Citation/Reference	Authors: Sibren Lagauw, Bart De Moor Title: Least Squares Projection Onto the Behavior for SISO LTI Models Journal/Conference/Book: Proc. of the 20th IFAC Symposium on System Identification (SYSID2024) Vol/Book Chapter: / Pages: /
Archived version	■ Submitted version ☐ Accepted (Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher) ☐ Published ☐ Published (Open Access)
Acknowledgements	In this manuscript, the following funding is acknowledged: ☐ EU H22020 ☐ EU Horizon Europe ■ FWO ■ KULeuven Internal Funds ☐ VLAIO ☐ AI ■ ERC ☐ Other
Published version	
Copyright statement	
Author contact	sibren.lagauw@esat.kuleuven.be bart.demoor@esat.kuleuven.be

Least Squares Projection Onto the Behavior for SISO LTI Models ★

S. Lagauw *, **B. De Moor** * *Fellow IEEE, IFAC, SIAM*

* *Center for Dynamical Systems, Signal Processing, and Data Analytics (STADIUS), Department of Electrical Engineering (ESAT), KU Leuven, Belgium*
Email addresses: {sibren.lagauw, bart.demoor}@esat.kuleuven.be

Abstract: We consider the least squares projection onto the behavior for linear time-invariant (LTI) single-input single-output (SISO) models, in which the observed input-output data are modified in a least squares (LS) sense by subtracting so-called misfits, so that the modified data satisfy a given linear dynamic relation. We show that the LS-criterion of the projection problem induces an orthogonal decomposition of the ambient data space and we characterize this decomposition by means of banded (block-) Toeplitz matrices, the elements of which are the coefficients of the difference equation describing the SISO LTI dynamics. We thereby generalize earlier results in the literature on autonomous LTI models to the more complicated SISO case. Additionally, we illustrate that the novel characterization is equivalent (up to a change of model representation) to results derived using (isometric) state space representations in the literature on behavioral systems theory.

Keywords: Modeling and identification, linear systems, parametric optimization.

1. INTRODUCTION

System identification translates observed time-series data into a mathematical model, which typically belongs to a user-specified model class. In practice, however, the observed data as such are almost never *model-compliant*, i.e., the data do not belong to the *behavior* of the model, which is the set of all trajectories that exactly satisfy the model dynamics [Willems, 1987, Heij, 1989]. There could be many reasons for this, e.g., measurement inaccuracies, missing data, observational errors, and model mismatch. Stochastics are often introduced to explain this discrepancy between the observed data and the model [Aoki and Yue, 1970, Ljung, 1999]. However, unless one has a priori information about the nature of the stochastic disturbances in the observed data, one might want to avoid making any statistical assumption(s). Instead, a more natural way to proceed is to modify the observed data as little as possible, by subtracting so-called *misfits*, so that the modified data become model-compliant. The smaller these misfits, the better the 'fit' between the model and the observed data.

We consider discrete-time causal single-input single-output (SISO) linear time-invariant (LTI) models of finite order n. For this model class, input data $\hat{\boldsymbol{u}} = [\hat{u}_0, \dots, \hat{u}_{N-1}]^\mathsf{T} \in \mathbb{R}^N$ and output data $\hat{\boldsymbol{y}} = [\hat{y}_0, \dots, \hat{y}_{N-1}]^\mathsf{T} \in \mathbb{R}^N$ are model-compliant if and only if they are exactly related by an nth order LTI input-output relation,

$$\sum_{i=0}^{n} a_i \hat{y}_{k-i} - \sum_{i=0}^{n} b_i \hat{u}_{k-i} = 0, \ \forall k \in \{n, \dots, N-1\}, \tag{1}$$

where we assume that $N > n$. The $2(n+1)$ coefficients that appear in (1) are the model parameters of the SISO LTI model, which we will group in the vectors $\boldsymbol{a} = [a_0, \dots, a_n]^\mathsf{T} \in \mathbb{R}^{n+1}$ and $\boldsymbol{b} = [b_0, \dots, b_n]^\mathsf{T} \in \mathbb{R}^{n+1}$. As such, given an observed input signal $\boldsymbol{u} = [u_0, \dots, u_{N-1}]^\mathsf{T} \in \mathbb{R}^N$ and output signal $\boldsymbol{y} = [y_0, \dots, y_{N-1}]^\mathsf{T} \in \mathbb{R}^N$, the identification problem consists out of the minimization of the 2-norm of the difference between the observed and model-compliant data sequences $\widetilde{\boldsymbol{u}} = \boldsymbol{u} - \hat{\boldsymbol{u}}$ and $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$, the so-called *misfits*, over the variables $\hat{\boldsymbol{u}}, \hat{\boldsymbol{y}}$, and the model parameters $\boldsymbol{a}, \boldsymbol{b}$,

$$\begin{aligned} \min_{\hat{\boldsymbol{u}}, \hat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{b}} \quad & J = \left\| \begin{bmatrix} \boldsymbol{y} - \hat{\boldsymbol{y}} \\ \boldsymbol{u} - \hat{\boldsymbol{u}} \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} \widetilde{\boldsymbol{y}} \\ \widetilde{\boldsymbol{u}} \end{bmatrix} \right\|_2^2, \\ \text{subject to} \quad & \text{Eq. (1) and } a_0 = 1, \end{aligned} \tag{2}$$

where the second constraint is a normalization constraint to ensure uniqueness of the difference relation in (1). This *misfit modeling* setup, which is depicted in Figure 1, corresponds to the *pure-misfit*-case of the more general misfit-versus-latency framework proposed in [Lemmerling and De Moor, 2001]. The problem is also known in the literature on behavioral systems theory [Willems, 1986], where it is described as a particular instance of the *global total least squares* (GTLS) problem [Roorda and Heij, 1995]: the GTLS formulation is more general than (2) in the sense that it considers the broader class of multiple-input multiple-output (MIMO) models.

★ This work was supported in part by the KU Leuven: Research Fund under Project iBOF/23/064 and Project C16/15/059, and several LRD bilateral industrial projects; in part by the Flemish Government agencies: FWO through the EOS Project under Grant G0F6718N (SeLMA) and EWI through the Flanders AI Research Program; and in part by the European Commission (ERC Adv. Grant) under Grant 885682. The work of Sibren Lagauw was supported by the FWO Fundamental Research Fellowship under Grant 11K5623N.

Plain nonlinear solvers could be used to tackle the constrained nonlinear least squares problem (2), but also dedicated solution approaches exist. The problem can be reformulated to a structured total least squares problem, where the (locally) optimal model parameters are to be retrieved via the non-trivial null space of a structured matrix approximation of the block-Hankel matrix containing the observed input-output data [De Moor, 1993, 1994, Markovsky et al., 2005, 2006].

Alternatively, the methods proposed in [Roorda, 1995] exploit the observation that (2) is a double minimization problem: 1. for given observed data and model parameters, the minimal norm misfits that make the observed data model-compliant are to be retrieved (the inner minimization), 2. the model parameters itself are to be optimized (the outer minimization). In particular, the fact that the optimal solution to the inner-problem is given by a linear transformation, allows one to substitute the closed-form solution into the outer problem, thereby eliminating the $2N$ decision variables $\hat{\boldsymbol{u}}, \hat{\boldsymbol{y}}$ from the optimization problem. A similar approach has been taken in [De Moor, 2020, Lagauw et al., 2023], where the misfit modeling problem for autonomous LTI models is considered: the optimal misfits are expressed as the orthogonal projection of the observed data onto a linear subspace, which is completely determined [1] by the parameters of the autonomous model.

In this paper, we study the inner minimization problem: the *least squares projection onto the behavior for SISO LTI models*. Our contribution is twofold:

(1) In Section 2, we show that the least squares criterion induces an orthogonal decomposition of the ambient data space. We completely characterize this decomposition using banded (block-) Toeplitz matrices, the elements of which are the coefficients of the difference equation describing the SISO LTI dynamics (1). We thereby generalize the results from [De Moor, 2020, Section 4] to the more complicated SISO case. Several optimality properties are obtained (e.g., the optimal misfits are structured, in the sense that they are itself generated by an nth order LTI model), which shed new light on results obtained in previous work [De Moor, 1994, De Moor and Roorda, 1994].

(2) In Section 3, we reconsider the projection onto the behavior using an (isometric) state space representation to model the SISO LTI dynamics, instead of the difference relation in (1), similar to the framework proposed in [Roorda and Heij, 1995]. We demonstrate that the orthogonal decomposition of the ambient space as established in Section 2 can also be characterized in this setting, thereby providing a state space based alternative to the results obtained in Section 2. The combination of both perspectives allows for a more intuitive understanding of the obtained results.

In practice, depending on the particular use-case, one of the two alternative characterizations of the optimal solution can be preferred. In future work, we want to exploit the obtained expression(s) to derive a novel, globally optimal approach for SISO LTI misfit modeling.

[1] The behavior of the model corresponds to the right null space of a banded Toeplitz matrix, which is constructed from the coefficients of the difference equation describing the autonomous LTI dynamics. The least squares optimal misfits are shown to belong to the orthogonal complementary subspace of this behavior, i.e., the row space of the banded Toeplitz matrix [De Moor, 2020].

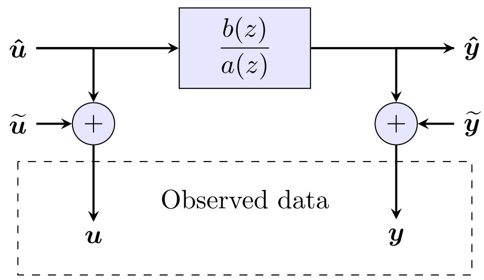

Fig. 1. The observed input-output signals $\boldsymbol{u}, \boldsymbol{y} \in \mathbb{R}^N$, respectively, are decomposed into a model-compliant part $\hat{\boldsymbol{u}}, \hat{\boldsymbol{y}}$, and so-called misfits $\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{y}}$, the 2-norm of which is to be minimized. The coefficients $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{n+1}$ of the nth degree polynomials $a(z)$ and $b(z)$ in the forward-shift z, are the model parameters.

Notation We denote scalars by lowercase letters, e.g., a, and tuples/vectors by boldface lowercase letters, e.g., $\boldsymbol{a}$. Matrices are characterized by boldface uppercase letters, e.g., $\boldsymbol{A}$. The transpose, and Hermitian transpose of $\boldsymbol{a}$ are indicated by $\boldsymbol{a}^\mathsf{T}$ and $\boldsymbol{a}^\mathsf{H}$, respectively. The 2-norm of a vector or matrix is denoted by $\|\cdot\|_2$ and the pseudo-inverse of the matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^\dagger$. Furthermore, to ease the notation, we introduce the observed *trajectory* $\boldsymbol{w} \in \mathbb{R}^{2N}$, which comprises the input-output data,

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{w}_0^\mathsf{T} & \dots & \boldsymbol{w}_{N-1}^\mathsf{T} \end{bmatrix}^\mathsf{T} = \begin{bmatrix} y_0 & u_0 & \dots & y_{N-1} & u_{N-1} \end{bmatrix}^\mathsf{T},$$

such that $\boldsymbol{w}_k = \begin{bmatrix} y_k & u_k \end{bmatrix}^\mathsf{T}$, and we define $\hat{\boldsymbol{w}}$ and $\widetilde{\boldsymbol{w}}$ similarly. The partial derivative of a scalar function $f(\dots)$ with respect to $\boldsymbol{w}_k$ will be denoted as,

$$\frac{\partial f(\dots)}{\partial \boldsymbol{w}_k} = \begin{bmatrix} \frac{\partial f(\dots)}{\partial y_k} & \frac{\partial f(\dots)}{\partial u_k} \end{bmatrix}^\mathsf{T}.$$

2. PROJECTION ONTO THE BEHAVIOR

In the projection problem, a given sequence of observed data $\boldsymbol{w}$ has to be projected onto the linear dynamics of a given model (i.e., the model parameters $\boldsymbol{a}, \boldsymbol{b}$ of (1) are known), so that the 2-norm of the obtained misfits $\widetilde{\boldsymbol{w}}$ is minimal. With $\boldsymbol{l} = [l_0, \dots, l_{N-n-1}]^\mathsf{T} \in \mathbb{R}^{N-n}$ and $\lambda \in \mathbb{R}$ Lagrange multipliers, the Lagrangian of this projection problem becomes,

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{l}, \lambda) = &\sum_{k=0}^{N-1} \|\boldsymbol{w}_k - \hat{\boldsymbol{w}}_k\|_2^2 + \lambda(a_0 - 1) \\ &+ \sum_{k=n}^{N-1} l_{k-n} \left(\sum_{i=0}^{n} a_i \hat{y}_{k-i} - \sum_{i=0}^{n} b_i \hat{u}_{k-i} \right). \quad (3)\end{aligned}$$

The first order necessary conditions for optimality, which can be obtained by equating all partial derivatives of (3) to zero, allow to derive the following results.

The behavior Writing out the linear relation from (1) for all $k = 0, \dots, N-1$, gives,

$$\widetilde{\boldsymbol{T}}^\mathsf{T} \hat{\boldsymbol{w}} = \boldsymbol{0}, \quad (4)$$

where $\widetilde{\boldsymbol{T}} \in \mathbb{R}^{(2N)\times(N-n)}$ a banded block-Toeplitz matrix, the transpose of which is defined [2] as,

$$\widetilde{\boldsymbol{T}}^{\mathsf{T}} = \begin{bmatrix} a_n & -b_n & a_{n-1} & -b_{n-1} & \dots & \dots & a_0 & -b_0 & & & & & & \\ & & a_n & -b_n & \dots & \dots & a_1 & -b_1 & a_0 & -b_0 & & & & \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & & & & a_n & -b_n & \dots & \dots & a_1 & -b_1 & a_0 & -b_0 \end{bmatrix}.$$

Equation (4) shows that the model-compliant data $\hat{\boldsymbol{w}}$ must lie in the (right) null space of $\widetilde{\boldsymbol{T}}^{\mathsf{T}}$, i.e., the behavior of the model from (1) corresponds to null($\widetilde{\boldsymbol{T}}^{\mathsf{T}}$). The latter is an $(N{+}n)$-dimensional subspace of $\mathbb{R}^{2N}$, since $a_0 = 1$ ensures that $\widetilde{\boldsymbol{T}}$ is of full-column rank. The banded block-Toeplitz structure of $\widetilde{\boldsymbol{T}}^{\mathsf{T}}$ allows us to construct a basis for its null space by hand, leading to the matrix $\hat{\boldsymbol{T}} \in \mathbb{R}^{(2N)\times(N+n)}$:

$$\hat{\boldsymbol{T}} = \left[\begin{array}{ccc|cccc|ccc} b_n & \dots & b_1 & b_0 & & & & & & \\ a_n & \dots & a_1 & a_0 & & & & & & \\ & \ddots & \vdots & b_1 & b_0 & & & & & \\ & \ddots & \vdots & a_1 & a_0 & \ddots & & & & \\ & & b_n & \vdots & \vdots & \ddots & b_0 & & & \\ & & a_n & \vdots & \vdots & & a_0 & & & \\ & & & b_n & b_{n-1} & & \vdots & b_0 & & \\ & & & a_n & a_{n-1} & \ddots & \vdots & a_0 & & \\ & & & & b_n & \ddots & b_{n-1} & \vdots & \ddots & \\ & & & & a_n & \ddots & a_{n-1} & \vdots & \ddots & \\ & & & & & \ddots & b_n & b_{n-1} & \dots & b_0 \\ & & & & & & a_n & a_{n-1} & \dots & a_0 \end{array}\right].$$

If we assume that the model in (1) is minimal, such that the polynomials $a(z)$ and $b(z)$ are coprime, the columns of $\hat{\boldsymbol{T}}$ are guaranteed to be linearly independent (this is a consequence of [Legat et al., 2023, Lemma 3.3]). Combined with (4), we find that,

$$\hat{\boldsymbol{w}} \in \text{null}(\widetilde{\boldsymbol{T}}^{\mathsf{T}}) = \text{range}(\hat{\boldsymbol{T}}).$$

We can conclude that the $(N{+}n)$-dimensional subspace range($\hat{\boldsymbol{T}}$) corresponds to the *behavior* associated to the model described by the input-output dynamics from (1). Unsurprisingly, the dimensionality of the behavior corresponds to the degrees of freedom (dofs) that one has to construct the trajectory $\hat{\boldsymbol{w}}$: when simulating an nth order SISO model for $k = 0, \dots, N{-}1$, there are n initial conditions and N inputs that can be chosen freely.

The misfit space It is straightforward, though notationally tedious, to show that equating $\partial\mathcal{L}(\dots)/\hat{\boldsymbol{w}}_k$ to zero for all $k = 0, \dots, N{-}1$, leads to,

$$\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{T}}\boldsymbol{l}, \tag{5}$$

indicating that the optimal output misfits $\widetilde{\boldsymbol{y}}$ can be expressed as the convolution of the vectors $\boldsymbol{a}$ and $\boldsymbol{l}$, and similarly that the optimal input misfits $\widetilde{\boldsymbol{u}}$ are obtained from the convolution of $-\boldsymbol{b}$ and $\boldsymbol{l}$. Furthermore, it can be seen that the misfit trajectory $\widetilde{\boldsymbol{w}}$ 'resides' in the $(N{-}n)$-dimensional subspace spanned by the columns of the matrix $\widetilde{\boldsymbol{T}}$. We will call this subspace the *misfit space*.

Orthogonality The orthogonal complementarity between the row and (right) null space of a matrix imply that the model- and misfit spaces are orthogonal complementary subspaces in $\mathbb{R}^{2N}$. Indeed, the matrix product $\widetilde{\boldsymbol{T}}^{\mathsf{T}}\hat{\boldsymbol{T}}$ is zero by construction. By consequence, the first-order necessary conditions for optimality of the data projection problem, and thus also of (2), induce an orthogonal decomposition of the ambient data space,

$$\mathbb{R}^{2N} = \underbrace{\text{range}(\hat{\boldsymbol{T}})}_{\text{behavior}} \oplus \underbrace{\text{range}(\widetilde{\boldsymbol{T}})}_{\text{misfits}}. \tag{6}$$

This decomposition implies that the optimal trajectories $\hat{\boldsymbol{w}}$ and $\widetilde{\boldsymbol{w}}$ are mutually orthogonal,

$$\hat{\boldsymbol{w}} \perp \widetilde{\boldsymbol{w}}. \tag{7}$$

As shown in [De Moor, 1994], one can alternatively derive (7) by substituting (5) in (4), after multiplying (4) from the left with $\boldsymbol{l}^{\mathsf{T}}$.

Structured misfits Counterintuitively, a natural consequence of the data orthogonality (7) is that besides the model-compliant data, also the optimal input-output misfits $\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{y}}$, are structured. In particular, the $(n{+}1)$th equation up until the Nth equation [3] in,

$$\hat{\boldsymbol{T}}^{\mathsf{T}}\widetilde{\boldsymbol{w}} = \boldsymbol{0}, \tag{8}$$

show that the input-output misfits satisfy a linear difference relation of the following form,

$$\sum_{i=0}^{n} b_{n-i}\widetilde{y}_{k-i} + \sum_{i=0}^{n} a_{n-i}\widetilde{u}_{k-i} = 0, \ \forall k \in \{n, \dots, N{-}1\}.$$

Characterizing these misfit dynamics in the z-domain (where z denotes the forward-shift: $zy_k = y_{k+1}$) leads to,

$$\widetilde{H}(z) = -\frac{a_r(z)}{b_r(z)} = \frac{b_n z^n + \dots + b_1 z + b_0}{a_n z^n + \dots + a_1 z + a_0}, \tag{9}$$

with the 'reversed-coefficients'-polynomials defined as $a_r(z) = z^n a(z^{-1})$ and $b_r(z) = z^n b(z^{-1})$. The transfer function in (9) mapping the inputs $\widetilde{\boldsymbol{u}}$ to the outputs $\widetilde{\boldsymbol{y}}$ can be interpreted as the 'inverse' of the transfer function of the model governed by (1),

$$\hat{H}(z) = \frac{b(z)}{a(z)} = \frac{b_0 z^n + \dots + b_{n-1} z + b_n}{a_0 z^n + \dots + a_{n-1} z + a_n}.$$

Alternatively, assuming that the given model (1) is stable, one can derive (9) immediately from (7) by expressing the orthogonality [4] in the z-domain, giving,

$$\begin{aligned} &\langle \frac{b(z)}{a(z)}\hat{U}(z), \widetilde{Y}(z)\rangle + \langle \hat{U}(z), \widetilde{U}(z)\rangle = 0 \\ \iff\ & \frac{1}{2\pi i}\oint_{|z|=1}\left[\widetilde{Y}(z)\frac{b(z^{-1})}{a(z^{-1})} + \widetilde{U}(z)\right]\hat{U}(z^{-1})\mathrm{d}z = 0 \\ \iff\ & \widetilde{Y}(z) = \frac{-a(z^{-1})}{b(z^{-1})}\widetilde{U}(z) = \frac{-a_r(z)}{b_r(z)}\widetilde{U}(z). \end{aligned}$$

[2] Notice that this banded block-Toeplitz matrix is the SISO generalization of the banded Toeplitz matrix derived in [De Moor, 2020].

[3] Notice the difference compared to the model-compliant data: the first and last n equations in (8) additionally constrain $\widetilde{\boldsymbol{w}}$. We will see in Section 3 that these constraints ensure that the initial ($k = 0$) and final ($k = N{-}1$) state of the misfit model is zero.

[4] The transfer (matrix) function of causal, stable, discrete-time LTI models can be considered as an element of the Hardy space $\mathcal{H}_2$ of functions analytic on the exterior of the unit-disc. The inner-product $\langle \boldsymbol{F}(z), \boldsymbol{G}(z)\rangle$ of two (matrix) functions $\boldsymbol{F}(z), \boldsymbol{G}(z) \in \mathcal{H}_2$ is defined as:

$$\langle \boldsymbol{F}(z), \boldsymbol{G}(z)\rangle = \frac{1}{2\pi i}\oint_{\mathbb{T}} \boldsymbol{F}^{\mathsf{H}}(1/z^{\mathsf{H}})\boldsymbol{G}(z)\frac{dz}{z}, \tag{10}$$

with $\mathbb{T}$ the unit circle: $|z| = 1$. Note that in the context of LTI systems, the function $\boldsymbol{F}(z)$ corresponds to a transfer (matrix) function with real-valued impulse response, such that $\boldsymbol{F}^{\mathsf{H}}(1/z^{\mathsf{H}}) = \boldsymbol{F}^{\mathsf{T}}(1/z)$ [Wahlberg, 2003].

The relation in (9) was also obtained in [De Moor and Roorda, 1994, De Moor, 1994], albeit without proof.

Optimal projection It is a consequence of the orthogonal decomposition of the ambient space in (6) that the least-squares optimal decomposition of the observed input-output data $\boldsymbol{w} = \hat{\boldsymbol{w}} + \widetilde{\boldsymbol{w}}$ for a given model $(\boldsymbol{a}, \boldsymbol{b})$ boils down to an orthogonal projection problem: the minimal norm misfits $\widetilde{\boldsymbol{w}}$ are obtained via the orthogonal projection of $\boldsymbol{w}$ onto the misfit space, range$(\widetilde{\boldsymbol{T}})$,

$$\widetilde{\boldsymbol{w}} = (\widetilde{\boldsymbol{T}}^{\top})^{\dagger}\widetilde{\boldsymbol{T}}^{\top}\boldsymbol{w} = \widetilde{\boldsymbol{T}}(\widetilde{\boldsymbol{T}}^{\top}\widetilde{\boldsymbol{T}})^{-1}\widetilde{\boldsymbol{T}}^{\top}\boldsymbol{w},$$

or equivalently, the optimal regular data $\hat{\boldsymbol{w}}$ can be computed via the orthogonal projection of $\boldsymbol{w}$ onto range$(\hat{\boldsymbol{T}})$,

$$\hat{\boldsymbol{w}} = (\boldsymbol{I} - (\widetilde{\boldsymbol{T}}^{\top})^{\dagger}\widetilde{\boldsymbol{T}}^{\top})\boldsymbol{w}.$$

3. THE STATE SPACE DUAL

Inspired by the framework developed in [Roorda and Heij, 1995], in which the least-squares misfit modeling problem has been studied from a 'behavioral' point of view, we reconsider the projection onto the behavior in this Section using a state space model to parametrize the SISO LTI dynamics, instead of the difference equation from (1). Unsurprisingly, we find the same results as in Section 2, albeit that they appear in a different form due to the state space representation.

Behavioral system theoretic models Alternative to the relation in (1), we can enforce the nth order SISO LTI dynamics on the model-compliant data $\hat{\boldsymbol{w}}$ by requiring the data to satisfy the following behavioral state space recurrence relations,

$$\begin{aligned}\hat{\boldsymbol{x}}_{k+1} &= \boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{B}\hat{v}_k \\ \hat{\boldsymbol{w}}_k &= \boldsymbol{C}\hat{\boldsymbol{x}}_k + \boldsymbol{D}\hat{v}_k,\end{aligned} \qquad \text{for } k = 0, \ldots, N-1, \qquad (11)$$

with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n}$, $\boldsymbol{C} \in \mathbb{R}^{2\times n}$, $\boldsymbol{D} \in \mathbb{R}^{2}$. Contrary to a standard input-output state space representation, which maps inputs $\hat{\boldsymbol{u}}$ to outputs $\hat{\boldsymbol{y}}$, the model in (11) maps a sequence of auxiliary variables $\hat{\boldsymbol{v}} \in \mathbb{R}^N$ to the trajectory $\hat{\boldsymbol{w}}$. The auxiliary variables $\hat{\boldsymbol{v}}$ have no immediate physical interpretation, and one should not consider them as 'inputs' [Willems, 1986]. Rather, they represent the degree of freedom in the trajectory $\hat{\boldsymbol{w}}$, at each time step k, that originates from the SISO structure of (1).

Besides the standard basis change of the state vector $\hat{\boldsymbol{x}}_k$, which makes up the similarity transform of standard input-output state space models, also the addition of static feedback and a change of basis [5] of the auxiliary variables $\hat{v}_k$ can be performed without altering the behavior of (11) [Roorda and Heij, 1995]. In particular, it has been shown that, assuming minimality and stability of the original input-output relation modeled by (11), one can always perform a similarity transformation on the behavioral model (11) so that the resulting model becomes *isometric* [Roorda and Heij, 1995],

$$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{bmatrix}^{\top}\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{bmatrix} = \boldsymbol{I}_{n+1}.$$

[5] Note that the auxiliary variables are scalar in the SISO case, thereby excluding the possibility of this change of basis.

Therefore, to simplify our derivations, we can assume without loss of generality that the behavioral state space model $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ in (11) is isometric.

Given an observed trajectory $\boldsymbol{w}$ and model $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$, the projection onto the behavior now consists of finding the minimal norm ***widetildew*** such that the obtained $\hat{\boldsymbol{w}} = \boldsymbol{w} - \widetilde{\boldsymbol{w}}$ is model-compliant with respect to (11). Using the Lagrange multipliers $\boldsymbol{\lambda}_k \in \mathbb{R}^n$ and $\boldsymbol{\mu}_k \in \mathbb{R}^2$ for $k = 0, ..., N-1$, we can write the Lagrangian of the projection problem as,

$$\mathcal{L}(\hat{\boldsymbol{w}}, \hat{\boldsymbol{x}}, \hat{\boldsymbol{v}}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}\sum_{k=0}^{N-1}\|\hat{\boldsymbol{w}}_k - \boldsymbol{w}_k\|_2^2$$
$$+\sum_{k=0}^{N-1}\boldsymbol{\lambda}_k^{\top}(\hat{\boldsymbol{x}}_{k+1} - \boldsymbol{A}\hat{\boldsymbol{x}}_k - \boldsymbol{B}\hat{v}_k) + \sum_{k=0}^{N-1}\boldsymbol{\mu}_k^{\top}(\hat{\boldsymbol{w}}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k - \boldsymbol{D}\hat{v}_k).$$

Behavior The structure of the behavioral state space model allows us to easily construct a basis for the behavior in which $\hat{\boldsymbol{w}}$ resides. Writing out the recurrence relation in (11) for $k = 0, \ldots, N-1$, gives,

$$\hat{\boldsymbol{w}} = \hat{\boldsymbol{H}}\hat{\boldsymbol{v}} + \boldsymbol{\Gamma}\hat{\boldsymbol{x}}_0,$$

where $\hat{\boldsymbol{H}} \in \mathbb{R}^{2N\times N}$ is the matrix build from the Markov parameters of the model,

$$\hat{\boldsymbol{H}} = \begin{bmatrix} \boldsymbol{D} & 0 & \ldots\ldots & \ldots & 0\\ \boldsymbol{CB} & \boldsymbol{D} & 0 \ \ldots & \ldots & 0\\ \boldsymbol{CAB} & \boldsymbol{CB} & \boldsymbol{D} \ \ 0 & \ldots & 0\\ \vdots & \vdots & \ddots \ \ \ddots & \ddots & \vdots\\ \vdots & \vdots & \ddots & \ddots & 0\\ \boldsymbol{CA}^{N-2}\boldsymbol{B} & \boldsymbol{CA}^{N-3}\boldsymbol{B} & \ldots\ldots & \boldsymbol{CB} & \boldsymbol{D}\end{bmatrix} \qquad (12)$$

and $\boldsymbol{\Gamma} \in \mathbb{R}^{2N\times n}$ denotes its (extended) observability matrix,

$$\boldsymbol{\Gamma} = \left[\boldsymbol{C}^{\top} \ (\boldsymbol{CA})^{\top} \ \ldots \ (\boldsymbol{CA}^{N-1})^{\top}\right]^{\top}.$$

Because of the isometry and minimality of (11) we know that the rank of $\left[\boldsymbol{\Gamma}, \hat{\boldsymbol{H}}\right]$ is equal to $N+n$, indicating that the union of range$(\boldsymbol{\Gamma})$ and range$(\hat{\boldsymbol{H}})$ serves as a basis for the $(N+n)$-dimensional behavior of the model in (11),

$$\hat{\boldsymbol{w}} \in \left[\text{range}(\hat{\boldsymbol{H}}) \ \oplus \ \text{range}(\boldsymbol{\Gamma})\right]. \qquad (13)$$

Structured misfits Equating the partial derivatives of the Lagrangian with respect to the state vectors $\hat{\boldsymbol{x}}_k$ and the auxiliary inputs $\hat{v}_k$ to zero for all $k = 0, \ldots, N-1$, immediately shows that the sequence of Lagrange multipliers $\boldsymbol{\mu}_k$ equals the misfits $\widetilde{\boldsymbol{w}}_k = \boldsymbol{w}_k - \hat{\boldsymbol{w}}_k$. In turn, we can use the partial derivatives with respect to the state $\hat{\boldsymbol{x}}_{\boldsymbol{k}}$ and the auxiliary variables $\hat{v}_k$ to show that the Lagrange multipliers $\boldsymbol{\lambda}_{\boldsymbol{k}}$ satisfy a particular backward recursion,

$$\begin{bmatrix}\partial\mathcal{L}/\partial\hat{\boldsymbol{x}}_{\boldsymbol{k}}\\ \partial\mathcal{L}/\partial\hat{v}_k\end{bmatrix} = \boldsymbol{0} \iff \begin{bmatrix}\boldsymbol{\lambda}_{\boldsymbol{k-1}}\\ 0\end{bmatrix} = \begin{bmatrix}\boldsymbol{A}^{\top} & \boldsymbol{C}^{\top}\\ \boldsymbol{B}^{\top} & \boldsymbol{D}^{\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{\lambda}_{\boldsymbol{k}}\\ \widetilde{\boldsymbol{w}}_{\boldsymbol{k}}\end{bmatrix}, \qquad (14)$$

where we substituted $\boldsymbol{\mu}_{\boldsymbol{k}}$ by $\widetilde{\boldsymbol{w}}_{\boldsymbol{k}}$ and defined $\boldsymbol{\lambda}_{(\boldsymbol{-1})} = \boldsymbol{0}$, so that the relation in (14) holds for $k = 0, ..., N-1$. Additionally, $\partial\mathcal{L}/\partial\hat{\boldsymbol{x}}_{\boldsymbol{N}} = 0$ implies that $\boldsymbol{\lambda}_{\boldsymbol{N-1}} = 0$.

By exploiting the fact that (11) is isometric, all solutions of (14) can be described as,

$$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{bmatrix}\begin{bmatrix}\boldsymbol{\lambda}_{\boldsymbol{k-1}}\\ 0\end{bmatrix} + \begin{bmatrix}\widetilde{\boldsymbol{B}}\\ \widetilde{\boldsymbol{D}}\end{bmatrix}\widetilde{v}_k = \begin{bmatrix}\boldsymbol{\lambda}_{\boldsymbol{k}}\\ \widetilde{\boldsymbol{w}}_{\boldsymbol{k}}\end{bmatrix} \qquad (15)$$

where $\widetilde{\boldsymbol{v}} \in \mathbb{R}^N$ and the matrices $\widetilde{\boldsymbol{B}} \in \mathbb{R}^n$ and $\widetilde{\boldsymbol{D}} \in \mathbb{R}^2$ come from the unitary completion of the behavioral model in (11), such that,

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} & \widetilde{\boldsymbol{B}} \\ \boldsymbol{C} & \boldsymbol{D} & \widetilde{\boldsymbol{D}} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} & \widetilde{\boldsymbol{B}} \\ \boldsymbol{C} & \boldsymbol{D} & \widetilde{\boldsymbol{D}} \end{bmatrix}^\top = \boldsymbol{I}_{n+2}. \tag{16}$$

Substitution of $\widetilde{\boldsymbol{x}}_{\boldsymbol{k}} = \boldsymbol{\lambda}_{\boldsymbol{k-1}}$ for $k = 0, \dots, N-1$, and rewriting (15) leads to the state space representation of the *misfit model*,

$$\begin{bmatrix} \widetilde{\boldsymbol{x}}_{\boldsymbol{k+1}} \\ \widetilde{\boldsymbol{w}}_{\boldsymbol{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \widetilde{\boldsymbol{B}} \\ \boldsymbol{C} & \widetilde{\boldsymbol{D}} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{x}}_{\boldsymbol{k}} \\ \widetilde{v}_k \end{bmatrix}, \quad k = 0, \dots, N-1. \tag{17}$$

Similarly as in Section 2, we can conclude that the optimal misfit is highly structured. In particular, the misfit itself is generated by an nth order LTI model.

Misfit space Equation (17) allows us to derive a basis for the linear subspace in which the misfit $\widetilde{\boldsymbol{w}}_{\boldsymbol{k}}$ resides. Contrary to (13), we do not need to consider the observability matrix of the misfit model because we know that $\widetilde{\boldsymbol{x}}_{\boldsymbol{0}} = \boldsymbol{0}$. That leaves us with the matrix $\widetilde{\boldsymbol{H}} \in \mathbb{R}^{2N \times N}$, which is defined similarly as in (12) but using the matrices $\boldsymbol{A}, \widetilde{\boldsymbol{B}}, \boldsymbol{C}$ and $\widetilde{\boldsymbol{D}}$. However, we cannot take any linear combination of its columns to generate $\widetilde{\boldsymbol{w}}$ as there is an additional constraint that needs to be satisfied: $\widetilde{\boldsymbol{x}}_{\boldsymbol{N}} = 0$. For this constraint to be true, $\widetilde{\boldsymbol{v}}$ needs to be an element of the null space of $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}$, where $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}} \in \mathbb{R}^{n \times N}$ is the reversed extended controllability matrix of the misfit model,

$$\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}} = \begin{bmatrix} \boldsymbol{A}^{N-1}\widetilde{\boldsymbol{B}} & \dots & \boldsymbol{A}\widetilde{\boldsymbol{B}} & \widetilde{\boldsymbol{B}} \end{bmatrix}.$$

Because of the minimality of (11), the matrix $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}$ will be of rank n, such that its null space $\boldsymbol{Z} \in \mathbb{R}^{N \times (N-n)}$,

$$\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}} \boldsymbol{Z} = \boldsymbol{0},$$

is $(N-n)$-dimensional. Finally, we find the basis for the $(N-n)$-dimensional misfit space as range$(\widetilde{\boldsymbol{H}}\boldsymbol{Z})$.

We conclude that the first-order necessary conditions for optimality of the state space projection onto the behavior problem once more induce an orthogonal decomposition of the ambient space,

$$\mathbb{R}^{2N} = \underbrace{(\text{range}(\boldsymbol{H}) \oplus \text{range}(\boldsymbol{\Gamma}))}_{\text{behavior}} \oplus \underbrace{\text{range}(\widetilde{\boldsymbol{H}}\boldsymbol{Z})}_{\text{misfits}}. \tag{18}$$

The orthogonal complementarity of both subspaces follows from the isometric construction of the model (11) and the misfit model (17).

Optimal projection Because of (18), we can express the minimal norm $\widetilde{\boldsymbol{w}}$ as the orthogonal projection of $\boldsymbol{w}$ onto the column space of $\widetilde{\boldsymbol{H}}\boldsymbol{Z}$,

$$\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{H}}\boldsymbol{Z}[(\widetilde{\boldsymbol{H}}\boldsymbol{Z})^\top \widetilde{\boldsymbol{H}}\boldsymbol{Z}]^{-1}(\widetilde{\boldsymbol{H}}\boldsymbol{Z})^\top \boldsymbol{w} = \widetilde{\boldsymbol{H}}\boldsymbol{Z}(\widetilde{\boldsymbol{H}}\boldsymbol{Z})^\top \boldsymbol{w}, \tag{19}$$

where we used the fact that $\widetilde{\boldsymbol{H}}\boldsymbol{Z}$ is a semi-orthogonal matrix, such that: $(\widetilde{\boldsymbol{H}}\boldsymbol{Z})^\top \widetilde{\boldsymbol{H}}\boldsymbol{Z} = \boldsymbol{I}_{\boldsymbol{N-n}}$.

From (17) and the fact that $\widetilde{\boldsymbol{x}}_{\boldsymbol{0}} = \boldsymbol{0}$, we know that $\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{v}}$. Combine this with (19) to see that we can compute the sequence $\widetilde{\boldsymbol{v}}$ which minimizes the 2-norm of $\widetilde{\boldsymbol{w}}$ via,

$$\widetilde{\boldsymbol{v}} = \boldsymbol{Z}\boldsymbol{Z}^\top \widetilde{\boldsymbol{H}}^\top \boldsymbol{w}. \tag{20}$$

The adjoint model There is a alternative way to interpret (20). By combining (11) and (17) and taking $\boldsymbol{x}_{\boldsymbol{k+1}} = \hat{\boldsymbol{x}}_{\boldsymbol{k+1}} + \widetilde{\boldsymbol{x}}_{\boldsymbol{k+1}}$, we get that,

$$\begin{bmatrix} \boldsymbol{x}_{\boldsymbol{k+1}} \\ \boldsymbol{w}_{\boldsymbol{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} & \widetilde{\boldsymbol{B}} \\ \boldsymbol{C} & \boldsymbol{D} & \widetilde{\boldsymbol{D}} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{\boldsymbol{k}} \\ \hat{v}_k \\ \widetilde{v}_k \end{bmatrix}, \quad k = 0, \dots, N-1.$$

We can now use the isometry (16) to derive that the sequences of auxiliary variables $\hat{\boldsymbol{v}}$ and $\widetilde{\boldsymbol{v}}$ can be obtained from a backward recursion [6] using the observed data $\boldsymbol{w}$,

$$\begin{bmatrix} \boldsymbol{x}_k \\ \hat{v}_k \\ \widetilde{v}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}^\top & \boldsymbol{C}^\top \\ \boldsymbol{B}^\top & \boldsymbol{D}^\top \\ \widetilde{\boldsymbol{B}}^\top & \widetilde{\boldsymbol{D}}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{\boldsymbol{k+1}} \\ \boldsymbol{w}_{\boldsymbol{k}} \end{bmatrix}, \quad k = 0, \dots, N-1. \tag{21}$$

Writing out the backward recursion from (21) gives,

$$\widetilde{\boldsymbol{v}} = \widetilde{\boldsymbol{H}}^\top \boldsymbol{w} + (\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^\top \hat{\boldsymbol{x}}_{\boldsymbol{N}}, \tag{22}$$

where we substituted $\boldsymbol{x}_{\boldsymbol{N}}$ with $\hat{\boldsymbol{x}}_{\boldsymbol{N}}$ since $\widetilde{\boldsymbol{x}}_{\boldsymbol{N}} = 0$. Since $\|\widetilde{\boldsymbol{w}}\|_2^2 = \|\widetilde{\boldsymbol{v}}\|_2^2$ due to the isometry, the optimal $\widetilde{\boldsymbol{v}}$, which generates the minimal norm misfit $\widetilde{\boldsymbol{w}}$, can be found by choosing $\hat{\boldsymbol{x}}_{\boldsymbol{N}}$ such that $\|\widetilde{\boldsymbol{v}}\|_2^2$ is minimal,

$$\begin{aligned} \boldsymbol{0} &= \frac{\partial \|\widetilde{\boldsymbol{v}}\|_2^2}{\partial \hat{\boldsymbol{x}}_{\boldsymbol{N}}} \\ \iff \boldsymbol{0} &= \widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^\top \hat{\boldsymbol{x}}_{\boldsymbol{N}} + \widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}} \widetilde{\boldsymbol{H}}^\top \boldsymbol{w} \\ \iff \hat{\boldsymbol{x}}_{\boldsymbol{N}} &= -(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^\top)^{-1} \widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}} \widetilde{\boldsymbol{H}}^\top \boldsymbol{w}, \end{aligned} \tag{23}$$

where we used that,

$$\|\widetilde{\boldsymbol{v}}\|_2^2 = \boldsymbol{w}^\top \widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{H}}^\top \boldsymbol{w} + \hat{\boldsymbol{x}}_{\boldsymbol{N}}^\top \widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^\top \hat{\boldsymbol{x}}_{\boldsymbol{N}} + 2\boldsymbol{w}^\top \widetilde{\boldsymbol{H}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^\top \hat{\boldsymbol{x}}_{\boldsymbol{N}}.$$

Substituting (23) in (22) shows that the optimal $\widetilde{\boldsymbol{v}}$ is equal to,

$$\begin{aligned} \widetilde{\boldsymbol{v}} &= \widetilde{\boldsymbol{H}}^\top \boldsymbol{w} - (\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^\top (\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^\top)^{-1} \widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}} \widetilde{\boldsymbol{H}}^\top \boldsymbol{w}, \\ &= \left(\boldsymbol{I} - (\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^\top (\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^\top)^{-1} \widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}\right) \widetilde{\boldsymbol{H}}^\top \boldsymbol{w}. \end{aligned}$$

Thus, the optimal $\widetilde{\boldsymbol{v}}$ can be obtained as the orthogonal projection of the backwards-filtered observed data $\boldsymbol{w}$ onto range$(\boldsymbol{Z})$, which corresponds to the orthogonal complement of row$(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})$, leading to the same conclusion as (20).

The z-domain It is straightforward to show that when the isometric model in (11) describes the same dynamics as the input-output relation in (1), the transfer matrix $\hat{\boldsymbol{H}}(z)$ of the isometric model (11), mapping the z-transform of the auxiliary inputs $\boldsymbol{v}$ to the z-transform of the model-compliant trajectory $\hat{\boldsymbol{w}}$, is given as,

$$\hat{\boldsymbol{H}}(z) = \frac{c}{d(z)} \begin{bmatrix} b(z) \\ a(z) \end{bmatrix},$$

where $c \in \mathbb{R}$ and $d(z)$ a monic polynomial of degree n, satisfying $d(z)d_r(z) = c^2(a_r(z)a(z) + b_r(z)b(z))$. The latter ensures that $\hat{\boldsymbol{H}}(z)$ is *lossless*, in the sense that,

$$\langle \hat{\boldsymbol{H}}(z), \hat{\boldsymbol{H}}(z) \rangle = 1,$$

with the inner-product defined as in (10). It now comes as no surprise that, similarly as in Section 2, the 'reversed polynomial' relation between the z-transforms of $\widetilde{\boldsymbol{y}}$ and $\widetilde{\boldsymbol{u}}$ is also found when one constructs the transfer matrices of the state space models discussed above. The transfer

[6] We call the state space models

$$(\boldsymbol{A}^\top, \boldsymbol{C}^\top, \boldsymbol{B}^\top, \boldsymbol{D}^\top) \quad \text{and} \quad (\boldsymbol{A}^\top, \boldsymbol{C}^\top, \widetilde{\boldsymbol{B}}^\top, \widetilde{\boldsymbol{D}}^\top),$$

the *adjoint model* and *adjoint misfit model*, respectively.

function of the misfit model (17) follows from the *lossless embedding* $\overline{\boldsymbol{H}}(z)$ [Genin et al., 1983] of the rational matrix function $\hat{\boldsymbol{H}}(z)$,

$$\overline{\boldsymbol{H}}(z) = \begin{bmatrix} \hat{\boldsymbol{H}}(z) & \widetilde{\boldsymbol{H}}(z) \end{bmatrix} = \frac{c}{d(z)} \begin{bmatrix} b(z) & a_r(z) \\ a(z) & -b_r(z) \end{bmatrix},$$

such that,

$$\langle \overline{\boldsymbol{H}}(z), \overline{\boldsymbol{H}}(z) \rangle = \boldsymbol{I}_2.$$

Since $\widetilde{\boldsymbol{H}}(z)$ describes the mapping from the auxiliary variables to the misfit $\widetilde{\boldsymbol{w}}$, the mapping from $\widetilde{\boldsymbol{u}}$ to $\widetilde{\boldsymbol{y}}$ can be derived as:

$$\begin{bmatrix} \widetilde{Y}(z) \\ \widetilde{U}(z) \end{bmatrix} = \widetilde{\boldsymbol{H}}(z)\widetilde{V}(z) = \frac{c}{d(z)} \begin{bmatrix} a_r(z) \\ -b_r(z) \end{bmatrix} \widetilde{V}(z)$$

$$\iff \widetilde{Y}(z) = -\frac{a_r(z)}{b_r(z)} \widetilde{U}(z),$$

which agrees with our findings from Section 2.

4. CONCLUSIONS AND FUTURE WORK

In this paper we considered the projection onto the behavior for linear time-invariant (LTI) single-input single-output (SISO) models, which arises in the context of misfit modeling. We characterized the orthogonal decomposition of the ambient data space, which is induced by the least squares optimality criterion, using banded (block-) Toeplitz matrices (Section 2). Then, in Section 3, we showed that similar results are obtained when an (isometric) behavioral state space representation is used to model the SISO LTI dynamics instead of the difference equation that was used in Section 2, thereby establishing two equivalent (up to a change in the model parametrization) descriptions of the optimal solution. Depending on the particular use-case, one of these two alternative formulations can be preferred: e.g., the difference equation considered in Section 2 has, if one fixes $a_0 = 1$, a minimal number of parameters $(2n+1)$ which can be favorable in the context of optimization. On the other hand, whereas a state space model is inherently overparametrized, the characterization developed in Section 3 allows for a more intuitive interpretation of the results, compared to the formulation given in Section 2.

Due to the non-convex nature of the misfit modeling problem (2), the solution approaches described in the literature [De Moor, 1993, 1994, Roorda, 1995, Markovsky et al., 2005, 2006] are heuristic in the sense that convergence to the globally optimal model(s) can generally not be guaranteed. In future work, we want to exploit the results of this paper to develop a methodology that allows to identify the globally optimal solutions of (2), thereby generalizing results from [De Moor, 2020], where it has been shown that the globally optimal solution(s) of the misfit modeling problem for autonomous LTI models can be retrieved via a (large) multiparameter eigenvalue problem (MEVP), to the SISO case. Additionally, since the characterization obtained in [Roorda and Heij, 1995] holds for the broader class of MIMO models, we believe that one must be able to generalize the 'banded block-Toeplitz' framework presented in Section 2 to this broader class of models.

REFERENCES

Aoki, M. and Yue, P.C. (1970). On a priori error estimates of some identification methods. *IEEE Transactions on Automatic Control*, 15(5), 541–548.

De Moor, B. (1993). Structured total least squares and L_2 approximation problems. *Linear Algebra an Its Applications*, 188-189, 163 – 205.

De Moor, B. (1994). Dynamic total linear least squares. In *Proc. of the 10th IFAC Symposium on System Identification*, volume 3, 159 – 164. Copenhagen, Denmark.

De Moor, B. (2020). Least squares optimal realisation of autonomous LTI systems is an eigenvalue problem. *Communications in Information and Systems*, 20(2), 163 – 207.

De Moor, B. and Roorda, B. (1994). L_2-optimal linear system identification: Structured total least squares for SISO systems. In *Proc. of the 33rd IEEE Conference on Decision and Control (CDC)*, 2874 – 2879. Lake Buena Vista, Florida, USA.

Genin, Y., Van Dooren, P., Kailath, T., Delsome, J.M., and Morf, M. (1983). On σ-lossless transfer functions and related questions. *Linear Algebra and its Applications*, 50, 251 – 275.

Heij, C. (1989). *Deterministic Identification of Dynamical Systems*, volume 127 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag Berlin Heidelberg.

Lagauw, S., Van Poucke, L., and De Moor, B. (2023). Exact characterization of the global optima of least squares realization of autonomous LTI models as a multiparameter eigenvalue problem. Technical report, ESAT-STADIUS, KU Leuven. `https://ftp.esat.kuleuven.be/pub/stadius/slagauw/23-140.pdf`.

Legat, B., Yuan, C., and Parrilo, P. (2023). Low-rank univariate sum of squares has no spurious local minima. *SIAM Journal on Optimization*, 33(3), 2041–2061.

Lemmerling, P. and De Moor, B. (2001). Misfit versus latency. *Automatica*, 37, 2057 – 2067.

Ljung, L. (1999). *System Identification: Theory for the User*. Prentice-Hall Information and System Sciences Series. Prentice-Hall, Upper Saddle River, NJ, USA, 2nd edition.

Markovsky, I., Willems, J.C., Van Huffel, S., and De Moor, B. (2006). *Exact and Approximate Modeling of Linear Systems: A Behavioral Approach*. Monographs on Mathematical Modeling and Computation. SIAM.

Markovsky, I., Willems, J.C., Van Hufffel, S., De Moor, B., and Pintelon, R. (2005). Application of structured total least squares for system identification. *IEEE Transactions on Automatic Control*, 50(10), 1490–1500.

Roorda, B. (1995). Algorithms for global total least square modelling of finite multivariable time series. *Automatica*, 31(3), 391–404.

Roorda, B. and Heij, C. (1995). Global total least squares modeling of multivariable time series. *IEEE Transactions on Automatic Control*, 40(1), 50 – 63.

Wahlberg, B. (2003). Orthogonal rational functions: A transformation analysis. *SIAM Review*, 45(4), 689 – 705.

Willems, J.C. (1986). From time series to linear system– part I. Finite dimensional linear time invariant systems. *Automatica*, 22(5), 561 – 580.

Willems, J.C. (1987). From time series to linear system– part III. Approximate modelling. *Automatica*, 23(1), 87 – 115.