Aspects and Experiences of User Choices in Subspace Identification

Lennart Ljung,

Division of Automatic Control
Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden
WWW: http://www.control.isy.liu.se
E-mail: ljung@isy.liu.se

3rd December 2003

Report no.: LiTH-ISY-R-2564
Submitted to The 13th IFAC Symposium on System Identification, Rotterdam, The Netherlands

Technical reports from the Control & Communication group in Linköping are available at http://www.control.isy.liu.se/publications.
ASPECTS AND EXPERIENCES OF USER CHOICES IN SUBSPACE IDENTIFICATION METHODS

Lennart Ljung

Abstract: Subspace identification methods, such as N4SID, MOESP, CVA etc have proven to be very successful for identification of multivariable, linear dynamic systems. These methods are associated with a number of design variables, or user choices. These include prediction horizons, weighting matrices and ways to perform the estimation. It is known that these choices may have a substantial influence on the model quality, at the same time as there is no comprehensive theory for rational decision making. In this contribution we study certain aspects of the choices, in particular their influence on both bias and variance. We also illustrate some aspects using a larger example, the Tennessee-Eastman identification challenge.

1. INTRODUCTION

Subspace methods for identification of linear multivariable dynamical systems have been very successful. They can be traced back to the realization algorithms of Kalman and Ho, (Ho and Kalman 1965). One of the first descriptions in an identification context was given by (Akaike 1976). The methods were further developed by Larimore in a series of papers, e.g. (Larimore 1983), (Larimore 1990), by van Overschee and de Moor, see e.g. the book (Van Overschee and DeMoor 1996), by Verhaegen, e.g. (Verhaegen 1994) and others. The literature is quite extensive.

The methods estimate a state-space model of the type (in standard notation)

\[ x(t+1) = Ax(t) + Bu(t) + Ke(t) \]  
\[ y(t) = Cx(t) + Du(t) + e(t) \]  

(1a)  
(1b)

Some variants estimate covariance matrices of process and measurement noise rather than the Kalman gain \( K \).

The subspace algorithms are known under several names and acronyms, like CVA (Canonical Variate Analysis, Larimore), MOESP (Verhaegen), N4SID (van Overschee/de Moor) and have been incorporated in several commercial software packages, like System Identification Toolbox for use with MATLAB, (Ljung 2000) and ADAPTx, (Larimore 1997).

Despite the undeniable good performance of the algorithms, there is still no comprehensive theory for the (asymptotic) properties of the resulting estimates. This is a drawback, since the methods contain several important design issues and choices, so a theory for how these should be chosen is lacking. Important partial results are given in (among many papers), (Bauer et al. 1999), (Jansson and Wahlberg 2000), (Knudsen 2001), (Bauer and Ljung 2002), (Bauer 2000), (Chiuso and Picci 2003), and (Dahlen and Scherrer 2003). These results all deal with variance issues. The question of bias distribution when more complex systems are approximated by lower order ones has apparently not been dealt with in a comprehensive manner.

In this paper we shall review some of the basic choices in the algorithms and, by simulation, illustrate some aspects and experiences of how to select them. We shall then discuss both variance and bias aspects.

2. BASIC NOTION OF SUBSPACE METHODS

A conceptually fairly correct (but very simplified) first view of subspace methods is to say that they build a higher order ARX-model, which is then subjected to model reduction of weighted Hankel-norm type. The design variables are then associated with the orders (both number of past
The choices that have to be made can be listed as follows:

1. The number, \( r \) of forward predictions used in the model reduction step.
2. The number of past inputs, \( s_u \) and the number of past outputs, \( s_y \) used in the underlying ARX model. If \( s_y = 0 \) we may talk about an output error model.
3. The matrices \( P_1 \) and \( P_2 \) that are used for pre- and post-multiplication of the predictor matrix in the model reduction step.
4. The procedure for selecting model order. This amounts to analyzing the singular values of the above mentioned matrix.
5. Once \( A \) and \( C \) are determined from the factorized matrix, the \( B \) and \( D \) matrices can be estimated in different ways.

2.2 The subspace identification code in the System Identification Toolbox

In this paper we shall perform the comparisons using the System Identification Toolbox, version 5, (Ljung 2000). The command is `N4SID`. This contains an implementation of a subspace algorithm with the following features:

- The horizons (property `N4Horizon`), \([r \ s_y \ s_u]\) can be chosen independently. Setting this property to `'auto'` (default) means that the Akaike criterion (AIC) is used to determine a suitable horizon, by estimating many ARX models of different orders. Then \( s_u = s_y \), unless `DisturbanceModel` has been set to `'none'` (i.e. \( K \) set to zero in (1a)), in which case \( s_y = 0 \).
- Two different weightings (property `N4Weight`) \((P_1, P_2)\), namely those that are known as CVA and MOESP are supported. (see page 351 in (Ljung 1999)). Earlier versions of the toolbox contained a third choice of weighting matrices, known as `N4SID`. These weightings, however seem to be inferior to CVA and MOESP. The routine has kept is name, `N4SID`, though.
- The \( A \) and \( C \) matrices are estimated from the shift properties of the estimated observability matrix (one of the factors obtained by the SVD step). The matrix \( K \) (in case `DisturbanceModel = Estimate`) is also directly estimated from the SVD-factors. For fixed \( A \) and \( C \) the estimation of \( B \) and \( D \) in (1a) is a linear regression problem. This is used in `N4SID`. The only question is whether \( K \) should be used or not:

\[
y = (C(qI - A)^{-1}B + D)u \quad (2a)
\]

or

\[
y = (C(qI - A + KC)^{-1} + D)u \quad (2b)
\]

Both these expressions are linear regressions in \( B \) and \( D \) (and the initial state \( x(0) \) if estimated) for given \( A, C, K \). The second case corresponds to a pre-whitening of the noise in the linear regression and is chosen if `focus = 'prediction'`. The former choice is made in case `focus = 'simulation'`.

2.3 Format of comparisons

In the following a number of comparisons of model qualities are made. The format for all these comparisons are as follows:

250 different, stable SISO systems or order 8 have been randomly generated. To the system was added a noise description, corresponding to a Kalman gain \( K \). The systems were simulated with a random binary input signal \((u = \text{idinput}(N, 'rbs', [0 0.2]))\). \( N = 1200 \) points were used with added noise according to the noise model. The size of the noise source was always adjusted to that a signal-to-noise ratio of 10 (amplitude-wise) was obtained at the output.

For each of these 250 data sets a model was generated using various design options. The quality of each model was evaluated as the norm of the difference between the true frequency function and the model one. (\( f_0 = \text{freqfunc}(m0) \), \( f_1 = \text{freqfunc}(m) \)). `fit = norm(squeeze(f0-f1))`. To illustrate the influence of the design variables, they were selected in pairs giving two models at a time to compare, one for each choice of design variable. The fit for the models for each data set were plotted against each other, thus creating a scatter plot with 250 points. The axes scale was chosen so that about 90% of these points were shown (to avoid strange scalings due to outliers.) A point below the bisecting line thus indicates that for that data set, the “y-axis-method” gave a better model than the “x-axis-method”.

2.4 The Prediction error method

We shall occasionally compare models estimated using various subspace methods with those using prediction error methods. Such methods use a parameterization of the state-space matrices (canonical or full parameterizations) and minimize the prediction errors (the determinant of the sample covariance matrix of the prediction errors) with respect to these parameters. See (Ljung 1999) for a comprehensive treatment.
3. BIAS CONSIDERATIONS

When estimating models using data from complex systems, the resulting model suffers from errors of two different types:

- The *bias error*, which stems from the fact that the true system cannot be represented within the chosen model class. If the true system happens to be linear, one can think of the bias error as a difference between the true frequency function and that of the “best model” available in the model class.
- The *variance error*, which originates from the disturbances that affect the data. For noise free data, the variance error would be zero.

The total (mean square) model error is thus the sum of the variance error and (the square of) the bias error.

For many methods it is more difficult to analyze the bias error than the variance error. For prediction error methods, cf. Section 2.4 the distribution of the bias error over frequencies can be characterized, see e.g. (Ljung 1999), but for subspace methods this topic has apparently not been discussed. One comment can be made though: If the $B$ and $D$ matrices are identified separately (for fixed $A, C, K$) as a linear regression, this indeed is a prediction error method, and the bias distribution result can be applied. In particular it means that the choice (2a) corresponds to a weighting for the approximation that is given by the input spectrum, while (2b) puts more weight to frequencies where the predictor is important.

The simulations described in this section are performed as in Section 2.3. The true systems are of 8th order and the models are of second order.

3.1 Influence of weighting matrices

The first test is to see if there is a systematic difference between the CVA and MOESP choices of weighting matrices. The result is shown in Figure 1. This shows a clear advantage for CVA (by 68 %)

3.2 Influence of Prediction Horizons

There is no particular reason to believe that the prediction horizons ($s_y$ and $s_u$) should affect the approximation qualities. In Figure 2 the use of long horizons ([30 30 30]) is compared to short horizons ([4 4 4]). The long horizon performs better. A possible explanation could be that the underlying ARX model with these orders have a better chance of picking up the dynamics of the complex system.

3.3 Influence of ‘focus’

Setting ‘focus’ = ‘estimation’, that is estimating $B$ from (2a) rather than from (2b), should favor the fit of the low order model, according to our earlier discussion. This is confirmed in Figure 3. Note that the fit also depends on the input spectrum. If we had evaluated a fit between model and true system, that was weighted according to the input spectrum, the effect would have been even more pronounced.

Fig. 1. CVA weighting (x-axis) versus MOESP weighting (y-axis). CVA is the better choice in 67 % of the cases.

Fig. 2. A comparison between “long” and “short” prediction horizons. The long horizon is the better choice in 80% of the cases.

Fig. 3. A comparison between ‘focus’=‘prediction’ and ‘focus’=‘simulation’. The simulation choice performs better in 79% of the cases.
### 3.4 Influence of Subspace/PEM

A further question is whether using further PEM iterations for the model obtained with the "simulate" focus will improve the approximation. It is expected that this should be the case, since the also the $A$ and $C$ matrices are adjusted to improve the fit as an optimization problem. As seen in Figure 4, the effect quite clear. Even more pronounced is a comparison between CVA/predict and PEM, where PEM performs better in 90% of the cases.

![Graph showing comparison between CVA and PEM](image)

Fig. 4. A comparison between the CVA-subspace method (with "focus" = "simulation") and the model obtained by prediction error identification. PEM is the better choice in 82% of the cases.

### 4. VARIANCE CONSIDERATIONS

There have been many investigations of the statistical efficiency of subspace methods. The general experience is that the work well most of the time and may give models with variances that are not far from the Cramer-Rao bound. No general theory for how the variance of the estimate is affected by the various design variables has yet, been given though. A number of partial results are available, though. They indicate that the CVA-weightings have certain theoretical advantages, and will be optimal in certain cases, like the time-series case (no input) (Bauer 2000), and the case with a white noise input, (Bauer and Ljung 2002). Other relevant studies include (Gustafsson 2002) and (Deistler et al. 1995).

In this section we instigate the variance properties only by simulation. For that purpose, we ran Monte-Carlo simulations over 250 random models as described in Section 2.3. The simulated systems were of 8th order, and so were the models. In all cases, the direct term was set to zero ($D = 0$), i.e. a delay of 1 from input to output.

### 4.2 Influence of Prediction Horizons

The issue of the choice of the prediction horizons ($s_y$ and $s_u$) is illustrated in Figure 6. The fixed choice $[10 \ 10 \ 10]$ ("short" horizons) is compared to $[30 \ 30 \ 30]$ ("long" horizons). The short horizons behave better than the long ones. It also turns out that both of them are (slightly) inferior to the AIC-based, automatic choice described in Section 2.2.

![Graph showing comparison between short and long prediction horizons](image)

Fig. 6. A comparison between short (x-axis) and long (y-axis) prediction horizons. The short horizons are better in 60% of the cases.

### 4.3 Influence of 'focus'

Setting $\text{focus} = \text{simulation}$ favors the fit of the frequency functions with weights that correspond to the input spectrum. This is at the possible loss of statistical efficiency, since the weighting/prewhitening in (2b) corresponds to the Markov (BLUE) estimate of a linear regression model. In Figure 7 the corresponding comparison
is given. It is quite clear that the prediction choice gives estimates with better accuracy. It is interesting to compare Figure 7 with Figure 3. Depending on the degree of approximation involved in the model estimation, one might favor different foci for a fit of the model frequency function from $u$ to $y$: If the bias error dominates the model error, it is better to choose 'focus'='simulation', otherwise 'focus'='prediction' is the natural choice. Note also the comments on the input spectrum that we gave in connection with the bias error.

4.4 Influence of Subspace/PEM

A key question for subspace methods is whether they really achieve the Cramer-Rao bound. Maximum-likelihood methods are know to do that asymptotically, so applying the prediction error method should give optimal accuracy (asymptotically). Now, that requires that the somewhat demanding minimization routine indeed gives the globally minimizing models, and numerical inaccuracies may hamper that. In Figure 8 the comparison is shown between the CVA-subspace estimate and the result from PEM (with 'lim'=0 and 'SSparameterization' = 'Canonical') (no robustification and canonical form parameterization). It is seen that PEM has a clear edge over CVA in these simulations. The result may be quite dependent on the input chosen. For a white noise Gaussian input, the two methods behave similarly. This is in line with the result of (Bauer and Ljung 2002). On the other hand, for band-limited (but still persistently exciting) inputs, PEM is considerably better than CVA.

5. THE TENNESSEE EASTMAN CHALLENGE PROCESS

The Tennessee Eastman identification and control challenge is described in (Downs and Vogel 1993). It consists of data from a chemical process, which is quite complex. The process has 7 inputs and 10 outputs and several non-linear features. The data themselves are simulated from a full scale nonlinear model, and the challenge is to obtain good models from these data.

The data consists of 13141 samples. In what follows, the first 6570 values were used for identification and the remaining 6571 values for validation purposes. A small portion of the data is shown in Figure 9.

![Fig. 7](image_url)

**Fig. 7.** The influence of the Focus: x-axis: Focus = 'prediction', y-axis: Focus='simulation'. 'Prediction' is the better choice in 65% of the cases.

![Fig. 8](image_url)

**Fig. 8.** A comparison between CVA-subspace method (x-axis) and the PEM-estimate (y-axis). PEM is the better choice in 64% of the cases.

![Fig. 9](image_url)

**Fig. 9.** A portion of input 1 and output 1 of the Tennessee Eastman data.

The data has been analyzed, e.g. in (Juricek et al. 2001). They found that a CVA model of order 23 works well.

We have applied subspace methods with different options to the data, as well as the prediction error method. For direct comparisons with (Juricek et al 2001), we also chose the order $n = 23$. For the prediction horizons we used the defaults ($N4hor = auto$) in the code which gave the horizon [24 10 10]. For the different models we have tested their ability to reproduce the validation data in simulation. That is, we computed the fit in $[yh, fits] = compare(datv, mod)$.

The results are shown in Table 1. The different models correspond to subspace estimates with CVA and MOESP weightings, with focus='prediction' or focus='simulation' (CVA(s) and MOE(s) for simulation focus). The table
also includes the prediction error estimated model (PEM) (with canonical parameterization and no robustification, ‘lim’=0 and a simulation focus), as well as the model estimated in (Juricek et al 2001), (“Juri”). The measure shown is 

\[
\text{fit} = (1 - \frac{\text{norm}(y - \bar{y})}{\text{norm}(y - \text{mean}(y))}) \times 100
\]

It thus differs from the common $R^2$ measure in that it is the norm and not the squared norm that is used. This generally gives lower figures for the fit.

<table>
<thead>
<tr>
<th>CVA</th>
<th>MOESP</th>
<th>CVA(s)</th>
<th>MOE(s)</th>
<th>PEM</th>
<th>Juri</th>
</tr>
</thead>
<tbody>
<tr>
<td>72.98</td>
<td>70.94</td>
<td>79.62</td>
<td>79.67</td>
<td>81.30</td>
<td>80.38</td>
</tr>
<tr>
<td>73.42</td>
<td>69.09</td>
<td>79.78</td>
<td>79.44</td>
<td>81.22</td>
<td>80.72</td>
</tr>
<tr>
<td>72.57</td>
<td>50.57</td>
<td>89.24</td>
<td>89.39</td>
<td>88.92</td>
<td>82.87</td>
</tr>
<tr>
<td>84.97</td>
<td>81.82</td>
<td>91.44</td>
<td>89.52</td>
<td>91.93</td>
<td>91.87</td>
</tr>
<tr>
<td>79.51</td>
<td>67.82</td>
<td>90.85</td>
<td>89.68</td>
<td>91.32</td>
<td>88.54</td>
</tr>
<tr>
<td>72.74</td>
<td>50.80</td>
<td>89.23</td>
<td>89.37</td>
<td>88.92</td>
<td>82.89</td>
</tr>
<tr>
<td>72.74</td>
<td>51.03</td>
<td>89.50</td>
<td>89.64</td>
<td>89.13</td>
<td>83.15</td>
</tr>
<tr>
<td>73.38</td>
<td>20.36</td>
<td>65.88</td>
<td>51.29</td>
<td>88.84</td>
<td>88.13</td>
</tr>
<tr>
<td>76.04</td>
<td>78.48</td>
<td>82.40</td>
<td>83.54</td>
<td>86.76</td>
<td>85.40</td>
</tr>
<tr>
<td>81.66</td>
<td>76.74</td>
<td>87.37</td>
<td>88.55</td>
<td>89.72</td>
<td>89.15</td>
</tr>
</tbody>
</table>

Table 1. The simulation precision for the 10 outputs (the rows) for models corresponding to different design variables.

6. CONCLUSIONS

We have illustrated some aspects of user choices in subspace identification methods. The results are based on simulations and examples, and one should be careful to draw far-reaching conclusions. However, the features shown here could be of some guidance in choosing options and methods.

7. ACKNOWLEDGMENTS

This research was supported by the Swedish National Research Council (VR).

Ben Juricek kindly provided the Tennessee Eastman data, as well as the model (“Juri”) that was estimated in the paper (Juricek et al 2001). I am also grateful to Dietmar Bauer who helped with several comments.

8. REFERENCES


Abstract

Keywords:
**Titel**  
Aspects and Experiences of User Choices in Subspace Identification

**Författare**  
Lennart Ljung

**Sammanfattning**  
.

**Nyckelord**  