Some Facts about the Choice of the Weighting Matrices in Larimore Type of Subspace Algorithms

Dietmar Bauer and Lennart Ljung

Department of Electrical Engineering
Linkping University, S-581 83 Linkping, Sweden
WWW: http://www.control.isy.liu.se
Email: ljung@isy.liu.se

2000-10-13

Report no.: LiTH-ISY-R-2305
Submitted to Automatica

Technical reports from the Automatic Control group in Linkping are available by anonymous ftp at the address ftp.control.isy.liu.se. This report is contained in the pdf file 2305.pdf.
Some facts about the choice of the weighting matrices
in Larimore type of subspace algorithms

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

Lennart Ljung
Division of Automatic Control, Department of Electrical Engineering,
Linköping University, SE-581 83 Linköping, Sweden,
e-mail: ljung@isy.liu.se

October 4, 2000
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.

Keywords: linear systems, discrete time systems, subspace methods, asymptotic
variance

*Corresponding author: e-mail: Dietmar.Bauer@tuwien.ac.at, fax.: +43 1 58801 11999.
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 MOESP (multivariable output error state space) 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: (Peterrell et al., 1996) derive the consistency, (Bauer et al., 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; 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 et al., 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 organised 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. Section 4 demonstrates the results in some numerical examples. Finally section 5 concludes.

Throughout the paper the following notation will be used: $I_n$ denotes the $n \times n$ identity matrix, $0^{a \times b}$ the nullmatrix of respective dimensions. Further $f_T = O(g_T)$ means that
\[ \limsup_{T \to 0} \| f_T / g_T \| \leq M \text{ almost sure (a.s.)}. \] Also \( f_T = o(g_T) \) means that \( \lim_{T \to \infty} \| f_T / g_T \| = 0 \text{ a.s.} \) 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 \( \hat{=} \) 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

\[
\begin{align*}
    x_{t+1} &= Ax_t + Bu_t + K\varepsilon_t \\
    y_t &= Cx_t + Du_t + \varepsilon_t
\end{align*}
\]

where \( y_t \in \mathbb{R}^s \) denotes the observed output process, \( u_t \in \mathbb{R}^m \) denotes the observed input process and \( \varepsilon_t \in \mathbb{R}^l \) 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 \( \mathcal{M}_n \) 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 \( \varepsilon_t \) is for simplicity assumed to be i.i.d. with mean zero, nonsingular variance and finite fourth moments. More general assumptions in a martingale difference framework can be found in (Bauer et al., 1999). The input is assumed to be i.i.d. with mean zero and nonsingular variance, also having finite fourth moments. Input and noise are assumed to be independent.

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 \( \varepsilon_t \) respectively in the place of \( y_t \). Let \( Z_{t,p} = [y_{t-1}, u_{t-1}^p, \ldots, y_{t-p}^p, u_{t-p}^p]' \). 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 \([CA^{i-2}B, \ldots, CB, D, 0^{i\times(f-j)m}]\) as its \( j \)-th block row and \( \mathcal{E}_f \) contains \([CA^{i-2}K, \ldots, CK, I, 0^{i\times(f-j)m}]\) as its \( j \)-th 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 \((\hat{\Gamma}_p^-\Pi)^{1/2}\), where \(\hat{\Gamma}_p^- = \frac{1}{T} \sum_{t=p+1}^T Z_{t,p} (Z_{t,p})'\) denotes the sample variance of \(Z_{t,p}\) and \(X^{1/2}\) denotes the uniquely defined symmetric square root of a matrix \(X\). Further \(\hat{\Gamma}_p^-\Pi = \hat{\Gamma}_p^- - \hat{\Gamma}_{u,z} \hat{\Gamma}_{u,z}^{-1} \hat{\Gamma}_{u,z}\). 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 \(\Gamma_p = E(\hat{\Gamma}_p^-\Pi)\) denote the expectation of the weighting matrix. It follows from the assumptions on the inputs and the noise stated above, that for any fixed \(p\) the estimate \(\hat{\Gamma}_p^-\Pi\) converges to \(\Gamma_p\). 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.

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

\[
\hat{\Gamma}_f^+\Pi = \hat{\Gamma}_f^+ - \hat{\Gamma}_{y,u} \hat{\Gamma}_{u,u}^{-1} \hat{\Gamma}_{u,y}
\]  

(2)

where \(\hat{\Gamma}_f^+\) 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 \(\|\hat{W}_f^+ - W_f^+\| = O(Q_T)\) 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. 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 bounded away from zero for \(f = O((\log T)^a), a < \infty\). The name CCA (canonical correlation analysis) will be reserved for the procedure using

\[
\hat{W}_p = (\hat{\Gamma}_p^-\Pi)^{1/2} \quad \text{and} \quad \hat{W}_f^+ = (\hat{\Gamma}_f^+\Pi)^{-1/2}
\]  

(3)

In the third step the difference between the two classes of procedures appears: Whereas the Larimore type of procedures use \(\hat{\mathcal{K}}_p\) to continue, the MOESP type of procedures uses \(\hat{\mathcal{O}}_f\) (for detail 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

\[ \hat{K}_p = \hat{S} \hat{V}_n' (\hat{W}_p^-)^{-1}. \]

Here \( \hat{S} = [\hat{V}_n' (\hat{\Gamma}_p^-)^{-1}]_n \) appears to be a convenient choice of \( \hat{S} \), where \( [X]_n \) denotes the submatrix containing the first \( n \) columns of \( X \). This is possible (a.s. asymptotically), if the first \( n \) columns of \( \mathcal{K}_p \) are linearly independent (in one and thus in any representation). This holds true on a generic subset of \( M_n \), which is denoted with \( M_n^+ \). 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 \( [\mathcal{K}_p]_n = I_n \) 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 vectorisation of the estimated system in a canonical form. This is done in the main result of this paper:

**Theorem 3.1** Let \( y_t \) be generated by a system \( (A_c, B_c, C_c, D_c, K_c) \), such that the corresponding pair of transfer functions are in \( M_n^+ \). The noise is assumed to be i.i.d. with zero mean, nonsingular variance and finite fourth moments. The observed input is restricted to be i.i.d. with zero mean, nonsingular variance and finite fourth moments. Furthermore it is assumed that noise and observed input are independent. Assume that the Larimore type of procedure using \( \hat{W}_p^- = (\hat{\Gamma}_p^-)^{1/2} \) is used to estimate the system and no delay is assumed, i.e. the entries in \( D_c \) are estimated and not restricted to be zero. Additionally it is assumed, that \( p \geq -d \log T/(2 \log |\rho_0|), 1 < d < \infty \) 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 \( f \) 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 equation (2). Then the asymptotic variance of \( \text{vec} \left[ \hat{A}_c - A_c, \hat{B}_c - B_c, \hat{C}_c - C_c, \hat{D}_c - D_c, \hat{K}_c - K_c \right] \) is of the form

\[
M_1 M'_1 + M_2 \left[ \Gamma_z \otimes \left\{ (\mathcal{O}_f W_2 \mathcal{O}_f)^{-1} \mathcal{O}_f' W_2 [E_f (I_f \otimes \Omega) E_f'] W_2 \mathcal{O}_f (\mathcal{O}_f' W_2 \mathcal{O}_f)^{-1} \right\} \right] M'_2
\]  

(4)
where $W_2 = \lim_{T \to \infty} (W_f^+) f W_f^+$. Here the matrices $M_1 \in \mathbb{R}^{(n+m)(n+m)+ns \times (n+m)}$ and $M_2 \in \mathbb{R}^{(n+m)(n+m)+ns \times \infty}$ do not depend on $f$ or $W_f^+$. Finally $\Omega$ denotes the noise covariance matrix and $\Gamma_z = \mathbb{E}Z_t \in \mathbb{R}^{(n+m)+ns}$. 

The 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.

The implications of the theorem are that it is always (i.e. for any choice of $f$) optimal to use the CCA weighting scheme. The theorem also suggests to use $f \to \infty$ 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. Furthermore the theorem provides a measure of how much of attainable accuracy one looses by using any other method than the optimal. 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 part of the proof to demonstrate that this limit exists.

**Proof:** Consider the estimation of the system matrices using the estimate of the state sequence $\hat{x}_t = \hat{K}_p Z_{t_p}$: This is done using ordinary least squares. 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 $[\hat{K}_p]_n = I_n$ and so does the true state $x_t$. Let $\Delta_t = \hat{x}_t - x_t$. Introduce the notation $\langle a_t, b_t \rangle = T^{-1} \sum_{t=p+1}^{T-f} a_t b_t$. Then:

$$
\begin{bmatrix}
\hat{C} - C, \hat{D} - D
\end{bmatrix}
= \begin{bmatrix}
\langle \hat{\varepsilon}_t - C \Delta_t, \hat{x}_t \rangle, \langle \hat{\varepsilon}_t - C \Delta_t, u_t \rangle
\end{bmatrix}
\begin{bmatrix}
\langle \hat{x}_t, \hat{x}_t \rangle, \langle \hat{x}_t, u_t \rangle
\end{bmatrix}^{-1}
$$

$$
\begin{bmatrix}
\hat{A} - A, \hat{B} - B
\end{bmatrix}
= \begin{bmatrix}
\langle \hat{\Delta}_{t+1} + K \varepsilon_t - A \Delta_t, \hat{x}_t \rangle, \langle \hat{\Delta}_{t+1} + K \varepsilon_t - A \Delta_t, u_t \rangle
\end{bmatrix}
\begin{bmatrix}
\langle \hat{x}_t, \hat{x}_t \rangle, \langle \hat{x}_t, u_t \rangle
\end{bmatrix}^{-1}
$$

$$
\begin{bmatrix}
\hat{K} - K
\end{bmatrix}
= \langle \hat{\Delta}_{t+1} - A \Delta_t + Bu_t - K (\hat{\varepsilon}_t - \varepsilon_t), \hat{\varepsilon}_t \rangle \langle \hat{\varepsilon}_t, \hat{\varepsilon}_t \rangle^{-1}
$$

(5)

Since $\hat{\varepsilon}_t - \varepsilon_t = - (\hat{C} - C) \hat{x}_t - C \Delta_t$ it is observed, that a number of terms appear in these expressions: $\langle \hat{\varepsilon}_t, \hat{x}_t \rangle, \langle \hat{\Delta}_t, \hat{x}_t \rangle, \langle \hat{\Delta}_{t+1}, \hat{x}_t \rangle, \langle \hat{\Delta}_{t+1}, u_t \rangle, \langle \Delta_t, u_t \rangle, \langle \hat{\Delta}_t, \hat{\varepsilon}_t \rangle, \langle \hat{\Delta}_{t+1}, \hat{\varepsilon}_t \rangle$ and $\langle \varepsilon_t, u_t \rangle$. 

7
Note that \( \Delta_t = (\hat{\mathcal{K}}_p - \mathcal{K}_p)Z_{t,p} - (A - KC)^p x_{t-p} \). Let \( P_{\mathcal{K}} = I_{p(s+m)} - [I_n, 0^{n \times (p^{(m+s)}-n)}]' \mathcal{K}_p \). Here all system matrices are assumed to be in the canonical form introduced by the restriction \( [\mathcal{K}_p]_n = I_n \). Further let \( \mathcal{O}_f = (\mathcal{O}_f W_2 \mathcal{O}_f)^{-1} \mathcal{O}_f W_2 \) denote a generalized left inverse of \( \mathcal{O}_f \). Recall that \( W_2 = (W_f^+)^/W_f^+ \). In the notation we refer to the case \( f \) fixed and finite. In the CCA case with \( f \to \infty \) we obtain \( W_2 = (\Gamma^{++}_\infty)^{-1} \). This introduces no further complications however, since it is straightforward to show, that all necessary limits exist. For details on this see (Bauer, 1998). Then it has been shown in (Bauer et al., 2000) that

\[
(\hat{\mathcal{K}}_p - \mathcal{K}_p) = \mathcal{O}_f^t(\hat{\beta}_z - \beta_z)P_{\mathcal{K}} + O(\|\hat{\beta}_z - \beta_z\|\|\hat{W}_p^+ - W_p^+\|) + O(\|\hat{\beta}_z - \beta_z\|^2)
\]

where \( \beta_z = \mathbb{E}Y_{t,f}(Z_{t,p})^/-\mathcal{G}_{\mathcal{P}} \). For any of the proposed weighting matrices, \( \|\hat{W}_p^+ - W_p^+\| = O(Q_T f p) \). 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 \( \mathcal{O}_f^t(\hat{\beta}_z - \beta_z)P_{\mathcal{K}} \) is the essential one, the remaining terms do not show up in the asymptotic distribution, as they are \( o(T^{-1/2}) \). Here asymptotic distribution is to be understood for the vectorisation of the matrix in the sense of (Lewis and Reinsel, 1985), as the dimensions of the matrix increase according to \( p \to \infty \): i.e. a zero mean vector \( x_T \in \mathbb{R}^{p(T)} \) is said to be distributed asymptotically normal, if for any vector \( l_T \in \mathbb{R}^{p(T)} \), such that

- \( \|l_T\|_2 \leq M \) for some \( M < \infty \)
- \( \|l_T, 0\|_2 \to 0 \) for some vector \( l \in \ell_2 \)
- \( \mathbb{E}(l_T, x_T)^2 \to c \) for some \( 0 \leq c < \infty \)

the scalar product \( l_T^t x_T \) converges in distribution to a normal random variable.

\( P_{\mathcal{K}} \) 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 regarding the weighting \( \hat{W}_f^+ \) only the expectation \( W_f^+ \) has an influence. Thus it is seen, that the only essential stochastic terms in the equations above are \( \langle \varepsilon_t, x_t \rangle, \langle \varepsilon_t, u_t \rangle \) and \( \hat{\mathcal{K}}_p - \mathcal{K}_p \doteq \mathcal{O}_f^t(\hat{\beta}_z - \beta_z)P_{\mathcal{K}} \), where \( \doteq \) denotes equality up to terms of order \( o(T^{-1/2}) \). Since \( p \to \infty \) as a function of the sample
size it follows that $\| \beta_z, 0_{n \to \infty}^T - \mathcal{O}_f \mathcal{K} \|_2 = O(p|\rho_0|^p) = o(T^{-1/2})$, where $\rho_0$ denotes a zero of $H(q)$ of maximum modulus. Then let $Z_{t,p}^{-\Pi} = Z_{t,p} - \hat{\Gamma}_z u \hat{\Gamma}_u^{-1} U_{t,f}^\top$. Therefore

$$
\beta_z - \beta_z \doteq \langle \mathcal{E}_f E_{t,f}^+; Z_{t,p}^{-\Pi} \rangle \langle Z_{t,p}^{-2\Pi}; Z_{t,p}^{-\Pi} \rangle^{-1}
$$

as follows from straightforward calculations. The main idea of the proof now is to find matrices $M_1 \in \mathbb{R}^{[n+s][n+m]+ns \times (n+m)}$ and $M_2 \in \mathbb{R}^{[n+s][n+m]+ns \times np(m+s)}$ such that

$$
\text{vec} \left[ \hat{A}_c - A, \hat{B}_c - B, \hat{C}_c - C, \hat{D}_c - D, \hat{K}_c - K \right] = M_1 \text{vec} [\langle \varepsilon_t,\begin{bmatrix} x_t \\ u_t \end{bmatrix} \rangle] + M_2 \text{vec}[\mathcal{O}_f \mathcal{E}_f E_{t,f}^+; Z_{t,p}^{-\Pi}]
$$

where $[M_{2,p}, 0_{[n+s][n+m]+ns \times \infty}]$ converges in $\ell_1$ norm to a matrix whose entries decrease exponentially with column index. In this case the sufficient conditions for asymptotic normality given in (Lewis and Reinsel, 1985) are fulfilled (see e.g. Bauer et al., 1999). It will be clear from the derivation below, that $M_1$ and $M_2 = \lim_{T \to \infty} [M_{2,p}, 0_{[n+s][n+m]+ns \times \infty}]$ do not depend on $f$ or $W_f^\top$. Note that

$$
\mathbb{E} \text{vec} [\langle \varepsilon_t, x_t \rangle] \text{vec} [\langle \varepsilon_t, x_t \rangle]^\top = \frac{1}{T^2} \sum_{t,s=p+1}^{T-f} \mathbb{E} [x_t x'_s \otimes \varepsilon_t \varepsilon'_s] = \frac{1}{T} (\Sigma_x \otimes \Omega)
$$

where $\Sigma_x = \mathcal{K} \mathcal{G} \mathcal{K}' = \mathbb{E} x_t x'_t$ and $\doteq$ again denotes equality up to terms of order $o(T^{-1/2})$. Analogously $\mathbb{E} \text{vec} [\langle \varepsilon_t, u_t \rangle] \text{vec} [\langle \varepsilon_t, u_t \rangle]^\top = (\Omega_u \otimes \Omega)$, where $\Omega_u = \mathbb{E} u_t u'_t$. Also

$$
\mathbb{E} \text{vec} [\langle \varepsilon_t, x_t \rangle] \text{vec} [\langle \varepsilon_t, u_t \rangle]^\top = 0
$$

Further note that for any component of $\text{vec} [\langle \varepsilon_t, x_t \rangle] \doteq \text{vec} [\langle \varepsilon_t, \mathcal{K}_p Z_{t,p} \rangle]$ and any linear combination $x'_t (\mathcal{K}_p - \mathcal{K}_p) X_p \doteq x'_t \mathcal{O}_f \mathcal{E}_f (E_{t,f}^+; Z_{t,p}^{-\Pi}) \Gamma_p \mathcal{P}_K X_p$ for some vectors $x \in \mathbb{R}^n$ and $X_p \in \mathbb{R}^{(s+m)}$ such that $||X'_p, 0_{1 \times \infty}^\times - X'||_1 \to 0$, where $X$ is a vector in $\ell_1$ having elements decreasing exponentially, one obtains that

$$
\mathbb{E} \frac{1}{T^2} \sum_{t,s=p+1}^{T-f} \varepsilon_{t,i} (\mathcal{K}_{p,j} Z_{t,p}^{-\Pi}) (x'_t \mathcal{O}_f \mathcal{E}_f E_{s,f}) (X'_p \mathcal{P}_K \Gamma_p^{-1} Z_{s,p}^{-\Pi})
$$

$$
= \frac{1}{T^2} \sum_{t,p+1}^{T-f} \sum_{s=1}^{T-t} \mathbb{E} \varepsilon_{t,i} (x'_t \mathcal{O}_f \mathcal{E}_f E_{s,f}^+ \mathcal{P}_K \Gamma_p^{-1} Z_{s,p}^{-\Pi})
$$

$$
= \frac{1}{T^2} \sum_{t=p+1}^{T-f} \sum_{s=1}^{T-t} \mathbb{E} \varepsilon_{t,i} (x'_t \mathcal{O}_f \mathcal{E}_f E_{s,f}^+ \mathcal{P}_K \Gamma_p^{-1} Z_{s,p}^{-\Pi}) [\mathcal{K}_{p,j}, 0_{1 \times [m+s]}(t-s)] \begin{bmatrix} \mathcal{H}_{t-s,p} \\ \Gamma_p \end{bmatrix} \Gamma_p^{-1} \mathcal{P}_K X_p
$$

$$
= \frac{1}{T^2} \sum_{t=p+1}^{T-f} \sum_{s=1}^{T-t} \mathbb{E} \varepsilon_{t,i} (x'_t \mathcal{O}_f \mathcal{E}_f E_{s,f}^+ \mathcal{P}_K \Gamma_p^{-1} Z_{s,p}^{-\Pi}) \begin{bmatrix} \mathcal{K}_{p,j} \Gamma_p \end{bmatrix} \Gamma_p^{-1} \mathcal{P}_K X_p + o(T^{-1}) = o(T^{-1})
where $\mathcal{H}_{j,p} = \mathbb{E}Z_{t,j}(Z_{t-j,p})' = \bar{O}_j \mathcal{K}_p \Gamma_p + o(T^{-1/2})$ and $\bar{s} = \max\{p+1, t-f+1\}$. Here mostly $\mathcal{K}_p P_{\mathcal{K}} = 0$ and $\|\mathcal{H}_{j,p} \Gamma_p^{-1} - \bar{O}_j \mathcal{K}_p\| = o(T^{-1/2})$ 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 $X_p$ and the analogous property of $\mathcal{K}_p$. 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. Thus let $\mathcal{H}_j$ denote $\mathcal{H}_{j,\infty}$. For the covariance of elements of $\langle \varepsilon_t, u_t \rangle$ with $x'(\bar{K}_p - \mathcal{K}_p)X_p$ analogous arguments hold. Therefore the two terms are asymptotically uncorrelated. The first term due to $\langle \varepsilon_t, x_t \rangle$ and $\langle \varepsilon_t, u_t \rangle$ has been shown to have variance $M_1 M_1'$, where $M_1$ can be shown to be independent of $f$ and $W^+_f$ because of equation (5), see also below. Thus the only effect of these choices is hidden in the second term. Consider the variance of $\mathcal{O}'_f \langle \mathcal{E}_f E^+_t z^+_{t,p} \rangle_{\mathbb{P}} \Gamma_p^{-1} P_{\mathcal{K}} x_p$ for some vector $x_p \in \mathbb{R}^{p[m + s]}$ such that $\|x'_p, 0^{1 \times \infty}\| - x'_p \|1 \to 0$ for some vector $x$ in $\ell_1$ having elements decreasing exponentially:

$$
\frac{1}{T} \sum_{s,t=p+1}^{T-f} \mathbb{E} \mathcal{O}'_f \mathcal{E}_f (E^+_t z^+_{t,p} \Gamma_p^{-1} P_{\mathcal{K}} x_p x'_p P_{\mathcal{K}}^{-1} z^+_{s,p} \Gamma_p^{-1} (E^+_s f)') \mathcal{E}_f (\mathcal{O}'_f)'
$$

$$
= \frac{1}{T} \sum_{l=1}^{T-f} \left[ \mathcal{O}'_f \mathcal{E}_f \left( E^+_t z^+_{t+l,f} \right) \mathcal{E}_f (\mathcal{O}'_f) \right] (x'_p P_{\mathcal{K}}^{-1} \mathbb{E} Z^+_{t,p} (z^+_{t+l,p})' \Gamma_p^{-1} P_{\mathcal{K}} x_p) + o(T^{-1})
$$

(7)

Note that for $l = 0$ the part due to $E^+_t f$ is equal to $\mathcal{O}'_f \mathcal{E}_f (I \otimes \Omega) \mathcal{E}_f (\mathcal{O}'_f)'. This is the central term in the expression for the asymptotic variance given in the theorem. From the description of the estimation error in equation (5) it follows, that a matrix $M_2$ as used above exists. The construction of this matrix will be clarified below. The theorem then is proved, if for any vectors $x_p, y_p$ postmultiplying $\bar{K}_p - \mathcal{K}_p$ in equations (5) it holds, that $\mathbb{E} x'_p P_{\mathcal{K}}^{-1} \Gamma_p^{-1} z^+_{t,p} (z^+_{t-j,p})' \Gamma_p^{-1} P_{\mathcal{K}} y_p \to 0, j \neq 0$. This will be done below.

Consider the case $n < (s+m)$ first. Examining the expressions for the estimation errors given in equation (5), one observes that only a number of terms are multiplying $\bar{K}_p - \mathcal{K}_p$: These terms converge to $\Gamma_z \mathcal{K}_z, [\mathcal{K}_z, \Gamma_z]^t \mathcal{K}_z', \left[ 0^{m \times s}, I_m, 0^{m \times \infty} \right]'$ and $[I_s, 0^{s \times \infty}]'$, where convergence is in $\ell_1$ norm as required above for the vectors $x_p$. This follows from the facts, that $\mathbb{E} Z^+_{t,\infty} u'_t = 0, \mathbb{E} Z^+_{t,\infty} \varepsilon'_t = 0$, the error $\|\bar{K}_p - \mathcal{K}_p\| = O(Q_T p f)$ and the uniform convergence of the sample covariances. All these matrices have elements decreasing exponentially. Note
that

$$\mathbb{E}[P^l_{\mathcal{K}} \Gamma^{-1}_{z} Z_{t-\infty}^{-1}(Z^{-1}_{t-j,\infty})^{T} P_{\mathcal{K}} = P_{\mathcal{K}}^{l} \Gamma^{-1}_{z}\begin{bmatrix} \mathcal{H}_j \\ \Gamma^{-1}_{z} \end{bmatrix} \Gamma^{-1}_{z} P_{\mathcal{K}}$$

$$= P_{\mathcal{K}}^{l} \Gamma^{-1}_{z} \begin{bmatrix} 0^{j(s+m)\times \infty} \\ I_{\infty} \end{bmatrix} P_{\mathcal{K}} = P_{\mathcal{K}}^{l} \begin{bmatrix} 0^{j(s+m)\times \infty} \\ \Gamma^{-1}_{z} \end{bmatrix} P_{\mathcal{K}} = \begin{bmatrix} 0^{j(s+m)\times \infty} \\ \Gamma^{-1}_{z} \end{bmatrix} P_{\mathcal{K}}$$

(8)

evaluating the expression for $p = \infty$ rather than dealing with the limit. Using the exponential decrease however it is straightforward show that in all situation, 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

$$\Gamma^{-1}_{z} = \begin{bmatrix} 0^{j(s+m)\times (s+m)} & 0^{j(s+m)\infty} \\ 0^{\infty\times (s+m)} & \Gamma^{-1}_{z} \end{bmatrix} + \begin{bmatrix} I_{s+m} \\ -\Gamma^{-1}_{z} \mathcal{H}_1' \end{bmatrix} (\gamma_z(0) - \mathcal{H}_1 \Gamma^{-1}_{z} \mathcal{H}_1')^{-1} [I_{s+m}, -\mathcal{H}_1 \Gamma^{-1}_{z}]$$

(9)

The projection $P_{\mathcal{K}}^l = I_{\infty} - \mathcal{K}'[I_{n}, 0^{n\times\infty}]$ and thus the last equality follows from $n < (s+m)$. Premultiplying with the above mentioned terms from the left shows, that the terms for $l \neq 0$ in the equation (7) do not matter in the case $n < (s+m)$. Take e.g.

$$\mathbb{E}[\mathcal{K} [\mathcal{H}_1', \Gamma_2] P_{\mathcal{K}}^{l} \Gamma^{-1}_{z} Z_{t-\infty}^{-1}(Z^{-1}_{t-j,\infty})^{T} P_{\mathcal{K}} = \mathcal{K} [\mathcal{H}_1', \Gamma_2] \begin{bmatrix} 0^{j(s+m)\times \infty} \\ \Gamma^{-1}_{z} \end{bmatrix} P_{\mathcal{K}} = 0$$

This shows, that in the case $n < (s+m)$ only the term for $l = 0$ in equation (7) is nonzero and thus the theorem holds in this case. Therefore assume $n \geq (s+m)$ from now on.

In the theorem the asymptotic variance of $(\hat{A}_c, \hat{B}_c, \hat{C}_c, \hat{D}_c, \hat{K}_c)$ is given. In order to show that only the term for $l = 0$ in equation (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 $(\hat{A}_c, \hat{B}_c, \hat{C}_c, \hat{D}_c, \hat{K}_c)$, where $\hat{A}_c = \hat{A}_c - \hat{K}_c \hat{C}_c$, $\hat{B}_c = \hat{B}_c - \hat{K}_c \hat{C}_c$. 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 $[\mathcal{K}_p]_n = I_n$, which is done using the transformation matrix $\hat{S}$, which is defined as $\hat{S} = [[\hat{K}, \hat{B}], \cdots, \hat{A}^{n-1}[\hat{K}, \hat{B}]]_n$. Then e.g. $\hat{C}_c - C = \hat{C} \hat{S} - C = (\hat{C} - C) \hat{S} + C(\hat{S} - I_n)$, where $I_n$ is the limit of $\hat{S}$ as follows
in a straightforward fashion from the consistency results for \( \hat{K}_p \) and the sample covariances used in the regression.

Thus consider the estimation errors in the system matrix estimates more closely: Due to the white noise assumption on the inputs one obtains \( \langle \hat{x}_t, u_t \rangle \to 0 \) and thus the estimation errors of \( \hat{C} \) and \( \hat{D} \) can be treated independently. Consider \( \hat{C} \) first:

\[
\hat{C} - C = \langle \varepsilon_t - C \Delta_t, \hat{x}_t \rangle (\hat{x}_t, \hat{x}_t)^{-1} = \langle \varepsilon_t, x_t \rangle \Sigma_x^{-1} - \langle C \Delta_t, x_t \rangle \Sigma_x^{-1} \\
= \langle \varepsilon_t, x_t \rangle \Sigma_x^{-1} - C (\hat{K}_p - K_p) \Gamma_p K_p' \Sigma_x^{-1}
\]

Here the error bound \( \| \hat{K}_p - K_p \| = o(Tp f) \) has been used to show e.g. that \( \langle \varepsilon_t, \hat{x}_t \rangle = \langle \varepsilon_t, x_t \rangle \). Next deal with \( \hat{K} \):

\[
\hat{K} - K = \langle \Delta_{t+1}, \varepsilon_t \rangle \Omega^{-1} + \langle -A \Delta_t - K (\hat{\varepsilon}_t - \varepsilon_t), \varepsilon_t \rangle \Omega^{-1} = (\hat{K}_p - K_p) \left[ \begin{array}{c} I_s \\ 0_{[p(s+m)-s] \times s} \end{array} \right]
\]

This follows from the error bound cited above and the uniform convergence of the sample covariances, as e.g.

\[
\langle \Delta_t, \varepsilon_t \rangle = (\hat{K}_p - K_p) \langle Z_{t+p}, \varepsilon_t \rangle - \hat{A} \langle x_{t-p}, \varepsilon_t \rangle \equiv 0.
\]

Also the fact that \( \langle u_t, \hat{\varepsilon}_t \rangle = 0 \) has been used. For \( n \geq (s+\mu) \) this is zero. In this case simple manipulations lead to

\[
\hat{A} - A = \langle \Delta_{t+1}, \Delta_t, x_t \rangle \Sigma_x^{-1} = (\hat{K}_p - K_p, 0_{n \times (m+s)}) \left[ \begin{array}{c} H_{1+p} \\ \Gamma_p \end{array} \right] K' \Sigma_x^{-1} - A (\hat{K}_p - K_p) \Gamma_2 K' \Sigma_x^{-1}
\]

where again \( \equiv \) denotes equality up to terms of order \( o(T^{-1/2}) \). The same arguments show that \( \hat{B} = \hat{B} = K \hat{D} \equiv \hat{B} = B - K D \), where again \( n \geq (s+m) \) is used. Therefore denoting \( \hat{T}_i = \hat{A}^i [\hat{K}, \hat{B}], T_i = A^i [K, B] \) the following recursion is obtained for \( i = 1, 2, \cdots \):

\[
\hat{T}_{i+1} - T_{i+1} = \hat{A} (\hat{T}_i - T_i) + (\hat{A} - A) T_i \\
= \Delta_\kappa \left[ \begin{array}{c} H_1 \\ \Gamma_2 \end{array} \right] K' \Sigma_x^{-1} T_i - A \Delta_\kappa \Gamma_2 K' \Sigma_x^{-1} T_i + A (\hat{T}_i - T_i) \\
= \Delta_\kappa \left[ \begin{array}{c} H_1 \\ \Gamma_2 \end{array} \right] K' \Sigma_x^{-1} T_i - \hat{A} \Delta_\kappa \Gamma_2 K' \Sigma_x^{-1} T_0 \\
+ \sum_{j=1}^{i} \hat{A}^j \Delta_\kappa \left[ \begin{array}{c} H_1 \\ \Gamma_2 \end{array} \right] K' \Sigma_x^{-1} - \Gamma_2 K' \Sigma_x^{-1} A \right] T_{i-j}
\]

12
where $\Delta_{\mathcal{K}} = [\hat{K}_p, 0^{n \times \infty}] - \mathcal{K}$ and where $\mathcal{K} = \mathcal{K}_\infty$ is used. The recursion is started at $\hat{T}_0 = T_0$. Note that due to the normalisation of $\mathcal{K}_n$ 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 $T_i, 0 \leq i < \bar{n}$ and the first $\bar{m}$ columns of $T_{\bar{n}}$ are vectors of the canonical basis. Therefore the columns of $\hat{C}_c - C$ are equal to the columns of the following expressions for the respective integers $i$:

\begin{equation}
(\hat{C} - C)T_{i-1} + C(T_{i-1} - T_{i-2}) = \langle \varepsilon_t, x_t \rangle \sum_{x=1}^{\bar{n}^{-1}T_{i-1} - C \hat{A}^{-1} \Delta_{\mathcal{K}} \Gamma_z \mathcal{K}' \Sigma_x^{-1} T_0 + \sum_{j=0}^{i-2} C \hat{A} \Delta_{\mathcal{K}} \begin{bmatrix}
\mathcal{H}_1 \\
\Gamma_z
\end{bmatrix} \mathcal{K}' \Sigma_x^{-1} - \Gamma_z \mathcal{K}' \Sigma_x^{-1} \hat{A} \begin{bmatrix}
\mathcal{T}_{i-j-2}
\end{bmatrix}
\end{equation}

Thus the columns of $\begin{bmatrix}
\mathcal{H}_1 \\
\Gamma_z
\end{bmatrix} \mathcal{K}' \Sigma_x^{-1} - \Gamma_z \mathcal{K}' \Sigma_x^{-1} \hat{A}$ and of $\Gamma_z \mathcal{K}' \Sigma_x^{-1} T_0$ appear as the vectors $x_p$ or $y_p$ in equation (7). Now for $j > 0$ consider

\begin{equation}
\mathbb{E} T'_0 \Sigma_x^{-1}\mathcal{K} \Gamma_z P'_z \Gamma_z^{-1} Z_{t,\infty} (Z_{t-j,\infty})' \Gamma_z^{-1} P_{\mathcal{K}} = T'_0 \Sigma_x^{-1}\mathcal{K} \Gamma_z P'_z \Gamma_z^{-1} \begin{bmatrix}
0^{j(s+m) \times \infty} \\
I_{\infty}
\end{bmatrix} P_{\mathcal{K}}
\end{equation}

which is zero due to the block matrix inversion (9) since $T'_0[I_n, 0^{n \times \infty}] = [I_{s+m}, 0^{j(s+m) \times \infty}]$. Here $\mathcal{K} P_{\mathcal{K}} = 0$ is used throughout.

The second term appearing above has the same property:

\begin{equation}
\mathbb{E} \{ \Sigma_x^{-1}\mathcal{K}[\mathcal{H}', \Gamma_z] - \hat{A} \Sigma_x^{-1}\mathcal{K} \Gamma_z \} P'_z \Gamma_z^{-1} Z_{t,\infty} (Z_{t-j,\infty})' \Gamma_z^{-1} P_{\mathcal{K}} = \{ \Sigma_x^{-1}\mathcal{K}[\mathcal{H}', \Gamma_z] - \hat{A} \Sigma_x^{-1}\mathcal{K} \Gamma_z \} P'_z \Gamma_z^{-1} \begin{bmatrix}
0^{j(s+m) \times \infty} \\
I_{\infty}
\end{bmatrix} P_{\mathcal{K}}
\end{equation}

\begin{equation}
= \{ \Sigma_x^{-1}\mathcal{K}[\mathcal{H}', \Gamma_z] - \hat{A} \Sigma_x^{-1}\mathcal{K} \Gamma_z \} (I_{\infty} - \mathcal{K}'[I_n, 0^{n \times \infty}] \Gamma_z^{-1} \begin{bmatrix}
0^{j(s+m) \times \infty} \\
I_{\infty}
\end{bmatrix} P_{\mathcal{K}}
\end{equation}

\begin{equation}
= \{ \Sigma_x^{-1}\mathcal{K}[\mathcal{H}', \Gamma_z] - \hat{A} \Sigma_x^{-1}\mathcal{K} \Gamma_z \} \mathcal{K}'[I_n, 0^{n \times \infty}] \Gamma_z^{-1} \begin{bmatrix}
0^{j(s+m) \times \infty} \\
I_{\infty}
\end{bmatrix} P_{\mathcal{K}}
\end{equation}

\begin{equation}
= - [C', 0^{n \times m}] T'_0[I_n, 0^{n \times \infty}] \Gamma_z^{-1} \begin{bmatrix}
0^{j(s+m) \times \infty} \\
I_{\infty}
\end{bmatrix} P_{\mathcal{K}} = 0
\end{equation}
Here $\mathcal{K}[\mathcal{H}_1, \Gamma_z] \mathcal{K}' = \mathbb{E}x_t(x_{t+1})' = \Sigma_x A'$ has been used. Corresponding to $D$ the situation is much easier, since it is easily seen, that $\hat{D}_c - D = \hat{D} - D \doteq (\varepsilon_t, u_t) (\mathbb{E}u_t u_t')^{-1}, \text{ since } \langle \Delta_t, u_t \rangle \doteq 0$. This follows from the bound on $||\hat{K}_p - \mathcal{K}_p||$ and the uncorrelatedness of $u_t$ and $Z_{t,p}$ in the white noise case. Thus the asymptotic distribution of $\hat{D}_c$ is not affected by the choice of $f$ or $W_f^*$.

Due to the definition of $\hat{K}_c$ and $\hat{B}_c = \hat{B}_c - \hat{K}_c \hat{D}_c$ it follows, that $\hat{K}_c = K$, $\hat{B}_c = \hat{B}$ for $n \geq (s + m)$. Thus it remains to consider the estimation of $\hat{A}_c$. Note that $\hat{A}_c - \bar{A} = -(\hat{S} - I_n)\bar{A} + (\hat{A} - \bar{A}) + \bar{A}(\hat{S} - I_n)$. If $i$ is such that $T_i$ is a blockcolumn of the identity matrix, then $\hat{S}T_i = \hat{T}_i$ and thus

$$-(\hat{S} - I_n) \bar{A}T_{i-1} + (\hat{A} - \bar{A}) T_{i-1} + \bar{A}(\hat{S} - I_n) T_{i-1} = -(\hat{S} - I_n) T_i + (\hat{A} - \bar{A}) T_{i-1} + \bar{A}(\hat{T}_{i-1} - T_{i-1}) = 0$$

This is also true for the first $\hat{m}$ columns of $T_{n-1}$. Therefore it remains to deal with the matrix $X = [x_0', \ldots, x_{n-1}', \bar{x}_n']'$, where $X$ denotes the matrix built of the last $s + m$ columns of $\bar{A}$. Here $x_i \in \mathbb{R}^{(s+m) \times (s+m)}$, $i = 0, \ldots, \bar{n} - 1$, $\bar{x}_n \in \mathbb{R}^{\hat{m} \times (s+m)}$. In order to unify the notation let $x_n = [\bar{x}_n', 0^{(s+m) \times (s+m-m)}]'$. Let $s_{n-1} = [0^{(s+m-m)} \times \hat{m}, I_{s+m-m}]'$, $s_n = [I_n, 0^{\hat{m} \times (s+m-m)}]'$, $\bar{s}_{n-1} = [I_{s+m-m}, I_{s+m-m}]'$ and $\bar{s}_n = [0^{\hat{m} \times (s+m-m)} \times \hat{m}]'$ respectively. Finally define $Y = \begin{bmatrix} \mathcal{H}_1 \\ \Gamma_z \end{bmatrix} \mathcal{K}' \Sigma_x^{-1}$. Then for $i = \bar{n} - 1$ and $i = \bar{n}$ one obtains

$$(\hat{T}_i - T_i - (\hat{S} - I_n) \bar{A} T_{i-1}) s_i = \sum_{j=0}^{\bar{n}} (\hat{T}_j - T_j) x_j \bar{s}_i + \hat{T}_i - T_i$$

$$(\hat{D}_i - D_i - (\hat{S} - I_n) \bar{A} T_{i-1}) s_i = \sum_{j=1}^{\bar{n}} \Delta_{k} Y T_{j-1} x_j \bar{s}_i + \hat{\bar{A}} \Delta_{k} \Sigma_x^{-1} T_{0} x_j \bar{s}_i$$

$$(\hat{D}_i - D_i - (\hat{S} - I_n) \bar{A} T_{i-1}) s_i - \sum_{j=1}^{\bar{n}} \sum_{l=1}^{i} \hat{\bar{A}} \Delta_{k} \Sigma_x^{-1} \bar{A} \Delta \{ Y - \Gamma_{z} \Sigma_x^{-1} \bar{A} \} T_{i-l-1} x_j \bar{s}_i$$

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

\[
\Delta_k Y \left( T_{i-1}s_i - \sum_{j=1}^{n} T_{j-1}x_j s_i \right) = \Delta_k Y \left( \begin{bmatrix} 0^{(s+m)\times(s+m)} \\ 0^{(s+m)\times(s+m)} \\ \vdots \\ 0^{m\times(s+m)} \\ I_{s+m} \end{bmatrix} + \begin{bmatrix} -x_1 \\ -x_2 \\ \vdots \\ -\tilde{x}_n \end{bmatrix} \right) s_i
\]

Denoting the matrix in brackets on the right hand side with \( \tilde{X} \) we obtain \( A \tilde{X} = -T_0 x_1 \).

Also \( A \tilde{X} = \tilde{A} \tilde{X} + T_0 [C', 0^{n \times m}] \tilde{X} = T_0 V \) for some matrix \( V \). Therefore

\[
\mathbb{E} \tilde{X}' \Sigma_x^{-1} \mathcal{K}[\mathcal{H}', \Gamma_z] P'_{\mathcal{K}} \Gamma_z^{-1} P_{\mathcal{K}}^{-1} Z_{t,\infty}(Z_{t-j,\infty}) \Gamma_z^{-1} P_{\mathcal{K}}^{-1} \mathcal{K}[\mathcal{H}', \Gamma_z]
\]

\[
= \tilde{X}' \Sigma_x^{-1} \mathcal{K}[\mathcal{H}', \Gamma_z] P'_{\mathcal{K}} \Gamma_z^{-1} \begin{bmatrix} 0^{j(s+m)\times\infty} \\ I_{\infty} \end{bmatrix} P_{\mathcal{K}}^{-1} \mathcal{K}[\mathcal{H}', \Gamma_z]
\]

\[
= \tilde{X}' A'[I_n, 0^{n \times \infty}] \Gamma_z^{-1} \begin{bmatrix} 0^{j(s+m)\times\infty} \\ I_{\infty} \end{bmatrix} P_{\mathcal{K}}^{-1} = 0
\]

Here the second last equation follows from \( \mathcal{K}[\mathcal{H}', \Gamma_z] \mathcal{K}' = \mathbb{E} x_t (x_{t+1})' = \Sigma_x A' \).

Summing up the findings up to now it follows that

\[
\text{vec} \left[ \tilde{A}_c - A, \tilde{B}_c - B, \tilde{C}_c - C, \tilde{D}_c - D, \tilde{K}_c - K \right] = \tilde{M}_1 \text{vec} \left( \varepsilon_t, \begin{bmatrix} x_t \\ u_t \end{bmatrix} \right) + M_{2,p} \text{vec} \left[ \mathcal{O}^1 \mathcal{E}_f (E_{t,f}, Z_{t,p}) \right]
\]

(11)

Here the matrices \( \tilde{M}_1 \) and \( M_{2,p} \) can be found by tracing the computations so far. Additionally it has been shown, that

\[
\mathbb{E} \left[ \tilde{M}_1 \left( \begin{bmatrix} x_t \\ u_t \end{bmatrix} \otimes \varepsilon_t \right) \right] \left[ M_{2,p}(P'_{\mathcal{K}} \Gamma_p^{-1} Z_{t+j,p} \otimes \mathcal{O}^1 \mathcal{E}_f E_{t,j})' \right]' \to 0, \forall j
\]

\[
\mathbb{E} \left[ M_{2,p}(P'_{\mathcal{K}} \Gamma_p^{-1} Z_{t,p} \otimes \mathcal{O}^1 \mathcal{E}_f E_{t,j}) \right] \left[ M_{2,p}(P'_{\mathcal{K}} \Gamma_p^{-1} Z_{t+j,p} \otimes \mathcal{O}^1 \mathcal{E}_f E_{t+j,j})' \right]' \to 0, \forall j \neq 0
\]

From the definition of \( M_1 \) and \( M_{2,p} \) it follows that these matrices do not depend on \( W_2 \) or \( f \).

For the case \( f \) fixed 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 \( f = O((\log T)^u) \).
Here the form of the weighting matrices is used to show e.g. that the variance of \( \mathcal{O}_f \mathcal{E}_f \mathcal{E}_f^+ \)
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.

Note that
\[
\mathcal{O}_f \mathcal{E}_f (I_f \otimes \Omega) \mathcal{E}_f^+(\mathcal{O}^+)' = (\mathcal{O}_f W_2 \mathcal{O}_f)^{-1} \mathcal{O}_f W_2 \mathcal{E}_f (I_f \otimes \Omega) \mathcal{E}_f^+ W_2 \mathcal{O}_f (\mathcal{O}_f W_2 \mathcal{O}_f)^{-1}.
\]
This is minimized by \( W_2 = [\mathcal{E}_f (I_f \otimes \Omega) \mathcal{E}_f^+]^{-1} \) with minimum \( (\mathcal{O}_f (\mathcal{E}_f^+)^{-1}(I_f \otimes \Omega^{-1}) \mathcal{E}_f^{-1} \mathcal{O}_f)^{-1} \).
Some matrix algebra shows, that this gives the identical variance with any choice \( W_2 = W_2' + W_2 \mathcal{O}_f X \mathcal{O}_f^+ W_2 \), such that \( W_2 \) is invertible. Thus the CCA weightings minimize the
variance of the estimated system, since in this case \( W_2 = (\Gamma_{y+}\Pi)^{-1} \), where \( \Gamma_{y+}\Pi = \mathcal{E}_f (I_f \otimes \Omega) \mathcal{E}_f^+ \mathcal{O}_f \sum_x \mathcal{O}_f^+ \). Therefore due to the matrix inversion lemma
\[
(\Gamma_{y+}\Pi)^{-1} = (\mathcal{E}_f (I_f \otimes \Omega) \mathcal{E}_f^+)^{-1} + (\mathcal{E}_f (I_f \otimes \Omega) \mathcal{E}_f^+)^{-1} \mathcal{O}_f X \mathcal{O}_f^+ (\mathcal{E}_f (I_f \otimes \Omega) \mathcal{E}_f^+)^{-1}
\]
for suitable matrix \( X \). The lower diagonal block Toeplitz structure of \( \mathcal{E}_f \) then shows, that
this minimum variance decreases monotonically in \( f \), since it ensures, that \( \mathcal{E}_f^{-1} \mathcal{O}_f \) is a
submatrix of \( \mathcal{E}_f^{-1} \mathcal{O}_f \). This completes the proof. \( \Box \)

This theorem clarifies a long standing question about the optimal choices of the weight-
ing matrices for the algorithms dealt with in this contribution. It shows clearly, that CCA
is the optimal choice in any situation, where no input is present or the observed input
is white noise. It also shows, that \( f \) finite leads to non efficient estimation in the generic
case. 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. The amount of accuracy, which is lost by using a small \( f \) is
thus in the case of using optimal weights determined by the magnitude of the noise zeros,
as they govern the rate of exponential decrease in the matrix \( \mathcal{E}_f^{-1} \mathcal{O}_f \).

The theorem also has implications for different weightings: On the basis of the theorem
it is possible to assess the loss of efficiency due to using a nonoptimal weighting. One
implication of the theorem could be the formulation of a two step algorithm: In a first
step an arbitrary weighting is chosen, leading to consistent results for the various matrices. In a second step for a particular weighting the optimal $f$ is calculated by minimizing the matrix given in the theorem. The benefits of such methods must be doubted however, since the derivation of the theorem have been given on the grounds of knowing the order of the system. In this case it does not seem to make sense to use any other weighting than the optimal.

Also another question is quite natural: The observed input has been assumed to be white. Some of the analysis did not need this assumption and thus also holds for more general inputs. In particular the equation (7) holds for more general inputs. This equation in principle could be used to obtain optimal weightings with respect to specific measures:

Note, that the equation defines a matrix and there need not be a minimum for the matrix minimization problem. However we could for example minimize the trace of the matrix, such that a scalar minimization problem is obtained, which could be solved using nonlinear search methods. However such a process includes optimisation issues, which one wants to avoid in the first place by using subspace rather than prediction error methods. Thus also such a procedure is not recommended, although possible. It is believed however, that the expressions given in this paper might be valuable also for the analysis of this general input case.

4 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:

$$ A = \begin{bmatrix} -0.532 & 0.4639 & 0.2855 \\ 1 & 0 & -0.2568 \\ 0 & 1 & 0.0054 \end{bmatrix}, K = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, C = \begin{bmatrix} -0.532 & 0.4639 & -0.0413 \end{bmatrix} $$

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
Figure 1: The no input case: The left plot shows the measure $E_f$ for the two weighting schemes CCA and N4SID for the range of values $f = 3, \ldots, 20$. The right picture shows a plot of the determinant of $\langle \mathbf{O}_f' \mathbf{W}_2 \mathbf{O}_f \rangle^{-1} \mathbf{O}_f' \mathbf{W}_2 \mathbf{E}_f \mathbf{E}_f' \mathbf{W}_2 \mathbf{O}_f (\mathbf{O}_f' \mathbf{W}_2 \mathbf{O}_f)^{-1}$ for the same two procedures and the same range of integers $f$.

the Fisher information matrix with respect to the echelon canonical form in this case. Then it is well known, that the Cramer Rao bound for the estimation is equal to $F_f^{-1}$. 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 $E_f = \text{tr}[V_f(W_f^+) F_f] - 2ns - (n + s)m$ is used. For an efficient estimation method this is equal to zero, otherwise positive. The left plot in Figure 1 shows this measure for the two weighting schemes CCA (i.e. $\hat{W}_f^+ = (\hat{\Gamma}_f)^{-1/2}$) and N4SID (i.e. $W_f^+ = I_{fs}$). The right plot of this figure shows $\text{det}[(\mathbf{O}_f' \mathbf{W}_2 \mathbf{O}_f)^{-1} \mathbf{O}_f' \mathbf{W}_2 \mathbf{E}_f \mathbf{E}_f' \mathbf{W}_2 \mathbf{O}_f (\mathbf{O}_f' \mathbf{W}_2 \mathbf{O}_f)^{-1}]$, where $W_2 = (W_f^+) W_f^+$.

The plots clearly reveal the identical behaviour of the two measures. It also can 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 also is in accordance with the theory.

The second example is second order single-input single-output system with one addi-
Figure 2: The white noise input case: The left plot shows the measure $E_f$ for the two weighting schemes CCA and N4SID for the range of values $f = 2, \cdots, 20$. The right picture shows a plot of the determinant of $(O_f^r W_2 O_f)^{-1} O_f^r W_2 C_f C_f^r W_2 O_f (O_f^r W_2 O_f)^{-1}$ for the same two procedures and the same range of integers $f$.

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. Analogously to the case of no inputs define $E_f = \text{tr} [V_f (W_f^+) F_f] - 7$. Figure 2 shows the result of the calculation. The pictures demonstrate again 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 the significance of the expressions found in this paper to assess the relative efficiency of various weighting schemes.
5 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 $W_j^+$ has been explored. 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 $(O_j'W_2O_j)^{-1}O_j'W_2\mathcal{E}_f(I \otimes \Omega)\mathcal{E}_f'W_2O_f(O_j'W_2O_f)^{-1}$ as has been shown in Theorem 3.1. 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 optimise the choice of $f$. Finally the new expressions for the asymptotic variance might 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 the first author was holding a post doc position at the Division of Automatic Control in Linköping. Financial support from the EU TMR project 'SI' is gratefully acknowledged.

References


