\mathbf{D}]\|^2 \\ &= [\hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}}^{\mathrm{T}} (\mathbf{I} \otimes \hat{\mathbf{\Gamma}}_\alpha^\perp (\hat{\mathbf{\Gamma}}_\alpha^\perp)^{\mathrm{T}}) \hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}}]^{-1} \hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}}^{\mathrm{T}} \\ &\quad \mathrm{vec}[\hat{\mathbf{\Gamma}}_\alpha^\perp (\hat{\mathbf{\Gamma}}_\alpha^\perp)^{\mathrm{T}} \langle \mathbf{Y}_{t,\alpha}, \mathbf{U}_{t,\alpha} \rangle \langle \mathbf{U}_{t,\alpha}, \mathbf{U}_{t,\alpha} \rangle^{-1}], \quad (3)\end{aligned}$$

where $\otimes$ denotes the Kronecker product and $\|\cdot\|$ the Euclidean norm. $\hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}}$ denotes the matrix $\mathbf{L}_{\mathbf{B},\mathbf{D}}$, where the estimate $\hat{\mathbf{\Gamma}}_\alpha$ is used rather than the matrix $\mathbf{\Gamma}_\alpha$. Note that the estimates of $\mathbf{B}$ and $\mathbf{D}$ depend on the choice of $\hat{\mathbf{\Gamma}}_\alpha^\perp$. Also note that instead of using the estimate $\hat{\mathbf{\Gamma}}_\alpha$ in the definition of $\hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}}$ the estimates $\hat{\mathbf{A}}_T$ and $\hat{\mathbf{C}}_T$ could be used. The above approach of estimating $\mathbf{B}$ and $\mathbf{D}$ can be given an instrumental variable (IV) interpretation. Since $\mathbf{u}_t$ and $\boldsymbol{\varepsilon}_t$ are assumed to be uncorrelated, the IV vector $\xi_t = \langle \mathbf{U}_{t,\alpha}, \mathbf{U}_{t,\alpha} \rangle^{-1} \mathbf{U}_{t,\alpha}$ can be used to correlate out the noise in the equation $(\mathbf{\Gamma}_\alpha^\perp)^{\mathrm{T}} \mathbf{N}_{t,\alpha} = (\mathbf{\Gamma}_\alpha^\perp)^{\mathrm{T}} \mathbf{Y}_{t,\alpha} - (\mathbf{\Gamma}_\alpha^\perp)^{\mathrm{T}} \mathbf{\Phi}_\alpha \mathbf{U}_{t,\alpha}$. Indeed, minimizing the IV criterion $\sum_{t=\beta+1}^{T-\alpha} \|((\hat{\mathbf{\Gamma}}_\alpha^\perp)^{\mathrm{T}} \mathbf{Y}_{t,\alpha} - (\mathbf{U}_{t,\alpha}^{\mathrm{T}} \otimes [\hat{\mathbf{\Gamma}}_\alpha^\perp]^{\mathrm{T}}) \hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}} \mathrm{vec}[\mathbf{B},\mathbf{D}]) \xi_t^{\mathrm{T}}\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm, with respect to $\mathbf{B}$ and $\mathbf{D}$ leads to exactly the same solution as given in (3). This is the original MOESP procedure proposed in Verhaegen (1994) Another, maybe more natural, choice of the IV vector is $\boldsymbol{\xi}_t = (\langle \mathbf{U}_{t,\alpha}, \mathbf{U}_{t,\alpha} \rangle)^{-1/2} \mathbf{U}_{t,\alpha}$. However, the estimate given in (3) is chosen in the analysis that follows. All the alternative approaches mentioned above are easily analyzed using the tools presented in this paper.

No attempt will be made to estimate the remaining matrices $\mathbf{E}$ and $\mathbf{K}$, the discussion rather concentrates on the estimation of $\mathbf{l}$. This is done mainly for two reasons: First, most of the proposed methods for the estimation of $\mathbf{E}$ and $\mathbf{K}$ result in consistent estimates only for $\beta \to \infty$. In this case, the analysis becomes much more complex, and the assumptions on the input sequence have to be adapted. Second, the original MOESP algorithm (Verhaegen, 1994) was developed for the estimation of $\mathbf{l}$ only. Therefore, the analysis is restricted to the estimation of $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$. In the next section, the asymptotic properties of the estimates $(\hat{\mathbf{A}}_T, \hat{\mathbf{B}}_T, \hat{\mathbf{C}}_T, \hat{\mathbf{D}}_T)$ obtained by this algorithm are investigated.

[2] Note that $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^{\mathrm{T}} \otimes \mathbf{A})\mathrm{vec}(\mathbf{B})$ for matrices of compatible dimensions.

4. Asymptotic properties

The first part of this section will focus on the question of consistency of the estimates. There will be two different concepts concerning the consistency, depending on whether the estimate of the transfer function is concerned, or whether the convergence of the system matrix estimates is investigated. From the description of the algorithm it can be seen that the system matrix estimates are a nonlinear function of the sample covariances of the joint process $\mathbf{z}_t = [\mathbf{y}_t^{\mathrm{T}}, \mathbf{u}_t^{\mathrm{T}}]^{\mathrm{T}}$ up to lag $\alpha + \beta - 1$. Up to now, no assumptions on the input process have been introduced, except for the independence of the noise. The assumptions needed for the consistency result are as follows:

Definition 4 (*Weak assumptions on the inputs*). The process $\mathbf{u}_t$ is pseudostationary, fulfilling (where these equations define $\boldsymbol{\mu}$ and $\gamma_{u,u}(j)$):

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{u}_t = \boldsymbol{\mu},$$

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-j} \mathbf{u}_t \mathbf{u}_{t+j}^{\mathrm{T}} = \gamma_{u,u}(j)$$

for $j \geq 0$. Furthermore, the input process is assumed to be persistently exciting of order $\alpha + \beta$, i.e. the block Toeplitz matrix

$$\mathbf{\Gamma}_{u,u} = \begin{bmatrix} \gamma_{u,u}(0) & \gamma_{u,u}(1) & \cdots & \gamma_{u,u}(\alpha+\beta-1) \\ \gamma_{u,u}(-1) & \gamma_{u,u}(0) & \ddots & \\ \ddots & & \ddots & \gamma_{u,u}(1) \\ \gamma_{u,u}(1-\alpha-\beta) & \cdots & \gamma_{u,u}(-1) & \gamma_{u,u}(0) \end{bmatrix}$$

is of full rank $(\alpha + \beta)m$. From these assumptions it follows (cf. e.g. Hannan & Deistler, 1988, Theorem 4.1.1) that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-j} \mathbf{u}_t \boldsymbol{\varepsilon}_{t+j}^{\mathrm{T}} = 0 = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1+j}^{\mathrm{T}} \mathbf{u}_t \boldsymbol{\varepsilon}_{t-j}^{\mathrm{T}}, \quad j \geq 0$$

In this case, $\mathbf{u}_t$ and $\boldsymbol{\varepsilon}_t$ will be called uncorrelated with slight abuse of terminology.

Let $\gamma_{z,z}(j) = \mathbb{E}\mathbf{z}_t \mathbf{z}_{t+j}^{\mathrm{T}}$, where $\mathbb{E}$ has to be interpreted as expectation for random variables and as a limit of the sample covariances for expressions involving $\mathbf{u}_t$. It is well known (see e.g. Hannan & Deistler, 1988, Chapter 4.1) that the given assumptions are sufficient for these limits to exist and also for the almost sure convergence of the sample covariances $\hat{\gamma}_{z,z}(j) = (1/T)\sum_{t=1}^{T-j} \mathbf{z}_t \mathbf{z}_{t+j}^{\mathrm{T}}$. This convergence result will be the basis for the consistency proof. Note that this result holds, e.g. if the input process is a trajectory generated by an i.i.d. sequence of random variables with finite variance, which is filtered using a linear filter $\mathbf{k}_u(z) = \sum_{j=0}^{\infty} \mathbf{K}_u(j) z^{-j}$, having the property that $\sum_{j=0}^{\infty} \|\mathbf{K}_u(j)\| < \infty$. Additionally, also a term of the form

$\sum_{l=1}^{h} \mathbf{c}_l e^{i\lambda_l t}$ may be added, where $\mathbf{c}_l \in \mathbb{C}^m$, $-\pi \le \lambda_l < \pi$, $1 \le l \le h$ are such that the corresponding process is real valued. Note that this additional term includes a mean value term as well as sinusoids. Thus the class of input signals for which the given assumptions hold seems to include many typical situations. However, the authors want to emphasize that the examples given are by no means the only signals satisfying the weak assumptions.

A central matrix in the evaluations is the matrix $\hat{\mathbf{W}}_\alpha^+ \hat{\mathbf{H}}_{\alpha,\beta} \hat{\mathbf{W}}_\beta^-$, on which the SVD is performed. It will be shown below that for the choices of $\hat{\mathbf{W}}_\alpha^+$ and $\hat{\mathbf{W}}_\beta^-$ given in Section 3 the weighting matrices converge to the corresponding matrices where sample estimates are replaced by population moments. The same is true for the estimates obtained in the regression, i.e. $\hat{\mathbf{H}}_{\alpha,\beta} \to \mathbf{H}_{\alpha,\beta}$ a.s., where $\mathbf{H}_{\alpha,\beta} = \mathbf{\Gamma}_\alpha [\mathbb{E}\mathbf{x}_t (\mathbf{P}_{t,\beta}^{\Pi})^{\mathrm{T}}][\mathbb{E}\mathbf{P}_{t,\beta}^{\Pi}(\mathbf{P}_{t,\beta}^{\Pi})^{\mathrm{T}}]^{-1}$. Here $\mathbf{P}_{t,\beta}^{\Pi}$ denotes the residual from a regression of $\mathbf{P}_{t,\beta}$ onto $\mathbf{U}_{t,\alpha}$, i.e., $\mathbf{P}_{t,\beta}^{\Pi} = \mathbf{P}_{t,\beta} - \mathbb{E}[\mathbf{P}_{t,\beta}, \mathbf{U}_{t,\alpha}]\mathbb{E}[\mathbf{U}_{t,\alpha}, \mathbf{U}_{t,\alpha}]^{-1}\mathbf{U}_{t,\alpha}$. Using the persistence of excitation assumption on the input sequence, it follows that the estimated covariance matrix, $\hat{\mathbf{\Pi}}_\beta$ say, of $\mathbf{P}_{t,\beta}^{\Pi}$ converges to the population analog. Moreover, the assumptions on the noise ensure that $\hat{\mathbf{\Pi}}_\beta$ is nonsingular almost sure for T large enough. Thus, in order to assess the rank of $\mathbf{H}_{\alpha,\beta}$ only the rank of $\mathbb{E}\mathbf{x}_t(\mathbf{P}_{t,\beta}^{\Pi})^{\mathrm{T}}$ has to be considered. It is easy to see that $\mathbb{E}\mathbf{x}_t(\mathbf{P}_{t,\beta}^{\Pi})^{\mathrm{T}}$ is of full rank iff the following matrix is of full rank:

$$\mathscr{R}_{\alpha,\beta} = \mathbb{E}\begin{pmatrix} \mathbf{x}_t \\ \mathbf{U}_{t,\alpha} \end{pmatrix}\begin{pmatrix} \mathbf{P}_{t,\beta} \\ \mathbf{U}_{t,\alpha} \end{pmatrix}^{\mathrm{T}}.$$

From Jansson and Wahlberg (1997) it follows that the weak assumptions on the input sequence are not sufficient for this matrix to be of full rank. Jansson and Wahlberg (1997) actually constructs an ARMA input process (which is persistent of any order) and a system, such that the rank of $\mathscr{R}_{\alpha,\beta}$ is smaller than the order of the system. It follows from the arguments in Peternell et al. (1996) that the rank of $\mathscr{R}_{\alpha,\beta}$ will be equal to the order of the true system if α and β are taken sufficiently large (some sufficient conditions for the rank constraint to hold are given in Jansson & Wahlberg, 1998). Another reference in this respect is Chui (1997). However, in this paper another route will be followed, by showing that the set of transfer functions for which the full rank condition is not satisfied is 'thin'. Here sets of transfer functions are equipped with the pointwise topology (see e.g., Hannan & Deistler, 1988), sets of finite-dimensional vectors with the Euclidean metric. Let $\bar{S}_n$ denote the set of all system matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{K})$, where the state dimension is equal to n, $\mathbf{E}$ is lower triangular with positive entries on the main diagonal, and $\mathbf{A}$ and $\mathbf{A} - \mathbf{K}\mathbf{E}^{-1}\mathbf{C}$ are stable. Let $S_n \subset \bar{S}_n$ denote the subset of $\bar{S}_n$, where additionally $(\mathbf{A}, \mathbf{C})$ is observable and $(\mathbf{A}, [\mathbf{B}, \mathbf{K}])$ is reachable. Note that $\bar{S}_n$ is not the closure of S_n in the corresponding Euclidean space, since the stability, the strict minimum-phase condition and the nonsingularity of $\mathbf{E}$ is maintained also for the systems in $\bar{S}_n$. Let π denote the mapping attaching transfer functions to system matrices. Finally, let $M_n = \pi(S_n)$ denote the set of all pairs of transfer functions corresponding to S_n. Then a 'thin' set is a set whose complement in M_n is open and dense in M_n.

Lemma 5. *The set $M_n(\mathbf{u}_t, \alpha, \beta) \subset M_n$ of pairs of transfer functions $(\mathbf{k}, \mathbf{l}) \in M_n$, such that the corresponding matrix $\mathscr{R}_{\alpha,\beta}$ is of full rank $(n + \alpha m)$, is open and dense in M_n.*

Proof. The proof uses similar techniques as have been used in Bauer et al. (1999): First, it will be shown that given the sequence of population covariances of the input sequence, $\gamma_{u,u}(j)$, the finite-dimensional matrix $\mathscr{R}_{\alpha,\beta}$ is an analytic function of the system matrices on $\bar{S}_n$, which is an open and pathwise connected set in the embedding Euclidean space. Thus the determinant of $\mathscr{R}_{\alpha,\beta}\mathscr{R}_{\alpha,\beta}^{\mathrm{T}} \in \mathbb{R}^{(n+\alpha m)\times(n+\alpha m)}$ is an analytic function of the system matrix entries. This shows that the determinant is either identically zero on $\bar{S}_n$, or generically nonzero. The existence of a single pair of transfer functions such that $\mathscr{R}_{\alpha,\beta}$ is of full rank then proves the lemma.

Thus consider the entries in $\mathscr{R}_{\alpha,\beta}$ more closely. Four types of entries have to be considered: $\mathbb{E}\mathbf{x}_t\mathbf{u}_{t+j}^{\mathrm{T}}, \mathbb{E}\mathbf{x}_t\mathbf{y}_{t+j}^{\mathrm{T}}$, $\mathbb{E}\mathbf{y}_t\mathbf{u}_{t+j}^{\mathrm{T}}$ and $\mathbb{E}\mathbf{u}_t\mathbf{u}_{t+j}^{\mathrm{T}}$. Here $|j| \le \alpha + \beta - 1$ in all cases. $\mathbb{E}\mathbf{u}_t\mathbf{u}_{t+j}^{\mathrm{T}}$ is independent of the system matrices and hence is an analytic function of the entries in $(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\mathbf{E},\mathbf{K}) \in \bar{S}_n$. Next consider $\mathbb{E}\mathbf{x}_t\mathbf{u}_{t+j}^{\mathrm{T}}$. Since $\boldsymbol{\varepsilon}_t$ and $\mathbf{u}_t$ are assumed to be completely uncorrelated, one obtains: $\mathbb{E}\mathbf{x}_t\mathbf{u}_{t+j}^{\mathrm{T}} = \mathbb{E}\sum_{i=1}^{\infty}\mathbf{A}^{i-1}\mathbf{B}\mathbf{u}_{t-i}\mathbf{u}_{t+j}^{\mathrm{T}}$, which is an analytic function of the entries in $\mathbf{A}$ and $\mathbf{B}$ due to the assumed stability of $\mathbf{A}$. This also shows the analyticity of $\mathbb{E}\mathbf{y}_t\mathbf{u}_{t+j}^{\mathrm{T}}$. Thus, it remains to show the result for terms of the form $\mathbb{E}\mathbf{x}_t\mathbf{y}_{t+j}^{\mathrm{T}} = \mathbb{E}\mathbf{x}_t(\mathbf{C}\mathbf{x}_{t+j} + \mathbf{D}\mathbf{u}_{t+j} + \mathbf{E}\boldsymbol{\varepsilon}_{t+j})^{\mathrm{T}}$. Now $\mathbb{E}\mathbf{x}_t\mathbf{u}_{t+j}^{\mathrm{T}}$ has been treated already, $\mathbb{E}\mathbf{x}_t\boldsymbol{\varepsilon}_{t-j}^{\mathrm{T}} = \mathbf{A}^{j-1}\mathbf{K}\delta_{j>0}$, where $\delta_{j>0}$ is equal to 1 for $j > 0$ and zero else. Thus the remaining term is equal to

$$\begin{aligned}\mathbb{E}\mathbf{x}_t\mathbf{x}_{t+j}^{\mathrm{T}} &= \mathbb{E}\left[\sum_{i=1}^{\infty}\mathbf{A}^{i-1}\mathbf{B}\mathbf{u}_{t-i} + \mathbf{A}^{i-1}\mathbf{K}\boldsymbol{\varepsilon}_{t-i}\right] \\ &\quad\times\left[\sum_{i=1}^{\infty}\mathbf{A}^{i-1}(\mathbf{B}\mathbf{u}_{t+j-i} + \mathbf{K}\boldsymbol{\varepsilon}_{t+j-i})\right]^{\mathrm{T}} \\ &= \sum_{r,s=1}^{\infty}[\mathbf{A}^{r-1}\mathbf{B}\mathbb{E}\mathbf{u}_{t-r}\mathbf{u}_{t+j-s}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}(\mathbf{A}^{s-1})^{\mathrm{T}} \\ &\quad + \mathbf{A}^{r-1}\mathbf{K}\mathbb{E}\boldsymbol{\varepsilon}_{t-r}\boldsymbol{\varepsilon}_{t+j-s}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}(\mathbf{A}^{s-1})^{\mathrm{T}}]\end{aligned}$$

due to the assumed orthogonality of $\mathbf{u}_t$ and $\boldsymbol{\varepsilon}_s$. Now again, the analyticity of this expression as a function of the entries in $\mathbf{A}$ follows from the stability of $\mathbf{A}$. Therefore, each entry in $\mathscr{R}_{\alpha,\beta}$ is an analytic function of the entries in the system matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{K}) \in \bar{S}_n$. As has been stated already, the lemma then follows from the existence of one system with the property that the corresponding matrix $\mathscr{R}_{\alpha,\beta}$ has full rank. This follows, for example, by

choosing $\mathbf{l} = 0$ and $\mathbf{k}$ as an arbitrary transfer function of order n. This completes the proof. □

As has been mentioned before, this is the strongest result one can hope to achieve, since the results in Jansson and Wahlberg (1997) show that it is possible to construct examples where consistency fails. Also note that the generic set on which the consistency condition is fulfilled depends on the (noncentral) covariances of the input process and on the choice of the truncation indices α,β.

Concerning the consistency, two different results will be presented: First, consistency of the transfer functions in the sense of the pointwise topology for the true pair $(\mathbf{k}_0, \mathbf{l}_0) \in M_n(\mathbf{u}_t, \alpha, \beta)$ will be stated. Then, under somewhat stronger assumptions, the consistency for the system matrix estimates is established. A central matrix in the derivation of the results will prove to be $\hat{\mathbf{X}}_{\alpha,\beta} = \hat{\mathbf{W}}_\alpha^+ \hat{\mathbf{H}}_{\alpha,\beta} \hat{\mathbf{W}}_\beta^- (\hat{\mathbf{W}}_\beta^-)^{\mathrm{T}} \hat{\mathbf{H}}_{\alpha,\beta}^{\mathrm{T}} (\hat{\mathbf{W}}_\alpha^+)^{\mathrm{T}}$. Note that $\hat{\mathbf{U}}_n$ is the matrix of the eigenvectors to the largest n eigenvalues of this matrix. Thus the properties of the eigenvalue decomposition of this matrix will be crucial. The analysis is analogous to the presentation given in Bauer et al. (1999).

First note that, due to the convergence of the sample covariances to the population counterparts, the matrices $\hat{\mathbf{W}}_\alpha^+ = (\langle \mathbf{Y}_{t,\alpha} - \hat{\mathbf{\Phi}}_\alpha \mathbf{U}_{t,\alpha}, \mathbf{Y}_{t,\alpha} - \hat{\mathbf{\Phi}}_\alpha \mathbf{U}_{t,\alpha} \rangle)^{-1/2}$ and $\hat{\mathbf{W}}_\beta^- = (\langle \mathbf{P}_{t,\beta}, \mathbf{P}_{t,\beta} \rangle)^{1/2}$ are easily seen to be consistent for some limits $\mathbf{W}_\alpha^+$ and $\mathbf{W}_\beta^-$, using, e.g. the Cholesky decomposition to define a unique square root. The assumptions on the noise and the nonsingularity of $\mathbf{E}$ ensure the invertibility of these matrices and thus of $\hat{\mathbf{W}}_\alpha^+$ and $\hat{\mathbf{W}}_\beta^-$ a.s. for T large enough. Similarily the a.s. convergence of $\hat{\mathbf{H}}_{\alpha,\beta}$ to $\mathbf{H}_{\alpha,\beta}$ follows. Thus the a.s. convergence of $\hat{\mathbf{X}}_{\alpha,\beta} = \hat{\mathbf{W}}_\alpha^+ \hat{\mathbf{H}}_{\alpha,\beta} \hat{\mathbf{W}}_\beta^- (\hat{\mathbf{W}}_\alpha^+ \hat{\mathbf{H}}_{\alpha,\beta} \hat{\mathbf{W}}_\beta^-)^{\mathrm{T}}$ to $\mathbf{X}_{\alpha,\beta} = \mathbf{W}_\alpha^+ \mathbf{H}_{\alpha,\beta} \mathbf{W}_\beta^- (\mathbf{W}_\alpha^+ \mathbf{H}_{\alpha,\beta} \mathbf{W}_\beta^-)^{\mathrm{T}}$ is obtained. Now the next lemma can be formulated.

Lemma 6. *Define the set* $M_n^+(\mathbf{u}_t, \alpha, \beta) \subset M_n(\mathbf{u}_t, \alpha, \beta)$ *as follows*: $(\mathbf{k}, \mathbf{l}) \in M_n^+(\mathbf{u}_t, \alpha, \beta)$ *if the corresponding matrix* $\mathbf{X}_{\alpha,\beta}$ *has exactly n distinct nonzero eigenvalues. Then* $M_n^+(\mathbf{u}_t, \alpha, \beta)$ *is open and dense in* $M_n(\mathbf{u}_t, \alpha, \beta)$ *and thus also in* M_n.

Proof. Consider the matrix $\mathbf{X}_{\alpha,\beta}$. The matrix contains products of the following three matrices: $\mathbf{H}_{\alpha,\beta}, \mathbf{W}_\alpha^+, \mathbf{W}_\beta^-$. It is straightforward to prove that all these matrices are analytic functions of $\mathbb{E}\mathbf{z}_t \mathbf{z}_{t+j}^{\mathrm{T}}$. These terms have been shown to be analytic functions of the entries in the system matrices in the proof of Lemma 5. Then using standard arguments for analytic functions (see e.g. Dieudonné, 1969) the analyticity of the entries of $\mathbf{X}_{\alpha,\beta}$ as a function of the entries of $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{K}) \in \bar{S}_n$, which is open and pathwise connected, follows.

Next note that $\mathbf{H}_{\alpha,\beta} = \mathbf{\Gamma}_\alpha \mathbf{\Xi}_{\alpha,\beta}$, where this equation defines $\mathbf{\Xi}_{\alpha,\beta}$. Then the nonzero eigenvalues of $\mathbf{X}_{\alpha,\beta}$ coincide with the eigenvalues of

$$\mathbf{\Lambda}_n = [\mathbf{\Gamma}_\alpha^{\mathrm{T}} (\mathbf{W}_\alpha^+)^{\mathrm{T}} \mathbf{W}_\alpha^+ \mathbf{\Gamma}_\alpha][\mathbf{\Xi}_{\alpha,\beta} \mathbf{W}_\beta^- (\mathbf{W}_\beta^-)^{\mathrm{T}} \mathbf{\Xi}_{\alpha,\beta}^{\mathrm{T}}] \in \mathbb{R}^{n \times n}.$$

It is straightforward to see that also the entries of $\mathbf{\Lambda}_n$ are analytic functions of the entries of $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{K})$. Thus it is sufficient to show that the property that the eigenvalues of $\mathbf{\Lambda}_n$ are distinct is generic in M_n. In order to show this, as in Bauer et al. (1999) consider the Sylvester matrix associated with the characteristic polynomial $\det(\mathbf{\Lambda}_n - \lambda \mathbf{I})$ and its derivative with respect to λ. The determinant of the Sylvester matrix is nonzero if and only if all eigenvalues are distinct. Again, the determinant of the Sylvester matrix is analytic in the entries of $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{K}) \in \bar{S}_n$. Thus it is sufficient to show that the set $M_n^+(\mathbf{u}_t, \alpha, \beta)$ is nonempty for each n.

However, this can be shown by using induction and a continuity argument: For $n = 1$ the conjecture is obvious. Thus assume that there exists a pair $(\mathbf{k}_0, \mathbf{l}_0)$ of order $n-1$ such that the corresponding matrix $\mathbf{\Lambda}_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$ has $n-1$ distinct nonzero eigenvalues. It then follows that for any non-minimal realization $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{K}) \in \bar{S}_n$ of $(\mathbf{k}_0, \mathbf{l}_0)$, the corresponding matrix $\mathbf{\Lambda}_n$ has $n-1$ distinct eigenvalues plus a zero eigenvalue. The proof of Lemma 5 shows that the set of all realizations of pairs of transfer functions $(\mathbf{k}, \mathbf{l}) \in M_n(\mathbf{u}_t, \alpha, \beta)$ is open and dense in $\bar{S}_n$. Thus there exist realizations arbitrarily close to every realization of $(\mathbf{k}_0, \mathbf{l}_0)$ in $\bar{S}_n$. The continuity of the mapping attaching the matrices $\mathbf{\Lambda}_n$ to realizations in $\bar{S}_n$ together with the continuity of the eigenvalues (see e.g. Chatelin, 1983) concludes the proof. □

For a comprehensive discussion of the results, also the following technical lemma will be useful, which can be found e.g. in the textbook (Chatelin, 1983). Similar results may be found in Anderson (1963).

Lemma 7. *Let* $\mathbf{T}_n$ *be a sequence of symmetric matrices converging to* $\mathbf{T}_0$, *where the rank of* $\mathbf{T}_0$ *be denoted with r. Then the following statements hold*:

(i) *The set of the r largest eigenvalues* $\{\lambda_{i,n}, i = 1, \ldots, r\}$ *of* $\mathbf{T}_n$ *converges to the set of nonzero eigenvalues of* $\mathbf{T}_0$, $\{\lambda_i, i = 1, \ldots, r\}$. *Here convergence is with respect to the Hausdorff metric induced by the Euclidean metric on* $\mathbb{R}$ (*for a definition of the Hausdorff metric see e.g. Chatelin*, 1983). *For each i, the span of all eigenspaces of* $\mathbf{T}_n$ *corresponding to eigenvalues* $\lambda_{i,n}$ *converging to* λ_i, *converges to the eigenspace of* $\mathbf{T}_0$ *corresponding to* λ_i. *Convergence takes place in the gap metric. For a definition of the gap metric, see e.g. Chatelin* (1983).

(ii) *For an eigenvalue* λ_i *of* $\mathbf{T}_0$ *of multiplicity equal to one, the eigenvalue* $\lambda_{i,n}$ *of* $\mathbf{T}_n$, *where* $\lambda_{i,n} \to \lambda_i$, *fulfills the following equation* (*for* $\|\mathbf{T}_n - \mathbf{T}_0\|$ *small*):

$$\lambda_{i,n} = \lambda_i + \mathbf{u}_i^{\mathrm{T}} (\mathbf{T}_n - \mathbf{T}_0) \mathbf{u}_i + o(\|\mathbf{T}_n - \mathbf{T}_0\|). \tag{4}$$

Here $\mathbf{u}_i$ *denotes an eigenvector of length one of* $\mathbf{T}_0$ *corresponding to the eigenvalue* λ_i. *Furthermore there exists a sequence of eigenvectors* $\mathbf{u}_{i,n}$ *of length one of* $\mathbf{T}_n$, *such that*

$$\mathbf{u}_{i,n} = \mathbf{u}_i + \sum_{j\,:\,\lambda_j \neq \lambda_i} \frac{\mathbf{u}_j^{\mathrm{T}}(\mathbf{T}_n - \mathbf{T}_0)\mathbf{u}_i}{\lambda_i - \lambda_j}\mathbf{u}_j + o(\|\mathbf{T}_n - \mathbf{T}_0\|). \tag{5}$$

For a proof see e.g. Chatelin (1983). Note that point (i) ensures the convergence of the eigenvalues and eigenspaces and thus will be the interesting result for the consistency results, whereas point (ii) refers to a linearization of the eigenvalues and eigenvectors and thus is useful in the derivation of a central limit theorem, which will be discussed at the end of this section.

In order to formulate consistency results for the system matrix estimates, the limiting realization of the true transfer function, $\mathbf{l}_0$, has to be determined. This realization corresponds to the eigenvalue decomposition of $\mathbf{X}_{\alpha,\beta}$: Choosing $\mathbf{\Gamma}_\alpha = (\mathbf{W}_\alpha^+)^{-1}\mathbf{U}_n$, where the matrix $\mathbf{U}_n \in \mathbb{R}^{\alpha s \times n}$ contains the eigenvectors of $\mathbf{X}_{\alpha,\beta}$ corresponding to the n largest eigenvalues ordered in size as columns, fixes a state basis and thus a particular realization of the true transfer function. This particular realization (which depends only on the true system, but not on the particular realization of the noise and the inputs) will be denoted with $(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0, \mathbf{D}_0)$ in the following. Note, however, that the eigenvalue decomposition is nonunique even for distinct eigenvalues due to the choice of the orientation of the eigenvectors. Lemma 7 states, that there exists a special choice, such that for the nonzero eigenvalues, the eigenvalue decomposition is continuous (and even differentiable for eigenvalues of multiplicity one). It will always be assumed that the actual implementation of the eigenvalue decomposition (respectively the SVD) has this type of continuity property, i.e. that the orientation of the singular vectors is chosen such that the corresponding SVD is continuous in this sense at the true system. In practice, this fact might have to be taken into account for the implementation of the SVD. The first result of this paper states the consistency properties of the `MOESP` type of algorithms.

Theorem 8 (Consistency of the `MOESP` type of algorithms). *Let the process* $\mathbf{y}_t$ *be generated by a system of the form* (1), *which fulfills the standard assumptions. Let the input sequence fulfill the weak assumptions, and let* $\alpha \geq n+1$ *and* $\beta \geq n$ *be user defined choices. Also let* $\hat{\mathbf{W}}_\alpha^+$ *and* $\hat{\mathbf{W}}_\beta^-$ *be user defined weightings subject to the restrictions presented in Section* 3. *Then the following holds*:

- *If the true pair* $(\mathbf{k}_0, \mathbf{l}_0) \in M_n(\mathbf{u}_t, \alpha, \beta)$, *then there exist orthonormal matrices* $\mathbf{S}_T$ *such that*

$$\|\mathrm{vec}[\mathbf{S}_T\hat{\mathbf{A}}_T\mathbf{S}_T^{\mathrm{T}} - \mathbf{A}_0, \mathbf{S}_T\hat{\mathbf{B}}_T - \mathbf{B}_0, \hat{\mathbf{C}}_T\mathbf{S}_T^{\mathrm{T}} - \mathbf{C}_0, \hat{\mathbf{D}}_T - \mathbf{D}_0]\| \to 0 \quad a.s.,$$

i.e. the estimate of the corresponding transfer function $\mathbf{l}$ *is a.s. consistent.*

- *If* $(\mathbf{k}_0, \mathbf{l}_0) \in M_n^+(\mathbf{u}_t, \alpha, \beta)$, *then*

$$\|\mathrm{vec}[\hat{\mathbf{A}}_T - \mathbf{A}_0, \hat{\mathbf{B}}_T - \mathbf{B}_0, \hat{\mathbf{C}}_T - \mathbf{C}_0, \hat{\mathbf{D}}_T - \mathbf{D}_0]\| \to 0 \quad a.s.,$$

i.e. the estimates of the system matrices are a.s. consistent.

Proof. The main technical issues have been given already before the theorem. Note that from the assumptions on $\mathbf{y}_t, \mathbf{u}_t$ and $(\mathbf{k}_0, \mathbf{l}_0)$ it follows that the sample covariances converge to their population analogs. As has been stated already, it follows that $\hat{\mathbf{X}}_{\alpha,\beta}$ converges to $\mathbf{X}_{\alpha,\beta}$ a.s. Then Lemma 7 implies that the eigenspaces corresponding to the n largest eigenvalues of $\hat{\mathbf{X}}_{\alpha,\beta}$ converge to the eigenspaces of the n nonzero eigenvalues of $\mathbf{X}_{\alpha,\beta}$, where convergence is in the gap metric. Recall that $\hat{\mathbf{\Gamma}}_\alpha = (\hat{\mathbf{W}}_\alpha^+)^{-1}\hat{\mathbf{U}}_n$. Note that $\hat{\mathbf{W}}_\alpha^+ \to \mathbf{W}_\alpha^+$ a.s. and, thus, the consistency of $\hat{\mathbf{\Gamma}}_\alpha$ is implied by the convergence of $\hat{\mathbf{U}}_n$ to $\mathbf{U}_n$. The columns of $\hat{\mathbf{U}}_n$ are identical to the eigenvectors of $\hat{\mathbf{X}}_{\alpha,\beta}$. Now for $(\mathbf{k}_0, \mathbf{l}_0) \in M_n^+(\mathbf{u}_t, \alpha, \beta)$ the eigenvalue decomposition is continuous at $\mathbf{X}_{\alpha,\beta}$, since then all eigenvalues are distinct, and the orientation of the eigenvectors is fixed so as to ensure the continuity (cf. the discussion before the theorem). Thus $\hat{\mathbf{\Gamma}}_\alpha \to \mathbf{\Gamma}_\alpha$ follows in this case. Note that $\hat{\mathbf{A}}_T$ and $\hat{\mathbf{C}}_T$ are nonlinear continuous functions of $\hat{\mathbf{\Gamma}}_\alpha$ (using the full rank property of $\mathbf{\Gamma}_{\alpha-1}$) and therefore the consistency for $\hat{\mathbf{\Gamma}}_\alpha$ implies consistency for $\hat{\mathbf{A}}_T$ and $\hat{\mathbf{C}}_T$, respectively. Here the restriction $\alpha \geq n+1$ is used. For $(\mathbf{k}_0, \mathbf{l}_0) \in M_n(\mathbf{u}_t, \alpha, \beta)$ only the existence of orthonormal matrices $\mathbf{S}_T$ such that $\hat{\mathbf{\Gamma}}_\alpha\mathbf{S}_T^{\mathrm{T}} \to \mathbf{\Gamma}_\alpha$ for $T \to \infty$ can be obtained. This follows in a straightforward manner from the convergence of the eigenspaces in the gap metric.

Thus it remains to prove the consistency for $\hat{\mathbf{B}}_T$ and $\hat{\mathbf{D}}_T$. Recall that $\mathbf{B}$ and $\mathbf{D}$ are estimated using the structure of $\mathbf{\Phi}_\alpha$. In addition the matrix $\hat{\mathbf{\Gamma}}_\alpha^\perp = (\hat{\mathbf{W}}_\alpha^+)^{\mathrm{T}}\hat{\mathbf{U}}_2$ has been introduced. As has been stated already, $\hat{\mathbf{W}}_\alpha^+ \to \mathbf{W}_\alpha^+$. Examining the least squares solution of Eq. (3) it follows that

$$\mathrm{vec}[\hat{\mathbf{B}}_T, \hat{\mathbf{D}}_T] = [\hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}}^{\mathrm{T}}(\mathbf{I}\otimes\hat{\mathbf{\Gamma}}_\alpha^\perp(\hat{\mathbf{\Gamma}}_\alpha^\perp)^{\mathrm{T}})\hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}}]^{-1}\hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}}^{\mathrm{T}} \mathrm{vec}[\hat{\mathbf{\Gamma}}_\alpha^\perp(\hat{\mathbf{\Gamma}}_\alpha^\perp)^{\mathrm{T}}\langle\mathbf{Y}_{t,\alpha}, \mathbf{U}_{t,\alpha}\rangle\langle\mathbf{U}_{t,\alpha}, \mathbf{U}_{t,\alpha}\rangle^{-1}].$$

The assumptions imply that $\langle\mathbf{Y}_{t,\alpha}, \mathbf{Y}_{t,\alpha}\rangle \to \mathbb{E}\mathbf{Y}_{t,\alpha}\mathbf{Y}_{t,\alpha}^{\mathrm{T}}$ and $\langle\mathbf{Y}_{t,\alpha}, \mathbf{U}_{t,\alpha}\rangle \to \mathbb{E}\mathbf{Y}_{t,\alpha}\mathbf{U}_{t,\alpha}^{\mathrm{T}}$. Furthermore $\hat{\mathbf{\Gamma}}_\alpha^\perp(\hat{\mathbf{\Gamma}}_\alpha^\perp)^{\mathrm{T}} = (\hat{\mathbf{W}}_\alpha^+)^{\mathrm{T}}\hat{\mathbf{U}}_2\hat{\mathbf{U}}_2^{\mathrm{T}}\hat{\mathbf{W}}_\alpha^+ = (\hat{\mathbf{W}}_\alpha^+)^{\mathrm{T}}(\mathbf{I} - \hat{\mathbf{U}}_n\hat{\mathbf{U}}_n^{\mathrm{T}})\hat{\mathbf{W}}_\alpha^+$. Since both the consistency of $\hat{\mathbf{W}}_\alpha^+$ and the consistency of $\hat{\mathbf{U}}_n$ have been established, the a.s. convergence of $\hat{\mathbf{\Gamma}}_\alpha^\perp(\hat{\mathbf{\Gamma}}_\alpha^\perp)^{\mathrm{T}}$ follows. From the consistency of $\mathbf{S}_T\hat{\mathbf{A}}_T\mathbf{S}_T^{\mathrm{T}}$ and $\hat{\mathbf{C}}_T\mathbf{S}_T^{\mathrm{T}}$ it is also clear that $\hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}}[\mathbf{I}_{m+s}\otimes\mathrm{diag}(\mathbf{S}_T^{\mathrm{T}}, \mathbf{I}_s)] \to \mathbf{L}_{\mathbf{B},\mathbf{D}}$ a.s., where $\mathbf{S}_T = \mathbf{I}_n$ is used for $(\mathbf{k}_0, \mathbf{l}_0) \in M_n^+(\mathbf{u}_t, \alpha, \beta)$. Here $\mathbf{I}_l$ denotes the $l \times l$ identity matrix. It remains to show the asymptotic nonsingularity of

$$\hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}}^{\mathrm{T}}(\mathbf{I}\otimes\hat{\mathbf{\Gamma}}_\alpha^\perp(\hat{\mathbf{\Gamma}}_\alpha^\perp)^{\mathrm{T}})\hat{\mathbf{L}}_{\mathbf{B},\mathbf{D}}.$$

It is sufficient to show that there exists no vector $\mathbf{x} \in \mathbb{R}^{mn+sm}$ different from zero such that $(\mathbf{I} \otimes (\mathbf{\Gamma}_\alpha^\perp)^{\mathrm{T}}) \mathbf{L}_{\mathbf{B},\mathbf{D}} \mathbf{x} = 0$. Recall that $\mathbf{L}_{\mathbf{B},\mathbf{D}} \operatorname{vec}[\mathbf{B}, \mathbf{D}] = \operatorname{vec}[\mathbf{\Phi}_\alpha]$. Defining two matrices $\mathbf{B}$ and $\mathbf{D}$ of appropriate dimensions via $\operatorname{vec}[\mathbf{B}, \mathbf{D}] = \mathbf{x}$, then clearly it is sufficient to show that there exist no two matrices $\mathbf{B}$ and $\mathbf{D}$ different from zero such that

$$(\mathbf{\Gamma}_\alpha^\perp)^{\mathrm{T}} \mathbf{\Phi}_\alpha = (\mathbf{\Gamma}_\alpha^\perp)^{\mathrm{T}} \begin{bmatrix} \mathbf{D} & 0 & \cdots & 0 \\ \mathbf{C}_0\mathbf{B} & \mathbf{D} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \mathbf{C}_0\mathbf{A}_0^{\alpha-2}\mathbf{B} & \cdots & \mathbf{C}_0\mathbf{B} & \mathbf{D} \end{bmatrix} = 0.$$

By studying the last block column of this expression it can be seen that this is only possible if

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{D} \end{bmatrix} = \mathbf{\Gamma}_\alpha \mathbf{T}$$

for some matrix $\mathbf{T} \in \mathbb{R}^{n \times s}$. Since $\mathbf{\Gamma}_{\alpha-1}$ is full rank, $\mathbf{T}$ has to be zero, which in turn implies that $\mathbf{D} = 0$. Using the same arguments on the second to last block column leads to the conclusion that $\mathbf{C}_0\mathbf{B} = 0$. Continuing the same reasoning on all block columns finally shows that $\mathbf{C}_0\mathbf{A}_0^i\mathbf{B} = 0$, $i = 0, 1, \ldots, \alpha - 2$, or equivalently $\mathbf{\Gamma}_{\alpha-1}\mathbf{B} = 0$. This implies that $\mathbf{B} = 0$ since $\mathbf{\Gamma}_{\alpha-1}$ is full rank which proves the conjecture.

Finally observe that in the case that only $(\mathbf{k}_0, \mathbf{l}_0) \in M_n(\mathbf{u}_t, \alpha, \beta)$ is imposed, the matrix $\mathbf{S}_T$ also appears in the convergence result for the least-squares estimate $\hat{\mathbf{B}}_T$, which is easily seen from Eq. (3). This concludes the proof. □

Remark 9. The result shown above holds equally well for the constrained regresssion approach, i.e., when the lower triangular block Toeplitz structure of $\mathbf{\Phi}_\alpha$ is imposed on $\hat{\mathbf{\Phi}}_\alpha$ in the regression computed in the first step of the algorithm. This can easily be seen from the consistency of $\hat{\mathbf{H}}_{\alpha,\beta}$ also in this case. Clearly the definition of $M_n(\mathbf{u}_t, \alpha, \beta)$ and $M_n^+(\mathbf{u}_t, \alpha, \beta)$ has to be adapted to the constrained regression approach.

Remark 10. As has been pointed out by an anonymous referee, results of this kind are sometimes termed *generic consistency* in the context of instrumental variable methods (see e.g. Söderström & Stoica, 1989). This observation is interesting since an instrumental variable interpretation of the class of subspace algorithms treated in this paper has been given e.g. in Viberg (1995). In view of this, the above consistency result is quite expected.

Note that this result implies the convergence of the system matrix estimates to the possibly nonminimal realization $(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0, \mathbf{D}_0)$. The key argument in the derivation proved to be the convergence of the sample covariances to the true ones, since the estimates of the system matrices have been shown to be continuous functions of the sample covariances under the conditions on the true pair $(\mathbf{k}_0, \mathbf{l}_0)$. For the asymptotic normality part, note that this nonlinear mapping can be linearized: The most crucial part of the proof of this statement is the linearization of the singular vectors contained in $\hat{\mathbf{U}}_n$. This property follows from Lemma 7. The remaining steps in the nonlinear mapping consist only of matrix inversions and Cholesky decompositions, which clearly can be linearized. For the proof of the latter, see e.g. Golub and Van Loan (1989). Thus, it remains to impose conditions on the exogenous inputs, such that the sample covariances for the joint process $\mathbf{z}_t$ fulfill a central limit theorem. For given α and β the following assumptions on the input sequence are imposed in order to ensure the asymptotic normality of the system matrix estimates:

Definition 11 (*Strong assumptions on inputs*). The input process $\mathbf{u}_t$ admits the decomposition $\mathbf{u}_t = \mathbf{v}_t + \mathbf{s}_t$, where $\mathbf{v}_t = \sum_{j=0}^{\infty} \mathbf{K}_u(j)\boldsymbol{\eta}_{t-j}$, $\|\mathbf{K}_u(j)\| \le C\rho^j$ for some $C < \infty$, $0 < \rho < 1$, and $\mathbf{s}_t = \sum_{l=1}^{h} \mathbf{c}_l e^{i\lambda_l t}$ for some integer h, for some vectors $\mathbf{c}_l \in \mathbb{C}^m$ and frequencies $-\pi \le \lambda_l < \pi$, $1 \le l \le h$ such that the corresponding process is real. Here $\boldsymbol{\eta}_t$ denotes an i.i.d. sequence having mean zero and variance unity and finite fourth moments, which is independent of $\boldsymbol{\varepsilon}_t$. Furthermore $\mathbf{u}_t$ is persistently exciting in the sense that the matrix $\mathbf{\Gamma}_{u,u}$, defined in the weak assumptions, is nonsingular.

Remark 12. The authors want to emphasize that this is by no means the only scenario, where the results below hold. See the proof of Theorem 13 for the crucial properties of the input process.

The following result is immediate from the discussion above.

Theorem 13 (Asymptotic normality). *Let the process* $\mathbf{y}_t$ *fulfill the standard assumptions, where the input process fulfills the strong assumptions given above. Let* $\alpha \ge n + 1$ *and* $\beta \ge n$. *Furthermore, let the weighting matrices* $\hat{\mathbf{W}}_\alpha^+$ *and* $\hat{\mathbf{W}}_\beta^-$ *be chosen according to the restrictions stated in Section* 3. *Finally, let the true pair* $(\mathbf{k}_0, \mathbf{l}_0) \in M_n^+(\mathbf{u}_t, \alpha, \beta)$. *Then*

$$\sqrt{T} \operatorname{vec}[\hat{\mathbf{A}}_T - \mathbf{A}_0, \hat{\mathbf{B}}_T - \mathbf{B}_0, \hat{\mathbf{C}}_T - \mathbf{C}_0, \hat{\mathbf{D}}_T - \mathbf{D}_0] \xrightarrow{\mathrm{d}} \mathbf{Z},$$

where $(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0, \mathbf{D}_0)$ *denotes the particular realization of the true pair* $(\mathbf{k}_0, \mathbf{l}_0)$ *described before Theorem* 8, *and* $\mathbf{Z}$ *denotes a multivariate Gaussian random vector with mean zero and variance equal to* $\mathbf{V}$.

Proof. From the discussion before the theorem, it only remains to prove the asymptotic normality of the sample

covariances up to lag $\alpha+\beta-1$ for the joint process $\mathbf{z}_t=[\mathbf{y}_t^{\mathrm{T}},\mathbf{u}_t^{\mathrm{T}}]^{\mathrm{T}}$ under the strong assumptions on the input process. This result is not new (cf. e.g. Hannan & Deistler, 1988 Lemma 4.3.4). The authors however decide to give the details of the proof, since this reveals the sufficient properties of the input sequence, which in fact guarantee the asymptotic normality. The main tool in the proof is what is sometimes called Bernstein's lemma (see e.g. Hannan & Deistler, 1988, Lemma 4.3.3): If $\mathbf{x}_T$ is a sequence of random vectors and for every $\zeta>0,\varepsilon>0$, $\eta>0$ there exist sequences $\mathbf{a}_T(\varepsilon),\mathbf{b}_T(\varepsilon)$ so that $\mathbf{x}_T=\mathbf{a}_T(\varepsilon)+\mathbf{b}_T(\varepsilon)$ and $\mathbf{a}_T(\varepsilon)$ has a distribution converging to the normal distribution with mean zero and variance $\boldsymbol{\Sigma}(\varepsilon)\to\boldsymbol{\Sigma}$ for $\varepsilon\to 0$ and $\mathbb{P}\{\mathbf{b}_T(\varepsilon)^{\mathrm{T}}\mathbf{b}_T(\varepsilon)>\zeta\}<\eta$, $\forall T>T_0$, then $\mathbf{x}_T$ is asymptotically normal with variance $\boldsymbol{\Sigma}$. Here $\mathbb{P}$ is used to denote probability. This lemma is used, where $\mathbf{x}_T=(1/\sqrt{T})\sum_{t=1}^{T}(\mathbf{z}_t\mathbf{z}_{t+j}^{\mathrm{T}}-\boldsymbol{\gamma}_{z,z}(j))$ and $\mathbf{a}_T(\varepsilon)=(1/\sqrt{T})\sum_{t=1}^{T}(\mathbf{z}_t(\varepsilon)\mathbf{z}_{t+j}^{\mathrm{T}}(\varepsilon)-\boldsymbol{\gamma}_{z,z}(j))$,$\mathbf{z}_t(\varepsilon)=[\mathbf{y}_t(\varepsilon)^{\mathrm{T}},\mathbf{u}_t(\varepsilon)^{\mathrm{T}}]^{\mathrm{T}}$, $\mathbf{y}_t(\varepsilon)=\sum_{i=0}^{m}\mathbf{K}(i)\boldsymbol{\varepsilon}_{t-i}+\mathbf{L}(i)\mathbf{u}_{t-i}(\varepsilon)$,$\mathbf{u}_t(\varepsilon)=\sum_{i=0}^{m}\mathbf{K}_u(i)\boldsymbol{\eta}_{t-i}+\mathbf{s}_t$, i.e. the infinite sums are truncated at m. The point of truncation m can be chosen to make the probability $\mathbb{P}\{\|\mathbf{u}_t-\mathbf{u}_t(\varepsilon)\|>\zeta\}$ arbitrarily small for all $\zeta>0$. This follows from straightforward calculations using the Chebycheff inequality and the exponential decrease of the coefficients $\mathbf{K}_u(j)$. The same arguments show that the condition also holds for $\mathbf{z}_t-\mathbf{z}_t(\varepsilon)$ and suitably chosen m. Using the Bernstein lemma again, it is observed that the joint asymptotic normality of the covariance estimates of $\mathbf{z}_t$ is proved, if the joint asymptotic normality of the terms $(1/\sqrt{T})\sum_{t=1}^{T}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_{t+j}^{\mathrm{T}}-\delta_{0,j}\mathbf{I})$,$(1/\sqrt{T})\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t\boldsymbol{\eta}_{t-j}^{\mathrm{T}}$,$(1/\sqrt{T})\sum_{t=1}^{T}(\boldsymbol{\eta}_t\boldsymbol{\eta}_{t-j}^{\mathrm{T}}-\delta_{0,j}I)$,$(1/\sqrt{T})\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t\mathbf{s}_{t-j}^{\mathrm{T}}$,$(1/\sqrt{T})\sum_{t=1}^{T}\mathbf{s}_t\boldsymbol{\eta}_{t-j}^{\mathrm{T}}$ and $(1/\sqrt{T})\sum_{t=1}^{T}\mathbf{s}_t\mathbf{s}_{t-j}^{\mathrm{T}}$ is shown for $|j|\le m$. Here $\delta_{0,j}=1$ for $j=0$ and zero else. Clearly this holds for the first three terms. The central limit theorem for $(1/T)\sum_{t=1}^{T}\mathrm{e}^{\mathrm{i}\lambda_t t}[\boldsymbol{\eta}_t^{\mathrm{T}},\boldsymbol{\varepsilon}_t^{\mathrm{T}}]$ follows, e.g. from Anderson (1971), Theorems 8.4.1. and 8.4.3). Also the joint asymptotic normality follows. Finally,

$$\sum_{r,s=1}^{h}\mathbf{c}_r\mathbf{c}_s^{\mathrm{H}}\left[\frac{1}{T}\sum_{t=1}^{T-j}\mathrm{e}^{\mathrm{i}(\lambda_r-\lambda_s)t}\right]\mathrm{e}^{-\mathrm{i}\lambda_s j}$$

$$=\sum_{r,s\,:\,\lambda_r\neq\lambda_s}\mathbf{c}_r\mathbf{c}_s^{\mathrm{H}}\left[\frac{\mathrm{e}^{\mathrm{i}(\lambda_r-\lambda_s)}(1-\mathrm{e}^{\mathrm{i}(\lambda_r-\lambda_s)(T-j)})}{T(1-\mathrm{e}^{\mathrm{i}(\lambda_r-\lambda_s)})}\right]\mathrm{e}^{-\mathrm{i}\lambda_s j}$$

$$+\sum_{r=1}^{h}\mathbf{c}_r\mathbf{c}_r^{\mathrm{H}}\frac{T-j}{T}\mathrm{e}^{-\mathrm{i}\lambda_s j}\to\sum_{r=1}^{h}\mathbf{c}_r\mathbf{c}_r^{\mathrm{H}}\mathrm{e}^{-\mathrm{i}\lambda_s j},$$

where the difference between the sample moments and the limit is $o(1/\sqrt{T})$. Here $\mathbf{c}_r^{\mathrm{H}}$ denotes the complex conjugate of the transposed vector. This completes the proof. □

As a byproduct also the asymptotic distribution of various invariants may be obtained:

Corollary 14. *Let the assumptions of Theorem* 13 *hold and let g be a differentiable mapping attaching the vector* $\mathbf{x}\in\mathbb{R}^l$ *to system matrices* $(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})$. *Denote the Jacobian matrix of g at* $(\mathbf{A}_0,\mathbf{B}_0,\mathbf{C}_0,\mathbf{D}_0)$ *with respect to the entries of* $(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})$ *with* $\mathbf{J}$. *Then*

$$\sqrt{T}[g(\hat{\mathbf{A}}_T,\hat{\mathbf{B}}_T,\hat{\mathbf{C}}_T,\hat{\mathbf{D}}_T)-g(\mathbf{A}_0,\mathbf{B}_0,\mathbf{C}_0,\mathbf{D}_0)]\xrightarrow{\mathrm{d}}\mathbf{Z},$$

where $\mathbf{Z}$ *is multivariate Gaussian with zero mean and variance* $\mathbf{J}\mathbf{V}\mathbf{J}^{\mathrm{T}}$. *Here*, $\mathbf{V}$ *is defined in Theorem* 13.

Of course, the variance $\mathbf{V}$ depends on the true pair $(\mathbf{k}_0,\mathbf{l}_0)$, the weighting matrices and the indices α,β, however, this has not been emphasized notationally. The expressions for the asymptotic variances are quite complicated and thus have not yet contributed to an analytical analysis, to the best of the authors' knowledge. However, it is possible to approximate them on a computer (cf. the next section). The corollary can then be used to compare how different choices of the truncation indices and weighting matrices affect the estimation accuracy. For example, the accuracies of the pole estimates or the estimates of the zeros can be assessed, if the poles or, respectively, the zeros are distinct. Another example concerns the variance of the frequency function $\mathbf{C}(\mathrm{e}^{\mathrm{i}\omega}\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D}$. For a given frequency, this function can be differentiated with respect to the system matrices whereafter the result of the corollary can be applied. Clearly, these results also can be used to compute approximative confidence intervals around the estimated quantities. Some illustrations of the above are given in the next section.

Remark 15. Note that the tools used in this paper also can be used to analyze the PI scheme (see Verhaegen & Dewilde, 1993). This scheme differs from the PO scheme only in the fact that $\mathbf{P}_{t,\beta}=\mathbf{U}_{t,\beta}$ is used instead of both the past of the input and the output process. In the case of the PI algorithm, the definition of the set $M_n^+(\mathbf{u}_t,\alpha,\beta)$ will be different. Then the set $M_n(\mathbf{u}_t,\alpha,\beta)$ e.g. can be chosen to be the set of all pairs of transfer functions $(\mathbf{k},\mathbf{l})$, where $\mathbf{k}$ is rational, stable and strictly minimumphase with nonsingular constant term, and where $\mathbf{l}$ is rational, stable and of order n. Note that in fact two different minimality concepts are used: Using the whole past (inputs and outputs) corresponds to parametrizing $\mathbf{k}$ and $\mathbf{l}$ jointly, i.e. mixing the dynamics as indicated in Eq. (1). Thus, in general the realization $(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})$ of $\mathbf{l}$ is not ensured to be minimal. (There may be modes in $\mathbf{A}$ corresponding to $\mathbf{k}$ which are not shared by $\mathbf{l}$. Such modes cancel when forming $\mathbf{l}$ from $(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})$.) On the other hand, using $\mathbf{P}_{t,\beta}=\mathbf{U}_{t,\beta}$ corresponds to parametrizing $\mathbf{k}$ and $\mathbf{l}$ independently and the corresponding realization $(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})$ of $\mathbf{l}$ will be minimal. However, this may not lead to a system of the form (1) of the same dimension.

5. Numerical examples

In the previous section, the asymptotic normality of the MOESP algorithm has been derived. In Theorem 13 the variance of the limiting normal distribution has been denoted with $\mathbf{V}$. As has been stated already, $\mathbf{V}$ depends on the covariance sequence of the inputs, the choice of the weighting matrices and the choice of the indices α, β. The theorem also shows that $\mathbf{V}$ can be calculated from the knowledge of the covariances of the covariance estimates of the joint process $\mathbf{z}_t = [\mathbf{y}_t^{\mathrm{T}}, \mathbf{u}_t^{\mathrm{T}}]^{\mathrm{T}}$. This merely amounts to calculating the linearization of the nonlinear mapping, which is induced by the algorithm, attaching system matrix estimates to estimates of the covariance sequence. The major steps in this nonlinear mapping are as follows: First a regression in Eq. (2) is performed, which is a function of the sample covariances of the input and the output process of lags up to $\alpha + \beta - 1$. The linearization of the mapping attaching the estimates $[\hat{\mathbf{H}}_{\alpha,\beta}, \hat{\mathbf{\Phi}}_\alpha]$ to the covariance estimates is straightforward to derive. The linearizations of the mappings attaching the weighting matrices $\hat{\mathbf{W}}_\alpha^+$ and $\hat{\mathbf{W}}_\beta^-$, respectively, can be calculated by using the Cholesky factors as the required square roots. In the next step, the SVD of $\hat{\mathbf{W}}_\alpha^+ \hat{\mathbf{H}}_{\alpha,\beta} \hat{\mathbf{W}}_\beta^-$ is calculated. In order to linearize the mapping attaching the matrix $\hat{\mathbf{U}}_n$ to the covariance estimates, the results given in Lemma 7 are used. The remaining steps consist of matrix inversions and multiplications only. Note, that these linearizations can be calculated without any approximation. In order to obtain the asymptotic variance of the covariance matrix estimates, the truncation techniques used in the proof of Theorem 13 may be used, which leads to an approximation of the asymptotic covariances. The approximation error depends on the impact of the truncated part of the infinite sum, and thus is directly related to the magnitude of $|\lambda_{\max}(\mathbf{A})|$.

As can be seen from the previous paragraph, the resulting expressions seem to be too complicated to be evaluated analytically. However, for a given system, the expressions can be approximated on a computer. Thus it is possible for any system to compare the asymptotic variance of estimates of system invariants as, e.g. the system poles, the system zeros or the transfer function at some frequency points. This will be done in the following. The discussion below is provided mainly to show a potential use of the theory presented above. It is not intended to investigate thoroughly the effects of various choices in subspace algorithms. Therefore, the reader should note that the statements below always refer only to a number of examples and are not general statements.

Consider the system

$$\mathbf{A} = 0.5, \quad \mathbf{B} = 1, \quad \mathbf{C} = 1, \quad \mathbf{D} = 0, \quad \mathbf{E} = 1, \quad \mathbf{K} = 1,$$

where the input process is either unit variance white noise, or unit variance white noise filtered with the filter having system matrices

$$\mathbf{A}_u = \begin{bmatrix} 0 & 1 \\ -0.7 & 0.5 \end{bmatrix}, \qquad \mathbf{B}_u = \begin{bmatrix} 1.3 \\ 0.3 \end{bmatrix},$$

$$\mathbf{C}_u = [1,0], \qquad \mathbf{D}_u = 1.$$

In a first example, the effect of various choices of the indices α and β is investigated. Fig. 1 shows the result: For the given system using white noise inputs, the asymptotic variance of the transfer function estimates at 100 equally spaced frequency points in the interval $(-\pi, \pi)$ is

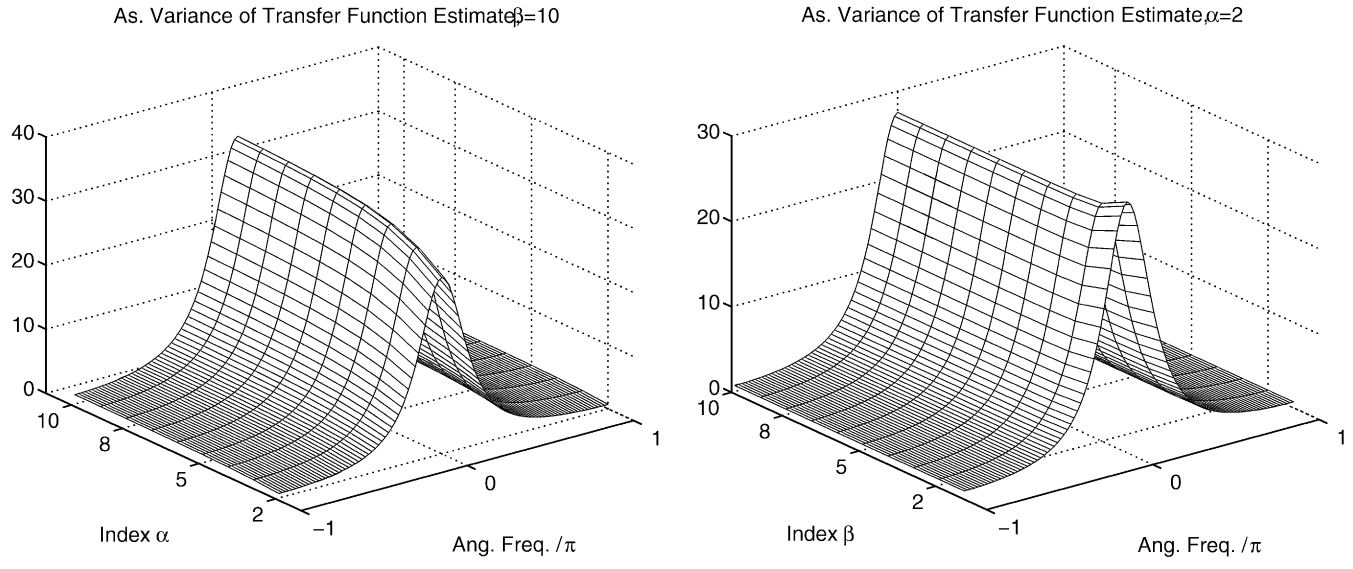


Fig. 1. The left plot shows the asymptotic variance of the estimated transfer function at 100 equally spaced frequency points, using $\hat{\mathbf{W}}_\alpha^+ = \mathbf{I}$ and $\beta = 10$ for various values of α. The right plot shows the same picture for $\alpha = 2$ and various values of β. In both plots the input is white noise.

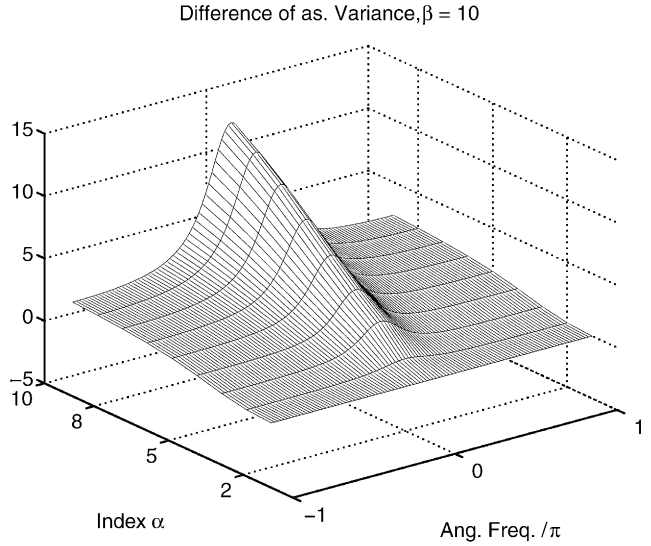


Fig. 2. This figure shows the asymptotic variance of the transfer function estimate obtained by using the CCA weight minus the corresponding variance when the MOESP weight is employed. The colored input is used and $\beta = 10$.

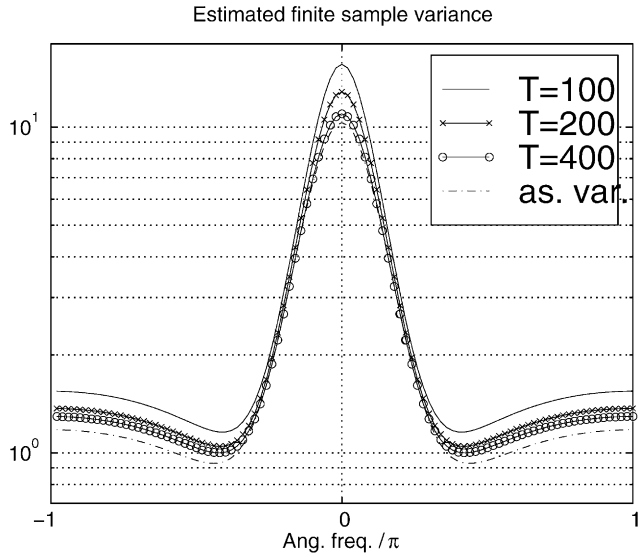


Fig. 3. In this figure the sample variance of the transfer function estimates and the true asymptotic variance are plotted for sample sizes $T = 100, 200$ and 400, respectively. $\alpha = 2, \beta = 10$ and MOESP weighting were used in all cases. 1000 replications of the colored input and the noise sequences were used to produce each of the curves.

plotted for various values of α and $\beta = 10$ (left plot) and for various values of β, where $\alpha = 2$ is used (right plot). The plots indicate, that while for β the asymptotic accuracy increases with increasing index, for the index α the contrary seems to be true. For the colored noise input the plots are qualitatively the same. A similar behavior has been noted for the accuracies of the pole estimates in Jansson and Wahlberg (1996). However, this is not true for all systems, i.e., for some other systems the performance improves with increasing α. In Jansson and Wahlberg (1996) it is shown that the choice of the weighting matrix $\hat{\mathbf{W}}_\alpha^+$ does not influence the asymptotic accuracy of the pole estimate of the system. Fig. 2 shows that an analogous statement for the transfer function estimate is not true: Two popular choices for the weighting are $\hat{\mathbf{W}}_\alpha^+ = \mathbf{I}$ (which will be called MOESP in the following) and $\hat{\mathbf{W}}_\alpha^+ = (\langle \mathbf{Y}_{t,\alpha}, \mathbf{Y}_{t,\alpha} \rangle - \langle \mathbf{Y}_{t,\alpha}, \mathbf{U}_{t,\alpha} \rangle (\langle \mathbf{U}_{t,\alpha}, \mathbf{U}_{t,\alpha} \rangle)^{-1} \langle \mathbf{U}_{t,\alpha}, \mathbf{Y}_{t,\alpha} \rangle)^{-1/2}$ (which uses the same weights as the CCA procedure of Larimore, 1983). Fig. 2 shows the difference between the two asymptotic variances of the transfer function estimates obtained by applying the CCA and, respectively, the MOESP weights for various values of α with $\beta = 10$ and using colored noise inputs. It can be seen that in this example the MOESP weight performs better than the CCA weight for all choices of indices. However, in the case with white noise inputs, the opposite is true. Then the CCA weight is to be preferred. In order to demonstrate the finite sample properties and to illustrate the asymptotic result, a simulation study was performed. For the given system with the colored noise inputs, 1000 replications of the noise and the input were generated for each of the sample sizes $T = 100$, 200 and 400, respectively. The system was estimated using the MOESP weighting and $\alpha = 2, \beta = 10$. Fig. 3 shows the sample variance (scaled by T) for the various values of the sample size T. Additionally, the theoretical variance is plotted at 100 equally spaced frequency points. The picture clearly reveals the convergence of the estimates to the true asymptotic values. Finally, the MOESP class of algorithms may be compared to another class of algorithms, called CCA in Peternell et al. (1996): This class was originally proposed by Larimore (1983) and refined by Peternell et al. (1996). The idea of these methods is to estimate the state in a first step from the SVD of $\hat{\mathbf{W}}_\alpha^+ \hat{\mathbf{H}}_{\alpha,\beta} \hat{\mathbf{W}}_\beta^-$ and to obtain estimates of the system matrices from regression in the system equations (1), once an estimate of the state is known. Simulations in that paper showed that in some cases a procedure, which will be called CCAI in the following, is close to optimal. This procedure uses a preliminary estimate of the transfer function $\mathbf{l}$ in order to eliminate the effect of the future of the inputs in Eq. (2) (For a detailed description of the algorithm see Peternell et al. 1996.) However, Fig. 4 shows that, for the present example, the transfer function $\mathbf{l}$ can be estimated more accurately using MOESP in the case of colored inputs. In the case of white noise inputs the two procedures show no significant difference. Note that in this case the asymptotic accuracy of CCA is indistinguishable from the Cramér Rao lower bound. The figures also show that the benefit from using the more complicated method CCAI is only marginal even in the case of colored noise inputs (in the case of white inputs this fact can be shown analytically, see e.g. Peternell et al. (1996)). These results indicate that the choice of the identification procedure seems to depend heavily on

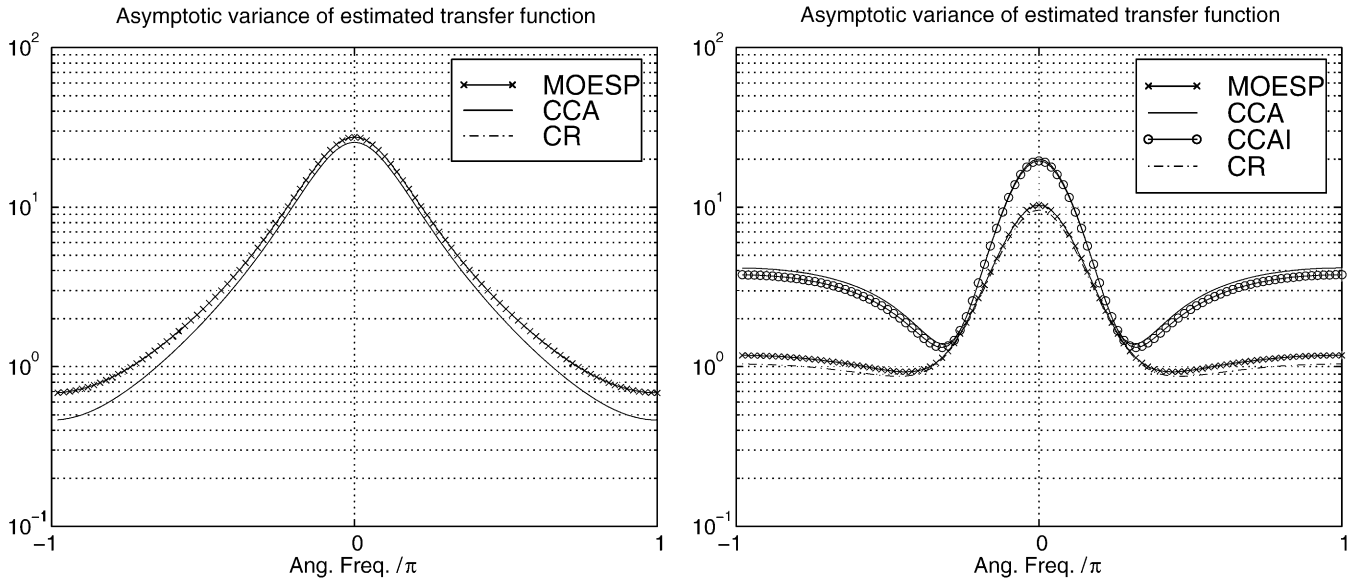


Fig. 4. In this figure the asymptotic variance of the estimates of the transfer function **l** at 100 equally spaced frequency points is plotted for MOESP ($\alpha = 2, \beta = 10$) and the algorithms denoted with CCA and CCAI ($\alpha = \beta = 15$) (see text for an explanation). The left plot refers to white inputs (in this case CCA and CCAI are asymptotically equivalent and thus only CCA is plotted), the right plot corresponds to colored inputs. In both plots CR denotes the Cramér Rao bound.

the input characteristics. In the examples it was observed that for white noise inputs CCA of Larimore (1983) performed better than MOESP, whereas in the colored noise case the opposite was true. Also the choice of the indices α and β seems to be crucial for the accuracy of the various algorithms.

6. Conclusions

In this paper the asymptotic performance of a special class of subspace algorithms has been investigated. The estimate of the transfer function from the exogenous inputs to the outputs has been shown to be a.s. consistent for a generic set of linear systems. The results in Jansson and Wahlberg (1997) show that this actually is the best result that can be expected. Furthermore, for a smaller generic set also the consistency for the system matrices has been shown, as well as asymptotic normality using suitable assumptions on the input process. This result can be used to compare various procedures on the basis of their asymptotic variance. Also, confidence regions for estimates of different system related quantities, e.g. the Markov parameters, can be computed using this asymptotic theory.

Acknowledgements

Support by the Austrian 'Fonds zur Förderung der wissenschaftlichen Forschung' Projekt P11213-MAT, the foundation BLANCEFLOR Boncompagni-Ludovisi, née Bildt, and the Swedish Foundation for International Cooperation in Research and Higher Education is gratefully acknowledged.

References

Anderson, T. W. (1971). *The statistical analysis of time series*. New York: Wiley.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematics and Statistics*, *34*, 122–148.

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

Bauer, D., Deistler, M., & Scherrer, W. (1999). Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs. *Automatica*, *35*(7), 1243–1254.

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

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

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

Dieudonné, J. (1969). *Foundations of modern analysis*. New York: Academic Press.

Golub, G., & Van Loan, C. (1989). *Matrix computations*. (2nd ed.). Maryland: John Hopkins University Press.

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

Jansson, M., & Wahlberg, B. (1996). A linear regression approach to state-space subspace system identification. *Signal Processing*, *52*(2), 103–129.

Jansson, M., & Wahlberg, B. (1998). On consistency of subspace methods for system identification. *Automatica*, *34*(12), 1507–1519.

Jansson, M., & Wahlberg, B. (1997). Counterexample to general consistency of subspace system identification methods. *Proceedings of SYSID'97*, Fukuoka, Japan (pp. 1677–1682).

Larimore, W. E. (1983). System identification, reduced order filters and modeling via canonical variate analysis. In H. S. Rao, & P. Dorato. editors, *Proceedings of the 1983 American Control Conference 2*. Piscataway, NJ (pp. 445–451). IEEE Service Center.

Lovera, M., Falcetti, A., & Bittanti, S. (1998). On the estimation of the A matrix in subspace model identification. *Proceedings of the MTNS'98 Conference*.

Peternell, K., Scherrer, W., & Deistler, M. (1996). Statistical analysis of novel subspace identification methods. *Signal Processing*, *52*, 161–177.

Söderström, T., & Stoica, P. (1989). *System identification*. Englewood Cliffs, NJ: Prentice-Hall.

VanOverschee, P., & DeMoor, B. (1994). N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, *30*(1), 75–93.

Van Overschee, P., & De Moor, B. (1996). *Subspace identification for linear systems: Theory*, *implementation*, *applications*. Dordrecht: Kluwer Academic Publishers.

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.

Verhaegen, M., & Dewilde, P. (1992a). Subspace model identification: Part 1. The output-error state-space identification class of algorithms. *International Journal of Control*, *56*(5), 1187–1210.

Verhaegen, M., & Dewilde, P. (1992b). Subspace model identification: Part 2. Analysis of the elementary output-error state-space model identification algorithm. *International Journal of Control*, *56*(5), 1211–1241.

Verhaegen, M., & Dewilde, P. (1993). Subspace model identification: Part 3. Analysis of the ordinary output-error state-space model identification algorithm. *International Journal of Control*, *58*(3), 555–586.

Viberg, M. (1995). Subspace-based methods for the identification of linear time-invariant systems. *Automatica*, *31*(12), 1835–1851.

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

Viberg, M., Wahlberg, B., & Ottersten, B. (1997). Analysis of state space system identification methods based on instrumental variables and subspace fitting. *Automatica*, *33*(9), 1603–1616.

Dietmar Bauer was born in St. Pölten, Austria, in 1972. He received his masters and Ph.D. degrees in Applied Mathematics from the Technical University of Vienna in 1995 and 1998 respectively. From 1995 until 1998 he was with the Institute for Econometrics, Operations Research and System Theory, Technical University of Vienna. Currently he is visiting the Department of Electrical and Computer Engineering, University of Newcastle, Australia. His research interests include system identification in particular subspace algorithms and parametrisation of linear systems, and economic applications of time series analysis. For a recent photograph of Dietmar Bauer please refer to Automatica 35(7) 1243–1254.

Magnus Jansson was born in Enköping, Sweden, in 1968. He received the Master of Science, Technical Licentiate, and Ph.D. degrees in electrical engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 1992, 1995 and 1997, respectively. From September 1998 he spent one year at the Department of Electrical and Computer Engineering, University of Minnesota, USA. He is currently a Research Associate at the Department of Signals, Sensors and Systems, Royal Institute of Technology.

His research interests include sensor array signal processing, time series analysis, and system identification.

For a recent photograph of Magnus Jansson please refer to Automatica 34(12) 1507–1519.

PERGAMON

Automatica 38 (2002) 763–773

automatica

www.elsevier.com/locate/automatica

Some facts about the choice of the weighting matrices in Larimore type of subspace algorithms ☆

Dietmar Bauer[a,*], Lennart Ljung[b]

[a] *Institute f. Econometrics, Operations Research and System Theory, TU Wien, Argentinierstrasse 8, A-1040 Wien, Austria*
[b] *Division of Automatic Control, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden*

Received 5 October 2000; received in revised form 7 May 2001; accepted 30 October 2001

Abstract

In this paper the effect of some weighting matrices on the asymptotic variance of the estimates of linear discrete time state space systems estimated using subspace methods is investigated. The analysis deals with systems with white or without observed inputs and refers to the Larimore type of subspace procedures. The main result expresses the asymptotic variance of the system matrix estimates in canonical form as a function of some of the user choices, clarifying the question on how to choose them optimally. It is shown, that the CCA weighting scheme leads to optimal accuracy. The expressions for the asymptotic variance can be implemented more efficiently as compared to the ones previously published. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Linear systems; Discrete time systems; Subspace methods; Asymptotic variance

1. Introduction

Subspace algorithms are used for the estimation of linear, time invariant, discrete time, finite dimensional black box state space models. The algorithms can be roughly divided into Larimore type of algorithms (Larimore, 1983) (one algorithm in this class is usually called CCA, canonical correlation analysis), which estimate the state in the first step and then extract the estimates of the system matrices from these estimates, and multivariable output error state space (MOESP) type of algorithms (Verhaegen, 1994), which estimate the observability matrix and use this estimate to obtain estimates of the system matrices. The asymptotic properties of the Larimore type of approach have been derived in a series of papers: Peternell, Scherrer, and Deistler (1996) derive the consistency, Bauer, Deistler, and Scherrer (1999) prove asymptotic normality in the case of no observed inputs, Bauer (1998) deals with the general case. For the MOESP type of procedure consistency and asymptotic normality are dealt with in Bauer and Jansson (2000), while preliminary results on consistency can also be found in Jansson and Wahlberg (1998) and Verhaegen (1994). The asymptotic normality proof is very constructive in both cases, which led to formulas for the asymptotic variance. However, these expressions were too complicated in order to directly provide some insight into the effect of certain user choices. Recently, simplifications of these formulas have been found independently in Jansson (2000) for the MOESP case and in Bauer, Deistler, and Scherrer (2000) for the Larimore type of procedures. These simpler expressions lie at the heart of this paper, which derives the corresponding variance expressions as a function of a certain weighting matrix. This expression can be used in order to optimize the user choice with respect to asymptotic accuracy of the estimated system.

The paper is organized as follows: In the next section the model set and the assumptions are stated and also a short overview of the estimation algorithms is given. Section 3 presents the main results, which are proved in Section 4. Section 5 demonstrates the results in some numerical examples. Finally Section 6 concludes.

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Brett Ninness under the direction of Editor Torsten Söderström.

* Corresponding author. Fax: +43-1-58801-11999.
E-mail address: dietmar.bauer@tuwien.ac.at (D. Bauer).

PII: S0005-1098(01)00261-8

Throughout the paper the following notation will be used: I_n denotes the $n \times n$ identity matrix, $0^{a\times b}$ the null-matrix of respective dimensions. Further, $f_T = O(g_T)$ means that $\limsup_{T>0} \|f_T/g_T\| \leqslant M$ almost sure (a.s.). Also, $f_T = o(g_T)$ means that $\lim_{T\to\infty} \|f_T/g_T\| = 0$ a.s. Here T is used to denote the sample size. Convergence is denoted as usual with $\to$ and is always meant to be a.s. if not stated explicitly. Prime is used to denote transposition of matrices. The Kronecker product between two matrices A and B is denoted as $A \otimes B$. Finally, $Q_T = \sqrt{T^{-1}\log\log T}$ is used and $\doteq$ denotes equality up to terms of order $o(T^{-1/2})$.

2. Model Set, assumptions and algorithm

This paper deals with linear, finite dimensional, discrete time, time invariant, state space systems of the form

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

where $y_t \in \mathbb{R}^s$ denotes the observed output process, $u_t \in \mathbb{R}^m$ denotes the observed input process and $\varepsilon_t \in \mathbb{R}^s$ the unobserved white noise sequence. $x_t \in \mathbb{R}^n$ is the state sequence. Thus, the true order of the system is denoted by 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 system is assumed to be stable, i.e. all eigenvalues of A are assumed to lie inside the unit circle, and strictly minimum phase, i.e. the eigenvalues of $A - KC$ are assumed to lie inside the unit circle. The system matrices correspond to a pair of transfer functions: Let $H(q)=I_s+C(qI_n-A)^{-1}K$ and let $G(q)=D+C(qI_n-A)^{-1}B$, where q denotes the forward shift operator. Furthermore, let M_n denote the set of all pairs of transfer functions that permit a state space representation of the form (1) fulfilling the stability and the strict minimum-phase assumption on $H(q)$.

The white noise ε_t is for simplicity assumed to be independently identically distributed (i.i.d.) with mean zero, nonsingular variance matrix $\Omega > 0$ and finite fourth moments. The results also hold under more general assumptions in a martingale difference framework, which can be found in Bauer et al. (1999). The input is assumed to be i.i.d. with mean zero and nonsingular variance $\Omega_u > 0$, also having finite fourth moments. Input and noise are assumed to be independent. These set of assumptions on the noise and the input will be termed *standard assumptions* in the following.

The basic structure of the algorithm can be outlined as follows (for a detailed description see e.g. Bauer, 1998, Chapter 3): Let $Y_{t,f}^+ = [y_t', y_{t+1}', \ldots, y_{t+f-1}']'$ and let $U_{t,f}^+$ and $E_{t,f}^+$, respectively, be constructed analogously using u_t and ε_t, respectively, in the place of y_t. Let $Z_{t,p}^- = [y_{t-1}', u_{t-1}', \ldots, y_{t-p}', u_{t-p}']'$. Here f and p are two integer parameters, which have to be chosen by the user. See below for assumptions on the choice of these integers.

Then it follows from the system equations (1) that

$$Y_{t,f}^+ = \mathcal{O}_f \mathcal{K}_p Z_{t,p}^- + \mathcal{U}_f U_{t,f}^+ + \mathcal{E}_f E_{t,f}^+ + \mathcal{O}_f (A-KC)^p x_{t-p}.$$

Here $\mathcal{O}_f' = [C', A'C', \ldots, (A^{f-1})'C']$ and $\mathcal{K}_p = [[K, B - KD], (A-KC)[K, B-KD], \ldots, (A-KC)^{p-1}[K, B-KD]]$. Further $\mathcal{U}_f$ is the matrix containing

$$[CA^{j-2}B, \ldots, CB, D, 0^{s\times(f-j)m}]$$

as its jth block row and $\mathcal{E}_f$ contains

$$[CA^{j-2}K, \ldots, CK, I_s, 0^{s\times(f-j)s}]$$

as its jth block row. This equation builds the basis for all subspace algorithms, which can be described as follows:

(1) Regress $Y_{t,f}^+$ onto $U_{t,f}^+$ and $Z_{t,p}^-$ to obtain an estimate $\hat{\beta}_z$ of $\mathcal{O}_f \mathcal{K}_p$ and an estimate $\hat{\beta}_u$ of $\mathcal{U}_f$, respectively. Due to finite sample effects $\hat{\beta}_z$ will typically be of full rank.
(2) For given n find a rank n approximation of $\hat{\beta}_z$ by using the SVD of $\hat{W}_f^+ \hat{\beta}_z \hat{W}_p^- = \hat{U}_n \hat{\Sigma}_n \hat{V}_n' + \hat{R}$. Here $\hat{\Sigma}_n$ denotes the diagonal matrix containing the largest n singular values in decreasing order. $\hat{U}_n$ contains the corresponding left singular vectors as columns and $\hat{V}_n$ the corresponding right singular vectors. Finally, $\hat{R}$ accounts for the neglected singular values. The matrices $\hat{W}_f^+$ and $\hat{W}_p^-$ are weighting matrices, which are chosen by the user. Further details are given below, for the moment it is sufficient to note, that these possibly data-dependent matrices are assumed to be nonsingular (a.s.). This leads to an approximation $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p = (\hat{W}_f^+)^{-1} \hat{U}_n \hat{\Sigma}_n \hat{V}_n' (\hat{W}_p^-)^{-1}$. The actual decomposition of this matrix into $\hat{\mathcal{O}}_f$ and $\hat{\mathcal{K}}_p$ has no influence on the estimated transfer functions.
(3) Using the estimates $\hat{\mathcal{O}}_f$, $\hat{\mathcal{K}}_p$ and $\hat{\beta}_u$ obtain the system matrix estimates.

In the second step an order has to be specified. Also, the matrices $\hat{W}_f^+$ and $\hat{W}_p^-$ have to be provided by the user. In the literature several different choices have been proposed. For the matrix $\hat{W}_p^-$ typical choices are $(\hat{\Gamma}_p^-)^{1/2}$ and $(\hat{\Gamma}_p^{-,II})^{1/2}$, where $\hat{\Gamma}_p^- = (1/T)\sum_{t=p+1}^T Z_{t,p}^- (Z_{t,p}^-)'$ denotes the sample variance of $Z_{t,p}^-$ and $X^{1/2}$ denotes the uniquely defined symmetric square root of a matrix X. Further $\hat{\Gamma}_p^{-,II} = \hat{\Gamma}_p^- - \hat{\Gamma}_{z,u} \hat{\Gamma}_u^{-1} \hat{\Gamma}_{u,z}$. Here $\hat{\Gamma}_u$ denotes the sample variance of $U_{t,f}^+$ and $\hat{\Gamma}_{u,z}$ the sample covariance of $U_{t,f}^+$ and $Z_{t,p}^-$. Let $\Gamma_p^- = \mathbb{E}\hat{\Gamma}_p^-$ denote the expectation of the covariance matrix. It follows from the assumptions on the inputs and the noise stated above that for any fixed p it holds that $\hat{\Gamma}_p^{-,II} - \Gamma_p^-$ converge to zero. Furthermore, the results stated e.g. in Hannan and Deistler (1988) imply, that the two norm of these matrices is bounded from below and from above a.s. uniformly for $p = O((\log T)^a)$, $a < \infty$, i.e. for moderately growing size. In this situation also $\|\hat{\Gamma}_p^{-,II} - \Gamma_p^-\| \to 0$. It has been shown in Bauer et al. (2000) that subject to mild condi-

tions ensuring the convergence and invertibility of $\hat{W}_p^-$ the choice of the weighting matrix $\hat{W}_p^-$ does not influence the asymptotic variance of the estimates. Therefore, this choice is not critical and only $\hat{W}_p^- = (\hat{\Gamma}_p^{-,\Pi})^{1/2}$ will be considered.

Corresponding to $\hat{W}_f^+$ typical choices include the identity matrix and $(\hat{\Gamma}_f^{+,\Pi})^{-1/2}$ using

$$\hat{\Gamma}_f^{+,\Pi} = \hat{\Gamma}_y - \hat{\Gamma}_{y,u}\hat{\Gamma}_u^{-1}\hat{\Gamma}_{u,y}, \tag{2}$$

where $\hat{\Gamma}_y$ stands for the sample variance of $Y_{t,f}^+$ and $\hat{\Gamma}_{y,u}$ denotes the sample covariance of $Y_{t,f}^+$ and $U_{t,f}^+$. In this paper the choice of the weighting $\hat{W}_f^+$ will be restricted depending on the choice of the integer f: If f is chosen to be fixed and finite, then $\hat{W}_f^+$ is assumed to be chosen such that $\|\hat{W}_f^+ - W_f^+\| = O(Q_T)$ for some nonsingular matrix W_f^+. For $f \to \infty$ only $\hat{W}_f^+ = (\hat{\Gamma}_f^{+,\Pi})^{-1/2}$ or a weighting matrix attached to a frequency weighting transfer function (cf. e.g. Bauer, 1998) are considered. Let the expectation be denoted with $\Gamma_f^{+,\Pi}$. Then analogous results hold true: The error $\|\hat{\Gamma}_f^{+,\Pi} - \Gamma_f^{+,\Pi}\|_2 \to 0$ and the two norm of $\Gamma_f^{+,\Pi}$ and thus of $\hat{\Gamma}_f^{+,\Pi}$ is bounded and its smallest singular value is bounded away from zero for $f = O((\log T)^a)$, $a < \infty$. The name canonical correlation analysis (CCA) will be reserved for the procedure using

$$\hat{W}_p^- = (\hat{\Gamma}_p^{-,\Pi})^{1/2} \quad \text{and} \quad \hat{W}_f^+ = (\hat{\Gamma}_f^{+,\Pi})^{-1/2}. \tag{3}$$

In the third step the difference between the two classes of procedures appears: Whereas the Larimore type of procedures use $\hat{\mathscr{K}}_p$ to continue, the MOESP type of procedures use $\hat{\mathscr{O}}_f$ (for details see Bauer, 1998, Chapter 3). In this paper only the Larimore type of procedures is dealt with.

3. Main results

The main idea of the considered class of algorithms is to estimate the state in a first step and to obtain the estimate of the system using this state estimate. Consider the estimate $\hat{\mathscr{K}}_p = \hat{S}\hat{V}_n'(\hat{W}_p^-)^{-1}$. Here $\hat{S} = [\hat{V}_n'(\hat{W}_p^-)^{-1}]_n^{-1}$ appears to be a convenient choice of $\hat{S}$, where $[X]_n$ denotes the submatrix containing the first n columns of X. Note that the only function of $\hat{S}$ is to change the coordinate system of the state. The estimated transfer function is identical for any choice of nonsingular $\hat{S}$. For the choice given above this is true (a.s. asymptotically), if the first n columns of $\mathscr{K}_p$ are linearily independent (in one and thus in any representation). This holds true on a generic subset of M_n, which is denoted by M_n^+. Let $(\hat{A}_c, \hat{B}_c, \hat{C}_c, \hat{D}_c, \hat{K}_c)$ denote the estimated system, which has been converted into the canonical form induced by the restriction that $[\mathscr{K}_p]_n = I_n$ and let $(A_c, B_c, C_c, D_c, K_c)$ denote the corresponding representation of the true system. Since the entries in a canonical form are system invariants, the estimation accuracy of two procedures can be assessed by comparing the asymptotic covariance matrix of the vectorization of the estimated system in a canonical form. This is done in the main result of this paper:

Theorem 1. *Let the output process y_t be generated by a system $(A_c, B_c, C_c, D_c, K_c)$, such that the corresponding pair of transfer functions is in M_n^+. The noise and the input sequence are assumed to fulfil the standard assumptions. Assume that the Larimore type of procedure using $\hat{W}_p^- = (\hat{\Gamma}_p^{-,\Pi})^{1/2}$ is used to estimate the system, where the true order is assumed to be known. Furthermore, it is assumed that no time delay is present, i.e. the entries in D_c are estimated and not restricted to zero. Additionally, it is assumed, that $p \geqslant -d \log T/(2 \log|\rho_0|), 1 < d < \infty$ and $p = o((\log T)^a)$ holds for some $a < \infty$, where T denotes the sample size and $\rho_0 = \lambda_{\max}(A_c - K_c C_c)$, where $\lambda_{\max}$ denotes an eigenvalue of maximum modulus. Corresponding to $\hat{W}_f^+$ it is assumed, that either $f \geqslant n$ is fixed and $\hat{W}_f^+$ is chosen such that there exists a nonsingular matrix W_f^+, where $\|\hat{W}_f^+ - W_f^+\| = O(Q_T)$, or that $f \to \infty$ and $\hat{W}_f^+$ is chosen according to Eq. (2). Then the asymptotic variance of* $\mathrm{vec}[\hat{A}_c - A_c, \hat{B}_c - B_c, \hat{C}_c - C_c, \hat{D}_c - D_c, \hat{K}_c - K_c]$ *is of the form*

$$M_1 M_1' + M_2[\Gamma_\infty^- \otimes \{W^\dagger[\mathscr{E}_f(I_f \otimes \Omega)\mathscr{E}_f'](W^\dagger)'\}]M_2', \tag{4}$$

where $W^\dagger = (\mathscr{O}_f' W_2 \mathscr{O}_f)^{-1}\mathscr{O}_f' W_2$, $W_2 = \lim (W_f^+)' W_f^+$ for $T \to \infty$. The matrices $M_1 \in \mathbb{R}^{[(n+s)(n+m)+ns] \times s(n+m)}$ and $M_2 \in \mathbb{R}^{[(n+s)(n+m)+ns] \times \infty}$ do not depend on f or W_f^+.

The theorem also has an immediate consequence, which is stated in the following corollary:

Corollary 2. *Expression (4) as a function of W_f^+ is minimized by the CCA choice of the weighting $W_f^+ = (\Gamma_f^{+,\Pi})^{-1/2}$ for each value of f. The minimum variance decreases monotonically in f for the CCA case.*

This theorem clarifies a long standing question about the optimal choices of the weighting matrices for the algorithms dealt with in this contribution. The implications of the theorem are that in the situation of known system order it is always (i.e. for any choice of f) optimal to use the CCA weighting scheme in any situation, where no input is present or the observed input is white noise. The theorem also suggests the use of $f \to \infty$ at some rate, which is in accordance with earlier simulation studies (cf. Bauer, 1998). It also shows that no choice of f finite can achieve the optimal accuracy in all cases, since the decrease with respect to f is in general strict. The subset of M_n, where finite f also leads to optimal estimates consists of ARX systems, as is easily seen from the form of the essential term of the asymptotic variance of the parameter estimates. Furthermore, the theorem provides a measure of how much of attainable accuracy one loses by using any method other than the optimal. The amount of accuracy, which is lost by using a small f is determined in

the case of using optimal weights by the magnitude of the noise zeros, as they govern the rate of exponential decrease in the matrix $\mathscr{E}_f^{-1}\mathscr{O}_f$.

Note that in the theorem it has been assumed that p tends to infinity as a function of the sample size. Therefore, the above expression should be viewed as the limit of the respective quantities for $p \to \infty$. It will be a part of the proof to demonstrate that this limit exists.

4. Proof of Theorem 1

The main structure of the proof is as follows: First, the problem of calculating the asymptotic variance of the estimated system matrices is reduced to the corresponding problem for the terms $\langle \varepsilon_t, x_t\rangle, \langle \varepsilon_t, u_t\rangle$ and $\hat{\mathscr{K}}_p - \mathscr{K}_p$. Here and below, the notation $\langle a_t, b_t\rangle = T^{-1}\sum_{t=p+1}^{T-f} a_t b_t'$ will be used with slight abuse of notation neglecting the dependence on the sample size T in the notation and using the same symbol for both the series $\{a_t\}_{t\in\mathbb{Z}}$ and the vector random variable a_t. Note that the last matrix is of size $n \times p(s+m)$ and thus the number of columns increases with the sample size under the assumptions of the theorem. Therefore, it is necessary to define the notion of asymptotic normality for vectors of growing size. Here asymptotic distribution is to be understood for the vectorization of the matrix in the sense of (Lewis & Reinsel, 1985): A zero mean vector $v_T \in \mathbb{R}^{p(T)}$ is said to be distributed asymptotically normal, if for any vector $l_T \in \mathbb{R}^{p(T)}$, such that

- $\sup_{T>0} \|l_T\|_1 \leqslant M$ for some $M < \infty$,
- $\|[l_T', 0] - l'\|_1 \to 0$ for $T \to \infty$ for some vector $l \in \ell_1$,
- $\mathbb{E}(l_T' v_T)^2 \to c$ for $T \to \infty$ for some $0 \leqslant c < \infty$,

the scalar product $l_T' v_T$ converges in distribution to a normal random variable.

In the next step it is shown that these three terms are uncorrelated and asymptotically normally distributed. The essential term will turn out to be the last one, as this is the only one depending on the user choices. The rest of the proof then deals with this term. The main steps are summarized in lemmas. The first lemma deals with the reduction of the problem to the three terms mentioned before:

Lemma 3. *Let the assumptions of Theorem* 1 *hold. Then*

$$\begin{aligned}\operatorname{vec}[\hat{A}_c - A_c, \hat{B}_c - B_c, \hat{C}_c - C_c, \hat{D}_c - D_c, \hat{K}_c - K_c] &= \bar{M}_1 \operatorname{vec}\left[\left\langle \varepsilon_t, \begin{bmatrix} x_t \\ u_t \end{bmatrix}\right\rangle\right] + \bar{M}_{2,p} \operatorname{vec}[\hat{\mathscr{K}}_p - \mathscr{K}_p] \\ &\quad + o_P(T^{-1/2})\end{aligned}$$

where $f_T = o_P(T^{-1/2})$ *means that* $T^{1/2} f_T \to 0$ *in probability. Here* $[\mathscr{K}_p]_n = I_n$ *and* $[\hat{\mathscr{K}}_p]_n = I_n$ *is assumed. Further* $\sup_{p>0} \|\bar{M}_{2,p}\|_1 < \infty$, $\|[\bar{M}_{2,p}, 0] - \bar{M}_2\|_1 \to 0$ *and* $\|\bar{M}_2\|_{Fr} < \infty$. $\bar{M}_1$ *and* $\bar{M}_2$ *do not depend on* f.

Proof. Consider the estimation of the system matrices using the estimate of the state sequence $\hat{x}_t = \hat{\mathscr{K}}_p Z_{t,p}^-$: This is done using ordinary least squares. Let $(\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{K})$ denote these estimates and let (A, B, C, D, K) denote the corresponding limits. In order to obtain the estimates $(\hat{A}_c, \hat{B}_c, \hat{C}_c, \hat{D}_c, \hat{K}_c)$ a state space transformation has to be applied. However, since this transformation is a nonlinear differentiable mapping of the system matrix estimates, it is sufficient to prove the result for $(\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{K})$.

The expressions for the estimation error are easily derived to be the following. Here (A, B, C, D, K) denotes the true system in the representation according to $[\mathscr{K}_p]_n = I_n$ and so does the true state x_t. Let $\Delta_t = \hat{x}_t - x_t$. Then

$$\begin{aligned}[\hat{C} - C, \hat{D} - D] &= [\langle \varepsilon_t - C\Delta_t, \hat{x}_t\rangle, \langle \varepsilon_t - C\Delta_t, u_t\rangle]\hat{M}^{-1}, \\ [\hat{A} - A, \hat{B} - B] &= [\langle \tilde{\Delta}_{t+1}, \hat{x}_t\rangle, \langle \tilde{\Delta}_{t+1}, u_t\rangle]\hat{M}^{-1}, \\ [\hat{K} - K] &= \langle \tilde{\Delta}_{t+1} + Bu_t - K\hat{\varepsilon}_t, \hat{\varepsilon}_t\rangle\langle \hat{\varepsilon}_t, \hat{\varepsilon}_t\rangle^{-1}, \end{aligned} \tag{5}$$

where

$$\hat{M} = \begin{bmatrix} \langle \hat{x}_t, \hat{x}_t\rangle & \langle \hat{x}_t, u_t\rangle \\ \langle u_t, \hat{x}_t\rangle & \langle u_t, u_t\rangle \end{bmatrix}$$

and $\tilde{\Delta}_{t+1} = \Delta_{t+1} + K\varepsilon_t - A\Delta_t$. Since $\sqrt{T}\langle \varepsilon_t - C\Delta_t, \hat{x}_t\rangle$ and $\sqrt{T}\langle \varepsilon_t - C\Delta_t, u_t\rangle$ converge in distribution (see e.g. Bauer, 1998), it follows that the inverse can be replaced with its expectation without changing the asymptotic distribution. Now $\langle \hat{x}_t, u_t\rangle \to 0$ due to the white noise assumption on the inputs and thus the estimation errors in $\hat{C}$ and in $\hat{D}$, respectively, can be treated separately. The same arguments hold for $\hat{A}$ and $\hat{B}$. Thus consider $\hat{C} - C$ first:

$$\begin{aligned}\hat{C} - C &\doteq \langle \varepsilon_t - C\Delta_t, \hat{x}_t\rangle\langle \hat{x}_t, \hat{x}_t\rangle^{-1} \\ &\doteq \langle \varepsilon_t, x_t\rangle\Sigma_x^{-1} - \langle C\Delta_t, x_t\rangle\Sigma_x^{-1} \\ &\doteq \langle \varepsilon_t, x_t\rangle\Sigma_x^{-1} - C(\hat{\mathscr{K}}_p - \mathscr{K}_p)\Gamma_p^- \mathscr{K}_p' \Sigma_x^{-1},\end{aligned}$$

where $\Sigma_x = \mathbb{E}x_t x_t'$. Here the error bound $\|\hat{\mathscr{K}}_p - \mathscr{K}_p\| = o(Q_T pf)$ has been used to show e.g. that $\langle \varepsilon_t, \hat{x}_t\rangle \doteq \langle \varepsilon_t, x_t\rangle$. Next deal with K:

$$\begin{aligned}\hat{K} - K &\doteq \langle \Delta_{t+1}, \varepsilon_t\rangle\Omega^{-1} + \langle -A\Delta_t - K(\hat{\varepsilon}_t - \varepsilon_t), \varepsilon_t\rangle\Omega^{-1} \\ &\doteq (\hat{\mathscr{K}}_p - \mathscr{K}_p)\begin{bmatrix} I_s \\ 0^{[p(s+m)-s]\times s} \end{bmatrix}.\end{aligned}$$

This follows from the error bound cited above and the uniform convergence of the sample covariances, as e.g. $\langle \Delta_t, \varepsilon_t\rangle = (\hat{\mathscr{K}}_p - \mathscr{K}_p)\langle Z_{t,p}^-, \varepsilon_t\rangle - \bar{A}^p\langle x_{t-p}, \varepsilon_t\rangle \doteq 0$ and also the fact that $\langle u_t, \hat{\varepsilon}_t\rangle = 0$ has been used. Corresponding to the estimation error in A it turns out to be more convenient to consider $\bar{A} = A - KC$ instead. The result for A then is immediate, of course:

$$\begin{aligned}\hat{\bar{A}} - \bar{A} &\doteq \langle \Delta_{t+1} - \bar{A}\Delta_t, x_t\rangle\Sigma_x^{-1} + \langle \Delta_{t+1}, C'\Omega^{-1}(\varepsilon_t + Du_t)\rangle \\ &\doteq [\hat{\mathscr{K}}_p - \mathscr{K}_p, 0^{n\times(m+s)}]\begin{bmatrix} \mathscr{H}_1, p \\ \Gamma_p^- \end{bmatrix}\mathscr{K}_p'\Sigma_x^{-1}\end{aligned}$$

$$-\bar{A}(\hat{\mathscr{K}}_p-\mathscr{K}_p)(\Gamma_p^-)\mathscr{K}_p'\Sigma_x^{-1}$$
$$+[\hat{\mathscr{K}}_p-\mathscr{K}_p]_{m+s}\begin{bmatrix}C\\ \Omega_u D'\Omega^{-1}C\end{bmatrix},$$

where again $\doteq$ denotes equality up to terms of order $o(T^{-1/2})$. Further,

$$\hat{B}-B\doteq\langle\tilde{\Delta}_{t+1},u_t\rangle\langle u_t,u_t\rangle^{-1}$$
$$\doteq(\hat{\mathscr{K}}_p-\mathscr{K}_p)\mathbb{E}Z_{t+1,p}^-u_t'(\mathbb{E}u_tu_t')^{-1}$$
$$+K\langle\varepsilon_t,u_t\rangle(\mathbb{E}u_tu_t')^{-1}.$$

Finally,

$$\hat{D}-D\doteq\langle\varepsilon_t-C\Delta_t,u_t\rangle\langle u_t,u_t\rangle^{-1}\doteq\langle\varepsilon_t,u_t\rangle(\mathbb{E}u_tu_t')^{-1}.$$

The remaining claims of the proof follow easily from these representations. In particular, the convergence properties for $M_{2,p}$ are derived using the exponential decrease in the elements of $\mathscr{K}_p$. This completes the proof of the lemma. □

The next lemma deals with the second order properties of the essential terms of the last lemma:

Lemma 4. *Under the assumptions of Theorem* 1 $\sqrt{T}\,\mathrm{vec}[\langle\varepsilon_t,x_t\rangle]$, $\sqrt{T}\,\mathrm{vec}[\langle\varepsilon_t,u_t\rangle]$ *and* $\sqrt{T}\,\mathrm{vec}[\hat{\mathscr{K}}_p-\mathscr{K}_p]$ *are asymptotically uncorrelated.*

Proof. In the proof again all system matrices are assumed to be in the canonical form. It has been shown in Bauer et al. (2000) that

$$(\hat{\mathscr{K}}_p-\mathscr{K}_p)=\mathscr{O}_f^\dagger(\hat{\beta}_z-\beta_z)P_{\mathscr{K}}+\mathrm{O}(\|\hat{\beta}_z-\beta_z\|\,\|\hat{W}_p^--W_p^-\|+\|\hat{\beta}_z-\beta_z\|^2),\tag{6}$$

where $\beta_z=\mathbb{E}Y_{t,f}^+(Z_{t,p}^-)'(\Gamma_p^-)^{-1}$ and $\mathscr{O}_f^\dagger=(\mathscr{O}_f'W_2\mathscr{O}_f)^{-1}\mathscr{O}_f'W_2$. Here $W_2=\lim_{T\to\infty}(W_f^+)'W_f^+$, where the limit also includes the possibility of f tending to infinity in the CCA case. For any of the proposed weighting matrices, $\|\hat{W}_p^--W_p^-\|=O(Q_Tfp)$. This follows from the uniform convergence of the sample covariances as stated e.g. in Hannan and Deistler (1988, Theorem 5.3.2). It also follows that $\|\hat{\beta}_z-\beta_z\|^2=o(T^{-1/2})$. Therefore for the asymptotic distribution the term $\mathscr{O}_f^\dagger(\hat{\beta}_z-\beta_z)P_{\mathscr{K}}$ is the essential one, the remaining terms do not show up in the asymptotic distribution, as they are $o(T^{-1/2})$.

$P_{\mathscr{K}}$ depends on p but this is not reflected in the notation. Note the fact, that this expression does not depend on the weighting $\hat{W}_p^-$ and that with respect to the weighting $\hat{W}_f^+$ only the expectation W_f^+ has an influence. Since $p\to\infty$ as a function of the sample size it follows that $\|[\beta_z,0^{fs\times\infty}]-\mathscr{O}_f\mathscr{K}\|_2=O(p|\rho_0|^p)=o(T^{-1/2})$, where ρ_0 denotes a zero of $H(q)$ of maximum modulus. Then let $Z_{t,p}^{-,\Pi}=Z_{t,p}^--\hat{\Gamma}_{z,u}\hat{\Gamma}_u^{-1}U_{t,f}^+$. Therefore,

$$\hat{\beta}_z-\beta_z\doteq\langle\mathscr{E}_fE_{t,f}^+,Z_{t,p}^{-,\Pi}\rangle\langle Z_{t,p}^{-,\Pi},Z_{t,p}^{-,\Pi}\rangle^{-1}$$
$$\doteq\mathscr{E}_f\langle E_{t,f}^+,Z_{t,p}^-\rangle(\Gamma_p^-)^{-1}$$

as follows from straightforward calculations using $\|\langle E_{t,f}^+,U_{t,f}^+\rangle\|=O(Q_Tf)$, $\|\langle Z_{t,p}^+,U_{t,f}^+\rangle\|=O(Q_T\sqrt{fp})$. Here the white noise assumption on the input is used in the last equation.

Note that

$$\mathbb{E}\,\mathrm{vec}[\langle\varepsilon_t,x_t\rangle]\,\mathrm{vec}[\langle\varepsilon_t,x_t\rangle]'=\frac{1}{T^2}\sum_{t,s=p+1}^{T-f}\mathbb{E}(x_tx_s'\otimes\varepsilon_t\varepsilon_s')$$

which essentially is equal to $1/T(\Sigma_x\otimes\Omega)$, where $\Sigma_x=\mathscr{K}\Gamma_\infty^-\mathscr{K}'=\mathbb{E}x_tx_t'$ and essentially again indicates equality up to terms of order $o(T^{-1/2})$. Analogously, it follows that $T\mathbb{E}\,\mathrm{vec}[\langle\varepsilon_t,u_t\rangle]\,\mathrm{vec}[\langle\varepsilon_t,u_t\rangle]'\doteq(\Omega_u\otimes\Omega)$, where $\Omega_u=\mathbb{E}u_tu_t'$ and

$$T\mathbb{E}\,\mathrm{vec}[\langle\varepsilon_t,x_t\rangle]\,\mathrm{vec}[\langle\varepsilon_t,u_t\rangle]'=0.$$

Next, consider the cross moments between $\langle\varepsilon_t,x_t\rangle$ and $\hat{\mathscr{K}}_p-\mathscr{K}_p$. Note that for the (i,j)th component of $\mathrm{vec}[\langle\varepsilon_t,x_t\rangle]\doteq\mathrm{vec}[\langle\varepsilon_t,\mathscr{K}_pZ_{t,p}^-\rangle]$ and any linear combination $\sqrt{T}v'\mathscr{O}_f^\dagger\mathscr{E}_f\langle E_{t,f}^+,Z_{t,p}^-\rangle(\Gamma_p^-)^{-1}P_{\mathscr{K}}V_p$ for some vectors $v\in\mathbb{R}^n$ and $V_p\in\mathbb{R}^{p(s+m)}$ such that $\|[V_p',0^{1\times\infty}]-V'\|_1\to0$, where V is a vector in ℓ_1 having elements decreasing exponentially, one obtains that

$$\mathbb{E}\sum_{t,s=p+1}^{T-f}\varepsilon_{t,i}(\mathscr{K}_{p,j}Z_{t,p}^-)(v'\mathscr{O}_f^\dagger\mathscr{E}_fE_{s,f}^+)(V_p'P_{\mathscr{K}}'(\Gamma_p^-)^{-1}Z_{s,p}^-)$$
$$=\sum_t\sum_{s=\bar{s}}^t\mathbb{E}\varepsilon_{t,i}(v'\mathscr{O}_f^\dagger\mathscr{E}_fE_{s,f}^+)$$
$$\times(\mathbb{E}\mathscr{K}_{p,j}Z_{t,p}^-(Z_{s,p}^-)')(\Gamma_p^-)^{-1}P_{\mathscr{K}}V_p$$
$$=\sum_t\sum_{s=\bar{s}}^t\mathbb{E}\varepsilon_{t,i}(v'\mathscr{O}_f^\dagger\mathscr{E}_fE_{s,f}^+)$$
$$\times\tilde{\mathscr{K}}_{p,j}\begin{bmatrix}\mathscr{H}_{t-s,p}\\ \Gamma_p^-\end{bmatrix}(\Gamma_p^-)^{-1}P_{\mathscr{K}}V_p$$
$$=\sum_t\sum_{s=\bar{s}}^t\mathbb{E}\varepsilon_{t,i}(v'\mathscr{O}_f^\dagger\mathscr{E}_fE_{s,f}^+)$$
$$\times\tilde{\mathscr{K}}_{p,j}\begin{bmatrix}\tilde{\mathscr{O}}_{t-s}\mathscr{K}_p\\ I_p\end{bmatrix}P_{\mathscr{K}}V_p+o(T)$$
$$=o(T),$$

where $\tilde{\mathscr{K}}_{p,j}=[\mathscr{K}_{p,j},0^{1\times(m+s)(t-s)}]$, the sum is over $t=p+1,\dots,T-f$ and where $\mathscr{H}_{j,p}=\mathbb{E}Z_{t,j}^-(Z_{t-j,p}^-)'=\tilde{\mathscr{O}}_j\mathscr{K}_p\Gamma_p^-+o(T^{-1/2})$ and $\bar{s}=\max\{p+1,t-f+1\}$. Here mostly $\mathscr{K}_pP_{\mathscr{K}}=0$ and $\|\mathscr{H}_{j,p}(\Gamma_p^-)^{-1}-\tilde{\mathscr{O}}_j\mathscr{K}_p\|=$

$o(T^{-1/2})$ for p as specified in the theorem is used (for a proof of the latter statement see e.g. Bauer, 1998). The latter fact is used in the replacement involved in the third equality. The convergence follows from the convergence assumptions on V_p and the analogous property of $\mathscr{K}_p$. These properties allow the replacement of the limit for $p \to \infty$ by the expression obtained for $p = \infty$, which will be done frequently in the following in order to simplify notations. For the covariance of elements of $\langle \varepsilon_t, u_t \rangle$ with $v'(\hat{\mathscr{K}}_p - \mathscr{K}_p)V_p$ analogous arguments hold. Therefore, the three terms are asymptotically uncorrelated. This completes the proof of the lemma. □

Thus the only effect of the user choices f and $\hat{W}_f^+$ is hidden in the term $\hat{\mathscr{K}}_p - \mathscr{K}_p$, which can be examined independent of the other terms due to the uncorrelatedness. Consider the variance of $\mathscr{O}_f^\dagger \langle \mathscr{E}_f E_{t,f}^+, Z_{t,p}^- \rangle (\Gamma_p^-)^{-1} P_{\mathscr{K}} v_p$ for some vector $v_p \in \mathbb{R}^{p(m+s)}$ such that $\|[v_p', 0^{1\times\infty}] - v'\|_1 \to 0$ for some vector v in ℓ_1 having elements decreasing exponentially:

$$\frac{1}{T^2} \sum_{s,t=p+1}^{T-f} \mathbb{E}\tilde{E}_{t,f}^+ (Z_{t,p}^{-,\Pi})' \tilde{v}_p \tilde{v}_p' Z_{s,p}^{-,\Pi} (\tilde{E}_{s,f}^+)'$$

$$= \frac{1}{T} \sum_{l=1-f}^{f-1} \mathbb{E}\tilde{E}_{t,f}^+ (\tilde{E}_{t+l,f}^+)' (\tilde{v}_p' \mathbb{E} Z_{t,p}^- (Z_{t+l,p}^-)' \tilde{v}_p) + o(T^{-1}). \tag{7}$$

Here $\tilde{E}_{t,f}^+ = \mathscr{O}_f^\dagger \mathscr{E}_f E_{t,f}^+$ and $\tilde{v}_p = (\Gamma_p^-)^{-1} P_{\mathscr{K}} v_p$. Note that for $l = 0$ the part due to $E_{t,f}^+$ is equal to $\mathscr{O}_f^\dagger \mathscr{E}_f (I_f \otimes \Omega) \mathscr{E}_f' (\mathscr{O}_f^\dagger)'$. This is the central term in the expression for the asymptotic variance given in the theorem. From Lemma 3 it follows, that a matrix $\bar{M}_2$ as used above exists. The construction of this matrix will be clarified below. The theorem then is proved, if for any vectors $v_p, \bar{v}_p$ postmultiplying $\hat{\mathscr{K}}_p - \mathscr{K}_p$ in the equations for the transformed system $(\hat{A}_c, \hat{B}_c, \hat{C}_c, \hat{D}_c, \hat{K}_c)$ analogous to Eqs. (5) it holds that $\mathbb{E} v_p' P_{\mathscr{K}}' (\Gamma_p^-)^{-1} Z_{t,p}^- (Z_{t-j,p}^-)' (\Gamma_p^-)^{-1} P_{\mathscr{K}} \bar{v}_p \to 0$, $j \neq 0$. It is in this part of the proof, where the white noise assumption on the input is essential. Most of the arguments used up to now hold also for more general inputs, in particular, an analogue to the variance expression given above exists. It is convenient to split the proof into two separate cases.

4.1. The case $n \leqslant (s+m)$

On examining the expressions given in Lemma 3, one observes that only a number of terms are multiplying $\hat{\mathscr{K}}_p - \mathscr{K}_p$: These terms converge to $\Gamma_\infty^- \mathscr{K}'$, $[\mathscr{H}_1', \Gamma_\infty^-]' \mathscr{K}'$, $[0^{m\times s}, I_m, 0^{m\times\infty}]'$ and $[I_s, 0^{s\times\infty}]'$, where convergence is in ℓ_1 norm as required above for the vectors v_p. All these matrices have elements decreasing exponentially. Note that

$$\mathbb{E} P_{\mathscr{K}}' (\Gamma_\infty^-)^{-1} Z_{t,\infty}^- (Z_{t-j,\infty}^-)' (\Gamma_\infty^-)^{-1} P_{\mathscr{K}}$$

$$= P_{\mathscr{K}}' (\Gamma_\infty^-)^{-1} \begin{bmatrix} \mathscr{H}_j \\ \Gamma_\infty^- \end{bmatrix} (\Gamma_\infty^-)^{-1} P_{\mathscr{K}}$$

$$= P_{\mathscr{K}}' (\Gamma_\infty^-)^{-1} \begin{bmatrix} 0^{j(s+m)\times\infty} \\ I_\infty \end{bmatrix} P_{\mathscr{K}}$$

$$= P_{\mathscr{K}}' \begin{bmatrix} 0^{j(s+m)\times\infty} \\ (\Gamma_\infty^-)^{-1} \end{bmatrix} P_{\mathscr{K}} = \begin{bmatrix} 0^{j(s+m)\times\infty} \\ (\Gamma_\infty^-)^{-1} \end{bmatrix} P_{\mathscr{K}} \tag{8}$$

evaluating the expression for $p = \infty$ rather than dealing with the limit. Here $\mathscr{H}_j = \mathscr{H}_{j,\infty}$. Using the exponential decrease, however, it is straightforward to show that in all situations, where the expression occurs, the limit and the expression for $p = \infty$ coincide. The next to last equality follows from the block matrix inversion formula, which gives

$$(\Gamma_\infty^-)^{-1} = \begin{bmatrix} 0^{(s+m)\times(s+m)} & 0^{(s+m)\times\infty} \\ 0^{\infty\times(s+m)} & (\Gamma_\infty^-)^{-1} \end{bmatrix} + \begin{bmatrix} I_{s+m} \\ -(\Gamma_\infty^-)^{-1} \mathscr{H}_1' \end{bmatrix} \Pi_\gamma^{-1} [I_{s+m}, -\mathscr{H}_1 (\Gamma_\infty^-)^{-1}], \tag{9}$$

where $\Pi_\gamma = (\gamma_z(0) - \mathscr{H}_1 (\Gamma_\infty^-)^{-1} \mathscr{H}_1') > 0$. The projection $P_{\mathscr{K}}' = I_\infty - \mathscr{K}'[I_n, 0^{n\times\infty}]$ and thus the last equality follows from $n \leqslant (s+m)$. Premultiplying with the above-mentioned terms from the left shows, that the terms for $l \neq 0$ in Eq. (7) do not matter in the case $n \leqslant (s+m)$. Take for e.g.

$$\mathbb{E}\mathscr{K}[\mathscr{H}_1', \Gamma_\infty^-] P_{\mathscr{K}}' (\Gamma_\infty^-)^{-1} Z_{t,\infty}^- (Z_{t-j,\infty}^-)' (\Gamma_\infty^-)^{-1} P_{\mathscr{K}}$$

$$= \mathscr{K}[\mathscr{H}_1', \Gamma_\infty^-] \begin{bmatrix} 0^{j(s+m)\times\infty} \\ (\Gamma_\infty^-)^{-1} \end{bmatrix} P_{\mathscr{K}} = 0.$$

This shows, that in the case $n \leqslant (s+m)$ only the term for $l = 0$ in Eq. (7) is nonzero and thus the theorem holds in this case.

4.2. The case $n > (s+m)$

The theorem gives the asymptotic variance of the system matrix estimates $(\hat{A}_c, \hat{B}_c, \hat{C}_c, \hat{D}_c, \hat{K}_c)$. In order to show that only the term for $l = 0$ in Eq. (7) is nonzero, it is sufficient to show this fact for any invertible (possibly nonlinear) transformation of these matrices. It proves to be convenient to consider $(\hat{\tilde{A}}_c, \hat{\tilde{B}}_c, \hat{C}_c, \hat{D}_c, \hat{K}_c)$, where $\hat{\tilde{A}}_c = \hat{A}_c - \hat{K}_c \hat{C}_c$, $\hat{\tilde{B}}_c = \hat{B}_c - \hat{K}_c \hat{C}_c$. These estimates are obtained by transforming the estimates of the subspace algorithm $(\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{K})$ into the particular canonical form defined by $[\mathscr{K}_p]_n = I_n$, which is done using the transformation matrix $\hat{S}$, which is defined as $\hat{S} = [[\hat{K}, \hat{\tilde{B}}], \ldots, \hat{\tilde{A}}^{n-1} [\hat{K}, \hat{\tilde{B}}]]_n$. Then e.g. $\hat{C}_c - C_c = \hat{C}\hat{S} - C = (\hat{C} - C)\hat{S} + C(\hat{S} - I_n)$, where I_n is the limit of $\hat{S}$ as follows in a straightforward fashion from the consistency results for $\hat{\mathscr{K}}_p$ and the sample covariances used in the regression. It follows from the normalization $[\hat{\mathscr{K}}_p]_n = [\mathscr{K}_p]_n = I_n$ that in

the current case the contribution of $\hat{\mathscr{K}}_p - \mathscr{K}_p$ to the error in $\hat{B}_c$, $\hat{D}_c$ and $\hat{K}_c$ converges to zero (see Lemma 3). Therefore only $\hat{C}_c$ and $\hat{A}_c$ have to be dealt with.

Note that the columns of $\hat{S}$ follow a recursive pattern. Therefore, denoting $\hat{T}_i = \hat{\bar{A}}^i[\hat{K}, \hat{\bar{B}}]$, $T_i = \bar{A}^i[K, \bar{B}]$ the following recursion is obtained for $i = 0, 1, \ldots$:

$$\begin{aligned}
&\hat{T}_{i+1} - T_{i+1} \\
&\quad = \hat{\bar{A}}(\hat{T}_i - T_i) + (\hat{\bar{A}} - \bar{A})T_i \\
&\quad \doteq \Delta_{\mathscr{K}} \begin{bmatrix} \mathscr{H}_1 \\ \Gamma_\infty^- \end{bmatrix} \mathscr{K}' \Sigma_x^{-1} T_i - \bar{A}\Delta_{\mathscr{K}} \Gamma_\infty^- \mathscr{K}' \Sigma_x^{-1} T_i + \bar{A}(\hat{T}_i - T_i) \\
&\quad \doteq \Delta_{\mathscr{K}} \begin{bmatrix} \mathscr{H}_1 \\ \Gamma_\infty^- \end{bmatrix} \mathscr{K}' \Sigma_x^{-1} T_i - \bar{A}^{i+1} \Delta_{\mathscr{K}} \Gamma_\infty^- \mathscr{K}' \Sigma_x^{-1} T_0 \\
&\qquad + \sum_{j=1}^{i} \bar{A}^j \Delta_{\mathscr{K}} \left\{ \begin{bmatrix} \mathscr{H}_1 \\ \Gamma_\infty^- \end{bmatrix} \mathscr{K}' \Sigma_x^{-1} - \Gamma_\infty^- \mathscr{K}' \Sigma_x^{-1} \bar{A} \right\} T_{i-j},
\end{aligned}$$

where $\Delta_{\mathscr{K}} = [\hat{\mathscr{K}}_p, 0^{n\times\infty}] - \mathscr{K}$ and where $\mathscr{K} = \mathscr{K}_\infty$ as well as $\Delta_{\mathscr{K}}[I_{s+m}, 0]' = 0$ are used. The recursion is started at $\hat{T}_0 = T_0$.

The next lemma shows that for the second and the third term in this recursion only the term corresponding to $l = 0$ in Eq. (7) is of relevance:

Lemma 5. *Under the conditions of Theorem* 1 *for* $j > 0$ *holds that*

$$\mathbb{E} T_0' \Sigma_x^{-1} \mathscr{K} \Gamma_\infty^- \tilde{Z}_{t,\infty}^- (\tilde{Z}_{t-j,\infty}^-)' = o(1),$$

$$\begin{aligned}
&\mathbb{E}\{\Sigma_x^{-1} \mathscr{K}[\mathscr{H}_1', \Gamma_\infty^-] - \bar{A}' \Sigma_x^{-1} \mathscr{K} \Gamma_\infty^-\} \tilde{Z}_{t,\infty}^- (\tilde{Z}_{t-j,\infty}^-)' \\
&\quad = o(1),
\end{aligned}$$

where $\tilde{Z}_{t,\infty}^- = P_{\mathscr{K}}'(\Gamma_\infty^-)^{-1} Z_{t,\infty}^-$. *Here* $\mathscr{K}[\mathscr{H}_1', \Gamma_\infty^-]\mathscr{K}' = \mathbb{E} x_t (x_{t+1})' = \Sigma_x A'$ *has been used.*

Proof. For the first term note that

$$\begin{aligned}
&\mathbb{E} T_0' \Sigma_x^{-1} \mathscr{K} \Gamma_\infty^- \tilde{Z}_{t,\infty}^- (\tilde{Z}_{t-j,\infty}^-)' \\
&\quad = T_0' \Sigma_x^{-1} \mathscr{K} \Gamma_\infty^- P_{\mathscr{K}}'(\Gamma_\infty^-)^{-1} \begin{bmatrix} 0^{j(s+m)\times\infty} \\ I_\infty \end{bmatrix} P_{\mathscr{K}} \\
&\quad = T_0' \Sigma_x^{-1} (\mathscr{K} - \Sigma_x [I_n, 0^{n\times\infty}](\Gamma_\infty^-)^{-1}) \begin{bmatrix} 0^{j(s+m)\times\infty} \\ I_\infty \end{bmatrix} P_{\mathscr{K}} \\
&\quad = -T_0' [I_n, 0^{n\times\infty}](\Gamma_\infty^-)^{-1} \begin{bmatrix} 0^{j(s+m)\times\infty} \\ I_\infty \end{bmatrix} P_{\mathscr{K}} \\
&\quad = -[I_{s+m}, 0^{(s+m)\times\infty}](\Gamma_\infty^-)^{-1} \begin{bmatrix} 0^{j(s+m)\times\infty} \\ I_\infty \end{bmatrix} P_{\mathscr{K}} = 0.
\end{aligned}$$

Here again the matrix inversion lemma has been used together with the property that $T_0 = [I_n]_{s+m}$. The expressions are evaluated at $p = \infty$ rather than dealing with the limit, which is possible due to the exponential decrease in $\mathscr{K}$.

The conjecture for the second term follows in a similar manner from:

$$\begin{aligned}
&\mathbb{E}\{\Sigma_x^{-1} \mathscr{K}[\mathscr{H}_1', \Gamma_\infty^-] - \bar{A}' \Sigma_x^{-1} \mathscr{K} \Gamma_\infty^-\} \tilde{Z}_{t,\infty}^- (\tilde{Z}_{t-j,\infty}^-)' \\
&\quad = \{\Sigma_x^{-1} \mathscr{K}[\mathscr{H}_1', \Gamma_\infty^-] - \bar{A}' \Sigma_x^{-1} \mathscr{K} \Gamma_\infty^-\} \\
&\qquad \times P_{\mathscr{K}}'(\Gamma_\infty^-)^{-1} \begin{bmatrix} 0 \\ I_\infty \end{bmatrix} P_{\mathscr{K}} \\
&\quad = \{\Sigma_x^{-1} \mathscr{K}[\mathscr{H}_1', \Gamma_\infty^-] - \bar{A}' \Sigma_x^{-1} \mathscr{K} \Gamma_\infty^-\} \\
&\qquad \times \mathscr{K}'[I_n, 0](\Gamma_\infty^-)^{-1} \begin{bmatrix} 0 \\ I_\infty \end{bmatrix} P_{\mathscr{K}} \\
&\quad = -[C', 0^{n\times m}] T_0' [I_n, 0^{n\times\infty}](\Gamma_\infty^-)^{-1} \begin{bmatrix} 0 \\ I_\infty \end{bmatrix} P_{\mathscr{K}} = 0,
\end{aligned}$$

where the dimensions of the zero blocks are omitted for notational simplicity. This completes the proof of the lemma. □

Therefore, the strategy in the remaining part of the proof will be to isolate these terms in all the occurring expressions. Note that due to the normalization of $\mathscr{K}_p$ it follows that each column of I_n is equal to a column of T_i for some index i. Let $\bar{n} = \lfloor n/(s+m) \rfloor$, where $\lfloor x \rfloor$ denotes the greatest integer smaller than x. Further let $\bar{m} = n - \bar{n}(s+m)$. Then the columns of T_i, $0 \leqslant i < \bar{n}$ and the first $\bar{m}$ columns of $T_{\bar{n}}$ are vectors of the canonical basis. Therefore, the columns of $\hat{C}_c - C_c$ are equal to the columns of the following expressions for the respective integers i:

$$\begin{aligned}
&(\hat{C} - C)\hat{T}_{i-1} + C(\hat{T}_{i-1} - T_{i-1}) \\
&\quad \doteq \langle \varepsilon_t, x_t \rangle \Sigma_x^{-1} T_{i-1} - C\bar{A}^{i-1} \Delta_{\mathscr{K}}(\Gamma_\infty^-) \mathscr{K}' \Sigma_x^{-1} T_0 \\
&\qquad + \sum_{j=0}^{i-2} C\bar{A}^j \Delta_{\mathscr{K}} \left\{ \begin{bmatrix} \mathscr{H}_1 \\ \Gamma_\infty^- \end{bmatrix} \mathscr{K}' \Sigma_x^{-1} \right. \\
&\qquad \left. - \Gamma_\infty^- \mathscr{K}' \Sigma_x^{-1} \bar{A} \right\} T_{i-j-2}. \qquad (10)
\end{aligned}$$

Application of the results of the last lemma also shows the result for $\hat{C}_c - C_c$ and thus only $\hat{A}_c - A_c$ is left for investigation.

Note that $\hat{\bar{A}}_c - \bar{A}_c \doteq -(\hat{S} - I_n)\bar{A} + (\hat{\bar{A}} - \bar{A}) + \bar{A}(\hat{S} - I_n)$. If i is such that T_i is a block column of the identity matrix, then $\hat{S}T_i = \hat{T}_i$ and thus

$$\begin{aligned}
&-(\hat{S} - I_n)\bar{A}T_{i-1} + (\hat{\bar{A}} - \bar{A})T_{i-1} + \bar{A}(\hat{S} - I_n)T_{i-1} \\
&\quad = -(\hat{S} - I_n)T_i + (\hat{\bar{A}} - \bar{A})T_{i-1} + \bar{A}(\hat{T}_{i-1} - T_{i-1}) = 0.
\end{aligned}$$

This is also true for the first $\bar{m}$ columns of $T_{\bar{n}-1}$. Therefore, it remains to deal with the matrix $W = [w_0', \ldots, w_{\bar{n}-1}', \tilde{w}_{\bar{n}}']'$, where W denotes the matrix built of the last $s+m$ columns of $\bar{A}$. Here $w_i \in \mathbb{R}^{(s+m)\times(s+m)}$, $i = 0, \ldots, \bar{n} - 1$, $\tilde{w}_{\bar{n}} \in \mathbb{R}^{\bar{m}\times(s+m)}$.

In order to unify the notation let $w_{\bar{n}} = [\tilde{w}'_{\bar{n}}, 0^{(s+m)\times(s+m-\bar{m})}]'$, $s_{\bar{n}-1} = [0^{(s+m-\bar{m})\times\bar{m}}, I_{s+m-\bar{m}}]'$, $s_{\bar{n}} = [I_{\bar{m}}, 0^{\bar{m}\times(s+m-\bar{m})}]'$, $\bar{s}_{\bar{n}-1} = [I_{s+m-\bar{m}}, 0^{(s+m-\bar{m})\times\bar{m}}]'$ and $\bar{s}_{\bar{n}} = [0^{\bar{m}\times(s+m-\bar{m})}, I_{\bar{m}}]'$, respectively. Finally define

$$Y = \begin{bmatrix} \mathscr{H}_1 \\ \Gamma_\infty^- \end{bmatrix} \mathscr{K}' \Sigma_x^{-1}.$$

Then for $i = \bar{n} - 1$ and $i = \bar{n}$ one obtains

$$\begin{aligned}
&(\hat{T}_i - T_i - (\hat{S} - I_n)\bar{A}T_{i-1})s_i \\
&\quad = -\sum_{j=0}^{\bar{n}} (\hat{T}_j - T_j) w_j \bar{s}_i + (\hat{T}_i - T_i) s_i \\
&\quad = -\sum_{j=1}^{\bar{n}} \Delta_{\mathscr{K}} Y T_{j-1} w_j \bar{s}_i + \bar{A}^j \Delta_{\mathscr{K}} (\Gamma_\infty^-) \mathscr{K}' \Sigma_x^{-1} T_0 w_j \bar{s}_i \\
&\qquad + \Delta_{\mathscr{K}} Y T_{i-1} s_i - \bar{A}^i \Delta_{\mathscr{K}} (\Gamma_\infty^-) \mathscr{K}' \Sigma_x^{-1} T_0 s_i \\
&\qquad - \sum_{j=1}^{\bar{n}} \sum_{l=1}^{j-1} \bar{A}^l \Delta_{\mathscr{K}} \{ Y - \Gamma_\infty^- \mathscr{K}' \Sigma_x^{-1} \bar{A} \} T_{j-l-1} w_j \bar{s}_i \\
&\qquad + \sum_{j=1}^{i-1} \bar{A}^j \Delta_{\mathscr{K}} \{ Y - \Gamma_\infty^- \mathscr{K}' \Sigma_x^{-1} \bar{A} \} T_{i-j-1} s_i.
\end{aligned}$$

It has been shown before, that for the terms postmultiplied by T_0 and the terms including $\{Y - \Gamma_\infty^- \mathscr{K}' \Sigma_x^{-1} \bar{A}\}$ only the covariance of the respective term matters in the asymptotic variance as stated in Eq. (7), but not the covariances at lags $l \neq 0$. The only terms of concern are the remaining ones, which are equal to

$$\begin{aligned}
&\Delta_{\mathscr{K}} Y \left(T_{i-1} s_i - \sum_{j=1}^{\bar{n}} T_{j-1} w_j \bar{s}_i \right) \\
&\quad = \Delta_{\mathscr{K}} Y \left(\begin{bmatrix} 0^{(s+m)\times(s+m)} \\ 0^{(s+m)\times(s+m)} \\ \vdots \\ 0^{\bar{m}\times(s+m)} \\ I_{s+m} \end{bmatrix} + \begin{bmatrix} -w_1 \\ -w_2 \\ \vdots \\ -\tilde{w}_{\bar{n}} \\ 0^{(s+m)\times(s+m)} \end{bmatrix} \right) \bar{s}_i.
\end{aligned}$$

Denoting the matrix in brackets on the right-hand side with $\tilde{W}$ we obtain $\bar{A}\tilde{W} = -T_0 w_1$. Also, $A\tilde{W} = \bar{A}\tilde{W} + T_0[C', 0^{n\times m}]'\tilde{W} = T_0 V$ for some matrix V. Therefore

$$\begin{aligned}
&\mathbb{E}\tilde{W}' \Sigma_x^{-1} \mathscr{K} [\mathscr{H}'_1, \Gamma_\infty^-] P'_{\mathscr{K}} (\Gamma_\infty^-)^{-1} Z_{t,\infty}^- (Z_{t-j,\infty}^-)' (\Gamma_\infty^-)^{-1} P_{\mathscr{K}} \\
&\quad = \tilde{W}' \Sigma_x^{-1} \mathscr{K} [\mathscr{H}'_1, \Gamma_\infty^-] P'_{\mathscr{K}} (\Gamma_\infty^-)^{-1} \begin{bmatrix} 0 \\ I_\infty \end{bmatrix} P_{\mathscr{K}} \\
&\quad = \tilde{W}' \Sigma_x^{-1} \mathscr{K} [\mathscr{H}'_1, \Gamma_\infty^-] (I_\infty - \mathscr{K}' [I_n, 0]) \\
&\qquad \times (\Gamma_\infty^-)^{-1} \begin{bmatrix} 0 \\ I_\infty \end{bmatrix} P_{\mathscr{K}} \\
&\quad = -\tilde{W}' A' [I_n, 0] (\Gamma_\infty^-)^{-1} \begin{bmatrix} 0 \\ I_\infty \end{bmatrix} P_{\mathscr{K}} = 0,
\end{aligned}$$

again omitting the dimensions of the zero blocks. Here the second last equation follows from $\mathscr{K}[\mathscr{H}'_1, \Gamma_\infty^-]\mathscr{K}' = \mathbb{E}x_t(x_{t+1})' = \Sigma_x A'$. Summing up the findings up to now it follows that

$$\begin{aligned}
&\operatorname{vec}[\hat{\bar{A}}_c - \bar{A}, \hat{\bar{B}}_c - \bar{B}, \hat{C}_c - C, \hat{D}_c - D, \hat{K}_c - K] \\
&\quad \doteq \bar{M}_1 \operatorname{vec} \left\langle \varepsilon_t, \begin{bmatrix} x_t \\ u_t \end{bmatrix} \right\rangle + \bar{M}_{2,p} \operatorname{vec}(\hat{\mathscr{K}}_p - \mathscr{K}_p). \qquad (11)
\end{aligned}$$

Here the matrices $\bar{M}_1$ and $\bar{M}_{2,p}$ can be found by tracing the computations so far. Additionally, it has been shown that

$$\begin{aligned}
&\mathbb{E}\left[\bar{M}_1 \left(\begin{bmatrix} x_t \\ u_t \end{bmatrix} \otimes \varepsilon_t \right) \right] [(\tilde{Z}_{t+j,p}^- \otimes \mathscr{O}^\dagger \mathscr{E}_f E_{t+j,f}^+)]' \\
&\quad (\bar{M}_{2,p})' \to 0, \\
&\bar{M}_{2,p} \mathbb{E}(\tilde{Z}_{t,p}^- \otimes \mathscr{O}^\dagger \mathscr{E}_f E_{t,f}^+)(\tilde{Z}_{t+j,p}^- \otimes \mathscr{O}^\dagger \mathscr{E}_f E_{t+j,f}^+)' \\
&\quad (\bar{M}_{2,p})' \to 0,
\end{aligned}$$

where the first limit holds for all j and the second for $j \neq 0$. From the definition of $\bar{M}_1$ and $\bar{M}_2$ it follows that these matrices do not depend on W_2 or f. Using the expressions given above and the arguments given in the proof of Lemma 5 the corresponding expressions for M_1 and M_2 follow in a straightforward fashion. For the fixed case f it follows directly, that the asymptotic covariance matrix is of the form given in the theorem. In the case $f \to \infty$ it can be shown in straightforward but tedious operations, that the terms given above are of order $o(T^{-1/2})$ uniformly in $f = O((\log T)^a)$. Here the form of the weighting matrices is used to show e.g. that the variance of $\mathscr{O}_f^\dagger \mathscr{E}_f E_{t,f}^+$ is bounded uniformly in f. Therefore, the expression for the asymptotic variance of the estimated system matrices also holds in the case $f \to \infty$ for `CCA` weights. This completes the proof of the theorem. □

4.3. Proof of Corollary 2

Note that

$$\begin{aligned}
&\mathscr{O}^\dagger \mathscr{E}_f (I_f \otimes \Omega) \mathscr{E}'_f (\mathscr{O}^\dagger)' \\
&\quad = (\mathscr{O}'_f W_2 \mathscr{O}_f)^{-1} \mathscr{O}'_f W_2 [\mathscr{E}_f (I_f \otimes \Omega) \mathscr{E}'_f] W_2 \mathscr{O}_f (\mathscr{O}'_f W_2 \mathscr{O}_f)^{-1}.
\end{aligned}$$

This is minimized by $W_2^\circ = [\mathscr{E}_f (I_f \otimes \Omega) \mathscr{E}'_f]^{-1}$ with minimum $(\mathscr{O}'_f (\mathscr{E}'_f)^{-1} (I_f \otimes \Omega^{-1}) \mathscr{E}_f^{-1} \mathscr{O}_f)^{-1}$. Some matrix algebra shows, that this gives the identical variance with any choice $W_2 = W_2^\circ + W_2^\circ \mathscr{O}_f W \mathscr{O}'_f W_2^\circ$, such that W_2 is invertible. Thus the `CCA` weightings minimize the variance of the

estimated system, since in this case $W_2 = (\Gamma_y^{+,I})^{-1}$, where $\Gamma_y^{+,I} = \mathscr{E}_f(I_f \otimes \Omega)\mathscr{E}'_f + \mathscr{O}_f \Sigma_x \mathscr{O}'_f$. Therefore, due to the matrix inversion lemma

$$(\Gamma_y^{+,I})^{-1} = (\mathscr{E}_f(I_f \otimes \Omega)\mathscr{E}'_f)^{-1} + (\mathscr{E}_f(I_f \otimes \Omega)\mathscr{E}'_f)^{-1}\mathscr{O}_f W \mathscr{O}'_f(\mathscr{E}_f(I_f \otimes \Omega)\mathscr{E}'_f)^{-1}$$

for suitable matrix W. The lower diagonal block Toeplitz structure of $\mathscr{E}_f$ then shows, that this minimum variance decreases monotonically in f, since it ensures, that $\mathscr{E}_f^{-1}\mathscr{O}_f$ is a submatrix of $\mathscr{E}_\infty^{-1}\mathscr{O}_\infty$. This completes the proof. □

5. Numerical illustration

In this section two examples will be given, which illustrate the findings of the last section. As a first case consider the following single-input single-output system without exogenous inputs having state dimension three:

$$A = \begin{bmatrix} -0.532 & 0.4639 & 0.2855 \\ 1 & 0 & -0.2568 \\ 0 & 1 & 0.0054 \end{bmatrix}, \quad K = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

$$C = [-0.532 \quad 0.4639 \quad -0.0413].$$

The noise is assumed to be white with variance equal to 1. For this system the asymptotic variance is compared to the Cramer–Rao bound using the following measure: Let F_I denote the Fisher information matrix with respect to the particular canonical form used in this paper. Then it is well known, that the Cramer–Rao bound for the estimation is equal to F_I^{-1}. Thus, let $V_f(W_f^+)$ denote the asymptotic variance of parameter estimates obtained from the subspace procedure using the integer f and the weighting matrix W_f^+. Then the measure $E_f = \mathrm{tr}[V_f(W_f^+)F_I] - 2ns - (n+s)m$ is used. For an efficient estimation method this is equal to zero, otherwise positive. The upper plot in Fig. 1 shows this measure for the two weighting schemes denoted as `CCA` (i.e. $\hat{W}_f^+ = (\hat{\Gamma}_f^+)^{-1/2}$) and `N4SID` (i.e. $W_f^+ = I_{fs}$). The authors want to emphasize that `N4SID` is only used as a label for the weighting scheme as indicated above. This is not to be confused with the algorithm called `N4SID` by VanOverschee and De Moor (1994). The lower plot of this figure shows $\det[(\mathscr{O}'_f W_2 \mathscr{O}_f)^{-1}\mathscr{O}'_f W_2 \mathscr{E}_f \mathscr{E}'_f W_2 \mathscr{O}_f (\mathscr{O}'_f W_2 \mathscr{O}_f)^{-1}]$, where $W_2 = \lim_{T\to\infty}(\hat{W}_f^+)'\hat{W}_f^+$. This is the central term in Eq. (4). The plots clearly reveal the identical behaviour of the two measures. It can also be seen, that for the `CCA` weights the measure E_f decreases to zero for $f \to \infty$, whereas for the `N4SID` weights the choice of $f = n$ is optimal. For both weightings a converging behaviour is observed for large f, which is also in accordance with the theory.

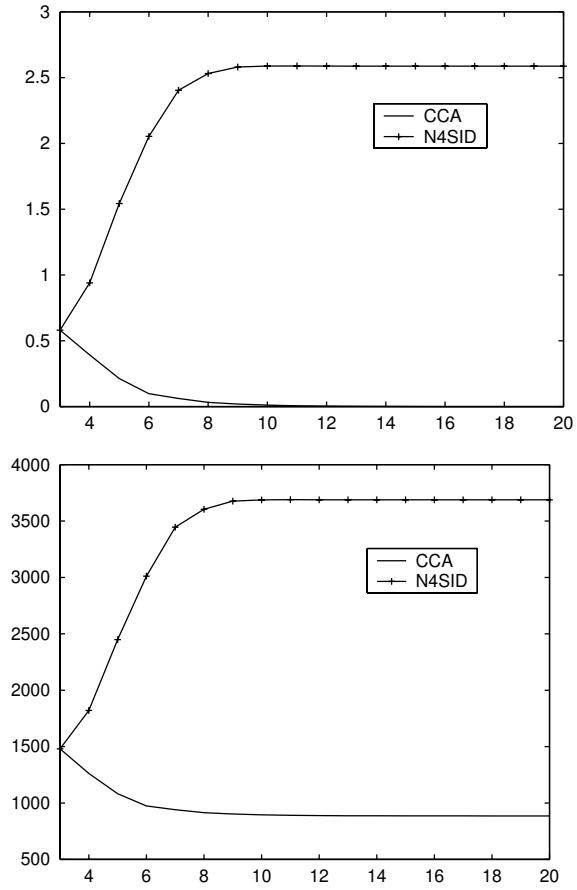


Fig. 1. The no input case: The upper plot shows the measure E_f for the two weighting schemes `CCA` and `N4SID` for the range of values $f = 3,\ldots,20$. The lower figure shows a plot of the determinant of $(\mathscr{O}'_f W_2 \mathscr{O}_f)^{-1}\mathscr{O}'_f W_2 \mathscr{E}_f \mathscr{E}'_f W_2 \mathscr{O}_f (\mathscr{O}'_f W_2 \mathscr{O}_f)^{-1}$ for the same two procedures and the same range of integers f.

The second example is a second order single- input single-output system with one additional observed white noise input given by the system matrices

$$A = \begin{bmatrix} 0.393 & 2.022 \\ -0.208 & -0.685 \end{bmatrix}, \quad B = \begin{bmatrix} 0.95 \\ 1.00 \end{bmatrix},$$

$$C = [0.326 \quad -0.743], \quad D = 0.95, \quad K = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The observed and the unobserved noise are assumed to have mean zero and variance 1. Thus there are a total of 7 parameters to be estimated. Analogous to the case of no inputs define $E_f = \mathrm{tr}[V_f(W_f^+)F_I] - 7$. Fig. 2 shows the result of the calculation. The figures again demonstrate identical behaviour of the two measures of accuracy. Again, the `CCA` weighting scheme is superior to the `N4SID` weighting scheme and again it reaches the Cramer–Rao lower bound for $f \to \infty$. This illustrates

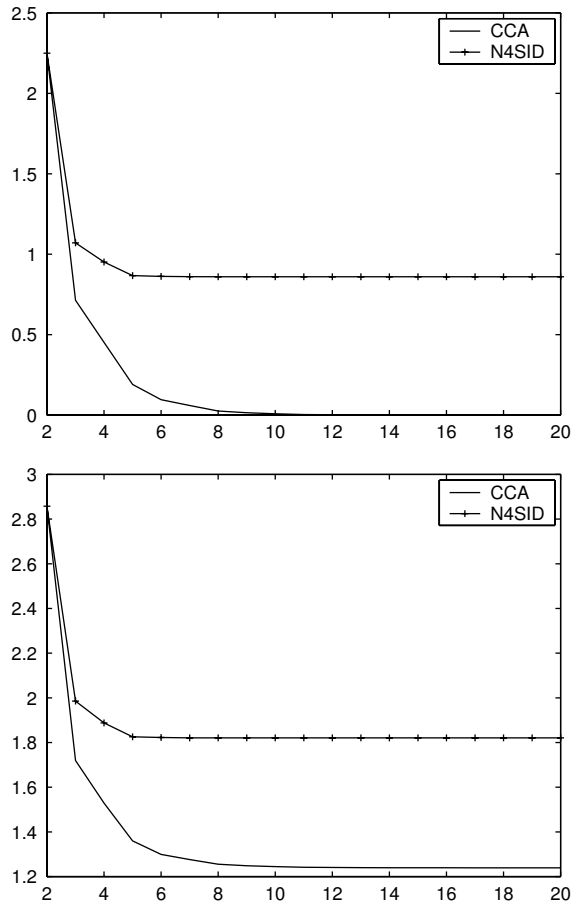


Fig. 2. The white noise input case: The upper plot shows the measure E_f for the two weighting schemes CCA and N4SID for the range of values $f = 2,\ldots,20$. The lower picture shows a plot of the determinant of $(\mathscr{O}_f'W_2\mathscr{O}_f)^{-1}\mathscr{O}_f'W_2\mathscr{E}_f\mathscr{E}_f'W_2\mathscr{O}_f(\mathscr{O}_f'W_2\mathscr{O}_f)^{-1}$ for the same two procedures and the same range of integers f.

the significance of the expressions found in this paper in assessing the relative efficiency of various weighting schemes.

6. Conclusions

In this paper the dependence of the asymptotic accuracy of the Larimore type of subspace methods with respect to the choice of the integer f and the weighting matrix W_f^+ has been explored in the situation, where the true system order is known. It has been shown, that the effects of these choices in the case of no observed inputs or white observed inputs can be summarized in the term $(\mathscr{O}_f'W_2\mathscr{O}_f)^{-1}\mathscr{O}_f'W_2\mathscr{E}_f(I \otimes \Omega)\mathscr{E}_f'W_2\mathscr{O}_f(\mathscr{O}_f'W_2\mathscr{O}_f)^{-1}$ as has been shown in Theorem 1. This term shows, that the CCA choice of the weighting according to (3) is optimal with respect to the asymptotic variance for each f. It also follows that for this optimal choice the variance decreases with increasing f, achieving the optimal accuracy for the choice $f \to \infty$. For other weighting procedures the expression can be used to optimize the choice of f. Finally, the new expressions for the asymptotic variance also lead to an efficient implementation of the computation of the asymptotic variance, which could be used for practical implementation rather than only for academic purposes.

Acknowledgements

This work has been done, while Dietmar Bauer was holding a post doc position at the Division of Automatic Control in Linköping. Dietmar Bauer acknowledges financial support in part by the European Commission through the program Training and Mobility of Researchers — Research Networks and through project System Identification (FMRX CT98 0206) and acknowledges contacts with the participants in the European Research Network System Identification (ERNSI).

References

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

Bauer, D., & Jansson, M. (2000). Analysis of the asymptotic properties of the MOESP type of subspace algorithms. *Automatica*, *36*(4), 497–509.

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

Bauer, D., Deistler, M., & Scherrer, W. (2000). On the impact of weighting matrices in subspace algorithms. *Proceedings of the IFAC conference 'SYSID' 2000*, Santa Barbara, CA.

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

Jansson, M. (2000). Asymptotic variance analysis of subspace identification methods. *Proceedings of the SYSID'2000 conference*. Santa Barbara, CA.

Jansson, M., & Wahlberg, B. (1998). On consistency of subspace methods for system identification. *Automatica*, *34*(12), 1507–1519.

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). Piscataway, NJ: IEEE Service Center.

Lewis, R., & Reinsel, G. (1985). Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis*, *16*, 393–411.

Peternell, K., Scherrer, W., & Deistler, M. (1996). Statistical analysis of novel subspace identification methods. *Signal Processing*, *52*, 161–177.

Van Overschee, P., & DeMoor, B. (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.

Lennart Ljung received his Ph.D. in Automatic Control from Lund Institute of Technology in 1974. Since 1976 he is Professor of the chair of Automatic Control In Linkoping, Sweden, and is currently Director of the Competence Center "Information Systems for Industrial Control and Supervision" (ISIS). He has held visiting positions at Stanford and MIT and has written several books on System Identification and Estimation. He is an IEEE Fellow and an IFAC Advisor as well as a member of the Royal Swedish Academy of Sciences (KVA), a member of the Royal Swedish Academy of Engineering Sciences (IVA), and an Honorary Member of the Hungarian Academy of Engineering. He has received honorary doctorates from the Baltic State Technical University in St Petersburg, and from Uppsala University.

Dietmar Bauer received his masters (1995) and Ph.D. degree (1998) in applied mathematics from the Technical University Wien. Since 1995 he is with the Institute for Econometrics, Operations Research and System Theory at the TU Wien. In 1999 he held a post doc position at the University of Newcastle, Australia and in 2000 he spent 6 months as a post doc at Linköping University, Sweden. His main areas of research are in system identification, in particular with an emphasis on economical applications. Recent topics include the applicability of subspace algorithms in the area of finance.

PERGAMON

Automatica 37 (2001) 1561–1573

automatica

www.elsevier.com/locate/automatica

Order estimation for subspace methods☆

Dietmar Bauer*,[1]

Institute f. Econometrics, Operations Research and System Theory TU Wien, Argentinierstr. 8, A-1040 Wien, Austria

Received 5 June 2000; received in final form 20 March 2001

Three different order estimation criteria in the context of subspace algorithms are introduced and sufficient conditions for strong consistency are derived. A simulation study points to open questions.

Abstract

In this paper the question of estimating the order in the context of subspace methods is addressed. Three different approaches are presented and the asymptotic properties thereof derived. Two of these methods are based on the information contained in the estimated singular values, while the third method is based on the estimated innovation variance. The case with observed inputs is treated as well as the case without exogenous inputs. The two methods based on the singular values are shown to be consistent under fairly mild assumptions, while the same result for the third approach is only obtained on a generic set. The former can be applied to Larimore type of procedures as well as to `MOESP` type of procedures, whereas the third is only applied to Larimore type of algorithms. This has implications for the estimation of the order of systems, which are close to the exceptional set, as is shown in a numerical example. All the estimation methods involve the choice of a penalty term. Sufficient conditions on the penalty term to guarantee consistency are derived. The effects of different choices of the penalty term are investigated in a simulation study.

Keywords: Subspace methods; System order; Estimation; Asymptotic properties

1. Introduction

There exists an extensive literature for order estimation algorithms for linear, dynamical, state space systems. Probably the most important contribution can be attributed to Akaike (1969) for introducing the information criteria. These criteria compare the model fit on the estimation data as measured by a function of the estimated innovation variance to some penalty term, which punishes high model orders. In other words, the higher model order is only chosen, if the increase in the accuracy is higher than a certain threshold, which depends on the sample size. Alternatively they can be seen as a sequence of tests to identify the model order, where the size of the tests is adjusted to the sample size. The properties of these estimation methods are well studied (Shibata, 1980; Akaike, 1969; Rissanen, 1978) and the effects of the choice of the penalty term are well understood (see e.g. Hannan & Deistler, 1988) for a comprehensive discussion of the known properties. All these estimation methods however rely on the use of the maximum likelihood estimate for the system for each order. Thus in practice a large number of systems has to be estimated using numerical search procedures to optimise the likelihood for given system order, leading to a sometimes prohibitive amount of computations.

For subspace algorithms the situation is different. Although subspace methods have been proposed quite some time ago, there exist only few references dealing with the estimation of the order in the context of subspace methods. The first contribution seems to be due to (Peternell, 1995). This method relies on the information of the estimated canonical correlations, which are estimated in the subspace methods. This leads to a very economical (in terms of computations) method, which has been shown to lead to almost sure (a.s.) consistent estimates under the usual assumptions. See below for

☆This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor H. Hjalmarsson under the direction of Editor Torsten Söderström.

[1] Parts of this work have been done while the author was holding a post-doc position at the Department of Automatic Control, Linköping University, Linköping, Sweden.

* Corresponding author. Tel.: +43-1-58801-11944; fax: +43-1-58801-11999.

E-mail address: dietmar.bauer@tuwien.ac.at (D. Bauer).

0005-1098/01/$ - see front matter
PII: S0005-1098(01)00118-2

more details on this. It has been observed in Bauer (1998), that this method seems to be relatively sensitive to the choice of certain user parameters, which can deteriorate the performance of the method considerably (see the simulations section). This motivates the development of an alternative, which is a small adaptation of the criterion given in Peternell (1995) and seems to be less sensitive. Bauer (1998) introduces another criterion for the Larimore type of procedures, which is much in the spirit of Akaikes information criteria, as it uses the estimated innovation variance. These three procedures will be presented and analysed below. It will be clear from the proofs however, that the proposed estimation method is only one possibility, as the main problem boils down to estimate the rank of a matrix. For this problem there are well established testing methods, which however rely on the distribution of the matrix, whose rank is estimated. Such procedures are presented in Sorelius (1999): There the rank of the crucial matrix is found by increasing the dimensions of the matrix by one in one step and performing a test on the newly introduced smallest singular value. This procedure however has the disadvantage of simultaneous tests, since in practice a sequence of tests will have to be performed, where the number of the tests and the dependency of the tests is unknown at the start of the tests.

The organisation of the paper is as follows: In the next section the model set is stated and the main assumptions are presented. The estimation algorithms are briefly reviewed in Section 3, where also the various order estimation algorithms are discussed. Section 4 then states the main results of this paper and provides proofs for them. A simulation study is performed in Section 5. Finally Section 6 concludes.

2. Model set and assumptions

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

$$x_{t+1} = Ax_t + Bu_t + K\varepsilon_t,$$
$$y_t = Cx_t + Du_t + \varepsilon_t, \tag{1}$$

are considered, where $y_t \in \mathbb{R}^s$ denotes the observed output process, $u_t \in \mathbb{R}^m$ denotes the observed input process and $\varepsilon_t \in \mathbb{R}^s$ the unobserved white noise sequence. $x_t \in \mathbb{R}^n$ is the state sequence. Here the true order of the system is denoted by n. The matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{s\times n}$, $D \in \mathbb{R}^{s\times m}$, $K \in \mathbb{R}^{n\times s}$ determine the system. In the case an input delay is postulated, D is restricted to zero. The system is assumed to be stable, i.e. all eigenvalues of A are assumed to lie inside the unit circle, and strictly minimum-phase, i.e. the eigenvalues of $A - KC$ are assumed to lie inside the unit circle. The system matrices correspond to a pair of transfer functions: Let $k(z) = I + zC(I - zA)^{-1}K$ and let $l(z) = D + zC(I - zA)^{-1}B$, where z denotes the backward shift operator. Furthermore let M_n denote the set of all pairs of transfer functions (k, l) that permit a minimal state space representation of the form (1) fulfiling the stability and the strict minimum-phase assumption.

The white noise ε_t is assumed to be an ergodic martingale difference sequence satisfying the following conditions:

$$\mathbb{E}\{\varepsilon_t | \mathscr{F}_{t-1}\} = 0, \quad \mathbb{E}\{\varepsilon_t\varepsilon_t' | \mathscr{F}_{t-1}\} = \Omega = \mathbb{E}\varepsilon_t\varepsilon_t' > 0,$$
$$\mathbb{E}\{\varepsilon_{t,a}\varepsilon_{t,b}\varepsilon_{t,c} | \mathscr{F}_{t-1}\} = \omega_{a,b,c}, \quad \mathbb{E}\{\varepsilon_{t,a}^4\} < \infty. \tag{2}$$

Here $\mathbb{E}$ denotes expectation, $\mathscr{F}_t$ denotes the σ-algebra spanned by $(y_s, s \leqslant t)$ and $\varepsilon_{t,a}$ denotes the ath component of the vector ε_t. Note that these assumptions coincide with the assumptions used in the analysis of the order estimation methods in the case of maximum likelihood estimation in Hannan and Deistler (1988, Theorem 4.3.2). Corresponding to the input two different sets of assumptions will be introduced for the Larimore type of procedures and the MOESP type of procedures.

Assumption 1 (*Larimore type of procedure*). The process $(u_t; t \in \mathbb{Z})$ is filtered white noise of the form $u_t = \sum_{j=0}^{\infty} K_u(j)\eta_{t-j}$, where η_t is an ergodic, martingale difference sequence with innovation covariance matrix $\Omega_\eta > 0$ fulfiling the assumptions stated in Eq. (2) and being independent of ε_t, and where $\|K_u(j)\| \leqslant c_u\rho_u^j$ for some $0 < c_u < \infty$, $0 < \rho_u < 1$. Furthermore $\Phi_u(\omega) = (\sum_{j=0}^{\infty} K_u(j)e^{i\omega j})\Omega_\eta(\sum_{j=0}^{\infty} K_u(j)e^{i\omega j})'$ is assumed to fulfil $0 < \underline{c}I \leqslant \Phi_u(\omega) \leqslant \bar{c}I < \infty$ for $-\pi < \omega \leqslant \pi$.

Assumption 2 (*MOESP type of procedures*). The input process $(u_t; t \in \mathbb{Z})$ is of the form $u_t = cv_t + \sum_{j=1}^{h} c_j e^{i\lambda_j t}$ where v_t fulfils Assumptions 1 and $c_j \in \mathbb{R}^m$ are zero mean random variables with finite mean square such that the corresponding process u_t is real valued. Further $0 \leqslant c < \infty$ is a constant. Furthermore the process u_t is assumed to be persistently exciting of order α (to be specified later) in the sense of Ljung (1999).

Note that the assumptions for the inputs in the Larimore procedure are more severe, as is apparent from the choice $c = 0$: In this case the input is just a sum of sinusoids and thus only persistent of finite degree, whereas the Larimore type of assumptions imply, that the input is persistent of any order. The reason for this lies in the fact, that for the Larimore type of procedures a necessary condition for consistency is that the integer parameter p tends to infinity (see below for details). For the MOESP type of procedures note that the assumptions are similar to the assumptions imposed in the proof of the asymptotic normality in Bauer and Jansson (2000). It will be clear from the proof given below, which properties for the input signal are really needed in this respect. Also note

that the conditions given in Bauer and Jansson (2000) permit certain pseudostationary sequences, i.e. sums of sinusoids. However in this case it is necessary to impose the necessary restrictions (i.e. the existence of certain limits, which appear in the proof) directly on the sequence rather than using sufficient conditions on the underlying random variables.

3. Estimation algorithms

In this section a brief review of the main steps in the considered subspace procedures is given and the estimation algorithms are motivated. For a more detailed description of subspace methods see Larimore (1983), Verhaegen (1994) or Bauer (1998, Chapter 3). Let $Y_{t,f}^+ = [y_t', y_{t+1}', \dots, y_{t+f-1}']'$ and let $U_{t,f}^+$ and $E_{t,f}^+$, respectively, be constructed analogously using u_t and ε_t, respectively, in the place of y_t. Let $Z_{t,p}^- = [y_{t-1}', u_{t-1}', \dots, y_{t-p}', u_{t-p}']'$. Here f and p are two integer parameters, which have to be chosen by the user. See below for assumptions on the choice of these integers. Then it follows from the system equations (1) that

$$Y_{t,f}^+ = \mathcal{O}_f \mathcal{K}_p Z_{t,p}^- + \mathcal{U}_f U_{t,f}^+ + \mathcal{E}_f E_{t,f}^+ + \mathcal{O}_f (A - KC)^p x_{t-p}.$$

Here $\mathcal{O}_f' = [C', A'C', \dots (A^{f-1})'C']$ and $\mathcal{K}_p = [[K, B - KD], (A - KC)[K, B - KD], \dots, (A - KC)^{p-1}[K, B - KD]]$. Further $\mathcal{U}_f$ and $\mathcal{E}_f$ are block Toeplitz matrices containing the impulse response sequences. The actual form of these two matrices is of no importance here and thus it is referred to the original articles for details. This equation builds the basis for all subspace algorithms, which can be described as follows:

(1) Regress $Y_{t,f}^+$ onto $U_{t,f}^+$ and $Z_{t,p}^-$ to obtain an estimate $\hat{\beta}_z$ of $\mathcal{O}_f \mathcal{K}_p$ and an estimate $\hat{\beta}_u$ of $\mathcal{U}_f$, respectively. Due to finite sample effects $\hat{\beta}_z$ will typically be of full rank.
(2) For given n find a rank n approximation of $\hat{\beta}_z$ by using the SVD of $\hat{W}_f^+ \hat{\beta}_z \hat{W}_p^- = \hat{U}_n \hat{\Sigma}_n \hat{V}_n' + \hat{R}$. Here $\hat{\Sigma}_n$ denotes the diagonal matrix containing the largest n singular values in decreasing order. $\hat{U}_n$ contains the corresponding left singular vectors as columns and $\hat{V}_n$ the corresponding right singular vectors. Finally $\hat{R}$ accounts for the neglected singular values. This leads to an approximation $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p = (\hat{W}_f^+)^{-1} \hat{U}_n \hat{\Sigma}_n \hat{V}_n' (\hat{W}_p^-)^{-1}$. The actual decomposition of this matrix into $\hat{\mathcal{O}}_f$ and $\hat{\mathcal{K}}_p$ has no influence on the estimated transfer functions.
(3) Using the estimates $\hat{\mathcal{O}}_f$, $\hat{\mathcal{K}}_p$ and $\hat{\beta}_u$ obtain the system matrix estimates.

In the second step an order has to be specified. Also the matrices $\hat{W}_f^+$ and $\hat{W}_p^-$ have to be provided by the user. In the literature several different choices have been proposed. For the matrix $\hat{W}_p^-$ the choices are restricted to $(\hat{\Gamma}_p^-)^{1/2}$ and $(\hat{\Gamma}_p^{-,\Pi})^{1/2}$, where $\hat{\Gamma}_p^- = 1/T \sum_{t=p+1}^{T-f} Z_{t,p}^- (Z_{t,p}^-)'$ denotes the sample variance of $Z_{t,p}^-$. Further $\hat{\Gamma}_p^{-,\Pi} = \hat{\Gamma}_p^- - \hat{\Gamma}_{z,u} \hat{\Gamma}_u^{-1} \hat{\Gamma}_{u,z}$. Here $\hat{\Gamma}_u$ denotes the sample covariance of $U_{t,f}^+$ and $\hat{\Gamma}_{u,z}$ the sample covariance of $U_{t,f}^+$ and $Z_{t,p}^-$. Corresponding to $\hat{W}_f^+$ two choices will be considered: $(\hat{\Gamma}_f^{+,\Pi})^{-1/2} = \hat{\Gamma}_y^+ - \hat{\Gamma}_{y,u} \hat{\Gamma}_u^{-1} \hat{\Gamma}_{u,y}$ using obvious notation, where y stands for $Y_{t,f}^+$, and $\hat{W}_f^+ = [K_W(i-j)]_{i,j}$, where $w(z) = \sum_{j=0}^{\infty} K_W(j) z^j$ denotes a frequency weighting. $K_W(j) = 0$, $j < 0$ and $K_W(0)$ is assumed nonsingular. Furthermore $w(z)$ is assumed to be stable and strictly minimum phase. The intuition of this special choice of the weighting is to emphasize some frequency range via specifically designing $w(z)$ to be a band pass filter (see e.g. McKelvey, 1995). The idea of this step is essentially to discriminate between the nonzero 'signal' singular values and the noise contained in $\hat{R}$, which is influenced by the weighting, since this scales different directions. Using the information contained in the estimated singular values will be the basis for two of the estimation methods.

For the Larimore type of methods also an order estimation algorithm will be given, which relies on the estimated innovation variance. Thus it is necessary to give more details on the estimation of the system matrices in this case. Note, that from step 2 an estimate $\hat{\mathcal{K}}_p$ is obtained. This is used to estimate the state sequence as $\hat{x}_t = \hat{\mathcal{K}}_p Z_{t,p}^-$. Let $\langle a_t, b_t \rangle = 1/T \sum_{t=p+1}^{T-f} a_t b_t'$. Inserting the estimated state into the system equations (1) one obtains estimates of (A, B, C, D) from the least squares solution:

$$[\hat{A}_T, \hat{B}_T] = [\langle \hat{x}_{t+1}, \hat{x}_t \rangle \ \langle \hat{x}_{t+1}, u_t \rangle] \begin{bmatrix} \langle \hat{x}_t, \hat{x}_t \rangle & \langle \hat{x}_t, u_t \rangle \\ \langle u_t, \hat{x}_t \rangle & \langle u_t, u_t \rangle \end{bmatrix}^{-1},$$

$$[\hat{C}_T, \hat{D}_T] = [\langle y_t, \hat{x}_t \rangle \quad \langle y_t, u_t \rangle] \begin{bmatrix} \langle \hat{x}_t, \hat{x}_t \rangle & \langle \hat{x}_t, u_t \rangle \\ \langle u_t, \hat{x}_t \rangle & \langle u_t, u_t \rangle \end{bmatrix}^{-1}.$$

If a delay is postulated, then in the second least squares problem u_t is omitted. The matrix K and the innovation sequence are estimated from the residuals of these equations as follows: Let $\hat{\varepsilon}_t = y_t - \hat{C}_T \hat{x}_t - \hat{D}_T u_t$. Then $\hat{\Omega} = \langle \hat{\varepsilon}_t, \hat{\varepsilon}_t \rangle$ and $\hat{K}_T = \langle \hat{x}_{t+1}, \hat{\varepsilon}_t \rangle \hat{\Omega}^{-1}$.

Following the discussion given above there are a couple of rather obvious algorithms to estimate the order. These will be presented in the following.

3.1. Using the information contained in the singular values

From standard theory it follows, that $\hat{X}_{f,p} = \hat{W}_f^+ \hat{\beta}_z \hat{W}_p^-$ converges a.s. to the limit $X_\infty = W_f^+ \mathcal{O}_f \mathcal{K}_p W_p^-$, where W_f^+ and W_p^- denote the a.s. limits of $\hat{W}_f^+$ and $\hat{W}_p^-$, respectively. Here convergence occurs in the operator norm acting on ℓ^2 almost surely, where the matrices occurring are seen as operators by adding zeros

in the infinite matrix representation corresponding to the operator. Therefore it follows from the results of operator theory (see e.g. Chatelin, 1983) that the singular values also converge. Since $X_\circ$ has rank n, only the first n singular values of $X_\circ$ are nonzero, the rest being zero. Therefore, the problem boils down to the assessment of the rank of a noisy matrix. The problem gets complicated, since the distribution of the noise acting on the matrix is hard to quantify. Therefore this paper resorts to estimation algorithms as opposed to methods of obtaining the order via a sequence of tests (cf. Sorelius, 1999). These algorithms share the idea of the information criteria of comparing the significance of the inclusion of another coordinate in the state to a penalty term, which is chosen such that the resulting estimates possess desirable properties, such as consistency. Define the following two criteria:

$$NIC(n) = \sum_{j=n+1}^{M} \hat{\sigma}_j^2 + C(T)d(n)/T, \tag{3}$$

$$SVC(n) = \hat{\sigma}_{n+1}^2 + C(T)d(n)/T. \tag{4}$$

Here $d(n) = n(m+s) + ns + sm$ denotes the number of parameters of a state space system of order n (see e.g. Hannan and Deistler, 1988, Theorem 2.5.3). $C(T) > 0$, $C(T)/T \to 0$ is a penalty term, which will be described below in more detail. In the definition $M = \min\{fs, p(s+m)\}$, the number of estimated singular values. The estimated order $\hat{n}$, say, is obtained as the minimising argument of these criterion functions. NIC has been introduced and analysed in Peternell (1995). In the definition Peternell (1995) used a different choice of $d(n)$, which however can be reformulated to fit into the present setting. Also Peternell (1995) only dealt with f and p fixed and finite, while the following discussion holds for general choices. SVC stands for *singular value criterion* and has been proposed as a refinement of NIC in Bauer (1998). The main difference lies in the fact, that NIC uses the Frobenius norm of the matrix $\hat{R}$, whereas SVC uses the two norm to measure the size of the neglected singular values. For both criteria the order estimate is obtained by minimizing the above expression. Note, that these order estimation techniques do not depend on whether MOESP or the Larimore type of methods is used and thus can be used in all these procedures. The author wants to stress, that these are just two algorithms, however many more seem possible, since in principle all that is done is to compare the size of $\hat{R}$ measured in some norm to some sample size dependent penalty term. Also note, that the choice of the weighting matrices $\hat{W}_f^+$ and $\hat{W}_p^-$ is very influential for the outcome of the estimation, as will be demonstrated in Section 5. This might indeed be desirable, since special weightings can be given a somewhat heuristic interpretation as frequency shaping filters (cf. McKelvey, 1995). In this case it follows, that the weighting matrices serve as a tool to stress the important frequencies for the identification, and thus these directions might be upweighed, whereas other directions are downweighed.

Note, that both criterion functions can be implemented with almost no computational load. The singular values are estimated in the algorithms, therefore only the addition of the penalty term and the minimization over a small range of integers has to be performed.

3.2. Using the estimated innovation covariance

A second intuitive idea would be to estimate the order using the estimated innovation covariance in the Larimore type of procedures. Recall that given the state sequence of dimension n, say $\hat{x}_t^n$, the innovation variance is estimated as $\hat{\Omega}_n = \langle y_t - \hat{C}_T^n \hat{x}_t^n - \hat{D}_T^n u_t, y_t - \hat{C}_T^n \hat{x}_t^n - \hat{D}_T^n u_t \rangle$. Here $[\hat{C}_T^n, \hat{D}_T^n]$ denotes the estimates of C and D using the estimated state $\hat{x}_t^n$. Then it is tempting to use the criterion function used also in the information criteria as follows:

$$IVC(n) = \log\det \hat{\Omega}_n + C(T)d(n)/T, \tag{5}$$

where $d(n)$ and $C(T)$ are identical to the definition of SVC and NIC. Again the order estimate is obtained by minimizing this function over the integers $0 \leqslant n \leqslant \min\{fs, p(s+m)\}$. Here IVC stands for *innovation variance criterion*. The author wants to stress, that this is *not* the standard information criterion, since the estimates $\hat{\Omega}_n$ are not the maximum likelihood estimates of the innovation sequence. In fact it will be shown, that this estimation algorithm may perform poor in some situations.

From a computational point of view this criterion is very attractive in the case of no exogenous inputs present in the read out equation, i.e. in the case $y_t = Cx_t + \varepsilon_t$, and additionally the choice of the weighting $\hat{W}_p^- = (\hat{\Gamma}_p^-)^{1/2}$. In this case the choice $\hat{\mathscr{K}}_p^n = \hat{V}_n'(\hat{W}_p^-)^{-1} Z_{t,p}^-$ leads to $\langle \hat{x}_t^n, \hat{x}_t^n \rangle = I$, i.e. the components of the state are orthogonal and thus the regressions can be performed independently. The estimation algorithm then amounts to estimating the matrix C for the maximal state dimension, max say, and then only additions and multiplications have to be performed. Let $\hat{C}_T^{\max} = [\hat{C}_T^n, \hat{C}_T^{n-\max}]$ then

$$\hat{\Omega}_n = \hat{\Omega}_{\max} + \hat{C}_T^{n-\max}(\hat{C}_T^{n-\max})'.$$

In the case of exogenous inputs present or a different choice of the weighting $\hat{W}_p^-$ on the contrary each regressions has to be performed separately. Note however, that normally these will be low dimensional regression in general and also of not too big numerical load. It is possible to implement the subspace procedures such that only the estimated covariances are used rather than the data itself. In this case the necessary covariance estimates are already calculated and thus only matrix inversions have to be calculated. Otherwise also in this step the

necessary covariances could be calculated in order to minimize the number of necessary calculations. It will be shown in the next section, that although this procedure seems appealing on first sight, it is not a recommended procedure. Thus in this respect the result of this paper is rather to show, that using this method may lead to problems, which are somewhat unexpected.

4. Main results

In this section the properties of the various estimation algorithms will be derived. The discussion draws heavily from Bauer (1998) and Peternell (1995). Some results for the MOESP case have been presented in (Bauer, 1999). The following notation will be used widely: Let $f_T = O(g_T)$ mean that $\|f_T\|_2/g_T \leqslant M$ a.s. Further $f_T = o(g_T)$ implies $\|f_T\|_2/g_T \to 0$ a.s.

The results are mainly based on the following lemmas: The first deals with the accuracy of the estimation of sample covariances under the given assumptions on the system and the input. The second one deals with the linearization of the SVD or SVD related quantities, which will be of importance mainly for the NIC and SVC cases.

Lemma 1. *Let* $(y_t;\ t\in\mathbb{Z})$ *be generated by a system of the form* (1), *where the noise fulfils the assumptions of Section* 2. *Let* $\hat{\gamma}_{z,z}(j) = T^{-1}\sum_{j=1}^{T-j} z_t z'_{t+j}$ *and let* $\gamma_{z,z}(j) = \mathbb{E} z_t z'_{t+j}$, *where* $z_t = [y'_t, u'_t]'$. *Furthermore let* $H_T = o((\log T)^a)$ *for some* $0 < a < \infty$.

If u_t *fulfils Assumptions* 1 *then*

$$\max_{|j|\leqslant H_T} \|\hat{\gamma}_{z,z}(j) - \gamma_{z,z}(j)\|_2 = O(Q_T), \tag{6}$$

where $Q_T = \sqrt{\log\log T/T}$. *If* u_t *only fulfils Assumptions* 2, *then the statement is true for* $H_T = M < \infty$.

This lemma follows from Hannan and Deistler (1988, Theorem 5.3.2, Chapter 5). The lemma provides relatively sharp bounds for the estimation error of the covariance sequences. In fact it follows from the law of the iterated logarithm for the estimated covariance sequences that—except for the exact evaluation of the constant involved in the $O(Q_T)$ statement—the bound is tight.

Lemma 2 (Chatelin). *Let* $\mathscr{T}_T$ *denote a sequence of symmetric, compact operators acting on* ℓ^2, *which converges in norm to the operator* $\mathscr{T}_\circ$. *Then it follows, that the set of eigenvalues of* $\mathscr{T}_T$ *converges to the set of eigenvalues of* $\mathscr{T}_\circ$. *Also the corresponding eigenspaces converge in the gap metric. Let* P_i° *denote the orthonormal projection matrix onto the space of eigenvectors corresponding to the eigenvalue* λ_i° *of* $\mathscr{T}_\circ$ *and let* P_i^T *and* λ_i^T *denote the corresponding quantities of* $\mathscr{T}_T$. *Here for a multiple eigenvalue* λ_i° *of* $\mathscr{T}_\circ$ *the quantities* P_i^T *refer to the orthonormal projection matrix onto the space spanned by the eigenvectors to all eigenvalues of* $\mathscr{T}_T$ *converging to* λ_i°. *Then*

$$P_i^T = P_i^\circ + \sum_{j\neq i} P_j^\circ \frac{\mathscr{T}_T - \mathscr{T}_\circ}{\lambda_i^\circ - \lambda_j^\circ} P_i^\circ + P_i^\circ \frac{\mathscr{T}_T - \mathscr{T}_\circ}{\lambda_i^\circ - \lambda_j^\circ} P_j^\circ + o(\|\mathscr{T}_T - \mathscr{T}_\circ\|). \tag{7}$$

The lemma implies, that the eigenspaces converge, and in particular the projections on the eigenspaces converge at the same rate as the error in the approximation.

It has been shown in Bauer and Jansson (2000), that the MOESP type of methods lead to consistent estimates for the system matrix estimates only in generic cases. Therefore also the SVC criterion can only produce consistent estimates in these cases. Let $\Phi_u(\omega)$ denote the spectrum of the stationary process u_t and assume, that the integers f and p are used for the estimation. Further denote the noise variance with Ω. Then it has been shown in (Bauer & Jansson, 2000) that there exists a set $U_n(f,p,\Phi_u,\Omega) \subset M_n$, such that the MOESP procedure provides consistent estimates of the pair of transfer functions. It is also shown, that this set is generic in M_n. However as the example given in Jansson and Wahlberg (1997) shows, the set is not identical to M_n in general. In the case $\min\{f,p\} > 3n$ it has been shown in (Chui, 1997) that this is the case, i.e. the consistency holds for every pair $(k,l)\in M_n$. In fact the sufficient conditions stated in (Chui, 1997) are much sharper.

Theorem 3. *Let the process* $(y_t;\ t\in\mathbb{Z})$ *be generated by a system of form* (1), *where the true system order is equal to* $n_\circ$, *and where the white noise process* $(\varepsilon_t;\ t\in\mathbb{Z})$ *fulfils the Assumptions of Section* 2. *Let the input fulfil the Assumptions* 1, *further* $\min\{f,p\} \geqslant n_\circ$ *and* $\max\{f,p\} = o((\log T)^a)$ *is assumed for some* $a < \infty$. *In this case the conditions* $C(T) > 0$, $C(T)/T \to 0$, $C(T)/(fp\log\log T) \to \infty$ *are sufficient for the a.s. consistency of the order estimate obtained by minimizing* $SVC(n)$.

If the input fulfils the Assumptions 2 *with* $\alpha = f + p - 1$, *then for each fixed pair* f *and* p *there exists a set* $U_n(f,p,\Phi_u(\omega),\Omega) \subset M_n$, *where* $\Phi_u(\omega)$ *denotes the pseudo-spectrum of the input sequence, such that for* $(k,l)\in U_n(f,p,\Phi_u(\omega),\Omega)$ *the* SVC *method leads to a.s. consistent estimates of the order under the assumption* $C(T) > 0$, $C(T)/T \to 0$, $C(T)/\log\log T \to \infty$. *If* $(k,l)\notin U_n(f,p,\Phi_u(\omega),\Omega)$ *then consistency fails for the same choice of the penalty term* $C(T)$, *i.e.* $\lim_{T\to\infty} \hat{n} < n_\circ$ *a.s.*

Proof. Note, that under both sets of assumptions the error in the estimation of the first $f+p-1$ covariances $\gamma_z(j)$ is of order $O(Q_T)$ uniformly due to the Lemma 1. The estimation uses the singular values of $\hat{X}_{f,p} = \hat{W}_f^+ \hat{\beta}_z \hat{W}_p^-$, which converges to $X_\circ = W_f^+ \mathscr{O}_f \mathscr{K}_p W_p^-$ a.s. as has been shown e.g. in Peternell,

Scherrer and Deistler (1996). Here convergence is in operator norm in the embedding ℓ^2. Consider the estimation error in $\hat{\beta}_z$ first: Introduce the notation $\langle a_t, b_t\rangle = T^{-1}\sum_{j=p+1}^{T-f} a_t b_t'$. Then

$$[\hat{\beta}_z,\hat{\beta}_u] = \left\langle Y_{t,f}^+, \begin{pmatrix} Z_{t,p}^- \\ U_{t,f}^+ \end{pmatrix} \right\rangle \left\langle \begin{pmatrix} Z_{t,p}^- \\ U_{t,f}^+ \end{pmatrix}, \begin{pmatrix} Z_{t,p}^- \\ U_{t,f}^+ \end{pmatrix} \right\rangle^{-1}.$$

The estimation error in each entry of these matrices is of the order $O(Q_T)$ as follows from the Lemma 1 together with Hannan and Deistler (1988) Theorem 6.6.11, which assures the summability of the columns of the inverse uniformely in f and p. Thus consider the weighting matrices: Recall that the weighting matrices are restricted to be either deterministic or chosen as the square roots of matrices like $\langle Y_{t,f}^+, Y_{t,f}^+\rangle - \langle Y_{t,f}^+, U_{t,f}^+\rangle\langle U_{t,f}^+, U_{t,f}^+\rangle^{-1}\langle U_{t,f}^+, Y_{t,f}^+\rangle$. Using the same arguments as have been used above shows that the estimation error in the entries of these matrices are of order $O(Q_T)$. Therefore also the error in the positive definite symmetric square root is of the same order, as can be seen from a Taylor series expansion of the square root, which can be used to define the symmetric square root of an operator.

It thus follows, that $\hat{X}_{f,p} \to X_\circ$, where $\|\hat{X}_{f,p} - X_\circ\|_2 = O(Q_T\sqrt{fp})$. Therefore the singular values converge at the same rate. This shows, that underestimation of the order is not possible asymptotically, if $\sigma_{n_\circ} > 0$, where σ_i denotes the singular values of $X_\circ$ ordered decreasingly, as

$$\begin{aligned} SVC(n) &= \hat{\sigma}_{n+1}^2 + \frac{d(n)C(T)}{T} \\ &= \sigma_{n+1}^2 + (\hat{\sigma}_{n+1}^2 - \sigma_{n+1}^2) + \frac{d(n)C(T)}{T} \\ &= \sigma_{n+1}^2 + O\left(\sqrt{fp}Q_T + \frac{d(n)C(T)}{T}\right). \end{aligned}$$

Since the second term tends to zero, the minimum cannot be attained at $n < n_\circ$. In the case of the MOESP procedure and $(k,l)\notin U_n(f,p,\Phi_u,\Omega)$ the nth singular value is zero and thus consistency fails, as follows from the same arguments given below, since in that case, the same arguments show that the asymptotic state dimension is equal to the number of nonzero singular values for the limiting matrix.

Therefore it needs to be shown, that the true order $n_\circ$ will be preferred to $n > n_\circ$ asymptotically. Thus for $n \geqslant n_\circ$ consider

$$\begin{aligned} \hat{\sigma}_{n+1} &= \|\hat{W}_f^+ \hat{\beta}_z \hat{W}_p^- - \hat{U}_n\hat{\Sigma}_n\hat{V}_n'\|_2 \\ &= \|\hat{U}'(\hat{X}_{f,p} - \hat{U}_n\hat{\Sigma}_n\hat{V}_n')\|_2 \\ &\leqslant \|\hat{U}_2\hat{U}_2'\hat{X}_{f,p} - U_2U_2'X_\circ\|_2. \end{aligned}$$

Here $\hat{U} = [\hat{U}_n,\hat{U}_{2,n}]$, $\hat{U}_n \in \mathbb{R}^{fs\times n}$, $\hat{U}_{2,n} \in \mathbb{R}^{fs\times(fs-n)}$ and $\hat{U}_2 = \hat{U}_{2,n_\circ}$, which together with $U_2'X_\circ = 0$ explains the last inequality. Since the entries of $\hat{X}_{f,p} - X_\circ$ have been shown to be of order $O(Q_T)$ the norm

$$\begin{aligned} \|\hat{U}_2\hat{U}_2'(\hat{X}_{f,p} - X_\circ)\|_2 &\leqslant \|\hat{U}_2\hat{U}_2'\|_2\|\hat{X}_{f,p} - X_\circ\|_{Fr} \\ &= O(Q_T\sqrt{fp}). \end{aligned}$$

Therefore it remains to obtain a bound on $\|\hat{U}_2\hat{U}_2' - U_2U_2'\|_2$. But this follows from Lemma 2, using $\mathscr{T}_T = \hat{X}_{f,p}\hat{X}_{f,p}'$, $\mathscr{T}_\circ = X_\circ X_\circ'$ and the exponential decrease in elements in the rows of $U_{n_\circ}$ (cf. Bauer, 1998). Indeed from this the result follows, since $n \geqslant n_\circ$

$$\begin{aligned} SVC(n_\circ) - SVC(n) &= \hat{\sigma}_{n_\circ+1}^2 - \hat{\sigma}_{n+1}^2 + (d(n_\circ) - d(n))\frac{C(T)}{T} \\ &= \frac{C(T)}{T}[O(fp\log\log T/C(T)) \\ &\quad + d(n_\circ) - d(n)] < 0, \end{aligned}$$

since $fp\log\log T/C(T) \to 0$ and $d(n) > d(n_\circ)$.

Note, that the result also proves the consistency of the NIC criterion for the same restrictions on the penalty term. Also note, that concerning the penalty term only a sufficient condition is given. The bound is obtained by rather brute force arguments, bounding the two norm with the Frobenius norm. In the case, where f and p tend to infinity at a rate $\log T$ it seems to be desirable to use a lower penalty term, as will be argued in the numerical examples.

For the estimation criteria, which are based on an estimate of the innovation variance, the situation is somewhat different. Note that this procedure only applies for the Larimore type of procedures. Therefore assume, that the input process fulfils Assumptions 1. Note that if no delay is postulated

$$\begin{aligned} \hat{\Omega}_n &= \langle y_t, y_t\rangle - \left\langle y_t, \begin{bmatrix}\hat{x}_t \\ u_t\end{bmatrix}\right\rangle \left\langle \begin{bmatrix}\hat{x}_t \\ u_t\end{bmatrix}, \begin{bmatrix}\hat{x}_t \\ u_t\end{bmatrix}\right\rangle^{-1} \left\langle \begin{bmatrix}\hat{x}_t \\ u_t\end{bmatrix}, y_t\right\rangle \\ &= \langle y_t, y_t\rangle - \langle y_t, Z_{t+1,p+1}^-\rangle\hat{L}_n' \\ &\quad \times(\hat{L}_n\langle Z_{t+1,p+1}^-, Z_{t+1,p+1}^-\rangle\hat{L}_n')^{-1}\hat{L}_n\langle Z_{t+1,p+1}^-, y_t\rangle \\ &= \langle y_t, y_t\rangle - \hat{\bar{\mathscr{H}}}_1\bar{\hat{L}}_n'(\bar{\hat{L}}_n\bar{\hat{L}}_n')^{-1}\bar{\hat{L}}_n\hat{\bar{\mathscr{H}}}_1', \end{aligned}$$

where

$$\hat{L}_n = \begin{bmatrix} 0 & 0 & \hat{\mathscr{K}}_p(n) \\ 0 & I & 0 \end{bmatrix}, \quad \hat{\mathscr{K}}_p(n) = \hat{V}_n'(\hat{W}_p^-)^{-1}.$$

Further $\bar{\hat{L}}_n = \hat{L}_n\langle Z_{t+1,p+1}^-, Z_{t+1,p+1}^-\rangle^{1/2}$ and $\hat{\bar{\mathscr{H}}}_1 = \langle y_t, Z_{t+1,p+1}^-\rangle\langle Z_{t+1,p+1}^-, Z_{t+1,p+1}^-\rangle^{-'/2}$.

First consider the problem of underestimating the order. Let Ω_n denote the limit of $\hat{\Omega}_n$. Then $\det[\Omega_{n_\circ}] < \det[\Omega_{n_\circ-1}]$ is a sufficient condition to avoid asymptotic underestimation of the order. This follows from $C(T)/T \to 0$. This condition has been analyzed in more

detail in Bauer (1998): In the special case, where no input is present in the readout equation, i.e. $D = 0$ and where $\hat{W}_p^- = (\hat{\Gamma}_p^-)^{1/2}$ this condition is equivalent to $C_{0,n_\circ} \neq 0$, where $C_{0,n_\circ}$ denotes the last column of the limiting realization of the true system. It has been found, that this condition is fulfiled on a generic subset in some special cases. It is referred to the original work for details. In general however the implications of this condition are unknown.

Next, consider the question of overestimation: For $n \geqslant n_\circ$ one obtains $\Omega_n = \Omega_{n_\circ}$ and thus the estimation error has to be analysed more closely. According to the equation above one obtains

$$\hat{\Omega}_n - \hat{\Omega}_{n_\circ} = \hat{\mathscr{H}}_1(\bar{\bar{L}}'_{n_\circ}(\bar{\bar{L}}_{n_\circ}\bar{\bar{L}}'_{n_\circ})^{-1}\bar{\bar{L}}_{n_\circ} - \bar{\bar{L}}'_n(\bar{\bar{L}}_n\bar{\bar{L}}'_n)^{-1}\bar{\bar{L}}_n)\hat{\mathscr{H}}'_1.$$

Using the matrix inversion lemma for partitioned matrices one obtains

$$\bar{\bar{L}}'_n(\bar{\bar{L}}_n\bar{\bar{L}}'_n)^{-1}\bar{\bar{L}}_n = \bar{\bar{L}}'_{n_\circ}(\bar{\bar{L}}_{n_\circ}\bar{\bar{L}}'_{n_\circ})^{-1}\bar{\bar{L}}_{n_\circ} + \hat{P}^\perp_{n_\circ}\bar{\bar{L}}'_{n,2}(\bar{\bar{L}}_{n,2}\hat{P}^\perp_{n_\circ}\bar{\bar{L}}'_{n,2})^{-1}\bar{\bar{L}}_{n,2}\hat{P}^\perp_{n_\circ},$$

where $\hat{P}^\perp_{n_\circ} = I - \bar{\bar{L}}'_{n_\circ}(\bar{\bar{L}}_{n_\circ}\bar{\bar{L}}'_{n_\circ})^{-1}\bar{\bar{L}}_{n_\circ}$ and where $\bar{\bar{L}}'_n = [\bar{\bar{L}}'_{n_\circ}, \bar{\bar{L}}'_{n,2}]$. Since the second term is a projection operator it has norm one. Thus the essential term is $\hat{\mathscr{H}}_1\hat{P}^\perp_{n_\circ}$, which converges to zero, since $\hat{\mathscr{H}}_1 \to [C, D]\bar{L}_{n_\circ}$ and $\hat{P}^\perp_{n_\circ} \to I - \bar{L}'_{n_\circ}(\bar{L}_{n_\circ}\bar{L}'_{n_\circ})^{-1}\bar{L}_{n_\circ} = P^\perp_{n_\circ}$. The estimation errors are derived using the uniform convergence of the covariance estimates: The main emphasis here lies on $\hat{\mathscr{K}}_p - \mathscr{K}_p$. It is straightforward to show, using Lemmas 1 and 2 that there exists a matrix $\hat{S}_T$ such that $\|\hat{S}_T\hat{\mathscr{K}}_p - \mathscr{K}_p\|_2 = O(Q_T\sqrt{p})$. Applying Lemma 1 to $\langle y_t, Z^-_{t+1,p+1}\rangle$ this also implies $\|\hat{\mathscr{H}}_1 - [C, D]\bar{L}_{n_\circ}\|_2 = O(Q_T\sqrt{p})$ as well as $\|\hat{P}^\perp_{n_\circ} - P^\perp_{n_\circ}\|_2 = O(Q_T\sqrt{p})$. Therefore consider

$$\begin{aligned}
IVC(n) - IVC(n_\circ) &= (d(n) - d(n_\circ))\frac{C(T)}{T} + \log(\det\hat{\Omega}_n/\det\hat{\Omega}_{n_\circ}) \\
&= (d(n) - d(n_\circ))\frac{C(T)}{T} + \log(\det[I + (\hat{\Omega}_n - \hat{\Omega}_{n_\circ})\hat{\Omega}^{-1}_{n_\circ}]) \\
&= (d(n) - d(n_\circ))\frac{C(T)}{T} + \operatorname{tr}[(\hat{\Omega}_n - \hat{\Omega}_{n_\circ})\hat{\Omega}^{-1}_{n_\circ}] \\
&\quad + o(\|\hat{\Omega}_n - \hat{\Omega}_{n_\circ}\|) \\
&= (d(n) - d(n_\circ))\frac{C(T)}{T} + O(Q_T^2 p)
\end{aligned}$$

as follows from a Taylor series expansion of $\log(1 + x)$. Thus

$$\frac{T}{C(T)}(IVC(n) - IVC(n_\circ)) = d(n) - d(n_\circ) + O(p\log\log T/C(T)).$$

This shows the following result:

Theorem 4. *Let the process $(y_t;\ t \in \mathbb{Z})$ be generated by a system of the form* (1), *where the true system $(k, l) \in M_{n_\circ}$, where $n_\circ$ denotes the true order. Let the noise fulfil the assumptions of Section* 2 *and let the input fulfil Assumptions* 1. *Let the system be estimated according to the Larimore type of procedure using $f \geqslant n_\circ$ and $p = p(T) \to \infty$, where $\max\{f, p\} = O((\log T)^a)$ for $a < \infty$.*

Then the order estimate obtained as the minimizing argument of $IVC(n)$ using a penalty term $C(T) > 0$, $C(T)/T \to 0$ and $C(T)/(p\log\log T) \to \infty$ is a.s. consistent, if $\det[\Omega_{n_\circ - 1}] > \det[\Omega_{n_\circ}]$. If $\det[\Omega_{n_\circ - 1}] = \det[\Omega_{n_\circ}]$ then the order is underestimated a.s. asymptotically.

The theorem leads to a penalty term, which has to be slightly higher than $p\log\log T$ and therefore the choice $\log T$ seems to be a reasonable choice for the usual choice of p (see the simulation section) noting that $\log\log T$ is small even for relatively large T, although not theoretically justified for the Larimore type of methods, where f and p tend to infinity. This result is new, as in Bauer (1998) much more severe restrictions on the penalty term have been used. The restriction $\det[\Omega_{n_\circ - 1}] > \det[\Omega_{n_\circ}]$ is worth being investigated further. The fundamental difference of the criterion $IVC(n)$ as compared to the information criteria, although formally defined analogously, is that the innovation variance is calculated for truncated states only, rather than newly computed states. However the first n components of $x_t = \mathscr{K}_\infty Z^-_{t,\infty}$ need not be generated by a state space equation of order n for $n < n_\circ$, i.e. the matrix $\mathscr{K}_\infty$ might not have the shift invariance structure $\mathscr{K}_{2:\infty} = \bar{A}_n\mathscr{K}_{1:\infty}$ for any matrix $\bar{A}_n$ of dimension $n \times n$, using obvious notation to denote submatrices. Therefore the criterion only measures the direct influence of the state coordinates on the prediction of y_t, but it does not take into account the dynamical generation of the state. Thus in the case, where a state does not contribute to the present of the output, but only to the future, it will be neglected according to the criterion given above. As the cited results show, this might be an extremely rare situation. The main concern in this respect is, that in situations, where the contribution is small, the same behaviour is expected, i.e. many observation will be needed in order to detect this state component. In the next section an example for this will be given.

5. Numerical examples

In this section three different examples are presented in order to compare the various proposed order estimation methods. The candidate order estimation algorithms will be $SVC(n)$, $IVC(n)$ as presented above, $NIC(n)$ as presented by Peternell (1995) and $MOE(n)$, which is implemented in the `N4SID` procedure of the system identification toolbox of `MATLAB` (Ljung, 1991): The idea here is to formalise the search for a "gap" in the singular values.

The order is estimated as

$$\hat{n} = \max\{n\colon \log\hat{\sigma}_n > \tfrac{1}{2}(\log\hat{\sigma}_1 + \log\hat{\sigma}_M)\},$$

i.e. the largest integer, such that the corresponding singular value is greater than the geometric mean of the largest and the smallest nonzero estimated singular value. The three examples include a low order single input single output system without exogenous inputs, where the order is expected to be easy to find, another SISO system without exogenous inputs, where the order is expected to be hard to identify, and finally a MIMO system with a two dimensional observed input. The main points of interest are the effects of the choice of the penalty term, the weighting matrices, the integer parameters f and p, and of course a comparison between the various procedures.

5.1. Example I

As a first example consider the system defined by the following matrices:

$$A = \begin{bmatrix} 0 & 1 \\ -0.7 & 0.5 \end{bmatrix}, \quad K = \begin{bmatrix} 1.3 \\ 0.3 \end{bmatrix}, \quad C = [1, 0].$$

This system has Lyapunov balanced Gramian of roughly $\Sigma = \mathrm{diag}(2.55, 1.78)$. The system poles are at $0.25 \pm 0.7984i$ and the zeros at $-0.4 \pm 0.4359i$. In the estimation two different weighting schemes are used: CCA uses $\hat{W}_f^+ = (\hat{\Gamma}_f^+)^{-1/2}$ and the method using $\hat{W}_f^+ = I$ will be labelled N4SID. The restriction to these two choices is arbitrary and only justified by the fact, that these choices seem to be the most widely used ones, in the following example different weightings will be used. Note that the label N4SID is not to be confused with the procedure N4SID introduced by Van Overschee and DeMoor (1994). Here only the same weighting scheme is treated, the actual algorithm however is not used. The indices $f = p = \hat{p}_{\mathrm{AIC}}$ are used. Here $\hat{p}_{\mathrm{AIC}}$ denotes the order estimate in a long autoregression to explain y_t as $y_t = a_1 y_{t-1} + \cdots + a_p y_{t-p} + e_t$, where the order is estimated using AIC. It is well known, that in the present setting $\hat{p}_{\mathrm{AIC}}$ tends to infinity at the rate $\log T$. From theoretical considerations as well as from practical point of view thus $f = p = \lfloor d\hat{p}_{\mathrm{AIC}} \rfloor$ for small values of $d > 1$ seems to be an appropriate choice (for a discussion on this see e.g. Bauer, 1998). For each of the weighting schemes, the order of the state space system is estimated using four different methods: IVC and SVC with $C(T) = \log T$ (denoted with IVC1 and SVC1, respectively, in the sequel), IVC and SVC with $C(T) = fp\log T$ (denoted with IVC2 and SVC2, respectively). Note, that only for the last two procedures the consistency results have been derived. One thousand time series of length 100, 1000 and 5000, respectively, have been generated and used for estimation. Table 1 shows the results for $T = 100$, $T = 1000$ and $T = 5000$ respectively. They show, that the performance of the order estimation procedure depends heavily on the weighting scheme: For CCA the IVC1 method works well, whereas it shows problems to estimate the true order, when used with N4SID. This is due to the fact, that in the Lyapunov balanced realization of the true system, the entry $C_{1,2}$ is equal to -0.0146 and thus close to zero. This leads to a high risk of underestimating the order using IVC together with N4SID in this example. For CCA it is observed, that as has been expected, the higher penalty term results in a high risk of underestimation, while reducing the risk of overestimation. For N4SID the SVC method outperforms IVC and it is also observed, that for $C(T) = fp\log T$ the accuracy increases with the sample size, whereas the lower penalty term does not seem to lead to consistent order estimates. In the CCA case it is seen, that the higher penalty term leads to a big risk of underestimating the order for small sample sizes. On the other hand for the N4SID weighting the smaller penalty leads to a high risk of overestimation. Therefore no clear decision about the choice of the penalty has been found. Both

Table 1
Here the probability of estimating the indicated order for 1000 time series of sample size T is shown for two different weighting schemes (CCA and N4SID) and 4 different estimation methods: $IVC(n)$ with $C(T) = \log T$ (IVC1) and $C(T) = fp\log T$ (IVC2) and SVC with $C(T) = \log T$ (SVC1) and $C(T) = fp\log T$ (SVC2). $f = p = \hat{p}_{\mathrm{AIC}}$ has been used

	T	100			1000			5000		
	Est. order	<2	2	>2	<2	2	>2	<2	2	>2
CCA	IVC1	0.00	0.83	0.17	0.00	0.77	0.23	0.00	0.67	0.33
	IVC2	1.00	0.00	0.00	0.63	0.37	0.00	0.05	0.95	0.00
	SVC1	0.00	0.94	0.06	0.00	0.93	0.07	0.00	0.94	0.06
	SVC2	1.00	0.00	0.00	0.86	0.14	0.00	0.09	0.91	0.00
N4SID	IVC1	0.82	0.03	0.15	0.68	0.06	0.26	0.50	0.13	0.37
	IVC2	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
	SVC1	0.00	0.66	0.34	0.00	0.40	0.60	0.00	0.37	0.63
	SVC2	0.45	0.55	0.00	0.04	0.96	0.00	0.00	1.00	0.00

Table 2
This table shows the estimated means of the various order estimation procedures as a function of sample size and different weighting matrices $\hat{W}_f^+$. Here SVC, NIC and IVC use the penalty term $C(T) = \log T$. The table has been produced using 1000 replications in each case

$\hat{W}_f^+$	T	Method $d = 2$				Method $d = 4$			
		IVC	SVC	NIC	MOE	IVC	SVC	NIC	MOE
CCA	100	3.50	2.30	5.50	3.67	2.07	2.31	6.32	4.03
	250	5.80	2.78	7.06	5.23	5.23	2.87	15.25	10.89
	500	6.53	3.37	7.64	6.03	5.55	3.43	17.23	13.48
	1000	7.53	4.03	7.74	6.65	6.63	4.11	17.90	15.14
low pass	100	5.12	5.23	5.91	5.24	3.99	5.92	6.85	5.93
	250	8.43	6.91	8.04	6.92	11.10	13.12	16.41	13.15
	500	9.61	7.68	9.10	7.69	12.64	15.26	19.27	15.34
	1000	10.43	8.51	10.10	8.52	14.32	16.68	21.09	16.76
high pass	100	3.97	5.60	6.30	5.64	2.57	6.43	7.21	6.50
	250	6.93	7.55	8.75	7.69	6.89	14.50	17.83	15.11
	500	8.12	8.38	9.92	8.61	7.82	16.83	20.89	17.66
	1000	9.08	9.30	11.04	9.55	9.86	18.34	22.85	19.36

choices used in this example are heuristic and not motivated by additional arguments. A theoretical justification seems to be needed.

5.2. Example II

Next, the various order estimation procedures will be tested on an eight order system with poles at $z = 0.8e^{\pm 0.2i\pi}$, $z = 0.7e^{\pm 0.3i\pi}$, $z = 0.5e^{\pm 0.5i\pi}$, $z = 0.6e^{\pm 0.4i\pi}$ and zeros at $z = 0.8e^{\pm 0.1i\pi}$, -0.4755,0.1,0.3,0. Using this example extensive simulations comparing the order estimation criteria have been performed. The system order is hard to estimate and consistent estimates of the order are not the main goal in this example. The Lyapunov balanced Gramian is equal to diag(6.85, 4.46, 1.08, 0.39, 0.045, 0.015, 0.0004, 0.0002) and thus the system is expected to be approximated well using a fourth order system. A couple of different setups have been tested. In a first simulation study 1000 replications of time series of sample lengths $T = 100$, 250, 500 and 1000 have been generated. In the subspace algorithms three different weighting matrices $\hat{W}_f^+$ have been applied: The CCA weights, a low-pass filter, generated using a 6th order butterworth filter with cutoff frequency 0.5π and the corresponding high pass filter have been incorporated. The choice of the cutoff frequency is arbitrary and not problem oriented. The only purpose of using these weighting schemes is to investigate their effects on the estimated order and the estimated transfer functions. Further $f = p = d\hat{p}_{\mathrm{AIC}}$ has been used in all cases, where $d = 2$ and 4 are tried. The choice $d = 4$ leads to comparably large values of f and p. These two different choices are used to investigate the sensitivity of the order estimation criteria on the size of the matrix, which is decomposed in the algorithm. The average values of the corresponding order estimates are given in Table 2. It can

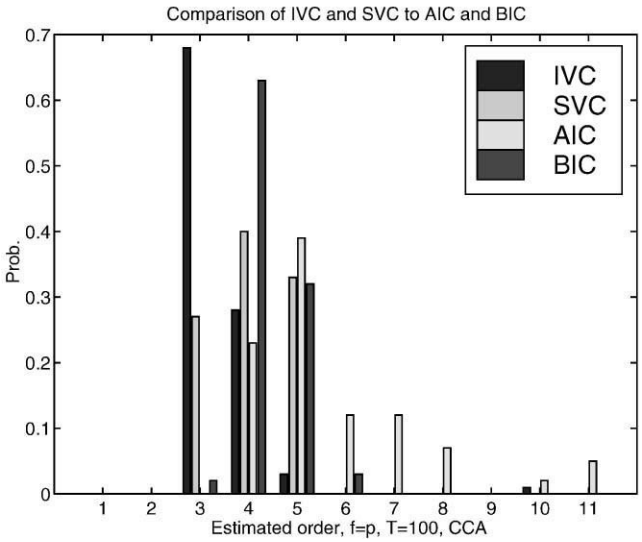


Fig. 1. In this figure the order estimates obtained by SVC and IVC using $C(T) = \log T$ are compared to the estimates obtained in the ML framework using AIC and BIC. $T = 100$ and $f = p = 2\hat{p}_{\mathrm{AIC}}$ are used together with the CCA weighting scheme. The plots have been obtained using 100 replications.

be seen, that the behaviour of the various algorithms is very different for different weightings $\hat{W}_f^+$. For the CCA weighting NIC gives values close to the true order for $d = 2$, while it results in overly large estimates for $d = 4$. Also MOE seems to suffer from the bigger choice of d, whereas both SVC and IVC are relatively robust with respect to this choice. For the low pass weighting all estimation procedures show a tendency to overestimate the system order by a factor of two for $d = 4$, and also the results for $d = 2$ are large compared to the CCA case. The same result also holds for the high pass weighting, except that the estimates of IVC for $d = 4$ are better than the

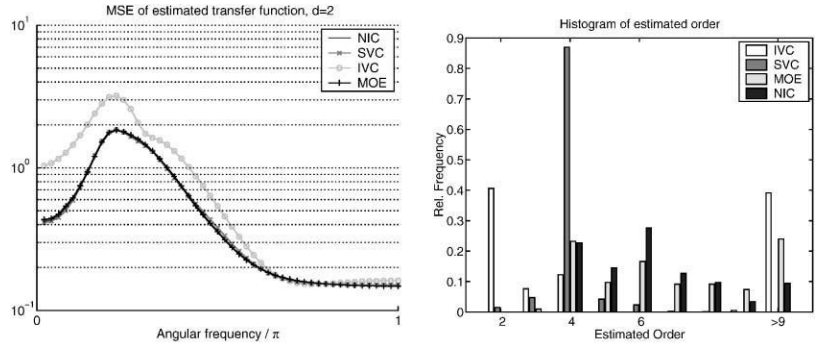


Fig. 2. These plots show the result of the simulation for $T = 1000$. The left picture shows the average mean square error of the transfer function estimates at 50 equally spaced frequencies in the angular frequency range $\omega \in [0, \pi)$ obtained using the various order estimation procedures with the CCA weighting scheme and $f = p = 2\hat{p}_{\mathrm{AIC}}$. The right plot shows the corresponding histogram for the estimated orders. The plots have been obtained using 1000 replications.

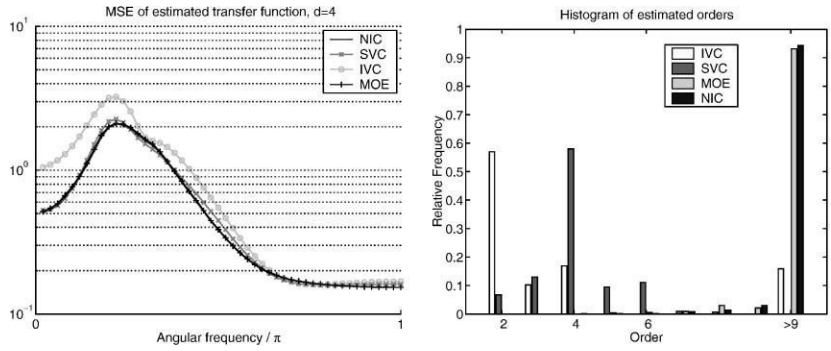


Fig. 3. These are the same plots as shown in Fig. 2 with respect to the choice $f = p = 4\hat{p}_{\mathrm{AIC}}$.

respective estimates of the other order estimation procedures. This indicates, that $d = 2$ is the favourable choice, as compared to $d = 4$.

The order estimation procedur es are also compared to the more traditional maximum likelihood based information criteria. In Fig. 1 a histogram for the estimated orders using IVC, SVC, AIC and BIC is given. The latter two criteria estimate the order as the minimizing argument of the following function:

$$IC(n) = \log\det(\hat{\Omega}_n) + C(T)d(n)/T,$$

where $d(n)$ denotes the number of parameters as in IVC. $\hat{\Omega}_n$ denotes the pseudo maximum likelihood estimates of the innovation covariance specifying the system order as n. AIC uses $C(T) = 2$, whereas BIC uses $C(T) = \log T$. Here $T = 100$ and $f = p = 2\hat{p}_{\mathrm{AIC}}$ have been used. It can be observed, that BIC tends to choose $n = 4$ with a high probability, while AIC selects relatively large orders. The two subspace order estimates lead to slightly smaller order estimates. Especially the results for SVC and BIC seem to be comparable.

However, the order estimate might be seen to be not the only interesting indicator. Therefore also the resulting estimates of the system are considered. The right plot of Fig. 2 shows the square root of the mean squared error of the estimated transfer function (estimated from 1000 replications) in the angular frequency range $[0, \pi]$ obtained by CCA using $d = 2$ for the four subspace based procedures. Here the sample size is equal to $T = 1000$ and $C(T) = \log T$ is used for SVC, IVC and NIC. The figures show, that the IVC estimates are worse, despite the fact, that the average estimated order seems to be the best for this scenario. This is explained in the right plot of Fig. 2, which gives the histogram of the order estimates: In the IVC case there is a relatively high portion of low order systems (over 50% are less than $n = 4$), as well as a high number of overly large estimates (35% larger than $n = 10$). This combines to a high bias, which shows in the mean square error. The NIC and MOE perform about equal, due to the very similar distribution of the order estimates. The SVC method leads to a mean square error almost identical to the one obtained by using NIC or MOE, while choosing smaller orders on average, which might be seen as an advantage. The results for $d = 4$ are similar with one exception: Contrary to what has been said before, the SVC method reacts much larger to the change of d than NIC and MOE with respect to the mean square error. This is due to the fact, that there is a higher percentage of low orders estimated in this case leading to a high bias error. The results for NIC and MOE are not

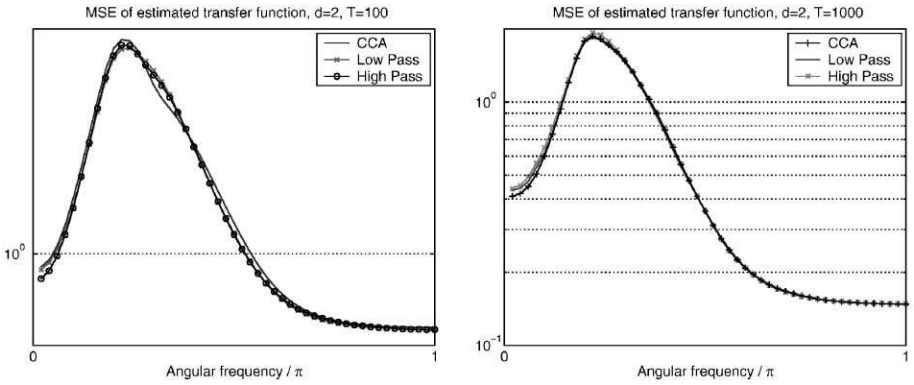


Fig. 4. These plots show the result of the simulation for $T = 100$ (left picture) and $T = 1000$ (right picture). The pictures shows the average mean square error of the transfer function at 50 equally spaced frequencies in the angular frequency range $\omega \in [0, \pi)$ obtained using the IVC procedure (left plot) and the MOE procedure (right plot) for the three different weighting scheme and $f = p = 2\hat{p}_{\mathrm{AIC}}$. The plots are based on 1000 replications.

that sensitive, although on average much higher orders are chosen. The corresponding pictures are given in Fig. 3. Finally also the effect of the weightings on the mean square error is discussed: Fig. 4 shows two plots, where the left one refers to $T = 100$ and the order estimate according to IVC with penalty $\log T$. The right plot shows the result for MOE and $T = 1000$. In both cases there is somewhat surprisingly hardly any difference due to the choice of the weighting matrices. A similar picture holds for the other cases as well. This observation is in contrast to the observations in the results of simulations with fixed order, where an influence of the weighting matrices with respect to the mean square error has been observed (see e.g. Bauer, 1998). It is remarked, that this observation might not be typical, but is certainly worth to be investigated further.

5.3. Example III

As a last example also a system with observed exogenous inputs is treated: Consider the system given by the following matrices:

$$A = \begin{bmatrix} 0.8 & 0.2 \\ -0.4 & -0.5 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & -1 \\ 1 & 0.5 \end{bmatrix},$$

$$K = \begin{bmatrix} 1.5 & 0 \\ -0.2 & -0.8 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and $C = I_2$, the two × two identity matrix. The poles of the system are 0.7352 and -0.4352, the zeros of k_0 are at -0.6583 and 0.2583, whereas the zeros of l_0 are at $-0.1000 \pm 0.9327i$. Fig. 5 shows the probabilities of the order estimates for different noise covariances $\Omega = s^2 I_2$, where the input is i.i.d uniformely distributed white noise with zero mean and unit variance. In the 1000 replications of sample size $T = 200$ the choices $f = p = 2\hat{p}_{\mathrm{AIC}}$ and the CCA weighting scheme have been used. The penalty term was equal to $C(T) = \log T$ in all cases. The

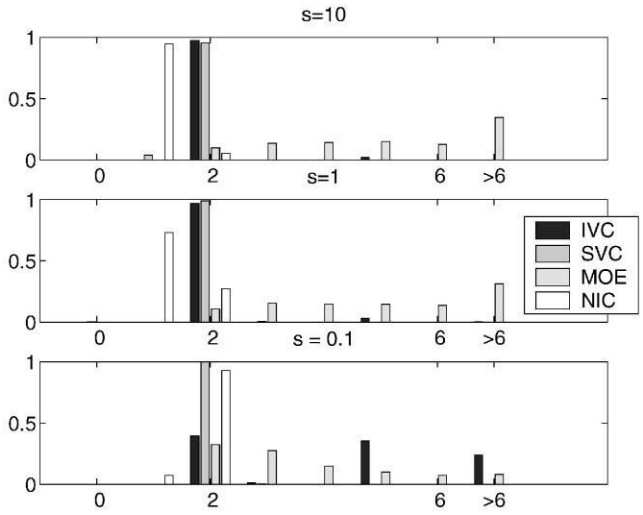


Fig. 5 . This figure shows the result of 1000 simulation runs using the noise level $s = 10$ (top row), $s = 1$ (middle row), $s = 0.1$ (bottom row). Each picture shows the probabilities of estimating the order using the four estimation algorithms IVC, SVC, MOE and NIC: The weightings have been chosen according to CCA. The truncation indices have been chosen as $f = p = 2\hat{p}_{\mathrm{AIC}}$. The inputs are i.i.d. uniformely distributed white noise normalized to zero mean and unit variance. For SVC and IVC the penalty term $C(T) = \log T$ is used.

three different noise levels were $s = 10$, 1 and 0.1, respectively. It can be seen, that IVC has a tendency to overestimate the order for small noise contribution, whereas NIC underestimates the order for high noise level frequently. Note, that for the case of additional exogenous inputs Peternell (1995) suggests to use $d(n) = (p(s + m) - n)(fs - n)$, which essentially leads to a bigger penalty term of the form $n(sf + p(m + s))C(T)$, which is used for this order estimation scheme. The order estimation procedure implemented in MATLAB, i.e. MOE, shows a tendency to overestimate the order in all cases. In this example SVC shows the best performance, however, further undocumented simulations show, that this is sensitive to the choice of f and p. This example was

chosen merely to illustrate that the proposed methods also work in the case of exogenous inputs. It should be noted again, that the order estimation techniques, except for the IVC method, apply equally to the MOESP type of methods.

5.4. Summary of simulations

The points investigated in the simulation section were the comparison in between the various estimation methods, the sensitivity with respect to the choice of the indices f and p and the choice of the penalty term. In this respect the first example showed, that the choice $C(T) = \log T$ for SVC and IVC leads to a serious threat of overestimation of the order, while leading to accurate estimates for small samples. The choice $C(T) = fp \log T$ on the other hand showed clearly convergent behaviour and a high rate of misspecifications for small samples. Thus a reasonable choice of the penalty term could lie in between. Further work is required to find motivations for particular choices. The example also showed, that for IVC there are systems, where the estimation leads to a high risk of underestimation even for high sample sizes. The second example tried to evaluate the effect of different choices of f and p, reassuring that for the CCA weighting scheme both the order estimates obtained by SVC and IVC react less sensitive with respect to these values, whereas both NIC and the procedure implemented in MATLAB show a high dependency on these parameters. These findings motivate the choice of $1 \leqslant d \leqslant 2$ with respect to order estimation. The results of this second example also show a large impact of the choice of the weighting scheme on the estimated order. However no systematic behaviour has been observed and also this point seems to be worth to be investigated further. Finally the third example simply shows, that also in the case of exogenous inputs present the order estimation procedures are capable of delivering suitable estimates, which show consistent behaviour. Summing up it can be stated, that no single criterion can be isolated as the best choice.

6. Conclusions

In this paper the question of order estimation in the context of subspace methods has been addressed. Two new procedures have been proposed and analysed. Lower bounds on the penalty term in order for the estimates to be (strongly) consistent have been given. The method using the innovation variance has been shown to suffer from severe theoretical disadvantages and thus the use of this intuitively appealing procedure is discouraged. For the SVC criterion the advantages certainly are the possibility to obtain an estimate of the order with almost no computational costs, as only the properties of the estimates of the singular values, which are estimated in any case, are used. In a simulation study it has been demonstrated, that the methods lead to reasonable results. It has been shown, that SVC is less sensitive to the choice of the truncation integers f and p than the criterion introduced by Peternell (1995) or the method used in the system identification toolbox of Ljung (1991). However the SVC criterion also contains a subjective component in the choice of the penalty term. In the simulations no clear picture on how this should be chosen could be obtained and no heuristical motivation for any particular choice has been found. This seems to be a rewarding question for future research.

Acknowledgements

The author acknowledges financial support in part by the European Commission through the program Training and Mobility of Researchers—Research Networks and through project System Identification (FMRX CT98 0206) and acknowledges contacts with the participants in the European Research Network System Identification (ERNSI).

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, *21*, 243–247.

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

Bauer, D. (1999). Order estimation in the context of MOESP subspace identification methods, *Proceedings of the ECC '99 conference*, Karlsruhe, Germany.

Bauer, D., & Jansson, M. (2000). Analysis of the asymptotic properties of the MOESP type of subspace algorithms. *Automatica*, *36*(4), 497–509.

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

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

Hannan, E. J., & Deistler, M. (1988). *The statistical theory of linear systems*. New York: John Wiley.

Jansson, M., & Wahlberg, Bo. (1997). Counterexample to general consistency of subspace system identification methods, *Proceedings of SYSID '97*. Fukuoka, Japan (pp. 1677–1682).

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

Ljung, L. (1991). *System identification toolbox*, *user's guide*. The MathWorks.

Ljung, L. (1999). *System identification: theory for the user* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

McKelvey, T. (1995). *Identification of state-space models from time and frequency data*. Ph.D. thesis, Department of Electrical Engineering, Linköping.

Peternell, K. (1995). *Identification of linear dynamic systems by subspace and realization-based algorithms*. Ph.D. thesis, TU Wien.

Peternell, K., Scherrer, W., & Deistler, M. (1996). Statistical analysis of novel subspace identification methods. *Signal Processing*, *52*, 161–177.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics*, *8*(1), 147–164.

Sorelius, J. (1999). *Subspace-based parameter estimation problems in signal processing.* Ph.D. thesis, Uppsala University.

Van Overschee, P., & DeMoor, B. (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.

Dietmar Bauer was born in St. Pölten, Austria, in 1972. He received his masters and Ph.D. degrees in Applied mathematics from the Technical University Vienna in 1995 and 1998, respectively. Since 1995 he was with the Institute for Econometrics, Operations Research and System Theory at the Technical University of Vienna. In late 1999 he held a postdoc position at the University of Newcastle, Australia and in 2000 he spent 6 months as a postdoc at Linköping University, Sweden. Since June 2000, he also holds a position in the Industrial Mathematics Competence Center, Linz, Austria. His main research interest lies in system identification, in particular exploring the capabilities of subspace algorithms with respect to the analysis of economic data.

ELSEVIER

Journal of Econometrics 111 (2002) 47–84

JOURNAL OF Econometrics

www.elsevier.com/locate/econbase

Estimating cointegrated systems using subspace algorithms

Dietmar Bauer[a,1], Martin Wagner[b,*]

[a] *Institute for Econometrics, Operations Research and System Theory, TU Wien, Argentinierstr. 8, A-1040 Vienna, Austria*

[b] *Department of Economics, University of Bern, Gesellschaftsstrasse 49, CH-3012 Bern, Switzerland*

Abstract

The properties of the so-called subspace algorithms, up to now used almost only for stationary processes, are investigated in the context of cointegrated processes of order 1. It is shown for one of these algorithms that it can be adapted to deliver consistent estimates of all system parameters in the case of general I(1) VARMA models and mild conditions on the underlying noise. Estimates of the cointegrating space are derived and several test procedures for the cointegrating rank are proposed. Consistent estimation of the system order is also discussed. A simulation study shows the usefulness of subspace algorithms for estimation of and testing in cointegrated systems.

JEL classification: C13; C32

Keywords: Cointegration; Subspace algorithms; State space representation

1. Introduction

In econometrics it is common practice to analyze integrated or cointegrated processes using their vector autoregressive (VAR) or vector autoregressive moving average (VARMA) representation. For linear processes the state space representation is an alternative and equivalent representation, which turns out to be very convenient for the analysis of cointegration. Based on the discussion of state space models for integrated processes (of order 1) we propose a simple estimation procedure for all

* Corresponding author. Tel.: +41-31-6314778; fax: +41-31-6313992.

E-mail address: martin.wagner@vwi.unibe.ch (M. Wagner).

[1] Support by the Austrian FWF under the project number P-14438-INF is gratefully acknowledged. Part of this work has been done while this author was on leave at the University of Newcastle, Australia.

0304-4076/02/$ - see front matter
PII: S0304-4076(02)00119-7

system parameters, two tests for the cointegrating rank and one estimation procedure for the cointegrating rank.

The estimate is based on the so-called subspace algorithms. These have been developed in the engineering literature over the past 15 years for stationary processes. The computational cost amounts to performing OLS regressions and one singular value decomposition. Thus, especially for VARMA processes, the computational cost is much lower than for the nonlinear optimization problem that has to be solved in (pseudo) maximum likelihood estimation.

The only application of subspace algorithms for nonstationary processes so far is Aoki (1990, Chapter 9). His procedure however lacks a thorough statistical foundation including issues of estimating integer parameters like the system order and the cointegrating rank. Aoki's procedure furthermore can be shown to be inefficient for stationary processes.

A couple of variants of subspace algorithms have been developed, e.g. Larimore (1983), Van Overschee and DeMoor (1994) or Verhaegen (1994). In recent years, the asymptotic theory has been developed for the stationary case in several papers: Deistler et al. (1995) and Peternell (1995) discuss consistency. Bauer (1998) and Bauer et al. (1999) establish central limit theorems for the estimates and also derive consistent order estimation procedures. Bauer (2002) shows that Larimore's (1983) `CCA` algorithm is asymptotically equivalent to pseudo-maximum likelihood analysis for stationary systems. I.e. for stationary processes this method results in estimates that have the same asymptotic variance as those obtained by maximizing the Gaussian likelihood function.

Given the above-mentioned result of Bauer (2002) and the fact that the `CCA` algorithm is especially suited for the analysis of multivariate time series without exogenous variables, this paper is confined to this procedure, or to an adapted version for integrated processes to be precise. Based on this procedure we derive tests for the number of unit roots and therefore for the dimension of the cointegrating space. These tests are based on the estimated singular values from the singular value decomposition performed in the algorithm, or on the estimated eigenvalues of the matrix describing the state transition respectively (the details are given in Section 3). The eigenvalue-based test is relying on arguments in the spirit of Stock and Watson (1988).

Let us state once again that the analysis is restricted to processes where the only unit roots are located at one and where also the integration order is restricted to one, thus, e.g. I(2) processes or processes with seasonal unit roots are excluded up to now. As will be seen below, the restriction of an integration order 1 corresponds in the state space representation to the assumption that the eigenvalues of the system at one are simple.

Our work is of course not the first to deal with cointegration analysis in the context of VARMA processes. Yap and Reinsel (1995) derive the maximum likelihood estimator for cointegrated Gaussian VARMA processes integrated of order one. Saikkonen (1992) derives consistency of Johansen (1995) type estimates for cointegrated VARMA processes if the lag length of an autoregressive approximation is increased with the sample size at a sufficient rate. For both of the above approaches, the asymptotic null distributions of the tests for the cointegrating rank are the same as for the Johansen

procedure for VAR processes. Wagner (1999) shows that the Johansen procedure applied with a fixed lag on an underlying VARMA process results in consistent estimates of the cointegrating space. The short-run parameters however are not estimated consistently anymore in this case. Another strand of the literature is based on (static or dynamic) regressions, with possible correction factors for serial correlation. These approaches are in a way nonparametric in that they focus on the testing for and estimation of cointegrating relationships. They usually neglect the estimation of the other system parameters, these however may usually be recovered in a second step, due to usual super-consistency of the estimated cointegrating relationships.[2] See e.g. Phillips (1991, 1995), Stock and Watson (1988), Bewley and Yang (1995) or Poskitt (2000) from a long list of contributions.

We believe that our approach represents a valuable additional tool for several reasons. First, it introduces the state space representation (in a canonical form) of integrated processes and highlights its properties. Second, it brings to the attention of the econometrics community results that have up to now been almost exclusively discussed in a stationary setting in the systems engineering literature. The consistency of one of these procedures also for integrated processes that is derived in this paper thus points to the potential usefulness of these developments also for econometric analysis. The applicability to VARMA processes and the computational simplicity are further advantages. The results thus allow at least for a cheap "cross-validation" of results derived with standard methods, like e.g. the Johansen procedure.

The estimates could also be used as consistent initial values to obtain efficient estimates of the parameters performing one Newton step for pseudo-maximum likelihood estimation, as presented in Yap and Reinsel (1995) or in Bauer and Wagner (2000b). If only used as initial values, the subspace estimates still have the additional advantage of providing also initial (and consistent) information concerning the structure of the system (i.e. the system order and the cointegrating rank).

Both, the simulation results and first applications (see e.g. Bauer and Wagner (2000a), for an application to interest rate data), indicate that the method performs at least comparable to standard methods like e.g. the Johansen method, with the additional advantage of providing consistent estimates of all system parameters for VARMA processes in a computationally simple fashion.

The paper is organized as follows: In Section 2 the state space framework is introduced and its relation to the VARMA representation of linear stochastic processes is discussed. In Section 3, subspace algorithms are introduced and discussed. Section 4, states the theoretical results for the method presented in Section 3 for the stationary and, which constitutes one of the main results of the paper, for integrated processes. Section 5 is devoted to derive consistent estimates of the system order and to the development of tests for the cointegrating rank. In Section 6 simulation results to assess the finite sample properties of our methods are presented and Section 7 summarizes and concludes. All proofs are collected in Appendix A and in Appendix B the gap metric (used in some of the proofs) is defined and the simulated systems are given.

[2] Thus, these regression-based approaches may often be seen as descendants of the seminal Engle and Granger (1987) 2-step procedure.

2. State space models

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

$$x_{t+1} = Ax_t + K\varepsilon_t, \quad y_t = Cx_t + E\varepsilon_t, \tag{1}$$

where y_t denotes the s-dimensional output series observed for $t = 0, \ldots, T$. ε_t denotes an s-dimensional white noise sequence. $A \in \mathbb{R}^{n\times n}$, $K \in \mathbb{R}^{n\times s}$, $C \in \mathbb{R}^{s\times n}$, $E \in \mathbb{R}^{s\times s}$ and $x_t \in \mathbb{R}^n$ denote the n-dimensional state sequence. Throughout the paper ε_t is assumed to be an ergodic strictly stationary martingale difference sequence for which the following conditions hold:

$$\mathbb{E}\{\varepsilon_t | \mathscr{F}_{t-1}\} = 0, \quad \mathbb{E}\{\varepsilon_t \varepsilon_t' | \mathscr{F}_{t-1}\} = \mathbb{E}\{\varepsilon_t \varepsilon_t'\} = I_s, \tag{2}$$

$$\mathbb{E}\{\varepsilon_{t,a}\varepsilon_{t,b}\varepsilon_{t,c} | \mathscr{F}_{t-1}\} = \omega_{a,b,c}, \quad \mathbb{E}\varepsilon_{t,a}^4 < \infty, \tag{3}$$

where $\varepsilon_{t,a}$ denotes the ath component of the vector ε_t and $\mathscr{F}_{t-1}$ denotes the σ-algebra spanned by the past, i.e. by $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_0$ and x_0. $\omega_{a,b,c}$ is a constant and I_s denotes the $s \times s$ identity matrix. The matrix E is assumed to be nonsingular and lower triangular with positive entries on the diagonal. This restriction is necessary to ensure identifiability of E and ε_t. The above conditions will be referred to as *standard assumptions* throughout the paper. The assumptions concerning the noise exclude ARCH effects. The extension of the method to cover also heteroskedastic innovations is a topic of further research.

In the systems theory literature it is well known that state space systems and VARMA systems are just two different representations of the same object. To acquaint the reader with the properties of state space and VARMA representations, the main facts relevant for this paper are collected in this section. For a more detailed discussion see e.g. Hannan and Deistler (1988, Chapter 1). Define the mapping π as $\pi(A,K,C,E) = k(z) = E + zC(I_n - zA)^{-1}K$, i.e. π maps the state space system (A,K,C,E) to its corresponding transfer function $k(z)$. Here z denotes both, a complex variable and the backward shift operator. From the above definition of $k(z)$, it is directly seen that $k(z)$ is a rational transfer function. It can furthermore be easily verified that also conversely for each rational transfer function $k(z)$ there exists (at least one) state space system (A,K,C,E) such that $\pi(A,K,C,E) = k(z)$. The same type of relationship also holds between VARMA representations of systems and the corresponding transfer function. So we can analogously define a mapping $\bar{\pi}$ as $\bar{\pi}(a,b) = k(z) = a^{-1}(z)b(z)$ for $a(z)$ and $b(z)$ matrix polynomials. Neither the VARMA nor the state space representations are unique. For fixed transfer function $k(z)$ the sets $\{(a,b)\colon \bar{\pi}(a,b) = k(z)\}$ and $\{(A,K,C,E)\colon \pi(A,K,C,E) = k(z)\}$ are called *equivalence sets*. For VARMA systems equivalence sets are described using polynomial matrices (as $\bar{\pi}(a,b) = \bar{\pi}(pa, pb)$ holds for all polynomial matrices p with $p(0)$ nonsingular) for state space representations nonsingular matrices have the same function: It is obvious, that $\pi(A,K,C,E) = \pi(TAT^{-1}, TK, CT^{-1}, E)$. In this case the state space systems (A,K,C,E) and $(TAT^{-1}, TK, CT^{-1}, E)$ are called *observationally equivalent*. A state

space representation of a transfer function $k(z)$ is called *minimal*, if no other state space representation with smaller state dimension exists. Under the assumption of minimality, all observationally equivalent state space systems are described by transformations with nonsingular matrices, as above.

The concept of minimality is linked to three matrices: To the *observability* matrix $\mathcal{O} = [C', A'C', (A^2)'C', \dots]'$, to the *controllability* matrix $\mathcal{C} = [K,\ AK,\ A^2K, \dots]$ and to the *Hankel* matrix $\mathcal{H} = \mathcal{O}C = [CA^{i+j-2}K]_{i,j=1,\dots}$. For a minimal system all three matrices have rank equal to the system dimension or system order, usually denoted as n. The nonsingular matrices T that generate observationally equivalent state space systems, correspond to a change in the state space basis and result in a different factorization of the Hankel matrix as $\mathcal{H} = [\mathcal{O}T^{-1}][T\mathcal{C}]$.

Assume e.g. that a system representation leads to an observability matrix $\mathcal{O}$ not having full column rank. In this case there exists a state vector, $\bar{x}$ say, such that $\mathcal{O}\bar{x} = 0$. This implies that components of the state in this direction $\bar{x}$ have no influence on y_t and can therefore be omitted from the system description. In other words (after an appropriate basis change) the state dimension can be reduced without changing the input–output characteristics of the model, when the observability matrix does not have full rank. Similar lines of thought can be applied to the controllability matrix.

Note at this point that for a minimal system, the eigenvalues of the matrix A correspond to the poles of $k(z)$ and this follows from $(I_n - zA)^{-1} = \det(I_n - zA)^{-1}\tilde{A}(z)$, with $\tilde{A}(z)$ denoting the matrix of cofactors. Thus, for integration to occur, some of the eigenvalues of A must be equal to one. Consider e.g. the case $A = I_n$. Then $x_{t+1} = x_t + K\varepsilon_t$ shows that x_t is integrated of order 1 and in this case the state is a vector random walk. Thus, for this example $x_t = K\sum_{j=0}^{t-1}\varepsilon_j + x_0$. Since $y_t = Cx_t + E\varepsilon_t$ it follows that $y_t = CK\sum_{j=0}^{t-1}\varepsilon_j + CE\varepsilon_t + Cx_0$, with $C \in \mathbb{R}^{s\times n}$ and $K \in \mathbb{R}^{n\times s}$. The number of common trends in y_t is thus given by the rank of the matrix $CK \in \mathbb{R}^{s\times s}$. This rank is at most s, which reflects the fact that at most s common trends can be present for y_t (see also the discussion below).

Seasonal unit roots analogously correspond to complex eigenvalues of the matrix A with modulus 1. A minimal system (1) thus generates output that is integrated if all the eigenvalues of A are inside the open unit disc or at one. The integration order of y_t is determined by the structure of the eigenvalues at one (see Bauer and Wagner (2001), for a detailed discussion). The integration order is equal to 1 if the algebraic multiplicity of the eigenvalue one equals its geometric multiplicity.

As a final assumption we restrict attention to systems that are *strictly minimum phase*, i.e. to systems where in the state space representation all eigenvalues of the matrix $(A - KE^{-1}C)$ have absolute value smaller than one. This corresponds to the assumption of all zeros of $k(z) = \pi(A, K, C, E)$ being outside the closed unit disc. Denote by M_n the set of all transfer functions $k(z)$, which fulfill the above conditions on the poles and the zeros and where the minimal state dimension of a state space representation of $k(z)$ is equal to n. Note that in M_n, stationary systems and integrated systems with different cointegrating ranks are included.

Since the representation of a transfer function $k(z) \in M_n$ in state space form is not unique, further restrictions have to be imposed on the matrices (A, K, C, E) in order to achieve uniqueness. This is achieved by the definition or construction of a

canonical form. There are many ways of imposing the required restrictions to ensure uniqueness. One way is to choose the controllability matrix $\mathscr{C}$ equal to the first n linearly independent rows of $\mathscr{H}$. The row indices describing these rows can be represented by a multi-index called Kronecker index. The Kronecker index is unique to each transfer function in M_n. The subset of transfer functions $k(z) \in M_n$, where the first n rows of the corresponding Hankel matrix $\mathscr{H}$ are linearly independent, is called the *generic neighborhood* of the echelon canonical form. The name derives from the fact that this set is generic in M_n and allows for a continuous parametrization, i.e. a homeomorphic mapping attaching parameter vectors $\tau \in \mathbb{R}^{2ns+s(s+1)/2}$ to transfer functions $k(z) \in M_n$ for appropriately defined topologies. The advantage of the echelon canonical forms lies in the fact that the parameter values occurring in the echelon state space representation are closely linked to the parameters occurring in the corresponding echelon VARMA representation of the system (for the exact relation see Hannan and Deistler, 1988, Section 2.6). In particular, there exists a homeomorphic bijection between these two different sets of parameter vectors. Also the sets of transfer functions, which can be parametrized continuously, are identical, so that the user is completely free to choose the setup she is more familiar with. In particular, any estimated state space system can be identified with the corresponding echelon VARMA system, and consistency results derived for echelon state space systems also hold for the echelon VARMA parameters (on generic subsets of M_n, to be precise).

A companion paper, Bauer and Wagner (2001), develops a different canonical form for state space systems of form (1) containing an arbitrary number of unit roots located at any point on the unit circle. This canonical form reveals the relationship between the integration orders (corresponding to the different unit roots) and the structure of the corresponding eigenvalues of A in a minimal representation. In this paper we are going to draw from these results, and the canonical form, on which the results derived in this paper are based, is a special case and is of the following form:

$$A = \begin{bmatrix} I_c & 0 \\ 0 & A_{\mathrm{st}} \end{bmatrix}, \quad K = \begin{bmatrix} K_1 \\ K_{\mathrm{st}} \end{bmatrix}, \quad C = [C_1 \quad C_{\mathrm{st}}].$$

Here c denotes the number of common trends in the minimal state x_t and $(A_{\mathrm{st}}, K_{\mathrm{st}}, C_{\mathrm{st}})$ denotes a state space realization of the stationary subsystem. Note that there can be no more than s common trends in a minimal state x_t, which is seen as follows: Due to the structure of the canonical form in the present case, the observability matrix $\mathscr{O}$ takes the form

$$\mathscr{O} = \begin{bmatrix} C_1 & C_{\mathrm{st}} \\ C_1 & C_{\mathrm{st}} A_{\mathrm{st}} \\ \vdots & \vdots \end{bmatrix}$$

and thus the first block column has rank equal to the rank of C_1, which is less or equal to s. This also shows that for a minimal representation, C_1 is of full column rank c. Analogous arguments show that in a minimal representation, the row rank of K_1 is equal to c. Thus, for minimal systems $c \leqslant s$ denotes the number of common trends present in both y_t and x_t, irrespective of the system order n. From the structure of the

state space representation it follows that

$$y_t = C_1 K_1 \sum_{j=0}^{t-1} \varepsilon_j + k_{\mathrm{st}}(z)\varepsilon_t, \tag{4}$$

where $k_{\mathrm{st}}(z) = E + zC_{\mathrm{st}}(I_{n-c} - zA_{\mathrm{st}})^{-1}K_{\mathrm{st}}$, assuming zero initial conditions for the nonstationary part of the state. This immediately shows that if r denotes the number of linearly independent cointegrating relations for y_t, the equation $c = s - r$ holds.

The representation given above is not unique. The set of observationally equivalent state space systems, which also have a block-diagonal A matrix, where the $(1,1)$ block is equal to the identity matrix, is characterized by $S = \mathrm{diag}(S_1, S_{\mathrm{st}})$, where both matrices $S_1 \in \mathbb{R}^{c\times c}$ and $S_{\mathrm{st}} \in \mathbb{R}^{(n-c)\times(n-c)}$ are nonsingular. Thus, further restrictions have to be imposed in order to reduce the set of observationally equivalent systems obeying all restrictions to a singleton. In other words, to achieve identification of the parameters, a unique representative has to be selected of the set of (observationally equivalent minimal) state space systems that represent the transfer function $k(z)$.

In the canonical form presented in Bauer and Wagner (2001) C_1 is chosen to be part of an orthonormal matrix, i.e. $C_1 \in \mathbb{R}^{s\times c}, C_1'C_1 = I_c$ is assumed.[3] Therefore there exists a matrix $C_1^{\perp}$ with $(C_1^{\perp})'C_1^{\perp} = I_r$ and $(C_1^{\perp})'C_1 = 0$, i.e. $C_1^{\perp}$ spans the orthogonal complement of C_1. Let $\bar{C}' = [C_1, C_1^{\perp}]$. Since all the eigenvalues of A_{st} are, by construction, restricted to be inside the unit circle, it is easily seen that $k_{\mathrm{st}}(z)$ is analytic in the closed unit disc. Note that the representation given in Eq. (4) coincides with that of Granger. It is immediate that the first component in (4) corresponds to the common trends and that the columns of $C_1^{\perp}$ span the space of the cointegrating relations. Therefore, the cointegrating rank is equal to r. The number of common trends is equal to the number of eigenvalues of A at one, denoted with c, and $c = s - r$ holds. In the case of higher integration orders, the relationship between the eigenvalues and the integration orders (at the different frequencies, i.e. corresponding to the different unit roots) is more complicated. Still however, the eigenvalue structure (for details see Bauer and Wagner, 2001) determines the integration orders and the numbers of components with different integration orders.

3. Subspace algorithms

Subspace algorithms originated in the engineering literature in the 1980s. They provide an alternative to classical (pseudo) maximum likelihood estimation of linear time-invariant systems, like e.g. VARMA systems. In the meantime, a variety of algorithms is available, e.g. `CCA` (Larimore, 1983), `N4SID` (Van Overschee and DeMoor, 1994) or `MOESP` (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.

[3] These restrictions are not sufficient for identifiability in the general case and some further restrictions are needed. However, these restrictions are not important for the present setting and thus we refer to Bauer and Wagner (2001) for details.

The main idea of this algorithm lies in the interpretation of the state: Consider the problem of predicting $y_{t+j}, j \geqslant 0$ from its finite past up to time $t-1$, i.e. from $y_{t-1}, y_{t-2}, \ldots, y_0$ and x_0.[4] From system equations (1) it follows that

$$y_{t+j} = CA^j x_t + \sum_{i=0}^{j-1} CA^i K \varepsilon_{t+j-i-1} + E\varepsilon_{t+j}. \tag{5}$$

Now, since

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

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 $y_{t-1}, \ldots, y_0, x_0$. Thus, the state x_t is a basis for the predictor space for the whole future of y_t, i.e. for $y_{t+j}, j \geqslant 0$, and is contained in the past of the time series. For notational brevity let $\bar{A} = (A - KE^{-1}C)$. Then, after substituting for x_t the above expression giving x_t as a function of x_0 and past y_t's, we can re-write Eq. (5) in stacked matrix format for all $j \geqslant 0$ as

$$\begin{aligned} \begin{pmatrix} y_t \\ y_{t+1} \\ \vdots \\ y_{t+j} \\ \vdots \end{pmatrix} &= \begin{pmatrix} CKE^{-1} & C\bar{A}KE^{-1} & \cdots & C\bar{A}^{t-1}KE^{-1} \\ CAKE^{-1} & CA\bar{A}KE^{-1} & \cdots & CA\bar{A}^{t-1}KE^{-1} \\ \vdots & \vdots & \vdots & \vdots \\ CA^jKE^{-1} & CA^j\bar{A}KE^{-1} & \cdots & CA^j\bar{A}^{t-1}KE^{-1} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_0 \end{pmatrix} \\ &\quad + \begin{pmatrix} C \\ CA \\ \vdots \\ CA^j \\ \vdots \end{pmatrix} \bar{A}^t x_0 + \begin{pmatrix} E & 0 & 0 & \cdots & 0 & \cdots \\ CK & E & 0 & \cdots & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ CA^{j-1}K & CA^{j-2}K & \cdots & CK & E & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \varepsilon_t \\ \varepsilon_{t+1} \\ \vdots \\ \varepsilon_{t+l} \\ \vdots \end{pmatrix} \\ &= \beta \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_0 \end{pmatrix} + \begin{pmatrix} C \\ CA \\ \vdots \\ CA^j \\ \vdots \end{pmatrix} \bar{A}^t x_0 + \mathscr{E} \begin{pmatrix} \varepsilon_t \\ \varepsilon_{t+1} \\ \vdots \\ \varepsilon_{t+l} \\ \vdots \end{pmatrix}. \end{aligned}$$

[4] In the case that x_0 is not known, x_0 is estimated using the Kalman filter and the resulting estimate is used for the prediction. This, however, does not change the asymptotic properties.

The above equation describes the future of the process, $y_{t+j}, j \geqslant 0$, as the sum of three components: The first term is due to the (finite) past of the process $y_{t-1}, \ldots, y_0$, the second term describes the impact of the initial state x_0 and the third term shows the impact of the future of the noise process $\varepsilon_{t+j}, j \geqslant 0$. The latter term is orthogonal to the former two. Note that for $t \to \infty$ the second term vanishes, since due to the strict minimum-phase assumption the matrix $\bar{A}^t = (A - KE^{-1}C)^t$ converges to zero.

The matrix β in the above equation connecting the past of the output y_t (i.e. terms for $s < t$) to the future values ($s \geqslant t$) carries structural information about the system, i.e. it contains relevant information about the system matrices (A, K, C, E). The idea of subspace algorithms is to use this information to obtain estimates of the system matrices.

For estimation only a finite set of observations is available, hence the above equation is utilized in a truncated version. Note that the matrix β in the equation above has rank equal to n, the system order. Thus, choose two indices f and p, both larger or equal to n, and define

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

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

and

$$E_{t,f}^{+} = [\varepsilon_t', \varepsilon_{t+1}', \ldots, \varepsilon_{t+f-1}']'.$$

Furthermore, let

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

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

and let $\mathcal{E}_f$ denote the matrix with the ith block row $[CA^{i-2}K, \ldots, CK, E, 0]$ for $i \geqslant 2$ and $[E, 0, \ldots, 0]$ as its first block row. As a second change to the above equation, the state p-periods ahead is employed, since only a finite past $y_{t-1}, \ldots, y_{t-p}$ is used. With this notation the truncated equation can compactly be written as

$$Y_{t,f}^{+} = \mathcal{O}_f \mathcal{K}_p Y_{t,p}^{-} + \mathcal{O}_f (A - KE^{-1}C)^p x_{t-p} + \mathcal{E}_f E_{t,f}^{+}. \tag{6}$$

The above observations lead to the following procedure:

(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 $\hat{\beta}_{f,p}$ is of full rank, whereas $\mathcal{O}_f \mathcal{K}_p$ is of rank n for $f, p \geqslant n$, where n denotes the true order and f and p are user-chosen integers. Thus, approximate $\hat{\beta}_{f,p}$ by a rank n matrix with decomposition $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$; see below for details on the approximation.

(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 (1) 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 $\hat{C}$ and residuals $\tilde{\varepsilon}_t$. Then $\hat{\Omega} = (1/T) \sum_{t=1}^{T} \tilde{\varepsilon}_t \tilde{\varepsilon}_t'$ is an estimate for the innovation variance. Thus, $\hat{E}$ can be calculated as the lower triangular Cholesky factor of $\hat{\Omega}$ and $\hat{\varepsilon}_t = \hat{E}^{-1} \tilde{\varepsilon}_t$. Finally regress $\hat{x}_{t+1}$ on $\hat{x}_t$ and $\hat{\varepsilon}_t$ to obtain estimates $\hat{A}$ and $\hat{K}$ respectively.

The approximation performed in step 2 is not performed on $\hat{\beta}_{f,p}$ directly, but on a weighted matrix $\hat{W}_f^+\hat{\beta}_{f,p}\hat{W}_p^-$. The mentioned subspace algorithms differ i.a. in their choices of weighting matrices. Let $\hat{\Gamma}_f^+ = \sum_{t=1}^{\mathrm{T}} Y_{t,f}^+(Y_{t,f}^+)'$ and $\hat{\Gamma}_p^- = \sum_{t=1}^{\mathrm{T}} Y_{t,p}^-(Y_{t,p}^-)'$ denote the (noncentered and unnormalized) sample covariances of $Y_{t,f}^+$ and $Y_{t,p}^-$. Here $y_t = 0$ for $t < 0$ and $t > T$ is used.[5] Then in the algorithm `CCA`, $\hat{W}_f^+ = (\hat{\Gamma}_f^+)^{-1/2}$ and $\hat{W}_p^- = (\hat{\Gamma}_p^-)^{1/2}$ respectively.[6] For `MOESP` and `N4SID`, $\hat{W}_f^+ = I$ and $\hat{W}_p^-$ is as for `CCA`. The `MOESP` type algorithms are differing from e.g. `CCA` algorithms in that they are uncovering the system matrix estimates based on $\hat{\mathcal{O}}_f$, whereas `CCA` is exploiting the structure of $\hat{\mathcal{K}}_p$, since the state is estimated as $\hat{x}_t = \hat{\mathcal{K}}_p Y_{t,p}^-$. `MOESP` algorithms are considered to be more suitable for the estimation of systems containing exogenous variables (see e.g. Bauer, 1998). The choice of weighting matrices in `CCA` also explains the name, *canonical correlation analysis*: The algorithm amounts to an estimation of the canonical correlations between $Y_{t,f}^+$ and $Y_{t,p}^-$.

Bauer (2002) shows that for stationary systems, the `CCA` algorithm results in estimates that have the same asymptotic properties as pseudo-maximum likelihood estimates. Thus, we focus attention on the weighting matrices as specified in the `CCA` algorithm, which however will have to be modified for integrated processes to ensure consistency also then.

Let $\hat{W}_f^+\hat{\beta}_{f,p}\hat{W}_p^- = \hat{U}\hat{\Sigma}\hat{V}'$ be the singular value decomposition where $\hat{U}$ contains the left singular vectors, $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_{\min(f,p)s})$ contains the singular values ordered decreasing in size and $\hat{V}$ contains the right singular vectors. For a system of order n, exactly n singular values are larger than zero. Of course, only estimates $\hat{\sigma}_1 \geqslant \cdots \geqslant \hat{\sigma}_n \geqslant \hat{\sigma}_{n+1} \geqslant \cdots \geqslant \hat{\sigma}_{\min(f,p)s}$ are available. The estimated singular values $\hat{\sigma}_{n+1}, \ldots, \hat{\sigma}_{\min(f,p)s}$ will be nonzero due to small sample and noise effects. Asymptotically $\hat{\sigma}_{n+1}, \ldots, \hat{\sigma}_{\min(f,p)s}$ converge to zero. Now, paralleling the stationary case (Bauer, 1998), the order estimation is based on considering the size of the first neglected singular value, $\hat{\sigma}_{n+1}$, exploiting the asymptotic behavior of the estimates. Define the following criterion:

$$SVC(n) = \hat{\sigma}_{n+1}^2 + 2nsH_T/T. \tag{7}$$

Here $H_T > 0, H_T/T \to 0$ denotes a penalty term, which determines the asymptotic properties of the estimated order. The number of parameters in a model with state dimension n is equal to $2ns$ (for the generic neighborhood, to be precise), excluding the parameters in E, see e.g. Hannan and Deistler (1988, Theorem 2.6.3). The order estimate, $\hat{n}$ say, is then given by the minimizing argument of the criterion function $SVC(n)$. It will later be shown that this procedure is consistent, for suitable choices of H_T, also for integrated processes (see Theorem 3 in Section 5).

[5] Alternatively, the summation can be limited to the range $p + 1 \leqslant t \leqslant T - f$, this does not change the asymptotic results of this paper. However, it may well influence the finite sample properties.

[6] $X^{1/2}$ denotes the Cholesky factor of the positive definite matrix X such that $X^{1/2}(X^{1/2})' = X$.

Thus, for the specified rank n, where e.g. $n = \hat{n}$, decompose the SVD in two parts:

$$\hat{W}_f^+ \hat{\beta}_{f,p} \hat{W}_p^- = \hat{U}\hat{\Sigma}\hat{V}' = \hat{U}_n \hat{\Sigma}_n \hat{V}_n' + \hat{R},$$

where $\hat{U}_n \in \mathbb{R}^{fs \times n}$, $\hat{V}_n \in \mathbb{R}^{ps \times n}$ and $\hat{\Sigma}_n \in \mathbb{R}^{n \times n}$. Here $\hat{\Sigma}_n = \mathrm{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_n)$ contains the n dominant singular values ordered decreasing in size, i.e. $1 \geqslant \hat{\sigma}_1 \geqslant \cdots \geqslant \hat{\sigma}_n > 0$. The matrices $\hat{U}_n$ and $\hat{V}_n$ contain the corresponding left and right singular vectors. The remaining singular values and vectors are attributed to $\hat{R}$ and are neglected. The rank n approximation to $\hat{\beta}_{f,p}$ is now 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 $\hat{\mathcal{K}}_p = \hat{\Sigma}_n \hat{V}_n' (\hat{W}_p^-)^{-1}$. Observe, that $\hat{\mathcal{O}}_f$ and $\hat{\mathcal{K}}_p$ will, in general, not have the structure of their population counterparts $\mathcal{O}_f$ and $\mathcal{K}_p$, e.g. there will exist no matrix $\hat{A}$, such that $[I_{(f-1)s}, 0^{(f-1)s,s}]\hat{\mathcal{O}}_f \hat{A} = [0^{f(s-1)s,s}, I_{(f-1)s}]\hat{\mathcal{O}}_f$.

For integrated processes we are going to modify the above procedure for the estimation of $\hat{\mathcal{K}}_p$ and therefore $\hat{x}_t$. Let, as before, the true cointegrating rank be denoted by r, then $c = s - r$ common trends drive the system. Assume that a consistent estimate $\hat{C}_1$ of C_1, as defined in the canonical form, is given. There are several ways known to obtain such an estimate, and below it is shown that the subspace algorithm in its standard form can be used to obtain the required estimate. A computationally simple way to obtain an estimate of C_1 is to regress y_t on the first c components of the state estimated by the standard `CCA` algorithm. The subspace estimates of the cointegrating space achieve super-consistency, i.e. $T^{\gamma}(\hat{C}_1 - C_1) \to 0$ for $0 < \gamma < 1$. Now denote (paralleling the discussion in Section 2) with $\hat{\tilde{C}} = [\hat{C}_1, \hat{C}_1^{\perp}]'$, where $\hat{C}_1^{\perp} \in \mathbb{R}^{s \times r}$, $\hat{C}_1' \hat{C}_1^{\perp} = 0$ and $(\hat{C}_1^{\perp})' \hat{C}_1^{\perp} = I_r$. Thus, the columns of $\hat{C}_1^{\perp}$ span the estimated cointegrating space. Now define a new weighting matrix $\hat{W}_{f,C_1}^+ = [(I \otimes \hat{\tilde{C}}) \sum_{t=1}^{T} Y_{t,f}^+ (Y_{t,f}^+)' (I \otimes \hat{\tilde{C}})']^{-1/2} (I \otimes \hat{\tilde{C}})$, using the Cholesky decomposition as the square root. In combination with the modified weighting matrix also the estimate for $\hat{\mathcal{K}}_p$ has to be modified: For any choice of weighting matrices, the estimated matrix $\hat{\mathcal{K}}_p = \hat{\Sigma}_n \hat{V}_n' (\hat{W}_p^-)^{-1}$ can alternatively be written as $\hat{\mathcal{K}}_p = \hat{U}_n' \hat{W}_f^+ \hat{\beta}_{f,p}$. Now, if the modified weighting matrix $\hat{W}_{f,C_1}^+$ is used, the corresponding matrix of left singular vectors $\hat{U}_n$ has to be changed to $\hat{U}_{n,c}$, where

$$\hat{U}_{n,c} = \begin{bmatrix} I_c & 0^{c \times (n-c)} \\ 0^{(fs-c) \times c} & \hat{U}(2,2) \end{bmatrix}.$$

$\hat{U}(2,2)$ denotes the $(2,2)$ block of $\hat{U}_n$. This modification is motivated by the fact that (as shown in Appendix A) $\hat{U}_n \to U_0$, with

$$U_0 = \begin{bmatrix} I_c & 0^{c \times (n-c)} \\ 0^{(fs-c) \times c} & U_0(2,2) \end{bmatrix}.$$

Thus, under the assumption of a correctly specified number of common trends, c, and the availability of a consistent estimate of the common trends space, the subspace procedure can be modified as follows: Use the corresponding modified weighting matrix $\hat{W}_{f,C_1}^+$ and the modified matrix $\hat{U}_{n,c}$ to obtain an adapted estimate of $\hat{\mathcal{K}}_p$, which is given by $\hat{U}_{n,c}' \hat{W}_{f,C_1}^+ \hat{\beta}_{f,p}$. The replacement of $\hat{U}_n$ by $\hat{U}_{n,c}$ changes the asymptotic properties of the estimates and guarantees that also the estimates corresponding to the stationary part

of the transfer function are consistent (see Theorem 2). The approach just described above is called *adapted procedure* throughout the paper.

For stationary processes, i.e. when $r = s$ and thus $c = 0$, the adapted procedure coincides with the standard `CCA` procedure.

4. Main results

Let $M_n^{\mathrm{st}} \subset M_n$ denote the set of all transfer functions $k(z) \in M_n$ without poles on the unit circle, i.e. which describe stationary systems. Further denote with $M_n^{\mathrm{st},+}$ the set of all transfer functions $k(z) \in M_n^{\mathrm{st}}$, such that the a.s. limit for $T \to \infty$ and $p = p(T) \to \infty$ (whose existence is guaranteed under the assumptions of Theorem 1) $W_f^+ \beta W_\infty^-$ of $\hat{W}_f^+ \hat{\beta}_{f,p} \hat{W}_p^-$ has n distinct nonzero singular values. Bauer et al. (1999) show that $M_n^{\mathrm{st},+}$ is a generic subset of M_n^{st}. The following results that clarify the asymptotic properties in the stationary case can be found in Bauer et al. (1999) and Bauer (1998, 2002).

Theorem 1. *Let y_t be generated by a system of form (1), where the white noise ε_t fulfills the standard assumptions and where $k(z) \in M_n^{\mathrm{st}}$ denotes the true transfer function of order n. If $f = f(T) \geqslant n$ is a user-supplied integer and $p(T) \geqslant -(d/2)(\log T/\log|\rho_0|)$, where ρ_0 is an eigenvalue of $A - KE^{-1}C$ of maximum modulus and $d > 1$ is some real value, and if $\max\{f(T), p(T)\} = o((\log T)^a)$ for some $a > 0$, then*

- *For $k(z) \in M_n^{\mathrm{st}}$ the estimate of the transfer function is a.s. consistent, i.e. $\hat{k}(z) \to k(z)$ a.s., where convergence is in the pointwise topology (see e.g. Hannan and Deistler, 1988).*
- *For $k(z) \in M_n^{\mathrm{st},+}$ the estimate of the system matrices is a.s. consistent, i.e. there exists a realization (A, K, C, E) of the true transfer function $k(z) \in M_n^{\mathrm{st},+}$, such that $\|vec[\hat{A} - A, \hat{K} - K, \hat{C} - C, \hat{E} - E]\| \to 0$ a.s. Here vec denotes the operator stacking the vectorizations of the various matrices.*
- *For $k \in M_n^{\mathrm{st},+}$ a central limit theorem for the system matrix estimates holds, i.e.*

$$\sqrt{T}[vec(\hat{A} - A, \hat{K} - K, \hat{C} - C, \hat{E} - E)] \xrightarrow{\mathrm{d}} \mathcal{N}(0, V),$$

  where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution and $\mathcal{N}(0, V)$ is a Gaussian random variable with zero mean and variance V.
- *The choice $\hat{W}_f^+ = (\hat{\Gamma}_f^+)^{-1/2}$ and $\hat{W}_p^- = (\hat{\Gamma}_p^-)^{1/2}$ and $f = p \to \infty$ according to the restrictions imposed above implements a generalized pseudo-maximum likelihood procedure and thus in the Gaussian case achieves optimal asymptotic variance.*
- *If $H_T/(f(T)p(T)\log\log T) \to \infty$ and $H_T/T \to 0$, the order estimated using SVC is a.s. consistent.*

The usual choices concerning f are $f(T) = f$ constant or $f(T) = p(T)$. For $p(T)$ often $2\hat{p}_{\mathrm{AIC}}$ is used, where $\hat{p}_{\mathrm{AIC}}$ denotes the order estimate obtained in an autoregressive approximation of the system. The above results, that clarify the asymptotic

properties of the algorithm for the stationary case, also motivate the specific choice of the weighting matrices and therefore the specific choice of the algorithm in this paper. Expressions for the variance V can be given, see e.g. Bauer and Ljung (2002). The expression given there shows that indeed CCA is the optimal choice of the weighting sequence for each fixed $f \geqslant n$.

The algorithm as it is described above, only leads to a system that is close to a cointegrated system, a precise meaning of this statement is contained in the formulation of Theorem 2. The estimation problem can however also easily be reformulated to result in an estimated system that corresponds to an exactly cointegrated system. A reduced rank regression approach delivers the required result: Remember the third step in the description of subspace algorithms, where, after $\hat{x}_t$, $\hat{\varepsilon}_t$, $\hat{C}$ and $\hat{E}$ have been estimated, $\hat{x}_{t+1}$ is regressed on $\hat{x}_t$ and $\hat{\varepsilon}_t$ to obtain estimates of $\hat{A}$ and $\hat{K}$ (i.e. the equation $x_{t+1} = Ax_t + K\varepsilon_t$ is estimated). Now, if the cointegrating rank of y_t is r, the rank of the matrix $A - I_n$ equals $n - c$ which due to minimality furthermore is equal to $n - s + r$. Thus, alternative estimates $\tilde{A}$ and $\tilde{K}$ of A and K can be obtained from a reduced rank regression $\hat{x}_{t+1} - \hat{x}_t = (\tilde{A} - I_n)\hat{x}_t + \tilde{K}\hat{\varepsilon}_t$ under the constraint that $rank(\tilde{A} - I_n) = n - c$. This approach results by construction in an estimated system that is exactly cointegrated with cointegrating rank r. In order to separate the two approaches notationally, the latter approach is referred to as *reduced rank regression approach* and the least-squares method for obtaining estimates of A and K is called *unrestricted regression approach*.

The first main result of this paper is now concerned with the properties of the described algorithm (and its adaptation) for integrated processes of order 1. The proof of the theorem is given in Appendix A.

Theorem 2. *Let the s-dimensional output y_t be generated by a system of form* (1) *with the ergodic noise ε_t fulfilling the standard assumptions. Assume that the true order n of the transfer function $k(z)$ is known. Concerning the indices f and p the following assumptions are made*: $p = p(T) = o((\log T)^a)$ *for some* $0 < a < \infty$, $p \geqslant -d \log T / \log|\rho_0|$, *where ρ_0 is an eigenvalue of $A - KE^{-1}C$ of maximum modulus*, $d > 1$ *and* $f \geqslant n$ *is fixed.*

Given the true cointegrating rank r, the standard CCA *subspace algorithm delivers consistent estimates of order T of the cointegrating space as follows*: *Denote by C_1 the first c columns of C. Then consistent estimates $\hat{C}_1$ of C_1, with $T^\gamma \|\hat{C}_1 - C_1\| \to 0$ in probability for $0 < \gamma < 1$, are obtained by an OLS regression of y_t on the first c components of the state estimated by the standard* CCA *procedure, say $\hat{x}_{t,1}$. As it is $C_1^\perp$ that is spanning the cointegrating space, note at this point that $T^\gamma \|\hat{C}_1 - C_1\| \to 0$ implies that also $T^\gamma \|\hat{C}_1^\perp - C_1^\perp\| \to 0$.*

Assume again that r is correctly specified and that the adapted subspace procedure as described above is used with an estimate $\hat{C}_1$ that is consistent of order T. Then the estimate $\hat{k}(z) = \hat{E} + z\hat{C}(I - z\hat{A})^{-1}\hat{K}$ converges in probability to the true transfer function $k(z)$ for the unrestricted regression approach.

The same result concerning the consistency of the transfer function estimate holds also if the reduced rank regression approach is used, where the estimate $\tilde{A}$ of A is constructed to fulfill the rank restriction $rank(\tilde{A} - I_n) = n - c$.

The consistency result in the theorem is stated in terms of the transfer functions. To reformulate the consistency result in terms of the system matrices (A, K, C, E), it is required to transform the estimates to a canonical form (on generic pieces of M_n). The next question, after having established consistency of the estimates, refers to the asymptotic distribution of the estimates. This is left as a topic of future research.

A difference between the algorithm for the stationary case and the integrated case is the different choice of the weighting matrices, $\hat{W}_f^+$ or $\hat{W}_{f,C_1}^+$ respectively. It follows from the above theorem that the former weighting matrix may be used in the estimation of the cointegrating space and this estimate may then be employed in the construction of the weighting matrix $\hat{W}_{f,C_1}^+$. Consistent estimates of the parameters corresponding to the stationary part of the transfer function are derived by using the adapted version of the subspace algorithm.

A second difference between the results for stationary and integrated processes is the stronger restriction for the increase of p as a function of the sample size for the case of integrated processes. This stronger restriction is introduced to guarantee that $(A-KE^{-1}C)^p = o(T^{-1})$ rather than only $(A-KE^{-1}C)^p = o(T^{-1/2})$, which is sufficient in the stationary case.

Related to the consistent estimation of the cointegrating space, also tests for its dimension are developed in the following section, where two different tests are provided.

5. Estimating the structure indices

In this section we discuss the determination of the structure indices, i.e. of the system order n and of the cointegrating rank r, or equivalently of the number of common trends c. These problems are tackled by employing the properties of the singular values of $\hat{W}_f^+ \hat{\beta}_{f,p} \hat{W}_p^-$ and of the eigenvalues of $\hat{A}$. It has been seen in Section 3 that the system order is equal to the number of nonzero singular values of the limit of $\hat{W}_f^+ \hat{\beta}_{f,p} \hat{W}_p^-$. This property of the singular values $\hat{\sigma}_{n+j}$ for $j > 0$ to converge to zero has been exploited for stationary systems in the criterion $SVC(n)$. The following theorem shows that this order estimation procedure is consistent also for integrated processes.

Theorem 3. *Under the conditions of Theorem* 2 *the estimate of the order obtained by SVC as defined in* (7) *is weakly consistent for* $H_T/(p(T)\log\log T) \to \infty$ *and* $H_T/T \to 0$, *i.e.* $\hat{n} \to n$ *in probability*.

The proof of Theorem 3 is given in Appendix A.

Concerning the estimation of (or testing for) the cointegrating rank we propose two tests. One based on $\hat{\sigma}_i$, the estimated singular values. This test procedure will turn out to be related to the Bewley and Yang (1995) approach. The other test is based on the eigenvalues of $\hat{A}$. The eigenvalue-based test is in the spirit of Stock and Watson (1988) and employs the fact that the number of eigenvalues of the matrix A equal to one equals the number of common trends. Let us analyze each test in turn and start with the singular value based test.

Our singular value based test for the number of common trends is based on correlations between $Y_{t,f}^{+}$ and $Y_{t,p}^{-}$. Thus, there are similarities to other tests that are based on canonical correlations between the observations. In particular, the test proposed by Poskitt (2000) turns out to use a special case of our framework, i.e. when $f = p = 1$ are chosen, as Poskitt calculates the correlations between y_t and y_{t-1}. The difference thus is that our test takes the short-run dynamics into account. Furthermore, in our proof we derive sharp bounds on the estimation error leading to the distribution of the test statistic, whereas Poskitt (2000) only derives an upper bound. This comes at the expense of not proving strong consistency, but only in probability statements. It should be noted, however, that Poskitt (2000) uses a slightly different test statistic, which could easily be adapted to the present case. For the special case of AR(1) processes and for $f = p = 1$ and for the null hypothesis of no cointegration, the test of Bewley and Yang (1995) coincides with the test statistic presented in Theorem 4.[7]

Let the process y_t be generated by a minimal system of form (1) with order n and true cointegrating rank r. Then (asymptotically) exactly c (where again $c = s - r$) estimated singular values are equal to one, whereas the remaining $n-c$ nonzero singular values are converging to their limits smaller than one. This relationship between the number of singular values equal to one and the number of common trends only holds true if no zeros of the transfer function are admitted on the unit circle[8] and the only poles of the transfer function on the unit circle, i.e. the only unit roots, occur at $z = 1$. Any of these other cases also introduces unit singular values and therefore the following test is not robust against the presence of e.g. seasonal unit roots. Consistency of the estimated singular values has been established in the proof of Theorem 2. More precisely, there it is shown that given a number of common trends c, the largest c estimated singular values converge to one at rate T, whereas the remaining $n - c$ converge to their limits only at rate $T^{1/2}$. Thus, a test for the cointegrating rank, $r = s - c$, or more directly for the number of common trends c, may be based on the asymptotic distribution of the first c estimated singular values.

Theorem 4. *Let the process y_t be generated by a system of form* (1), *where the true noise satisfies the standard assumptions. Let $\hat{\sigma}_i$ denote the estimate of the ith singular value* (*which are assumed to be ordered decreasing in size*) *and let c denote the true number of common trends. Then $T(1 - 1/c \sum_{j=1}^{c} \hat{\sigma}_j^2)$ converges in distribution to*

$$\frac{1}{c} \operatorname{tr} \left[C_1' \Omega C_1 \left(\int_0^1 W(u) W(u)' \, \mathrm{d}u \right)^{-1} \right]. \tag{8}$$

Here $\int_0^1 W(u)W(u)' \, \mathrm{d}u$ denotes a mixture of Brownian motions, where the covariance associated with $W(u)$ is equal to $K_1 K_1'$. $\Omega = EE'$ denotes the innovation covariance matrix.

The proof of Theorem 4 is given in Appendix A.

[7] For the general situation, i.e. for higher order processes or for values of f and p larger than one, there seems to be no connection.

[8] Thus, we exclude in terms of a left coprime VARMA representation unit roots in the MA polynomial.

Thus, for all values of c, the system can be estimated by the adapted CCA procedure to obtain estimates of C_1, K_1 and Ω and these estimates could then be inserted in the test statistic. However, this already shows the main disadvantage of this test, its dependence on nuisance parameters. This drawback could in principle be overcome or at least mitigated by bootstrapping the test statistic to decrease the finite sample effects (see e.g. Bauer and Wagner (2000a), for a first application of the bootstrapped version of this test).

Another idea is to follow the arguments developed in Poskitt (2000) and to ignore the above distributional result and use only the implied in probability bounds in a procedure to estimate the cointegrating rank. Again exploit the fact that as many singular values as there are common trends converge to one at rate T. Hence, simply take as an estimate of the number of common trends the largest integer, say c, such that the cth singular value $\hat{\sigma}_c$ is the smallest one for which $1-\hat{\sigma}_c^2 < h_T/T$, where $h_T \to \infty$ and $h_T/T^{1/4} \to 0$ as $T \to \infty$. This leads to a weakly consistent estimation of c due to the results concerning the asymptotic distribution of the singular values. The specific choice of the threshold h_T influences the finite sample properties of the estimation procedure and also the asymptotic properties of the estimated cointegrating rank derived with this approach. It is common in the literature to choose the penalty h_T close to the lower bound of possible values for which a.s. consistency is obtained (compare e.g. the order estimation criterion BIC in the stationary case). Choices close to the lower bound then ensure even strong consistency for the procedure. E.g. for the special case $c=1$ in the present setup it can be shown that $h_T=(\log T)^2$ is a crude lower bound to achieve almost sure consistency (see the remark in Appendix A).

We advocate the use of the preceding result to obtain preliminary intuition concerning possible cointegrating vectors and to combine the results of this procedure either with the above test or with the nuisance parameter free test based on the eigenvalues of $\hat{A}$ presented below. Nuisance parameter free test statistics can be based on the eigenvalues of $\hat{A}$. This approach is very much in the spirit of Stock and Watson (1988): In that paper a test for the number of common trends in processes z_t having a representation of the form $\Pi(z)D'\Delta z_t=\varepsilon_t$ is derived. Here ε_t is white noise, $\Pi(z)\in\mathbb{R}^{c\times c}$ is a matrix polynomial and $D\in\mathbb{R}^{s\times c}$. A test for the number of common trends is based on an estimated first order autoregressive coefficient matrix of the integrated process $\Pi(z)D'z_t$. It is shown in Stock and Watson (1988) that the test statistic is asymptotically unchanged, whether $\Pi(z)$ and D are known or estimated consistently.

This setup fits to our problem very well, taking into account a few changes. We are not testing on the observations themselves, but on the state x_t corresponding to the canonical form presented in Section 2. And in our case it is the state that is unknown and of which only an estimate is available. The matrices $\Pi(z)$ and D themselves have a very simple form in our case, since the state transition equation is an autoregression of order one. Also in our canonical form the first c components of the state are the integrated ones, i.e. $[I_c,0^{c\times(n-c)}]\Delta x_t=K_1\varepsilon_{t-1}$, hence $\Pi(z)=I_c$ and $D=[I_c,0^{c\times(n-c)}]'$. Now, to derive a nuisance parameter free test for the number of common trends, we have to show that the replacement of the state with an estimate $\hat{x}_t=\hat{\mathscr{K}}_p Y_{t,p}^-$ does not change the asymptotic distribution of the test statistic. This is the content of the following theorem, whose proof is given in Appendix A.

Theorem 5. *Let the assumptions of Theorem* 2 *hold and let the true number of common trends be denoted as c. Assume that the adapted subspace procedure for cointegrated processes under the hypothesis of a correctly specified cointegrating rank is used.*

Then the asymptotic distribution of the largest c eigenvalues of $T(\sum_{t=0}^{T-1} \hat{x}_{t+1}, \hat{x}_t')(\sum_{t=0}^{T-1} \hat{x}_t, \hat{x}_t')^{-1} - I_n)$ *is equal to the distribution of the eigenvalues of* $\int_0^1 W(u)\,\mathrm{d}W(u)'(\int_0^1 W(u)W(u)'\mathrm{d}u)^{-1}$, *where* $W(u)$ *denotes the c-dimensional standard Brownian motion.*

Since $\hat{A} = (\sum_{t=0}^{T-1} \hat{x}_{t+1}, \hat{x}_t')(\sum_{t=0}^{T-1} \hat{x}_t, \hat{x}_t')^{-1}$, *this result directly leads to nuisance parameter free tests based on the eigenvalues of* $\hat{A}$.

Thus, given the above result, the eigenvalues of $\hat{A}$ or respectively of $T(\hat{A} - I_n)$ can be used to construct tests for the number of common trends (and therefore for the number of cointegrating relationships) in a straightforward way. Tests based on Theorem 5 have, as opposed to tests based on Theorem 4, the advantage of being nuisance parameter free.

Analogous to this result, it seems tempting to investigate also different schemes for obtaining the number of cointegrating relations: Theorem 5 basically shows that the replacement of the state x_t by an estimate $\hat{x}_t$ does not change the usual asymptotics. The question whether an analogous statement also holds for the Johansen approach applied to the state equation is currently under investigation.

There are, as in Stock and Watson (1988), several possibilities for constructing tests: For a given null hypothesis of c common trends, the test statistic can be based on the cth estimated eigenvalue alone, or on the c largest estimated eigenvalues of $T(\hat{A} - I_n)$. For the latter choice the sum of the c largest eigenvalues may be considered (compare e.g. the construction of the trace and the max tests in the Johansen framework). The second choice to be made is whether one bases the test statistic on the real parts of the eigenvalues or on the absolute values. Following again Stock and Watson we decide to use the real part of the eigenvalues, which can then be ordered according to decreasing size. Thus, in the simulations presented in the following section the test statistic is constructed from the real part of the cth largest eigenvalue (according to the size of the real part) of $T(\hat{A} - I_n)$.[9]

The test is one sided, with the alternative that the (real part of the) cth largest eigenvalue of $T(\hat{A} - I_n)$ is smaller than 0, as explosive systems are not of great concern. Thus, the test is performed by comparing the cth largest real part of the eigenvalues of $T(\hat{A} - I_n)$ with the real part of the cth largest eigenvalue of the functional of Brownian motions given in the formulation of the theorem. The asymptotic distributions of the (real parts of the ordered) eigenvalues have been simulated and critical values are available from the authors upon request.

The test sequence we propose is to start with an upper bound for the number of common trends as initial null hypothesis. For a chosen null hypothesis of c common trends, the matrix $\hat{A}$ is then estimated by applying the adapted subspace procedure

[9] In further simulations also the properties of other choices concerning the construction of test statistics have been investigated.

with the corresponding number of common trends. This in turn means nothing but the construction of the weighting matrix W^+_{f,C_1} based on the first c columns of an initial estimate of C. If the null hypothesis is rejected, the test sequence is continued with the null of $c-1$ common trends and is stopped when the corresponding null hypothesis cannot be rejected anymore. The sequence of course stops also after the rejection of the null hypothesis $c=1$.

An open question at this point is the determination of the initial null hypothesis. There are basically two possibilities: The first is to start the testing sequence with the maximal possible number of common trends, which is (remember the discussion in Section 2) given by the minimum of s, the dimension of the observed time series, and n, the system order. In the simulations this test sequence will be labeled as *eigenvalue test sequence*. The second possibility is to use a threshold-based estimate of the number of common trends as an initial guess (see the discussion below Theorem 4). To ensure that this approach works well, it is required to find a threshold that implies a low probability of underestimating the number of common trends. This is achieved by using large values for the penalty term h_T. The test procedure that starts with an initial guess derived from a threshold estimate of the number of common trends and uses the eigenvalue test sequence thereafter is called *combined test procedure* in the following section. In principle it is furthermore possible, as indicated before, to bootstrap the singular value based test statistic to obtain an initial guess for the number of common trends.

6. A small simulation study

In this section the theoretical results obtained in the previous sections are tested on simulated data. Two aspects of the presented methods are investigated: the order estimation and the performance of the proposed test sequences.

The systems we simulate have been previously investigated by Saikkonen and Luukkonen (1997). A precise description of the systems is given in Appendix B. All three systems generate three-dimensional outputs. The three scenarios include the cases of a two-dimensional cointegrating space (Scheme 1), of a one-dimensional cointegrating space (Scheme 2) and of an integrated system without cointegration (Scheme 3).

For each system 1000 time series of lengths $T=100$, 500 and 1000 have been generated using Gaussian white noise with covariance matrix as specified in Appendix B. We report the estimated order of the system and the cointegrating rank as determined by the different proposed testing procedures. The integers f and p are chosen to equal $2\hat{p}_{\mathrm{AIC}}$, where $\hat{p}_{\mathrm{AIC}}$ denotes the order estimate obtained by using AIC in an autoregressive approximation. It can be shown that under the assumptions in this paper, the probability that $\mathrm{d}\,\hat{p}_{\mathrm{AIC}} \geqslant -\log T/\log|\rho_0|$ tends to one for $d>2$. The system order is estimated by using the criterion $SVC(n)$, as described in the previous sections with penalty $H_T=\log T$. For all three systems the true system order is given by $n=3$.

In the right panel of Table 1 the estimated orders are reported. For all three systems for $T=500$ and 1000 the order is estimated correctly in more than 99% of the replications. For $T=100$ the order estimation turns out to be inaccurate, with a substantial bias toward underestimation of the system order. In the left panel of Table 1

Table 1
Frequency distributions of the test results for the dimensions of the cointegrating space and of the estimated system orders for Schemes 1 to 3 and sample sizes 100, 500 and 1000. This table displays the test results for the *eigenvalue test sequence*

Scheme	Sample size	Dim. of coint. space				System order			
		0	1	2	3	1	2	3	4
1	$T = 100$	0.082	0.106	0.705	0.107	0.511	0.288	0.201	0
	$T = 500$	0	0	0.962	0.038	0	0	0.996	0.004
	$T = 1000$	0	0	0.976	0.024	0	0	0.998	0.002
2	$T = 100$	0.141	0.486	0.351	0.022	0.335	0.378	0.287	0
	$T = 500$	0	0.918	0.082	0	0	0	0.995	0.005
	$T = 1000$	0	0.962	0.038	0	0	0	0.994	0.006
3	$T = 100$	0.625	0.061	0.31	0.004	0.306	0.056	0.638	0
	$T = 500$	0.961	0.035	0.004	0	0	0	0.995	0.005
	$T = 1000$	0.968	0.031	0.001	0	0	0	1	0

the distributions of the test results for the dimension of the cointegrating spaces using the *eigenvalue test sequence* are reported. For 100 observations the results are only satisfactory for Scheme 1, with a correct result in about 70% of the replications. For Schemes 2 and 3 the results are not so good for $T = 100$, with a correct decision in only about 49% and 63% of the replications respectively. This unsatisfactory behavior of the tests is also a consequence of the imprecise estimation of the system order for this small sample. It may be necessary to consider different values for H_T to improve the small sample performance of the order estimation criterion $SVC(n)$.[10] For the larger two sample sizes the performance of this test procedure is very good, and the nominal size corresponds to the actual size quite accurately. All tests in this section are performed at a nominal size of 5%.

In Table 2 the results for the *threshold estimate* (left panel) and the *combined test sequence* (right panel) are reported. The underlying penalty term is given by $h_T = (\log T)^2$. Some differences between the different proposed tests can be observed. For $T=100$ the eigenvalue test sequence is better than the other two, except for Scheme 2, where the threshold test sequence is slightly better than the combined test sequence and both of these are better than the eigenvalue test sequence. For this system also for the larger sample sizes the actual size of the eigenvalue test sequence is slightly lower than for the other two, where however all three tests do not have problems in detecting the correct number of cointegrating relationships.

The good performance of the threshold $(\log T)^2$ may heuristically be explained by the fact that in this particular example it happens to be of the same magnitude as the 95% percentile of the asymptotic distribution of the singular value based test statistic presented in Theorem 4. By bootstrapping, this percentile can be estimated to be e.g.

[10] For the systems reported, the behavior of both the estimation of the system order and the test procedures can be improved for $T = 100$ by using smaller values for f and p than $2\hat{p}_{\mathrm{AIC}}$. This seems to make the estimation more precise but is lacking an asymptotic argument.

Table 2
Frequency distributions of the test results for the dimensions of the cointegrating space for Schemes 1–3 and sample sizes and $T = 100, 500, 1000$

Scheme	Sample size	Threshold				Combined			
		0	1	2	3	0	1	2	3
1	$T = 100$	0.039	0.442	0.519	0	0.031	0.142	0.702	0.125
	$T = 500$	0	0.003	0.997	0	0	0	0.977	0.023
	$T = 1000$	0	0	1	0	0	0	0.977	0.023
2	$T = 100$	0.068	0.609	0.323	0	0.053	0.579	0.342	0.026
	$T = 500$	0	0.994	0.006	0	0	0.933	0.067	0
	$T = 1000$	0	1	0	0	0	0.956	0.044	0
3	$T = 100$	0.574	0.1	0.326	0	0.567	0.1	0.325	0.008
	$T = 500$	0.961	0.039	0	0	0.946	0.047	0.007	0
	$T = 1000$	0.984	0.016	0	0	0.967	0.028	0.005	0

The left panel displays the results from the *threshold estimate* and the right panel displays the results from the *combined test sequence*, based on the threshold estimate. The penalty underlying these results is given by $h_T = (\log T)^2$.

approximately 35 for Scheme 2 with cointegrating rank $r = 1$. Comparing this number with $(\log 100)^2 = 21.2$, $(\log 500)^2 = 38.6$ and $(\log 1000)^2 = 47.7$ explains the good performance. Thus, by specifying the initial null hypothesis with this threshold, the combined test sequence starts "closer" to the true number of common trends in a large number of replications. Other employed thresholds did not deliver results as favorable as $h_T = (\log T)^2$. Note once again that only for Scheme 2 and $T = 100$ the threshold estimate and the combined test sequence lead to improvements over the eigenvalue test sequence. We want to stress again that the discussion above is only valid for this particular example and cannot be seen as a justification for choosing the threshold $(\log T)^2$ in all cases. The main message from this example lies in the fact that the singular values based test may be used to obtain a first insight into the number of possible common trends, rather than delivering a precise test.

Another possibility for determining the cointegrating rank, e.g. employed in Bauer and Wagner (2000a), is to bootstrap the singular value based test statistic. Also this approach, although computationally more costly, can easily be implemented and leads to satisfactory results as well. Note that the singular value based test leads to test sequences for determining the number of common trends in the same way as the eigenvalue-based tests.

Turning back to the order estimation it is remarkable that for $T = 500$ and 1000 the correct system order is detected in almost every replication. The good performance for the larger sample sizes can be explained visually by inspecting the estimated singular values.

Fig. 1 shows (for one replication) the estimated singular values for Scheme 1 (left plot) and for Scheme 2 (right plot) for 100 and 1000 observations. It can clearly be seen that for sample size $T = 1000$ the gap between the third and the fourth estimated

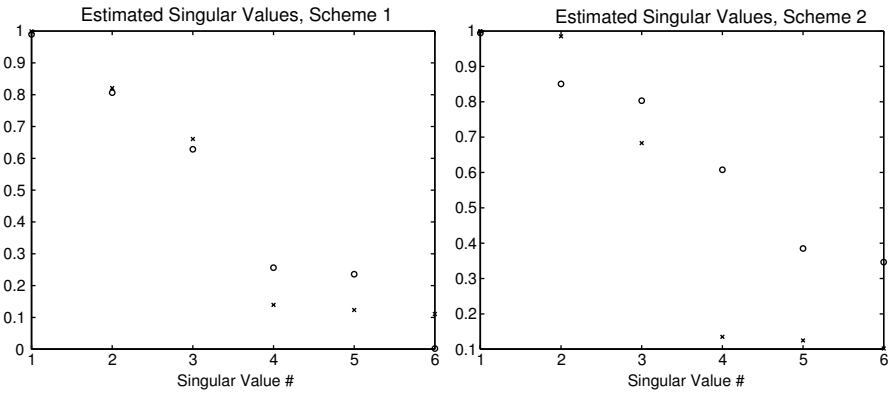


Fig. 1. Estimated singular values for Scheme 1 (left plot) and Scheme 2 (right plot) for one example and sample size $T = 100$ (o) and $T = 1000$ (x) respectively.

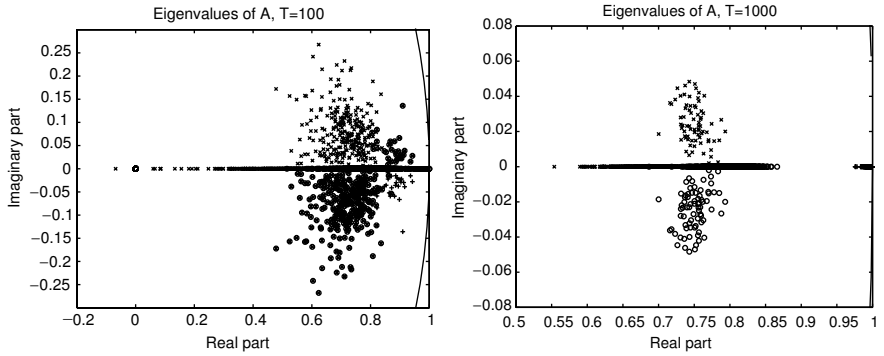


Fig. 2. Estimated eigenvalues for Scheme 1 and sample size $T = 100$ (left plot) and sample size $T = 1000$ (right plot).

singular value is very pronounced, which is reflected by the accuracy of the order estimates. The graphical information presented in Fig. 1 also gives, especially again for $T = 1000$, a clear indication about the number of singular values equal to 1, and therefore about the number of common trends. Note however the similarity of the two plots for 100 observations. This explains the difficulty of estimating the cointegrating rank for this sample size, at least for the singular value based procedures. We can also inspect the eigenvalues of $\hat{A}$. For Scheme 1, with true eigenvalues 1, 0.8 and 0.7, this is done in Fig. 2 for all replications. For $T = 100$ the plot is quite ambiguous, whereas for $T = 1000$ the plot clearly reveals the single unit root at $z = 1$.

For comparison we also report the results obtained by applying the widely used Johansen method on the simulated data. The results are documented in Table 3. The order of the approximating autoregressive model is chosen according to AIC in each

Table 3
Frequency distributions of the test results for the dimensions of the cointegrating space for Schemes 1–3 and sample sizes and $T = 100, 500, 1000$ using the Johansen trace test in an autoregressive approximation of the systems

Sample	$T = 100$				$T = 500$				$T = 1000$			
Coint.	0	1	2	3	0	1	2	3	0	1	2	3
Sch. 1	0.028	0.481	0.425	0.066	0	0	0.93	0.07	0	0	0.922	0.078
Sch. 2	0.085	0.813	0.092	0.001	0	0.913	0.082	0.005	0	0.926	0.07	0.004
Sch. 3	0.496	0.458	0.037	0.009	0.722	0.266	0.009	0.003	0.721	0.273	0.003	0.003

repetition, the test employed is the trace test, the significance level is again 95% and the usual test sequence that starts with an initial null of no cointegration is employed.

Comparing the results in Table 3 with the results for the subspace procedure in Tables 1 and 2 we observe that for Schemes 1 and 3 and $T = 100$ all the subspace tests deliver better results than the Johansen procedure. The ordering is reversed for Scheme 2, where the Johansen trace test produces better results for $T=100$. For Scheme 3 the subspace procedures exhibit a better performance than the Johansen procedure also for $T = 500$ and even for $T = 1000$. It has to be noted however that for Scheme 3 Johansen's max test delivers results that are comparable to the results obtained by the subspace cointegration analysis. For the other systems and the larger sample sizes both procedures have no difficulties in detecting the correct number of cointegrating relationships.

Thus, when we base the decision concerning the cointegrating rank on the eigenvalue test sequence or on the combined test sequence, the subspace cointegration analysis yields in these examples results that are at least comparable to results obtained by applying the Johansen procedure. In addition to that, subspace cointegration analysis has the further advantage of delivering consistent estimates of all system parameters. For VARMA systems the approach based on the Johansen procedure only delivers estimates of an autoregressive approximation, where the orders have to be increased with the sample size to ensure consistency. Modelling the VARMA system with a state space system is of course much more parsimonious than by an infinite order autoregression.

It has to be stressed that the results are so far only based on a couple of simulation exercises. Further work has to be done to gain further understanding about the properties of the proposed method and the test procedures based on it.

7. Summary and conclusions

This paper establishes consistency for an adapted version of the `CCA` subspace algorithm. Based on this result, several methods for testing for (or estimating) the number of common trends and thus equivalently the number of cointegrating relationships have been introduced and analyzed. Furthermore, also a consistent order estimation procedure is provided.

The significance of these results lies in the fact that the method provides consistent estimates of all system parameters, including the cointegrating space, for VARMA processes. This general applicability is an advantage compared to some other methods, like e.g. the method proposed by Johansen for Gaussian VAR processes. As the subspace estimates are computationally very cheap, they can also be used for "cross-validation" of results obtained by the application of standard methods. The subspace estimates can also be used as consistent initial values to obtain efficient estimates of the parameters performing one Newton step for pseudo-maximum likelihood estimation as described in Yap and Reinsel (1995) or Bauer and Wagner (2000b).

The limited simulation evidence presented in this paper indicates that the performance of the method is as good as that of the Johansen procedure. However, further understanding concerning an optimal construction of the tests (e.g. whether the tests should be based on the real parts or the absolute values of the eigenvalues of $\hat{A}$) and an optimal choice of the penalty in deciding about the number of singular values equal to one (if one wants to use the combined test procedure) has to be gained. One advantage of the state space framework is that the user is directly provided with easily accessible information on the cointegrating rank and the system order. The estimated singular values and the eigenvalues of $\hat{A}$ provide the required information that can also be inspected graphically.

In addition to an investigation of the properties of the tests and estimates, 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. This second question 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 subspace algorithms for processes having arbitrary unit roots, i.e. processes with seasonal unit roots as well as processes integrated of higher orders, as the main step of the algorithm is an autoregression, which is known to provide consistent estimates in all these contexts (see Lai and Wei, 1982a, b).

Acknowledgements

The authors would like to thank two anonymous referees, the editor, participants of the International Workshop in Statistical Modelling in Bilbao, the CEMAPRE Workshop in Lisboa and seminar participants in Vienna for valuable comments that helped to improve the paper significantly.

Appendix A. Proofs

Proof of Theorem 2. 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 appropriate transformations

of $Y_{t,f}^{+}$ and $Y_{t,p}^{-}$, as 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, under the assumptions of the theorem, that $y_t = C_1 K_1 \sum_{j=1}^{t} \varepsilon_{t-j} + k_{\mathrm{st}}(z)\varepsilon_t$, where $k_{\mathrm{st}}(z)$ denotes the stable part of the transfer function and where $C_1 \in \mathbb{R}^{s\times c}$, $K_1 \in \mathbb{R}^{c\times s}$, $C_1' C_1 = I_r$. In this representation C_1 is not unique. Bauer and Wagner (2001) show how a unique choice for C_1 can be obtained. Note again that the cointegrating space does not depend on the specific choice of C_1. If $C_1^{\perp} \in \mathbb{R}^{s\times r}$, where again $c=s-r$, is such that $(C_1^{\perp})' C_1^{\perp} = I_c, (C_1^{\perp})' C_1 = 0$, then, with $\bar{C}$ as defined in Section 2, in $\bar{C} y_t$ the first c components are equal to $\sum_{j=1}^{t} K_1 \varepsilon_{t-j} + v_t$, where $v_t = C_1' k_{\mathrm{st}}(z)\varepsilon_t$ is stationary. The remaining r components are stationary and the dimension of the cointegrating space is equal to r. Let $\bar{Q}_f$ be of the form

$$\bar{Q}_f = \begin{bmatrix} I_c & 0^{c\times r} & & & & \\ 0^{r\times c} & I_r & 0 & & & \\ -I_c & 0^{c\times r} & I_c & & & \\ & \ddots & \ddots & \ddots & & \\ & & -I_c & 0^{c\times r} & I_c & 0^{c\times r} \\ & & & & 0^{r\times c} & I_r \end{bmatrix}.$$

Then $Z_{t,f}^{+} = \bar{Q}_f (I \otimes \bar{C}) Y_{t,f}^{+}$ can be represented as

$$Z_{t,f}^{+} = \begin{bmatrix} I \\ 0 \\ \vdots \\ \\ \\ \vdots \\ 0 \end{bmatrix} \left(\sum_{j=0}^{t-2} K_1 \varepsilon_j \right) + \begin{bmatrix} K_1 \varepsilon_{t-1} \\ 0 \\ K_1 \varepsilon_t \\ \\ \vdots \\ \\ K_1 \varepsilon_{t+f-2} \\ 0 \end{bmatrix} + \begin{bmatrix} C_1' k_{\mathrm{st}}(z)\varepsilon_t \\ C_2' k_{\mathrm{st}}(z)\varepsilon_t \\ C_1' k_{\mathrm{st}}(z)\Delta\varepsilon_{t+1} \\ \\ \vdots \\ \\ C_1' k_{\mathrm{st}}(z)\Delta\varepsilon_{t+f-1} \\ C_2' k_{\mathrm{st}}(z)\varepsilon_{t+f-1} \end{bmatrix}.$$

Here $\Delta = (1 - z)$ denotes, as usual, the first difference operator. Analogously, the vector $Y_{t,p}^{-}$ can be transformed to $Z_{t,p}^{-} = \bar{Q}_p (I \otimes \bar{C}) Y_{t,p}^{-}$, which is given by

$$Z_{t,p}^{-} = \begin{bmatrix} I \\ 0 \\ \vdots \\ \\ \\ \vdots \\ 0 \end{bmatrix} \left(\sum_{j=0}^{t-2} K_1 \varepsilon_j \right) - \begin{bmatrix} 0 \\ 0 \\ K_1 \varepsilon_{t-2} \\ \vdots \\ \\ K_1 \varepsilon_{t-p} \\ 0 \end{bmatrix} - \begin{bmatrix} C_1' k_{\mathrm{st}}(z)\varepsilon_{t-1} \\ C_2' k_{\mathrm{st}}(z)\varepsilon_{t-1} \\ C_1' k_{\mathrm{st}}(z)\Delta\varepsilon_{t-1} \\ \vdots \\ \\ C_1' k_{\mathrm{st}}(z)\Delta\varepsilon_{t-p+1} \\ C_2' k_{\mathrm{st}}(z)\varepsilon_{t-p} \end{bmatrix}.$$

Let $D_T = \mathrm{diag}(T^{-1} I_c, T^{-1/2} I_{fs-c})$, where T denotes sample size. From the construction of $Z_{t,f}^{+}$ and $Z_{t,p}^{-}$ it follows that only the first c components are nonstationary, whereas the

remaining components are stationary. Hence the normalization. Let $\langle a_t, b_t\rangle = \sum_{t=1}^{\mathrm{T}} a_t b_t'$. [11] Furthermore we use the following notation: For a scalar random variable f_T let $f_T = \mathrm{o}(g_T)$ mean $\lim f_T/g_T = 0$ a.s. $f_T = \mathrm{O}(g_T)$ means that $|f_T/g_T| \leqslant M$ a.s. for some finite constant M. For matrix-valued random variables the notation is used for the maximum of all entries, i.e. $f_T = \mathrm{O}(g_T)$ means that $\max_{i,j}|f_{T,i,j}/g_T| \leqslant M$ a.s., where $f_{T,i,j}$ denotes the (i,j) element of f_T. $\mathrm{o_P}$ and $\mathrm{O_P}$ denote the corresponding in probability statements. Then the following lemma will be used widely in order to derive the asymptotic properties of the estimates.

Lemma 1. *Let $n_t = \sum_{i=1}^{t-1} \varepsilon_i$ and let $v_t = \sum_{j=0}^{\infty} K_j \varepsilon_{t-j}$, where $k(z) = \sum_{j=0}^{\infty} K_j z^j$ denotes a rational transfer function, whose poles are strictly outside the unit circle and ε_t denotes a strictly stationary martingale difference sequence fulfilling the standard assumptions. Thus, n_t is integrated and v_t is stationary. Then*

- *If $\hat{\gamma}_j = (1/T)\sum_{t=j+1}^{\mathrm{T}} v_t v_{t-j}'$ and $\gamma_j = \mathbb{E} v_t v_{t-j}'$, then $\max_{|j|\leqslant F_T} \|\hat{\gamma}_j - \gamma_j\| = \mathrm{O}(\sqrt{\log\log T/T})$ for $F_T = \mathrm{o}((\log T)^a), a < \infty$.*
- *$\langle n_t, n_t\rangle^{-1/2}\langle n_t, \varepsilon_{t+j}\rangle = \mathrm{o}(\sqrt{\log T}), j \geqslant 0$.*
- *$\langle n_t, n_t\rangle^{-1} = \mathrm{o}(T^{-1})$.*
- *$\langle n_t, n_t\rangle^{-1/2}\langle n_t, v_t\rangle = \mathrm{O_P}(1), T^{-1}\langle n_t, v_t\rangle = \mathrm{O_P}(1), \langle n_t, v_t\rangle = \mathrm{o}(T\log T)$.*
- *$T^{-2}\langle n_t, n_t\rangle \Rightarrow \int_0^1 W(u)W(u)'\,\mathrm{d}u$, where $\Rightarrow$ denotes weak convergence of measures and the Brownian motion $W(u)$ is the limit of $(1/\sqrt{T})\sum_{t=1}^{\lfloor Tu\rfloor} \varepsilon_t$, where $\lfloor x\rfloor$ denotes the smallest integer equal or larger than x.*
- *$\langle n_t, n_t\rangle = \mathrm{o}(T^2\log(T))$.*

Proof. The first point follows from Hannan and Deistler (1988, Theorem 5.3.2). The second and the third points can be found in Lai and Wei (1982a) and Lai and Wei (1982b), respectively. The last three points follow from standard evaluations, see e.g. Johansen (1995), except for $\langle n_t, v_t\rangle = \mathrm{o}(T\log T)$: Note that $v_t = k(z)\varepsilon_t = k(1)\varepsilon_t + (k(z) - k(1))\varepsilon_t$. Thus, $\langle n_t, v_t\rangle = \langle n_t, \varepsilon_t\rangle k(1)' + \langle n_t, (k(z) - k(1))\varepsilon_t\rangle$. The first term is $\mathrm{o}(T\log T)$ according to the results of Lai and Wei (1982a, Corollary 2). For the second term, note that $n_t = \sum_{j=1}^{t-1} \varepsilon_j$ and that $k(z) - k(1) = (1-z)k_d(z)$. This shows that $\langle n_t, (k(z) - k(1))\varepsilon_t\rangle = (\sum_{j=1}^{\mathrm{T}} \varepsilon_j)(k_d(z)\varepsilon_T)' - \sum_{j=1}^{T-1} \varepsilon_j(k_d(z)\varepsilon_j)' = \mathrm{o}(T\log T)$. This follows from the (conditional) zero mean assumption on ε_t, the fact that $k_d(z)\varepsilon_T = \mathrm{O}(T^{1/2})$ and the standard convergence orders for stationary processes. □

Let $\Phi_T = (1/T^2)\sum_t (\sum_{j=0}^{t-2} K_1\varepsilon_j)(\sum_{j=0}^{t-2} K_1\varepsilon_j)'$. Clearly, Φ_T is the (appropriately scaled) dominant term in the nonstationary component of both $Z_{t,f}^+$ and $Z_{t,p}^-$. Further let $\tilde{D}_T = \mathrm{diag}(\Phi_T^{-1/2}, I)D_T$. [12] Then $\tilde{D}_T\langle Z_{t,f}^+, Z_{t,f}^+\rangle\tilde{D}_T$ converges to $\mathrm{diag}(I, \tilde{\Gamma}_f^+)$ in

[11] Alternatively, the summations could be in the range $t = p+1, \ldots, T-f$ without changing the results.

[12] 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.

probability. This can be seen by considering the difference

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

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

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

Here v_t stands for all stationary contributions. Thus, we obtain $T^{-2}\sum_{t=0}^{\mathrm{T}}\Phi_T^{-1/2}(n_t v_t' + v_t n_t' + v_t v_t')(\Phi_T^{-1/2})'$. This matrix converges, when multiplied by T, in distribution to a random variable, since $T^{-1}\sum_{t=0}^{\mathrm{T}} n_t v_t'$ converges in distribution (see e.g. Johansen, 1995, Theorem B.13). The $(2,1)$ (and the $(1,2)$ block, which is the transpose thereof) are of the form $T^{-3/2}\sum_{t=0}^{\mathrm{T}}\Phi_T^{-1/2} n_t v_t' + \mathrm{O_P}(T^{-1/2})$. 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, see also Lemma 1. Finally the $(2,2)$ term is the sample covariance of a stationary process and thus the error converges in distribution, when multiplied by $T^{1/2}$, as follows from standard arguments, see e.g. Hannan and Deistler (1988, Chapter 4). 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, except for the (1,2) block, which is identically zero due to the lower triangular structure of the Cholesky factor.

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

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

where the second expression can be analyzed more easily due to the fact that in $Z^+_{t,f}$ and $Z^-_{t,p}$ the coordinates corresponding to the stationary and the nonstationary parts are separated. Note however that in this equality the square roots are defined differently: If $\langle Z^+_{t,f}, Z^+_{t,f}\rangle^{-1/2}$ corresponds to the Cholesky factor, the expression involving $Y^+_{t,f}$ will not and vice versa, since the transformation mapping $Y^+_{t,f}$ into $Z^+_{t,f}$ is not lower triangular in general. Replacing $Y^+_{t,f}$ with $\hat{Y}^+_{t,f} = (I\otimes\hat{\bar{C}})Y^+_{t,f}$ and $Z^+_{t,f}$ with $\hat{Z}^+_{t,f} = \bar{Q}_f\hat{Y}^+_{t,f}$ this is true and thus for the Cholesky factors $\bar{Q}_f\langle\hat{Y}^+_{t,f}, \hat{Y}^+_{t,f}\rangle^{1/2} = \langle\hat{Z}^+_{t,f}, \hat{Z}^+_{t,f}\rangle^{1/2}$ holds. Note that the error in the replacement of $Z^+_{t,f}$ by $\hat{Z}^+_{t,f}$ is of order $\mathrm{O_P}(T^{-1})$ and thus can be neglected for our purposes. This substitution has however an effect on the asymptotic distribution of the estimates, which is not further analyzed here.

Recall that $Y^+_{t,f} = \mathscr{O}_f\mathscr{K}_p Y^-_{t,p} + \mathscr{O}_f(A - KE^{-1}C)^p x_{t-p} + \mathscr{E}_f E^+_{t,f}$ and thus $Z^+_{t,f} = \tilde{\mathscr{O}}_f\tilde{\mathscr{K}}_p Z^-_{t,p} + \tilde{\mathscr{O}}_f(A - KE^{-1}C)^p x_{t-p} + \tilde{\mathscr{E}}_f E^+_{t,f}$, where $\tilde{\mathscr{O}}_f = \bar{Q}_f(I\otimes\bar{C})\mathscr{O}_f$ and $\tilde{\mathscr{K}}_p$ and

$\tilde{\mathscr{E}}_f$ are defined analogously. Therefore $Z^+_{t,f} = \tilde{\mathscr{O}}_f \tilde{\mathscr{K}}_p Z^-_{t,p} + N^+_{t,f}$, where $\langle N^+_{t,f}, N^+_{t,f}\rangle = o(T^{1+\varepsilon})\ \forall \varepsilon > 0$ due to the bound on the increase of $p(T)$, which implies that $(A - KE^{-1}C)^p = o(1/T)$. Also note that

$$\tilde{\mathscr{O}}_f = \begin{bmatrix} I_c & \tilde{\mathscr{O}}_f^{\mathrm{st},1} \\ 0 & \tilde{\mathscr{O}}_f^{\mathrm{st},2} \end{bmatrix}, \quad \tilde{\mathscr{K}}_p = \begin{bmatrix} I_c & \tilde{\mathscr{K}}_p^{\mathrm{st},1} \\ 0 & \tilde{\mathscr{K}}_p^{\mathrm{st},2} \end{bmatrix},$$

where $\tilde{\mathscr{O}}_f^{\mathrm{st},i}$ and $\tilde{\mathscr{K}}_p^{\mathrm{st},i}$ for $i = 1,2$ correspond to the stationary part. From the results given above, it will be shown that $\hat{X}$ converges to

$$X = \begin{bmatrix} I & 0 \\ 0 & (\tilde{\Gamma}_f^+)^{-1/2} \tilde{\mathscr{H}}_{f,\infty} (\tilde{\Gamma}_\infty^-)^{-1} \tilde{\mathscr{H}}'_{f,\infty} ((\tilde{\Gamma}_f^+)^{-1/2})' \end{bmatrix}. \tag{A.1}$$

In order to do this, it remains to consider the estimation error in the term $\tilde{D}_T \langle Z^+_{t,f}, Z^-_{t,p}\rangle \langle Z^-_{t,p}, Z^-_{t,p}\rangle^{-1} \langle Z^-_{t,p}, Z^+_{t,f}\rangle \tilde{D}'_T$, since

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

Here r_T accounts for the error from replacing estimates with limits in the first and second difference. The first and the last summands have been dealt with above and shown to be convergent in distribution to a constant and thus convergent in probability. Here in the $(1,1)$ sub-block convergence is of order $O_P(T^{-1})$, in the remaining blocks

of order $O_P(T^{-1/2})$. Note that $\langle Z_{t,f}^+, Z_{t,p}^- \rangle = \tilde{\mathscr{O}}_f \tilde{\mathscr{K}}_p \langle Z_{t,p}^-, Z_{t,p}^- \rangle + \langle N_{t,f}^+, Z_{t,p}^- \rangle$. Thus,

$$\begin{aligned}
&\tilde{D}_T \langle Z_{t,f}^+, Z_{t,p}^- \rangle \langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1} \langle Z_{t,p}^-, Z_{t,f}^+ \rangle \tilde{D}_T' \\
&\quad = \tilde{D}_T \tilde{\mathscr{O}}_f \tilde{\mathscr{K}}_p \langle Z_{t,p}^-, Z_{t,p}^- \rangle \tilde{\mathscr{K}}_p' \tilde{\mathscr{O}}_f' \tilde{D}_T' + \tilde{D}_T (\tilde{\mathscr{O}}_f \tilde{\mathscr{K}}_p \langle Z_{t,p}^-, N_{t,f}^+ \rangle \\
&\qquad + \langle N_{t,f}^+, Z_{t,p}^- \rangle \tilde{\mathscr{K}}_p' \tilde{\mathscr{O}}_f' + \langle N_{t,f}^+, Z_{t,p}^- \rangle \langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1} \langle Z_{t,p}^-, N_{t,f}^+ \rangle) \tilde{D}_T'.
\end{aligned}$$

The first term converges in probability to $\mathrm{diag}[I, \tilde{\mathscr{H}}_{f,\infty} (\tilde{\Gamma}_\infty^-)^{-1} \tilde{\mathscr{H}}_{f,\infty}']$, as follows from standard arguments considering the structure of $\tilde{\mathscr{O}}_f$ and $\tilde{\mathscr{K}}_p$. Moreover the error terms in the $(1,1)$ element are of order $O_P(T^{-1})$ and in the remaining blocks of order $O_P(T^{-1/2})$ using standard arguments as above. Recall that $N_{t,f}^+ = \tilde{\mathscr{E}}_f E_{t,f}^+ + \tilde{\mathscr{O}}_f (A - KE^{-1}C)^p x_{t-p}$. Then the convergence of the second and the third terms is straightforward to show using Lemma 1 and the fact that $\tilde{\mathscr{K}}_p Z_{t,p}^- = x_t - (A - KE^{-1}C)^p x_{t-p}$. Here also $\|(A - KE^{-1}C)^p\| = o(1/T)$ is used. For the fourth term apply the matrix inversion lemma to the matrix $\langle Z_{t,p}^-, Z_{t,p}^- \rangle$ to separate the effects of the stationary and the nonstationary part: Let $Z_{t,p}^{-,1}$ and $Z_{t,p}^{-,\mathrm{st}}$ denote the nonstationary and the stationary parts of $Z_{t,p}^-$ respectively. Then

$$\begin{aligned}
\langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1} = &\begin{bmatrix} 0 & 0 \\ 0 & \langle Z_{t,p}^{-,\mathrm{st}}, Z_{t,p}^{-,\mathrm{st}} \rangle^{-1} \end{bmatrix} + \begin{bmatrix} I \\ -\langle Z_{t,p}^{-,\mathrm{st}}, Z_{t,p}^{-,\mathrm{st}} \rangle^{-1} \langle Z_{t,p}^{-,\mathrm{st}}, Z_{t,p}^{-,1} \rangle \end{bmatrix} \\
&\langle Z_{t,p}^{-,II}, Z_{t,p}^{-,II} \rangle^{-1} \left[I, -\langle Z_{t,p}^{-,1}, Z_{t,p}^{-,\mathrm{st}} \rangle \langle Z_{t,p}^{-,\mathrm{st}}, Z_{t,p}^{-,\mathrm{st}} \rangle^{-1} \right], \qquad \text{(A.2)}
\end{aligned}$$

where $\langle Z_{t,p}^{-,II}, Z_{t,p}^{-,II} \rangle = \langle Z_{t,p}^{-,1}, Z_{t,p}^{-,1} \rangle - \langle Z_{t,p}^{-,1}, Z_{t,p}^{-,\mathrm{st}} \rangle \langle Z_{t,p}^{-,\mathrm{st}}, Z_{t,p}^{-,\mathrm{st}} \rangle^{-1} \langle Z_{t,p}^{-,\mathrm{st}}, Z_{t,p}^{-,1} \rangle$. Thus, the fourth term is equal to

$$\begin{aligned}
&\tilde{D}_T (\langle N_{t,f}^+, Z_{t,p}^{-,\mathrm{st}} \rangle \langle Z_{t,p}^{-,\mathrm{st}}, Z_{t,p}^{-,\mathrm{st}} \rangle^{-1} \langle Z_{t,p}^{-,\mathrm{st}}, N_{t,f}^+ \rangle + \langle N_{t,f}^+, Z_{t,p}^{-,II} \rangle \\
&\quad \langle Z_{t,p}^{-,II}, Z_{t,p}^{-,II} \rangle^{-1} \langle Z_{t,p}^{-,II}, N_{t,f}^+ \rangle) \tilde{D}_T'.
\end{aligned}$$

Therefore in the first term only stationary variables and $\tilde{\mathscr{O}}_f (A - KE^{-1}C)^p x_{t-p}$ occur and thus the term is of order $o(p(T)/T)$ a.s. Here also the fact that $\langle x_{t-p}, x_{t-p} \rangle = o(T^2 \log T)$ is used, see e.g. Lemma 1. Corresponding to the second term note that due to Lemma 1 $\langle Z_{t,p}^{-,II}, Z_{t,p}^{-,II} \rangle = \langle Z_{t,p}^{-,1}, Z_{t,p}^{-,1} \rangle + o(T(\log T)^2 p(T))$ and $\langle N_{t,f}^+, Z_{t,p}^{-,II} \rangle = \langle N_{t,f}^+, Z_{t,p}^{-,1} \rangle + o(\sqrt{T}(\log T)^2 p(T))$. Thus, the crucial term can be seen to be essentially equal to $\tilde{D}_T \langle N_{t,f}^+, Z_{t,p}^{-,1} \rangle \langle Z_{t,p}^{-,1}, Z_{t,p}^{-,1} \rangle^{-1} \langle Z_{t,p}^{-,1}, N_{t,f}^+ \rangle \tilde{D}_T'$ and is also seen to converge to zero. Considering the various error terms, it is straightforward but cumbersome to show that the $(1,1)$ block of r_T is of order $o_P(T^{-1})$. The remaining blocks of r_T are of order $o_P(T^{-1/2})$, as they are the sum of products of matrices having the same property. Note that the increase in $p(T)$ only affects the stationary part $Z_{t,p}^{-,\mathrm{st}}$ and thus especially the inverse $\langle Z_{t,p}^{-,\mathrm{st}}, Z_{t,p}^{-,\mathrm{st}} \rangle$ can be dealt with using well-understood theory for stationary processes, showing e.g. that the infinity norm of the inverse is bounded uniformly in $p(T)$, for $0 \leqslant p(T) \leqslant (\log(T))^a$ and $a < \infty$.

In subspace algorithms an eigenvalue decomposition is performed on $\hat{X}$. For the limit X the first $c = s - r$ eigenvalues are equal to one, assuming the cointegrating rank to be equal to r. The corresponding eigenvectors span the space corresponding to the first c vectors of the canonical basis. With regard to the remaining eigenvalues and

eigen-vectors note that the term $(\tilde{\Gamma}_f^+)^{-1/2}\tilde{\mathscr{H}}_{f,p}((\tilde{\Gamma}_p^-)^{-1/2})'$ corresponds to the stationary transfer function

$$\tilde{k}(z) = \begin{bmatrix} (1-z)I_c & 0 \\ 0 & I_r \end{bmatrix} \bar{C}k(z)$$

as can be shown from the definition of $Z_{t,f}^+$ and $Z_{t,p}^-$. The transfer function $\tilde{k}(z)$ is of order smaller or equal to n. This can be seen by considering the nonminimal representation

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

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

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

From the expressions for $Z_{t,f}^+$ and $Z_{t,p}^-$ and realization theory for the stationary case it follows that $\tilde{\mathscr{H}}_{f,\infty}((\tilde{\Gamma}_\infty^-)^{-1/2})'$ 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 $\tilde{\Gamma}_\infty^-$. Now some algebraic computations show that the Markov parameters $\tilde{K}(j)$ of $\tilde{k}(z)$ are equal to

$$\tilde{K}(1) = \begin{bmatrix} C_1'C_{\mathrm{st}}K_{\mathrm{st}} + K_1 - C_1'E \\ C_2'C_{\mathrm{st}} \end{bmatrix}, \quad \tilde{K}(j) = \begin{bmatrix} C_1'C_{\mathrm{st}}(A_{\mathrm{st}} - I_{n-c})A_{\mathrm{st}}^{j-2}K_{\mathrm{st}} \\ C_2'C_{\mathrm{st}}A_{\mathrm{st}}^{j-1} \end{bmatrix}, \ j \geqslant 2.$$

It follows that $\tilde{\mathscr{H}}_{f,\infty}((\tilde{\Gamma}_\infty^-)^{-1/2})'$ is of rank $n-c$, since it is, up to an orthonormal transformation, essentially the Hankel matrix of the coefficients $\tilde{K}(j)$, where the first c rows have been omitted. The typical element of this matrix is equal to $C_1'C_{\mathrm{st}}(A_{\mathrm{st}} - I_{n-c})A_{\mathrm{st}}^{j-2}K_{\mathrm{st}}$ or $C_2'C_{\mathrm{st}}A_{\mathrm{st}}^{j-1}K_{\mathrm{st}}$. The Hankel matrix with the first c rows omitted can be factored into

$$\begin{bmatrix} C_2'C_{\mathrm{st}} \\ C_1'C_{\mathrm{st}}(A_{\mathrm{st}} - I_{n-c}) \\ C_2'C_{\mathrm{st}}A_{\mathrm{st}} \\ C_1'C_{\mathrm{st}}(A_{\mathrm{st}} - I_{n-c}) \\ \vdots \end{bmatrix} [K_{\mathrm{st}} \quad A_{\mathrm{st}}K_{\mathrm{st}} \quad A_{\mathrm{st}}^2K_{\mathrm{st}} \quad \cdots],$$

where the second matrix is equal to the controllability matrix of the stationary part and thus of full rank $n-c$. Thus, consider the case, where the first matrix above has not full column rank. Then, there exists a vector x such that $C_1'C_{\mathrm{st}}(A_{\mathrm{st}} - I_{n-c})A_{\mathrm{st}}^j x = 0, j > 0$, which implies that $C_1'C_{\mathrm{st}}A_{\mathrm{st}}^{j+1}x = C_1'C_{\mathrm{st}}A_{\mathrm{st}}^j x, j > 0 \Rightarrow C_1'C_{\mathrm{st}}A_{\mathrm{st}}^j x = 0, j > 0$. Also $C_2'C_{\mathrm{st}}A_{\mathrm{st}}^j x = 0, j > 0$ must hold in this case. Thus, $x = 0$ follows from minimality of the stationary state space system $(A_{\mathrm{st}}, K_{\mathrm{st}}, C_{\mathrm{st}})$. This implies that the number of nonzero

singular values of the limit of $\hat{X}=(\langle Y_{t,f}^+, Y_{t,f}^+\rangle)^{-1/2}\langle Y_{t,f}^+, Y_{t,p}^-\rangle(\langle Y_{t,p}^-, Y_{t,p}^-\rangle^{-1/2})'$ is (generically) equal to n, the order of the system. From Eq. (A.1) it thus follows that the SVD leads to a factorization

$$\begin{bmatrix} I_c & 0 \\ 0 & (\tilde{\Gamma}_f^+)^{-1/2}\bar{\mathcal{O}}_f \end{bmatrix}\begin{bmatrix} I_c & 0 \\ 0 & \bar{\mathcal{K}}_p(\tilde{\Gamma}_p^-)^{1/2} \end{bmatrix}.$$

Here $\bar{\mathcal{O}}_f$ and $\bar{\mathcal{K}}_p$ correspond to the decomposition of the stationary part. From the definitions of $\tilde{A}$ and $\tilde{C}$ it follows that $\tilde{C}\tilde{A}=[0,\bar{C}C_{\mathrm{st}}A_{\mathrm{st}}-[{}_0^I]C_1'C_{\mathrm{st}},0]$ and thus only the $n-c$ columns in the middle of $\tilde{\mathcal{O}}_f$ contribute to $\bar{\mathcal{H}}_{f,p}((\tilde{\Gamma}_p^-)^{-1/2})'$. It follows from the definitions of $\tilde{A}$ and $\tilde{K}$ that the middle rows of $\tilde{\mathcal{C}}$ correspond to the controllability matrix corresponding to k_{st}. Therefore, $\tilde{\mathcal{C}}_\infty(\tilde{\Gamma}^-)^{-1/2}Z_{t,\infty}^-=\bar{\mathcal{K}}_\infty Z_{t,\infty}^-=x_{t,\mathrm{st}}$, the stationary part of the state. Which particular realization $A_{\mathrm{st}},K_{\mathrm{st}}$ is used is determined by the SVD. Furthermore, the convergence of the matrix $\hat{X}$ implies the convergence of the eigenvalues and also the eigenspaces, as follows from the next lemma. Thus, let $\hat{U}_n$ denote the matrix, whose columns correspond to the eigenvectors to the n dominant eigenvalues of $\hat{X}$. The following lemma (see e.g. Chatelin, 1983) provides tools to assess the estimation error.

Lemma 2. *Let $\mathcal{T}$ denote a symmetric, positive definite compact linear operator and let $\hat{\mathcal{T}}$ denote a sequence of symmetric, positive definite compact operators converging to $\mathcal{T}$. Let $\lambda_1>\cdots>\lambda_k>0$ denote the k, say, distinct nonzero eigenvalues of $\mathcal{T}$ 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 $\mathcal{T}$. Furthermore, let $\hat{\lambda}_{i,j}$ and $\hat{P}_i$ denote the corresponding approximating quantities calculated from $\hat{\mathcal{T}}$. 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 (for a definition see Chatelin (1983), or Appendix B).*

Furthermore, the following first order approximations hold:

$$\frac{1}{k_i}\sum_{j=1}^{k_i}\hat{\lambda}_{i,j}=\lambda_i+\frac{1}{k_i}\operatorname{tr}[(\hat{\mathcal{T}}-\mathcal{T})P_i], \tag{A.3}$$

$$\hat{P}_i=P_i+\sum_{\lambda_j\neq\lambda_i}\left\{\frac{1}{\lambda_i-\lambda_j}P_j[\hat{\mathcal{T}}-\mathcal{T}]P_i+\frac{1}{\lambda_i-\lambda_j}P_i[\hat{\mathcal{T}}-\mathcal{T}]P_j\right\}. \tag{A.4}$$

From the lemma it follows that the probability that there exists a nonsingular matrix $\hat{S}_T$, such that $\tilde{U}_n=\hat{U}_n\hat{S}_T=[{}_{\hat{U}_{n,1}}^{I}\ {}_{\hat{U}_{n,2}}^{0}]$ converges to one. Further $\hat{U}_n\hat{S}_T$ converges in probability to $U_0=[{}_0^I\ {}_{\hat{U}_0}^0]$. Here again $\tilde{U}_0$ corresponds to the stationary part. In the adapted procedure the zero blocks are explicitly imposed, which is only possible, since the Cholesky factors of $\langle\hat{Z}_{t,f}^+,\hat{Z}_{t,f}^+\rangle$ can be calculated. Note that the Cholesky factor of $\langle Z_{t,f}^+,Z_{t,f}^+\rangle$ cannot be calculated in general. 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 $\hat{X}$ and thus in particular power series expansions exist, ensuring the validity of linearization arguments. Consider the estimate

$$\begin{aligned}\hat{x}_t &= \tilde{U}_n'\langle Y_{t,f}^+, Y_{t,f}^+\rangle^{-1/2}\langle Y_{t,f}^+, Y_{t,p}^-\rangle\langle Y_{t,p}^-, Y_{t,p}^-\rangle^{-1}Y_{t,p}^- \\ &= \tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\langle Z_{t,f}^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- \\ &= \tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\langle (W_f^+)^{-1}U_0x_t + \Gamma x_{t,\mathrm{st}} + \mathscr{E}E_{t,f}^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^-.\end{aligned}$$

Here the limit of $\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\tilde{D}_T^{-1}$ is denoted with W_f^+. Furthermore note that $\Gamma = [\Gamma_1, 0^{c\times(fs-c)}]'$ and thus only affects the components due to the common trends. Note that $U_0'W_f^+(W_f^+)^{-1}U_0 = I_n$ and thus contrary to the stationary case, the estimate $\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\langle Z_{t,f}^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}$ is not consistent for $\tilde{\mathscr{K}}_p$, since it neglects the term $U_0'W_f^+[0,\Gamma]\tilde{\mathscr{K}}_p$ due to the fact that the common trends dominate the first components of the state. Let $G = (I_n + U_0'W_f^+[0,\Gamma])^{-1}$. Recall that $x_t = \tilde{\mathscr{K}}Z_{t,\infty}^- = \tilde{\mathscr{K}}_pZ_{t,p}^- + (A - KE^{-1}C)^px_{t-p}$. Therefore

$$\begin{aligned}&G\tilde{D}_T^{-1}\hat{x}_t - x_t \\ &\quad= G\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\langle \tilde{\mathscr{O}}_fx_t + \tilde{\mathscr{E}}E_{t,f}^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- - x_t \\ &\quad= (G\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\tilde{\mathscr{O}}_f - I)\langle \tilde{\mathscr{K}}_pZ_{t,p}^-, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- \\ &\qquad+ G\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\langle \tilde{\mathscr{E}}_fE_{t,f}^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- - (A - KE^{-1}C)^px_{t-p} \\ &\qquad+ G\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\langle \tilde{\mathscr{O}}_f(A - KE^{-1}C)^px_{t-p}, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- \\ &\quad= (G\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\tilde{\mathscr{O}}_f - I)\tilde{\mathscr{K}}_pZ_{t,p}^- - (A - KE^{-1}C)^px_{t-p} \\ &\qquad+ G\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\tilde{\mathscr{O}}_f(A - KE^{-1}C)^px_{t-p}, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^- \\ &\qquad+ G\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}\langle \tilde{\mathscr{E}}_fE_{t,f}^+, Z_{t,p}^-\rangle\langle Z_{t,p}^-, Z_{t,p}^-\rangle^{-1}Z_{t,p}^-.\end{aligned}$$

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:

$$\begin{aligned}\hat{C}\tilde{D}_TG^{-1} - C &= \left(\sum_{t=1}^{\mathrm{T}}(y_t - CG\tilde{D}_T^{-1}\hat{x}_t)\hat{x}_t'\right)\left(\sum_{t=1}^{\mathrm{T}}\hat{x}_t\hat{x}_t'\right)^{-1}\tilde{D}_TG^{-1} \\ &= \left(\sum_t[C(x_t - G\tilde{D}_T^{-1}\hat{x}_t) + E\varepsilon_t]\hat{x}_t'\right)\left(\sum_t\hat{x}_t\hat{x}_t'\right)^{-1}\tilde{D}_TG^{-1}.\end{aligned}$$

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. The above evaluations apply for both, the standard and the adapted, procedures. In both cases the block matrix inversion of $\langle Z_{t,p}^-, Z_{t,p}^-\rangle$ is used analogous to Eq. (A.2). Considering the expression given above, one can show that for the adapted

procedure $\langle G\tilde{D}_T^{-1}\hat{x}_t - x_t, \hat{x}_t\rangle$ converges in distribution, if $p = p(T) \geqslant -d\log T/\log|\rho_0|$, where $d > 1$. We impose this 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}. Here all evaluations are standard, except for the term $(G\tilde{D}_T^{-1}\tilde{U}_n'\langle Z_{t,f}^+, Z_{t,f}^-\rangle^{-1/2}\tilde{\mathscr{O}}_f - I)$ which is equal to

$$[G\tilde{D}_T^{-1}\tilde{U}_n'\tilde{D}_T\tilde{D}_T^{-1}\langle Z_{t,f}^+, Z_{t,f}^-\rangle^{-1/2} - GU_0'(W_f^+)][(W_f^+)^{-1}U_0 + [0, \Gamma]].$$

Noting that for the adapted procedure $\tilde{D}_T^{-1}\langle Z_{t,f}^+, Z_{t,f}^+\rangle^{-1/2}$ converges to W_f^+ and that $\tilde{D}_T^{-1}\tilde{U}_n'\tilde{D}_T$ converges to U_0' in probability, shows the convergence of this expression to zero. Evaluating the errors in each sub-block shows that the expression times $\tilde{D}_T^{-1}$ converges in distribution for the adapted procedure. For the algorithm not taking the cointegrating rank into account, this statement does no longer hold, since in this case the $(1,2)$ block of $\tilde{D}_T^{-1}\tilde{U}_n'\tilde{D}_T$ is of order $O_P(1)$ and thus does not converge to zero. In this case the matrix post-multiplied by $\mathrm{diag}(TI_c, I_{fs-c})$ converges in distribution. Similar reasoning for the remaining terms shows the assertion that $\langle G\tilde{D}_T^{-1}\hat{x}_t - x_t, \hat{x}_t\rangle$ converges in distribution for the case of the adapted procedure. Therefore, $\hat{C}\tilde{D}_T G^{-1}$ converges in probability to C and furthermore $(\hat{C}\tilde{D}_T G^{-1} - C)G\tilde{D}_T^{-1}$ converges in distribution, establishing the familiar convergence of order T for the complement of the cointegrating space (and thus also for the cointegrating space) for the adapted procedure.

For the standard subspace procedure one can show that the order of convergence of the cointegrating space still applies, whereas the remaining columns of $\hat{C}\tilde{D}_T - C$ only converge in distribution and thus are not estimated consistently by the standard `CCA` procedure. Also note that the first columns of G are equal to the corresponding columns of the identity matrix, showing that the first c columns of $\tilde{\mathscr{O}}_f$ and $\tilde{\mathscr{O}}_f G^{-1}$ are identical and thus the same result applies for the transformed system, which is realizable from the data.

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

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

$$\begin{aligned} G\tilde{D}_T^{-1}\hat{A}\tilde{D}_T G^{-1} - A &= \langle G\tilde{D}_T^{-1}\hat{x}_{t+1} - AG\tilde{D}_T^{-1}\hat{x}_t, \hat{x}_t\rangle\langle\hat{x}_t, \hat{x}_t\rangle^{-1}\tilde{D}_T G^{-1} \\ &= \langle G\tilde{D}_T^{-1}\hat{x}_{t+1} - x_{t+1}, \hat{x}_t\rangle\langle\hat{x}_t, \hat{x}_t\rangle^{-1}\tilde{D}_T G^{-1} \\ &\quad + A\langle x_t - G\tilde{D}_T^{-1}\hat{x}_t, \hat{x}_t\rangle\langle\hat{x}_t, \hat{x}_t\rangle^{-1}\tilde{D}_T G^{-1} \\ &\quad + \langle K\varepsilon_t, \hat{x}_t\rangle\langle\hat{x}_t, \hat{x}_t\rangle^{-1}\tilde{D}_T G^{-1}. \end{aligned}$$

It follows from the arguments given above that all these terms converge to zero in probability (using the expression for $G\tilde{D}_T^{-1}\hat{x}_t - x_t$ and the analogous expression for $G\tilde{D}_T^{-1}\hat{x}_{t+1} - x_{t+1}$).

Finally also consistency of $\hat{K}$ is shown. Note that for $\hat{\varepsilon}_t = \hat{E}^{-1}(y_t - \hat{C}\hat{x}_t)$ it holds that $\langle \hat{\varepsilon}_t, \hat{x}_t \rangle = 0$, since $\tilde{\varepsilon}_t = \hat{E}\hat{\varepsilon}_t$ and $\langle \tilde{\varepsilon}_t, \hat{x}_t \rangle = 0$, as $\tilde{\varepsilon}_t$ denotes the residuals of the first regression, where $\hat{x}_t$ were used as regressors. The relevant quantity in accordance with the results for $\hat{A}$ and $\hat{C}$ is equal to $G\tilde{D}_T^{-1}\hat{K}$. Therefore, consider

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

Tedious but straightforward calculations show that this expression converges to K in probability. It finally remains to show consistency also for the reduced rank regression approach, i.e. the procedure where the system is estimated under a constraint on the number of cointegrating relationships. The proof of consistency follows directly from using the consistency of the state estimation (apparent e.g. from the equation for $G\tilde{D}_T^{-1}\hat{x}_t - x_t$) and the well-known consistency of e.g. the Johansen procedure. The latter is a reduced rank regression problem itself. This completes the proof. □

Remark. Note however that the proof only shows consistency for the transfer function estimates. The system description $(\hat{A}, \hat{K}, \hat{C}, \hat{E})$ itself 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). The proof given above then shows the consistency for the estimated system matrices on a generic subset. Note that the state space echelon canonical form can easily be transformed to a VARMA representation, if this is the preferred system representation (see Section 2).

Proof of Theorem 3. Consider $\hat{Y} = \langle Y_{t,f}^+, Y_{t,f}^+ \rangle^{-1/2} \langle Y_{t,f}^+, Y_{t,p}^- \rangle (\langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1/2})'$ and the corresponding limit of the matrix sequence $\hat{Y}$, denoted by Y. The relevant quantity for order estimation is $\|\hat{Y} - Y\|$. In order to see this, first consider the probability of underestimating the order, i.e. choosing the order $n < n_0$. Simple manipulations show that $\mathbb{P}\{SVC(n) < SVC(n_0)\} = \mathbb{P}\{\hat{\sigma}_{n+1}^2 - \hat{\sigma}_{n_0+1}^2 < H_T(d(n_0) - d(n))/T\}$. Since $d(n) < d(n_0)$ for $n < n_0$ and $H_T/T \to 0$ this probability converges to zero, if $\hat{\sigma}_i \to \sigma_i$, $\forall i \leqslant n_0$. On the other hand, overestimation occurs, if $\min_{n>n_0} SVC(n) < SVC(n_0)$. The probability for this to occur is equal to

$$
\begin{aligned}
&\mathbb{P}\{\hat{\sigma}_{n_0+1}^2 + H_T d(n_0)/T - \min(\hat{\sigma}_{n+1}^2 + d(n)H_T/T) > 0\} \\
&\quad \leqslant \mathbb{P}\{\hat{\sigma}_{n_0+1}^2 + H_T d(n_0)/T - d(n_0+1)H_T/T) > 0\}
\end{aligned}
$$

$$= \mathbb{P}\{\hat{\sigma}^2_{n_0+1} > (d(n_0+1) - d(n_0))H_T/T\}$$
$$= \mathbb{P}\left\{\frac{T}{H_T}\hat{\sigma}^2_{n_0+1} > d(n_0+1) - d(n_0)\right\}.$$

Therefore the results is proven, if it can be shown that $\sqrt{(T/H_T)}\hat{\sigma}_{n_0+1} \to 0$ in probability. Note that $\hat{\sigma}_{n_0+1} \leqslant C\|\hat{P}_2 - P_2\|_2 + \|\hat{Y} - Y\|_2$ (see Bauer (1998) for a proof). Here $\hat{P}_2$ denotes the orthogonal projection onto the orthogonal complement of $\hat{U}_{n_0}$ and P_2 the corresponding limit. Thus, it follows from Lemma 2 that $\hat{\sigma}_{n_0+1} = \mathrm{O_P}(\|\hat{Y} - Y\|_2)$ and it is sufficient to show that $\sqrt{T/H_T}\|\hat{Y} - Y\|_2 \to 0$ in probability.

In this respect take $\tilde{D}_T\langle Z^+_{t,f}, Z^-_{t,p}\rangle\tilde{D}'_T - \mathrm{diag}(I, \tilde{\mathscr{H}})$ as an example. The $(1,1)$ block has been shown to be of order $\mathrm{O_P}(T^{-1})$ and is thus $\mathrm{o_P}(\sqrt{H_T/T})$. The $(2,1)$ block is of fixed finite size and has been shown to converge in distribution, when multiplied with $\sqrt{T}$. Therefore it is also $\mathrm{o_P}(\sqrt{H_T/T})$. The $(2,2)$ block corresponds to stationary variables and therefore can be dealt with using standard methods, showing that it is $\mathrm{O}(Q_T\sqrt{p})$, where $Q_T = \sqrt{\log\log T/T}$, showing that also the norm of this component is $\mathrm{o_P}(\sqrt{H_T/T})$ under the assumptions imposed upon H_T. Finally consider the $(1,2)$ block. This is equal to $(T^{-2}\langle n_t, n_t\rangle)^{-1/2}T^{-3/2}\langle z^+_t, Z^{-,\mathrm{st}}_{t,p}\rangle$, where $n_t = \sum_{j=0}^{t-1} K_1\varepsilon_j, z^+_t = C'_1 y_t$ and $Z^{-,\mathrm{st}}_{t,p}$ denotes the stationary part of $Z^-_{t,p}$ as before. According to the results of Lemma 1 it follows that

$$\langle z^+_t, Z^{-,\mathrm{st}}_{t,p}\rangle = \langle n_t + v_t, Z^{-,\mathrm{st}}_{t,p}\rangle = \langle n_t, \varepsilon_t\rangle K_V(1)' + \left(\sum_{j=1}^{\mathrm{T}}\varepsilon_j\right)V'_T + \langle \varepsilon_t, V_t\rangle + \mathrm{o}(T\sqrt{\log\log Tp(T)}),$$

where $Z^{-,\mathrm{st}}_{t,p} = K_V(z)\varepsilon_t = K_V(1)\varepsilon_t + (1-z)V_t$. Note that the entries of $K_V(1)$ are bounded uniformly in $p(T)$. This decomposition holds for each fixed p. Thus

$$\|(T^{-2}\langle n_t, n_t\rangle)^{-1/2}\langle z^+_t, Z^{-,\mathrm{st}}_{t,p}\rangle\|_2$$
$$\leqslant \|(T^{-2}\langle n_t, n_t\rangle)^{-1/2}\langle n_t, \varepsilon_t\rangle\|_2\|K_V(1)\|_2 + \|(T^{-2}\langle n_t, n_t\rangle)^{-1/2}\left(\sum_{j=1}^{\mathrm{T}}\varepsilon_j\right)V'_T\|_2$$
$$+ \|(T^{-2}\langle n_t, n_t\rangle)^{-1/2}\langle \varepsilon_t, V_t\rangle\|_2 + \mathrm{o_P}(T\sqrt{H_T}).$$

The first term is $\mathrm{o_P}(T\sqrt{H_T})$ and thus of correct order also for $p \to \infty$ as indicated in the theorem. The same holds true for the last term due to stationarity arguments. Finally note that $\mathbb{E}\|V_T\|_2^2 = \mathbb{E}\sum_{j=1}^{p} V_T(j)^2 \leqslant Cp$, since the entries of V_T have bounded variance (uniformly in p). Therefore also this term is of the required order. Thus, $\sqrt{(T/H_T)}\|\tilde{D}_T\langle Z^+_{t,f}, Z^-_{t,p}\rangle\tilde{D}'_T - \mathrm{diag}(I, \tilde{\mathscr{H}})\|_2 \to 0$ in probability. Similar arguments for the remaining terms show that $\sqrt{T/H_T}\|\hat{Y} - Y\|_2 = \mathrm{o_P}(1)$. This completes the proof. □

Remark. For $c = 1$ an a.s. consistency result could be obtained using the above techniques and the bound $\limsup\langle n_t, n_t\rangle^{-1} = \mathrm{O}(T^{-2}\sqrt{\log\log T})$, see e.g. Lai and Wei (1982a). In this case Lemma 1 may be strengthened to $\langle n_t, v_t\rangle = \mathrm{o}(T\sqrt{\log T})$. Therefore it follows that $\langle Z^+_{t,f}, Z^+_{t,f}\rangle^{-1/2} = W^+_f + \mathrm{o}((\log T)T^{-1/2})$. Similar arguments show

that $\|\hat{Y}-\hat{Y}\|_{\mathrm{Fr}}^2 = \mathrm{O}(\max\{(\log T)^2/T, (p(T)\log T)/T\})$. Here $\|\ \|_{\mathrm{Fr}}$ denotes the Frobenius norm. This gives a (somewhat heuristic) motivation for a penalty term $H_T = (\log T)^2$ for the choice $p(T) = \mathrm{O}(\log T)$.

Proof of Theorem 4. The asymptotic properties of the eigenvalues (or equivalently of the singular values) have already been stated in Eq. (A.3) in the proof of Theorem 2 in this appendix. Thus, we have to evaluate $\mathrm{tr}[P_1(\hat{X}-X)]$, which can easily be seen to equal $\mathrm{tr}[\hat{X}^{1,1}-X^{1,1}]$, where the superscript ^{1,1} denotes the $(1,1)$ block of the respective quantities. Let $z_t^+ = \sum_{j=0}^{t-1} K_1\varepsilon_j + C_1' k_{\mathrm{st}}(z)\varepsilon_t$ denote the vector of the first c components of $Z_{t,f}^+$. Then it is straightforward to see that the relevant quantity is equal to

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

Let the first c rows of $Z_{t,p}^-$ be denoted by $z_t^- = \sum_{j=0}^{t-2} K_1\varepsilon_j + C_1' k_{\mathrm{st}}(z)\varepsilon_{t-1}$. Then it follows that $z_t^+ = z_t^- + K_1\varepsilon_{t-1} + C_1' k_{\mathrm{st}}(z)\Delta\varepsilon_t = z_t^- + C_1'\Delta y_t$. 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

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

Using the decomposition of z_t^+ and Eq. (A.2) again the essential term in the second summand is seen to equal $\langle a_t, z_t^+\rangle - \langle a_t, a_t\rangle + \langle a_t, Z_{t,p}^{-,\mathrm{st}}\rangle\langle Z_{t,p}^{-,\mathrm{st}}, Z_{t,p}^{-,\mathrm{st}}\rangle^{-1}\langle Z_{t,p}^{-,\mathrm{st}}, a_t\rangle$, where $Z_{t,p}^{-,\mathrm{st}}$ denotes the stationary part of $Z_{t,p}^-$. Therefore up to first order approximation we obtain

$$\begin{aligned}T\left(1 - \frac{1}{c}\sum_{j=1}^{c}\hat{\sigma}_j^2\right) \doteq \frac{T}{c}\,\mathrm{tr}[&\langle z_t^+, z_t^+\rangle^{-1}\{\langle a_t, a_t\rangle - \langle a_t, Z_{t,p}^{-,\mathrm{st}}\rangle\\ &\langle Z_{t,p}^{-,\mathrm{st}}, Z_{t,p}^{-,\mathrm{st}}\rangle^{-1}\langle Z_{t,p}^{-,\mathrm{st}}, a_t\rangle\}].\end{aligned}$$

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

Proof of Theorem 5. The eigenvalues of $T(\langle K_1\varepsilon_t, x_{t,1}\rangle\langle x_{t,1}, x_{t,1}\rangle^{-1})$ converge in distribution to the distribution mentioned in the formulation of the theorem. Here $x_{t,1} \in \mathbb{R}^c$ denotes the first c coordinates of the state in the canonical form. Note that the eigenvalues do not depend on the choice of the basis for $x_{t,1}$, i.e. a transformation $Sx_{t,1}$ of $x_{t,1}$ leaves the eigenvalues unchanged. Using Eq. (A.2) it follows that

$$T\langle K_1\varepsilon_t, x_t\rangle\langle x_t, x_t\rangle^{-1}[I_c, 0^{c\times n}]' = T(\langle K_1\varepsilon_t, x_{t,1}\rangle\langle x_{t,1}, x_{t,1}\rangle^{-1}) + \mathrm{o_P}(1).$$

Noting that $K_1\varepsilon_t = x_{t+1,1} - x_{t,1}$, it remains to show that the replacement of x_t by $\tilde{x}_t = \hat{S}_T G\tilde{D}_T^{-1}\hat{\mathscr{K}}_p Z_{t,p}^-$ does not change the asymptotic distribution, where $\hat{S}_T$ denotes the

matrix transforming the estimates $(G\tilde{D}_T^{-1}\hat{A}\tilde{D}_T G^{-1}, G\tilde{D}_T^{-1}\hat{K}, \hat{C}\tilde{D}_T G^{-1})$ into the canonical form, where $G\tilde{D}_T^{-1}\hat{A}\tilde{D}_T G^{-1}$ is in Jordan normal form. Here $\hat{A}$ denotes the estimate from the unrestricted regression.

Note that due to the order of consistency it follows that $\hat{S}_T = \mathrm{diag}(\hat{S}_T^1, \hat{S}_T^{\mathrm{st}}) + O_{\mathrm{P}}(T^{-1/2}), \hat{S}_T^1 \in \mathbb{R}^{c\times c}$ and thus $\hat{S}_T G\tilde{D}_T^{-1}\hat{\mathscr{K}}_p - \mathscr{K}_p = \mathrm{O}_{\mathrm{P}}(T^{-1/2})$ (see the decomposition of $x_t - G\tilde{D}_T^{-1}\hat{\mathscr{K}}_p Z_{t,p}^-$). Therefore it follows that $T^{-2}(\langle x_{t,1}, x_{t,1}\rangle - \langle \tilde{x}_{t,1}, \tilde{x}_{t,1}\rangle) = \mathrm{o}_{\mathrm{P}}(1)$, where the additional subscript $_1$ denotes the first c components. Also $T^{-1}(\langle \tilde{x}_{t+1,1} - \tilde{x}_{t,1}, \tilde{x}_{t,1}\rangle - \langle x_{t+1,1} - x_{t,1}, x_{t,1}\rangle) = \mathrm{o}_{\mathrm{P}}(1)$. This shows that the asymptotic distributions of $T\langle \tilde{x}_{t+1,1} - \tilde{x}_{t,1}, \tilde{x}_{t,1}\rangle\langle \tilde{x}_{t,1}, \tilde{x}_{t,1}\rangle^{-1}$ and $T\langle x_{t+1,1} - x_{t,1}, x_{t,1}\rangle\langle x_{t,1}, x_{t,1}\rangle^{-1}$ coincide. Analogous arguments show that

$$T\langle \tilde{x}_{t+1,1} - \tilde{x}_{t,1}, \tilde{x}_t\rangle\langle \tilde{x}_t, \tilde{x}_t\rangle^{-1}[I_c, 0^{c\times n}]' = T(\langle \tilde{x}_{t+1,1} - \tilde{x}_{t,1}, \tilde{x}_{t,1}\rangle \langle \tilde{x}_{t,1}, \tilde{x}_{t,1}\rangle^{-1}) + \mathrm{o}_{\mathrm{P}}(1)$$

Note that due to the transformation of the system this $(1,1)$ block of the A-matrix contains the c eigenvalues of maximum modulus. This completes the proof. □

Appendix B. Gap metric and simulated systems

The gap metric is defined as follows. Let H be a Hilbert space and let M and N be two closed subspaces of H. Then the *gap* Θ between M and N is defined as follows:

$$\Theta(M,N) = \max\left(\sup_{x\in M, \|x\|=1} \|(I-Q)x\|, \sup_{x\in N, \|x\|=1} \|(I-P)x\|\right),$$

where Q denotes the orthogonal projection onto N and P is the orthogonal projection onto M. In the definition of the gap metric, $\|x\|$ denotes the norm induced by the inner product on H.

The simulated systems are taken from Saikkonen and Luukkonen (1997) and are the following three-dimensional VARMA(1,1) processes:

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

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

$$C_\gamma = \begin{pmatrix} -0.816 & -0.657 & -0.822 \\ -0.624 & -0.785 & 0.566 \\ -0.488 & 0.475 & 0.174 \end{pmatrix}, \tag{B.2}$$

$$\Sigma = \begin{pmatrix} 0.47 & 0.20 & 0.18 \\ 0.20 & 0.32 & 0.27 \\ 0.18 & 0.27 & 0.30 \end{pmatrix} \tag{B.3}$$

Table 4
Parameter values ϕ_i for Schemes 1–3

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

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

$$N^{-1} = \begin{pmatrix} -0.29 & -0.47 & -0.57 \\ -0.01 & -0.85 & 1.00 \\ -0.75 & 1.39 & -0.55 \end{pmatrix}. \tag{B.4}$$

The three sets of parameters ϕ_i are given in Table 4.

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

References

Aoki, M., 1990. State Space Modelling 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. Ph.D. Thesis, TU Wien.

Bauer, D., 2002. Comparing the CCA subspace method to pseudo maximum likelihood methods for the case of no exogenous inputs. Journal of Time Series Analysis, submitted for publication.

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

Bauer, D., Wagner, M., 2000a. Subspace algorithm cointegration analysis—an application to interest rate data. In: Nunez-Anton, V., Ferreira, E. (Eds.), Proceedings of the 15th International Workshop in Statistics, Bilbao, Spain, pp. 146–151.

Bauer, D., Wagner, M., 2001. A canonical form for unit root analysis in the state space framework. Journal of Econometrics, submitted for publication.

Bauer, D., Wagner, M., 2000b. Subspace algorithm cointegration analysis-an application to interest rate data, in: Nunez-Anton, V., Ferriera, E. (Eds.), Proceedings of the 15th International Workshop in Statistics, 146–151. Bilbao, Spain.

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

Bewley, R., Yang, M., 1995. Tests for cointegration based on canonical correlation analysis. Journal of the American Statistical Association 90, 990–996.

Chatelin, F., 1983. Spectral Approximation of Linear Operators. Academic Press, New York.

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

Engle, R.F., Granger, C.W.J., 1987. Co-integration and error correction: representation, estimation and testing Econometrica 55, 251–276.

Hannan, E., Deistler, M., 1988. The Statistical Theory of Linear Systems. Wiley, New York.

Johansen, S., 1995. Likelihood-based Inference in Cointegrated Vector Auto-regressive Models. Oxford University Press, Oxford.

Lai, T.L., Wei, C.Z., 1982a. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. The Annals of Statistics 40, 154–166.

Lai, T.L., Wei, C.Z., 1982b. Asymptotic properties of projections with applications to stochastic regression problems. Journal of Multivariate Analysis 12, 346–370.

Larimore, W.E., 1983. System identification, reduced order filters and modelling via canonical variate analysis. In: Rao, H.S., Dorato, P. (Eds.), Proceedings of the 1983 American Control Conference, Vol. 2, IEEE Service Center, Piscataway, NJ, pp. 445– 451.

Lütkepohl, H., Saikkonen, P., 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. Ph.D. Thesis, TU Wien.

Phillips, P.C.B., 1991. Optimal inference in cointegrated systems. Econometrica 59, 283–306.

Phillips, P.C.B., 1995. Fully modified least squares and vector autoregression. Econometrica 63, 1023–1078.

Poskitt, D.S., 2000. Strongly consistent determination of cointegrating rank via canonical correlations. Journal of Business and Economic Statistics 18, 77–90.

Saikkonen, P., 1992. Estimation and testing of cointegrated systems by an autoregressive approximation. Econometric Theory 8, 1–27.

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

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

Stock, J.H., Watson, M.W., 1988. Testing for common trends. Journal of the American Statistical Association 83, 1097–1107.

Van Overschee, P., DeMoor, B., 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., 1999. VAR Cointegration in VARMA Models, Economics Series 65. Institute for Advanced Studies, Vienna.

Yap, S.F., Reinsel, G.C., 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.

IDENTIFICATION OF STATE SPACE SYSTEMS WITH CONDITIONALLY HETEROSKEDASTIC INNOVATIONS

Dietmar Bauer *,[1]

* *Institute for Econometrics, Operations Research and System Theory, TU Wien, Argentinierstr. 8, A-1040 Vienna, Austria*

Abstract: In this paper consistency of estimates of linear dynamic systems obtained by using subspace algorithms under quite general assumptions on the innovations are derived. The assumptions include i.a. GARCH type of errors as well as E-GARCH. Also the consistent estimation of the model for the conditional variance is discussed. A small simulation study shows the potential of subspace algorithms in the context of GARCH modelling in comparison with the optimization based method implemented in MATLAB.

Keywords: subspace methods, GARCH models, finance, asymptotic properties

1. INTRODUCTION

The concept of heteroskedastic innovations has been introduced in the analysis of financial time series to explain the phenomenon of volatility clustering: Periods of high fluctuations alternate with periods of low fluctuation, which can be modelled via introducing a dependence of the conditional variances of the innovations. As a second property, GARCH models also helped to explain the 'fat tails' often observed in financial time series. The conditional first two moments build the basis of the most prominent portfolio selection methods, which are based on the assumption, that the investor measures his benefit using expected returns and his risk using the variance. Thus a model for the conditional first two moments is the core of any investment strategy building on these assumptions.

Since the introduction of ARCH models by (Engle, 1982) a number of different algorithms for the estimation have been proposed. Most of these procedures resort to optimization of some criterion function, such as the likelihood or the one step ahead prediction error. It is well known, that the prediction error approach neglecting the ARCH property of the errors leads to preliminary estimates, which are consistent but not efficient in the presence of ARCH effects (cf. e.g. Gourieroux, 1997). Also the asymptotic properties of maximum likelihood estimates in the ARMA case are known (cf. e.g. Gourieroux, 1997, for a discussion). However, in all situations, where the optimization of the criterion function is performed using standard numerical methods, the question of initial estimates is virulent. Especially in a multivariate context a good initial estimate is needed in order to achieve a low probability of being trapped in a local minimum. In the conventional homoskedastic case, where the conditional variance of the innovations is constant, it has been shown in (Bauer, 2000) that a particular subspace algorithm sometimes called `CCA`, which has been proposed by (Larimore, 1983), asymptotically is equivalent to a generalized pseudo maximum likelihood estimate, i.e. optimizing the Gaussian likelihood. Here equivalent means, that square root sample size times the difference of the two estimates converges to zero almost sure, so that the estimates tend to the same asymptotic distribution. In this paper it is shown, that the subspace

[1] Support by the Austrian FWF under the project number P14438-INF is gratefully acknowledged.

estimates possess some robustness properties with respect to the assumptions on the innovations.

2. MODEL SET AND ASSUMPTIONS

This paper deals with finite dimensional, discrete time, time invariant, linear, dynamical state space systems of the form

$$x_{t+1} = Ax_t + K\varepsilon_t, \quad y_t = Cx_t + \varepsilon_t \qquad (1)$$

where y_t denotes the s-dimensional observed output, x_t the n-dimensional state and ε_t the s-dimensional innovation sequence. $A \in \mathbb{R}^{n\times n}, K \in \mathbb{R}^{n\times s}, C \in \mathbb{R}^{s\times n}$ are real matrices. Note, that it is not assumed, that y_t is univariate. Throughout the paper it is assumed, that the system is stable, i.e. all the eigenvalues of A are assumed to lie within the open unit disc, and strictly minimumphase, i.e. all the eigenvalues if $A - KC$ are assumed to lie within the unit circle.

It is well known (cf. e.g. Hannan and Deistler, 1988) that state space models and ARMA models are just two representations of the same mathematical object, namely the transfer function: It is easy to verify (using some mild assumptions on the noise sequence ε_t) that one solution to the difference equation given above is of the form

$$y_t = \varepsilon_t + \sum_{j=1}^{\infty} K(j)\varepsilon_{t-j}$$

where $K(j) = CA^{j-1}K, j > 0$ and the infinite sum corresponds to a.s. convergence (or limit in mean square, according to the assumptions imposed upon ε_t). The transfer function $k(z)$, where z denotes the backward shift operator, then is defined as $k(z) = I + zC(I - zA)^{-1}K$. Further let $M(n)$ denote the set of all transfer functions of McMillan degree equal to n fulfilling the stability and the strict minimumphase assumption. $k(z) \in M(n)$ is a rational function in z seen as a complex variable. Therefore the transfer function has a representation as an ARMA system according to $k(z) = a^{-1}(z)b(z)$. A more detailed discussion on the relation between ARMA and state space systems can be found in (Hannan and Deistler, 1988).

The solution y_t as given above is stationary, if the noise ε_t is a stationary sequence. This statement holds both in the weak sense and the strict stationary setting. Throughout this paper it will always be assumed, that ε_t is a martingale difference sequence with respect to the sequence of increasing sigma fields $\mathcal{F}_t$, i.e. $\mathbb{E}\{\varepsilon_t|\mathcal{F}_{t-1}\} = 0$. Furthermore it is assumed, that ε_t is ergodic and of finite fourth moment, i.e. $\mathbb{E}\varepsilon_{t,i}^4 < \infty$, where the notation indicates the i-th component of ε_t. It is also assumed, that $\lim_{k\to\infty} \mathbb{E}\{\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-k}\} = \Sigma = \mathbb{E}\varepsilon_t\varepsilon_t'$ a.s. This property is sometimes referred to as *linear regularity*.

3. SUBSPACE ALGORITHMS

The subspace algorithm investigated in this paper is the CCA method proposed by (Larimore, 1983). Up to now a great number of results exist only for the case, where the innovations also fulfill $\mathbb{E}\{\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}\} = \Sigma$, i.e. where no heteroskedasty is present. For the CCA case consistency has been shown in (Deistler *et al.*, 1995), asymptotic normality in (Bauer *et al.*, 1999) and asymptotic equivalence to pseudo maximum likelihood estimation in (Bauer, 2000). Especially the last result seems to be valuable, since it shows, that the computationally advantageous subspace algorithms are a very good substitute for pseudo maximum likelihood estimation. We also note, that in (Bauer and Wagner, 2001) it is shown, that an adaptation of the algorithm is able to produce (weakly) consistent estimates also in the case of cointegrated processes, where also some tests for the number of cointegrating relations are presented. It is the aim of the present paper to show, that the consistency property of the subspace algorithm holds for an extended range of innovation sequences. The asymptotic distribution is a matter of future research.

The CCA algorithm builds on the properties of the state. In the following we will only give a brief outline. For a more detailed description, also of different subspace algorithms cf. e.g. (Bauer, 1998). Fix two integers f and p. Denoting $Y_{t,p}^- = [y_{t-1}', y_{t-2}', \cdots, y_{t-p}']'$ and $Y_{t,f}^+ = [y_t', y_{t+1}', \cdots, y_{t+f-1}']'$ we obtain the following equation:

$$Y_{t,f}^+ = \mathcal{O}_f\mathcal{K}_pY_{t,p}^- + \mathcal{O}_f(A - KC)^p x_{t-p} + N_{t,f}^+ \qquad (2)$$

where $N_{t,f}^+$ summarizes the effects of the future of the noise, which is orthogonal to the two other terms due to the assumptions on ε_t. Further $\mathcal{O}_f = [C', A'C', \cdots, (A^{f-1})'C']'$, $\mathcal{K}_p = [K, (A - KC)K, \cdots, (A - KC)^{p-1}K]$. Finally let $\langle a_t, b_t\rangle = T^{-1}\sum_{t=p+1}^{T-f} a_t b_t'$. Neglecting the second term in (2), since $(A - KC)^p$ tends to zero for $p \to \infty$, CCA obtains estimates of the system in the following three steps:

- Estimate $\mathcal{O}_f\mathcal{K}_p$ by LS regression in (2) as $\hat{\beta}_{f,p} = \langle Y_{t,f}^+, Y_{t,p}^-\rangle\langle Y_{t,p}^+, Y_{t,p}^-\rangle^{-1}$.
- $\hat{\beta}_{f,p}$ will be of full rank in general, whereas $\mathcal{O}_f\mathcal{K}_p$ is of rank n, where n denotes the system order. Thus approximate

$$\langle Y_{t,f}^+, Y_{t,f}^+\rangle^{-1/2}\hat{\beta}_{f,p}\langle Y_{t,p}^-, Y_{t,p}^-\rangle^{1/2} = \hat{U}\hat{\Sigma}\hat{V}' = \hat{U}_n\hat{\Sigma}_n\hat{V}_n' + \hat{R}_n$$

to obtain estimates $\hat{\mathcal{O}}_f = \langle Y^+_{t,f}, Y^+_{t,f}\rangle^{1/2}\hat{U}_n\hat{\Sigma}_n$ and $\hat{\mathcal{K}}_p = \hat{V}'_n\langle Y^-_{t,p}, Y^-_{t,p}\rangle^{-1/2}$. Here $\hat{U}\hat{\Sigma}\hat{V}'$ denotes the SVD of

$$\langle Y^+_{t,f}, Y^+_{t,f}\rangle^{-1/2}\hat{\beta}_{f,p}\langle Y^-_{t,p}, Y^-_{t,p}\rangle^{1/2}.$$

Thus e.g. $\hat{\Sigma}$ is the diagonal matrix containing the singular values ordered in decreasing size as diagonal entries. $\hat{U}_n \in \mathbb{R}^{fs\times n}, \hat{V}_n \in \mathbb{R}^{ps\times n}$ and $\hat{\Sigma}_n \in \mathbb{R}^{n\times n}$ correspond to the submatrices obtained by neglecting the singular values numbered $n+1$ and higher. Therefore in this step the order is specified.

- Given the estimate $\hat{\mathcal{K}}_p$ from the second step the state is estimated as $\hat{x}_t = \hat{\mathcal{K}}_p Y^-_{t,p}$ and the system matrices are obtained using least squares regressions in the system equations (1), where the estimated state takes the place of the state.

Estimation of the order can be performed using the information contained in the estimated singular values in a number of different ways (for a discussion see Bauer, 1998, Chapter 5). Here we will deal with the criterion SVC. Let

$$SVC(n) = \hat{\sigma}^2_{n+1} + \frac{C_T d(n)}{T}$$

where $d(n) = 2ns$ denotes the number of parameters and $C_T > 0, C_T/T \to 0$ denotes a penalty term. Here $\hat{\sigma}_i$ denotes the estimated singular values ordered decreasing in size. In the homoskedastic case it is known, that a penalty such that $C_T/(fp\log T) \to \infty$ leads to almost sure (a.s.) consistent estimates of the order $\hat{n} = \arg\min SVC(n), 0 \le n \le H_T, H_T = O((\log T)^a), a < \infty$.

4. RESULTS

The key to the results in this section lies in the uniform convergence of the estimated covariance sequence. The conditions in Theorem 5.3.2. of (Hannan and Deistler, 1988) require, that in order for the sequence of covariance estimates to converge uniformly of order $O(Q_T)$ the noise has to be homoskedastic. Here $g_T = O(f_T)$ means that there exists a constant M, such that $g_T/f_T < M$ a.s. and $Q_T = \sqrt{\log\log T/T}$. However, equation (5.3.7.) in the same book provides the result, that if the limiting covariance sequence is replaced with a sequence, where the innovation variance Σ is replaced with $T^{-1}\sum_{t=1}^{T}\varepsilon_t\varepsilon'_t$ the same results holds under weaker assumptions. This enables the results in the next theorem:

Theorem 1. Let the process $\{y_t\}$ be generated by a stable, strictly minimumphase system $k(z) \in M(n)$, where the innovation process is an ergodic, strictly stationary martingale difference sequence satisfying $\mathbb{E}\{\varepsilon_t|\mathcal{F}_{t-1}\} = 0, \mathbb{E}\varepsilon^4_{t,i} < \infty$ and $\lim_{k\to\infty}\mathbb{E}\{\varepsilon_t\varepsilon'_t|\mathcal{F}_{t-k}\} = \Sigma = \mathbb{E}\varepsilon_t\varepsilon'_t$. a.s. Let $(\hat{A}, \hat{K}, \hat{C})$ denote the estimates obtained via the CCA subspace algorithm using the true order n for the estimation, which have been transformed to the corresponding echelon canonical form. Then the following statements hold:

i) $I + z\hat{C}(I - z\hat{A})^{-1}\hat{K} \to k(z)$ a.s. for each fixed $z = \exp(i\omega)$, if $f \ge n, p = p(T) \to \infty, \max\{f,p\} = O((\log T)^a), a < \infty$. That is, the transfer function is estimated consistently.

ii) Let (A_0, K_0, C_0) denote the representation of the system in the echelon canonical form. Then for $k(z)$ in the generic neighbourhood of the echelon canonical form and if $p \ge -d\log T/(2\log\rho_0), d > 1$

$$\max\{\|\hat{A}-A_0\|, \|\hat{K}-K_0\|, \|\hat{C}-C_0\|\} = O(Q_T)$$

Here $0 < \rho_0 < 1$ denotes the maximal modulus of the eigenvalues of $A_0 - K_0C_0$.

iii) The order estimate $\hat{n}$ obtained by minimizing the SVC criterion is strongly consistent, i.e. $\hat{n} \to n$ a.s., for $C_T/(fp\log T) \to \infty$.

The three parts of the theorem state that with regard to consistency there is no major difference between the homoskedastic and the heteroskedastic case, as long as stationarity is preserved: The subspace estimates still are consistent, the estimation error can be bounded as in the homoskedastic case. Note that the result ii) has the form of a law of the iterated logarithm, except that the constant is not evaluated exactly. This result is only given for the generic neighbourhood of the echelon form, however, using overlapping forms (see e.g. Hannan and Deistler, 1988, Chapter 2) one can show, that an equivalent error bound is indeed valid for all $k \in M(n)$. The last result shows, that also the order estimation can be performed as in the homoskedastic case. This essentially means, that one can use the same code as in the homoskedastic case for the identification irrespective if the system is homo- or heteroskedastic. The derivation of the asymptotic distribution and the investigation of the comparison with prediction error methods is left as a topic of future research.

The theorem imposes an order of convergence for the integer p as a function of the sample size, which is only needed for the derivation of the error bound. This order of convergence includes system dependent quantities and thus might be seen as useless in practice. However, Theorem 6.6.3 in (Hannan and Deistler, 1988) shows, that if p is chosen as $\lfloor d\hat{p}_{AIC}\rfloor$ for $d > 1$, where $\lfloor x\rfloor$ denotes the largest integer smaller than x and where $\hat{p}_{AIC}$ is chosen as the order estimate of a long autoregression for approximating y_t using AIC, then p fulfills the assumption of part ii) a.s.

for large T.[2] Thus an algorithm using this choice of the integer p will lead to consistent estimates, where also the error bound on the estimation error holds.

In comparison to the homoskedastic case the theorem leaves out two important results: The asymptotic distribution of the estimates is not analyzed and secondly the consistency result should also be extended to the unit root case. Both questions are topics of future research.

4.1 *ARCH(p) innovations*

(Engle, 1982) introduced the class of ARCH(p) models, where the conditional variance h_t of the univariate innovations ε_t is modelled as a linear function of the last p squares of the innovations:

$$h_t = c + \sum_{j=1}^{p} a_j \varepsilon_{t-j}^2$$

where ε_t conditional on $\mathcal{F}_{t-1}$, the sigma algebra spanned by $\{\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots\}$, is identically Gaussian distributed with mean zero and variance h_t. Here $0 \leq a_j, c > 0$ is assumed. In order for the process to be strictly stationary with finite variance it is assumed that $\sum_{j=1}^{p} a_j < 1$. It follows from (Bougerol and Picard, 1992) that in this situation the process ε_t is ergodic. Furthermore it is assumed for ψ being equal to the matrix with typical element $\psi_{i,j} = a_{i+j} + a_{i-j}$, where $a_j = 0, j \notin \{1, \cdots, p\}$ that $3(a_1, \cdots, a_p)(I - \psi)^{-1}(a_1, \cdots, a_p)' < 1$. This condition is sufficient for the existence of fourth moments (see e.g. Gourieroux, 1997, Exercise 3.4). Thus the system estimates obtained using subspace methods are consistent. The assumption on Gaussianity of $\varepsilon_t | \mathcal{F}_{t-1}$ is not necessary and can be replaced by other assumptions, which imply the existence of the fourth moment of the process ε_t.

From the discussion given above it follows, that a regression of $\hat{\varepsilon}_t^2$ onto $[1, \hat{\varepsilon}_{t-1}^2, \cdots, \hat{\varepsilon}_{t-p}^2]$ results in consistent estimates of the model for the conditional variance. This follows from the finiteness of the fourth moment, the strict stationarity and ergodicity of ε_t and the consistency of $\hat{\varepsilon}_t$ for ε_t.

4.2 *GARCH(p,q) innovations*

(Bollerslev, 1986) extended the ARCH(p) specification to also include MA terms, leading to GARCH(p,q) systems: Let the conditional variance be denoted as $h_t = \mathbb{E}\{\varepsilon_t^2 | \mathcal{F}_{t-1}\}$, then the model assumes that

$$h_t = c + \sum_{j=1}^{p} a_j \varepsilon_{t-j}^2 + \sum_{j=1}^{q} b_j h_{t-j}$$

where again $c > 0, a_j \geq 0, b_j \geq 0$. (Bougerol and Picard, 1992) show, that the process ε_t is strictly stationary and ergodic, if $h_t^{-1/2}\varepsilon_t$ is identically standard normally distributed and if $\sum_{j=1}^{p} a_j + \sum_{j=1}^{q} b_j < 1$. In this case also the second moments exist and the process is also wide sense stationary. It remains to find a bound for the fourth moment: Conditions for this to hold are fairly complicated and can be found in (He and Teräsvirta, 1999). Thus in this case the result above shows the consistency of the transfer function estimates. Therefore also the estimated residuals are consistent. The estimation of the model for the innovations leads again to an ARMA model with heteroskedastic innovations. Thus in order to apply the results in this paper, the existence of an eighth moment has to be assumed: Although it follows from (Hannan and Deistler, 1988) that also in this case finite fourth moments are sufficient to achieve a uniform convergence of the sample covariances, no bound on the order of convergence can be given and thus the arguments given above fail for $p \to \infty$. Holding f and p fixed leads to consistent estimates in the sense, that the estimated system matrices converge to some constants a.s., but the estimated system will be asymptotically biased, where the bias depends on the magnitude of ρ_0^p.

4.3 *E-GARCH processes*

As a final example consider the exponential GARCH models considered in (Nelson, 1991): In order to guarantee positivity of the conditional variances the following model has been introduced:

$$\log h_t = \alpha_t + \sum_{j=1}^{\infty} \beta_k g(z_{t-k})$$

Here $\varepsilon_t = z_t h_t^{1/2}$, where z_t is assumed to be i.i.d. with mean zero and variance unity and α_t is a deterministic sequence e.g. constant. The function g is assumed to be of the lin-lin type: $g(z) = \theta z + \gamma(|z| - \mathbb{E}|z|)$ Further the distribution of z_t is assumed to be of the GED type with tail thickness parameter $\nu > 1$. Under these assumptions it follows that $\exp(-\alpha_t)\varepsilon_t$ is strictly stationary and ergodic with finite moments of all orders. Furthermore $\mathbb{E}\{\varepsilon_t^2 | \mathcal{F}_{t-k}\} \to \sigma^2$ a.s. for $k \to \infty$. Thus the assumptions of Theorem 1 are fulfilled and the subspace estimates are a.s. consistent.

5. SIMULATIONS

In this section a simple simulation study compares the properties of the subspace estimates to the

[2] This does not hold for AR(p) systems. In this case $\rho_0 = 0$ and $\hat{p}_{AIC}$ stays bounded. However, all results remain true.

estimates obtained by using a likelihood approach. The procedure, which serves as a benchmark, is the one provided in the MATLAB toolbox. The investigated properties are the accuracy of the estimates and the computation times as measured by the MATLAB function profile. It should be noted, that both the ML procedure as well as the subspace algorithm have not been trimmed to have minimum computations and there seems to be much potential of improving the subspace algorithms, but on the other hand also the ML approach uses some consistency checks on the data, which increase the computations as well.

The system we will use is an ARMA model with GARCH(1,1) innovations and thus very simple. The specification in full detail is as follows:

$$y_t = 0.8y_{t-1} + \varepsilon_t + 0.3\varepsilon_{t-1}$$
$$h_t = 0.3h_{t-1} + 0.2\varepsilon_{t-1}^2 + 1$$

The conditional distribution of the innovations is Gaussian. The processes are generated using the MATLAB function garchsim. For each sample size $T = 200, T = 500, T = 1000$ and $T = 2000$ a total of 1000 time series have been generated and the system estimated using the function garchfit and the correct specification. Also the subspace procedure is used with $f = p = 2\hat{p}_{AIC}$, where $\hat{p}_{AIC}$ denotes the lag length selected by the AIC criterion.

The summary statistics of the estimates can be seen in Table 2 for the ML procedure and in Table 3 for the subspace procedure: The better accuracy of the ML method is clearly visible, however the difference does not seem to be striking for the ARMA model for the output series. Especially for $T = 2000$ the difference in accuracy is minor, except for the occurrence of some outliers in the subspace case. The estimates for the variance model achieved using subspace procedures however, are not very reliable, and this is in particular true for the estimated zeros of the variance model. Even at sample size $T = 2000$ there seems to be a downward bias in the estimates. These facts are also visible in Figure 1: The upper plot here shows a scatter plot of the estimated autoregressive parameters, the lower plot shows the scatter plot for the zero of the estimated variance models, both for sample size $T = 2000$. The upper plot shows a high correlation between the estimates, whereas the lower plot indicates a number of aberrant estimates for the subspace algorithms.

Also in a number of cases some outliers occur, which inflate the estimated variability. This is the reason for using robust estimates of the root mean square and the mean. It should also be mentioned, that in a number of cases the MATLAB routine garchfit crashed, giving no resulting

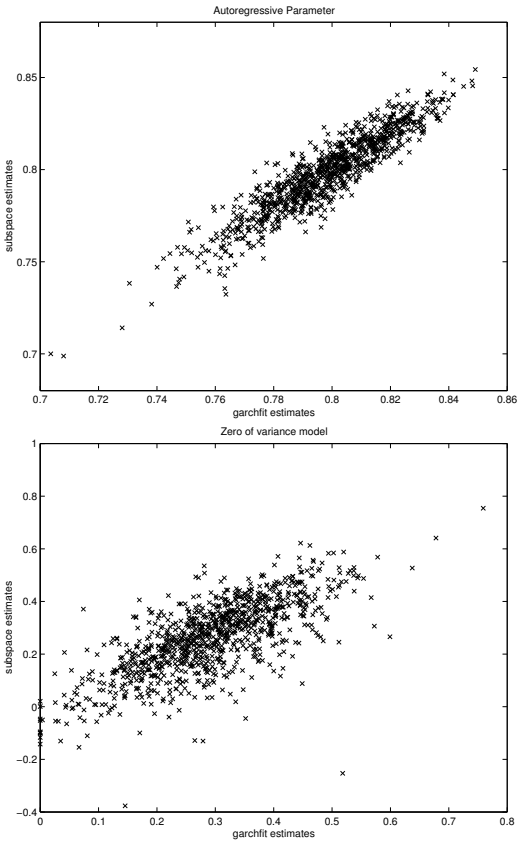


Fig. 1. Upper plot: estimates of the autoregressive parameter of the conditional mean model estimated using garchfit (x-axis) versus the estimates obtained using the subspace method (y-axis) for sample size $T = 2000$ in 1000 trials. Lower plot: analogous picture for the estimated zero of the variance model.

Method	$T = 200$	$T = 500$	$T = 1000$	$T = 2000$
garchfit	3.52	3.77	5.17	6.38
subspace	0.61	0.65	0.74	0.89
Quotient	5.77	5.8	7.0	7.2

Table 1. Mean computation time per identification experiment in seconds for the various sample sizes.

system at all. These cases have been taken out of the simulations, leading to some bias in the comparison.

Finally the computational time can be analyzed, which clearly shows a huge advantage for the (not even optimized) subspace methods (see Table 1). It is clearly visible, that the subspace method requires only a fraction of computations, while still providing reasonable estimates. The main conclusion of the small simulation study is that the subspace algorithms provide relatively good initial estimates for a subsequent pseudo ML approach in terms of the asymptotic statistical properties, while still keeping the amount of computations required at a low level.

6. CONCLUSIONS

In this paper the asymptotic properties of estimates of state space models using subspace methods with heteroskedastic innovations are investigated. Consistency is shown and a bound on the obtainable order of consistency is provided. The result is stated in a general fashion such that it applies for a wide range of models for the heteroskedasticity, including ARCH(p), GARCH(p,q) and E-GARCH(p,q) models. This shows, that the standard subspace algorithms provide consistent estimates of the system also in situations, where the model for the conditional variance might be doubted. This of course is due to the fact, that the subspace algorithms are based mainly on regression techniques, which are robust with respect to the variance structure of the innovations. With respect to the estimation of the model for the conditional variances consistency can be achieved in the ARCH(p) case, whereas no comparable results are given for the general case. A simulation study compares the estimates with the estimates obtained using the GARCH toolbox implemented in `MATLAB` both with respect to accuracy and computation time. The loss of efficiency in the estimation of the model for the heteroskedasticity is clearly visible, however, the accuracy of the model for the conditional mean seems to be acceptable. Finally the main power of subspace algorithms, namely their low computational load is demonstrated in comparison with a GARCH routine implemented in the `MATLAB` GARCH toolbox.

T	Meas.	a	b	a_v	b_v	σ^2
	True	0.8	-0.3	0.2	0.3	2.0
200	Mean	0.776	-0.283	0.188	0.293	1.97
	RMSE	0.070	0.111	0.100	0.261	0.28
500	Mean	0.788	-0.288	0.194	0.290	1.99
	RMSE	0.041	0.068	0.064	0.201	0.18
1000	Mean	0.796	-0.297	0.198	0.277	1.99
	RMSE	0.026	0.045	0.045	0.147	0.12
2000	Mean	0.798	-0.298	0.199	0.291	1.99
	RMSE	0.019	0.032	0.032	0.104	0.09

Table 2. Summary of estimation results for the ARMA model for the conditional mean (parameters a and b) and the ARMA model for the conditional variance (parameters a_v and b_v) and implied stationary variance σ^2 for various sample sizes and for `garchfit`. For each sample size the trimmed mean and the trimmed root mean square error (RMSE) neglecting the extreme 5%, are calculated.

T	Meas.	a	b	a_v	b_v	σ^2
	True	0.8	-0.3	0.2	0.3	2.0
200	Mean	0.778	-0.286	0.155	0.095	1.97
	RMSE	0.074	0.121	0.111	0.391	0.27
500	Mean	0.787	-0.286	0.177	0.196	1.99
	RMSE	0.044	0.075	0.078	0.264	0.18
1000	Mean	0.795	-0.296	0.187	0.233	1.99
	RMSE	0.031	0.054	0.065	0.219	0.13
2000	Mean	0.798	-0.298	0.192	0.265	1.99
	RMSE	0.020	0.036	0.042	0.132	0.09

Table 3. Summary of estimation results for the subspace procedure.

7. REFERENCES

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. (2000). Asymptotic efficiency of the CCA subspace method in the case of no exogenous inputs. *Submitted to Journal of Time Series Analysis.*

Bauer, D. and M. Wagner (2001). Estimating cointegrated systems using subspace algorithms. *to appear in Journal of Econometrics.*

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.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Bougerol, P. and N. Picard (1992). Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics* **52**, 115–127.

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

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of u.k. inflation. *Econometrica* **50**, 987–1008.

Gourieroux, Ch. (1997). *ARCH Models and Financial Applications.* Springer Series in Statistics.

Hannan, E. J. and M. Deistler (1988). *The Statistical Theory of Linear Systems.* John Wiley. New York.

He, C. and T. Teräsvirta (1999). Fourth moment structure of the GARCH(p,q) process. *Econometric Theory* **15**(6), 824-846.

Larimore, W. E. (1983). System identification, reduced order filters and modelling via canonical variate analysis. In: *Proc. 1983 Amer. Control Conference 2.*, Piscataway, NJ. pp. 445–451.

Nelson, D. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* **59**, 347–370.

Peternell, K., W. Scherrer and M. Deistler (1996). Statistical analysis of novel subspace identification methods. *Signal Processing* **52**, 161–177.