Analysis of State Space System Identification Methods Based on Instrumental Variables and Subspace Fitting*

Mats Viberg[†] Bo Wahlberg[‡] Björn Ottersten[‡]

October 20, 1995, Revised April 18, 1996

Abstract

Subspace-based State Space System IDentification (4SID) methods have recently been proposed as an alternative to more traditional techniques for multivariable system identification. The advantages are that the user has simple and few design variables, and that the methods have robust numerical properties and relatively low computational complexities. Though subspace techniques have been demonstrated to perform well in a number of cases, the performance of these methods is neither fully understood nor analyzed. Our principal objective is to undertake a statistical investigation of subspace based system identification techniques. The studied methods consist of two steps. The subspace spanned by the extended observability matrix is first estimated. The asymptotic properties of this subspace estimate are derived herein. In the second step, the structure of the extended observability matrix is used to find a system model estimate. Two possible methods are considered. The simplest one only uses a certain shift-invariance property, while in the other method a parametric representation of the null-space of the observability matrix is exploited. Explicit expressions for the asymptotic estimation error variances of the corresponding pole estimates are given.

*This work was supported by the Swedish Research Council for Engineering Sciences (TFR).
[†]Department of Applied Electronics, Chalmers University of Technology, S-421 86 Gothenburg, Sweden
[‡]Department of Signals, Sensors and Systems, Royal Institute of Technology, S-100 44 Stockholm, Sweden

1 Introduction

Advanced engineering applications require useful mathematical models. To extend physical knowledge and prior information of the system, practical experiments are often performed. System identification (SI) deals with the problem of obtaining models of dynamical systems from measured data. The field of SI is by now quite advanced. However, several important problems remain to be solved.

The family of Subspace State Space System Identification (4SID) methods consists of identification techniques which directly estimate a state space representation, without using a particular canonical (minimal) parameterization. See *e.g.* [27, 28, 38] for recent overviews of 4SID methods. These techniques have their origin in state space realization theory. In the classical approach of [7], a block Hankel matrix is constructed from the $n+1$ first impulse response matrices of the system, where n is the system order. As first noted by Kronecker, this Hankel matrix has rank n, and it can be factored as a product of the system observability and controllability matrices. Hence, these matrices can be retrieved from, *e.g.* the singular value decomposition (SVD) of the Hankel matrix. A state space description of the system can then be constructed by utilizing the structure of the observability and the controllability matrices. A problem with this technique is that the system impulse response matrices may be difficult to estimate directly from the input-output data only. See, *e.g.* [23] for a possible approach. The Kung algorithm [12] is similar to Ho and Kalman's state space realization technique, but it is based on using more than $n+1$ impulse response coefficients to form the Hankel matrix. A truncated SVD is then utilized to form estimates of the observability and controllability matrices. This results in more accurate estimates when the impulse response is not precisely known (*i.e.* estimated from data).

More recently, 4SID techniques that sidestep the need for estimating the impulse response have been proposed. Examples are basic projection 4SID techniques given in [5] and their more recent versions, including N4SID, summarized in [29, 30], the CVA algorithm described in [14], and the MOESP method given in [34, 36].

The first step in the 4SID approach to system identification is to estimate the observability matrix directly from the data. In general, the particular state-space realization that the estimated observability matrix belongs to is not specified, and only its range space is estimated. This range space plays the same role as the *signal subspace* in sensor array signal processing, see *e.g.* [22, 39]. Using this observation, we will analyze the performance of these techniques and to relate them to more traditional techniques for system identification.

In [40] some basic 4SID techniques due to [5], described in Section 3.1, were analyzed from a statistical perspective. A severe drawback of these methods is that they only give consistent estimates for systems with no process noise and spatially and temporally white[1] measurement noise. In essence, this fact limits the applicability of the basic 4SID method to scalar output error systems[2] and to high signal-to-noise ratio situations. The approaches of [15] and [34] are algebraically identical to the basic 4SID method, see [38].

As previously mentioned, 4SID methods that work under more general noise conditions have recently been proposed. We present a simple derivation of a variation of the technique of [30], using the concept of instrumental variables. Based on this insight, we undertake a statistical analysis of the method in question. Also, a method for extracting the system poles from the subspace estimate is presented. It is based on a reparameterization resulting in a linear problem in system parameters related to a matrix fraction description of the system. A statistical analysis of the proposed technique is also presented.

The rest of this manuscript is organized as follows: In Section 2 the data model and assumptions are given. The signal subspace estimation problem is studied in Section 3, and an expression for the statistical properties of the estimated extended observability subspace is derived. Section 4 deals with the problem of finding the system parameters from the estimated signal subspace. Two methods are analyzed. The simplest one is the IV-ESP approach, which only use the shift structure of the extended observability matrix to extract the A matrix. The statistical properties of the corresponding estimated eigenvalues, viz. the system poles, are given. Then a more complicated method based on a parametric model of the null-space of the extended observability matrix is derived. The advantage of the latter approach is that it takes the statistical properties of the estimated signal subspace into account in an optimal way, and thus can be used as an "yard stick" for other algorithms. Since it is based on estimation of the coefficients of the characteristic polynomial its numerical properties may be poor. A simple example is given in Section 5, while Section 6 concludes the paper.

[1] A signal is said to be spatially white if its covariance matrix is proportional to the identity matrix.

[2] A system is output error if there is no process noise, *i.e.* $w_t = 0$.

2 Problem Formulation

Consider a linear time-invariant finite-dimensional system with m input signals and l output signals. Assume that the dynamics of the system can be described by an n:th order state space realization

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t && (1) \\ y_t &= Cx_t + Du_t + v_t \ . && (2) \end{aligned}$$

Here, the system matrix A is $n \times n$, B is $n \times m$, C is $l \times n$, and D is $l \times m$. The system is driven by an observable input, $\{u_t\}$ $(m \times 1)$ as well as an unobservable disturbance, $\{w_t\}$ $(n \times 1)$. The output measurements are corrupted by the additive noise process $\{v_t\}$ $(l \times 1)$.

The 4SID methods are based on geometrical properties of a certain input-output relation implied by the description (1)–(2). Introduce the $(l\alpha \times 1)$ vector of stacked outputs as

$$\mathbf{y}(t) = \left[y_t^T, y_{t+1}^T, \ldots, y_{t+\alpha-1}^T\right]^T , \tag{3}$$

where $\alpha > n$ is a user specified integer. Similarly, define the vector of stacked inputs, disturbances, and measurement noises as $\mathbf{u}(t)$, $\mathbf{w}(t)$, and $\mathbf{e}(t)$, respectively. The following input-output equation, [4], is then easily derived from the system description (1)–(2)

$$\mathbf{y}(t) = \boldsymbol{\Gamma}_\alpha x_t + \boldsymbol{\Phi}_\alpha \mathbf{u}(t) + \mathbf{n}(t) , \tag{4}$$

where the term $\mathbf{n}(t) = \boldsymbol{\Psi}_\alpha \mathbf{w}(t) + \mathbf{v}(t)$ contains the noise, and where we have introduced the block matrices

$$\boldsymbol{\Gamma}_\alpha = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{\alpha-1} \end{bmatrix} \tag{5}$$

$$\boldsymbol{\Phi}_\alpha = \begin{bmatrix} D & 0 & \cdots & 0 & 0 \\ CB & D & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & 0 & 0 \\ \vdots & & & & \vdots \\ CA^{\alpha-2}B & CA^{\alpha-3}B & \cdots & CB & D \end{bmatrix} \tag{6}$$

$$\boldsymbol{\Psi}_\alpha = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ C & \ddots & 0 & 0 \\ \vdots & \ddots & 0 & 0 \\ CA^{\alpha-2} & \cdots & C & 0 \end{bmatrix} . \tag{7}$$

Suppose we are given observations of y_t and u_t for $t = 1, 2, \ldots, N + \alpha - 1$. Using (4), the observations can be modeled by the matrix equation

$$\mathbf{Y} = \boldsymbol{\Gamma}_\alpha \mathbf{X} + \boldsymbol{\Phi}_\alpha \mathbf{U} + \mathbf{N} , \tag{8}$$

where the block Hankel matrices of outputs, inputs, and noises, are defined as

$$\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(N)] \tag{9}$$

$$\mathbf{U} = [\mathbf{u}(1), \ldots, \mathbf{u}(N)] \tag{10}$$

$$\mathbf{N} = [\mathbf{n}(1), \ldots, \mathbf{n}(N)] , \tag{11}$$

and the state trajectory is collected in the matrix

$$\mathbf{X} = [x_1, \ldots, x_N] . \tag{12}$$

Note that $\boldsymbol{\Gamma}_\alpha$ represents the extended observability matrix of the system. Equation (8) forms the basis for the 4SID approach. One first attempts to estimate the matrix $\boldsymbol{\Gamma}_\alpha$ directly from the data. The matrices A and C are then retrieved from the structure of $\boldsymbol{\Gamma}_\alpha$ (5) as described in the next section. A number of techniques have been proposed to extract B and D following this step, but the major concern in the current study is estimation of the system matrix, A, which describes the dynamics.

The following assumptions will be used on the data set:

Assumption A.1. The eigenvalues of the matrix A are distinct[3], and they all lie strictly inside the unit circle. □

Assumption A.2. The disturbance and the measurement noises are assumed to be zero-mean, stationary, white Gaussian random processes, with joint covariance matrix

$$\mathrm{E} \begin{bmatrix} w_t \\ v_t \end{bmatrix} \begin{bmatrix} w_s \\ v_s \end{bmatrix}^T = \begin{bmatrix} r_{ww} & r_{wv} \\ r_{wv}^T & r_{vv} \end{bmatrix} \delta_{t,s} , \tag{13}$$

[3]The assumption that the eigenvalues of the matrix A are distinct is only used to simplify the analysis of the statistical properties of the corresponding estimates.

where $\delta_{t,s}$ represents the Kronecker delta. $\square$

Assumption A.3. The system input is modeled as an arbitrary, quasi-stationary, [16], deterministic sequence, and thus uncorrelated with the disturbance and noise terms. The auto-covariance sequence of the input is denoted

$$r_{uu}(\tau) = \bar{E}[u_{t+\tau} u_t^T] \ , \tag{14}$$

where the averaging operator $\bar{E}$ is defined as

$$\bar{E}[z_t] = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \mathrm{E}[z_t] \ . \tag{15}$$

The auto-covariance matrix of the stacked input vectors is similarly defined as

$$\mathbf{R}_{uu}(\tau) = \bar{E}[\mathbf{u}(t+\tau)\mathbf{u}^T(t)] \ . \tag{16}$$

The input is assumed to be persistently exciting in the sense that

$$\mathbf{R}_{uu}(0) \geq \epsilon \mathbf{I} \ , \tag{17}$$

for some $\epsilon > 0$, where the matrix inequality $\mathbf{A} \geq \mathbf{B}$ means that the difference, $\mathbf{A} - \mathbf{B}$ is positive semi-definite. $\square$

We remark that the condition (17) may be neither sufficient nor necessary to guarantee consistency of the proposed estimator, however, it simplifies the analysis. A more precise persistence-of-excitation condition is given in Section 3.3.

3 Signal Subspace Estimation

The first step in subspace-based methods is to estimate the "signal subspace", i.e. the range space of the extended observability matrix. In this section, two approaches for solving this problem are presented. In addition, the asymptotic properties of the instrumental variables approach are derived.

3.1 Basic 4SID Approach

Introduce the orthogonal projection matrix, $\boldsymbol{\Pi}^{\perp}_{\mathbf{U}^T}$, onto the null-space of $\mathbf{U}$. Provided $\mathbf{U}$ has full rank (guaranteed for "large enough" N under (17)), the projection matrix can be expressed as

$$\boldsymbol{\Pi}^{\perp}_{\mathbf{U}^T} = \mathbf{I} - \mathbf{U}^T(\mathbf{U}\mathbf{U}^T)^{-1}\mathbf{U} \ . \tag{18}$$

Let $\boldsymbol{\Pi}^{\perp}_{\mathbf{U}^T}$ operate on both sides of (8). Then, in the absence of noise

$$\mathbf{Y}\boldsymbol{\Pi}^{\perp}_{\mathbf{U}^T} = \boldsymbol{\Gamma}_{\alpha}\mathbf{X}\boldsymbol{\Pi}^{\perp}_{\mathbf{U}^T} \ . \tag{19}$$

Provided the matrix $\mathbf{X}\boldsymbol{\Pi}^{\perp}_{\mathbf{U}^T}$ has full rank (conditions for this will be discussed in Section 3.3), the space spanned by the columns of $\boldsymbol{\Gamma}_{\alpha}$, often termed the signal subspace, is spanned by $\mathbf{Y}\boldsymbol{\Pi}^{\perp}_{\mathbf{U}^T}$. In the basic 4SID approach, the effect of noise is reduced by using a large value of α (the number of block rows in $\mathbf{Y}$) and by estimating the signal subspace from the n principal left singular vectors of $\mathbf{Y}\boldsymbol{\Pi}^{\perp}_{\mathbf{U}^T}$. As shown in [35, 40], this procedure will produce a statistically consistent estimate of the range space of the extended observability matrix, $\Re\{\boldsymbol{\Gamma}_{\alpha}\}$, only if the covariance matrix of $\mathbf{n}(t)$ is proportional to the identity matrix.

3.2 Instrumental Variables Approach

In this section we will discuss how to use the instrumental variable concept in order to remove the problems in the basic 4SID approach. The connection between instrumental variable methods and 4SID methods has been pointed out in [1, 33, 34, 35, 36]. We will here extend the presentation in [19, 41], where an IV interpretation of (a variation of) the N4SID method of [30] is presented. The corresponding analysis also shows the somewhat implicit relation to the work of Larimore [13, 14].

Recall the output model (8). Our objective is to estimate the range space of the observability matrix given the input-output data only. Thus, we must eliminate both the effect of noise and that of the input signal from $\mathbf{Y}$. As seen earlier, the input signal is removed simply by projecting $\mathbf{Y}$ onto the null-space of $\mathbf{U}$. The trick to reduce the effect of noise is to correlate the projected output with an *instrumental variable* (IV) vector. For this to work, the noise variables in the output must be uncorrelated with those of the IV's. On the other hand, the informative part of $\mathbf{Y}$ (the signal subspace) must remain undestroyed. This can be achieved by partitioning the input and output vectors/matrices into the *past* and the *future* parts. We can then use the future output as an observation vector, and past outputs and inputs as IV's. Define the partitionings as

$$\mathbf{y}(t) = \begin{bmatrix} \mathbf{y}_p(t) \\ \mathbf{y}_f(t) \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} \mathbf{Y}_p \\ \mathbf{Y}_f \end{bmatrix}, \tag{20}$$

and similarly for $\mathbf{u}(t)$, $\mathbf{U}$, $\mathbf{n}(t)$ and $\mathbf{N}$. The number of block rows in $\mathbf{y}_p(t)$ and $\mathbf{Y}_p$ is denoted β, and the number of block rows in $\mathbf{y}_f(t)$ and $\mathbf{Y}_f$ is then $\gamma = \alpha - \beta$.

In selecting γ and β, we require that $\gamma > n$ and [38] gives the necessary condition $\beta > n/(m+l)$. In [8] the connection to dead-beat observers is studied, which for the case $l = 1$ leads to the choice $\beta = n$. The model for the future outputs is

$$\mathbf{Y}_f = \boldsymbol{\Gamma}_\gamma \mathbf{X}_f + \boldsymbol{\Phi}_\gamma \mathbf{U}_f + \mathbf{N}_f \ , \tag{21}$$

where $\mathbf{X}_f = [x_{\beta+1}, \ldots, x_{\beta+N}]$.

As in the basic 4SID approach, the first step is to eliminate the input from (21). Thus, define the orthogonal projection

$$\boldsymbol{\Pi}^\perp_{\mathbf{U}_f^T} = \mathbf{I} - \mathbf{U}_f^T(\mathbf{U}_f\mathbf{U}_f^T)^{-1}\mathbf{U}_f \ . \tag{22}$$

The projected data equation is

$$\mathbf{Y}_f\boldsymbol{\Pi}^\perp_{\mathbf{U}_f^T} = \boldsymbol{\Gamma}_\gamma\mathbf{X}_f\boldsymbol{\Pi}^\perp_{\mathbf{U}_f^T} + \mathbf{N}_f\boldsymbol{\Pi}^\perp_{\mathbf{U}_f^T} \ . \tag{23}$$

Now, by assumption we have

$$\mathrm{E}[\mathbf{n}_f(t)\mathbf{u}^T(t)] = 0 \tag{24}$$

$$\mathrm{E}[\mathbf{n}_f(t)\mathbf{y}_p^T(t)] = 0 \ , \tag{25}$$

which means that $\mathbf{u}(t)$ and $\mathbf{y}_p(t)$ are good candidates for IV's. However, due to the $\boldsymbol{\Pi}^\perp_{\mathbf{U}_f^T}$-factor in (23), it only makes sense to use $\mathbf{u}_p(t)$ and $\mathbf{y}_p(t)$. By ergodicity it follows from (24) and (25) that

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{N}_f\boldsymbol{\Pi}^\perp_{\mathbf{U}_f^T}\mathbf{U}_p^T = \bar{E}[\mathbf{n}_f\mathbf{u}_p^T] - \bar{E}[\mathbf{n}_f\mathbf{u}_f^T]\left(\bar{E}[\mathbf{u}_f\mathbf{u}_f^T]\right)^{-1}\bar{E}[\mathbf{u}_f\mathbf{u}_p^T] = 0 \tag{26}$$

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{N}_f\boldsymbol{\Pi}^\perp_{\mathbf{U}_f^T}\mathbf{Y}_p^T = \bar{E}[\mathbf{n}_f\mathbf{y}_p^T] - \bar{E}[\mathbf{n}_f\mathbf{u}_f^T]\left(\bar{E}[\mathbf{u}_f\mathbf{u}_f^T]\right)^{-1}\bar{E}[\mathbf{u}_f\mathbf{y}_p^T] = 0 \ , \tag{27}$$

where the limits are taken w.p.1. The relations (26)–(27) imply that the IV matrix

$$\mathbf{P} = \begin{bmatrix} \mathbf{U}_p \\ \mathbf{Y}_p \end{bmatrix} = [\mathbf{p}(1), \ldots, \mathbf{p}(N)] \tag{28}$$

can be used for asymptotically decorrelating the noise from $\mathbf{Y}_f \boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T}$. We have

$$\begin{aligned} &\lim_{N\to\infty} \frac{1}{N} \mathbf{Y}_f \boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T} \mathbf{P}^T = \\ &= \boldsymbol{\Gamma}_\gamma \left(\bar{E}[x_{f,t}\mathbf{p}^T(t)] - \bar{E}[x_{f,t}\mathbf{u}_f(t)] \left(\bar{E}[\mathbf{u}_f(t)\mathbf{u}_f^T(t)] \right)^{-1} \bar{E}[\mathbf{u}_f(t)\mathbf{p}^T(t)] \right) \\ &= \boldsymbol{\Gamma}_\gamma \left(\mathbf{r}_{x_f p} - \mathbf{r}_{x_f u_f} \mathbf{R}^{-1}_{u_f u_f} \mathbf{R}_{u_f p} \right) . \end{aligned} \tag{29}$$

Provided the matrix post multiplying $\boldsymbol{\Gamma}_\gamma$ in (29) has full rank, we conclude that a consistent estimate of the signal subspace can be obtained, for example, from the n principal left singular vectors of $\frac{1}{N} \mathbf{Y}_f \boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T} \mathbf{P}^T$. The required rank-condition is discussed in some detail in Section 3.3. It is also possible to generalize the estimate by introducing weighting matrices. The general signal subspace estimate will be based on the quantity

$$\hat{\mathbf{G}} = \frac{1}{N} \hat{\mathbf{W}}_r \; \mathbf{Y}_f \boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T} \mathbf{P}^T \; \hat{\mathbf{W}}_c \, . \tag{30}$$

The weighting matrices $\hat{\mathbf{W}}_r$ and $\hat{\mathbf{W}}_c$ have no effect in the perfect data case. However, they may influence the variance of the estimates due to noise as well as the bias emanating from under-modeling [9, 28, 31, 32]. The weightings are regarded as, possibly data-dependent, *user's choices*. For simplicity, we will define the weighting matrices as normalized, symmetric and of fixed size. In addition, the row weighting, $\hat{\mathbf{W}}_r$, must be positive definite. For some of the known methods, the number of columns of $\hat{\mathbf{W}}_c$ is allowed to increase with the number of data. However, since the resulting subspace estimate will depend on $\hat{\mathbf{W}}_c$ only through the Gramian matrix $\hat{\mathbf{W}}_c \hat{\mathbf{W}}_c^T$, we might as well define the column weighting as a symmetric square-root factor of the Gramian. In the data-dependent case, the sequences of weighting matrices $\hat{\mathbf{W}}_r$ and $\hat{\mathbf{W}}_c$ are assumed to converge (w.p.1) to the limiting matrices $\mathbf{W}_r$ and $\mathbf{W}_c$ respectively, with $\mathbf{W}_r > 0$. The column-weighting $\mathbf{W}_c$ is assumed to satisfy

$$\mathrm{rank}[\mathbf{Y}_f \boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T} \mathbf{P}^T] = \mathrm{rank}[\mathbf{Y}_f \boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T} \mathbf{P}^T \; \mathbf{W}_c \,] \, , \tag{31}$$

for which a necessary condition is that it has rank no less than n.

Different choices of weighting matrices correspond to known methods as follows (see also [31])

$$
\begin{array}{lll}
\hat{\mathbf{W}}_c = \left(\frac{1}{N}\mathbf{P}\boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T}\mathbf{P}^T\right)^{-1/2} & \hat{\mathbf{W}}_r = \mathbf{I} & \text{PO-MOESP [36]} \\
\hat{\mathbf{W}}_c = \left(\frac{1}{N}\mathbf{P}\boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T}\mathbf{P}^T\right)^{-1}\left(\frac{1}{N}\mathbf{P}\mathbf{P}^T\right)^{1/2} & \hat{\mathbf{W}}_r = \mathbf{I} & \text{N4SID [30]} \\
\hat{\mathbf{W}}_c = \left(\frac{1}{N}\mathbf{P}\mathbf{P}^T\right)^{-1/2} & \hat{\mathbf{W}}_r = \left(\frac{1}{N}\mathbf{Y}_f\boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f}\mathbf{Y}_f^T\right)^{-1/2} & \text{IVM choice [38]} \\
\hat{\mathbf{W}}_c = \left(\frac{1}{N}\mathbf{P}\boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f}\mathbf{P}^T\right)^{-1/2} & \hat{\mathbf{W}}_r = \left(\frac{1}{N}\mathbf{Y}_f\boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f}\mathbf{Y}_f^T\right)^{-1/2} & \text{CVA [14]}
\end{array}
$$

As can bee seen the PO-MOESP and N4SID weightings are very similar. The same also holds for CVA and the natural IVM choice.

Given $\hat{\mathbf{G}}$ of (30), the signal subspace estimate is computed from a truncated singular value decomposition (SVD) as follows: Let the SVD be partitioned according to

$$\hat{\mathbf{G}} = \hat{\mathbf{Q}}_s\hat{\mathbf{S}}_s\hat{\mathbf{V}}_s^T + \hat{\mathbf{Q}}_n\hat{\mathbf{S}}_n\hat{\mathbf{V}}_n^T \ , \tag{32}$$

where $\hat{\mathbf{S}}_s$ is a diagonal matrix containing the n largest singular values, and the columns of $\hat{\mathbf{Q}}_s$ are the corresponding left singular vectors. Let $\mathbf{Q}_s$ denote the limiting (as $N \to \infty$) value of $\hat{\mathbf{Q}}_s$ (w.p.1). To ensure the existence of a unique limit, one has to assume that the corresponding n singular values are distinct. Then, it follows from (29) that

$$\mathbf{W}_r\boldsymbol{\Gamma}_\gamma = \mathbf{Q}_s\mathbf{T} \ ,$$

for some full-rank $n \times n$ matrix $\mathbf{T}$. This matrix corresponds only to a change of basis in the state-space, and we will take the observability matrix estimate as

$$
\begin{aligned}
\widehat{\operatorname{span}\{\boldsymbol{\Gamma}_\gamma\}} &= \operatorname{span}\{\hat{\mathbf{Q}}_s^W\} \ , && (33)\\
\hat{\mathbf{Q}}_s^W &= \hat{\mathbf{W}}_r^{-1}\hat{\mathbf{Q}}_s \ . && (34)
\end{aligned}
$$

Under the rank condition pointed out in (29), this estimate is guaranteed to converge to the true observability matrix *for some* state space realization. This realization depends, among other things, upon the input sequence to the system, see [28, 32].

3.3 Asymptotic Properties of the Subspace Estimate

As mentioned in the previous section it is required that

$$\operatorname{rank}\left(\mathbf{r}_{xfp} - \mathbf{r}_{xfu_f}\mathbf{R}^{-1}_{u_fu_f}\mathbf{R}_{u_fp}\right) = n \ , \tag{35}$$

in order for the IV-4SID method to produce a consistent signal subspace estimate. This condition is a combined persistence of excitation (PE) and identifiability condition. For the basic 4SID approach the corresponding condition can be expressed

$$\bar{E}\begin{bmatrix}\mathbf{u}(t)\\ x_t\end{bmatrix}\begin{bmatrix}\mathbf{u}(t)\\ x_t\end{bmatrix}^T > 0\ . \tag{36}$$

This holds under some controllability condition on the system and PE condition on the input, see [15, 37]. It seems more difficult to derive a corresponding explicit condition for the IV-4SID method. Notice however that we can relate the condition (35) to the cross-covariance matrix

$$\bar{E}\begin{bmatrix}\mathbf{u}_f(t)\\ x_f\end{bmatrix}\begin{bmatrix}\mathbf{u}_f(t)\\ (\mathbf{t})\end{bmatrix}^T = \begin{bmatrix}\mathbf{R}_{u_fu_f} & \mathbf{R}_{u_fp}\\ \mathbf{r}_{x_fu_f} & \mathbf{r}_{x_fp}\end{bmatrix}\ . \tag{37}$$

Then, $\mathbf{r}_{x_fp} - \mathbf{r}_{x_fu_f}\mathbf{R}_{u_fu_f}^{-1}\mathbf{R}_{u_fp}$ is the Schur complement of the the matrix (37). Since,

$$\operatorname{rank}\left(\mathbf{r}_{x_fp} - \mathbf{r}_{x_fu_f}\mathbf{R}_{u_fu_f}^{-1}\mathbf{R}_{u_fp}\right) + \operatorname{rank}\left(\mathbf{R}_{u_fu_f}\right) = \operatorname{rank}\begin{bmatrix}\mathbf{R}_{u_fu_f} & \mathbf{R}_{u_fp}\\ \mathbf{R}_{pu_f} & \mathbf{r}_{x_fp}\end{bmatrix}\ , \tag{38}$$

we can rewrite the condition (35) as

$$\mathbf{R}_{u_fu_f} \;>\; 0 \tag{39}$$

$$\operatorname{rank}\begin{bmatrix}\mathbf{R}_{u_fu_f} & \mathbf{R}_{u_fp}\\ \mathbf{r}_{x_fu_f} & \mathbf{r}_{x_fp}\end{bmatrix} \;\geq\; n+\gamma l\ . \tag{40}$$

How to make this condition more explicit is an open problem. Also, notice that introduction of column weighting gives the extra condition (31). However, whenever these are satisfied, the relation (29) shows that the range space of the matrix $\hat{\mathbf{Q}}_s^W$ converges to that of $\boldsymbol{\Gamma}_\gamma$ w.p.1 as $N \to \infty$.

The following theorem, which generalizes previous results reported in [40, 41], gives the large sample accuracy of the signal subspace estimate. Only the part of $\hat{\mathbf{Q}}_s^W$ which is orthogonal to $\boldsymbol{\Gamma}_\gamma$ is of interest.

Theorem 1 *Let the deterministic vectors* $\boldsymbol{\xi}$ *and* $\boldsymbol{\eta}$ *satisfy*

$$\boldsymbol{\xi}^*\boldsymbol{\Gamma}_\gamma = \boldsymbol{\eta}^*\boldsymbol{\Gamma}_\gamma = 0\ , \tag{41}$$

where $(\cdot)^*$ *denotes complex conjugate transpose*[4]*. Then,* $\sqrt{N}\,[\hat{\mathbf{Q}}_s^W]^T\boldsymbol{\xi}$ *and* $\sqrt{N}\,[\hat{\mathbf{Q}}_s^W]^T\boldsymbol{\eta}$ *are asymptotically jointly Gaussian distributed with cross covariance elements*

$$\mathbf{C}(\boldsymbol{\xi},\boldsymbol{\eta}) = \sum_{|\tau|<\gamma}\mathbf{H}^T\mathbf{R}_{\zeta\zeta}(\tau)\mathbf{H}\ \boldsymbol{\eta}^*\mathbf{R}_{n_fn_f}(-\tau)\boldsymbol{\xi}\ , \tag{42}$$

[4] We will later deal with the diagonal form of the system, implying that complex matrices may appear. The input-output data is, however, always assumed to be real.

where

$$
\begin{aligned}
\mathbf{Z} &= \begin{bmatrix} \mathbf{U}_f \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_f \\ \mathbf{U}_p \\ \mathbf{Y}_p \end{bmatrix} = [\boldsymbol{\zeta}(1), \ldots, \boldsymbol{\zeta}(N)] && (43) \\
\mathbf{H} &= \begin{bmatrix} -\mathbf{R}_{u_f u_f}^{-1} \mathbf{R}_{u_f p} \\ \mathbf{I} \end{bmatrix} \mathbf{W}_c \mathbf{V}_s \mathbf{S}_s^{-1} \, . && (44)
\end{aligned}
$$

The proof is given in Appendix A.1. This result forms the basis for statistical analysis of subspace-based methods and it is based on extending ideas from the corresponding sensor array subspace estimation problem (see e.g. [3]). The case $\boldsymbol{\xi} = \boldsymbol{\eta}$ is of course a special case for which

$$\sqrt{N}\, [\hat{\mathbf{Q}}_s^W]^T \boldsymbol{\xi} \in \mathrm{As}\mathcal{N}(0, \mathbf{C}(\boldsymbol{\xi}, \boldsymbol{\xi})) \, . \tag{45}$$

4 Estimation of System Matrices

The previous section treated the estimation of the signal subspace. In the following, two different techniques for extracting the system matrices given the signal subspace estimate are presented. The first of these is the popular shift-structure based approach, whereas the second takes the statistical properties of $\hat{\mathbf{Q}}_s^W$, as presented in the previous section, into account.

4.1 Shift-Structure Approach

The major interest herein is estimation of the system matrix, A. The most often used technique in the 4SID methods is taken from the Kung algorithm [12]. It is based on the shift-invariance structure of the observability matrix (5). Clearly, the matrix C can be extracted from the first block-row (first l rows) of $\boldsymbol{\Gamma}_\gamma$. Introduce the selection matrices

$$\mathbf{J}_1 = [\mathbf{I}_{(\gamma-1)l} \; 0_{(\gamma-1)l\times l}], \quad \mathbf{J}_2 = [0_{(\gamma-1)l\times l} \; \mathbf{I}_{(\gamma-1)l}] \; . \tag{46}$$

Then $\mathbf{J}_1\boldsymbol{\Gamma}_\gamma$ is the upper $\gamma - 1$ blocks of $\boldsymbol{\Gamma}_\gamma$, whereas $\mathbf{J}_2\boldsymbol{\Gamma}_\gamma$ consists of the lower $\gamma - 1$ blocks. It follows from (5) that

$$\mathbf{J}_1\boldsymbol{\Gamma}_\gamma A = \mathbf{J}_2\boldsymbol{\Gamma}_\gamma \; . \tag{47}$$

Now, assume that an estimate of the observability matrix (for some state space realization), denoted $\hat{\mathbf{Q}}_s^W$, is available. Then, a natural estimate of A is to solve (47) in a least-squares (LS) sense

$$\hat{A} = [\mathbf{J}_1\hat{\mathbf{Q}}_s^W]^\dagger \mathbf{J}_2\hat{\mathbf{Q}}_s^W \; , \tag{48}$$

where $[\mathbf{J}_1\hat{\mathbf{Q}}_s^W]^\dagger$ denotes the pseudo-inverse of $\mathbf{J}_1\hat{\mathbf{Q}}_s^W$. This estimate will be called the IV-ESP estimate, since this approach for estimating A from $\hat{\mathbf{Q}}_s^W$ is closely related to the ESPRIT (estimation of signal parameters via rotational invariance techniques) algorithm for sensor array signal processing, [21]. Following [21], a total-least squares (TLS) solution [6] can also be applied for solving (48). While the TLS solution has the potential of improving the accuracy in small samples, it can be shown as in [20, 25] that the TLS and LS estimates have the same asymptotic (for large N) properties.

Given estimates of A and C, the problem of estimating B and D can be reduced to a linear regression problem [17, 28]. It is straightforward to show that the input-output relation can be expressed as

$$y_t = \hat{C}\hat{A}^t x_0 + (u_t^T \otimes \mathbf{I}_l)\mathrm{vec}(\mathrm{D}) + \left(\sum_{k=0}^{t-1} u_t^T \otimes (\hat{C}\hat{A}^{t-k-1})\right)\mathrm{vec}(\mathrm{B}) + n_t \; , \tag{49}$$

where $\mathbf{I}_l$ denotes the $l \times l$ identity matrix, $\otimes$ is the Kronecker product and n_t represents the noise contribution. Based on (49), an LS problem can be formulated, which leads to estimates of B and D. The asymptotic properties of the corresponding estimates remain to be investigated. Notice also that the LS problem will be ill-conditioned for large N, because the initial state cannot be consistently estimated. This effect can be removed by setting $x_0 = 0$ for large N. Another problem with the above estimate is that the residuals are generally not white. Hence, better accuracy can be obtained by pre-whitening n_t by a noise filter.

4.2 Asymptotic Distribution of IV-ESP Pole Estimates

The IV-ESP estimator is rather ad-hoc from a statistical point of view. It is therefore of great interest to compare the statistical properties of the algorithm with that of other subspace based and instrumental variable techniques, as well as with more traditional prediction error methods. Toward this end, we present an asymptotic analysis of the IV-ESP method. To obtain a canonical parameter set we choose to consider the accuracy of the system poles only. In the following, it is assumed that the system is given in its diagonal form, *i.e.* $A = A_d$ is diagonal with the system poles, μ_k on the diagonal.

Since Theorem 1 establishes the asymptotic distribution of the estimated signal subspace matrix, we only only need to relate the error in the estimated poles to that of $\hat{\mathbf{Q}}_s^W$. Let $\{\hat{\mu}_k\}$ denote the pole estimate corresponding to μ_k, and define the error as

$$\tilde{\mu}_k = \hat{\mu}_k - \mu_k \ . \tag{50}$$

We shall make one simplification, namely to assume that $\mathbf{W}_r = \mathbf{I}$. The reason is that in [9, 10] it is shown that this weighting does not influence the variance of the IV-ESP estimated poles.

Following [40] (see also [24]), we can then relate the error in the estimated system poles to the subspace estimate as

$$\tilde{\mu}_k \simeq \mathbf{f}_k^* \hat{\mathbf{Q}}_s \mathbf{Q}_s^T \boldsymbol{\Gamma}_{.k} \ , \tag{51}$$

where $\simeq$ denotes $o_p(1/\sqrt{N})$ approximation[5], $\boldsymbol{\Gamma}_{.k}$ is the k^{th} column of $\boldsymbol{\Gamma}_\gamma$ and

$$\mathbf{f}_k^* = \{\boldsymbol{\Gamma}_1\}_{k.}^{\dagger} (\mathbf{J}_2 - \mathbf{J}_1 \mu_k) \ . \tag{52}$$

[5]The symbols $O_p(\cdot)$ and $o_p(\cdot)$ denote "in-probability" versions of the corresponding deterministic notation.

Here, $\boldsymbol{\Gamma}_i = \mathbf{J}_i \boldsymbol{\Gamma}_\gamma, \; i = 1, 2$, where the selection matrices $\mathbf{J}_1$ and $\mathbf{J}_2$ are defined by (46). Note that $\mathbf{f}_k$ is orthogonal to $\boldsymbol{\Gamma}_\gamma$ and, hence, to $\mathbf{Q}_s$:

$$\mathbf{f}_k^* \boldsymbol{\Gamma}_\gamma = \{\boldsymbol{\Gamma}_1\}_{k.}^\dagger (\boldsymbol{\Gamma}_1 A_d - \boldsymbol{\Gamma}_1 \mu_k) = 0. \tag{53}$$

Taking $\boldsymbol{\xi} = \mathbf{f}_k$ and $\boldsymbol{\eta} = \mathbf{f}_l$ in Theorem 1, it follows that $\sqrt{N}\,\mathbf{f}_k^* \hat{\mathbf{Q}}_s$ has a limiting zero-mean Gaussian distribution with covariance elements

$$\begin{aligned} N\,\mathrm{E}[(\mathbf{f}_k^* \hat{\mathbf{Q}}_s)^T (\mathbf{f}_l^* \hat{\mathbf{Q}}_s)] &= \sum_{|\tau|<\gamma} \mathbf{H}^T \mathbf{R}_\zeta(\tau) \mathbf{H}\, \mathbf{f}_l^* \mathbf{R}_{n_f}(-\tau) \bar{\mathbf{f}}_k \\ &= \sum_{|\tau|<\gamma} \mathbf{H}^T \mathbf{R}_\zeta(\tau) \mathbf{H}\, \mathbf{f}_k^* \mathbf{R}_{n_f}(\tau) \bar{\mathbf{f}}_l \;, \end{aligned} \tag{54}$$

where $\mathbf{H}$ is defined in (44) and $\bar{(\cdot)}$ denotes complex conjugate. Define the vector $\mathbf{b}_k$ by

$$\mathbf{b}_k = \mathbf{H} \mathbf{Q}_s^T \boldsymbol{\Gamma}_{.k} = \begin{bmatrix} -\mathbf{R}_{uu}^{-1} \mathbf{R}_{u_f p} \\ \mathbf{I} \end{bmatrix} \mathbf{W}_c \mathbf{G}^\dagger \boldsymbol{\Gamma}_{.k} \;, \tag{55}$$

for $k = 1, \ldots, n$. Inserting (54)–(55) into (51), the covariance elements of the asymptotic distribution of the pole estimates are then obtained as

$$N\,\mathrm{E}[\tilde{\mu}_k \tilde{\mu}_l^*] = \sum_{|\tau|<\gamma} \mathbf{b}_k^T \mathbf{R}_\zeta(\tau) \bar{\mathbf{b}}_l\, \mathbf{f}_k^* \mathbf{R}_{n_f}(\tau) \mathbf{f}_l \tag{56}$$

$$N\,\mathrm{E}[\tilde{\mu}_k \tilde{\mu}_l] = \sum_{|\tau|<\gamma} \mathbf{b}_k^T \mathbf{R}_\zeta(\tau) \mathbf{b}_l\, \mathbf{f}_k^* \mathbf{R}_{n_f}(\tau) \bar{\mathbf{f}}_l \;. \tag{57}$$

It would be interesting to analyze these expressions with respect to column weighting to compare the different choices. This can of course be done on given examples, but a general conclusion has not yet been found. Notice also that the choice of input signal influences the estimation accuracy, and that for a white noise input all the proposed weightings give the same result.

4.3 Subspace-Fitting Approach

Note that the IV-ESP method does not exploit the full structure of the extended observability matrix. In [26] a technique is proposed which attempts to further use the shift invariance structure of $\boldsymbol{\Gamma}_\gamma$. However, these methods do not take the statistical properties of $\hat{\mathbf{Q}}_s^W$ into account. Below, a novel technique for statistically optimal estimation of A, based only on $\hat{\boldsymbol{\Gamma}}_\gamma$ is proposed. Note, however, that e.g. [30, 37] base the system matrix estimates not only on $\hat{\boldsymbol{\Gamma}}_\gamma$, but parts of the data are also re-used.

The transfer function matrix of the state-space model (1)–(2), without direct term D, equals

$$G(z) = C(zI - A)^{-1}B, \tag{58}$$

and can be represented by a matrix fraction description (MFD) [11]

$$G(z) = D^{-1}(z)\, N(z), \tag{59}$$

where

$$\begin{aligned} D(z) &= \sum_{k=0}^{r} D_{r-k} z^k, \quad \dim D_k = l \times l & (60) \\ N(z) &= \sum_{k=0}^{r} N_{r-k} z^k, \quad \dim N_k = l \times m\ . & (61) \end{aligned}$$

The coefficients of $D(z)$ are related to the observability matrix through

$$[D_r\ D_{r-1}\ \cdots\ D_0]\, \boldsymbol{\Gamma}_{r+1} = 0\ . \tag{62}$$

This means that the coefficients of $D(z)$ express certain linear dependence relations among the rows of $\boldsymbol{\Gamma}_{r+1}$. The result (62) follows by analyzing the relationship

$$C[zI - A]^{-1} = D(z)^{-1}\tilde{N}(z), \text{ where } N(z) = \tilde{N}(z)B\ . \tag{63}$$

A formal geometric series expansion of $[zI - A]^{-1}$ shows that

$$D(z)C[z^{-1}I + z^{-2}A + \cdots] = \tilde{N}(z) = \text{ a polynomial.} \tag{64}$$

Equating all coefficients of the left hand side corresponding to z^{-k}, $k > 0$, to zero leads to (62) (see also [11, page 479], where a similar relation with the controllability matrix is derived).

Note that by (62) the matrix coefficients D_k of $D(z)$ parameterize parts of the null-space of $\boldsymbol{\Gamma}_{r+1}$. The idea of the estimation procedure to be presented is now to extend the above relation so that we get a linear parameterization of the entire null-space of $\boldsymbol{\Gamma}_\gamma$! Briefly, we wish to obtain the following:

- Extend the relation (62) to

  $$\mathbf{D}_\gamma(\boldsymbol{\theta})\boldsymbol{\Gamma}_\gamma = 0\ ,$$

  where $\mathbf{D}_\gamma(\boldsymbol{\theta})$ is a full rank $(\gamma l - n) \times \gamma l$ matrix (because the dimension of the null-space of $\boldsymbol{\Gamma}_\gamma$ is $\gamma l - n$), which depends linearly upon some minimal set of parameters $\boldsymbol{\theta}$.

- Estimate the parameters $\boldsymbol{\theta}$ by solving

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{D}_\gamma(\boldsymbol{\theta})\hat{\boldsymbol{\Gamma}}_\gamma\| \ ,$$

  where the norm is selected to optimize the estimation accuracy, and also to allow an explicit solution w.r.t. $\boldsymbol{\theta}$.

For the single output case, $D(z)$ is the characteristic polynomial, $a(z)$, of A. An optimal subspace-fitting method for this case is presented in [19]. In an attempt to generalize to the multiple output case, we propose to use the Hermite form [11]. Denote the entries of $D(z)$ by $D_{ij}(z)$,

$$D(z) = \{D_{ij}(z)\} \quad i = 1, \ldots, l \quad j = 1, \ldots, l \ . \tag{65}$$

In the Hermite form, $D(z)$ is lower triangular, with monic polynomials as diagonal elements. In addition, each diagonal entry must have the highest degree in its column. The degrees in question are the so-called observability indices $\{k_1 \ldots k_l\}$, with $k_i = \deg(D_{ii}(z))$, and they are ordered in non-increasing order so that $r = k_1$. The order of the corresponding system equals $n = \sum_{i=1}^{l} k_i$. Given the indices, this parameterization is canonical and is based on as few parameters as possible. The coefficients of the scalar polynomials $D_{ij}(z)$ are denoted as

$$\begin{aligned} D_{ii}(z) &= z^{k_i} + d_{ii1}z^{k_i-1} + \cdots + d_{iik_i} & (66) \\ D_{ij}(z) &= d_{ij1}z^{k_j-1} + d_{ij2}z^{k_j-2} + \cdots + d_{ijk_j} \quad i \neq j \, . & (67) \end{aligned}$$

The relation (62) now gives a direct linear parameterization of parts of the null-space of $\boldsymbol{\Gamma}_r$, in terms of the coefficients d_{ijk}. Using the shift structure of the observability matrix, we can easily extend this parameterization to a general γ

$$\gamma - r \text{ block rows} \left\{ \begin{bmatrix} D_r & D_{r-1} & \ldots & D_0 & 0 & \ldots & 0 \\ 0 & D_r & \ldots & D_1 & D_0 & \ldots & 0 \\ \vdots & & \ddots & & & \ddots & \vdots \\ 0 & \ldots & & D_r & D_{r-1} & \ldots & D_0 \end{bmatrix} \right. \boldsymbol{\Gamma}_\gamma = 0 \, . \tag{68}$$

The question is whether or not a basis for the entire null-space is obtained. Assuming that the left hand side matrix has full rank (it has $(\gamma - r)l$ rows) gives the necessary condition $rl = n$. Let us illustrate the potential problems with the help of a simple example:

Example 4.1

Assume $l = 2$, $n = 2$, $k_1 = 1$ and $k_2 = 1$, which gives $r = 1$ and the necessary condition is satisfied. Then,

$$D(z) = \begin{bmatrix} z + d_{111} & 0 \\ d_{211} & z + d_{221} \end{bmatrix} \tag{69}$$

and (62) with $r = 1$ equals

$$\begin{bmatrix} d_{111} & 0 & 1 & 0 \\ d_{211} & d_{221} & 0 & 1 \end{bmatrix} \boldsymbol{\Gamma}_2 = 0 \;, \quad \boldsymbol{\Gamma}_2 = \begin{bmatrix} c_1 \\ c_2 \\ c_1 A \\ c_2 A \end{bmatrix} . \tag{70}$$

The rank of the null-space of $\boldsymbol{\Gamma}_2$ equals $4 - 2 = 2$, and is hence completely parameterized by the matrix

$$\begin{bmatrix} d_{111} & 0 & 1 & 0 \\ d_{211} & d_{221} & 0 & 1 \end{bmatrix} , \tag{71}$$

which is of rank 2! For a general γ we just use the shift structure to extend this matrix. For example, $\gamma = 4$ gives

$$\begin{bmatrix} d_{111} & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ d_{211} & d_{221} & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_{111} & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & d_{211} & d_{221} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{111} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & d_{211} & d_{221} & 0 & 1 \end{bmatrix} \boldsymbol{\Gamma}_4 = 0 \;. \tag{72}$$

The left matrix has rank 6, which equals the rank of the null-space of $\boldsymbol{\Gamma}_4$: $8 - 2 = 6$.

Let us instead consider the situation when $k_1 = 2$ and $k_2 = 0$. Then $r = 2$ and the necessary condition is not satisfied. Here,

$$D(z) = \begin{bmatrix} z^2 + d_{111} z + d_{112} & 0 \\ d_{211} z + d_{212} & 1 \end{bmatrix} \tag{73}$$

and (62) with $r = 2$ reads

$$\begin{bmatrix} d_{112} & 0 & d_{111} & 0 & 1 & 0 \\ d_{212} & 1 & d_{211} & 0 & 0 & 0 \end{bmatrix} \boldsymbol{\Gamma}_3 = 0 \;. \tag{74}$$

The rank of the null-space of $\boldsymbol{\Gamma}_3$ equals $6 - 2 = 4$, while

$$\begin{bmatrix} d_{112} & 0 & d_{111} & 0 & 1 & 0 \\ d_{212} & 1 & d_{211} & 0 & 0 & 0 \end{bmatrix} \tag{75}$$

is of rank 2! The rank cannot be increased by using the shift-structure as in (68), because $\gamma = r + 1$ in this case. Hence, we have not used all possible degrees of freedom. $\square$

As previously stated, we need to find a linear parameterization of the null-space of $\boldsymbol{\Gamma}_\gamma$. From the above example we can draw some important conclusions regarding this problem:

- The Hermite MFD gives a canonical parameterization of at least parts of the null-space.
- For some choices of indices, we can parameterize the entire null-space, but in general we do not use all degrees of freedom.

The latter problem can be overcome by adding more relations by making $\mathbf{D}_\gamma(\boldsymbol{\theta})$ nonlinear in $\boldsymbol{\theta}$, or by introducing some dummy variables. Since neither of these extensions are convenient for our purposes, we will in the following specialize to the case $k_1 = n$; which means that the system (1)-(2) is observable from the first component of the output vector[6] y_t. We can then use the Cayley-Hamilton theorem to alleviate the problems with the parameterization. More precisely, the entire null-space of $\boldsymbol{\Gamma}_\gamma$ will be obtained only if $\gamma \geq 2n$, which indeed may be a useful choice in practice. A description of how this can be done is given next.

If the system is observable from the first component of the output vector we have $k_1 = n = r$ and $k_i = 0,\ i = 2, \ldots, l$. The matrix coefficients appearing in (62) are then given by

$$D_0 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix}, \tag{76}$$

$$D_i = \begin{bmatrix} d_{11i} & 0 & \cdots & 0 \\ d_{21i} & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ d_{l1i} & 0 & \ldots & 0 \end{bmatrix} \quad i = 1, \ldots, n-1, \tag{77}$$

[6]This is not really a severe restriction, since we can use any *known* linear combination of the "physical outputs" as first output!

$$D_n = \begin{bmatrix} d_{11n} & 0 & \cdots & 0 \\ d_{21n} & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ d_{l1n} & 0 & \dots & 1 \end{bmatrix} . \tag{78}$$

Now, (62) reads

$$[D_n \; D_{n-1} \; \cdots \; D_0] \, \boldsymbol{\Gamma}_{n+1} = 0 \; . \tag{79}$$

Let c_i denotes the i:th row of C, $i = 1, \ldots, l$. The rows of the above equation are

$$d_{11n}c_1 + d_{11(n-1)}c_1 A + \cdots + d_{111}c_1 A^{n-1} + c_1 A^n = 0 \tag{80}$$

$$d_{j1n}c_1 + d_{j1(n-1)}c_1 A + \cdots + d_{j11}c_1 A^{n-1} + c_j = 0, \quad j = 2, \ldots, l \, . \tag{81}$$

The characteristic polynomial of A is the determinant of $D(z)$, which for our particular case is

$$A(z) = z^n + d_{111} z^{n-1} + \cdots + d_{11n} \; . \tag{82}$$

So, using Cayley-Hamilton's theorem,

$$A^n + d_{111} \, A^{n-1} + \cdots + d_{11n} \, I = 0 \; , \tag{83}$$

we can extend the relation (80) with, for example,

$$d_{11n}c_j + d_{11(n-1)}c_j A + \cdots + d_{111}c_j A^{n-1} + c_j A^n = 0 \; , \quad j = 2, \ldots, l \; . \tag{84}$$

Using (80) and (84) we can now construct a suitable $\mathbf{D}_\gamma(\boldsymbol{\theta})$. In order to clarify the rank constraint we will permute the rows of $\boldsymbol{\Gamma}_\gamma$ (this does not change its rank):

$$\tilde{\boldsymbol{\Gamma}}_\gamma = \begin{bmatrix} c_1 \\ c_1 A \\ \vdots \\ c_1 A^{\gamma-1} \\ c_2 \\ \vdots \\ c_2 A^{\gamma-1} \\ \vdots \\ c_l \\ \vdots \\ c_l A^{\gamma-1} \end{bmatrix} . \tag{85}$$

Define the $(\gamma - n) \times \gamma$ Toeplitz matrix

$$\mathcal{D}_1 = \begin{bmatrix} d_{11n} & d_{11(n-1)} & \cdots & d_{111} & 1 & 0 & \cdots & 0 \\ 0 & d_{11n} & \cdots & d_{112} & d_{111} & 1 & \cdots & 0 \\ \vdots & & \ddots & & & & \ddots & \vdots \\ 0 & & & d_{11n} & d_{11(n-1)} & \cdots & & 1 \end{bmatrix} \tag{86}$$

and the $n \times \gamma$ Toeplitz matrices

$$\mathcal{D}_j = \begin{bmatrix} d_{j1n} & d_{j1(n-1)} & \cdots & d_{j11} & 0 & 0 & \cdots & 0 & 0 \\ 0 & d_{j1n} & \cdots & d_{j12} & d_{j11} & 0 & \cdots & 0 & 0 \\ \vdots & & \ddots & & & \ddots & & \vdots & \vdots \\ 0 & & & d_{j1n} & d_{j1(n-1)} & \cdots & & d_{j11} & 0 \end{bmatrix}, \tag{87}$$

for $j = 2, \ldots, l$. The rightmost block of zeros in (87) extends over $\gamma - 2n + 1$ columns, which requires that $\gamma \geq 2n - 1$. Finally, define the $(\gamma l - n) \times (\gamma l)$ matrix

$$\tilde{\mathbf{D}}_\gamma(\theta) = \begin{bmatrix} \mathcal{D}_1 & 0 & 0 & 0 & \ldots & 0 \\ \mathcal{D}_2 & I_n & 0 & 0 & \ldots & 0 \\ 0 & \mathcal{D}_1 & 0 & 0 & \ldots & 0 \\ \mathcal{D}_3 & 0 & I_n & 0 & \ldots & 0 \\ 0 & 0 & \mathcal{D}_1 & 0 & \ldots & 0 \\ \vdots & & & & \ddots & \vdots \\ \mathcal{D}_l & 0 & 0 & \ldots & 0 & I_n \\ 0 & & \ldots & & 0 & \mathcal{D}_1 \end{bmatrix}. \tag{88}$$

Then, (80) and (84) imply that

$$\tilde{\mathbf{D}}_\gamma(\theta)\tilde{\boldsymbol{\Gamma}}_\gamma = 0 \ . \tag{89}$$

The important observation is that, by construction, $\tilde{\mathbf{D}}_\gamma(\theta)$ has full row-rank $\gamma l - n$, which exactly matches the dimension of the null-space of $\boldsymbol{\Gamma}_\gamma$. It is now straightforward to shift back the columns (does not change the rank) of $\tilde{\mathbf{D}}_\gamma(\theta)$ to give a $\mathbf{D}_\gamma(\theta)$ satisfying the desired relations

$$\begin{aligned} \mathbf{D}_\gamma(\theta)\boldsymbol{\Gamma}_\gamma &= 0 \qquad (90) \\ \operatorname{rank}(\mathbf{D}_\gamma(\theta)) &= \gamma l - n \ . \qquad (91) \end{aligned}$$

Thus, we have found the desired parameterization for the case of $k_1 = n$ and $\gamma \geq 2n - 1$. Notice that this covers the special case of a single output system $D(z) = a(z)$. Also, notice that the first n components of $\boldsymbol{\theta}$ are the coefficients of the characteristic polynomial. Thus, only this part has to be determined explicitly if we are interested in analyzing the poles of the system.

The estimation method is based on the equation

$$\mathbf{D}_\gamma(\boldsymbol{\theta})\hat{\mathbf{Q}}_s^W = \mathcal{E} \ . \tag{92}$$

The above relation may be vectorized as

$$\mathrm{vec}(\mathbf{D}_\gamma(\boldsymbol{\theta})\hat{\mathbf{Q}}_{\mathrm{s}}^{\mathrm{W}}) = \mathrm{vec}(\mathcal{E}) \ . \tag{93}$$

Let the covariance matrix of the normalized vectorized error be

$$\mathbf{R}_N(\boldsymbol{\theta}) = \left[\mathrm{Cov}\left(\sqrt{N}\mathrm{vec}(\mathcal{E})\right)\right] \ , \tag{94}$$

where the factor $\sqrt{N}$ has been introduced to make $\mathbf{R}_N(\boldsymbol{\theta}) = O(1)$. An asymptotically best consistent (ABC) estimate of $\boldsymbol{\theta}$ is then obtained by solving

$$\hat{\boldsymbol{\theta}}_{\mathrm{opt}} = \arg\min_{\boldsymbol{\theta}} \mathrm{vec}\left(\mathbf{D}_\gamma(\boldsymbol{\theta})\hat{\mathbf{Q}}_{\mathrm{s}}^{\mathrm{W}}\right)^{\mathrm{T}} \mathbf{R}_{\mathrm{N}}(\boldsymbol{\theta})^{-1}\mathrm{vec}\left(\mathbf{D}_\gamma(\boldsymbol{\theta})\hat{\mathbf{Q}}_{\mathrm{s}}^{\mathrm{W}}\right) \ . \tag{95}$$

An approximate expression for the optimal weighting matrix, $\mathbf{R}_N(\boldsymbol{\theta})^{-1}$, is provided by the following theorem.

Theorem 2 *Let $\mathbf{D}_\gamma$ be a deterministic matrix, such that $\mathbf{D}_\gamma\boldsymbol{\Gamma}_\gamma = 0$. Then, the vectorized error is asymptotically normal, with covariance matrix $\mathbf{R}(\boldsymbol{\theta})$, i.e.,*

$$\sqrt{N}\,\mathrm{vec}(\mathbf{D}_\gamma\hat{\mathbf{Q}}_{\mathrm{s}}^{\mathrm{W}}) \in \mathrm{As}\mathcal{N}(0, \mathbf{R}(\boldsymbol{\theta}))$$

where

$$\mathbf{R}(\boldsymbol{\theta}) = \sum_{|\tau|<\gamma} (\mathbf{H}^T\mathbf{R}_{\zeta\zeta}(\tau)\mathbf{H}) \otimes (\mathbf{D}_\gamma\mathbf{R}_{n_f n_f}(\tau)\mathbf{D}_\gamma^T) \tag{96}$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{U}_f \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_f \\ \mathbf{U}_p \\ \mathbf{Y}_p \end{bmatrix} = [\boldsymbol{\zeta}(1), \ldots, \boldsymbol{\zeta}(N)] \tag{97}$$

$$\mathbf{H} = \begin{bmatrix} -\mathbf{R}_{u_f u_f}^{-1}\mathbf{R}_{u_f p} \\ \mathbf{I} \end{bmatrix} \mathbf{W}_c\mathbf{V}_s\mathbf{S}_s^{-1} \ . \tag{98}$$

Proof See Appendix A.2. $\square$

Replacing $\mathbf{R}_N(\boldsymbol{\theta})$ in (95) with its asymptotic value $\mathbf{R}(\boldsymbol{\theta})$ in (96) will not change the asymptotic properties of the parameter estimate. Furthermore, to remove the $\boldsymbol{\theta}$ dependence of the weighting matrix, any $O(1/\sqrt{N})$ consistent estimate of the true $\mathbf{R} = \mathbf{R}(\boldsymbol{\theta})$ can be used without affecting the asymptotic optimality of (95), see Compliment C4.4 in [23]. We arrive at the following estimation criterion

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathrm{vec}(\mathbf{D}_\gamma(\boldsymbol{\theta})\hat{\mathbf{Q}}_{\mathrm{s}}^{\mathrm{W}})^{\mathrm{T}}\hat{\mathbf{R}}^{-1}\mathrm{vec}(\mathbf{D}_\gamma(\boldsymbol{\theta})\hat{\mathbf{Q}}_{\mathrm{s}}^{\mathrm{W}})\,, \tag{99}$$

where $\hat{\mathbf{R}}^{-1}$ denotes an estimate of the asymptotic covariance matrix.

A two-step (non-iterative!) optimal estimation procedure can now be devised as follows; obtain a consistent estimate of $\mathbf{D}_\gamma$ by using, for instance, $\hat{\mathbf{R}}^{-1} = \mathbf{I}$ in (99). The approach of Section 4.1 could also be used. Insert the resulting $\hat{\mathbf{D}}_\gamma$ in the expression for the optimal weighting matrix (96). The covariances involved in calculating (96) need to be replaced by the sample analogs. The matrix $\mathbf{D}_\gamma\mathbf{R}_{n_f n_f}(\tau)\mathbf{D}_\gamma^T$ can be estimated by noting that (23) gives

$$\mathbf{D}_\gamma\mathbf{N}_f \approx \mathbf{D}_\gamma\mathbf{Y}_f\boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T}\ . \tag{100}$$

Solve (99) using the estimated weighting, which results in ABC estimates of $\boldsymbol{\theta}$. If desired, find the roots of the characteristic polynomial, which are the system poles. For a detailed description of an implementation of this scheme, see [2]. It should, however, be noted that estimating polynomial coefficients is numerically difficult for high order systems.

The ABC theory also provides us with the asymptotic properties of $\hat{\boldsymbol{\theta}}$. Recall that

$$\boldsymbol{\theta} = (d_{1,1}\ldots d_{1,n}\; d_{2,1}\ldots d_{2,n}\;\ldots d_{l,n})\ . \tag{101}$$

Let,

$$\mathbf{M}_0 = \tilde{\boldsymbol{\Gamma}}_\gamma(n+1:\gamma l,:) \tag{102}$$

in the MATLAB notation[7] for describing submatrices. For $j = 1,\ldots,n$, set

$$\mathbf{M}_{1,j} = \begin{bmatrix} \tilde{\boldsymbol{\Gamma}}_\gamma(n+1-j:\gamma-j,:) \\ 0_{n\times n} \\ \tilde{\boldsymbol{\Gamma}}_\gamma(\gamma+n+1-j:2\gamma-j,:) \\ 0_{n\times n} \\ \vdots \\ \tilde{\boldsymbol{\Gamma}}_\gamma((l-1)\gamma+n+1-j:l\gamma-j,:) \end{bmatrix}, \tag{103}$$

[7] $a:b$ selects rows (columns) a through b, whereas a : means all rows (columns).

and for $i = 2, \ldots, l$

$$\mathbf{M}_{i,j} = \begin{bmatrix} 0_{(i-1)\gamma-n\times n} \\ \tilde{\boldsymbol{\Gamma}}_\gamma(n+1-j : 2n-j, :) \\ 0_{(l-i+1)\gamma-n\times n} \end{bmatrix} . \tag{104}$$

With the definitions

$$\mathbf{m} = \mathrm{vec}(\mathbf{M}_0) \tag{105}$$

$$\mathbf{M} = [\mathrm{vec}(\mathbf{M}_{1,1}) \ldots \mathrm{vec}(\mathbf{M}_{1,n}) \ldots \mathrm{vec}(\mathbf{M}_{l,n})] \tag{106}$$

we get the linear regression model

$$\mathrm{vec}(\mathbf{D}_\gamma(\boldsymbol{\theta})\boldsymbol{\Gamma}_\gamma) = \mathbf{M}\boldsymbol{\theta} + \mathbf{m} . \tag{107}$$

The corresponding relation for the estimated quantities is

$$\mathrm{vec}(\mathbf{D}_\gamma(\boldsymbol{\theta})\hat{\mathbf{Q}}_{\mathrm{s}}^{\mathrm{W}}) = \hat{\mathbf{M}}\boldsymbol{\theta} + \hat{\mathbf{m}} , \tag{108}$$

where $\hat{\mathbf{M}}$ and $\hat{\mathbf{m}}$ are constructed from the permuted version of $\hat{\mathbf{Q}}_s^W$, conformly with $\mathbf{M}$ and $\mathbf{m}$. The estimate (99) can now be expressed as

$$\hat{\boldsymbol{\theta}} = -[\hat{\mathbf{M}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{M}}]^{-1}\hat{\mathbf{M}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{m}} . \tag{109}$$

The asymptotic properties are given by (see e.g. [23])

$$\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \in \mathrm{As}\mathcal{N}(0, [\mathbf{M}^T\mathbf{R}^{-1}\mathbf{M}]^{-1}) . \tag{110}$$

Finally, it remains to determine the asymptotic distribution of the corresponding poles. Recall that the coefficients of the characteristic polynomial equal the first n elements of $\boldsymbol{\theta}$. Define

$$\boldsymbol{\Upsilon} = \begin{pmatrix} a_n & \ldots & a_1 \\ & \ddots & \vdots \\ 0 & & a_n \end{pmatrix} , \tag{111}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \mu_1^{-n} & \ldots & \mu_n^{-n} \\ \vdots & & \vdots \\ \mu_1^{-1} & \ldots & \mu_n^{-1} \end{pmatrix} , \tag{112}$$

$$\mathbf{R}_a = [\mathbf{M}^T\mathbf{R}^{-1}\mathbf{M}]^{-1}(1:n, 1:n) . \tag{113}$$

From [18] it then follows that the estimate of $\boldsymbol{\mu} = (\mu_1 \ldots \mu_n)^T$, for distinct poles, has the asymptotic distribution

$$\sqrt{N}\,(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \in \mathrm{As}\mathcal{N}(0, [\boldsymbol{\Upsilon}\boldsymbol{\Lambda}]^{-1}\mathbf{R}_a[\boldsymbol{\Upsilon}\boldsymbol{\Lambda}]^{-T}) . \tag{114}$$

5 Example

In this section an example is presented to illustrate the performance of the different techniques for extracting the poles from the subspace estimate. The agreement between the sample statistics and analytic expressions derived in the paper is also examined.

Consider the following second-order system with two outputs and a single input.

$$A = \begin{bmatrix} 1 & -0.5 \\ 1 & 0 \end{bmatrix} \tag{115}$$

$$C = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \tag{116}$$

$$B^T = [1 \;\; 0] \quad D^T = [0 \;\; 0] \; . \tag{117}$$

The measurement noise, e_t, process noise, w_t, and input sequence, u_t are all uncorrelated white Gaussian random processes. The covariance matrices are diagonal, with the variance of e_t equal to 0.1 and the variances of w_t and u_t set to unity. The poles of this system are located at $\mu_{1,2} = 0.5 \pm 0.5i$.

The parameters γ and β are equal to 5 resulting in an extended observability matrix which is 10×2. The space spanned by observability matrix is estimated from the input-output data according to Section 3 with the N4SID weighting, and for different number of vector samples N. Observe that the choice of column weighting does not affect the analytic results, since the input is white.

The asymptotic standard deviation of the estimation error of the pole, $\sqrt{\mathrm{E}\{\tilde{\mu}_1 \tilde{\mu}_1^*\}}$ is displayed in Figure 1. The solid line depicts the result for ABC estimator (99) and the dashed line shows the result for IV-ESP estimate (48). The corresponding sample statistics, based on 512 trails, is also in figure and show a good agreement with the analytic standard deviations. The conclusion is that using more structure and correct weighting improves the performance, as is seen by comparing the two results.

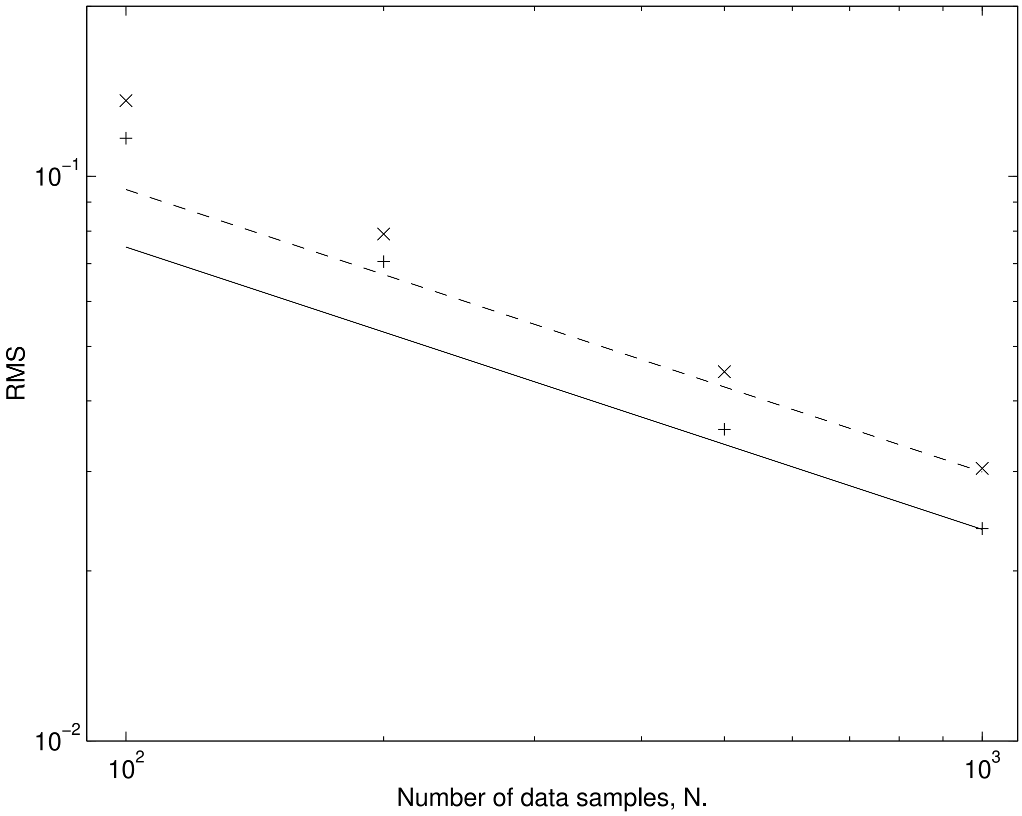

Figure 1: Standard deviation of pole estimate $\hat{\mu}_1$ versus number of samples. Lines shows the asymptotic results. (x) and dashed line are from the IV-ESP estimator, and (+) and solid line are for the ABC estimator.

6 Conclusions

The objective of this paper has been to analyze the performance of subspace-based state space system identification (4SID) methods. The connection to instrumental variable methods and sensor array signal processing has been used to derive the statistical properties of the signal subspace estimate employed by most current 4SID methods. Using this result, the statistical properties of an estimated dynamical system can be assessed. We derived the covariance elements of the system poles for the simple shift-invariance estimator. The connection to sensor array signal processing also motivated us to propose a novel estimation method based on a direct parameterization of the null-space of the extended observability matrix. The method uses the estimated signal subspace in a statistically optimal fashion, and hence gives a lower bound on achievable performance. However, the parameterization used corresponds to canonical forms and is thus sensitive to numerical errors. Further work is required to investigate more robust linear parameterizations of the null-space of observability matrices. Other topics for future studies include guidelines for optimally selecting user quantities such as γ, β and the signal-subspace weighting matrices. It is hoped that the statistical results presented herein, although somewhat involved, provide means of further improving and understanding the properties of the 4SID approach.

Acknowledgment

The authors would like to thank Lennart Ljung for his contributions to early versions of this work, Magnus Jansson for many constructive comments, and Klaus Bernpainter for the implementation of the numerical example.

A Appendix

A.1 Proof of Theorem 1

In this appendix, Theorem 1 giving the asymptotic distribution of the signal subspace matrix is derived. Let the deterministic vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ satisfy

$$\boldsymbol{\xi}^* \boldsymbol{\Gamma}_\gamma = \boldsymbol{\eta}^* \boldsymbol{\Gamma}_\gamma = 0 \ . \tag{118}$$

Post-multiply (32) by $\hat{\mathbf{V}}_s$, to obtain

$$\frac{1}{N}\hat{\mathbf{W}}_r \mathbf{Y}_f \boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T} \mathbf{P}^T \hat{\mathbf{W}}_c \hat{\mathbf{V}}_s = \hat{\mathbf{Q}}_s \hat{\mathbf{S}}_s \ . \tag{119}$$

Pre-multiplying (119) by $\boldsymbol{\xi}^* \hat{\mathbf{W}}_r^{-1}$, post-multiplying by $\hat{\mathbf{S}}_s^{-1}$ and using (118) and (23), we obtain

$$\frac{1}{N}\boldsymbol{\xi}^* \mathbf{N}_f \boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T} \mathbf{P}^T \hat{\mathbf{W}}_c \hat{\mathbf{V}}_s \hat{\mathbf{S}}_s^{-1} = \boldsymbol{\xi}^* \hat{\mathbf{W}}_r^{-1} \hat{\mathbf{Q}}_s \ = \boldsymbol{\xi}^* \hat{\mathbf{Q}}_s^W \ . \tag{120}$$

Inserting the definition $\boldsymbol{\Pi}^{\perp}_{\mathbf{U}_f^T} = \mathbf{I} - \mathbf{U}_f^T (\mathbf{U}_f \mathbf{U}_f^T)^{-1} \mathbf{U}_f$ into (120) leads to the expression

$$\begin{aligned}
\sqrt{N}\,\boldsymbol{\xi}^* \hat{\mathbf{Q}}_s^W \;&=\; \boldsymbol{\xi}^* \left(\frac{1}{\sqrt{N}} \mathbf{N}_f \mathbf{P}^T \right) \hat{\mathbf{W}}_c \hat{\mathbf{V}}_s \hat{\mathbf{S}}_s^{-1} \\
&\quad - \boldsymbol{\xi}^* \left(\frac{1}{\sqrt{N}} \mathbf{N}_f \mathbf{U}_f^T \right) (\mathbf{U}_f \mathbf{U}_f^T)^{-1} \mathbf{U}_f \mathbf{P}^T \hat{\mathbf{W}}_c \hat{\mathbf{V}}_s \hat{\mathbf{S}}_s^{-1} \ .
\end{aligned} \tag{121}$$

From Assumptions 2-3, the Central Limit Theorem shows that the elements of the matrices

$$\frac{1}{\sqrt{N}} \mathbf{N}_f \mathbf{P}^T \quad \text{and} \quad \frac{1}{\sqrt{N}} \mathbf{N}_f \mathbf{U}_f^T \tag{122}$$

are $O_p(1)$ and that they all have a limiting zero-mean Gaussian distribution. Hence, the asymptotic distribution of $\sqrt{N}\,\boldsymbol{\xi}^* \hat{\mathbf{Q}}_s^W$ is the same as that of the quantity

$$\begin{aligned}
\sqrt{N}\,\boldsymbol{\xi}^* \hat{\mathbf{Q}}_s^W \;&\simeq\; \boldsymbol{\xi}^* \left(\frac{1}{\sqrt{N}} \mathbf{N}_f \mathbf{P}^T \right) \mathbf{W}_c \mathbf{V}_s \mathbf{S}_s^{-1} - \boldsymbol{\xi}^* \left(\frac{1}{\sqrt{N}} \mathbf{N}_f \mathbf{U}_f^T \right) \mathbf{R}_{u_f u_f}^{-1} \mathbf{R}_{u_f p} \mathbf{W}_c \mathbf{V}_s \mathbf{S}_s^{-1} \\
&=\; \boldsymbol{\xi}^* \left(\frac{1}{\sqrt{N}} \mathbf{N}_f \mathbf{Z}^T \right) \mathbf{H} \ ,
\end{aligned} \tag{123}$$

where, $\mathbf{Z}$ and $\mathbf{H}$ are defined in (43) and (44) respectively. Note that $\mathbf{H}$ is a deterministic quantity and thus $\sqrt{N}\,\boldsymbol{\xi}^* \hat{\mathbf{W}}_r^{-1} \hat{\mathbf{Q}}_s$ has a limiting zero-mean Gaussian distribution. The covariance of the asymptotic distribution is calculated as

$$\mathbf{C}(\boldsymbol{\xi}, \boldsymbol{\eta}) = \lim_{N\to\infty} \mathrm{E} \left[\mathbf{H}^T \left(\frac{1}{\sqrt{N}} \mathbf{Z} \mathbf{N}_f^T \right) \boldsymbol{\xi}\, \boldsymbol{\eta}^* \left(\frac{1}{\sqrt{N}} \mathbf{N}_f \mathbf{Z}^T \right) \mathbf{H} \right] \ . \tag{124}$$

Since $\boldsymbol{\zeta}(t)$ and $\mathbf{n}_f(t)$ are both Gaussian random variables (the deterministic part of $\boldsymbol{\zeta}(t)$ being thought of as a Gaussian distribution with zero variance), the expectation can be calculated as

$$\begin{aligned}
\mathrm{E}[\mathbf{Z}\mathbf{N}_f^T\boldsymbol{\xi}\boldsymbol{\eta}^*\mathbf{N}_f\mathbf{Z}^T] &= \sum_{t,s,k,l} \mathrm{E}[\boldsymbol{\zeta}(t)\boldsymbol{\zeta}^T(s)\mathbf{n}_{f,k}(t)\boldsymbol{\xi}_k\boldsymbol{\eta}_l^*\mathbf{n}_{f,l}(s)] \\
&= \sum_{t,s,k,l} \Big(\mathrm{E}[\boldsymbol{\zeta}(t)\boldsymbol{\zeta}^T(s)]\,\mathrm{E}[\mathbf{n}_{f,k}(t)\mathbf{n}_{f,l}(s)] \\
&\quad +\mathrm{E}[\boldsymbol{\zeta}(t)\mathbf{n}_{f,k}(t)]\,\mathrm{E}[\boldsymbol{\zeta}^T(s)\mathbf{n}_{f,l}(s)] \\
&\quad +\mathrm{E}[\boldsymbol{\zeta}(t)\mathbf{n}_{f,l}(s)]\,\mathrm{E}[\boldsymbol{\zeta}^T(s)\mathbf{n}_{f,k}(t)]\Big)\,\boldsymbol{\xi}_k\boldsymbol{\eta}_l^* \; .
\end{aligned} \tag{125}$$

In the sum, t and s range from 1 to N, whereas k and l go from 1 to γl. Since $\mathrm{E}[\boldsymbol{\zeta}(t)\mathbf{n}_f^T(s)] = 0$ for $t \leq s$, only the first term of (125) is non-zero. Thus, (125) reduces to

$$\mathrm{E}[\mathbf{Z}\mathbf{N}_f^T\boldsymbol{\xi}\boldsymbol{\eta}^*\mathbf{N}_f\mathbf{Z}^T] = \sum_{t,s} \mathrm{E}[\boldsymbol{\zeta}(t)\boldsymbol{\zeta}^T(s)]\;\boldsymbol{\eta}^*\mathbf{R}_{n_f}(s-t)\boldsymbol{\xi} \; . \tag{126}$$

Changing the summation variables to t and $\tau = s-t$ and using (43) now gives

$$\lim_{N\to\infty}\frac{1}{N}\,\mathrm{E}[\mathbf{Z}\mathbf{N}_f^T\boldsymbol{\xi}\boldsymbol{\eta}^*\mathbf{N}_f\mathbf{Z}^T] = \lim_{N\to\infty}\sum_{\tau=1-N}^{N-1}\mathbf{R}_{\zeta\zeta}(-\tau)\boldsymbol{\eta}^*\mathbf{R}_{n_f n_f}(\tau)\boldsymbol{\xi} \; . \tag{127}$$

Using the structure

$$\mathbf{n}_f(t) = \boldsymbol{\Psi}_\gamma\mathbf{w}_f(t) + \mathbf{e}_f(t)$$

where $\boldsymbol{\Psi}_\gamma$ is defined in (7), it is easy to see that $\boldsymbol{\eta}^*\mathbf{R}_{n_f n_f}(\tau)\boldsymbol{\xi} = 0$ for $|\tau| \geq \gamma$. Inserting (127) into (124) then shows that

$$\mathbf{C}(\boldsymbol{\xi},\boldsymbol{\eta}) = \sum_{|\tau|<\gamma}\mathbf{H}^T\mathbf{R}_{\zeta\zeta}(\tau)\mathbf{H}\;\boldsymbol{\eta}^*\mathbf{R}_{n_f n_f}(-\tau)\boldsymbol{\xi} \; , \tag{128}$$

which agrees with (42).

A.2 Proof of Theorem 2

The asymptotic covariance (94) is

$$\mathbf{R} = \lim_{N\to\infty} N\,\mathrm{E}\left[\mathrm{vec}(\mathbf{D}_\gamma\hat{\mathbf{Q}}_\mathrm{s}^\mathrm{W})\mathrm{vec}(\mathbf{D}_\gamma\hat{\mathbf{Q}}_\mathrm{s}^\mathrm{W})^\mathrm{T}\right] \; . \tag{129}$$

By (123), we thus must evaluate

$$\mathrm{E}\left[\mathrm{vec}\left\{\mathbf{D}_\gamma\left(\frac{1}{\sqrt{\mathrm{N}}}\mathbf{N}_\mathrm{f}\mathbf{Z}^\mathrm{T}\right)\mathbf{H}\right\}\mathrm{vec}\left\{\mathbf{D}_\gamma\left(\frac{1}{\sqrt{\mathrm{N}}}\mathbf{N}_\mathrm{f}\mathbf{Z}^\mathrm{T}\right)\mathbf{H}\right\}^\mathrm{T}\right] \; . \tag{130}$$

Now, the kl:th block of the above matrix can be expressed as

$$\begin{aligned}
&\mathrm{E}\left[\mathrm{vec}\left\{\mathbf{D}_{\mathrm{k.}}\left(\tfrac{1}{\sqrt{\mathrm{N}}}\mathbf{N}_{\mathrm{f}}\mathbf{Z}^{\mathrm{T}}\right)\mathbf{H}\right\}\mathrm{vec}\left\{\mathbf{D}_{\mathrm{l.}}\left(\tfrac{1}{\sqrt{\mathrm{N}}}\mathbf{N}_{\mathrm{f}}\mathbf{Z}^{\mathrm{T}}\right)\mathbf{H}\right\}^{\mathrm{T}}\right] \\
&= \mathrm{E}\left[\mathbf{H}^{T}\left(\tfrac{1}{\sqrt{N}}\mathbf{Z}^{T}\mathbf{N}_{f}\right)\mathbf{D}_{k.}^{T}\mathbf{D}_{l.}\left(\tfrac{1}{\sqrt{N}}\mathbf{N}_{f}\mathbf{Z}^{T}\right)\mathbf{H}\right] \\
&\to \mathbf{C}(\mathbf{D}_{k.}^{T},\mathbf{D}_{l.}^{T})
\end{aligned} \tag{131}$$

as $N \to \infty$. From the expression (128) it follows that

$$\mathbf{R} = \sum_{|\tau|<\gamma} (\mathbf{H}^{T}\mathbf{R}_{\zeta}(\tau)\mathbf{H}) \otimes (\mathbf{D}_{\gamma}\mathbf{R}_{n_f}(\tau)\mathbf{D}_{\gamma}^{T}) \;, \tag{132}$$

which is (96).

References

[1] M. Aoki. *State Space Modeling of Time Series.* Springer Verlag, Berlin, second edition, 1990.

[2] Klaus Bernpainter. "A Study of Threee Subspace State-Space System Identification Methods for Multiple-Output Systems". Master's thesis, Royal Institute of Technology, Department of Signals, Sensors & Systems, Dec. 1995.

[3] H. Clergeot, S. Tressens, and A. Ouamri. "Performance of High Resolution Frequencies Estimation Methods Compared to the Cramér-Rao Bounds". *IEEE Trans. ASSP*, 37(11):1703–1720, November 1989.

[4] B. DeMoor. *Mathematical Concepts and Techniques for Modeling of Static and Dynamic Systems.* PhD thesis, Kath. Universiteit Leuven, Heverlee, Belgium, 1988.

[5] B. DeMoor, J. Vandewalle, L. Vandenberghe, and P. Van Mieghem. "A Geometrical Strategy for the Identification of State Space Models of Linear Multivariable Systems with Singular Value Decomposition". In *Proc. IFAC 88*, pages 700–704, Beijing, China, 1988.

[6] G.H. Golub and C.F. VanLoan. "An Analysis of the Total Least Squares Problem". *SIAM J. Num. Anal.*, 17:883–893, 1980.

[7] B.L. Ho and R.E. Kalman. "Efficient Construction of Linear State Variable Models from Input/output Functions". *Regelungstechnik*, 14:545–548, 1966.

[8] M. Jansson and B. Wahlberg. "4SID Linear Regression". In *Proc.* 33*rd CDC*, pages 2528–2863, Orlando, FL, 1994.

[9] M. Jansson and B. Wahlberg. On weighting in state-space subspace system identification. In *Proc. European Control Conference, ECC'95*, pages 435–440, Roma, Italy, 1995.

[10] M. Jansson and B. Wahlberg. "A Linear Regression Approach to State-Space Subspace System Identification". *Signal Processing*, 1996. To appear.

[11] T. Kailath. *Linear Systems.* Prentice-Hall, Englewood Cliffs, NJ, 1980.

[12] S.Y. Kung. "A New Identification and Model Reduction Algorithm via Singular Value Decomposition". In *Proc. 12th Asilomar Conf. on Circuits, Systems and Computers*, pages 705–714, Pacific Grove, CA, November 1978.

[13] W. E. Larimore. "System Identification, Reduced-Order Filtering and Modelling via Canonical Variate Analysis". In *Proc. ACC*, pages 445–451, June 1983.

[14] W. E. Larimore. "Canonical Variate Analysis in Identification, Filtering and Adaptive Control". In *Proc. 29th CDC*, pages 596–604, Honolulu, Hawaii, December 1990.

[15] K. Liu. "Identification of Multi-Input and Multi-Output Systems by Observability Range Space Extraction". In *Proc. CDC*, pages 915–920, Tucson, AZ, 1992.

[16] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ, 1987.

[17] T. McKelvey. "On State-Space Models in System Identification". Licentiate thesis, Linköping University, Sweden, May 1994.

[18] A. Nehorai and D. Starer. "Adaptive Pole Estimation". *IEEE Trans. ASSP*, ASSP-38:825–838, May 1990.

[19] B. Ottersten and M. Viberg. "A Subspace-Based Instrumental Variable Method for State-Space System Identification". In *Proc. 10th IFAC Symp. on Syst. Id.*, Copenhagen, Denmark, 1994.

[20] B. D. Rao and K. V. S. Hari. Performance analysis of ESPRIT and TAM in determining the direction of arrival of plane waves in noise. *IEEE Trans. ASSP*, ASSP-37(12):1990–1995, Dec. 1989.

[21] R. Roy and T. Kailath. "ESPRIT – Estimation of Signal Parameters via Rotational Invariance Techniques". *IEEE Trans. on ASSP*, ASSP-37(7):984–995, July 1989.

[22] R. O. Schmidt. *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*. PhD thesis, Stanford Univ., Stanford, CA, Nov. 1981.

[23] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, London, U.K., 1989.

[24] P. Stoica and T. Söderström. "Statistical Analysis of MUSIC and Subspace Rotation Estimates of Sinusoidal Frequencies". *IEEE Trans. ASSP*, ASSP-39:1836–1847, Aug. 1991.

[25] Petre Stoica and Mats Viberg. "Weighted LS and TLS Approaches Yield Asymptotically Equivalent Results". *Signal Processing*, 45(2), 1995.

[26] A. Swindlehurst, R. Roy, B. Ottersten, and T. Kailath. "A Subspace Fitting Method for Identification of Linear State Space Models". *IEEE Trans. AC*, March 1995.

[27] A.-J. VanDerVeen, E.F. Deprettere, and A.L. Swindlehurst. "Subspace-Based Signal Analysis Using Singular Value Decomposition". *Proceedings of the IEEE*, 81(9):1277–1308, Sept. 1993.

[28] P. VanOverschee. *Subspace Identification: Theory – Implementation – Application.* PhD thesis, Katholieke Universiteit Leuven, Feb. 1995.

[29] P. VanOverschee and B. DeMoor. "A Unifying Theorem for three Subspace System Identification Algorithms". In *Proc. 10th IFAC Symp. on Syst. Id.*, Copenhagen, Denmark, 1994.

[30] P. VanOverschee and B. DeMoor. **N4SID**: Subspace Algorithms for the Identification of Combined Deterministic-Stochastic Systems. *Automatica, Special Issue on Statistical Signal Processing and Control*, 30(1):75–93, Jan. 1994.

[31] P. VanOverschee and B. DeMoor. "A Unifying Theorem for three Subspace System Identification Algorithms". *Automatica, Special Issue on Treends in System Identification*, 31(12):1853–1864, Dec. 1995.

[32] P. VanOverschee and B. DeMoor. Choice of State-space Basis in Combined Deterministic-Stochastic Subspace Identification. *Automatica, Special Issue on Treends in System Identification*, 31(12):1877–1884, Dec. 1995.

[33] P. VanOverschee, B. DeMoor, and J. Suykens. "Subspace Algorithms for System Identification and Stochastic Realization". In *Proc. MTNS*, pages 589–595, Kobe, Japan, 1991. Mita Press.

[34] M. Verhaegen. "A Novel Non-iterative MIMO State Space Model Identification Technique". In *Proc.* 9^{th} *IFAC/IFORS Symposium on Identification and System Parameter Estimation*, pages 1453–1458, Budapest, Hungary, July 1991.

[35] M. Verhaegen. "Subspace Model Identification. Part III: analysis of the ordinary output-error state space model identification algorithm". *Int. J. Control*, 58:555–586, 1993.

[36] M. Verhaegen. "Identification of the Deterministic Part of MIMO State Space Models given in Innovations Form from Input-Output Data". *Automatica, Special Issue on Statistical Signal Processing and Control*, 30(1):61–74, Jan. 1994.

[37] M. Verhaegen and P. Dewilde. "Subspace Model Identification. Part I: the output-error state space model identification class of algorithms". *Int. J. Control*, 56:1187–1210, 1992.

[38] M. Viberg. "On Subspace-Based Methods for the Identification of Linear Time-Invariant Systems". *Automatica, Special Issue on Treends in System Identification*, 31(12):1835–1852, Dec. 1995.

[39] M. Viberg and B. Ottersten. "Sensor Array Processing Based on Subspace Fitting". *IEEE Trans. SP*, SP-39(5):1110–1121, May 1991.

[40] M. Viberg, B. Ottersten, B. Wahlberg, and L. Ljung. "A Statistical Perspective on State-Space Modeling Using Subspace Methods". In *Proc.* 30^{th} *IEEE Conf. on Decision & Control*, pages 1337–1342, Brighton, England, Dec. 1991.

[41] M. Viberg, B. Ottersten, B. Wahlberg, and L. Ljung. "Performance of Subspace-Based System Identification Methods". In *Proc. IFAC 93*, volume 7, pages 369–372, Sydney, Australia, July 1993.