On Canonical Forms, Parameter Identifiability and the Concept of Complexity

Rissanen, Jorma; Ljung, Lennart

1975

Document Version:
Publisher's PDF, also known as Version of record

ON CANONICAL FORMS, PARAMETER IDENTIFIABILITY AND THE CONCEPT OF COMPLEXITY

L. LJUNG
J. RISSANEN

Report 7527(C) November 1975
Department of Automatic Control
Lund Institute of Technology

For the 4th IFAC Symposium on Identification, Tbilisi, USSR 1976

ON CANONICAL FORMS, PARAMETER IDENTIFIABILITY AND THE CONCEPT OF COMPLEXITY

Lennart Ljung and Jorma Rissanen

Dept of Automatic Control
Lund Institute of Technology
S-220 07 LUND 7, SWEDEN

IBM Research
San Jose, California
USA

ABSTRACT

A method is presented to parameterize multi-output linear models for identification purposes. Ideas from realization theory are used as guidance in selecting a basis in a prediction space which in turn determines an appropriately parameterized model. The selection is interpreted as a data compression procedure, and the concept of complexity and entropy of a random variable is used as the criterion.

1. INTRODUCTION

During the past decade identification of single-output systems has become a routine as evidenced by numerous successful applications. In comparison, considerable difficulties arise when the same identification methods are attempted at the multi-output systems. The major source of trouble is due to the fact that there is no universal structure within which the systems could be parameterized. As a consequence, the identification problem includes the structure estimation as an essential part.

In this paper we shall discuss how a stationary, multi-output time series $\{y\}$ is modelled as an ARMA-process or in state space form, i.e.

$$y(t) + A_1 y(t-1) + \ldots + A_n y(t-n) = e(t) + C_1 e(t-1) + \ldots + C_n e(t-n) \quad (1)$$

or

$$\begin{aligned} x(t+1) &= F\,x(t) + K\,e(t) \\ y(t) &= H\,x(t) + e(t) \end{aligned} \quad (2)$$

where {e} is a sequence of independent random variables. The inclusion of an input signal to treat dynamical, stochastic systems is straightforward.

We shall treat the problem how to select a set of parameters describing the ARMA coefficients or the matrices F, K, H such that they may be uniquely determined in a parameter estimation procedure. We shall not discuss the parameter estimation part of the problem here, but let us assume that the selected parameters are determined using a prediction error identification method, such as the maximum likelihood method. See, e.g., [1], [2] and [3].

The choice of parameters is based on the idea of selecting a structure with its parameters in such a way as to make the parameter estimation a well-conditioned problem. The structure selection is interpreted as an information compression procedure, and the concept of complexity and entropy of a random variable (see [4]) will be used as a guiding criterion. From these preliminary studies in the present paper a perhaps more satisfying over-all approach to the identification problem was developed in [5] and [6].

2. CANONICAL FORMS AND THE HANKEL MATRIX OF THE IMPULSE RESPONSE

It is quite well known, see e.g. [7]-[11], that the Hankel matrix of the impulse response can be used as a starting point for describing the system as a state space or ARMA-representation.

For future reference, we shall briefly review the procedure here. Let

$$\bar{H}_S(z) = H_0 + H_1 z + H_2 z^2 + \ldots\ldots \quad ; H_0 = I \tag{3}$$

be the transfer function from {e} to {y}, and

$$H_{N,M} = \begin{pmatrix} H_1 & H_2 & H_3 & \ldots\ldots & H_M \\ H_2 & H_3 & H_4 & \ldots\ldots & H_{M+1} \\ \vdots & & & & \\ H_N & H_{N+1} & H_{N+2} & \ldots & H_{N+M} \end{pmatrix} \tag{4}$$

be the $Mn_y | Nn_y$ block Hankel matrix formed from the impulse response matrices $\{H_i\}$. Then if the process {y(t)} has a finite, say n-dimensional, representation (2) or an ARMA-representation (1), all matrices $H_{N,M}$ will have rank less or equal to n. The n first linearly indepen-

dent rows determine the observability or Kronecker indices, and they can be used as a basis for representations like (1) or (2).

From (3)

$$y(t) = \sum_{k=0}^{\infty} H_k \, e(t-k) \tag{5}$$

we obtain for $\hat{y}(t|t-r)$, the prediction of $y(t)$ based on $y(t-r)$, $y(t-r-1),\ldots$.

$$\hat{y}(t|t-r) = \sum_{k=r}^{\infty} H_k \, e(t-k) \tag{6}$$

Hence

$$Y_N(t) = \begin{pmatrix} \hat{y}(t+1|t) \\ \vdots \\ \hat{y}(t+N|t) \end{pmatrix} = H_{N,\infty} \begin{pmatrix} e(t) \\ e(t-1) \\ \vdots \end{pmatrix} \tag{7}$$

If the rank of $H_{N,\infty}$ is bounded when N increases, $\hat{y}(t+N|t)$ will eventually lie in the linear span of the other predictors. Therefore

$$\hat{y}(t+N|t) + A_1 \, \hat{y}(t+N-1|t) + \ldots + A_N \, \hat{y}(t+1|t) = 0 \tag{8}$$

Clearly, when the innovations are added to the predictions in order to obtain $y(t+N)$, the relation (8) gives an ARMA-representation of $\{y(t)\}$ with the LHS of (8) formed from $e(t+1),\ldots,e(t+N)$.

Analogously the linear relationship (8) can be expressed as

$$\begin{pmatrix} \hat{y}(t+2|t) \\ \vdots \\ \hat{y}(t+N|t) \end{pmatrix} = F \begin{pmatrix} \hat{y}(t+1|t) \\ \vdots \\ \hat{y}(t+N-1|t) \end{pmatrix} \tag{9}$$

which gives rise to the state space representation (2) in the following way:

Let $x(t+1) = [\hat{y}(t+1|t) \ \ldots \ \hat{y}(t+N-1|t)]^T$. Then

$$y(t) = [I, 0 \ 0 \ .. \ 0] \, x(t) + e(t)$$

and

$$x(t+1) = \begin{pmatrix} \hat{y}(t+1|t-1) \\ \vdots \\ \hat{y}(t+N-1|t-1) \end{pmatrix} + \begin{pmatrix} K_1 \, e(t) \\ \vdots \\ K_{N-1} e(t) \end{pmatrix} = F \, x(t) + K \, e(t)$$

If the components of $Y_N(t)$ that enter in (8) and (9) are restricted to a basis that spans $H_{N,\infty}$ then the representation will be unique (for

this basis). Consequently, the problem of parameterization or of representation of $\{y(t)\}$ can be seen as the problem of choosing of an appropriate basis in the "predictor space". This interpretation has very elegantly been persued by Akaike [11]. Usually, the "first" linearly independent rows are chosen as the basis.

Let us consider a simple example to which we shall return several times below.

Example 1. Consider a process with 2 components, $n_y = 2$, and suppose that the 1st, 2nd and 4th rows are the first linearly independent rows. Then the basis consists of $\hat{y}_1(t+1|t)$, $\hat{y}_2(t+1|t)$ and $\hat{y}_2(t+2|t)$ and (9) has the form

$$\begin{pmatrix} \hat{y}_1(t+2|t) \\ \hat{y}_2(t+2|t) \\ \hat{y}_2(t+3|t) \end{pmatrix} = \begin{pmatrix} x & x & 0 \\ 0 & 0 & 1 \\ x & x & x \end{pmatrix} \begin{pmatrix} \hat{y}_1(t+1|t) \\ \hat{y}_2(t+1|t) \\ \hat{y}_2(t+2|t) \end{pmatrix} \quad ; \quad K = \begin{pmatrix} x & x \\ x & x \\ x & x \end{pmatrix} \tag{10}$$

where "x" denotes a numerical value. The zero at the (1,3) place comes from the way basis is chosen; we know that $\hat{y}_1(t+2|t)$ lies in the span of the rows above it. Correspondingly, the ARMA-representation has the form

$$A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad A_1 = \begin{pmatrix} x & x \\ 0 & x \end{pmatrix} \qquad A_2 = \begin{pmatrix} 0 & 0 \\ x & x \end{pmatrix} \tag{11}$$

The matrices C_i may be fully parameterized.

3. PARAMETER IDENTIFIABILITY AND CANONICAL FORMS

The concept of parameter identifiability has been given some different definitions, but in essence we may say that a certain model parameterization, M, is Parameter Identifiable (PI) if the identification criterion (the prediction error loss function) has a unique global minimum, corresponding to a true description of the system.

In the previous section we described how a certain set of canonical forms may be constructed from the first linearly independent rows (i.e. the Kronecker or observability indices). It is true that canonical forms are by construction such that the corresponding model parameterizations are PI. This requires however, that certain a priori information (namely the Kronecker indices) about the system is avail-

able, and it has been suggested that these be estimated first, e.g. [9], [16], [17].

For identification the primary interest is in model parameterizations that yield PI rather than in canonical forms, i.e. we do not bother if a given system may be represented within several parameterizations, as long as it is uniquely represented within each. This point has also been stressed by Glover and Willems [12], and it makes it possible to extend the search for suitable structures outside the canonical forms.

It is clear from the expressions (8) and (9) that the cause of non-uniqueness within a given representation is that the components of $\hat{y}(t+k|t)$ $k=1,...N$ involved in the expressions may be linearly dependent. Therefore, by restricting these components to a linearly independent set, we do obtain unique representations. In the ARMA representations (1) the restriction is made by fixing parameters in A_i to zero, that correspond to components of y not belonging to this set. Let us formulate this rather straightforward result, which is closely related to Luenberger [13], as a theorem.

Theorem: Let $x(t) = \{\hat{y}_{i_k}(t+k|t);\ k=1,...,N,\ i_k \in I_k \subset \{1,...,n_y\}\}$ be a set of components of $Y_N(t)$, such that the corresponding rows of H_N are linearly independent, and such that $I_1 = \{1,...,n_y\}$. Assume further that if $i \in I_k$, then $i \in I_{k-1}$. Then the following parameterizations are PI-parameterizations:

° ARMA:

$A_0 = I$, $A_r = (a^r_{ts})$ where $a^r_{ts} = 0$ if $s \in I_{N-r+1}$; $r=1,..,N$

$C_0 = I$, $C_r = (c^r_{ts})$ $r=1,..,N$

° State Space:

$F = (f_{rs})$, where if $x_r(t)=\hat{y}_j(t+k|t)$ is the r:th component of x(t) (hence $j \in I_k$) and also $j \in I_{k-1}$, then $(f_{rs}, s=1,...,n) =$ $= (0,..,1,..,0)$ where the 1 is on the place corresponding to $\hat{y}_j(t+k-1|t)$.

$K = (k_{rs})$; $H = (I\ 0\ ...\ 0)$

If we return to Example 1, we could e.g. consider the rows 1, 2 and 4 as a basis for H_N, where row 3 not necessarily lies in the span of rows 1 and 2. Then the parameterization would be

$$F = \begin{pmatrix} x & x & x \\ 0 & 0 & 1 \\ x & x & x \end{pmatrix} \tag{12}$$

and

$$A_0 = I,\ A_1 = \begin{pmatrix} 0 & x \\ 0 & x \end{pmatrix},\quad A_2 = \begin{pmatrix} x & x \\ x & x \end{pmatrix} \tag{13}$$

instead of (10) and (11). Notice in particular that the zero in the (1,3) place in (12) can be replaced by a parameter without the PI-property being lost.

The conclusion of this simple example is that the parameterizations (12) and (13) yield PI even though they contain more parameters than the "canonical" ones (10) and (11). This raises the question whether the smallest possible parameterization for identification is of any interest at all. In a certain sense it is so, since if the (1,3) element in (12) really is zero, then the other five parameters can be estimated more accurately in (10) than in (12). However, since this parameter is identifiable, it helps to give a better fit to the measured data, and the identification criterion will take a smaller value for the structure (12) than for (10). The hypothesis that the parameter has the value zero can also be tested using either the estimate covariance or Akaike's [14] criterion for the number of parameters.

The procedure of testing whether certain parameters are zero in various structures is in fact quite common, but we claim that this is not done to minimize the number of estimated parameters, but to obtain the simplest possible model. To return to our example, it is indeed doubtful whether the model (12) is "simpler" if the (1,3) element is zero. We would say it is not, and that there is no point in setting this element to zero, since it does not effect the PI-property.

The choice of parameterization described in the theorem leaves in general several possible choices for a given system (i.e. they are not "canonical forms"). For example 1 it may very well be that not only rows 1, 2 and 4, but also the first three rows form a basis for H. The parameterization of F corresponding to the latter choice is then

$$F = \begin{pmatrix} 0 & 0 & 1 \\ x & x & x \\ x & x & x \end{pmatrix} \tag{14}$$

Now, it is not completely without importance whether (12) or (14) is chosen. For example, if the estimated value of the (1,3) element in (12) is small, this means that $\hat{y}_1(t+2|t)$ is "close" to the linear span of $\hat{y}_1(t+1|t)$ and $\hat{y}_2(t+1|t)$. Then (14) is a badly conditioned parameterization, i.e. relatively large changes in the parameters can be made without too much an effect on the transfer function coefficients. Moreover, (14) and (12) do not describe exactly the same set of linear systems as the parameters vary.

We may summarize the discussion of this section as follows.

- o A PI model parameterization is obtained as soon as a basis for H is chosen.
- o There is no need to elaborate on the structure of the basis (such as certain rows not in the bases being linearly dependent on the above ones) unless this gives simpler models.
- o In general we have a choice of bases (parameterizations).
- o For a given model order (n), the n "most linearly independent" rows of H should be chosen as a basis for the parameterization.

We shall return to the issue of how "the most linearly independent rows of H" can be interpreted, after what may seem to be a digression.

4. THE CONCEPT OF COMPLEXITY

The concept of complexity has been discussed in various contexts. We shall here follow the exposition by van Emden [4] and give a few details that are relevant for the discussion above.

Complexity is closely related to interaction. The more interaction there is in an object the more complex it is. Interaction can be measured in terms of entropy and along these lines a general definition can be given. We will here be concerned with the complexity of random vectors. Since interaction within a random vector can be measured by the correlations, the complexity can be expressed using the covariance matrix. Van Emden derives the following expressions:

$$C_1 = \max \left(H(x_1) + \ldots + H(x_n) - H(x_1,\ldots,x_n) \right) = -\frac{1}{2} \sum_{i=1}^{n} \log(n\lambda_i) \quad (15)$$

Here $(x_1,\ldots,x_n)$ is an n-dimensional random vector, $H(x_i)$ are the entropies of the component and $H(x_1,\ldots,x_n)$ is the entropy of the whole

vector. λ_i are the eigenvalues of the covariance matrix R_x, which has been normalized so that $\operatorname{tr} R_x = 1$. The maximum is taken over all orthonormal transformations.

By expanding the logarithm around $\lambda_i = 1/n$ up to second order terms, C_1 can be approximated by

$$C = 1/n \sum_{i=1}^{n} (\lambda_i^2 - 1/n^2) = 1/n \sum_{i,j=1}^{n} r_{ij}^2 - 1/n^2 \qquad (16)$$

where r_{ij} are the components of R_x and again $\sum_i r_{ii} = 1$. The measure C has the advantage of being computable directly from the matrix, without having to find the eigenvalues, but is less sensitive to complete correlation. Observe, that then measures C_1 and C are zero for $R_x = I$ and positive for all other matrices.

The complexity is closely related to data compression, van Emden [4]. If a random vector has high complexity, then there are strong interactions between the components and it can be compressed into a lower dimensional vector without loosing much information contained in the original vector.

The complexity is clearly also related to the condition number of the covariance matrix: If the complexity is large, then the covariance matrix is ill-conditioned.

5. CHOICE OF BASIS

Let us return to the random vector $Y_N(t)$ defined in (3.5). Its covariance matrix is

$$\tilde{R}_Y = E\, Y_N(t)\, Y_N(t)^T = H_N\, H_N^T \quad ; \quad H_N = H_{N,\infty} \qquad (17)$$

Hence linear dependence of the rows of H_N is connected with correlation between the corresponding random variables: If $\hat{y}_1(t+1|t)$ and $\hat{y}_1(t+2|t)$ in our example are completely correlated, then the first and third rows of H_N are linearly dependent and vice versa. The third row of H_N is "close" to the linear span of the two first ones if the three random variables $\hat{y}_1(t+1|t)$, $\hat{y}_2(t+1|t)$ and $\hat{y}_1(t+2|t)$ have strong interaction (and, consequently, the submatrix of $\tilde{R}_y$ corresponding to these variables has a large complexity).

We now see a possibility to interpret "the most linearly independent rows" stated in the end of Section 3. The n most linearly inde-

pendent rows of H_N are those, whose corresponding nxn submatrix of $\tilde{R}_y$ has the least complexity.

If we consider $Y_N(t)$ for some suitably large N as the available information about the past; as the "true" basis for the predictor space, then the problem of finding an n-dimensional representation is the problem of compressing the n_yxN vector $Y_N(t)$ into an n-vector x_n, the basis for the parameterization. This should naturally be done so that as much information as possible is retained. The choice of taking x_n so that $Ex_n x_n^T$ is the submatrix of $\tilde{R}_y$ that has the least complexity is then the same as choosing those n components of $Y_N(t)$ as a basis that are most difficult to compress further. This seems to be a sensible solution.

To summarize this discussion, we suggest that the basis for the parameterization should be chosen as consisting of those components of Y_N(say i_j; j=1,..,n) which give the nxn submatrix of $\tilde{R}_y$ (made up from the elements (i_j,i_k; j,k=1,...,n)) with the smallest complexity of all its nxn submatrices, subject to the constraint that if row i is chosen then also row i - n_y should be. Once the basis is chosen it is clear from the theorem how the model is to be parameterized.

6. A PROCEDURE FOR IDENTIFYING MULTIVARIABLE SYSTEMS

The procedure described in Section 5 requires that $H_N H_N^T$, or an estimate of it, is known. Since no structure must be imposed on H_N in the preliminary analysis, $H_N H_N^T$ should be estimated using basically a non-parametric method. Several possibilities exist, but we would suggest the following one. Adjust the parameters of a high order autoregressive model to the data using the method of least squares. In this way a fairly good estimate of the sequence of innovations is obtained by using fast and straightforward calculations. Subtract the innovations from the measured y, which yields $\hat{y}(t+1|t)$ t=1,... . Repeat N times which gives $\hat{y}(t+k|t)$ t=1,... ; k=1,..,N. Then form the estimate of $\tilde{R}_y$ ($=H_N H_N^T$) by simply taking for the (i j) block

$$r_{ij} = \frac{1}{M} \sum_{t=1}^{M} \hat{y}(t+i-1|t)\ \hat{y}(t+j-1|t)^T \qquad \text{(M=number of data)} \qquad (18)$$

It is known, see [15] or [16] that the observability indices can also be obtained from one "corner" of the covariance matrix of the

process y since this has the same rank properties as H. Therefore also this submatrix (times its transpose) may be utilized for the selection of a basis, using the complexity criterion. This procedure can be given an independent interpretation since it gives a well-conditioned way of determining the A-matrices using the method of instrumental variables.

We now suggest the following procedure for identification of multivariable systems:

- Form from the recorded output sequence y the estimates of $H_N H_N^T$ (or of the covariance matrix) as described above. The number N is chosen suitably large to reflect the largest time lag one is prepared to accept in the model.
- Choose the model order n and determine the nxn submatrix of $H_N H_N^T$ that has the smallest complexity according to Section 4. The corresponding components of Y_N are then chosen as the basis for the parameterization as given in the theorem.
- For the model order n, and the given parameterization then the prediction error estimation method is applied, the model that minimizes the prediction error covariance is determined and the corresponding value of the criterion, V_n, is calculated.
- The procedure is repeated for higher order models, n+1,n+2, ..., until the decrease in V_k is not significant any more. For instance, Akaike's [14] criterion $V_n + n_\theta$ where n_θ is the number of estimated parameters ($=n \cdot n_y \cdot 2$ for the parameterization of the theorem) could be minimized with respect to n.

7. CONCLUDING REMARKS

We have stressed the use of parameter identifiable model parameterizations rather than canonical forms. The reason can be expressed as follows: For a realization of a given order, say n, the number of parameters in a canonical form based on the observability indices depends on the actual values of these. This reflects the fact that the indices contain certain information about the structure of the basis. However, this information can be obtained also in the parameter estimation stage, and it corresponds to certain parameters having the

value zero. Now, it is usually an easier problem to estimate a parameter than to check for linear dependence in the (estimated and noisy) Hankel matrix of the impulse responses. Therefore it seems more natural to defer this analysis to the parameter estimation step than to extract this information from the Hankel matrix.

With this philosophy we have not only sought a parameterization that will be identifiable, but we have also tried to make it as well conditioned as possible, so that the estimation step will have good numerical properties. This feature is difficult to introduce in the canonical form approach; see also [5] and [6].

As a final remark, the rank testing procedures are difficult statistical problems. Our procedure does not involve rank testing or any decision-making threshold. All decisions are based on direct comparisons of real numbers (the complexities). The order of the model is determined by use of the prediction error criterion, which has successfully been used for single output systems, [1], [14].

References

1. Åström, K J and Bohlin, T: Numerical identification of linear dynamic systems from normal operating records. IFAC (Teddington) Symposium, 1965.
2. Caines, P and Rissanen, J: Maximum likelihood estimation of parameters in multivariable Gaussian stochastic processes. IEEE Trans IT-20, No 1, 1974.
3. Ljung, L: On the consistency of prediction error identification methods. In Advances in System Identification (R K Mehra and D G Lainiotis, Eds), Marcel Dekker Inc, New York, 1976.
4. van Emden, M: Analysis of complexity. Math Centr Tracts, 35, Amsterd. 1971.
5. Rissanen, J: Minimax entropy estimation of models for vector processes. In Advances in System Identification (R K Mehra, D G Lainiotis, Eds), Marcel Dekker, Inc, 1976.
6. Rissanen, J and Ljung, L: Estimation of optimum structures and parameters for linear systems. Presented at Symposium - Advanced School on Algebraic System Theory, Udine, Italy, June 1975. Proceedings by CISM and Springer Verlag (to appear).
7. Kalman, R E, Falb, P L and Arbib, M A: Topics in mathematical system theory. McGraw-Hill, New York, 1969.
8. Popov, V M: Invariant description of linear time-invariant controllable systems. SIAM J Contr 10, 252-264, 1972.
9. Rissanen, J: Basis of invariants and canonical forms for linear dynamic systems. Automatica 10, 175-182, 1974.

10. Dickinson, B W, Kailath, T and Morf, M: Canonical matrix fraction and state-space descriptions for deterministic and stochastic linear systems. IEEE Trans AC-19, 656-667, 1974.

11. Akaike, H: Stochastic theory of minimal realization. IEEE Trans AC-19, 667-674, 1974.

12. Glover, K and Willems, J C: Parameterization of linear dynamical systems - canonical forms and identifiability. IEEE Trans AC-19, 640-646, 1974.

13. Luenberger, D G: Canonical forms for linear multivariable systems. IEEE Trans AC-12, 290-293, 1967.

14. Akaike, H: A new look at the statistical model identification. IEEE Trans AC-19, 716-723, 1974.

15. Rissanen, J and Kailath, T: Partial realization of random systems. Automatica 8, 389-397, 1972.

16. Tse, E and Weinert, H: Structure determination and parameter identification for multivariable stochastic linear systems. Proc 1973 JACC, Colombus, Ohio. (See also IEEE Trans AC-20, 603-612, 1975.)

17. Guidorzi, R: Canonical structures in the identification of multivariable structures. Automatica 11, 361-374, 1975.