Communications in Information and Systems
Volume 20, Number 2, 163–207, 2020

# Least squares optimal realisation of autonomous LTI systems is an eigenvalue problem*

Bart De Moor†

We outline the solution of a long-standing open problem in system identification, on how to find the best least squares realisation of an autonomous linear time-invariant (LTI) dynamical system from given data. The global optimum is found among all stationary points of a least squares objective function, which we show to correspond to the eigen-tuples of a multi-parameter eigenvalue problem (MEVP). Such an MEVP can be solved by applying Forward (multi-) Shift Recursions to the given set of multivariate polynomial equations, generating so-called block Macaulay matrices, the null space of which can be modelled as the observability matrix of a multi-dimensional shift-invariant linear commutative singular system. The state equations of this system can be found from multi-dimensional realisation theory. From the corresponding eigen-tuples, one can then find the optimal parameters of the best LTI autonomous model. Our solution methodology uses ingredients from algebraic geometry, operator theory, multi-dimensional system theory and numerical linear algebra, and ultimately requires as basic building blocks only the singular value decomposition and eigen-solvers.

Surprisingly enough, the conclusion is that the globally optimal model in 1D least squares realisation, can be found exactly from multi-dimensional realisation. In addition, we describe several new, previously unknown, properties that characterise the optimal model and its behaviour.

*This work was supported by **KU Leuven**: Research Fund (projects C16/15/059, C32/16/013, C24/18/022), Industrial Research Fund (Fellowship 13-0260) and several Leuven Research and Development bilateral industrial projects; **Flemish Government Agencies: FWO** (EOS Project no 30468160 (SeLMA), SBO project I013218N, PhD Grants (SB/1SA1319N, SB/1S93918, SB/151622)), **EWI** (PhD and postdoc grants Flanders AI Impulse Program), **VLAIO** (City of Things (COT.2018.018), PhD grants: Baekeland (HBC.20192204) and Innovation mandate (HBC.2019.2209), Industrial Projects (HBC.2018.0405)); **European Commission** (EU H2020-SC1-2016-2017 Grant No. 727721: MIDAS), the European Research Council (ERC) (EU Horizon 2020 Advanced Grant Agreement No. 885682).

†Fellow IEEE & SIAM.

## 1. A tribute

This paper is dedicated to Thomas Kailath at the occasion of his 85th birthday. From our first encounter at a conference in Valencia, Spain in 1986 [20], over my postdoc stay at Stanford in 1988-1989 with Tom in the Information Systems Lab and with Gene Golub in the Computer Science department, to the many other events and occasions where we met: all of them have had a lasting impact on me as a scientist and an engineer. Our weekly meetings often started with the introductory *"There are 168 hours in a week. Bart, what have you done this week?*". Thirty years later, I must admit that, with my own PhD students, I use that line too! Tom's broad and deep scientific expertise in information theory, signal processing and system theory, with an eye towards applications, has definitely inspired me to pursue a career in what today is called *mathematical engineering*.

In this paper we will outline the exact solution to a long-standing open mathematical engineering problem in system identification, how to find the best linear time-invariant (LTI) model for a given set of data. In doing so we will borrow results from *(multi-)dimensional system theory*, *operator theory*, *algebraic geometry* and *numerical linear algebra*. For me as a PhD student, Tom's classic on *Linear Systems* [55] was a real eye-opener, and (not so) surprisingly, it contained many seeds for the current paper. I have always been fascinated by the discussion on Model reduction in Section 10.4, the last page of the last Chapter 10 of the book, where Tom announces, almost as a *'a note added in proof'*, the now famous result of Adamjan-Arov-Krein [2], mathematicians who were working in operator theory, on how to best approximate a rank deficient (double infinite) Hankel matrix with Markov parameters of a high order system, by a Hankel matrix of a lower rank, where 'best' is measured in the Hankel norm. The answer is in terms of the singular value decomposition (SVD) of the Hankel matrix. Later on these results were elegantly rephrased by Glover in the state space formalism [47]. That the data in rank deficient Hankel matrices can be 'realised' into a rational form, is an old result of Kronecker [61], who investigated the conditions for a Taylor series to be the expansion of a rational function. This is the case when the Hankel matrix with the coefficients of the expansion is rank deficient. The link with transfer functions is readily made, and the realisation into state space models was described by Ho and Kalman [52] and many others, with Kung [62] and Zeiger-McEwen [90] bringing in the SVD.

In Chapter 5 (pp. 322–325) of his book [55], Tom also hints at the problem of 'noisy' data, so that in practice one does not start from a Hankel matrix that is rank deficient. The 'noisy' problem has been tackled in the

past with a tsunami of papers (see Subsection 10.2 for some references), by heuristically applying realisation algorithms that are 'correct' on exact data, but that only deliver approximate results with 'noisy' data, where the quality of the approximation is not quantified. One of them is the paper we wrote at the occasion of Tom's sixtieth birthday, where we describe a nonlinear generalisation of the SVD, called the *Riemannian SVD* [34], to tackle the said 'noisy' realisation problem.

An obvious relevant question is therefore how to optimally approximate, in a least squares sense (so not in the Hankel norm), a given finite data sequence by the output of an autonomous LTI system?

This is exactly the problem we will be solving here: how to modify the given data in a least squares sense, so as to make the modified data compatible with an autonomous LTI system of a pre-specified given order.

We will use results from mathematical (sub-)disciplines that are also close to Tom's hart: *System theory* with Kalman [5] [57] [58] [59] as one of its founding fathers, with numerous contributions by Tom, and many applications, including system identification (see e.g. [67]); *Operator Theory*, comprising the study of operators and their algebras on infinite dimensional spaces (see e.g. [44] [45]); *Algebraic Geometry*, since Descartes the happy marriage between the classic geometry of the Greeks and the manipulations of equations by the algebraists, its central 'objects' being multivariate polynomial equations (ideals) and their sets of roots (varieties) (see e.g. [24] [25] [83]); Last but not least *Numerical Linear Algebra*, matured since the advent of advanced computing more than 60 years ago, dealing with finite precision floating point numerical algorithms (see e.g. [18] [27] [49] [82]).

One of the main ingredients in Tom's books [55] [56] is linear algebra, with a special role for the EVP [88]. Indeed, in system theory, eigenvalues and -vectors characterise stability, controllability and observability of LTI dynamical systems, but even so in optimal control the steady state versions of the LQR problem and the Kalman filter derive from Algebraic Riccati Equations, which are Hamiltonian eigenvalue problems in disguise. Similarly, the solutions to their $H_\infty$ counterparts derive from symplectic eigenvalue problems. This paper adds another explicit case to the list of system and control problems that can be solved as a(n) (series of) EVP(s): The explicit solution of the least squares realisation problem in terms of an Multi-parameter Eigenvalue Problem (MEVP) and multi-dimensional realisation theory, ultimately only requiring tools from numerical linear algebra, such as the eigenvalue and singular value decompositions.

For all of these reasons, we deem our contributions here to be a more than appropriate birthday present for Tom!

This paper is organised as follows: In Section 2, we discuss the classical *system identification loop*. We introduce the notion of *data misfit* and describe precisely what we mean by '*solving*' an identification problem. Section 3 introduces the notions of kernel, image and state space model representations of LTI systems, treats *Forward Shift Recursions* (FSRs) and how they induce shift-invariant subspaces, which is the topic of Section 4. In Section 5, we derive the first-order necessary conditions for a global minimum starting from a kernel model representation, which leads to a multiparameter eigenvalue problem (MEVP) that delivers the optimal model parameters. In Section 7 we turn to image model representations, which will provide additional insights that characterise the optimal models and their behaviour. In Section 8 we focus on the illuminating case of a first order approximation. In Section 9, we discuss some new properties of the optimal solution: A Beurling-Lax-Halmos (BLH) like property; The orthogonal decomposition of the given data vector into 'exact' data and a misfit while optimising a Riemannian metric; An optimality property, satisfied by the misfit, reminiscent of Walsh's Theorem but now for finite dimensional data (to be interpreted as a 'double' BLH property); Last but not least, the demonstration that the ambient data space can be partitioned in three complementary shift-invariant subspaces, generated by the optimal model poles. In Section 10 we discuss applications, briefly enumerate the heuristic algorithms that have been proposed in the past and discuss some potential extensions. Concluding remarks can be found in Section 11.

Space limits do not allow us to go into more detail on the numerical linear algebra challenges of and potential algorithms for our approach, which merit a full paper in their own right, nor can we demonstrate larger examples here. We will restrict ourselves to some simple didactical examples. Yet we hope that the steps we outline are sufficiently illuminating for the more general cases as well. Our language will be informal and expository, without providing formal proofs nor extensive derivations.

## 2. Challenges in the system identification loop

### 2.1. Steps in a typical modelling set-up

A typical mathematical modelling cycle or system identification loop proceeds by 1. Collecting and preprocessing data; 2. Selecting a pre-specified model class parametrised by unknown parameters; 3. Choosing an appropriate approximation criterion; 4. Numerically solving a nonlinear optimisation problem that outputs optimal parameters; 5. Validating the resulting model;

6. Re-iterating the whole loop as long as necessary with different models or even model classes.

In this paper, we will exclusively focus[1] on Steps 3 and 4. The model class of Step 2 consists of causal, autonomous single-output LTI minimal models, with a fixed predefined order (McMillan degree, number of poles), but otherwise unknown model parameters and initial states. In a certain sense, this is the simplest class of models for LTI systems, representing only a small subset. We will show that for this specific model class and for a least squares objective function (Step 3), the globally optimal solution can be found among the eigen-tuples of an MEVP, which can be solved exactly using multi-dimensional realisation theory.

### 2.2. Data misfit

*If the model does not fit the data*
*Let the data fit the model*

In general, it is not so difficult to decide whether a given data set is generated by a model of some specified model class. As an example, if given scalar data are effectively generated by an autonomous minimal LTI system of order $n$, a sufficiently large Hankel matrix with the data, will be of rank $n$. If that would be the case, we can derive the specific dynamical equations of that model, in which we 'realise' the model from the data, via 'realisation' algorithms. Realisation theory studies the transformation of one model representation into another one that is equivalent. We call the data 'exact' when they are generated by a model in the model class at hand. Phrased in Willems's *behavioural framework* [89]: They belong to the *behaviour* of that model. For a given model class exact data represent a model exactly and there is a bijective map from exact data to any of the equivalent model representations. Said in other words, realisation is about exact modelling. As we will see, for LTI systems, we can use either *kernel*, *image* or *state representations* for the mappings and the (in-)compatibilities between models and behaviours. Unfortunately, in most applications, the given data are inexact: they are not compatible with the selected model class, i.e. they do not belong to the behaviour of a model. So the basic 'engineering' approach to tackle this

[1] In particular, we assume that in Step 1 data have been properly collected and preprocessed, e.g. they are equidistantly sampled, the requirements induced by the Nyquist sampling theorem have been properly dealt with, etc. We will also not elaborate any further on Steps 5 and 6 of the identification loop but refer to standard books on system identification.

paradox, is to modify the data so that the modified data belong to the behaviour of the specified model class. The difference between the given data and the modified data will be called *the misfit*. Obviously, there is an infinite number of ways in which we could modify the observed data, but the modification that we will pick is the one that minimises the 2-norm (sum-of-squares) of the misfit. This will require a least squares optimal minimisation over the model parameters and the misfit sequences. Obviously, in this framework, exactness is a relative notion: the 'non-exactness' of the given data (the misfit) and its distance from the behaviour of the optimal model can be quantified, as we will see in Subsection 9.2 in a metric that is induced by the optimal model. When the complexity of the model – in this paper the order – is allowed to increase, we expect the misfit to decrease, and the model's sensitivity to perturbations in the data to increase, which is a manifestation of the so-called bias-variance trade-off, which however we are not going to consider here. We will assume the model's order to be specified and fixed.

### 2.3. What do we mean by a 'solution' of the least squares realisation problem?

Typically, in order to minimise constrained non-linear least squares objective functions as the ones that occur in Step 3 of the identification cycle, an iterative optimisation algorithm is required, with all due challenges: the need for appropriate initial guesses inducing issues of reproducibility, iteration step choices (all kinds of variations and accelerations of steepest descent) with potential (slow) convergences issues, and in most cases, the occurrence of local minima and the impossibility of testing which one of them is global. As a result, heuristics largely prevail, turning system identification (and machine learning) more into an art (with a large 'bag of tricks') than a science. Even for the identification of LTI dynamical models with 'established' approaches (e.g. *Subspace Identification* [84], *Prediction Error Methods* [67] or *Errors-in-Variables* [77]), one can not guarantee global optimality of the models. We do not suggest at all that these 'approximate' methods are not successful. On the contrary, they have been and continue to be hugely successful in important industrial challenges of simulation, prediction, filtering, control, monitoring, state estimation, soft-sensoring, fault detection, etc, and successful software suites and even companies, have been built on them. Yet, the 'solution' they provide, is at best approximative. Therefore, we need to clarify more precisely what we mean by 'solution'.

For instance, algebraic problems, like rooting a polynomial up to degree 4, are solved by formulas in terms of radicals. In analysis, we can solve certain integrals exactly using formulas. In mathematical engineering, we can consider a problem to be solved when it reduces to a set of linear equations, singular value decompositions (SVD) or eigenvalue problems (EVP), to an optimisation problem that is convex [19] or to a sequence of such algorithmic steps, from which we can guarantee global optimality. For each of these numerical linear algebra 'tools', there are efficient and numerically reliable algorithms that calculate all 'solutions' within a level of accuracy that is guaranteed by numerical analysis research results (e.g. forward and backward stability), on a machine with standardised floating point arithmetic. We call these problems 'solved' as no heuristics are involved and global optimality is guaranteed up to within machine precision. In this paper, we reduce the least squares realisation problem to an MEVP, which can be solved exactly, in the precise meaning as outlined above, via multidimensional realisation theory. The optimal solutions will derive from the spectra we calculate from several EVPs. Because the derivation that characterises optimality only contains linear algebra (sets of linear equations, eigenvalue and singular value problems), we consider the least squares realisation problem for autonomous single-output LTI models to be 'solved' in the spirit described above.

## 3. The behaviour of autonomous LTI systems

The model class we consider is that of autonomous LTI single-output systems and their corresponding behaviour for 'exact' data, i.e. data that are compatible with the models considered. We discuss kernel, state and image representations of these exact data. Let the $N$ scalar exact data $\hat{y}_k \in \mathbb{R}, k = 0, 1, \ldots, N-1$ satisfy an $n$-th order difference equation

$$\hat{y}_{k+n} + \alpha_1 \hat{y}_{k+n-1} + \ldots + \alpha_{n-1} \hat{y}_{k+1} + \alpha_n \hat{y}_k = 0 \,, \tag{1}$$

where $\alpha_i \in \mathbb{R}, i = 1, \ldots, n$ are the model coefficients and $k \in \mathbb{N}$ are the discrete time indices. Without loss of generality the leading coefficient $\alpha_0 = 1$. The 'hat' refers to the fact that the data $\hat{y}_k$ are exact, i.e. compatible with the model: they are generated by the difference equation and therefore belong to the behaviour (the allowed trajectories) of the model. We assume that the number $N$ of data is 'sufficiently large' (see below). We can recursively apply equation (1), a process that we call a *Forward Shift Recursion*

(FSR) and write the result as $T_a \cdot \hat{y} = 0$, where the vector $\hat{y} \in \mathbb{R}^N$ contains the consecutive data $\hat{y}_k$ and $T_a \in \mathbb{R}^{(N-n)\times N}$ is a banded Toeplitz matrix

$$T_a\,\hat{y} = \begin{pmatrix} \alpha_n & \alpha_{n-1} & \dots & \dots & \alpha_1 & 1 & 0 & \dots & 0 \\ 0 & \alpha_n & \alpha_{n-1} & \dots & \dots & \alpha_1 & 1 & \dots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \dots & \alpha_n & \alpha_{n-1} & \dots & \dots & \alpha_1 & 1 \end{pmatrix} \begin{pmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \vdots \\ \hat{y}_{N-1} \end{pmatrix} = 0.$$

This banded Toeplitz matrix $T_a$ contains the model information, while its kernel contains all possible vectors $\hat{y}$ that are compatible with the model, i.e. the behaviour. Therefore, we call this a *kernel representation* of the (exact) data. Because the leading coefficient of the difference equation $\alpha_0 = 1$, obviously $T_a$ is of full row rank: $\text{rank}(T_a) = N - n$. The nullity of $T_a$, i.e. the dimension of its null space, is $n$, the order of the difference equation. So the behaviour of the difference equation (1) is an $n$-dimensional subspace of the ambient vector space $\mathbb{R}^N$ and is the null space of the banded Toeplitz matrix $T_a$. It is straightforward to show that the difference equation (1) can also be modeled by a canonical state space model. Picking as a state vector $\hat{x}_k^{\text{ob}} = (\hat{y}_k\ \hat{y}_{k+1}\ \dots \hat{y}_{k+n-1})^T \in \mathbb{R}^n$, we easily find

$$\hat{x}_{k+1}^{\text{ob}} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \dots & 1 \\ -\alpha_n & -\alpha_{n-1} & \dots & \dots & -\alpha_1 \end{pmatrix} \cdot \hat{x}_k^{\text{ob}} = A_{\text{ob}} \cdot x_k^{\text{ob}}\ ,$$

$$\hat{y}_k = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \end{pmatrix} \cdot \hat{x}_k^{\text{ob}} = C_{\text{ob}} \cdot \hat{x}_k^{\text{ob}}\ , \tag{2}$$

in which the super- or sub-script '$^{\text{ob}}$' refers to the fact that this state space representation is in the observable canonical form. The recursiveness of the state space model (2) implies that $\hat{y}_k = C_{\text{ob}} \cdot A_{\text{ob}}^k \cdot \hat{x}_0^{\text{ob}}$, so that

$$\hat{y} = \begin{pmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \vdots \\ \hat{y}_{N-1} \end{pmatrix} = \begin{pmatrix} C_{\text{ob}} \\ C_{\text{ob}}A_{\text{ob}} \\ \vdots \\ C_{\text{ob}}A_{\text{ob}}^{N-1} \end{pmatrix} \cdot \hat{x}_0^{\text{ob}} = \Gamma_N^{\text{ob}} \cdot \hat{x}_0^{\text{ob}}\ . \tag{3}$$

Here, the behaviour is characterised as the image (range) of the (extended) observability matrix $\Gamma_N^{\text{ob}}$, reason why we call this model an *image representation*. All possible 'compatible' 'exact' data vectors $\hat{y}$ can be generated for

different choices of $\hat{x}_0^{\text{ob}}$. The dimension of the range (column space) of $\Gamma_N^{\text{ob}}$ is $n$ (provided that $N \geq n$) as can easily be seen from the first $n$ rows of $\Gamma_N^{\text{ob}}$ being the identity matrix. Hence, the LTI system is observable.

## 4. Single-shift-invariant subspaces

### 4.1. Single-output observability matrices

The 'cyclic' structure of the observability matrix $\Gamma_N^{\text{ob}}$, with the increasing integer powers of $A_{\text{ob}}$, induces a special so-called *backward shift property*:

$$\underline{\Gamma_N^{\text{ob}}} \cdot A_{\text{ob}} = \overline{\Gamma_N^{\text{ob}}} \,, \tag{4}$$

where a 'bar' beneath (on top of) the matrix denotes a new matrix obtained by omitting its last (first) row. The word *backward shift* comes from the observation that the rows in $\overline{\Gamma_N^{\text{ob}}}$ are the same as those of $\underline{\Gamma_N^{\text{ob}}}$ shifted up (i.e. backward in time) one position. If $\Gamma_N^{\text{ob}}$ were known, (4) is an overdetermined but consistent set of linear equations in the unknown matrix $A_{\text{ob}}$, the solution of which is unique, provided that $\text{rank}(\underline{\Gamma_N^{\text{ob}}}) = n$, a condition that is called the *partial realisation condition*. Obviously, the model output vector $C_{\text{ob}}$ is the first row of $\Gamma_N^{\text{ob}}$. State space models are only unique up to within a non-singular similarity transformation $T$: $(A_{\text{ob}}, C_{\text{ob}}, \hat{x}_0^{\text{ob}}) \rightarrow (T^{-1}A_{\text{ob}}T, C_{\text{ob}}T, T^{-1}\hat{x}_0^{\text{ob}}) = (A, C, \hat{x}_0)$, which will still generate the same output data $\hat{y}_k$. This will also induce a change of basis for the column space: $\Gamma_N^{\text{ob}} \rightarrow \Gamma_N = \Gamma_N^{\text{ob}} \cdot T$. In the new basis, a modified set of linear equations for the system matrix follows from $\underline{\Gamma_N} \cdot A = \overline{\Gamma_N}$. This implies that backward shift-invariance is essentially a property of a subspace (the column space of $\Gamma_N$) and not of the specific choice of basis in that space. Summarising, backward shift invariance is characterised by $n = \text{rank}(\Gamma_N) = \text{rank}(\underline{\Gamma_N}) = \text{rank}(\underline{\Gamma_N} \;\; \overline{\Gamma_N})$. The first equality refers to observability, the second one is the *partial realisation condition* to uniquely identify the underlying system matrix $A$ and the third equality expresses the fact that $\mathbf{R}(\overline{\Gamma_N}) \subseteq \mathbf{R}(\underline{\Gamma_N})$, as follows from equation (4), with equality when $A$ is non-singular ($\mathbf{R}(.)$ denotes the range/column space). The fact that eigenvalues are invariants for a similarity transformations, seems to suggest that the shift-invariance of a subspace is solely determined by the eigenvalues and their multiplicity structure. We can show that this is true indeed. Using the forward shift operator $z$ as $z(y_k) = y_{k+1}$, we can write equation (1) as $(z^n + \alpha_1 z^{n-1} + \ldots + \alpha_{n-1}z + \alpha_n)y_k = p(z)y_k = 0$. Obviously, the

zeros of $p(z)$ are the eigenvalues of the companion matrix $A_{\rm ob}$, and therefore the eigenvalues of any transformed matrix $T^{-1}A_{\rm ob}T$. Assume that $\lambda$ is a root of $p(z) = 0$, then $p(\lambda) = a^T v$ where $a^T = (\alpha_n \ \alpha_{n-1} \ldots \alpha_1 \ 1)$ and $v^T = (1 \ \lambda \ \ldots \ \lambda^n)$ is a 'Vandermonde' vector. When the algebraic multiplicity of $\lambda$ is $\mu$, it is well known that $\lambda$ will also be a root of the 'derivative' polynomials $d^k p(z)/dz^k = 0, k = 1, \ldots, \mu - 1$. We then find that $d^k p(z)/dz^k(\lambda) = a^T(d^k v/d\lambda^k)$. As an example, take $n = 6$ with three eigenvalues $\lambda_1, \lambda_2, \lambda_3$ with multiplicities $3, 1, 2$. Then:

$$a^T \cdot V = (\alpha_6 \ \alpha_5 \ \alpha_4 \ \alpha_3 \ \alpha_2 \ \alpha_1 \ 1) \left(\begin{array}{ccc|c|cc} 1 & 0 & 0 & 1 & 1 & 0 \\ \lambda_1 & 1 & 0 & \lambda_2 & \lambda_3 & 1 \\ \lambda_1^2 & 2\lambda_1 & 1 & \lambda_2^2 & \lambda_3^2 & 2\lambda_3 \\ \lambda_1^3 & 3\lambda_1^2 & 3\lambda_1 & \lambda_2^3 & \lambda_3^3 & 3\lambda_3^2 \\ \lambda_1^4 & 4\lambda_1^3 & 6\lambda_1^2 & \lambda_2^4 & \lambda_3^4 & 4\lambda_3^3 \\ \lambda_1^5 & 5\lambda_1^4 & 10\lambda_1^3 & \lambda_2^5 & \lambda_3^5 & 5\lambda_3^4 \\ \lambda_1^6 & 6\lambda_1^5 & 15\lambda_1^4 & \lambda_2^6 & \lambda_3^6 & 6\lambda_3^5 \end{array}\right) = 0$$

$V$ is called a 'confluent Vandermonde' matrix as some of its columns are derivatives of Vandermonde vectors. We can now use $p(\lambda) = a^T v = 0$ to explicitly write out the eigenvalue decomposition of the companion matrix for this example as

$$\begin{aligned} A_{\rm ob} \cdot V &= \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -\alpha_6 & -\alpha_5 & -\alpha_4 & -\alpha_3 & -\alpha_2 & -\alpha_1 \end{pmatrix} V \\ &= V \cdot J = V \cdot \left(\begin{array}{ccc|c|cc} \lambda_1 & 1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & \lambda_2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \lambda_3 & 1 \\ 0 & 0 & 0 & 0 & 0 & \lambda_3 \end{array}\right), \end{aligned} \tag{5}$$

where $J$ is in Jordan canonical form with 3 Jordan blocks of dimensions given by the algebraic multiplicities. When we now in (3) change the state space basis using the confluent Vandermonde matrix $V$, as $(A_{\rm ob}, C_{\rm ob}, \hat{x}_0^{\rm ob}) \rightarrow (V^{-1}A_{\rm ob}V, C_{\rm ob}V, V^{-1}\hat{x}_0^{\rm ob}) = (J, C_v, \hat{x}_0^v)$, the corresponding observability

matrix will transform as $\Gamma_N \to (\Gamma_N \cdot V) = \Gamma_N^v$, where $C_v = (1\ 0\ 0\ 1\ 1\ 0)$. The extended observability matrix $\Gamma_N^v$ will now be an 'extended' confluent Vandermonde matrix (i.e. a rectangular extension of the square confluent Vandermonde matrix $V$). This demonstrates that a backward shift invariant subspace is solely determined by the roots, and their multiplicities, of the difference equation of the underlying linear model (1).

### 4.2. Shift-invariant spaces in operator theory

It is straightforward to derive that $T_a \cdot \Gamma_N = 0$. Since $T_a$ is of full row and $\Gamma_N$ of full column rank, this implies:

$$\mathbf{R}(T_a^T) \oplus \mathbf{R}(\Gamma_N) = \mathbb{R}^N \ , \ \mathbf{R}(T_a^T) \perp \mathbf{R}(\Gamma_N) \ .$$

The ranges are perpendicular to one another and are complementary subspaces. We have characterised the range of $\Gamma_N$ to be backward shift-invariant, which is a finite dimensional translation of similar characterisations of backward and forward shift invariance in infinite dimensional spaces,[2] which have been intensively studied in operator theory [23] [38] [44] [45] [72] [75]. So far we assumed that there was only one single shift $\lambda$, and that the state space model in (2) was single-output. In what follows, we will generalise these single-shift scalar (= single-output) shift-invariant spaces towards multiple-output multi-shift-invariant subspaces, which in the infinite dimensional setting have been studied much less (see e.g. [11] [12] [13]).

## 5. Minimisation of a least squares misfit objective

### 5.1. Minimising the misfit

We now start from given, observed data $y = (y_0\ y_1\ \ldots\ y_{N-1})^T \in \mathbb{R}^N$, which are not exact, i.e. they do not belong to the behaviour of the models

[2]The notion of shift is ubiquitous in mathematics. For a function with Taylor series expansion $f(z) = a_0 + a_1 z + a_2 z^2 + \ldots$, the forward shift operator S can be characterised as $Sf(z) = zf(z) = a_0 z + a_1 z^2 + a_2 z^3 + \ldots$, and its backward version as $S^* f(z) = (f(z) - f(0))/z = a_1 + a_2 z + a_3 z^2 + \ldots$. In system theory, for a state sequence $x(k)$ with $z$-transform $X(z)$, we have for the forward shift in time, $q(x(k)) = x(k+1)$ as $z$-transform $Z(x(k+1)) = zX(z)$, so shifting forward in time correspond to a multiplication with the shift in the $z$-domain. In harmonic analysis, a shift in time $T^t f(x) = f(x+t)$ generates a multiplication by $\exp(itx)$ with its Fourier transform.

we discussed in Section 3. We will modify them additively by a misfit vector[3] $\tilde{y} \in \mathbb{R}^N$ so as to obtain the exact data $\hat{y} \in \mathbb{R}^N$: $y = \hat{y} + \tilde{y}$, where $\hat{y}$ belongs to the behaviour of the models of Section 3. For a pre-specified order $n$ in (1), we want to find the optimal parameters $\alpha_i \in \mathbb{R}, i = 1, \ldots, n$, that minimise the 2-norm of the misfit vector $\tilde{y}$.

$$\min_{\alpha_i \in \mathbb{R}\ ,\ i=1,\ldots,n} \sigma^2 = \|\tilde{y}\|_2^2 = \|y - \hat{y}\|_2^2 = \sum_{k=0}^{N-1} (y_k - \hat{y}_k)^2 \ , \tag{6}$$

where the exact data $\hat{y}_k$ satisfy the difference equation (1). We will show that the solution can be interpreted as a separable least squares problem: In a first step, one could consider the model coefficients $\alpha_i$ to be known, and then obtain the optimal misfit vector $\tilde{y}$ by an orthogonal projection so that the 'exact' data $\hat{y} = y - \tilde{y}$ belong to the behaviour of the model generated by the $\alpha_i$. A second step then consists of a non-linear optimisation over the $\alpha_i$, which turns out to be multivariate polynomial, and therefore, as we will show, is in essence an MEVP. We will start from the kernel representation, which immediately leads to $T_a \cdot y = T_a \cdot \hat{y} + T_a \cdot \tilde{y} = T_a \cdot \tilde{y}$. If $T_a$ were known, this would be an underdetermined set of linear equations in the unknown misfit $\tilde{y}$. The unique minimum norm solution follows from the pseudo-inverse of $T_a$:

$$\tilde{y} = T_a^\dagger T_a y = T_a^T (T_a T_a^T)^{-1} T_a y \ . \tag{7}$$

The second equality follows from the fact that $T_a$ is of full row rank, so that $T_a T_a^T$ is nonsingular and $T_a^\dagger = T_a^T (T_a T_a^T)^{-1}$. The matrix $\Pi_a = T_a^T (T_a T_a^T)^{-1} T_a$ is the orthogonal projector onto the row space of $T_a$. If $T_a$ were known, the given data vector $y$ could be decomposed into two mutually orthogonal vectors as $y = \hat{y} + \tilde{y} = (I_N - \Pi_a) y + \Pi_a y$ (see also Section 9), where $\hat{y}$ belongs to the behaviour of model (1).

### 5.2. Secular equations

The least squares objective function (6) can now be written as

$$\sigma^2 = \|\tilde{y}\|_2^2 = \|\Pi_a y\|_2^2 = y^T T_a^T (T_a T_a^T)^{-1} T_a y \ , \tag{8}$$

[3]The misfit $\tilde{y}$ could be called 'noise' (e.g. as in 'measurement noise'), suggesting inaccuracies that corrupt an otherwise clean exact signal. Often, a priori unverifiable additional assumptions (Kalman's 'prejudices') are invoked, e.g. statistical ones like Gaussianity, or whiteness, to 'model' these inaccuracies. In our framework however, the misfit is simply the correction needed to orthogonally project the given data $y$ onto the behaviour of the model, without any further unverifiable priors.

which is to be minimised over the coefficients $\alpha_i, i = 1, \ldots, n$. Define the matrix $D_a = T_a T_a^T$, which itself is a symmetric, positive definite, banded Toeplitz matrix. The first order optimality conditions are:

$$(9) \quad \frac{\partial \sigma^2}{\partial \alpha_i} = 0 = 2y^T T_a^T D_a^{-1} T_a^{\alpha_i} y - y^T T_a^T D_a^{-1} D_a^{\alpha_i} D_a^{-1} T_a y \ , \ \forall i = 1, \ldots, n,$$

where a superscript $\alpha_i$ denotes the partial derivative with respect to $\alpha_i$. We have used the fact that $\partial D_a^{-1} / \partial \alpha_i = -D_a^{-1} D_a^{\alpha_i} D_a^{-1}$. We will call equations (9) the $n$ *secular equations*. The observation that they are 'nonlinear' in the coefficients $\alpha_i$, has led to a lot of heuristic algorithms in the past (see Section 10). However, as $D_a^{-1} = \text{adj}(D_a)/\det(D_a)$, where $\text{adj}(D_a)$ is the adjugate of the matrix $D_a$ (the transpose of the matrix with the cofactors of all elements of $D_a$), and because $\det(D_a) \neq 0$, these $n$ equations (9) are equivalent to $n$ multivariate polynomials in the $n$ unknowns $\alpha_i$, after 'multiplying out' $\det(D_a)$. Their common roots are all global and local minima and maxima and saddle points of the objective function (8). So finding the critical points of the least squares realisation objective function is an exercise in rooting multivariate polynomials. The relation between common roots of sets of multivariate polynomials on the one hand and the matrix EVP on the other hand, is (not so) well known (see e.g. [24] [78]). In [40] we have developed a framework to find all common roots of a set of multivariate polynomials, based on Macaulay matrices, their null space being backward multi-shift invariant, and multi-dimensional realisation theory, that takes full advantage of powerful algorithms for the singular value and eigenvalue decomposition. We could follow that framework to directly solve the set of multivariate polynomials implicit in (9), but that would require the calculation of determinants for $\text{adj}(D_a)$, which we want to avoid. Instead, we will rewrite (9) as an MEVP, that can also be tackled with insights from multi-dimensional realisation theory (see also [36]).

### 5.3. Multi-parameter eigenvalue problem (MEVP)

Define the vector $f = D_a^{-1} T_a \, y \in \mathbb{R}^{N-n}$ and rewrite (8) as

$$(10) \qquad \begin{pmatrix} D_a & T_a y \\ y^T T_a^T & \sigma^2 \end{pmatrix} \begin{pmatrix} f \\ -1 \end{pmatrix} = 0 \ .$$

Taking partial derivatives with respect to all variables $\alpha_i, i = 1, \ldots, n$, and using the derivative chain rule, results in, $\forall i = 1, \ldots, n$:

$$(11) \quad \begin{pmatrix} D_a^{\alpha_i} & T_a^{\alpha_i} y \\ y^T (T_a^{\alpha_i})^T & 0 \end{pmatrix} \begin{pmatrix} f \\ -1 \end{pmatrix} + \begin{pmatrix} D_a & T_a y \\ y^T T_a^T & \sigma^2 \end{pmatrix} \begin{pmatrix} f^{\alpha_i} \\ 0 \end{pmatrix} = 0 \ .$$

Eqs. (10) and (11) contain $(n+1)(N-n+1)$ equations, and the number of unknowns is the same (namely $(N-n)$ in $f$, 1 in $\sigma$, $n(N-n)$ in all $f^{\alpha_i}$ and $n$ for all of the $\alpha_i$). Three observations can be made: **1.** The unknown vectors $f$ and $f^{\alpha_i}, i = 1, \ldots, n$ appear linearly in the equations. The variables $\alpha_i$ appear linearly in $T_a$ and quadratically in $D_a$. The matrices $T_a^{\alpha_i}$ are constant matrices. **2.** The last equation in (10) is the only one involving $\sigma$, because the last component of the last vector in (11) is 0. In what follows, we will omit the variable $\sigma$, which is the value of the objective function and the corresponding equation (but come back to this in Section 10). **3.** One can easily recover the secular eqs. (9) from (11) by eliminating $f^{\alpha_i}$ from the first block rows in (11), plugging it in into the second block row and using $f = D_a^{-1} T_a y$ from (10). Let's verify that these equations are equivalent to (9). We leave out the equation for $\sigma^2$. From (11), we find $D_a^{\alpha_i} f - T_a^{\alpha_i} y + D_a f^{\alpha_i} = 0$ and $y^T (T_a^{\alpha_i})^T f + y^T T_a^T f^{\alpha_i} = 0$. Hence $f^{\alpha_i} = -D_a^{-1} D_a^{\alpha_i} f + D_a^{-1} T_a^{\alpha_i} y$. Using from (10) $f = D_a^{-1} T_a y$, we then find $y^T (T_a^{\alpha_i})^T D_a^{-1} T_a y + y^T T_a^T D_a^{-1} T_a^{\alpha_i} y - y^T T_a^T D_a^{-1} D_a^{\alpha_i} D_a^{-1} T_a y, i = 1, \ldots, n$, which is exactly the same as (9). Leaving out the equation for $\sigma^2$, we can also regroup equations (10) and (11) as

$$\begin{pmatrix} D_a^{\alpha_i} & D_a & T_a^{\alpha_i} y \\ D_a & 0 & T_a y \\ y^T ((T_a^{\alpha_i})^T & y^T T_a^T & 0 \end{pmatrix} \begin{pmatrix} f \\ f^{\alpha_i} \\ -1 \end{pmatrix} = 0 \ , \ i = 1, \ldots, n. \tag{12}$$

In this form, there is a nice symmetry, but the equations $D_a f = T_a y$ are repeated $n$ times, so there is redundancy here. However, this form clearly shows that we have to find the coefficients $\alpha_i, i = 1, \ldots, n$ so that the square matrices between brackets are singular: Their determinants should be zero, and using some well known lemma's for the determinants of block matrices, we would again find back the secular equations (9). This separation in variables that appear linearly, and others that appear polynomially, is reminiscent of a similar separation property for the algebraic EVP.[4] We now

[4]For $A \in \mathbb{R}^{n \times n}$, both formulations $Ax = x\lambda, x \neq 0$, and $\det(\lambda I_n - A) = 0$ are equivalent. The fact that the eigenvectors $x$ appear linearly in the former allows one to write $(\lambda I_n - A)x = 0$, which only has a nontrivial solution $x \neq 0$ if and only if the characteristic polynomial $\chi(A) = \det(\lambda I_n - A) = 0$. This effectively separates out the unknown elements of $x$ from the unknown variables $\lambda$. Therefore, the secular equations (9) are the generalisations to $n$ variables of the characteristic equation in 1 variable, just like the MEVP is the generalisation to $n$ variables of the 1 parameter algebraic EVP.

combine (10)–(12) to find

(13)
$$
\begin{array}{c}
\\
N-n \\
N-n \\
N-n \\
\vdots \\
N-n \\
1 \\
\vdots \\
1
\end{array}
\begin{array}{c}
\begin{array}{cccccc} N-n & N-n & N-n & \dots & N-n & 1 \end{array} \\
\begin{pmatrix}
D_a & 0 & 0 & \dots & \dots & T_a y \\
D_a^{\alpha_1} & D_a & 0 & \dots & 0 & T_a^{\alpha_1} y \\
D_a^{\alpha_2} & 0 & D_a & \dots & 0 & T_a^{\alpha_2} y \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
D_a^{\alpha_n} & 0 & \dots & \dots & D_a & T_a^{\alpha_n} y \\
y^T (T_a^{\alpha_1})^T & y^T T_a^T & 0 & \dots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
y^T (T_a^{\alpha_n})^T & 0 & 0 & \dots & y^T T_a^T & 0
\end{pmatrix}
\end{array}
\begin{pmatrix} f \\ f^{\alpha_1} \\ f^{\alpha_2} \\ \vdots \\ f^{\alpha_n} \\ -1 \end{pmatrix} = 0 \, .
$$

The dimensions of the matrix are $((N-n)(n+1)+n) \times ((N-n)(n+1)+1)$, so it has $n-1$ more rows than columns. It is a function of the given data $y$ and the unknown coefficients $\alpha_i$, which appear quadratically in $D_a$ and linearly in $D_a^{\alpha_i}$ and $T_a$. We can collect the coefficients of all monomials[5] $1, \alpha_1, \alpha_2, \dots, \alpha_n, \alpha_1^2, \alpha_1\alpha_2, \dots, \alpha_{n-1,n}, \alpha_n^2$ in matrices $A_{(i_1,i_2,\dots,i_n)}$, where $(i_1, i_2, \dots, i_n) \in \mathbb{N}^n$ is a multi-index. To give an example, for $n = 2$, we can write (13) as

$$
(A_{00} + A_{10}\alpha_1 + A_{01}\alpha_2 + A_{20}\alpha_1^2 + A_{11}\alpha_1\alpha_2 + A_{02}\alpha_2^2) \cdot w = 0 \, , \tag{14}
$$

with $w = (f^T \ (f^{\alpha_1})^T \ (f^{\alpha_2})^T \ -1)^T$ and $A_{ij} \in \mathbb{R}^{(3N-4)\times(3N-5)}$, $i = 0, 1, 2$. The pair $(\alpha_1, \alpha_2)$ is called an eigenvalue-pair of the MEVP and $z$ the corresponding eigenvector. For general $n$, the $n$-tuple $(\alpha_1, \dots, \alpha_n)$ is called an eigen-tuple. In what follows, we will show that finding all solution of MEVPs as in (13) proceeds in several steps: Enlarge the number of equations using *Forward multi-Shift Recursions* (FmSRs), calculate the null space of the resulting block Macaulay matrix, which will be block backward multi-shift invariant, and then exploit that property to set up a multi-dimensional realisation problem that will result in $n$ commuting system matrices $A_i$, the eigenvalues of which will by the eigenvalues $\alpha_i$ in the eigen-tuples.

[5] As for a monomial ordering, we use the *degree negative lexicographic* ordering, by way of example shown here for 3 variables $\alpha, \beta, \gamma$: $1 < \alpha < \beta < \gamma < \alpha^2 < \alpha\beta < \alpha\gamma < \beta^2 < \beta\gamma < \gamma^2 < \alpha^3 < \alpha^2\beta < \alpha^2\gamma < \alpha\beta^2 < \alpha\beta\gamma < \beta^3 < \beta^2\gamma < \beta\gamma^2 < \gamma^3 < \alpha^4 < \dots$.

## 6. Solving multi-parameter eigenvalue problems (MEVP)

Contrary to one-parameter EVPs (including the Jordan, Weierstrass and Kronecker canonical forms, or EVPs polynomial or rational in one variable), MEVPs have been studied much less intensively in the literature. Early references include [7] [8] [9] [22] [87]. More recent papers include [53] [54] [70] [71] [73] [74]. If algorithms are discussed at all, the MEVP is often just considered to be a nonlinear optimisation problem, looking for one or some, but not all of the solutions, without exploiting its inherent structure that we will reveal here. Indeed, there does not seem to exist a general, unifying theoretical framework to tackle MEVPs. The approach we will outline here, is a happy symbiosys between the *Fundamental Theorem of Linear Algebra*, and the *Fundamental Theorem of Algebra*. The fundamental theorem of linear algebra describes the 4 fundamental subspaces of a matrix: (column, row, left and right null spaces) with their orthogonality and dimensional rank-based properties, as revealed by the SVD [79] [80]. The fundamental theorem of algebra states that a univariate polynomial of degree n with complex coefficients has n complex roots, counting multiplicities (the field of the complex numbers is algebraically closed) [14] [37] [66] [81]. For real coefficients, the roots are symmetric with respect to the real axis. Hilbert's Nullstellensatz [24] [25] is its generalisation to sets of polynomial equations. In our forthcoming paper [36], we elaborate extensively on this unifying framework to tackle MEVPs (see also [29]), the ingredients of which stem for linear algebra (FmSRs, SVD, EVP), algebraic geometry (Hilbert's Nullstellensatz), operator theory (forward and backward (multi-)shift invariant spaces) and system theory (multi-dimensional realisation theory). In what follows, we will only summarise the main insights based on the following table:

| **Fund. Thm.** | **EVP** | **MEVP** |
|---|---|---|
| **Lin. Alg.** | **Section 3**<br>Forward Shift Recursions<br>Row space banded Toeplitz | **Subsection 6.1**<br>Forward multi-Shift Recursions<br>Row space block Macaulay |
| **Algebra** | **Section 4**<br>Scalar backward<br>shift-invariant null space<br>1D observability matrix<br>1D realisation<br>eigenvalues | **Subsection 6.2 6.3 6.4**<br>Block backward<br>multi-shift-invariant null space<br>nD observability matrix<br>multi-dimensional realisation<br>eigen-tuples |

### 6.1. Forward multi-Shift Recursions (FmSR)

For the sake of clarity of exposition, we take the case $n = 2$. Generalisations for $n > 2$ are straightforward. First, we 'enlarge' the MEVP (14) by FmSRs with all monomials in $\alpha_1, \alpha_2$ of increasing degree:

$$\begin{pmatrix} A_{00} & A_{10} & A_{01} & A_{20} & A_{11} & A_{02} & 0 & \dots \\ 0 & A_{00} & 0 & A_{10} & A_{01} & 0 & A_{20} & \dots \\ 0 & 0 & A_{00} & 0 & A_{10} & A_{01} & 0 & \dots \\ 0 & 0 & 0 & A_{00} & 0 & 0 & A_{10} & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} w \\ w \cdot \alpha_1 \\ w \cdot \alpha_2 \\ w \cdot \alpha_1^2 \\ w \cdot \alpha_1\alpha_2 \\ w \cdot \alpha_2^2 \\ w \cdot \alpha_1^3 \\ \vdots \end{pmatrix} = 0 \, .$$

The matrix to the left is called a block Macaulay matrix. Observe how FmSRs force the vector(s) in its null space to adopt a generalised Vandermonde structure, generalised for two reasons: $w$ is a vector, and, instead of having one parameter, we now have 2 of them (or $n$ in the general case). The first rows of these vectors in the null space, corresponding to the vectors $w$, will be called the *degree* 0 *block*. The next ones, comprising the rows with $w.\alpha_1$ and $w.\alpha_2$ constitute the *degree* 1 *block*, etc...

### 6.2. Multi-output backward multi-shift invariant null space

Although the rank of the block Macaulay matrix keeps increasing as a function of the FmSRs, one can show that its nullity $n_M$ will stabilise whenever the eigen-tuples are isolated. Said in other words, in that case the set of eigen-tuples is zero dimensional. It will count the total number of roots in projective space, $n_M = n_M^a + n_M^\infty$, where $n_M^a$ is the number of affine ones and $n_M^\infty$ the number at infinity.[6] Next, one can show (see e.g. [36] [39] [40]) that this null space can be modelled as the observability matrix of a 2D-shift-invariant system (in the general case an $n$-dimensional shift-invariant system), generated by a 2-dimensional singular system:

$$x^R_{k+1,l} = A_1 x^R_{k,l} \quad , \quad x^S_{k-1,l} = E_1 x^S_{k,l} \, ,$$

[6] In projective space, for the roots of multivariate polynomials, zeros at infinity originate in much the same way as, when in 2 dimensions, 2 parallel lines intersect in the point at infinity.

$$
\begin{aligned}
& x^R_{k,l+1} = A_2 x^R_{k,l} \ , \ \ x^S_{k,l-1} = E_2 x^S_{k,l} \ , \\
& y_{k,l} = C_R x^R_{k,l} + C_S x^S_{k,l} \ .
\end{aligned} \tag{15}
$$

Here, $x^R_{k,l} \in \mathbb{R}^{n^a_M}$ is the regular part of the state, governed by two discrete indices $k, l \in \mathbb{N}^+_0$, where $n^a_M$ is the number of affine (finite) eigen-pairs. The 2D-grid state propagation for increasing $k$ is modelled by $A_1$, while that over increasing $l$ is governed by $A_2$. Together they generate the dynamics of the regular state $x^R_{k,l}$ as it moves causally over a 2D discrete grid, starting in its origin. Here, $A_1, A_2 \in \mathbb{R}^{n^a_M \times n^a_M}$ commute [10]: $A_1 A_2 = A_2 A_1$ (Intuitively, they should, as $x^R_{k+1,l+1}$ can be reached from $x^R_{k,l}$ in 2 different ways as $x^R_{k+1,l+1} = A_1 x^R_{k,l+1} = A_1 A_2 x^R_{k,l} = A_2 x^R_{k+1,l} = A_2 A_1 x^R_{k,l}$, which should hold for arbitrary $x^R_{k,l}$). The singular part of the state is $x^S_{k,l} \in \mathbb{R}^{n^\infty_M}$, which propagates backward – anti-causally- both in $k$ and $l$ via the matrices $E_1, E_2 \in \mathbb{R}^{n^\infty_M \times n^\infty_M}$, where $n^\infty_M$ is the number of eigen-pairs at infinity. The sum $n^a_M + n^\infty_M = n_M$ is the nullity of the Macaulay matrix that stabilises from a certain FmSR on. The matrices $E_1$ and $E_2$ also commute: $E_1 E_2 = E_2 E_1$ and in addition, $E_1$ and $E_2$ are nilpotent, i.e. when powered up, from a certain power on, called the nilpotency index, we get a zero matrix.

So two of the state equations are causal, they describe how the regular part of the state propagates 'forward' on a 2D discrete grid starting from a regular initial state 'in the past'. Two of them are anti-causal and describe how the singular part of the state propagates 'backward' on a 2D discrete grid, starting from a collection of initial states 'in the future'. The output is then a sum, via two output matrices $C_R$ and $C_S$, of the regular and the singular part of the state. Generalisations of such state space models to mD commutative systems with $m > 2$ are straightforward.[7]

Let us now discuss the block multi-shift invariant structure of the null space of the block Macaulay matrix, still for $n = 2$. We will depict the situation when there have already been a sufficient number of FmSRs, so that not only the nullity has stabilised, revealing the total number of isolated solutions, but that also the threefold structure in the null space that we explain below, appears. We use a special choice of basis in the null space (which always exists, see below), so that the threefold structure is clearly visualised in the following mixed causal – anti-causal nD observability matrix $\Gamma$ of full column rank $n_M = n^a_M + n^\infty_M$. This matrix $\Gamma$ is a concatenation of $\Gamma_R$,

[7] These mD singular systems are a generalisation of the 1D case described e.g. in [69], and one might conjecture that there exists a Weierstrass Canonical Form for 2D commutative systems (or even for mD, when we take the general case), which seems plausible but definitely is an open problem.

which is the observability matrix of the regular part of the state space model (15) of order $n_M^a$ and has $n_M^a$ columns, and $\Gamma_S$, which is the observability matrix of the singular part of (15) of order $n_M^\infty$ with $n_M^\infty$ columns.

$$\Gamma = \begin{array}{c} \begin{array}{cc} n_M^a & n_M^\infty \end{array} \\ \left( \begin{array}{cc} \Gamma_R & \Gamma_S \end{array} \right) \end{array} \tag{16}$$

$$\left( \begin{array}{cc}
C_R & 0 \\ \hline
C_R A_1 & 0 \\
C_R A_2 & 0 \\ \hline
C_R A_1^2 & 0 \\
C_R A_1 A_2 & 0 \\
C_R A_2^2 & 0 \\ \hline
\vdots & \vdots \\ \hline
C_R A_1^{\delta_g - 1} & 0 \\
C_R A_1^{\delta_g - 2} A_2 & 0 \\
\vdots & \vdots \\
C_R A_2^{\delta_g - 1} & 0 \\ \hline\hline
C_R A_1^{\delta_g} & 0 \\
\vdots & \vdots \\
C_R A_2^{\delta_g} & 0 \\ \hline
\vdots & \vdots \\ \hline\hline
C_R A_1^{\delta_s} & r_{\delta_s,0}^T \\
\vdots & \\
C_R A_2^{\delta_s} & r_{0,\delta_s}^T \\ \hline
C_R A_1^{\delta_s+1} & r_{\delta_s+1,0}^T \\
\vdots & \\
C_R A_2^{\delta_s+1} & r_{0,\delta_s+1}^T \\ \hline
C_R A_1^{\delta_s+2} & r_{0,\delta_s+2}^T \\
\vdots & \vdots \\
\vdots & \\
\vdots & r_{1,\delta_s+1}^T \\
C_R A_2^{\delta_s+2} & r_{0,\delta_s+2}^T
\end{array} \right) .$$

There are three degree block rows in $\Gamma$, separated by double lines:

**The regular (affine) zone:** The column space of the first zone is only generated by $\Gamma_R$. All corresponding rows in $\Gamma_S$ are zero. In this zone, when checking the linear independency of rows of $\Gamma_R$ from the top downwards, the rank increases at least by 1 for every degree block that is added. One could call the indices of the linear independent rows the *observability indices*.

**The mind-the-gap-zone:** The second zone starts when, for a certain degree $\delta_g$ (subscript 'g' for 'gap'), there is a degree block, with all of its rows linear dependent on rows in the previous degree blocks. If the FmSRs are progressed sufficiently far, there might be several of such degree blocks that are completely linear dependent on the previous ones. When checking the degree blocks from the top downwards, the rank has now stabilised to $n_M^a$, which is the number of affine eigen-tuples. One can see that the mind-the-gap zone separates the regular zone from the singular zone. This observation will allow us to deflate the eigen-tuples at infinity, with a column compression.

**The singular or 'A-bout-de-souffle'-zone:** From a certain degree $\delta_s$ (subscript 's' for 'singular'), the rank per degree block starts increasing again at least by one per degree block, until we reach a degree block where the total cumulative rank equals $n_M$.

The null space of the block Macaulay matrix is seen to be the union of the column space of a regular, causal observability matrix $\Gamma_R$, with increasing powers of the system matrices $A_i$, and a singular, anti-causal one, $\Gamma_S$, the structure of which we will explain a bit more in detail. Let's assume, as in Eq. (16), just by way of example, for $n = 2$, that at the degree $\delta_s$ block the rank starts increasing again after the mind-the-gap zone, and that it keeps increasing for two more degree blocks up to the degree $(\delta_s + 2)$ block. We represent the rows in the highest degree block of $\Gamma_S$ by row vectors of the form $r^T(i, j)$, where $i$ and $j$ are the degrees of the corresponding monomials. For the degree $(\delta_s + 2)$ block, $i + j = \delta_s + 2$, for the degree $(\delta_s + 1)$ block, $i + j = \delta_s + 1$ etc.... Just as in the regular case, there are relations between the rows of the degree blocks, with powers of the two $n_M^\infty \times n_M^\infty$ commuting matrices $E_1$ and $E_2$ of the singular part of the state space model:

$$\begin{aligned} r^T(i,j) &= r^T(i+1,j)E_1 = r^T(i,j+1)E_2 \\ &= r^T(i+2,j)E_1^2 = r^T(i+1,j+1)E_2E_1 \\ &= r^T(i+1,j+1)E_1E_2 = r^T(i,j+2)E_2^2 \,. \end{aligned}$$

These relations also induce algebraic constraints between the initial conditions at infinity (in our example the rows in the degree $(\delta_s + 2)$ block). In addition, there can be complicated multiplicity structures at infinity, as revealed by the Jordan forms of the singular matrices $E_i$ (which we will not elaborate on here). So we clearly see that the singular part $\Gamma_S$ of $\Gamma$, grows anti-causally with increasing powers of the singular matrices $E_i$, starting from the bottom upwards. All of the matrices $E_i$ are nilpotent, so after some degree starting from the bottom upwards, all rows in the degree blocks are zero. They run 'out-of-breath' ('a-bout-de-souffle'). In the visualisation above, there are 3 degree blocks over which the non-zero rows of $\Gamma_S$ develop upwards, so in this case the nilpotency index is 3. Such a singular behaviour only occurs when there are eigen-tuples at infinity (in the example there would be 3 zeros at infinity).

In this paper, however, there is no need to explore the fine structure of the zeros at infinity in more detail, as we will exploit the 'mind-the-gap'-zone to deflate them out. For more details about the behaviour at infinity, we refer to [36] where we will deploy 'singular perturbation' approaches to provide more insight.

### 6.3. Column compression for the affine zeros

When calculating the null space of the block Macaulay matrix, e.g. with an SVD, the basis of the null space will not reveal directly the observability structure as in (16). Instead, the numerical result will be a linear combination of the columns of the multi-dimensional observability matrix (16), mixing up the columns of $\Gamma_R$ and $\Gamma_S$. However, these linear combinations of columns do not modify the linear (in)-dependency of the rows of the null space matrix, when checking linear (in-)dependency from the top downwards. Said in other words, the observability indices are invariants, independent of the choice of basis for the null space. This observation will allow us to deflate the roots at infinity as explained in Figure 1.

### 6.4. Multi-dimensional realisation in the compressed null space

The column space of the regular observability matrix $\Gamma_R$ is a *block backward multi-shift-invariant subspace*. Denote by $\Gamma_1$ the sub-matrix of $\Gamma_R$ that contains the first degree blocks up to and including the degree $\delta_g - 1$ block (so all degree blocks of the regular zone). One can now verify that (recall that

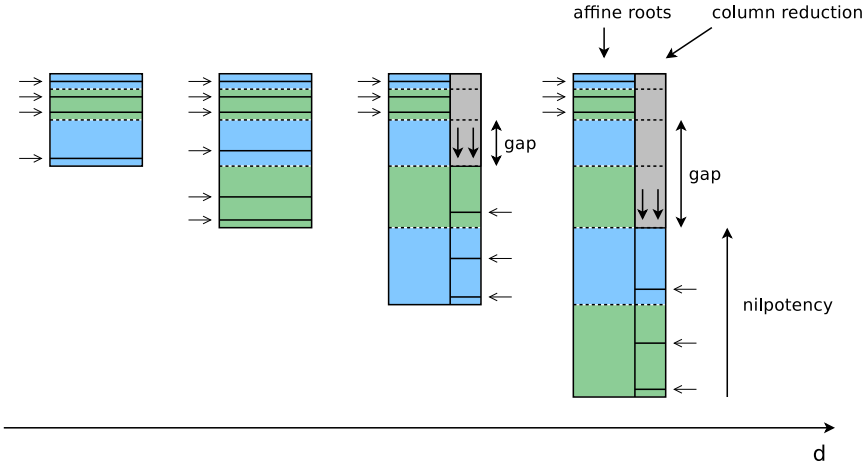


Figure 1: The figure shows the null space (in some arbitrary numerical basis), for increasing FmSRs represented by the long horizontal arrow for increasing block degrees $d$. As the recursion proceeds, there will be linear independent rows in certain degree blocks, indicated by arrows, when we start checking linear (in-)dependency of rows from top to bottom. One does not have to check row by row, but it can be verified degree block per degree block using SVDs (only the singular values are needed), whether the rank increases or not. When the set of eigen-tuples is zero dimensional, from a certain recursion on, there will be a complete degree block, where all of its rows will be linear dependent on previous ones. This indicates that we have found 'the-mind-the-gap'-zone. It is possible that below that gap downwards, there are still some rows that are linear independent, 'caused' by the eigen-tuples at infinity. In any case, in the null space, the sub-matrix formed with the degree blocks above the mind-the-gap zone has rank $n_M^a$. When $n_M^a = n_M$, the nullity of the Macaulay matrix, nothing special needs to be done as there are no roots at infinity in that case. When $n_M^a < n_M$, one has to deflate the eigen-tuples at infinity, which can be done by a so-called column compression (with an SVD) to the whole null space matrix, applied to the right. Then, all block rows, including those of the mind-the-gap zone, of the $n_M^a$ columns to the left, will exhibit the structure of a regular nD observability matrix as in (16), up to within a similarity transformation.

$A_1$ and $A_2$ commute):

$$\Gamma_1 A_1 = \left(\begin{array}{c} C_R \\ \hline C_R A_1 \\ C_R A_2 \\ \hline C_R A_1^2 \\ C_R A_1 A_2 \\ C_R A_2^2 \\ \hline \vdots \\ \hline C_R A_1^{\delta_g - 1} \\ C_R A_1^{\delta_g - 2} A_2 \\ \vdots \\ C_R A_2^{\delta_g - 1} \end{array}\right) A_1 = \left(\begin{array}{c} C_R A_1 \\ \hline C_R A_1^2 \\ C_R A_1 A_2 \\ \hline C_R A_1^3 \\ C_R A_1^2 A_2 \\ C_R A_1 A_2^2 \\ \hline \vdots \\ \hline C_R A_1^{\delta_g} \\ C_R A_1^{\delta_g - 1} A_2 \\ \vdots \\ C_R A_1 A_2^{\delta_g - 1} \end{array}\right) = S_1 \cdot \Gamma_R \tag{17}$$

where the selector matrix $S_1$ selects the appropriate block rows of $\Gamma_R$, that are 'hit' by the shift matrix $A_1$. A similar observation holds for the 'shift' matrix $A_2$ with $\Gamma_1 \cdot A_2 = S_2 \cdot \Gamma_R$, where the selector matrix $S_2$ selects the appropriate block rows of $\Gamma_R$ 'hit' by the shift matrix $A_2$. By construction, $\Gamma_1$ is of full column rank, so that we can now find $A_1$ and $A_2$ by exploiting the multi-shift-invariance property (17), which is independent of the specific choice of basis in the null space, using pseudo-inverses:

$$A_1 = \Gamma_1^{\dagger}(S_1 \Gamma_R) \text{ and } A_2 = \Gamma_1^{\dagger}(S_2 \Gamma_R) .$$

The eigenvalue pairs $(\alpha_1, \alpha_2)$ follow from the eigenvalues $\alpha_1$ of $A_1$ and $\alpha_2$ of $A_2$, and have to be matched with each other (details omitted here).

For the sake of clarity, we did the case $n = 2$, but the generalisation to the cases $n > 2$ is straightforward: Instead of having two commuting matrices $A_1, A_2$ and $E_1, E_2$ in the state space model (15), in the general case one has sets of pairwise commuting matrices $\{A_i, i = 1, \ldots, n\}$ and $\{E_i, i = 1, \ldots, n\}$, but basically all observations made above carry through.

Some final observations, which we do not treat in detail here, are: *1. Eigen-tuples with multiplicity larger than 1,* in which case we have to deal with generalised' confluent Vandermonde matrices, generalised because of the degree block structure, and because we have now multi-shifts, instead of single-shift (see e.g. [26] [39] [40]). *2. Generalised Cayley-Hamilton:* The fact that there is a mind-the-gap zone, implies that products of higher powers of the matrices $A_i$ can be written as linear combinations of products of lower power ones. This is a generalisation of the Theorem of Cayley-Hamilton to

multiple commuting matrices, but we will refer for this to some forthcoming publication. *3. The secular equations* (9) will only contain the affine solutions, not the ones at infinity. This is comparable to the algebraic generalised EVP of the form $Ax = Bx\lambda$ for $A, B \in \mathbb{R}^{n\times n}$. When $B$ is singular, there will be roots $\lambda$ at infinity, however in that case the characteristic equation, $\det(A - B\lambda) = 0$, will have a lower degree than $n$, because its roots do not contain the ones at infinity. *4. Non-zero dimensional varieties:* It may happen that the affine variety, corresponding to the finite eigen-tuples, is zero-dimensional, but that the variety of the eigen-tuples at infinity is non-zero-dimensional. In this case, the nullity will not stabilise, but the 'affine nullity' $n_M^a$ will, so that we can still do a column compression to deflate the roots at infinity. *5. Variations on the basic outline:* There are many variations on the basic outline presented here and numerical implementations that remain to be developed, one of which can already be found in [85]. Fast algorithms that exploit the 'quasi-Toeplitz'-structure of the block Macaulay matrices, and their sparsity, need to be developed as well (see e.g. [15]).

## 7. Image representation with simple poles

So far, we have been using the kernel representation of the autonomous LTI systems, but we can obtain complementary results by using an image presentation with the extended observability matrix in a state space basis that reveals it as an extended Vandermonde matrix. For reasons of conciseness and simplicity, we will assume that all roots of (1) are simple. The exact data can now be modelled as $\hat{y}_k = \sum_{i=1}^{n} \lambda_i^{k-1}\xi_i$, where $\lambda_i$ are the roots (i.e. the poles of the autonomous LTI system) and $\xi_i$ the components of the initial state. With obvious definitions for $\Lambda_N^v$ and $\hat{x}_0^v$, we can write ('v' stands for 'Vandermonde'):

$$\hat{y} = \Lambda_N^v \cdot \hat{x}_0^v = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \lambda_1 & \lambda_2 & \dots & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \dots & \lambda_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1^{N-1} & \lambda_2^{N-1} & \dots & \lambda_n^{N-1} \end{pmatrix} \cdot \begin{pmatrix} \xi_1 \\ \xi_2 \\ \dots \\ \xi_n \end{pmatrix}.$$

The least squares objective function is

$$\sigma^2 = \|\tilde{y}\|_2^2 = \|y - \Lambda_N^v \hat{x}_0^v\|_2^2 = y^T y + (\hat{x}_0^v)^* (\Lambda_N^v)^* \Lambda_N^v \hat{x}_0^v - y^T \Lambda_N^v \hat{x}_0^v - (\hat{x}_0^v)^* (\Lambda_N^v)^* y,$$

to be minimised over the $\lambda_i$ and $\xi_i$. Obviously, this is an unconstrained multivariate polynomial optimisation problem, and by putting all partial

derivatives equal to zero we will obtain a set of multivariate polynomials that is to be rooted.[8] We have to be a bit careful when dealing with derivatives of real functions (such as the objective function $\sigma^2$), with respect to complex variables like $\lambda_i$ and $\xi_i$ and their complex conjugates. Indeed, $\lambda_i$ as a complex number is 'parametrised' by two real numbers (its real and imaginary part) and its complex conjugate $\overline{\lambda}_i$ by the same two real numbers. Let's have a look at the objective function for $n = 2$ and $N = 3$, with two distinct complex conjugated roots $\lambda$ and $\overline{\lambda}$ and conjugated components $\xi$ and $\overline{\xi}$ of $x$, which is an example that is sufficiently general to make some points.[9] In essence, we treat $\lambda, \overline{\lambda}, \xi, \overline{\xi}$ as independent variables. We then find

$$\begin{aligned}
\frac{\partial \sigma^2}{\partial x} &= 0 = \quad (\Lambda_3^v)^* \Lambda_3^v x - (\Lambda_3^v)^* y = \frac{\partial \sigma^2}{\partial x^*} \ , \\
\frac{\partial \sigma^2}{\partial \lambda} &= 0 = \quad x^* (\frac{\partial (\Lambda_3^v)^*}{\partial \lambda} \Lambda_3^v + (\Lambda_3^v)^* \frac{\partial \Lambda_3^v}{\partial \lambda}) x - y^T \frac{\partial \Lambda_3^v}{\partial \lambda} x - x^* \frac{\partial (\Lambda_3^v)^*}{\partial \lambda} y \ , \\
\frac{\partial \sigma^2}{\partial \overline{\lambda}} &= 0 = \quad x^* (\frac{\partial (\Lambda_3^v)^*}{\partial \overline{\lambda}} \Lambda_3^v + (\Lambda_3^v)^* \frac{\partial \Lambda_3^v}{\partial \overline{\lambda}}) x - y^T \frac{\partial \Lambda_3^v}{\partial \overline{\lambda}} x - x^* \frac{\partial (\Lambda_3^v)^*}{\partial \overline{\lambda}} y \ .
\end{aligned}$$

With $\tilde{I}_2$ the $2 \times 2$ reverse identity matrix, we have the following relations:

$$\Lambda_3^v = \begin{pmatrix} 1 & 1 \\ \lambda & \overline{\lambda} \\ \lambda^2 & \overline{\lambda}^2 \end{pmatrix} \Longrightarrow \frac{\partial \Lambda_3^v}{\partial \lambda} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 2\lambda & 0 \end{pmatrix} = \overline{(\frac{\partial \Lambda_3^v}{\partial \overline{\lambda}})} \cdot \tilde{I}_2 = (\frac{\partial \overline{\Lambda_3^v}}{\partial \lambda}) \cdot \tilde{I}_2 = \overline{\frac{\partial \overline{\Lambda_3^v}}{\partial \overline{\lambda}}} \ .$$

We see that $x$ occurs linearly in the 'normal equations'. With $x^T = (\xi \ \overline{\xi})$ and recalling that $\Lambda_3^v x \in \mathbb{R}^N$, we then find

$$\frac{\partial \sigma^2}{\partial \lambda} = 2 \cdot \xi \cdot (0 \ 1 \ 2\lambda) \cdot (\Lambda_3^v x - y) = 0 \ , \ \frac{\partial \sigma^2}{\partial \overline{\lambda}} = 2 \cdot \overline{\xi} \cdot (0 \ 1 \ 2\overline{\lambda}) \cdot (\Lambda_3^v x - y) = 0 \ .$$

[8] Via Vieta's Theorem, the map between the coefficients in (1) and the roots $\lambda_i$ is of course multivariate polynomial. So it should come as no surprise that the optimisation problems in both model representations (the kernel and the image one) are 'equivalent' in the sense that one can be turned into the other via a multivariate polynomial transformation. Of course, the MEVP that solves them, has a different appearance. The set of multivariate polynomials is closed for many different operations applied to the variables: partial derivatives, multivariate polynomial transformations, etc.

[9] When $C$ is a complex matrix, $\overline{C}$ is the matrix obtained by replacing every element with its complex conjugate, $C^T$ is its transpose (without complex conjugation) and $C^* = (\overline{C})^T$ is its complex conjugate transpose

We see that in principle, $x = 0$ could also be a solution of these equations, which leads to $(\Lambda_3^v)^T y = 0$. In this case, $\lambda_1$ and $\lambda_2$ will be the roots of $y_2\lambda^2 + y_1\lambda + y_0 = 0$. Returning to the general case, for general $n$ and $N$, solutions with $\hat{x}_0^v = 0$, correspond to the $N-1$ roots of $\sum_{k=0}^{N-1} y_k\lambda^k = 0$, in which case $\sigma^2 = \|y\|_2^2$ and $\tilde{y} = 0$. All solutions here are maximising, not minimising. So we can assume that $x \neq 0$, and introduce the $N \times n$ matrix,

$$(\Lambda_N^v)^\lambda = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 1 & 1 & \dots & 1 \\ 2\lambda_1 & 2\lambda_2 & \dots & 2\lambda_n \\ \vdots & \vdots & \vdots & \vdots \\ (N-1)\lambda_1^{N-2} & (N-1)\lambda_2^{N-2} & \dots & (N-1)\lambda_n^{N-2} \end{pmatrix}.$$

Then we find

$$\begin{pmatrix} (\Lambda_N^v)^T\Lambda_N^v & (\Lambda_N^v)^T y \\ ((\Lambda_N^v)^\lambda)^T(\Lambda_N^v)^\lambda & ((\Lambda_N^v)^\lambda)^T y \end{pmatrix} \begin{pmatrix} x \\ -1 \end{pmatrix} = 0\,.$$

These are $2n$ equations in $2n$ unknowns, and as $x$ appears linearly, they form an MEVP. One can also eliminate $x$ all together by requiring that all $(n+1) \times (n+1)$ minors are zero, hence generating $n$ secular equations. Observe also that the misfit vector $\tilde{y} = y - \Lambda_N^v \hat{x}_0^v$ is perpendicular to the column spaces of both $\Lambda_N^v$ and $(\Lambda_N^v)^\lambda$, an observation to which we will return in Section 9.

## 8. Order $n = 1$ least squares realization

Let us work out, for didactics sake, the case $n = 1$ in some detail, in which case there is only 1 unknown $\alpha$ in the difference equation $\hat{y}_{k+1} + \alpha\hat{y}_k = 0$, and the root $\lambda = -\alpha$.

### 8.1. Kernel and image representations

In this case, eqs. (13) reduce to a single parameter polynomial EVP. Grouping powers of $\alpha$ then results in $(A_0 + A_1\alpha + A_2\alpha^2)w = 0$, where, for general $N$, $A_0, A_1, A_2 \in \mathbb{R}^{(2N-1)\times(2N-1)}$ and $w = (f^T\ (f^\alpha)^T\ -1)^T$.

There are now several ways to proceed. First, call $w_0 = w$ and $w_1 = w_0\alpha$, we can easily obtain the generalised EVP

$$\begin{pmatrix} 0 & I_{2N-1} \\ -A_0 & -A_1 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} I_{2N-1} & 0 \\ 0 & A_2 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} \alpha.$$

There might be eigenvalues at infinity, because typically $A_2$ is singular. The secular equations reduce here to one characteristic equation, given by

$$\det\left( \begin{pmatrix} 0 & I_{2N-1} \\ -A_0 & -A_1 \end{pmatrix} - \begin{pmatrix} I_{2N-1} & 0 \\ 0 & A_2 \end{pmatrix} \alpha \right) = 0 \, .$$

Another way to proceed, is to generate a block Toeplitz matrix with FSRs with increasing powers of $\alpha$ to get

$$\begin{pmatrix} A_0 & A_1 & A_2 & 0 & 0 & \dots \\ 0 & A_0 & A_1 & A_2 & 0 & \dots \\ 0 & 0 & A_0 & A_1 & A_2 & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} w_0 \\ w_0\alpha \\ w_0\alpha^2 \\ w_0\alpha^3 \\ \vdots \end{pmatrix} = 0.$$

The dimension of the null space will be the total number of projective roots, including the ones at infinity. The null space is a block backward shift-invariant subspace, in which we can first do a column compression as described in Figure 1, and then perform 1D realisation. A third way to proceed is to use the image representation with one pole:

$$\sigma^2 = \min_{\beta,\lambda} \sum_{k=0}^{N-1} (y_k - \beta\lambda^k)^2 \quad \Longrightarrow \quad \begin{array}{l} \frac{\partial \sigma^2}{\partial \beta} = 2 \sum_{k=0}^{N-1} (y_k - \beta\lambda^k)\lambda^k = 0 \, , \\ \frac{\partial \sigma^2}{\partial \lambda} = 2\beta \sum_{k=1}^{N} (y_k - \beta\lambda^k) k\lambda^{k-1} = 0 \, . \end{array}$$

Defining the vectors $y^T = (y_0 \; y_1 \; \dots \; y_{N-1})$, $\hat{y}^T = (\beta \; \beta\lambda \; \dots \beta\lambda^{N-1})$ and the misfit $\tilde{y} = y - \hat{y}$, we can rewrite these two equations as $\tilde{y}^T \hat{y} = 0$ and $\tilde{y}^T \frac{\partial \hat{y}}{\partial \lambda} = 0$: The misfit is orthogonal to the least squares approximating signal $\hat{y}$ and its derivative. For $\beta = 0$, solutions are the roots of $\sum_{k=0}^{N-1} y_k \lambda^k = 0$, and are the maximising ones. When $\beta \neq 0$, both equations are linear in $\beta$ so that

$$\beta = \frac{\sum_{k=0}^{N-1} y_k \lambda^k}{\sum_{k=0}^{N-1} \lambda^{2k}} = \frac{\sum_{k=1}^{N-1} k y_k \lambda^{k-1}}{\sum_{k=1}^{N-1} k \lambda^{2k-1}} \, ,$$

and $(\sum_{k=0}^{N-1} y_k \lambda^k)(\sum_{k=1}^{N-1} k\lambda^{2k-1}) = (\sum_{k=1}^{N-1} k y_k \lambda^{k-1})(\sum_{k=0}^{N-1} \lambda^{2k})$, which is the secular equation. It can be verified that the degree of this polynomial in $\lambda$ is $3N - 5$, with leading coefficient $y_{N-2}$.

### 8.2. A small numerical example for $n = 1$ and $N = 4$

As an illustrative example, for $N = 4$ data points, eqs. (13) reduce to

$$\left(\begin{array}{ccc|ccc|c} 2\alpha & 1 & 0 & 1+\alpha^2 & \alpha & 0 & y_0 \\ 1 & 2\alpha & 1 & \alpha & 1+\alpha^2 & \alpha & y_1 \\ 0 & 1 & 2\alpha & 0 & \alpha & 1+\alpha^2 & y_2 \\ \hline 1+\alpha^2 & \alpha & 0 & 0 & 0 & 0 & \alpha y_0 + y_1 \\ \alpha & 1+\alpha^2 & \alpha & 0 & 0 & 0 & \alpha y_1 + y_2 \\ 0 & \alpha & 1+\alpha^2 & 0 & 0 & 0 & \alpha y_2 + y_3 \\ \hline y_0 & y_1 & y_2 & \alpha y_0 + y_1 & \alpha y_1 + y_2 & \alpha y_2 + y_3 & 0 \end{array}\right) \left(\begin{array}{c} f \\ f^\alpha \\ -1 \end{array}\right) = 0.$$

The matrices $A_0, A_1, A_2$ are:

$$A_0 = \left(\begin{array}{ccc|ccc|c} 0 & 1 & 0 & 1 & 0 & 0 & y_0 \\ 1 & 0 & 1 & 0 & 1 & 0 & y_1 \\ 0 & 1 & 0 & 0 & 0 & 1 & y_2 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & y_1 \\ 0 & 1 & 0 & 0 & 0 & 0 & y_2 \\ 0 & 0 & 1 & 0 & 0 & 0 & y_3 \\ \hline y_0 & y_1 & y_2 & y_1 & y_2 & y_3 & 0 \end{array}\right), A_1 = \left(\begin{array}{ccc|ccc|c} 2 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 1 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & y_0 \\ 1 & 0 & 1 & 0 & 0 & 0 & y_1 \\ 0 & 1 & 0 & 0 & 0 & 0 & y_2 \\ \hline 0 & 0 & 0 & y_0 & y_1 & y_2 & 0 \end{array}\right),$$

$$A_2 = \left(\begin{array}{ccc|ccc|c} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Let us take data $y = (4\ 3\ 2\ 1)^T$. There will be $2(2N-1) = 14$ eigenvalues. The secular equation is $-4\lambda^{10} + 18\lambda^9 - 56\lambda^8 + 96\lambda^7 - 128\lambda^6 + 120\lambda^5 - 32\lambda^4 + 24\lambda^3 + 36\lambda^2 - 18\lambda + 24 = 0$, so there are 10 affine roots and 4 at infinity. The maximising roots are (rounded to 4 digits) 1.6506, $0.1747 \pm 1.5469$j and the ones corresponding to $\beta \neq 0$ are $1.3216 \pm 2.0058$j, $-0.6764$, $-0.1589 \pm 0.8080$j, $0.4209 \pm 0.6233$j. The minimising root is $\lambda = -0.6764$ and correspondingly, $\sigma^2 = 1.4868e-01$ with $\hat{x}_0 = 4.1155e+00$. The FSR with the polynomial EVP results in the following table:

| # | size | rank | nullity | linear independent monomials |
|---|---|---|---|---|
| 0 | $7 \times 21$ | 7 | 14 | $1,2,3,4,5,6,7\|8,9,10,11,12,14\|21$ |
| 1 | $14 \times 28$ | 14 | 14 | $1,2,3,4,5,6,7\|8,9,10,11,14\|21,28$ |
| 2 | $21 \times 35$ | 21 | 14 | $1,2,3,4,5,6,7\|8,9,10,14\|21,28,35$ |
| 3 | $28 \times 42$ | 28 | 14 | $1,2,3,4,5,6,7\|8,9,10\|21,28,35,42$ |
| 4 | $35 \times 49$ | 35 | 14 | $1,2,3,4,5,6,7\|8,9,10\|\text{gap}\|28,35,42,49$ |

In this table, the left column indicates the recursion number of the FSR. For recursion number 0, the block Toeplitz matrix has size $7 \times 21$ and is of rank 7, so has nullity 14, indicating that, when we look at a matrix the columns of which form a basis of the null space, 14 rows are linear independent. Their indices can be found in the right column. The little bar between 7 and 8 separates the degree 0 block from the degree 1 block. As the FSR proceeds, the nullity stabilises (actually for this example right from the beginning) to 14. The indices of the first 7 rows in the degree 0 block also stabilise right away, but in the degree 1 block, there are 5 linear independent monomials (recursion 1), then 4 (recursion 2) and from recursion 3 on, the indices in the degree 1 block stabilise to $8, 9, 10$. From recursion 4 we see for the first time in the recursion, the 'mind-the-gap'-zone, as all rows in the degree 2 block are linear dependent. There are now 4 linear independent monomials in the degree 3 block. The 10 linear independent rows above the gap are caused by the 10 affine roots, and the 4 below the gap, by the roots at infinity. We can now do a column compression of the matrix consisting of degree blocks 0 to 2, and then deploy 1D realisation to set up a $10 \times 10$ EVP, the eigenvalues of which will exactly correspond to the roots we have enumerated above. For the global minimum, we find

$$y = \begin{pmatrix} 4 \\ 3 \\ 2 \\ 1 \end{pmatrix} = \hat{y} + \tilde{y} = \begin{pmatrix} 4.1155e+00 \\ 2.7835e+00 \\ 1.8827e+00 \\ 1.2734e+00 \end{pmatrix} + \begin{pmatrix} -1.1549e-01 \\ 2.1645e-01 \\ 1.1732e-01 \\ -2.7336e-01 \end{pmatrix} .$$

## 9. New system theoretic properties of the optimal solutions

In this Section, the optimal solutions of the least squares realisation program, are characterised by four properties, which have not been described before in the literature: 1. The fact that the misfit is structured; 2. Orthogonality properties; 3. Optimal Riemannian metrics; 4. A canonical decomposition of the ambient data space in complementary forward and backward shift invariant subspaces, defined by the optimal eigen-tuples.

### 9.1. Beurling-Lax-Halmos: the misfit is structured

Assuming that the model coefficients $\alpha_i$ are known, we can rephrase the optimisation problem (6) using a vector of Lagrange multipliers $l \in \mathbb{R}^N$ (we add a factor $1/2$ for convenience): $\min_{\tilde{y}} \ \frac{1}{2}\|\tilde{y}\|_2^2$ subject to $T_a(y - \tilde{y}) = 0$, with the Langrangean function as $L(\tilde{y}, l) = \frac{1}{2}\|\tilde{y}\|_2^2 + l^T T_a(y - \tilde{y})$. Equating

all partial derivatives to zero, results in the set of equations: $\tilde{y} = T_a^T l$ and $T_a y = T_a \tilde{y}$. Hence $T_a y = T_a T_a^T l$, so that $l = (T_a T_a^T)^{-1} T_a y$ and

$$\tilde{y} = T_a^T (T_a T_a^T)^{-1} T_a y = \Pi_a y = \tilde{y} = T_a^T l \; . \tag{18}$$

We now also see from Subsection 5.3, that the vector $f$ we introduced there is the same as the vector $l$ of Lagrange multipliers. The interpretation of (18) is the following: the misfit signal $\tilde{y}$ is obtained by filtering the input sequence contained in $l$ through a finite impulse response (FIR) filter, the zeros of which are the reciprocals of the roots (poles) of the difference equation we started from: Eq. (1) can be written as $a(z)\hat{y}_k$ where $z$ is the forward shift operator. This implies that $\tilde{y}_k = [(a^{\text{rev}}(z))/z^n] f_k$, where the coefficients of $a^{\text{rev}}(z)$ are those of $a(z)$ in reversed order and driven with the 'input' sequence $f_k$ (appropriately padded with zeros). So we find that $\tilde{y}$ itself is generated by a FIR linear system, the zeros of which are the reciprocals of the roots that characterise the backward shift-invariant column space of $\Gamma_N$. Hence, the subspace of all vectors $\tilde{y}$ orthogonal to the backward shift-invariant subspace determined by the roots of $a(z)$, consists of 'input' vectors $f$ that are filtered through the FIR filter $a^{rev}(z)/z^n$, the zeros of which are the reciprocals of the roots of $a(z)$. This can be seen as a finite-dimensional vector space version of the operator-theoretic Theorem of Beurling-Lax-Halmos (see e.g. [44] [75]).[10]

### 9.2. Orthogonal decomposition à la Thales

From $T_a \cdot \hat{y} = 0$ and $\tilde{y} = T_a^T \cdot f$, we easily deduce that $\hat{y}$ and $\tilde{y}$ are perpendicular to each other, as $\hat{y}$ is orthogonal to the rows of $T_a$, and $\tilde{y}$ is

[10]The properties of the unilateral forward (right) and backward (left) shift operator have been intensively studied the last 50 years in operator theory [23] [72]. Beurling [16] characterized all forward shift invariant subspaces of the Hardy space $H_2$ as having the form $\theta(z)f(z)$ with $f \in H_2$ and $\theta(z)$ an inner function, the zeros of which 'characterise' the shift-invariant subspace. Lax [63] extended this result to finite-dimensional vector-valued functions (where the unit disc is replaced by the right half-plane) and Halmos [51] later proved this for the infinite-dimensional case. The orthogonal complement $K_\theta = (\theta(z)H_2)^\perp$ of a forward shift invariant space is invariant under the backward shift operator $S^*$ (adjoint of S), referred to as the model space [23] [38] [45], because in this space very general classes of Hilbert space contractions can be modelled by the backward shift operator $S^*$ acting on $K_\theta$. Famous results include Sz.-Nagy's Dilation Theorem (every contraction has a unitary dilation (i.e. a 'power preserving' unitary lifting)) and the Commutant Lifting Theorem (CLT) of the Sz.-Nagy Foias model space [44]. Backward multi-shift-invariant subspaces on the polydisc are studied in [11] [12] [13].

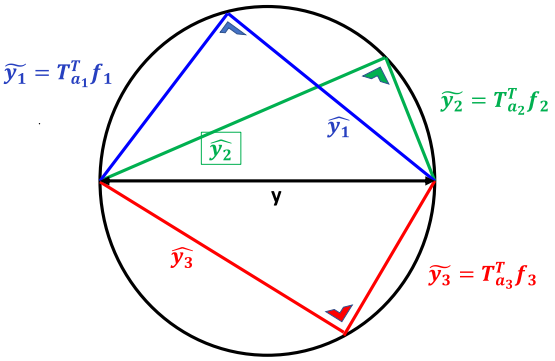


Figure 2: For every choice of model parameters $\alpha_i$ (not necessary optimal), there is an orthogonal decomposition of the given data vector $y$, in two orthogonal vectors: $\hat{y}$ belonging to the behaviour of the model (the orthogonal projection of the data vector $y$ onto the backward shift-invariant subspace generated by the roots of $a(z)$), and a vector $\tilde{y}$ orthogonal to it. We show three such possible orthogonal decompositions.

a linear combination of these rows. So $\hat{y}^T\tilde{y} = 0$. Furthermore, it follows that $\|y\|_2^2 = \|\hat{y}\|_2^2 + \|\tilde{y}\|_2^2$. Said in other words, if $T_a$ would be known, every data vector $y$ can be decomposed orthogonally into two vectors $\hat{y}$ and $\tilde{y}$ as $y = \hat{y} + \tilde{y} = (I_N - \Pi_a)y + \Pi_a y$ where $\Pi_a = T_a^T(T_aT_a^T)^{-1}T_a$, where the first term is the orthogonal projection of the data vector $y$ into the null space of $T_a$ and the second term into the row space of $T_a$. But here, out of this infinite set of orthogonal projectors, one for each thinkable choice of the parameters $\alpha_i, i = 1, \ldots, n$, we want to find that specific projector $\Pi_a = T_a^T(T_aT_a^T)^{-1}T_a$ that minimises $\|\tilde{y}\|_2^2$ over the coefficients of $a(z)$. Hence, we can interpret the objective function (8) as the problem of finding the optimal metric, represented by the nonnegative definite symmetric projection operator matrix $\Pi_a$, in the following sense: The set of all row spaces of $T_a$ over all possible vectors $a$, is a manifold. The misfit $\tilde{y}$ belongs to its tangent space, while $\hat{y}$ is orthogonal to it. We are looking for an optimal choice of $a$, so that the orthogonal decomposition of the data vector $y$ as $y = \hat{y} + \tilde{y}$ is such that the norm of $\tilde{y}$, as measured in the non-negative definite metric induced by $\Pi_a$, $\sigma^2 = \|\tilde{y}\|^2 = \tilde{y}^T\Pi_a\tilde{y}$, is minimised. Another interpretation arises by swapping the role of $y$ and $a$ as $T_a y = Ya = e$, where $e$ is an equation error and $Y$ is a Hankel matrix with the data, of appropriate dimensions. The objective function can now be written as $\sigma^2 = y^T T_a^T(T_aT_a^T)^{-1}T_a y = a^T Y(T_aT_a^T)^{-1}Ya = e^T(T_aT_a^T)^{-1}e$,

so it is a weighted quadratic function of the equation error to be minimised over the coefficients $\alpha_i$.

### 9.3. Walsh's Theorem as double Beurling-Lax-Halmos

Let us now show that the optimal vector of Lagrange multipliers $l$ itself is the output of a FIR filter (actually the same as the one we already described). Let's focus on eqs. (11). From the first block row, we find $f^{\alpha_i} = -D_a^{-1} D_a^{\alpha_i} f + D_a^{-1} T_a^{\alpha_i} y$. Substitute this in the second block row $y^T (T_a^{\alpha_i})^T f + y^T T_a^T f^{\alpha_i} = 0$ and using $f = D_a^{-1} T_a y$, we find $2 f^T T_a^{\alpha_i} y = f^T D_a^{\alpha_i} f$. Now, we use $D_a^{\alpha_i} = (T_a T_a^T)^{\alpha_i} = T_a^{\alpha_i} T_a^T + T_a (T_a^{\alpha_i})^T$ and (18) to find $f^T T_a^{\alpha_i} (y - \tilde{y}) = 0$. Writing this out for all $\alpha_i, i = 1, \ldots, n$ and using $\hat{y} = y - \tilde{y}$, we then obtain (illustrated here for $N = 6$ and $n = 2$):

$$f^T \begin{pmatrix} \hat{y}_0 & \hat{y}_1 \\ \hat{y}_1 & \hat{y}_2 \\ \hat{y}_2 & \hat{y}_3 \\ \hat{y}_3 & \hat{y}_4 \end{pmatrix} = 0 \ .$$

We deduce that $f$ itself belongs to the left null space of an observability matrix (coinciding with the column space of the Hankel matrix above), so that similarly to (18) and the reasoning that follows, $f$ itself must be the output of a FIR filter driven by an unknown signal $g$. These conclusions hold for general $n$ and $N$: Let $S_a \in \mathbb{R}^{(N-2n)\times(N-n)}$ be the banded Toeplitz with the model parameters $\alpha_i$. Then, there exists a vector $g \in \mathbb{R}^{N-2n}$ so that

$$f = S_a^T g \ \text{ and } \ \tilde{y} = T_a^T f = (T_a^T S_a^T) g \ .$$

Said in other words, the misfit $\tilde{y}$ is generated by filtering an unknown signal $g$ twice through the same FIR filter, the zeros of which are the reciprocals of the roots that characterise the backward shift-invariant subspace determined by the optimal roots: $f = (a^{rev}(z)/z^n)^2 g$ for some $g$. This is a finite-dimensional vector space version of what in the $H_2$-model reduction literature is known as Walsh's Theorem (see e.g. [75]). So, in a certain sense, Walsh's Theorem, which is a characterisation of the optimal roots, is a 'double' Beurling-Lax-Halmos theorem that characterise the shift-invariant subspace to which the optimal misfit $\tilde{y}$ belongs.

### 9.4. Structured decomposition of the ambient space

The previous results lead to a (new) canonical decomposition of the ambient space $R^N$. The matrix $S_a T_a$ is itself a $(N - 2n) \times N$ banded Toeplitz

matrix of rank $N-2n$, that corresponds to a difference equation of the form $(a^T p(z))^2 y_k = 0$. The polynomial equation $(a^T p(z))^2 = 0$ has the same roots as $a^T p(z)$, but all multiplicities are doubled. For simplicity, we assume that all roots of $p(z) = 0$ are simple, which implies that all multiplicities of the roots of $(a^T p(z))^2 = 0$ are 2. Indeed, it is not so difficult to show that

$$(S_a T_a) \cdot \begin{pmatrix} \Lambda_N^v & (\Lambda_N^v)^\lambda \end{pmatrix} = 0 \,,$$

which basically expresses the fact that the following column spaces, parameterised by the optimal coefficients $\alpha_i$, are complementary and orthogonal:

$$\mathbb{R}^N = \mathbf{R}(T_a^T S_a^T) \oplus \mathbf{R}(\Lambda_N^v) \oplus \mathbf{R}((\Lambda_N^v)^\lambda) \text{ and } \mathbf{R}(T_a^T S_a^T) \perp \mathbf{R} \begin{pmatrix} \Lambda_N^v & (\Lambda_N^v)^\lambda \end{pmatrix} .$$

The column space of $(\Lambda_N^v \ (\Lambda_N^v)^\lambda)$ is backward shift-invariant, and only depends on the original spectrum, the roots of $a^T p(z) = 0$, but now with double multiplicity. The corresponding shift matrix $A \in \mathbb{R}^{2n \times 2n}$ will be similar to a Jordan form, with for each root a $2 \times 2$ Jordan block. The orthogonal complement of this backward shift-invariant subspace is the forward shift-invariant *model space*, generated by the column space of the matrix $T_a^T S_a^T$. Let is illustrate this with some examples. For $n = 1$ and $N = 5$, we have $\alpha = -\lambda$, so that the canonical decomposition of $\mathbb{R}^5$ becomes:

$$\begin{pmatrix} (S_a T_a)^T & \Lambda_5^v & (\Lambda_5^v)^\lambda \end{pmatrix} = \left(\begin{array}{ccc|c|c} \alpha^2 & 0 & 0 & 1 & 0 \\ 2\alpha & \alpha^2 & 0 & \lambda & 1 \\ 1 & 2\alpha & \alpha^2 & \lambda^2 & 2\lambda \\ 0 & 1 & 2\alpha & \lambda^3 & 3\lambda^2 \\ 0 & 0 & 1 & \lambda^4 & 4\lambda^3 \end{array}\right)$$

One can readily verify that the columns of the first block column are orthogonal to the last two columns, and that the matrix is of full rank 5. For $n = 2$ and $N = 7$, we find

$$\begin{pmatrix} (S_a T_a)^T & \Lambda_7^v & (\Lambda_7^v)^\lambda \end{pmatrix} = \left(\begin{array}{ccc|cc|cc} \alpha_2^2 & 0 & 0 & 1 & 1 & 0 & 0 \\ 2\alpha_2\alpha_1 & \alpha_2^2 & 0 & \lambda_1 & \lambda_2 & 1 & 1 \\ \alpha_1^2 + 2\alpha^2 & 2\alpha_1\alpha_2 & \alpha_2^2 & \lambda_1^2 & \lambda_2^2 & 2\lambda_1 & 2\lambda_2 \\ 2\alpha_1 & \alpha_1^2 + 2\alpha_2 & 2\alpha_1\alpha_2 & \lambda_1^3 & \lambda_2^3 & 3\lambda_1^2 & 3\lambda_2^2 \\ 1 & 2\alpha_1 & \alpha_1^2 + 2\alpha_2 & \lambda_1^4 & \lambda_2^4 & 4\lambda_1^3 & 4\lambda_2^3 \\ 0 & 1 & 2\alpha_1 & \lambda_1^5 & \lambda_2^5 & 5\lambda_1^4 & 5\lambda_2^4 \\ 0 & 0 & 1 & \lambda_1^6 & \lambda_2^6 & 6\lambda_1^5 & 6\lambda_2^5 \end{array}\right) .$$

## 10. Applications, heuristics, extensions

### 10.1. Applications

Signals $\hat{y}_k$ that can be modelled by the kernel, image and state representations of Section 3 are sometimes called *Bohl functions*, which, for discrete time, are functions that are linear combinations of terms of the form $k^l exp(\lambda k)$, where $k, l \in \mathbb{N}$ and $\lambda \in \mathbb{C}$. These are basic signal modes that appear in many fields of science and engineering as 'eigenfunctions' of linear time-invariant difference equations. Interchanging the role of $\hat{y}$ and $a$, we find $T_a\ \hat{y} = \hat{Y}a = 0$, where $\hat{Y}$ is a $(N-n) \times (n+1)$ Hankel matrix generated from the sequence $\hat{y}_k$. This equation expresses the well-known fact that a Hankel matrix, generated from a sequence $\hat{y}_k$ generated by the difference equation (1), has rank $n$. Obviously, for the vector $a$ to be essentially unique, we require that the number of rows of $\hat{Y}$ is strictly larger than $n+1$, which implies that we require that $N \geq 2n+1$. The universality of Bohl functions reveals itself in the hundreds (if not thousands) of papers that deal with applications in which these functions appear, such as (we only give a small sample of references): high resolution frequency estimation and harmonic retrieval problems [1] [17] [21] [31], the shape from moments problem [41], direction-of-arrival problems [50] [76], realisation algorithms from impulse response samples to state space models [52] [62] [90]. In addition, there are many relations with classical moment and interpolation problems (e.g. Caratheodory, Hamburger, Nevanlinna-Pick, etc.) [4] [60] that remain to be explored.

### 10.2. Heuristic approaches

The results derived here for the globally optimal solution to the least squares realisation problem, could be used to assess the effectiveness of the tens of heuristic algorithms that have been described in the literature. Some, but not all of them, can be shown to converge to a local minimum. An algorithm that seems to converge is the following heuristic iteration [21], applied to the (short) data sequence $y = (4\ 3\ 2\ 1)^T$ we discussed before. The $3 \times 2$ Hankel matrix $H_0$ formed with it is of rank 2. The best least squares Frobenius norm matrix approximation of rank 1 is given by the first singular triplet, say the rank 1 matrix $H_1$, but it will not be Hankel. The closest Hankel matrix in Frobenius-norm can be found by 'averaging' the anti-diagonals. Call it $H_2$, which will be Hankel but not of rank 1. Use the SVD again to calculate the best rank 1 approximation $H_3$., which will not be Hankel.

Average the anti-diagonals to obtain $H_4$, etc. The even iterations in this alternating projections algorithm will be Hankel, the odd ones will have rank 1, but if this iteration converges, we have a rank 1 Hankel, which unfortunately does not satisfy the necessary conditions (13) for a (global) optimum:

$$H = \begin{pmatrix} 4 & 3 \\ 3 & 2 \\ 2 & 1 \end{pmatrix} \rightarrow H_1 \rightarrow H_2 \rightarrow \ldots H_\infty = \begin{pmatrix} 4.1593e+00 & 2.8441e+00 \\ 2.8441e+00 & 1.9448e+00 \\ 1.9448e+00 & 1.3299e+00 \end{pmatrix}.$$

Obviously, this is an example of a heuristic alternating least squares algorithm that seems to converge to some stationary point, that is not to a (local, let alone global) minimum. Other heuristics start from the correct necessary conditions for a (local) minimum, but then have to resort to plain non-linear optimisation algorithms. Examples are Iterative Quadratic Maximum Likelihood (IQML, see e.g. [64]), Steiglitz-McBride (see e.g. [43] [75]), plain numerical optimization (like in PEM in [67]), Constrained Total Least Squares [1], the Riemannian SVD [30] [31] [33] [35] (which is actually a heuristic method to find the minimising solution of (13)). A good survey is provided in the PhD thesis [46], which also provides another heuristic method, called weighted null space fitting. We hope that our results will shed some new light in understanding these heuristic approaches (e.g. the nature of their 'fixed points' if and to which they converge). A special mentioning deserves a 'classic' paper by Golub and Pereyra [48] on separable nonlinear least squares problems, describing a heuristic called VARPRO (Variable Projection), basically referring to the metric $\Pi_a$ that is updated in each step of an iteratively re-weighted least squares iteration.

### 10.3. Extensions

Before concluding, let us mention some potential extensions:

**Least squares:** Since the times of Gauss and Legendre, least squares has provided the basis for an uncountable number of estimation theories because it is analytically tractable with simple linear algebra. A possible extension could include weights in the objective function, that can be any positive real number, the inverse of which reflects the a priori confidence one would have concerning the relative size of the misfits on specific data points. Extreme cases are an infinite weight for an 'exact' (not to be modified), and a zero weight (modification does not matter, so a missing observation) data point.

Another example is the Hilbert-Schmidt-Hankel norm, the Frobenius norm of a Hankel matrix, which has also a very interesting system theoretical interpretation as the surface under the Nyquist plot [28] and in operator theory forms the definition of a Dirichlet space [42].

**Minimisation:** In (13), we omitted the equation for $\sigma^2$, and we derived a solution method to find all the coefficients of (1) or equivalently, the roots. However, an alternative is to keep $\sigma^2$ in the equations, and design iterative algorithms that only find those roots that correspond to a globally minimal $\sigma^2$, for instance some version of the inverse power method [88]. We will show how to do this in some future work.

**Including a priori information:** We did not invoke any statistical assumptions here, but of course, under the appropriate conditions of Gaussianity and whiteness, our approach could be interpreted in a maximum likelihood setting. However, the approach described here can never be asymptotically (as $N \to \infty$) efficient in the statistical sense, as we also estimate the complete misfit vector $\tilde{y}$ (implicitly or explicitly), variables which, in the literature on errors-in-variables, are called the 'incidental parameters' (see [77]). One could also start from (13) to consider the second order derivatives, leading to Hessians, that might be useful in characterising sensitivities and conditioning.

**Recursiveness in $n$ and $N$** is also a problem to be looked into, both in terms of recursive updating with respect to containing the computational complexity, as in trying to find an optimal order $n$ as a trade of between bias and variance. When the number of data $N \to \infty$, and the data $y$ are themselves generated by an autonomous LTI system of order $m > n$, the problem described here becomes the $H_2$ model reduction problem (see [6], and as an MEVP [3]).

## 11. Concluding remarks

In this paper, we could only schematically sketch an outline of the major results: Least squares optimal realisation of observed data is basically an MEVP, and hence a sequence of SVDs and EVPs. It is surprising that a (difficult) nonlinear problem in a 1D system theoretic setting, can in principle be solved exactly as a series of SVDs and EVPs, in an $n$-dimensional system theoretic setting. The required steps involve writing the first order optimality conditions as an MEVP, next, using FmSRs to generate a block Macaulay

matrix, finding in its null space the regular part that is multi-shift invariant. Then use a $n$-dimensional realisation step to calculate the 'shift' matrices $A_1, \ldots, A_n$, the eigenvalues of which will generate the $n$-tuples $(\alpha_1, \ldots, \alpha_n)$, one of which corresponds to the global minimum. In doing so, we also presented a new solution method for MEVPs (see also [29] [36]). Of course, many more details will be discussed elsewhere, but let's make some final observations. This work is a fascinating combination of several disciplines, like numerical linear algebra, system theory in one and more dimensions, (commutative) algebraic geometry, operator theory, etc. The results presented here are part of a larger system identification framework that was started with the papers [32] [65], inspired by Willems's behavioural framework [68] [89], where we not only deal with misfits, but also with 'latent' unobserved inputs). In [65] we have proposed the framework, but we did not have yet the insights of the combination of MEVPs, multi-shift invariant subspaces, multi-dimensional realisation theory that leads to the solution strategy described in this paper, but that also applies to more general misfit-latency models (see e.g. [86] for the ARMA case). All of these ideas will be pursued in future work, that will focus in particular on large scale numerical linear algebra algorithms.

## References

[1] T. J. Abatzoglou, J. M. Mendel, G. A. Harada, *The constrained total least squares technique and its applications to harmonic superresolution*, IEEE Trans. Signal Process., vol. 39, no. 5, pp. 1070–1087, 1991.

[2] V. M. Adamjan, D. Z. Srov, M. G. Krein, *Mat. USSR Sbornik*, 15, pp. 31–73, 1971 (English Translation).

[3] M. Agudelo, C. Vermeersch, B. De Moor, *Globally optimal $H_2$-norm model reduction: A numerical linear algebra approach*, Internal Report 20-07, ESAT-STADIUS, KU Leuven (Leuven, Belgium), 2020, submitted for publication.

[4] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis*. Oliver and Boyd, Edinburgh and London, 1965. MR0184042

[5] A. C. Antoulas, *Mathematical System Theory: The Influence of R. E. Kalman*. Springer, 2013.

[6] A. C. Antoulas, *Approximation of Large-scale Dynamical Systems*, Advanced in Design and Control. SIAM, 2005, 479 pp. MR2155615

[7] F. V. Atkinson, *Multiparameter spectral theory*, Bull. Am. Math. Soc., vol. 74, no. 1, pp. 1–27, 1968. MR0220078

[8] F. V. Atkinson, *Multiparameter Eigenvalue Problems Volume 1: Matrices and Compact Operators*. Academic Press, New York (N.Y.), 1972. MR0451001

[9] F. Atkinson, A. Mingarelli, *Multi-parameter Eigenvalue Problems; Sturm-Liouville Theory*. CRC Press, Taylor and Francis Group, 2011, 279 pp. MR2760763

[10] S. Attasi, *Modelling and recursive estimation for double indexed sequences*, in: System Identification Advances and Case Studies, R. K. Mehra and D. G. Lainiotis (eds.), pp. 289–348. Academic Press, 1976. MR0688160

[11] J. A. Ball, W. S. Li, D. Timotin, and T. T. Trent, *A commutant lifting theorem on the polydisc*, Indiana Univ. Math. J., vol. 48, no. 2, pp. 653–675, 1999. MR1722812

[12] J. A. Ball, G. Groenewald, T. Malakorn, *Structured noncommutative multidimensional linear systems*, SIAM J. Control Optim., vol. 44, no. 4, pp. 1474–1528, 2005. MR2178040

[13] J. A. Ball, V. Bolotnikov, Q. Fang, *Multivariable backward-shift-invariant subspaces and observability operators*, Multidimens. Syst. Signal Process., vol. 18, no. 4, pp. 191–248, 2007. MR2432062

[14] S. Basu S., D. J. Velleman, *On Gauss's first proof of the fundamental theorem of algebra*, Am. Math. Mon., vol. 124, no. 8, pp. 688–694, 2017. MR3706816

[15] K. Batselier, P. Dreesen, B. De Moor, *A fast iterative orthogonalization scheme for the Macaulay matrix*, Journal of Computational and Applied Mathematics, vol. 267, pp. 20–32, Sep. 2014. MR3181679

[16] A. Beurling, *On two problems concerning linear transformations in Hilbert space*, Acta Math., vol. 81, pp. 239–255, 1949. MR0027954

[17] Y. Bresler, A. Macovski, *Exact maximum likelihood parameter estimation of superimposed exponential signals in noise*, IEEE Trans. Acoust., vol. 1, no. 5, pp. 1081–1089, 1986.

[18] Å. Björck, *Numerical Methods in Matrix Computations*. Springer International Publishing, 2015. MR3288840

[19] S. Boyd, L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. MR2061575

[20] R. Bru, J. Vitkia, *Report on the International Conference on Linear Algebra and Applications, 28–30 September 1987, Universidad Politecnica de Valencia, Valencia, Spain*, Linear Algebra and its Applications, vol. 121, pp. 537–548, August 1989.

[21] J. Cadzow, *Signal enhancement: a composite property mapping algorithm*, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 36, no. 2, pp. 49–62, 1988.

[22] R. D. Carmichael, *Boundary Value Problems and Expansion Problems. Part I: Algebraic Basis of the Theory*: Amer. J. Math, vol. 43, pp. 69–101, 1921; *Part II: Formulation of Various Transcendental Problems*: Amer. J. Math, vol. 43, pp. 232–270, 1921; *Part III: Oscillatory, companion and expansion problems*: Amer. J. Math, vol. 44, pp. 129–152, 1922. MR1506450

[23] J. A. Cima, W. T. Ross, *The Backward Shift on the Hardy Space*, American Mathematical Society, 2000. MR1761913

[24] D. A. Cox, J. Little, D. O'Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, 4th ed. Springer, Cham, 2015. MR3330490

[25] D. A. Cox, J. Little, D. O'Shea, *Using Algebraic Geometry*, 2nd ed. Springer, 2005. MR2122859

[26] B. H. Dayton, Z. Zeng, *Computing the multiplicity structure in solving polynomial systems*, in: Proceedings of the 2005 International Symposium on Symbolic and Algebraic computation – ISSAC '05, pp. 116–123, 2005. MR2280537

[27] J. W. Demmel, *Applied Numerical Linear Algebra*. SIAM, Philadelphia, PA, 1997. MR1463942

[28] K. De Cock, B. Hanzon, B. De Moor, *On a cepstral norm for an ARMA model and the polar plot of the logarithm of its transfer function*, Signal Processing, vol. 83, no. 2, pp. 439–443, 2003.

[29] K. De Cock, B. De Moor, *Multiparameter Eigenvalue Problems and Shift-invariance*, Internal Report 20-06, ESAT-STADIUS, KU Leuven (Leuven, Belgium), 2020, submitted for publication.

[30] B. De Moor, *Structured total least squares and L2 approximation problems*, Linear Algebra Appl., vols. 188–189, pp. 163–205, 1993. MR1223460

[31] B. De Moor, *Total least squares for affinely structured matrices and the noisy realization problem*, IEEE Trans. on Signal Processing, vol. 42, no. 11, pp. 3104–3113, 1994.

[32] B. De Moor, B. Roorda, *L2-optimal linear system identification Structured Total Least Squares for SISO systems*, in: Proceedings of the 33rd IEEE Conference on Decision and Control, pp. 2874–2879, 1994. MR1447479

[33] B. De Moor, *The Riemannian singular value decomposition*, in: SVD and Signal Processing III: Algorithms, Architectures and Applications, M. Moonen and B. De Moor (eds.), pp. 61–78. Elsevier, 1995.

[34] B. De Moor, *Structured total least squares for Hankel matrices*, in: Communications, Computation, Control and Signal Processing, Paulray A., Roychowdhury V., and Schaper C.D. (eds.), A tribute to Thomas Kailath, pp. 243–258.. Kluwer Academic Publishers, 1997. MR1325926

[35] B. De Moor, *Linear system identification, structured total least squares and the Riemannian SVD*, in: Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling, S. Van Huffel (ed.). SIAM, Philadelphia, 1997, pp. 225–238. MR1447479

[36] B. De Moor, *The multi-parameter eigenvalue problem*. ESAT-STADIUS Technical Report, March 2020, submitted for publication.

[37] H. Derksen, *The fundamental theorem of algebra and linear algebra*, Am. Math. Mon., vol. 110, no. 7, pp. 620–623, 2003. MR2001152

[38] R. G. Douglas, H. S. Shapiro, *Cyclic vectors and invariant subspaces for the backward shift operator*, Ann. l'Institut Fourier, vol. 20, no. 1, pp. 37–76, 1970. MR0270196

[39] P. Dreesen, *Back to the Roots: Polynomial System Solving Using Linear Algebra*, PhD thesis, KU Leuven, Leuven, Belgium, 2013.

[40] P. Dreesen, K. Batselier, B. De Moor, *Multidimensional realisation theory and polynomial system solving*, Int. J. Control, vol. 91, no. 12, pp. 2692–2704, 2018. MR3894707

[41] M. Elad, P. Milanfar, G. H. Golub, *Shape from moments – an estimation theory perspective*, IEEE Trans. Signal Process., vol. 52, no. 7, pp. 1814–1828, 2004. MR2087688

[42] O. El-Fallah, K. Kellay, J. Mashreghi, T. Ransford, *A Primer on the Dirichlet Space*. Cambridge University Press, 2014 MR3185375

[43] N. Everitt, M. Galrinho, H. Hjalmarsson, *Open-loop asymptotically efficient model reduction with the Steiglitz–McBride method*, Automatica, vol. 89, pp. 221–234, 2018. MR3762050

[44] C. Foias, A. Frazho, *The Commutant Lifting Approach to Interpolation Problems*, Operator Theory: Advances and Applications, vol. 44. Birkhauser, 1990, 632 pp. MR1120546

[45] S. R. Garcia, J. Mashreghi, W. T. Ross, *Introduction to Model Spaces and Their Operators*. Cambridge University Press, 2016. MR3526203

[46] M. Galrinho, *System Identification with Multi-step Least-Squares Methods*. PhD Thesis, KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science, 2018, 307 pp.

[47] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their $L_\infty$ error bounds*, Int. J. Control, vol. 39, pp. 1115–1193, 1984. MR0748558

[48] G. H. Golub, V. Pereyra, *Separable nonlinear least squares: the variable projection method and its applications*, Inverse Probl., vol. 19, no. 2, pp. R1–R26, 2003. MR1991786

[49] G. H. Golub, C. F. Van Loan, *Matrix Computations*, 4th ed. Johns Hopkins University Press, Baltimore, MD, 2013. MR3024913

[50] M. Haardt, *Structured least squares to improve the performance of ESPRIT-type algorithms*, IEEE Trans. Signal Process., vol. 45, no. 3, pp. 792–799, 1997.

[51] P. R. Halmos, *Shifts on Hilbert spaces*, J. Reine Angew. Math., vol. 208, pp. 102–112, 1961. MR0152896

[52] B. L. Ho, R. E. Kalman, *Effective construction of linear, state-variable models from input/output functions*, Regelungstechnik, vol. 14, pp. 545–548, 1966. MR0245360

[53] M. E. Hochstenbach, A. Muhič, B. Plestenjak, *On linearizations of the quadratic two-parameter eigenvalue problem*, Linear Algebra Appl., vol. 436, no. 8, pp. 2725–2743, 2012. MR2908593

[54] M. E. Hochstenbach, B. Plestenjak, *Backward error, condition numbers, and pseudospectra for the multiparameter eigenvalue problem*, Linear Algebra Appl., vol. 375, pp. 63–81, 2003. MR2013456

[55] T. Kailath, *Linear Systems*. Prentice Hall, Englewood Cliffs, 1980. MR0569473

[56] T. Kailath, A. Sayed, B. Hassibi, *Linear Estimation*, Prentice Hall Information and System Science Series. Prentice Hall, 2000, 854 pp. MR0569473

[57] R. E. Kalman, *Mathematical descriptions of linear dynamical systems*, SIAM J. Control, vol. 1, pp. 152–192, 1963. MR0152167

[58] R. E. Kalman, *What is system theory?* Lecture given on the occasion of the Kyoto prize, 1985. URL: https://www.kyotoprize.org/wp-content/uploads/2019/07/1985_A.pdf

[59] R. E. Kalman, *Old and new directions of research in system theory*, Perspect. Math. Syst. Theory, Control. Signal Process., vol. 398, pp. 3–13, 2010. MR2647646

[60] M. G. Krein, A. A. Nudel'man, *The Markov Moment Problem and Extremal Problems*. American Mathematical Society, 1977. MR0458081

[61] L. Kronecker, *Zur Theorie der Elimination einer Variablen aus zwei algebraischen Gleichungen*, Monatsber. Konigl. Preussischen Acad. Wiss. (Berlin), 1881, pp. 535–600. Reprinted as pp. 113–192 in Mathematische Werke, vol. 2, B.G. Teubner, Lepizig, 1897 or Chelsea, New York, 1968.

[62] S.Y. Kung, *A new identification and model reduction algorithm via singular value decomposition*, in: Proceedings of the 12th Asilomar Conference on Circuits, Systems and Computers, Institute of Electrical and Electronics Engineers, New York, 1978, pp. 705–714.

[63] P. D. Lax, *Translation invariant spaces*, Acta Math., vol. 101, no. 3, pp. 163–178, 1959. MR0105620

[64] P. Lemmerling, L. Vanhamme, S. Van Huffel, B. De Moor, *IQML-like algorithms for solving structured total least squares problem*, Signal Processing, vol. 81, pp. 1935–1945, 2001. MR1952938

[65] P. Lemmerling, B. De Moor, *Misfit versus latency*, Automatica, vol. 37, no. 12, pp. 2057–2067, 2001.

[66] B. Q. Li, *A direct proof and a transcendental version of the fundamental theorem of algebra via Cauchy's theorem*, Am. Math. Mon., vol. 121, no. 1, pp. 75–77, 2014. MR3139585

[67] L. Ljung, *System Identification: Theory for the User, 2nd ed.*, Information and System Sciences Series. Prentice Hall, 1999. MR3444790

[68] I. Markovsky, J. C. Willems, S. Van Huffel, B. De Moor, *Exact and Approximate Modelling of Linear Systems: A Behavioral Approach*. SIAM, 2006. MR2207544

[69] M. Moonen, B. De Moor, J. Ramos, S. Tan, *A subspace identification algorithm for descriptor systems*, Syst. Control Lett., vol. 19, pp. 47–52, 1992. MR1170987

[70] A. Muhič, B. Plestenjak, *On the singular two-parameter eigenvalue problem*, Electron. J. Linear Algebr., vol. 18, pp. 420–437, 2009. MR2530144

[71] A. Muhič, B. Plestenjak, *On the quadratic two-parameter eigenvalue problem and its linearization*, Linear Algebra Appl., vol. 432, no. 10, pp. 2529–2542, 2010. MR2608173

[72] N. K. Nikol'skiĭ, *Treatise on the Shift Operator: Spectral Function Theory*. Springer-Verlag, Berlin–Heidelberg, 1986.

[73] B. Plestenjak, C. I. Gheorghiu, M. E. Hochstenbach, *Spectral collocation for multiparameter eigenvalue problems arising from separable boundary value problems*, J. Comput. Phys., vol. 298, pp. 585–601, 2015. MR3374566

[74] B. Plestenjak, M. E. Hochstenbach, *Roots of bivariate polynomial systems via determinantal representations*, SIAM J. Sci. Comput., vol. 38, no. 2, pp. A765–A788, 2016. MR3473599

[75] P. A. Regalia, *Adaptive IIR Filtering in Signal Processing and Control*. Marcel Dekker Inc., 1995, 678 pp.

[76] R. Roy, T. Kailath, *ESPRIT-estimation of signal parameters via rotational invariance techniques*, Acoust. Speech Signal Process. IEEE Trans., vol. 37, no. 7, pp. 984–995, 1989. MR1058066

[77] T. Soderstrom, *Errors-in-variables Methods in System Identification*, Communications and Control Engineering. Springer, 2018. MR3791479

[78] H. J. Stetter, *Numerical Polynomial Algebra*. SIAM, Philadelphia, 2004. MR2048781

[79] G. Strang, *The fundamental theorem of linear algebra*, Am. Math. Mon., vol. 100, no. 9, pp. 848–855, 1993. MR1247531

[80] G. Strang, *Introduction to Linear Algebra*, 5th ed. Cambridge Press, 2016.

[81] J. Suzuki, *Lagrange's proof of the fundamental theorem of algebra*, Am. Math. Mon., vol. 113, no. 8, pp. 705–714, 2006. MR2256531

[82] L. N. Trefethen, D. Bau, *Numerical Linear Algebra*. SIAM, 1997. MR1444820

[83] B. L. van der Waerden, *Algebra I, II*. Springer, 1991. MR1080173

[84] P. Van Overschee, B. De Moor, *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers, 1996, 254 pp. (See also ftp://ftp.esat.kuleuven.be/pub/stadius/ida/reports/96-26a.pdf). MR1427490

[85] C. Vermeersch, B. De Moor, *A Column Space Based Approach to Solve Systems of Multivariate Polynomial Equations*, Internal Report 20-08, ESAT-STADIUS, KU Leuven (Leuven, Belgium), 2020, submitted for publication.

[86] C. Vermeersch, B. De Moor, *Globally optimal least-squares ARMA model identification is an eigenvalue problem*, IEEE Control Systems Letters (L-CSS), vol. 2, no. 4, pp. 1062–1067, Oct. 2019.

[87] H. Volkmer, *Multi-parameter Eigenvalue Problems and Expansion Theorems*. Lecture Notes in Mathematics. Springer-Verlag, 1988, 157 pp. MR0973644

[88] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*. Oxford University Press, 1995. MR0950175

[89] J. Willems, *From time series to linear system*, Automatica, *Part I. Finite dimensional linear time invariant systems*, vol. 22, pp. 561–580, 1986; *Part II. Exact modelling*, vol. 22, pp. 675–694, 1986; *Part III. Approximate modelling*, vol. 23, pp. 87–115, 1987. MR0870059

[90] H. Zeiger, A. McEwen, *Approximate linear realizations of given dimension via Ho's algorithm*, IEEE Transactions on Automatic Control, vol. 19, no. 2, pp. 153–153, May 1974.

Bart De Moor
ESAT-STADIUS
KU Leuven
Kasteelpark Arenberg 10
B-3001 Leuven (Heverlee)
Belgium
*E-mail address:* bart.demoor@kuleuven.be

Received February 27, 2020