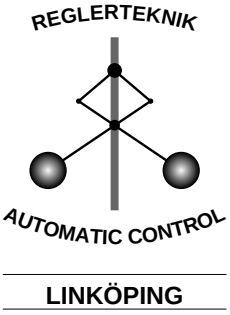
Order estimation for subspace methods

Dietmar Bauer

Department of Electrical Engineering
Linköping University, SE-581 83 Linköping, Sweden
WWW: http://www.control.isy.liu.se
Email: Dietmar.Bauer@tuwien.ac.at

June 6, 2000



Report no.: LiTH-ISY-R-2263
Submitted to Automatica

Order estimation for subspace methods

Dietmar Bauer*
Department of Electrical Engineering
Linköping University, SE-581 83 Linköping, Sweden
WWW: http://www.control.isy.liu.se
Email: Dietmar.Bauer@tuwien.ac.at

June 6, 2000

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 exogeneous 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 latter 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, 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, 1975) 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, 1975; Rissanen, 1978) and the effects of the choice of the penalty term are well understood (see e.g. Hannan and 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.

*Parts of this work have been done while the author was holding a post-doc position at the Department f. Automatic Control, Linköping University, Linköping, Sweden, financed by the EU TMR project 'SI'.

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 more details on this. However it has been observed, that this method seems to be relatively sensitive to the choice of certain user parameters in (Bauer, 1998). In this thesis also an alternative is analysed, which is a small adaptation of the criterion given in (Peternell, 1995), which 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. This 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 mostly 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. For this reason it is believed, that estimating the order is more appropriate than testing for the rank in the present setup.

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 an assumptions

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

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

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. 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 be 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 state space representation of the form (1) fulfilling 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:

$$\begin{array}{cc} \mathbb{E}\{\varepsilon_t|\mathcal{F}_{t-1}\} = 0, & \mathbb{E}\{\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}\} = \Omega = \mathbb{E}\varepsilon_t\varepsilon_t' > 0 \\ \mathbb{E}\{\varepsilon_{t,a}\varepsilon_{t,b}\varepsilon_{t,c}|\mathcal{F}_{t-1}\} = \omega_{a,b,c}, & \mathbb{E}\{\varepsilon_{t,a}^4\} < \infty \end{array} \tag{2}$$

Here $\mathbb{E}$ denotes expectation, $\mathcal{F}_t$ denotes the σ-algebra spanned by $(y_s, s \le t)$ and $\varepsilon_{t,a}$ denotes the a-th 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.

Assumptions 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 fulfilling the assumptions stated in equation (2) and being independent of ε_t, and where $\|K_u(j)\| \leq c_u \rho_u^j$ for some $0 < c_u < \infty, 0 < \rho_u < 1$. Furthermore $\Phi_u(\omega) = (\sum_{j=0}^{\infty} K_u(j)e^{i\omega j})(\sum_{j=0}^{\infty} K_u(j)e^{i\omega j})'$ is assumed to fulfill $0 < \underline{c}I \leq \Phi_u \leq \bar{c}I < \infty$ for $-\pi < \omega \leq \pi$.*

Assumptions 2 (`MOESP` type of procedures) *The input process $(u_t; t \in \mathbb{Z})$ is of the form $u_t = v_t + \sum_{j=1}^{h} c_j e^{i\lambda_j t}$ where v_t fulfills assumptions 1 and the random variables $c_j \in \mathbb{R}^m$ have zero mean and finite variances and are such that the corresponding process u_t is real valued. 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. The reason for this lies in the fact, that for this class of procedures a necessary condition for consistency is that the integer parameter p tends to infinity (see below for details). This amounts to the condition on the input to be of infinite persistency. For the `MOESP` type of procedures note that the assumptions are identical to the assumptions imposed in the proof of the asymptotic normality in (Bauer and Jansson, 2000). Note however, that this only occurs, since no minimal conditions are given in this paper. In fact 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 there 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 $Y_{t,f}^+ = [y_t', y_{t+1}', \cdots, y_{t+f-1}']'$ 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 $Z_{t,p}^- = [y_{t-1}', u_{t-1}', \cdots, y_{t-p}', u_{t-p}']'$. 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 $\mathcal{O}_f' = [C', A'C', \cdots (A^{f-1})'C']$ and $\mathcal{K}_p = [[K, B - KD], (A - KC)[K, B - KD], \cdots, (A - KC)^{p-1}[K, B - KD]]$. 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 for us and thus we refer 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 $\hat{\Gamma}_p^- = \frac{1}{T}\sum_{t=p+1}^{T} Z_{t,p}^-(Z_{t,p}^-)'$ denotes the sample variance of $Z_{t,p}^-$. Further $\hat{\Gamma}_p^{-,\Pi} = \hat{\Gamma}_p^{-1} - \hat{\Gamma}_{u,z}\hat{\Gamma}_u^{-1}\hat{\Gamma}_{u,z}$. 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. 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 diferent 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 $\langle a_t, b_t\rangle = \frac{1}{T}\sum_{t=p+1}^{T-f} a_t b_t'$. Inserting the estimated state into the system equations (1) one obtains estimates of (A, B, C, D) from the least squares solution:

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

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 $X_\circ = W_f^+\mathcal{O}_f\mathcal{K}_p W_p^-$, 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 we resort 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:

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

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

Here $d(n) = nm + ns + sm$ denotes the number of parameters of a state space system of order n. $C(T) > 0, C(T)/T \to 0$ 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` 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 not so important 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:

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

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 $0 \le n \le \min\{fs, p(s+m)\}$. 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 $\hat{K}_p^n = \hat{V}_n'(\hat{W}_p^-)^{-1} Z_{t,p}^-$ 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 and then only additions and multiplications have to performed. Let $\hat{C}_T^{max} = \left[\hat{C}_T^n, \quad \hat{C}_T^{n-max}\right]$ 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 regression 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 leads 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; Peternell, 1995). Some results for the `MOESP` case have been presented in (Bauer, 1999).

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 linearisation of the SVD or SVD related quantities, which will be of importance mainly for the `SVC` case.

Lemma 4.1 *Let $(y_t; t \in \mathbb{Z})$ be generated by a system of the form (1), where the noise fulfills 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 fulfills assumptions 1 then

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

If u_t only fulfills the assumptions 2, then the statement is true for $H_T = M < \infty$.

This lemma follows from (Hannan and Deistler, 1988, Theorem 5.3.2 and Chapter 4). 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 4.2 (Chatelin) *Let $\mathcal{T}_T$ denote a sequence of symmetric, compact operators acting on ℓ^2, which converges in norm to the operator $\mathcal{T}_\circ$. Then it follows, that the set of eigenvalues of $\mathcal{T}_T$ converges to the set of eigenvalues of $\mathcal{T}_\circ$. 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 $\mathcal{T}_\circ$ and let P_i^T and λ_i^T denote the corresponding quantities of $\mathcal{T}_T$. Here for a multiple eigenvalue λ_i° of $\mathcal{T}_\circ$ the quantities P_i^T refer to the orthonormal projection matrix onto the space spanned by the eigenvectors to all eigenvalues of $\mathcal{T}_T$ converging to λ_i°. Then*

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

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 leads 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 and 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 in fact 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 4.1 *Let the process $(y_t; t \in \mathbb{Z})$ be generated by a system of the form (1), where the true system order is equal to $n_\circ$, and where the white noise process $(\varepsilon_t; t \in \mathbb{Z})$ fulfills the assumptions of section 2. If the input fulfills the assumptions 1, then $\max\{f,p\} = o((\log T)^a)$ 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 fulfills the assumptions 2 with $\alpha = f + p - 1$, then for each f and p there exists a set $U_n(f,p,\Phi_u(\omega),\Omega) \subset M_n$ 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 set of assumptions the error in the estimation of the first $f+p-1$ covariances $\gamma_z(j)$ are of order $O(Q_T)$ uniformely due to the Lemma 4.1. The estimation uses the singular values of $\hat{X}_{f,p} = \hat{W}_f^+ \hat{\beta}_z \hat{W}_p^-$, which converges to $X_\circ = W_f^+ \mathcal{O}_f \mathcal{K}_p W_p^-$ a.s. as has been shown e.g. in (Peternell *et al.*, 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

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

The estimation error in each entry of these matrices is of the order $O(Q_T)$ as follows from the Lemma 4.1 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

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

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 n-th 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 we have to show, that the true order $n_\circ$ will be preferred to $n > n_\circ$ asymptotically. Thus consider

$$\begin{aligned} \hat{\sigma}_{n+1} &= \|\hat{W}_f^+ \hat{\beta}_z \hat{W}_p^- - \hat{U}_n \hat{\Sigma}_n \hat{V}_n'\|_2 = \|\hat{U}' \left(\hat{X}_{f,p} - \hat{U}_n \hat{\Sigma}_n \hat{V}_n' \right) \|_2 \\ &\le \|\hat{U}_2 \hat{U}_2' \hat{X}_{f,p} - U_2 U_2' X_\circ\|_2 \end{aligned}$$

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 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

$$\|\hat{U}_2\hat{U}_2'\left(\hat{X}_{f,p} - X_\circ\right)\|_2 \leq \|\hat{U}_2\hat{U}_2'\|_2\|\left(\hat{X}_{f,p} - X_\circ\right)\|_{Fr} = O(Q_T\sqrt{fp})$$

Therefore it remains to obtain a bound on $\|\hat{U}_2\hat{U}_2' - U_2U_2'\|_2$. But this follows from Lemma 4.2, using $\mathcal{T}_T = \hat{X}_{f,p}\hat{X}_{f,p}', \mathcal{T}_0 = X_\circ X_\circ'$ and the exponential decrease in elements in the rows of $U_{n_\circ}$ (cf. Bauer, 1998). From this the result follows, since

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

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

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 fulfills assumptions 1. Note that if no delay is postulated

$$\begin{aligned} \hat{\Omega}_n &= \langle y_t, y_t\rangle - \langle y_t, \begin{bmatrix} \hat{x}_t \\ u_t \end{bmatrix}\rangle\langle\begin{bmatrix} \hat{x}_t \\ u_t \end{bmatrix}, \begin{bmatrix} \hat{x}_t \\ u_t \end{bmatrix}\rangle^{-1}\langle\begin{bmatrix} \hat{x}_t \\ u_t \end{bmatrix}, y_t\rangle \\ &= \langle y_t, y_t\rangle - \langle y_t, Z^-_{t+1,p+1}\rangle\hat{L}_n'(\hat{L}_n\langle Z^-_{t+1,p+1}, Z^-_{t+1,p+1}\rangle\hat{L}_n')^{-1}\hat{L}_n\langle Z^-_{t+1,p+1}, y_t\rangle \\ &= \langle y_t, y_t\rangle - \hat{\bar{\mathcal{H}}}_1\bar{\hat{L}}_n'(\bar{\hat{L}}_n\bar{\hat{L}}_n')^{-1}\bar{\hat{L}}_n\hat{\bar{\mathcal{H}}}_1' \end{aligned}$$

where $\hat{L}_n = \begin{bmatrix} 0 & 0 & \hat{\mathcal{K}}_p(n) \\ 0 & I & 0 \end{bmatrix}, \hat{\mathcal{K}}_p(n) = \hat{V}_n'(\hat{W}_p^-)^{-1}$. Also $\bar{\hat{L}}_n = \hat{L}_n\langle Z^-_{t+1,p+1}, Z^-_{t+1,p+1}\rangle^{1/2}$ and $\hat{\bar{\mathcal{H}}}_1 = \langle y_t, Z^-_{t+1,p+1}\rangle\langle Z^-_{t+1,p+1}, Z^-_{t+1,p+1}\rangle^{-'/2}$.

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 fulfilled 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 $n \geq n_\circ$ 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:

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

Using the matrix inversion lemma for partitioned matrices one obtains

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

where $\hat{P}^\perp_{n_\circ} = I - \bar{\hat{L}}_{n_\circ}'(\bar{\hat{L}}_{n_\circ}\bar{\hat{L}}_{n_\circ}')^{-1}\bar{\hat{L}}_{n_\circ}$ and where $\bar{\hat{L}}_n' = [\bar{\hat{L}}_{n_\circ}', \bar{\hat{L}}_{n,2}']$. Since the second term is a projection operator it has norm one. Thus the essential term is $\hat{\bar{\mathcal{H}}}_1\hat{P}^\perp_{n_\circ}$, which converges to zero,

since $\widehat{\bar{\mathcal{H}}}_1 \to [C,D]\bar{L}_{n_\circ}\bar{L}'_{n_\circ}$ 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 $\hat{\mathcal{K}}_p - \mathcal{K}_p$. It is straightforward to show, using Lemma 4.1 and Lemma 4.2 that there exists a matrix $\hat{S}_T$ such that $\|\hat{S}_T\hat{\mathcal{K}}_p - \mathcal{K}_p\|_2 = O(Q_T\sqrt{p})$. Applying Lemma 4.1 to $\langle y_t, Z^-_{t+1,p+1}\rangle$ this also implies $\|\widehat{\bar{\mathcal{H}}}_1 - [C,D]\bar{L}_{n_\circ}\bar{L}'_{n_\circ}\|_2 = O(Q_T\sqrt{p})$ as well as $\|\hat{P}^{\perp}_{n_\circ} - P^{\perp}_{n_\circ}\|_2 = O(Q_T\sqrt{p})$. Therefore consider

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

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

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

This shows the following result:

Theorem 4.2 *Let the process $(y_t; t \in \mathbb{Z})$ be generated by a system of the form (1), where the true system $(k,l) \in M_n$. Let the noise fulfill the assumptions of section 2 and let the input fulfill the assumptions 1. Let the system be estimated according to the Larimore type of procedure.*

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 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 $x_t = \mathcal{K}_\infty Z^-_{t,\infty}$ need not be generated by a state space equation of order n for $n < n_\circ$, i.e. the matrix $\mathcal{K}_\infty$ might not have the shift invariance structure $\mathcal{K}_{2:\infty} = \bar{A}_n\mathcal{K}_{1:\infty}$ 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 we present three different examples 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

	T	100			1000		
	Est. Order	< 2	2	> 2	< 2	2	> 2
CCA	IVC1	0.00	0.83	0.17	0.00	0.77	0.23
	IVC2	1.00	0.00	0.00	0.82	0.18	0.00
	SVC1	0.00	0.96	0.04	0.00	0.93	0.07
	SVC2	1.00	0.00	0.00	0.86	0.14	0.00
N4SID	IVC1	0.82	0.03	0.15	0.68	0.06	0.26
	IVC2	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
	SVC2	0.45	0.55	0.00	0.04	0.96	0.00

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). $f = p = \hat{p}_{AIC}$ has been used.

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

$$\hat{n} = \max \left\{ n : \log \hat{\sigma}_n > \frac{1}{2}(\log \hat{\sigma}_1 + \log \hat{\sigma}_M) \right\}$$

i.e. the largest integer, which is greater than the arithmetic 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.

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

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

This system has Lyapunov balanced Gramian of roughly $\Sigma = diag(2.55, 1.78)$. The system poles are at $0.25 \pm 0.7984i$ and the zeros at $-0.4 \pm 0.4359i$. In the estimation we use two different weighting schemes: CCA uses $\hat{W}_f^+ = (\hat{\Gamma}_f^+)^{-1/2}$ and the method using $\hat{W}_f^+ = I$ will be labelled N4SID. The indices $f = p = \hat{p}_{AIC}$ are used. For each of the weighting schemes, the order 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. 1000 time series of length 100 and 1000 respectively have been generated and used for estimation. Table 1 shows the results for $T = 100$ and $T = 1000$ 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 we also observe, 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. On the other hand for the N4SID weighting the smaller penalty leads to a high risk of overestimation. Therefore no clear decision

$\hat{W}_f^+$	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

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 and IVC use the penalty term $C(T) = \log T$. The table has been produced using 1000 replications in each case.

about the choice of the penalty has been found. Both choices used in this example are heuristic and not motivated by additional arguments. A theoretical justification seems to be needed.

Next the various order estimation procedures will be tested on an eight order system with poles at $z = 0.8e^{\pm 0.2i\pi}, z = 0.7e^{\pm 0.3i\pi}, z = 0.5e^{\pm 0.5i\pi}, z = e^{\pm 0.4i\pi}$ and zeros at $z = 0.8e^{\pm 0.1i\pi}, z = -0.4755, z = 0.1, z = 0.3, z = 0$. Using this example extensive simulations comparing the four 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 $\mathrm{diag}(6.85, 4.46, 1.08, 0.39, 0.045, 0.015, 0.0004, 0.0002)$ 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 $T = 1000$ 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 6-th order butterworth filter with cutoff frequency 0.5π and the corresponding high pass filter have been incorporated. $f = p = d\hat{p}_{AIC}$ has been used in all cases, where $d = 2$ and $d = 4$ are tried. The average values of the corresponding order estimates are given in Table 2. It can 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 NIC for $d = 4$ are better than the 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 procedures are also compared to the more traditional maximum likelihood based information criteria. In Figure 1 a histogram for the estimated orders using IVC, SVC, AIC and BIC is given. Here $T = 100$ and $f = p = 2\hat{p}_{AIC}$ 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 figure 2 shows the square root of the mean squared error of the estimated transfer function in the angular frequency

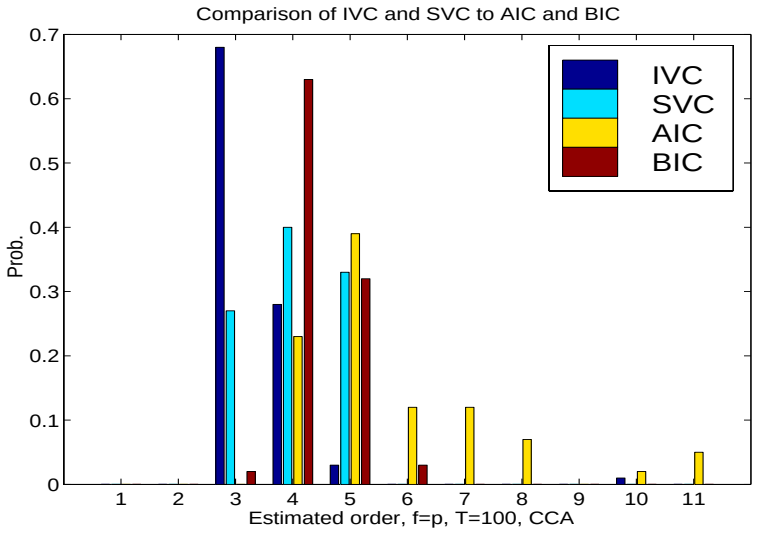


Figure 1: In this figure the order estimates obtained by `SVC` and `IVC` are compared to the estimates obtained in the ML framework using `AIC` and `BIC`. $T = 100$ and $f = p = 2\hat{p}_{AIC}$ are used together with the `CCA` weighting scheme. The plots have been obtained using 100 replications.

range $[0, \pi]$ obtained by `CCA` using $d = 2$ for the various procedures. Here the sample size is equal to $T = 1000$. 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 figure 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 $n = 4$) occurs, 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 comparable mean square error, while choosing smaller orders on average, which might be seen as an advantage for control design. 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 that sensitive, although on average much higher orders are chosen. The corresponding pictures are given in figure 3. Finally also the effect of the weightings on the mean square error is discussed: Figure 4 shows two plots, where the left one refers to $T = 100$ and the order estimate according to `IVC`. 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.

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

$$A = \begin{bmatrix} 0.8 & 0.2 \\ -0.4 & -0.5 \end{bmatrix}, B = \begin{bmatrix} 0 & -1 \\ 1 & 0.5 \end{bmatrix}, K = \begin{bmatrix} 1.5 & 0 \\ -0.2 & -0.8 \end{bmatrix}, D = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and $C = I_2$, the two-by-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 $-0.1000 \pm 0.9327i$. Figure 5 shows the probabilities of the order estimates for different sample sizes, where the input is i.i.d uniformely distributed white noise with zero mean and unit variance. It can be seen, that `IVC`

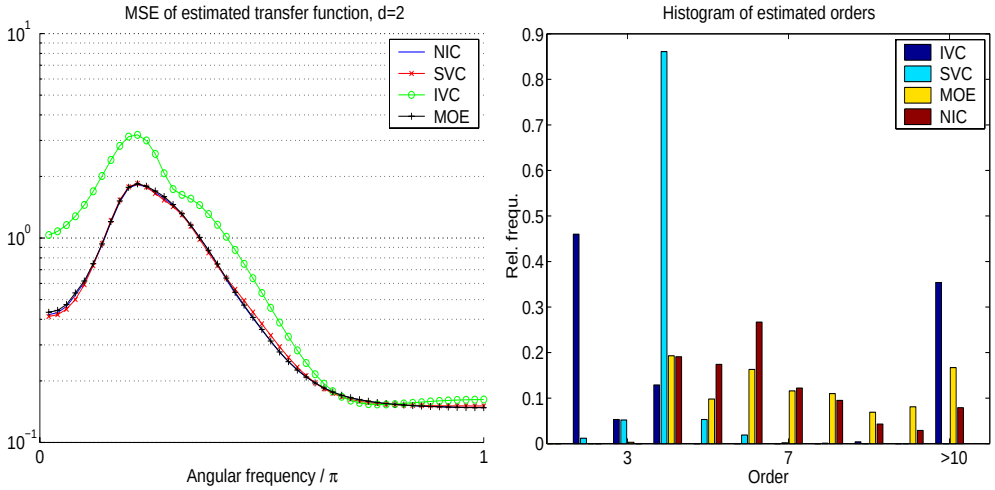


Figure 2: These plots show the result of the simulation for $T = 1000$. The left picture shows the average mean square error of the transfer function at 50 equally spaced frequencies in the angular frequency range $\omega \in [0, \pi)$ obtained using the various order estimation procedures with the `CCA`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

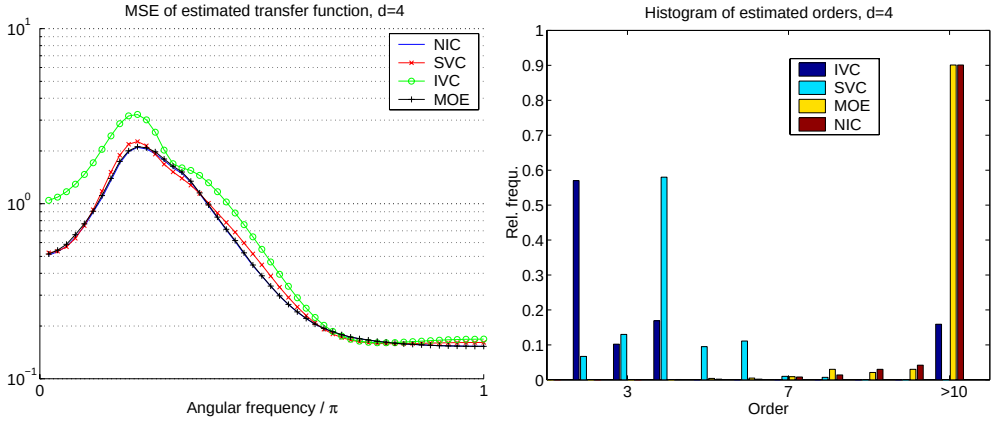


Figure 3: These plots show the result of the simulation for $T = 1000$. The left picture shows the average mean square error of the transfer function at 50 equally spaced frequencies in the angular frequency range $\omega \in [0, \pi)$ obtained using the various order estimation procedures with the `CCA` weighting scheme and $f = p = 4\hat{p}_{AIC}$. The right plot shows the corresponding histogram for the estimated orders. The plots have been obtained using 1000 replications.

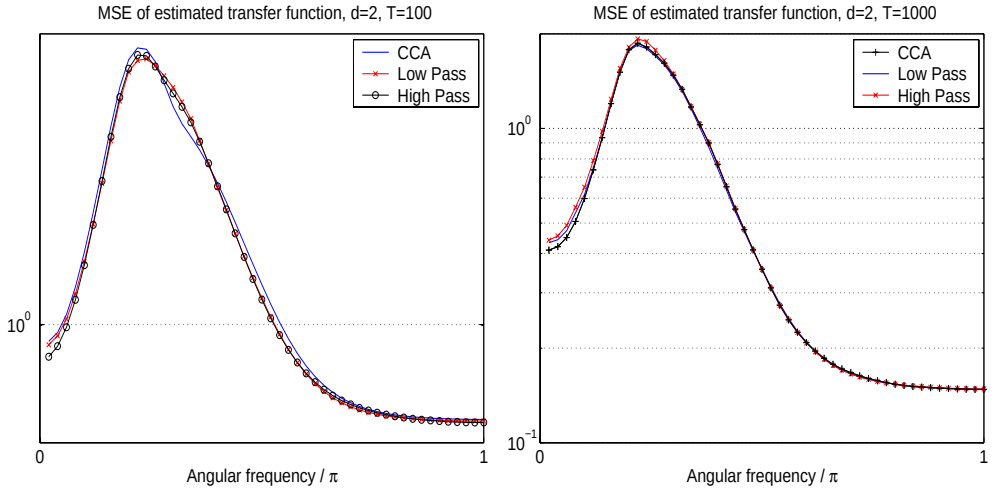


Figure 4: These plots show the result of the simulation for $T = 100$ (left picture) and $T = 1000$ (right picture). The pictures shows the average mean square error of the transfer function at 50 equally spaced frequencies in the angular frequency range $\omega \in [0, \pi)$ obtained using the `IVC`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

has a tendency to underestimate the order in all cases, `NIC` only for $T = 100$. In this example `SVC` and `NIC` show the best performance. 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 $n(sf + p(m+s))C(T)$. 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.

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 consistent have been given. Also the behaviour of one particular procedure in the case of infinite-dimensional models with summable second order modes has been investigated. 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.

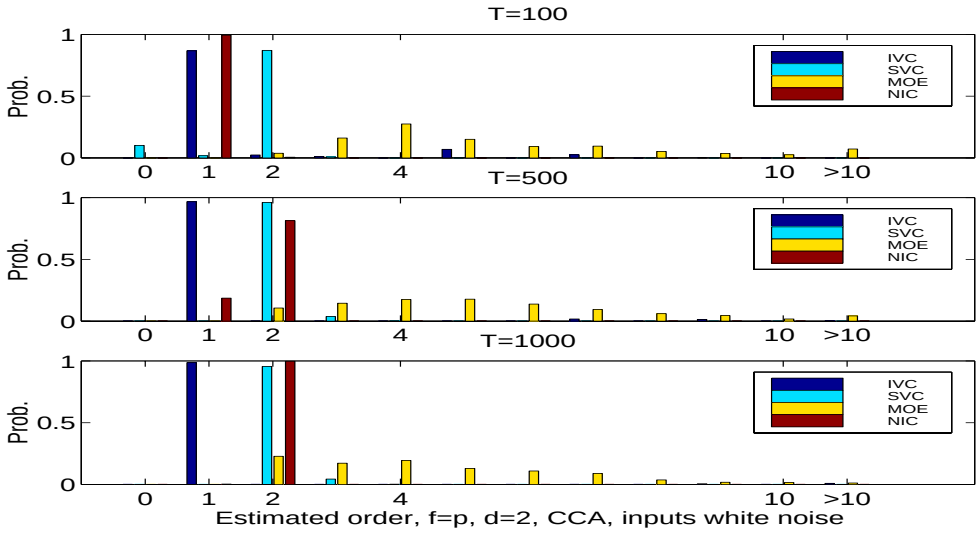


Figure 5: This figure shows the result of 1000 simulation runs using sample sizes $T = 100$ (top row), $T = 500$ (middle row), $T = 1000$ (bottom row). Each picture shows the probabilities of estimating the order using the four estimation algorithms: The weightings have been chosen according to `CCA`. The truncation indices have been chosen as $f = p = 2\hat{p}_{AIC}$. The inputs are i.i.d. uniformely distributed white noise normalized to zero mean and unit variance.

References

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

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. (1999). Order estimation in the context of `MOESP` subspace identification methods. In: *Proceedings of the ECC'99 Conference, Karlsruhe, Germany.*

Bauer, D. and M. Jansson (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.* Academic Press.

Chui, N. (1997). Subspace Methods and Informative Experiments for System Identification. PhD thesis. Cambridge University.

Hannan, E. J. and M. Deistler (1988). *The Statistical Theory of Linear Systems.* John Wiley. New York.

Jansson, M. and Bo Wahlberg (1997). Counterexample to general consistency of subspace system identification methods. In: *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: *Proc. 1983 Amer. Control Conference 2.* (H. S. Rao and P. Dorato, Eds.). Piscataway, NJ. pp. 445–451. 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. Prentice Hall.

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

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

Peternell, K., W. Scherrer and M. Deistler (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. *Ann. Statistics* **8**(1), 147–164.

Sorelius, J. (1999). Subspace-Based Parameter Estimation Problems in Signal Processing. PhD thesis. Uppsala University.

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.