Data-driven structured noise filtering via common dynamics estimation

Markovsky, Ivan; Liu, Tianxiang; Takeda, Akiko

Published in:
IEEE Transactions on Signal Processing

DOI:
10.1109/TSP.2020.2993676

Publication date:
2020

Document Version:
Submitted manuscript

Data-driven structured noise filtering via common dynamics estimation

Ivan Markovsky, Tianxiang Liu, and Akiko Takeda

***Abstract*—Classical signal from noise separation problems assume that the signal is a trajectory of a low-complexity linear time-invariant system and that the noise is a random process. In this paper, we generalize this classical setup to what we call data-driven structured noise filtering. In the new setup, the noise has two components: structured noise, which is also a trajectory of a low-complexity linear time-invariant system, and unstructured noise, which is a zero-mean white Gaussian process. The key assumption that makes the separation problem in the new setup well posed is that among several experiments, the signal's dynamics remains the same while the structured noise's dynamics varies. The data-driven structured noise filtering problem then becomes a problem of estimation of common linear time-invariant dynamics among several observed signals. We show that this latter problem is a structured low-rank approximation problem with multiple rank constraints and use a subspace identification approach for solving it. The resulting methods allow computationally efficient and numerically robust implementation and have the system theoretic interpretation of finding the intersection of autonomous linear time-invariant behaviors. Statistical analysis providing confidence bounds is a topic for future research.**

***Index Terms*—Hankel structured low-rank approximation, Subspace system identification, Behavioral approach.**

I. Introduction

The prototypical signal processing problem of signal from noise separation is ill-posed, *i.e.*, it has a nonunique solution unless prior information is given. The prior information is expressed in the form of assumptions about the signal and the noise. Different signal from noise separation methods are developed for different signal and noise assumptions. Stronger assumptions lead to more accurate but less general methods.

Despite the diversity of the assumptions, in all problems the signal is in some sense predictable while the noise is unpredictable. This is justified in practice by the fact that the signal satisfies natural laws which allow for its deterministic modeling while the noise is poorly understood which allows only a statistical description. Natural laws, such as the Newton's second law of dynamics in mechanical engineering and RLC-circuits in electrical engineering, are often given by linear constant coefficients differential equations. Therefore, the classical signal from noise separation setup assumes that the signal is a trajectory of a linear time-invariant system while the noise is a random process.

I. Markovsky is with the Department ELEC, Vrije Universiteit Brussel, 1050 Brussels, Belgium (e-mail: ivan.markovsky@vub.be)

T. Liu is with the RIKEN Center for Advanced Intelligence Project, Nihonbashi, Chuo-ku, Tokyo 103–0027, Japan, (e-mail: tianxiang.liu@riken.jp)

A. Takeda is with the Department of Creative Informatics, The University of Tokyo, Tokyo, 113-8656, Japan, (e-mail: takeda@mist.i.u-tokyo.ac.jp)

The classical setup declares that anything that is deterministically predictable is a signal. This is restrictive. Indeed, there are deterministic components in the data such as offsets, trends, and periodic disturbances that are not part of the signal but are also responses of linear time-invariant systems. Traditionally, in system identification [1] these components are removed in a preprocessing step. As pointed out in [2] and further elaborated in [3], this leads to inferior results than modeling them together with the system dynamics. In this paper, we generalize the classical signal from noise separation setup in order to have deterministic noise component.

The generalization proposed deals with noise consisting of two components: a trajectory of a low-complexity autonomous linear time-invariant system (*structured noise*) and a zero-mean white Gaussian process (*unstructured noise*). The structured noise is therefore a sum of polynomials-times-damped-complex-exponentials signal. The key assumptions that make separability in this new setup possible are:

- at least two data collection experiments are performed, and
- in different experiments, the structured noise models have different poles.

A generic example that shows how the assumptions occur in practice is approximation of a nonlinear system by a linear time-invariant model. Contrary to the classical assumption in system identification that the error is a stochastic process, the error of approximating the nonlinear system by a linear time-invariant model is deterministic. Moreover, it depends on the experimental conditions. In the case of autonomous systems, the experimental conditions are determined by the initial conditions. In different experiments when the system is excited by different initial conditions, the resulting error signals have different deterministic dynamics. However, the linear time-invariant model remains the same in all experiments. This leads to the problem of structured noise filtering (take the linear model as the "true signal dynamics" and the approximation error as the structured noise). In general, the structured noise has no linear time-invariant dynamics, however, in special cases, *e.g.*, when the nonlinear system is a Wiener system, it does [4]. When the structured noise has no linear time-invariant dynamics, it can be approximated by a linear time-invariant system, attributing the remaining approximation error to unstructured noise since it is unmodeled.

We show that the maximum likelihood estimation problem in the new setup is a structured low-rank approximation problem with multiple rank constraints. This is a nonconvex optimization problem, for which an analytical solution is

not known. Local optimization methods require an initial approximation obtained by alternative direct methods. The aim of this paper is to develop such direct methods, using results from the behavioral system theory and subspace identification. First, we consider the data-driven structured noise filtering problem when there is no unstructured noise. In this case, under the assumption that the structured noise models have no common poles and the true signals are persistently exciting of sufficiently high order, the methods developed in this paper achieve exact signal from noise separation. The observed signals' common dynamics is the true signals' dynamics. Computing the common dynamics from models of the observed signals is a greatest common divisor computational problem [5], [6], [7]. The method developed in the paper

1) identifies models for the observed signals and
2) computes the common dynamics of the models.

The generalization of the method to the case of unstructured as well as structured noise is done heuristically by incorporating approximation in each step and using prior knowledge about the data generating systems' orders. Using different methods for approximate model identification and approximate common divisor computation, we obtain a variety of methods for data-driven structured noise filtering.

Apart from data-driven structured noise filtering, the common dynamics problem occurs in biomedical signal processing [8], monitoring of material structures [9], and audio modeling [10], [11]. Methods for common dynamics estimation are proposed in [12], [13]. Note that the data may be collected in a single multi-channel experiment as well as in multiple experiments. For example, a real-life application, considered in [14], that leads to structured noise filtering is multi-channel EEG seizure detection. In this application, the aim is to retrieve the common epileptic seizure information among the recorded EEG channels, taking into account the fact that each channel may be affected by different artifact sources: muscle artefacts, eye blink artefacts, respiration artefacts, *etc*. This leads to different disturbance dynamics in the different channels.

To the best of our knowledge, the data-driven structured noise filtering problem considered in the paper is new. Note that the classical methods for sum-of-exponentials modeling, spectral estimation, and latent variable modeling, such as MUSIC [15], ESPRIT [16], and dynamic PCA [17] are not applicable to the data-driven structured noise filtering problem. Indeed, these classical methods correspond to step 1 of the data-driven structured noise filtering method developed in the paper. Without imposing the constraint that the observed signals have common dynamics (the true signal's dynamics) as well as different dynamics (the structured noise's dynamics), the identified poles can not be separated into signal poles and structured noise poles.

The main contributions of the paper are:

- a novel signal from noise separation setup, called data-driven structured noise filtering, where the noise has deterministic as well as stochastic components,
- identifiability conditions (Theorem 8) and equivalence of the maximum likelihood estimation problem in the new setup to a structured low-rank approximation problem with multiple rank constraints (Theorems 9 and 10),
- a class of subspace methods for solving the data-driven structured noise filtering problem (Section VI).

In Section II, we illustrate the data-driven structured noise filtering problem by a numerical example. For the problem formulation and derivation of solution methods, we use the behavioral approach to system theory [18], [19], [20]. The main difference between the behavioral and the classical approaches is that the model is viewed as a set of signals rather than an equation (such as a difference equation). This makes the behavioral approach particularly convenient for model-free signal reconstruction. The necessary background, notation, and basic results are given in Section III. Section IV presents the classical signal from noise separation setup and shows its connection to Hankel structured low-rank approximation [21], [22], [23], [3]. The maximum likelihood estimation problem in the new setup is defined in Section V and is shown to be equivalent to a generalized structured low-rank approximation problem. In Section VI, we present a general subspace approach for solving the problem. In the absence of unstructured noise, the method yields the exact data generating system; however, in the presence of unstructured noise it becomes a heuristic for solving the maximum likelihood estimation problem. The performance of the subspace methods resulting from the general subspace approach is empirically evaluated in Section VII. Conclusions and directions for future work are given in Section VIII.

II. Illustrative example

In this section, we show a numerical example of the data-driven structured noise filtering problem considered in the paper. Doing data collection experiments, we obtain a set of scalar discrete-time signals (the data)

$$y_i = \big(y_i(1), \ldots, y_i(T_i)\big), \quad \text{for } i = 1, \ldots, N.$$

In the example, $N = 2$ experiments are done and the signals have $T_1 = T_2 = 1000$ samples. The signals

$$y_i = s_i + d_i + e_i, \quad \text{for } i = 1, \ldots, N, \tag{1}$$

are sums of three components:

- s_i is the *true signal*

$$s_1(t) = 2\sin(0.05t + 100) + \sin(0.03t - 100)$$
$$s_2(t) = 2\sin(0.05t - 150) - \sin(0.03t + 50)$$

- d_i is the *structured noise*

$$d_1(t) = -2\sin(0.04t - 150)$$
$$d_2(t) = \quad 2\sin(0.07t + 150)$$

- e_i is the *unstructured noise* $e_1, e_2 \sim \mathrm{N}(0, \varsigma^2 I)$ that is zero-mean white Gaussian with standard deviations $\varsigma = 1$.

The aim of the data-driven structured noise filtering problem is to recover the true signals s_1, s_2 from the measurements y_1, y_2 and the prior knowledge that

- the data generating system of the true signals is autonomous linear time-invariant of order $n_{\mathrm{s}} = 4$,

- the data generating systems of the structured noises are autonomous linear time-invariant of order $n_{\mathrm{d}} = 2$, and
- the unstructured noises are zero-mean white Gaussian.

Figure 1, first column, shows the noisy data y_1, y_2 (solid red lines) and the approximation of the data (dashed blue lines), obtained with the subspace method presented in Section VI. The approximation is a sum of the true signal's estimate, the structured noise's estimate, and a residual. The second column in Figure 1 shows the true signals s_1, s_2 (solid red lines) and the corresponding approximations (dashed blue lines). Despite the low signal-to-noise ratio (SNR=-0.5dB), the fit of the true signals is good. The same good fit is observed in the third column that shows the true structured noise signals d_1, d_2 (solid red lines) and the approximations (dashed blue lines). The fourth column shows the unstructured noises e_1, e_2 (solid red lines) and the residuals (dashed blue lines).

III. Notation and preliminaries

We use the behavioral approach to systems theory [18], [19], [20]. A discrete-time dynamical system is a set $\mathscr{B}$ of signals $y = \big(y(1), y(2), \dots\big)$. (Contrast this to the classical approach that invariably defines the system by an equation.) The set $\mathscr{B}$ is called the *behavior* of the system that it describes. Since the behavior specifies completely the system, we will refer to $\mathscr{B}$ as the *system*. The notation $y \in \mathscr{B}$ is a convenient way of saying that the signal y is a trajectory of the system $\mathscr{B}$. It replaces writing the equation defining the system in the classical setting.

The restriction of the signal y to the interval $[1, L]$ is

$$y|_L := \big(y(1), \dots, y(L)\big).$$

Similarly, $\mathscr{B}|_L$ is the restriction of $\mathscr{B}$ to the interval $[1, L]$,

$$\mathscr{B}|_L := \{\, y|_L \mid y \in \mathscr{B} \,\}.$$

In this paper, we consider scalar linear time-invariant systems. The class of all such systems is denoted by $\mathscr{L}$. Let σ be the *shift operator*

$$(\sigma y)(t) := y(t+1), \quad \text{for all } t.$$

Acting on $\mathscr{B}$, σ shifts all signals in $\mathscr{B}$, $\sigma\mathscr{B} = \{\, \sigma y \mid y \in \mathscr{B} \,\}$. By definition, a system $\mathscr{B}$ is linear if $\mathscr{B}$ is a subspace and time-invariant if $\sigma\mathscr{B} = \mathscr{B}$. The dimension $\dim \mathscr{B}$ of $\mathscr{B} \in \mathscr{L}$ is the *order* $\mathtt{n}(\mathscr{B})$ of $\mathscr{B}$. The order is a measure of the system's *complexity*. The subclass of $\mathscr{L}$ consisting of systems with complexity bounded by n is denoted by $\mathscr{L}_n$. The statement "$\mathscr{B}$ is a scalar linear time-invariant system of order bounded by n" is then concisely written as $\mathscr{B} \in \mathscr{L}_n$.

A system $\mathscr{B} \in \mathscr{L}$ is an $\mathtt{n}(\mathscr{B})$-dimensional subspace. Similarly, $\mathscr{B}|_L$ is a subspace of $\mathbb{R}^L$. Its dimension is

$$\dim \mathscr{B}|_L = \begin{cases} L & \text{if } L \leq \mathtt{n}(\mathscr{B}) \\ \mathtt{n}(\mathscr{B}) & \text{if } L \geq \mathtt{n}(\mathscr{B}). \end{cases}$$

The system $\mathscr{B} \in \mathscr{L}_n$ admits different *representations*—state-space; polynomial, also called kernel; poles, also called sum-of-polynomials-times-damped-exponentials; *etc*. Representations bring parameterizations of the system and are unavoidable in computational methods. The *kernel representation* [18],

$$\mathscr{B} = \ker p(\sigma) := \{\, y \mid p_0 y + p_1 \sigma y + \cdots + p_n \sigma^n y = 0 \,\}, \quad (2)$$

is defined by a scalar univariate polynomial

$$p(z) = p_0 + p_1 z + \cdots + p_n z^n$$

with coefficients vector $p := \begin{bmatrix} p_0 & p_1 & \cdots & p_n \end{bmatrix} \neq 0$. The roots $z_1, \dots, z_n$ of $p(z)$ are invariant of the representation and are called the *poles* of the system $\mathscr{B}$. The set of the poles of $\mathscr{B} \in \mathscr{L}_n$ is denoted by $\lambda(\mathscr{B})$.

Identification of linear time-invariant systems and processing of signals that are trajectories of such systems is closely related to low-rank approximation of Hankel matrices [24], [25]. The *Hankel matrix* with $L \leq T$ rows, constructed from the signal $y = \big(y(1), \dots, y(T)\big)$ is denoted by

$$\mathscr{H}_L(y) := \begin{bmatrix} y(1) & y(2) & \cdots & y(T-L+1) \\ y(2) & y(3) & \cdots & y(T-L+2) \\ \vdots & \vdots & & \vdots \\ y(L) & y(L+1) & \cdots & y(T) \end{bmatrix}. \quad (3)$$

Of particular interest is whether the matrix $\mathscr{H}_L(y)$ is full row rank. This property is important enough to be given a name.

Definition 1 (persistency of excitation)**.** A signal y is *persistently exciting* of order L if $\operatorname{rank} \mathscr{H}_L(y) = L$. More generally, a set of signals $y = \{y_1, \dots, y_N\}$ is persistently exciting of order L if

$$\operatorname{rank} \begin{bmatrix} \mathscr{H}_L(y_1) & \cdots & \mathscr{H}_L(y_N) \end{bmatrix} = L. \quad (4)$$

The importance of the Hankel matrix $\mathscr{H}_L(y)$ in identification and signal processing stems from the following result.

Lemma 2. *Let $y = \big(y(1), \dots, y(T)\big)$ be a trajectory of a linear time-invariant system $\mathscr{B} \in \mathscr{L}_n$,* i.e.*, $y \in \mathscr{B}|_T$. Then, the image of the Hankel matrix $\mathscr{H}_L(y)$ is a subset of $\mathscr{B}|_L$,* i.e.*,*

$$\operatorname{image} \mathscr{H}_L(y) \subseteq \mathscr{B}|_L. \quad (5)$$

Equality holds in (5) when y is persistently exciting of order $\mathtt{n}(\mathscr{B})$ and $L \leq T - \mathtt{n}(\mathscr{B})$. More generally, let the signals $y_1, \dots, y_N$ be trajectories of a linear time-invariant system $\mathscr{B} \in \mathscr{L}_n$, i.e.*, $y_i \in \mathscr{B}|_{T_i}$, for $i = 1, \dots .N$. Then,*

$$\operatorname{image} \begin{bmatrix} \mathscr{H}_L(y_1) & \cdots & \mathscr{H}_L(y_N) \end{bmatrix} \subseteq \mathscr{B}|_L. \quad (6)$$

Proof. By the time-invariance property of $\mathscr{B}$, the columns of the Hankel matrix $\mathscr{H}_L(y)$, viewed as time series, are L-samples long trajectories of $\mathscr{B}$. Then by the linearity property of $\mathscr{B}$, a linear combination of the columns of $\mathscr{H}_L(y)$ is also an element of $\mathscr{B}_L$. This proves (5).

Under the persistency of excitation assumption,

$$\dim \operatorname{image} \mathscr{H}_L(y) = \begin{cases} L & \text{for } L < \mathtt{n}(\mathscr{B}) \\ \mathtt{n}(\mathscr{B}) & \text{for } \mathtt{n}(\mathscr{B}) \leq L \leq T - \mathtt{n}(\mathscr{B}). \end{cases}$$

On the other hand, since $\mathscr{B} \in \mathscr{L}$, $\dim \mathscr{B}|_L = L$, for $L \leq \mathtt{n}(\mathscr{B})$, and $\dim \mathscr{B}|_L = \mathtt{n}(\mathscr{B})$, for $L \geq \mathtt{n}(\mathscr{B})$, so that

$$\operatorname{image} \mathscr{H}_L(y) = \mathscr{B}|_L, \quad \text{for } L \leq T - \mathtt{n}(\mathscr{B}).$$

The generalization (6) for multiple trajectories follows from (5) and the fact that

$$\operatorname{image} \begin{bmatrix} \mathscr{H}_L(y_1) & \cdots & \mathscr{H}_L(y_N) \end{bmatrix} = \operatorname{image} \mathscr{H}_L(y_1) + \cdots + \operatorname{image} \mathscr{H}_L(y_N).$$

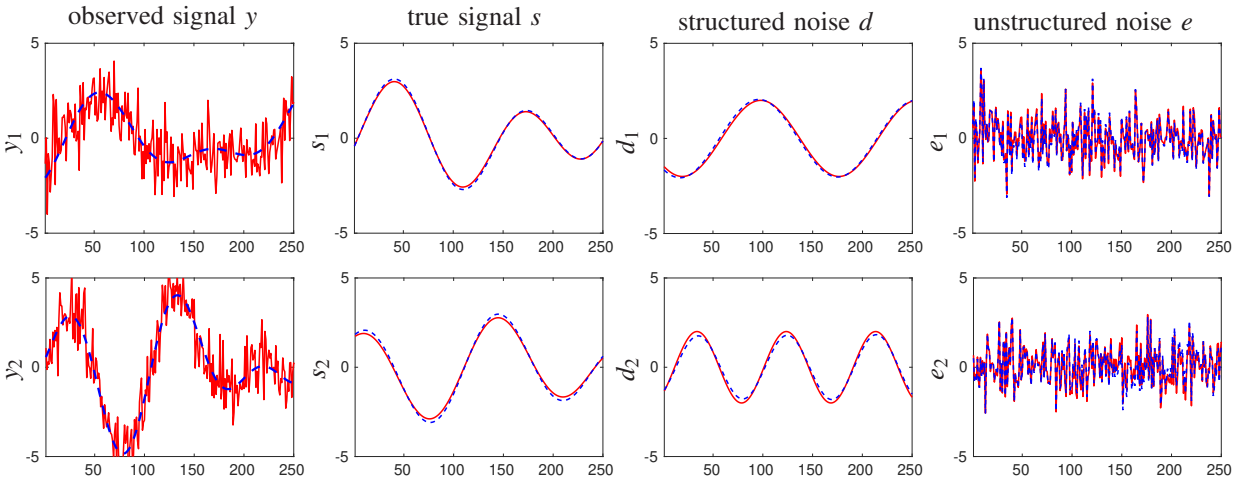

Fig. 1. Numerical illustration of the signal from structured noise separation problem. (solid red lines — observed and true signals, dashed blue lines — estimates obtained with the subspace method presented in Section VI.)

□

As a corollary of Lemma 2, we obtain a simple test for whether a given signal is a trajectory of a linear time-invariant system of bounded complexity by computing the rank of a Hankel matrix $\mathscr{H}_L(y)$.

Corollary 3. *The signal $y = \big(y(1),\ldots,y(T)\big)$ is a trajectory of a linear time-invariant system $\mathscr{B} \in \mathscr{L}_n$,* i.e., *$y \in \mathscr{B}|_T$ for some $\mathscr{B} \in \mathscr{L}_n$, if and only if $\operatorname{rank}\mathscr{H}_L(y) \leq n$ for some L such that $n+1 \leq L \leq T-n$. More generally, the signals $y_1,\ldots,y_N$ are trajectories of a linear time-invariant system $\mathscr{B} \in \mathscr{L}_n$,* i.e., *$y_i \in \mathscr{B}|_T$ for some $\mathscr{B} \in \mathscr{L}_n$ and for all $i = 1,\ldots,N$, if and only if*

$$\operatorname{rank}\begin{bmatrix}\mathscr{H}_L(y_1) & \cdots & \mathscr{H}_L(y_N)\end{bmatrix} \leq n, \tag{7}$$

for some L such that

$$n+1 \leq L \leq \textstyle\sum_{i=1}^{N}(T_i - n). \tag{8}$$

The inequality $\operatorname{rank}\mathscr{H}_L(y) \leq n$ is a fundamental relation between the rank of the Hankel matrix constructed from the data and the order of an exact linear time-invariant model for the data. For $L = 1,\ldots,n$ and $L = T-n+1,\ldots,T$, $\operatorname{rank}\mathscr{H}_L(y) \leq n$, irrespective of the data y. Otherwise, $\operatorname{rank}\mathscr{H}_L(y) \leq n$ if and only if $y \in \mathscr{B}|_T$ for a system $\mathscr{B} \in \mathscr{L}_n$.

Note 4 (On the choice of L). Using Corollary 3 for checking numerically whether $y \in \mathscr{B}|_T$ for some $\mathscr{B} \in \mathscr{L}_n$ requires choosing a priori the value of L. Any value in the range (8) is allowed, however; the boundary values $L = n+1$ and $L = \sum_{i=1}^{N}(T_i - n)$ are advantageous from a numerical computation point of view. For example, $L = n+1$ leads to the smallest number of variables when using a left kernel representation

$$p\begin{bmatrix}\mathscr{H}_L(y_1) & \cdots & \mathscr{H}_L(y_N)\end{bmatrix} = 0, \quad \text{where } p \in \mathbb{R}^{1\times(n+1)} \text{ and } p \neq 0.$$

For this reason, the value $L = n+1$ is used in the maximum likelihood optimization problems (10), (16), and (17).

By Corollary 3, a signal that is persistently exciting of order $n+1$ has no exact linear time-invariant model of order n or less. Vice versa, rank deficiency of the Hankel matrix $\mathscr{H}_L(y)$, *i.e.*, lack of persistency of excitation, implies the existence of an exact linear time-invariant model of bounded complexity.

The identity matrix is denoted by I. A zero-mean Gaussian process e with covariance matrix $\varsigma^2 I$ is denoted by $e \sim \mathrm{N}(0,\varsigma^2 I)$. With probability one, a realization $e = \big(e(1),\ldots,e(T)\big)$ of a Gaussian process e with nonsingular covariance matrix is persistently exciting of order $\lfloor T/2 \rfloor$, where $\lfloor a \rfloor$ is the largest integer smaller than a. Throughout the paper $\|\cdot\|$ denotes the 2-norm.

IV. Classical setup: zero-mean white Gaussian noise

Let y be the observed signal, s its true value, and e the measurement noise. The classical assumptions are that s is generated by a low-complexity linear time-invariant system and e is zero-mean white Gaussian process. Using the notation introduced in Section III, the classical data generating model is concisely written as

$$y = s + e, \quad \text{where} \quad s \in \mathscr{B}_{\mathrm{s}} \in \mathscr{L}_{n_{\mathrm{s}}} \text{ and } e \sim \mathrm{N}(0,\varsigma^2 I). \tag{9}$$

The maximum likelihood estimation problem for (9) is: Given y and n_{s}, find estimates $\widehat{s}$ of s and $\widehat{\mathscr{B}}_{\mathrm{s}}$ of $\mathscr{B}_{\mathrm{s}}$ as a solution of the optimization problem

$$\begin{aligned} &\text{minimize} \quad \text{over } \widehat{s} \text{ and } \widehat{\mathscr{B}}_{\mathrm{s}} \quad \|y - \widehat{s}\| \\ &\text{subject to} \quad \widehat{s} \in \widehat{\mathscr{B}}_{\mathrm{s}} \in \mathscr{L}_{n_{\mathrm{s}}}. \end{aligned} \tag{10}$$

In applications of simulation, prediction, and control, we are interested in the model $\widehat{\mathscr{B}}_{\mathrm{s}}$ rather than the signal $\widehat{s}$. In the case of noise free data, *i.e.*, $e = 0$ so that $y = s$, the signal from noise separation problem has a trivial solution $\widehat{s} = y$. Nevertheless, the data modeling problem $y \mapsto \widehat{\mathscr{B}}_{\mathrm{s}}$ is meaningful and nontrivial. The conditions under which the data generating system $\mathscr{B}_{\mathrm{s}}$ can be recovered back from the data y, *i.e.*, $\widehat{\mathscr{B}}_{\mathrm{s}} = \mathscr{B}_{\mathrm{s}}$, are called *identifiability conditions*.

Proposition 5. *If the true signal s is persistently exciting of order n_s and the data is exact,* i.e., *$y = s \in \mathscr{B}_s$, the solution $\widehat{\mathscr{B}}_s$ of (10) is the true system $\mathscr{B}_s$,* i.e., *$\widehat{\mathscr{B}}_s = \mathscr{B}_s$.*

Proof. By construction, the pair

$$(\widehat{s} = s,\ \widehat{\mathscr{B}}_{\mathrm{s}} = \mathscr{B}_{\mathrm{s}}) \tag{11}$$

satisfies the constraint of (10), *i.e.*, it is feasible. In the case of exact data, $y = s = \widehat{s}$ so that (11) achieves zero approximation error $\|y - \widehat{s}\| = 0$. Therefore, it is an optimal solution. Finally, by the persistency of excitation assumption there is no feasible solution in $\mathscr{L}_{n_{\mathrm{s}}-1}$, so that (11) is the unique solution of (10). □

As an application of Corollary 3 to the maximum likelihood problem (10), we have the following proposition.

Proposition 6. *The maximum likelihood estimation problem (10) is equivalent to the Hankel structured low-rank approximation problem*

$$\begin{aligned} &\textit{minimize} \quad \textit{over } \widehat{s} \quad \|y - \widehat{s}\| \\ &\textit{subject to} \quad \operatorname{rank} \mathscr{H}_{n_s+1}(\widehat{s}) \leq n_s. \end{aligned} \tag{12}$$

Proof. With $L = n_{\mathrm{s}} + 1$, by Corollary 3, $\widehat{s} \in \widehat{\mathscr{B}}_{\mathrm{s}} \in \mathscr{L}_{n_{\mathrm{s}}}$ if and only if $\operatorname{rank} \mathscr{H}_{n_{\mathrm{s}}+1}(\widehat{s}) \leq n_{\mathrm{s}}$. □

The Hankel structured low-rank approximation problem (12) is a nonconvex optimization problem, so it requires iterative optimization methods. An alternative suboptimal solution approach is subspace identification [26]. Subspace identification methods can be used to provide an initial approximation for local optimization methods.

The subproblem of (10), where the model $\widehat{\mathscr{B}}_{\mathrm{s}}$ is given,

$$\begin{aligned} &\text{minimize} \quad \text{over } \widehat{s} \quad \|y - \widehat{s}\| \\ &\text{subject to} \quad \widehat{s} \in \widehat{\mathscr{B}}_{\mathrm{s}}, \end{aligned} \tag{13}$$

is called the *smoothing problem*. An efficient way to solve (13) is the Kalman filter, which effectively employs a state space representation of $\widehat{\mathscr{B}}_{\mathrm{s}}$. Alternatively, the Kalman filter can be viewed as an estimation method for the initial condition. In case of a noise free data, (13) has the trivial solution $\widehat{s} = y$. However, the initial state estimation problem is nontrivial and meaningful; in fact, it is the observer design problem [27].

Note 7 (Order selection in the classical setup). In the statement of the maximum likelihood estimation problem (10), it is assumed that the order n_{s} of the true system $\mathscr{B}_{\mathrm{s}}$ is given. Therefore, n_{s} should be known a priori or estimated from the data y in advance. The estimation of n_{s} is a rank estimation problem. Indeed, by the persistency of excitation assumption and Corollary 3, we have that $\operatorname{rank} \mathscr{H}_L(s) = n_{\mathrm{s}}$ for all L satisfying (8). A classical approach for rank estimation is to compute the singular values of the Hankel matrix $\mathscr{H}_{\lceil T/2 \rceil}(y)$ and take as an estimate of the order the number of singular values larger than a given noise dependent threshold. If such a threshold is not given, statistical methods such as the Akaike information criterion, the minimum description length, or the L-curve heuristic can be used instead [28]. For more information about the order estimation problem and methods for its solution, see the overview paper [29].

V. New setup: data-driven structured noise filtering

The new setup for data-driven structured noise filtering was illustrated in Section II on an example with two observed signals. The general data generation model in the structured noise filtering problem is

$$\begin{aligned} y_i = s_i + d_i + e_i, \quad \text{where } s_i \in \mathscr{B}_{\mathrm{s}} \in \mathscr{L}_{n_{\mathrm{s}}},\ d_i \in \mathscr{B}_{\mathrm{d},i} \in \mathscr{L}_{n_{\mathrm{d}}}, \\ \text{and } e_i \sim \mathrm{N}(0, \varsigma^2 I), \quad \text{for } i = 1, \ldots, N. \end{aligned} \tag{14}$$

Here, s_i is the true signal, d_i is the structured noise, and e_i is the unstructured noise in the ith experiment. The problem is to estimate the true signals $s_1, \ldots, s_N$ and the true model $\mathscr{B}_{\mathrm{s}}$, given the noisy data $y_1, \ldots, y_N$ and the model orders n_{s} and n_{d}. The maximum likelihood data-driven structured noise filtering problem is

$$\begin{aligned} \text{minimize} \quad & \sqrt{\sum_{i=1}^{N} \|y_i - \widehat{s}_i - \widehat{d}_i\|^2} \\ \text{over} \quad & \widehat{s}_i,\ \widehat{\mathscr{B}}_{\mathrm{s}},\ \widehat{d}_i,\ \widehat{\mathscr{B}}_{\mathrm{d},i}, \text{ for } i = 1, \ldots, N \\ \text{subject to} \quad & \widehat{s}_i \in \widehat{\mathscr{B}}_{\mathrm{s}} \in \mathscr{L}_{n_{\mathrm{s}}}, \quad \text{for } i = 1, \ldots, N \quad \text{and} \\ & \widehat{d}_i \in \widehat{\mathscr{B}}_{\mathrm{d},i} \in \mathscr{L}_{n_{\mathrm{d}}}, \quad \text{for } i = 1, \ldots, N. \end{aligned} \tag{15}$$

The following proposition states identifiability conditions in the new setup.

Theorem 8. *For $e_1 = \cdots = e_N = 0$, under the assumptions that:*

- A1: *the set of signals $\{s_1, \ldots, s_N\}$ is persistently exciting of order n_s and*
- A2: *the structured noise models $\mathscr{B}_{d,1}, \ldots, \mathscr{B}_{d,N}$ have no common poles,*

the solution $\widehat{\mathscr{B}}$ of (15) coincides with the true data generating system, i.e., *$\widehat{\mathscr{B}}_s = \mathscr{B}_s$.*

Proof. Define the systems

$$\mathscr{B}_i := \mathscr{B}_{\mathrm{s}} + \mathscr{B}_{\mathrm{d},i}, \quad \widehat{\mathscr{B}}_i := \widehat{\mathscr{B}}_{\mathrm{s}} + \widehat{\mathscr{B}}_{\mathrm{d},i}, \quad \text{for } i = 1, \ldots, N.$$

Note that $\mathscr{B}_i \in \mathscr{L}_{n_{\mathrm{s}}+n_{\mathrm{d}}}$ and $\widehat{\mathscr{B}}_i \in \mathscr{L}_{n_{\mathrm{s}}+n_{\mathrm{d}}}$. Moreover, by assumption A2, $\mathscr{B}_1, \ldots, \mathscr{B}_N$ and $\widehat{\mathscr{B}}_1, \ldots, \widehat{\mathscr{B}}_N$ have n_s common poles:

- the common poles of $\mathscr{B}_1, \ldots, \mathscr{B}_N$ are $\lambda(\mathscr{B}_{\mathrm{s}})$ and
- the common poles of $\widehat{\mathscr{B}}_1, \ldots, \widehat{\mathscr{B}}_N$ are $\lambda(\widehat{\mathscr{B}}_{\mathrm{s}})$.

By assumption $e_i = 0$, so that $y_i \in \mathscr{B}_i$. Then, by assumption A1, y_i is persistently exciting of order (at least) n_{s}. Although $\mathscr{B}_i$ may not be identifiable from y_i, the persistency of excitation assumption A1 guarantees that $\mathscr{B}_{\mathrm{s}} \subset \widehat{\mathscr{B}}_i$, for $i = 1, \ldots, N$. Then the common poles of $\widehat{\mathscr{B}}_1, \ldots, \widehat{\mathscr{B}}_N$ must coincide with the poles $\lambda(\mathscr{B}_{\mathrm{s}})$ of $\mathscr{B}_{\mathrm{s}}$. Therefore, $\widehat{\mathscr{B}}_{\mathrm{s}} = \mathscr{B}_{\mathrm{s}}$. □

Proposition 6 shows that in the classical setup for signal from noise separation, the maximum likelihood estimation problem is equivalent to Hankel structured low-rank approximation. Similarly, applying Corollary 3 to (15), we have that in the new setup the maximum likelihood estimation problem is equivalent to a generalized Hankel structured low-rank approximation problem.

Theorem 9. *The maximum likelihood estimation problem (15) is equivalent to the following Hankel structured low-rank approximation problem with multiple rank constraints*

$$\begin{array}{lll} \textit{minimize} & \textit{over } \widehat{s}_i,\ \widehat{d}_i,\ i=1,\ldots,N & \sqrt{\sum_{i=1}^{N} \|y_i-\widehat{s}_i-\widehat{d}_i\|^2} \\ \textit{subject to} & \textit{rank}\left[\mathscr{H}_{n_s+1}(\widehat{s}_1) \quad \cdots \quad \mathscr{H}_{n_s+1}(\widehat{s}_N)\right] \leq n_s, & \\ & \textit{rank}\,\mathscr{H}_{n_d+1}(\widehat{d}_i) \leq n_d, \quad \textit{for } i=1,\ldots,N. & \end{array} \tag{16}$$

Another reformulation of maximum likelihood estimation problem (15) as a Hankel structured low-rank approximation is given in the following proposition.

Theorem 10. *Under assumptions A2 and*

A1': *for $i=1,\ldots,N$, d_i is persistently exciting of order n_d, the maximum likelihood estimation problem (15) is equivalent to the Hankel structured low-rank approximation problem with multiple rank constraints*

$$\begin{array}{ll} \textit{minimize} \quad \textit{over } \widehat{y}_1,\ldots,\widehat{y}_N & \sqrt{\sum_{i=1}^{N} \|y_i-\widehat{y}_i\|^2} \\ \textit{subject to rank}\,\mathscr{H}_{n_s+n_d+1}(\widehat{y}_i) \leq n_s+n_d, \quad i=1,\ldots,N, & \\ \textit{rank}\left[\mathscr{H}_{Nn_d+n_s+1}(\widehat{y}_1) \ \cdots \ \mathscr{H}_{Nn_d+n_s+1}(\widehat{y}_N)\right] \leq Nn_d+n_s. & \end{array} \tag{17}$$

Proof. With the definition $\widehat{y}_i = \widehat{s}_i + \widehat{d}_i$, (15) and (17) have the same cost functions. Therefore in order to prove their equivalence, we need to prove that their constraints are equivalent (*i.e.*, that they define the same feasible sets). The first N constraints of (17) impose the constraints that $\widehat{y}_1,\ldots,\widehat{y}_N$ are trajectories of scalar linear time-invariant systems of order at most $n_\mathrm{s}+n_\mathrm{d}$, *i.e.*, $\widehat{y}_i \in \widehat{\mathscr{B}}_i \in \mathscr{L}_{n_\mathrm{s}+n_\mathrm{d}}$, for $i=1,\ldots,N$. Without extra constraints, we then have

$$\text{rank}\left[\mathscr{H}_{Nn_\mathrm{d}+n_\mathrm{s}+1}(\widehat{y}_1) \quad \cdots \quad \mathscr{H}_{Nn_\mathrm{d}+n_\mathrm{s}+1}(\widehat{y}_N)\right] \leq N(n_\mathrm{d}+n_\mathrm{s}).$$

Equality holds when each of the signals $\widehat{y}_i$ are persistently exciting of the maximal possible order $(n_\mathrm{s}+n_\mathrm{d})$ and the systems $\widehat{\mathscr{B}}_1,\ldots,\widehat{\mathscr{B}}_N$ have no common poles. Existence of n_s common poles implies that

$$\text{rank}\left[\mathscr{H}_{Nn_\mathrm{d}+n_\mathrm{s}+1}(\widehat{y}_1) \ \cdots \ \mathscr{H}_{Nn_\mathrm{d}+n_\mathrm{s}+1}(\widehat{y}_N)\right] \leq Nn_\mathrm{d}+n_\mathrm{s}. \tag{18}$$

Vice versa, if (18) holds and the signals $\widehat{y}_1,\ldots,\widehat{y}_N$ are persistently exciting of the maximal possible order, the models $\widehat{\mathscr{B}}_1,\ldots,\widehat{\mathscr{B}}_N$ must have at least n_s common poles. The persistency of excitation assumption that makes (18) equivalent to the condition that $\widehat{\mathscr{B}}_1,\ldots,\widehat{\mathscr{B}}_N$ have n_s common poles is guaranteed by assumptions A2 and A1'. □

Problem (17) can be viewed as a preprocessing operation on the data imposing the prior knowledge that the data generating model is (14). Computing the common dynamics model $\widehat{\mathscr{B}}_\mathrm{s}$ from the signals $\widehat{y}_1,\ldots,\widehat{y}_N$ is then an exact identification problem. This is considered in the next section.

Note 11 (Order selection in the new setup). As in the classical setup, the maximum-likelihood problem (15) in the new setup assumes that the orders n_s and n_d are known. However, if they are not given as prior information, they can be estimation from the data. First, the selection of the order $n_\mathrm{s}+n_\mathrm{d}$ of the combined true signal and structured disturbance system is a classical order selection problem (see Note 7). Second, as shown in Note 14, the selection of the order n_s of the common dynamics model $\mathscr{B}_\mathrm{s}$ is equivalent to the choice of the rank of a Sylvester matrix.

Note 12. The smoothing problem (*i.e.*, solving (15) with given models) is the classical smoothing problem that can be solved by the Kalman filter.

VI. Subspace approach for data-driven structured noise filtering

In the absence of unstructured noise, the data-driven structured noise filtering problem can be solved by the following generic method:

1) identify the models $\widehat{\mathscr{B}}_1,\ldots,\widehat{\mathscr{B}}_N \in \mathscr{L}_{n_\mathrm{s}+n_\mathrm{d}}$ of the observed signals $y_1,\ldots,y_N$,
2) comput the common dynamics $\widehat{\mathscr{B}}_\mathrm{s}$ of $\widehat{\mathscr{B}}_1,\ldots,\widehat{\mathscr{B}}_N$.

As shown in Theorem 8, under assumptions A1 and A2, $\widehat{\mathscr{B}}_\mathrm{s} = \mathscr{B}_\mathrm{s}$. In this section, we consider implementation details of the generic method for data-driven structured noise filtering. Particular ways of implementing steps 1 and 2 lead to different algorithms. These algorithms are then used as heuristics for data-driven structured noise filtering in the presence of unstructured noise.

A. Step 1: Identification of the true signals plus structured noise models

In the absence of unstructured noise and under assumption A1, the restriction of the model is given by the image of a Hankel matrix constructed from the data (Lemma 2). In the presence of unstructured noise, step 1 involves approximation. This step can be viewed then as a data preprocessing step

$$(y_1,\ldots,y_N) \mapsto (\widehat{y}_1,\ldots,\widehat{y}_N)$$

that imposes the prior information $s_i+d_i \in \mathscr{B}_i \in \mathscr{L}_{n_\mathrm{s}+n_\mathrm{d}}$.

1) Using the "row" data (no preprocessing):

$$(y_1,\ldots,y_N) \mapsto (\widehat{\mathscr{B}}_1|_L,\ldots,\widehat{\mathscr{B}}_N|_L).$$

Denote by $\mathscr{B}_i := \mathscr{B}_\mathrm{s}+\mathscr{B}_{\mathrm{d},i}$, for $i=1,\ldots,N$, the exact models for $y_1,\ldots,y_N$. In the absence of unstructured noise, by Lemma 2,

$$\text{image}\,\mathscr{H}_L(y_i) =: \widehat{\mathscr{B}}_i|_L \subset \mathscr{B}_i|_L, \qquad \text{for } i=1,\ldots,N. \tag{19}$$

Moreover, under assumption A2, $\mathscr{B}_\mathrm{s}|_L \subset \mathscr{B}_i|_L$, for $i=1,\ldots,N$. Therefore, $\mathscr{B}_\mathrm{s}|_L$ is a common subspace of $\mathscr{B}_1|_L,\ldots,\mathscr{B}_N|_L$. Under assumption A1, $\mathscr{B}_\mathrm{s}|_L$ can be computed as an intersection of subspaces defined by the data $y_1,\ldots,y_N$:

$$\mathscr{B}_\mathrm{s}|_L = \widehat{\mathscr{B}}_\mathrm{s}|_L := \widehat{\mathscr{B}}_1|_L \cap \cdots \cap \widehat{\mathscr{B}}_N|_L. \tag{20}$$

Finally, parameters of a representation of $\widehat{\mathscr{B}}_\mathrm{s}$ can be computed from the basis of $\widehat{\mathscr{B}}_\mathrm{s}|_L$.

2) Kung's method $(y_1,\dots,y_N)\mapsto(\widehat{R}_1,\dots,\widehat{R}_N)$*:*

Kung's method [30] is an effective suboptimal procedure for solving (10). It solves (12) by ignoring the structure and doing unstructured low-rank approximation—{the approximating matrix's rank-$(n_\mathrm{s}+n_\mathrm{d})$} is enforced by truncating of the singular value decomposition. Finally, the model parameters are computed by solving approximately an overdetermined system of equations in the least squares sense.

Applied to step 1 of the generic method for data driven structured noise filtering, Kung's method yields the following algorithm:

- For $i=1,\dots,N$
  1) Using the singular value decomposition
  $$\mathscr{H}_L(y_i)=U\Sigma V^\top,$$
  compute the rank-$(n_\mathrm{s}+n_\mathrm{d})$ approximation
  $$\mathscr{H}_L(y_i)\approx\mathscr{O}\mathscr{C}$$
  with $\mathscr{O}\in\mathbb{R}^{L\times(n_\mathrm{s}+n_\mathrm{d})}$ and $\mathscr{C}^{(n_\mathrm{s}+n_\mathrm{d})\times(T_i-L)}$, where $\mathscr{O}=U\sqrt{\Sigma}$ and $\mathscr{C}=\sqrt{\Sigma}V^\top$.
  2) Let $\widehat{A}_i$ be the least-squares approximate solution of the system of linear equations $\overline{\mathscr{O}}\widehat{A}_i=\underline{\mathscr{O}}$, where $\overline{\mathscr{O}}$ is the matrix $\mathscr{O}$ with the first row removed, and $\underline{\mathscr{O}}$ is the matrix $\mathscr{O}$ with the last row removed.
  3) Compute the characteristic polynomial $\widehat{p}^i(z)$ of $\widehat{A}_i$.

The suboptimal solution to (10) obtained by the subspace method is $\widehat{\mathscr{B}}_i=\ker\widehat{p}^i(\sigma)$, see (2).

The method has as a hyper parameter the natural number L,

$$n_\mathrm{s}+n_\mathrm{d}+1\leq L\leq T_i-(n_\mathrm{s}+n_\mathrm{d}).$$

Empirical evidence shows that best performance is obtained for square matrix $\mathscr{H}_L(y_i)$, *i.e.*, $L=\lfloor T/2\rfloor$.

Note 13. In [12] the total least squares method [31] is used for the parameter estimation instead of the least square method.

B. Step 2: Common dynamics computation

The identified models $\widehat{\mathscr{B}}_1,\dots,\widehat{\mathscr{B}}_N$ on Step 1 are not constrained to have common dynamics. Therefore, in the presence of unstructured noise, step 2 also involves an approximation that imposes the property of common dynamics. This can be done using an (approximate) subspace intersection or using common factor computational methods.

1) Subspace intersection $(\widehat{\mathscr{B}}_1|_L,\dots,\widehat{\mathscr{B}}_N|_L)\mapsto\widehat{\mathscr{B}}_s|_L$*:*

Consider kernel representations of the subspaces $\widehat{\mathscr{B}}_i|_L$ with dimensions $n_\mathrm{s}+n_\mathrm{d}=\dim\widehat{\mathscr{B}}_i|_L$, $\widehat{\mathscr{B}}_i|_L=\ker\mathbf{R}_i\subset\mathbb{R}^L$, for $i=1,\dots,N$, where $\mathbf{R}_i\in\mathbb{R}^{(L-n_\mathrm{s}+n_\mathrm{d})\times L}$ is the kernel parameter of $\widehat{\mathscr{B}}_i|_L$. We aim to find a kernel representation of their intersection (20) with dimension $n_\mathrm{s}=\dim\widehat{\mathscr{B}}_\mathrm{s}|_L$

$$\widehat{\mathscr{B}}_\mathrm{s}|_L:=\widehat{\mathscr{B}}_1|_L\cap\cdots\cap\widehat{\mathscr{B}}_N|_L=\ker\mathbf{R}\subset\mathbb{R}^L,$$

where $\mathbf{R}\in\mathbb{R}^{(L-n_\mathrm{s})\times L}$ is a kernel parameter of $\widehat{\mathscr{B}}_\mathrm{s}|_L$. First, we solve the exact intersection problem. Then, we explain the modification of the method for approximate intersection.

The matrix $\mathbf{R}'=\begin{bmatrix}\mathbf{R}_1\\ \vdots\\ \mathbf{R}_N\end{bmatrix}$ defines an exact kernel representation of the intersection $\widehat{\mathscr{B}}_\mathrm{s}|_L$. Indeed,

$$\begin{aligned}y\in\widehat{\mathscr{B}}_\mathrm{s}\quad&\Longleftrightarrow\quad y\in\widehat{\mathscr{B}}_i,\ \text{for } i=1,\dots,N\\ &\Longleftrightarrow\quad \mathbf{R}_iy=0,\ \text{for } i=1,\dots,N\\ &\Longleftrightarrow\quad \mathbf{R}'y=0.\end{aligned}$$

$\mathbf{R}'$ however is not minimal, *i.e.*, it is not full row rank. Computing a minimal kernel parameter $\mathbf{R}$ of $\widehat{\mathscr{B}}_\mathrm{s}$ requires finding a nonsingular matrix U, such that $U\mathbf{R}'=\left[\begin{smallmatrix}\mathbf{R}\\0\end{smallmatrix}\right]$, with $\mathbf{R}$ full row rank. Computing a kernel representation of an approximate intersection with dimension n_s requires a rank-$(L-n_\mathrm{s})$ approximation of $\mathbf{R}'$.

2) Approximate common factor $(\widehat{R}_1,\dots,\widehat{R}_N)\mapsto\widehat{R}_s$*:*

The subspaces $\widehat{\mathscr{B}}_1|_L,\dots,\widehat{\mathscr{B}}_N|_L$ have the special property that they are behaviors of autonomous linear time-invariant systems. Exploiting this structure allows us to develop more efficient methods for common dynamics computation. One approach of exploiting the linear time-invariant structure is to use the polynomials $\widehat{p}^1(z),\dots,\widehat{p}^N(z)$ in kernel representations of the models. Then, the subspace intersection problem becomes a problem of computing the greatest common divisor of the set of polynomials $\widehat{p}^1(z),\dots,\widehat{p}^N(z)$.

There are existing methods for greatest common factor computation. In the case when the unstructured noise is present, generically, an exact common factor does not exist. In this case, the aim is to find an approximate common factor of degree n_s. Again, this is a well developed problem in computer algebra. Computing an approximate greatest common factor however is a nonconvex optimization problem. As shown in [7], this problem is a Sylvester structured low-rank approximation. This motivates an alternative suboptimal method that first computes an approximate intersection $\widehat{\mathscr{B}}_\mathrm{s}$ and then models the resulting subspace as a linear time-invariant behavior [32].

Note 14 (Order selection of the common dynamics model). The estimation of the order n_s of the true signal's data generating model in the second step of the method is equivalent to the estimation of the order of the common dynamics of the combined true signal and structured noise models $\widehat{\mathscr{B}}_1,\dots,\widehat{\mathscr{B}}_N$ estimated in the first step. The order of the common dynamics, in turn, is the degree of the greatest common divisor of the polynomials $R_1,\dots,R_N$ defining minimal kernel representations of $\widehat{\mathscr{B}}_1,\dots,\widehat{\mathscr{B}}_N$. It is a well known result that the greatest common divisor degree is the co-rank (dimension minus the rank) of the generalized Sylvester matrix [7]

$$\mathscr{S}(R_1,\dots,R_N):=\begin{bmatrix}\mathscr{M}(R_1) & \cdots & \mathscr{M}(R_N)\end{bmatrix},$$

where

$$\mathscr{M}(R):=\begin{bmatrix}R_0 & & & \\ R_1 & R_0 & & \\ \vdots & \ddots & \ddots & \\ R_n & & \ddots & R_0\\ & \ddots & & R_1\\ & & \ddots & \vdots\\ & & & R_n\end{bmatrix}.$$

VII. Numerical examples

In this section, we validate empirically the generic approach for structured noise filtering described in Section VI. First, in Section VII-A, we illustrate the order estimation procedure. In the absence of unstructured noise, under the conditions of Theorem 8, the true signal's and structured noise models are identifiable and are recovered exactly by the subspace method independently of the pole location (*i.e.*, the systems can be stable, unstable, or marginally stable). As validated empirically in Section VII-B, the subspace method has good performance in the presence of unstructured noise when the true signal's dynamics is marginally stable. In case of stable or unstable true signal's dynamics, however, consistent parameter estimation from a finite number of experiments is not possible [25]. This is illustrated in Section VII-C.

A. Order estimation

First, we illustrate and validate empirically the order estimation procedure outlined in Note 11. The simulation setup is described in Section II, in particular, the true system and structured noise models have orders $n_\mathrm{s} = 4$ and $n_\mathrm{d} = 2$.

For the estimation of the order $n = n_\mathrm{s} + n_\mathrm{d}$ of the combined true signal and structured noise model in step 1 of the algorithm, we compute the singular values $s_1, s_2, \ldots$ of the Hankel matrix $\mathscr{H}_{\lceil T/2 \rceil}(y_1)$, see Figure 2. The estimate $\widehat{n}$ of n is the number of singular values above a noise dependent threshold. In the simulation example, there is a clear separation between s_6—the smallest singular value related to the model—and s_7—the largest singular value related to the unstructured noise. This makes possible order selection by visual inspection. Automated procedures that do not depend on human decision making and a priori given threshold are described in [29].

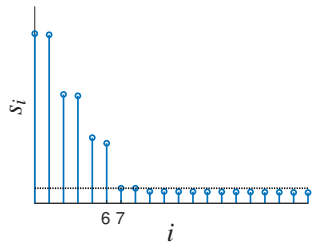

Fig. 2. The estimate $\widehat{n}$ of the order n of the combined true signal and structured noise model is obtained as the number of singular values s_i of the Hankel matrix $\mathscr{H}_{\lceil T/2 \rceil}(y_1)$ above a noise dependent threshold. In the example, there is a clear separation between s_6 (the smallest singular value related to the model) and s_7 (the largest singular value related to the unstructured noise).

For the estimation of the common dynamics order n_s in step 2 of the algorithm, we compute the singular values of the generalized Sylvester matrix $\mathscr{S}(\widehat{R}_1, \widehat{R}_2)$, where $\widehat{R}_1, \widehat{R}_2$ are the polynomials defining minimal kernel representations of the estimated models in step 1. In this case, the order estimate $\widehat{n}_\mathrm{s}$ is equal to the co-rank of $\mathscr{S}(\widehat{R}_1, \widehat{R}_2)$, *i.e.*, the number of singular values that are "close" to zero. Figure 3 shows the singular values of $\mathscr{S}(\widehat{R}_1, \widehat{R}_2)$ in the simulation example. There are $\widehat{n}_\mathrm{s} = 4$ singular values (s_{10}, s_{11}, s_{12}, and s_{13}) of the order of 10^{-10}, with $s_9 = 3 \times 10^{-4}$.

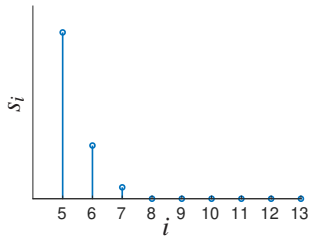

Fig. 3. The estimate $\widehat{n}_\mathrm{s}$ of the true signal's order n_s is equal to the number of singular values s_i of the Sylvester matrix $\mathscr{S}(\widehat{R}_1, \widehat{R}_2)$ that are close to zero. (In the example, "close to zero" means of the order of 10^{-10}.)

B. Comparison of the methods

Next, we compare the performance of the following methods for data-driven structured noise filtering:

- `alg1`: the subspace method without preprocessing followed by approximate subspace intersection,
- `alg2`: Kung's method using least squares (LS) for the parameter estimation followed by approximate subspace intersection,
- `alg3`: Kung's method using total least squares (TLS) for the parameter estimation followed by approximate subspace intersection (SI),
- `alg4`: Kung's method using least squares (LS) for the parameter estimation followed by approximate common factor (AGCD) computation with the method of [32].

The simulation setup is (14) with $N = 2$ trajectories, and model orders $n_\mathrm{s} = 4$, $n_\mathrm{d} = 4$. The signal lengths are $T_1 = T_2 = 150$, and the noise standard deviation is $\varsigma = 0.25$. The true models are randomly generated marginally stable linear time-invariant systems.

Let $\bar{p}(z)$ be a monic polynomial that defines a minimal kernel representation $\ker \bar{p}(\sigma)$ of the true model $\mathscr{B}_\mathrm{s}$ and let $\widehat{p}(z)$ be a monic polynomial that defines a minimal kernel representation $\ker \widehat{p}(\sigma)$ of the estimated model $\widehat{\mathscr{B}}_\mathrm{s}$. We define the following estimation errors

$$e_p = \frac{\|\bar{p} - \widehat{p}\|}{\|\bar{p}\|} \qquad \text{and} \qquad e_y = \frac{\|s - \widehat{y}\|}{\|s\|}$$

for the comparison of the methods: e_p is the relative *parameter error* and e_y is the relative *signal error*. The results of a Monte-Carlo simulation, averaged over 100 repetitions, are shown in Table I. The results show that

- no preprocessing gives worse performance,
- using ordinary least squares gives better performance than total least squares, and
- using common factor computation gives better performance than subspace intersection.

C. Other examples

The structured noise filtering problem has a special case the classical noise filtering problem. Although the subspace method proposed in the paper can handle this special case, it is less accurate than alternative methods, see, *e.g.*, [25]. The

TABLE I
RESULTS FROM A MONTE-CARLO SIMULATION, COMPARING THE SUBSPACE METHODS FOR DATA-DRIVEN STRUCTURED NOISE FILTERING.

	step 1	step 2	e_p	e_y
`alg1`	raw data	SI	0.3468	0.3163
`alg2`	[30] + LS	SI	0.2182	0.2084
`alg3`	[30] + TLS	SI	0.2460	0.2053
`alg4`	[30] + LS	AGCD	0.1368	0.1868

latter do not impose the common dynamics constraint in a postprocessing step but incorporates it in the first step.

Figure 4 shows that reducing the amplitude of the structured noise relative to the unstructured noise variance has no effect on the accuracy of the true signal model estimate, however, it has negative effect on the accuracy of the structured noise model estimate. The explanation for this fact is that reducing the amplitude of the unstructured noise has no effect on the SNR. However, structured noise to unstructured noise ratio (SNR with respect to estimation of the unstructured noise model) drops to zero.

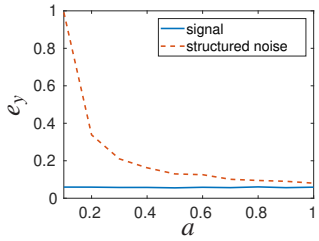

Fig. 4. Average relative estimation error e_y of the true signal (solid blue line) and structured noise (dashed red line) parameters as a function of the amplitude a of the structured noise for a fixed value of the unstructured noise variance. The estimation error of the true signal is independent of the structured noise amplitude, while the estimation error of the structured noise increases as the structured noise amplitude goes to zero.

In the introductory example of Section II (see Figure 1), the true signal and structured noises are selected as a periodic signals (corresponding to marginally stable data generating systems) in order to allow collection of data over an arbitrary long period of time. Obviously this is not possible in case of unstable data generating systems because of the exponential increase of the signals. The case of a stable data generating system is also problematic because in this case the signal-to-unstructured noise ratio converges exponentially to zero. This makes consistent estimation of the model not possible [25].

Figure 5 shows a simulation example with stable structured noise and true signal dynamics. Due to the exponential convergence of the signals the data collection period is limited. (In the example, just $T = 25$ samples are collected and used for solving the structured noise filtering problem.) Although the amount of information is limited and consistent estimation is not possible, the subspace method achieves reasonably good separation of the true signal, structured, and unstructured noise components.

VIII. CONCLUSIONS

Motivated by the need to deal with deterministic noise components, we considered a generalization of the classical data-driven noise filtering problem, where the noise has two components—structured noise, which is a trajectory of a low-complexity linear time-invariant system, and unstructured noise, which is a zero-mean white Gaussian process. The problem is well-posed when data is collected from multiple experiments and the structured noise models in the experiments have no common poles.

The maximum likelihood estimator in the new setup is a Hankel structured low-rank approximation problem with multiple rank constraints. We developed a generic subspace-type method that has the following steps: 1) model the observed signals, which serves a preprocessing role and 2) compute the intersection of the models obtained in step 1. The parameters of the common subspace yield then the parameters of interest—poles of the true data generating system.

The methods proposed achieve exact recovery in the absence of unstructured noise, *i.e.*, the methods separate the true signal from the structured noise exactly. In the presence of unstructured noise, the methods proposed are heuristics that yield suboptimal solutions to the maximum likelihood estimation problem. Simulation results comparing four variations of the generic subspace method show that parameter estimation using the ordinary least squares method yields more accurate results than parameter estimation using the total least squares method. Exploiting the linear time-invariant structure in the subspace intersection step by using methods for approximate common factor computation further improves the estimation accuracy.

The performance of the method in the case of unstructured noise and marginally stable true system's dynamics is validated empirically. The results show a gradual degradation of the performance as a function of the noise variance up to a threshold noise variance above which the method fails to yield good results. Statistical analysis of the method providing confidence bounds and prior estimate of the threshold noise level is a topic for future research.

Another direction for future work is extending the results to systems with driving inputs. This would generalize the setup considered in the paper from data generated by an autonomous linear-time-invariant system (sum-of-polynomials-times-damped-exponentials model) to data generated by a multi-input multi-output linear-time invariant system.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007–2013) / ERC Grant agreement number 258581 "Structured low-rank approximation: Theory, algorithms, and applications" and Fund for Scientific Research Vlaanderen (FWO) projects G028015N "Decoupling multivariate polynomials in nonlinear system identification" and G090117N "Block-oriented nonlinear identification using Volterra series"; Fonds de la Recherche Scientifique (FNRS) – FWO Vlaanderen under Excellence of Science (EOS) Project no 30468160 "Structured low-rank matrix / tensor approximation: numerical optimization-based algorithms and applications"; and JSPS KAKENHI Grant Numbers 17H01699 and 19H04069.

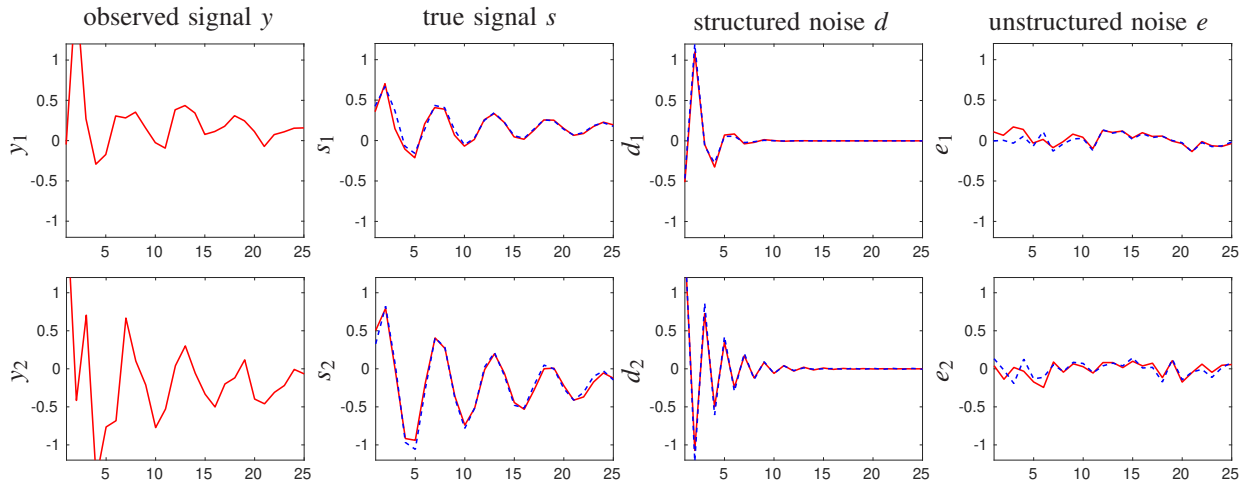

Fig. 5. Example with stable true signal and structured noise dynamics, *i.e.*, the true signal and structured noise converges to zero while the unstructured noise variance is constant. (solid red lines — observed and true signals, dashed blue lines — estimates obtained with the subspace method presented in Section VI.)

REFERENCES

[1] L. Ljung, *System identification: Theory for the user*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[2] M. Verhaegen, V. Verdult, and N. Bergboer, "Filtering and system identification: An introduction to using matlab software," 2007.

[3] I. Markovsky, *Low-Rank Approximation: Algorithms, Implementation, Applications*. Springer, 2019.

[4] ——, "On the behavior of autonomous Wiener systems," *Automatica*, vol. 110, p. 108601, 2019.

[5] N. Karmarkar and Y. Lakshman, "On approximate GCDs of univariate polynomials," in *J. Symbolic Comput.*, S. Watt and H. Stetter, Eds., vol. 26, 1998, pp. 653–666.

[6] M. Agarwal, P. Stoica, and T. P. Åhgren, "Common factor estimation and two applications in signal processing," *Signal Processing*, vol. 84, pp. 421–429, 2004.

[7] K. Usevich and I. Markovsky, "Variable projection methods for approximate (greatest) common divisor computations," *Theoretical Computer Science*, vol. 681, pp. 176–198, 2017.

[8] S. Van Huffel, "Enhanced resolution based on minimum variance estimation and exponential data modeling," *Signal Processing*, vol. 33, no. 3, pp. 333–355, 1993.

[9] L. Rippert, "Optical fibers for damage monitoring in carbon fiber reinforced plastic composite materials," Ph.D. dissertation, Katholieke Universiteit Leuven, 2005.

[10] R. Boyer and K. Abed-Meraim, "Audio modeling based on delayed sinusoids," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 110–120, 2004.

[11] Y. Haneda, S. Makino, and Y. Kaneda, "Common acoustical pole and zero modeling of room transfer functions," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 320–328, 1994.

[12] J.-M. Papy, L. De Lathauwer, and S. Van Huffel, "Common pole estimation in multi-channel exponential data modeling," *Signal Processing*, vol. 86, no. 4, pp. 846–858, Apr. 2006.

[13] I. Markovsky, T. Liu, and A. Takeda, "Subspace methods for multi-channel sum-of-exponentials common dynamics estimation," in *Proc. of the IEEE Conf. on Decision and Control*, 2019, pp. 2672–2675.

[14] W. De Clercq, B. Vanrumste, J. Papy, W. Van Paesschen, and S. Van Huffel, "Modeling common dynamics in multichannel signals with applications to artifact and background removal in EEG recordings," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 12, pp. 2006–2015, 2005.

[15] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[16] R. Roy and T. Kailath, "ESPRIT—estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[17] Y. Dong and J. Qin, "A novel dynamic PCA algorithm for dynamic data modeling and process monitoring," *Journal of Process Control*, vol. 67, pp. 1–11, 2018.

[18] J. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory*. New York: Springer-Verlag, 1998.

[19] J. C. Willems, "The behavioral approach to open and interconnected systems: Modeling by tearing, zooming, and linking," *Control Systems Magazine*, vol. 27, pp. 46–99, 2007.

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

[21] L. Scharf, "The SVD and reduced rank signal processing," *Signal Processing*, vol. 25, no. 2, pp. 113–133, 1991.

[22] I. Markovsky, "Structured low-rank approximation and its applications," *Automatica*, vol. 44, no. 4, pp. 891–909, 2008.

[23] ——, "Recent progress on variable projection methods for structured low-rank approximation," *Signal Processing*, vol. 96PB, pp. 406–419, 2014.

[24] ——, "A software package for system identification in the behavioral setting," *Control Eng. Practice*, vol. 21, no. 10, pp. 1422–1436, 2013.

[25] I. Markovsky and R. Pintelon, "Identification of linear time-invariant systems from multiple experiments," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3549–3554, 2015.

[26] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Boston: Kluwer, 1996.

[27] D. Luenberger, "An introduction to observers," *IEEE Trans. Automat. Contr.*, vol. 16, pp. 596–602, 1972.

[28] P. Hansen, *The L-Curve and Its Use in the Numerical Treatment of Inverse Problems*, 01 2001, vol. 4, pp. 119–142.

[29] P. Stoica and Y. Selén, "Model-order selection: A review of information criterion rules," *IEEE Signal Proc. Magazine*, vol. 21, pp. 36–47, 2004.

[30] S. Kung, "A new identification method and model reduction algorithm via singular value decomposition," in *Proc. 12th Asilomar Conf. Circuits, Systems, Comp.*, 1978, pp. 705–714.

[31] I. Markovsky and S. Van Huffel, "Overview of total least squares methods," *Signal Proc.*, vol. 87, pp. 2283–2302, 2007.

[32] W. Qiu, Y. Hua, and K. Abed-Meraim, "A subspace method for the computation of the GCD of polynomials," *Automatica*, vol. 33, no. 4, pp. 741–743, 1997.