Non-negative Observation-based Decomposition of Operators

Manuel Lopez Radcenco, Ronan Fablet, Abdeldjalil Aissa El Bey

Non-negative Observation-based Decomposition of Operators

Manuel Lopez-Radcenco, *Student Member, IEEE,* Ronan Fablet, *Senior Member, IEEE* and Abdeldjalil Aïssa-El-Bey, *Senior Member, IEEE*

***Abstract*—The problem of observation-based characterization of operators, closely related to the well-studied problem of blind source separation, remains nonetheless considerably less studied. Inspired by the recent success of non-negative and sparse blind source separation, we aim at extending constrained blind source separation models to the data-driven characterization of operators. We introduce a novel non-negative decomposition model for linear operators and investigate different parameter estimation algorithms. We study and compare the proposed algorithms in terms of identification and reconstruction performance in a variety of experimental settings, in order to gain insight into the robustness and limitations of the proposed algorithms. We further discuss the main contribution of our approach compared with state-of-the-art methods for the analysis and decomposition of operators.**

I. Introduction

The problem of observation-based separation of different contributions associated with multiple sources or processes, formally known as blind source separation [1], [2], has been extensively studied in signal and image processing. Considerable advancements have been made in the last few years with the introduction of non-negativity [3]–[6] and sparsity [7]–[9] constraints, which allow for the development of part-based representations and more interpretable models.

Non-negativity constraints have been exploited extensively in science and engineering. They were first introduced to account for situations where an inherent non-negativity exists within the problem solution (e.g. physical measurements, pixel intensities, frequency counts, etc), to constraint models in order to avoid physically impossible or absurd results [4]. Classical non-negatively constrained problems include Non-negative Least Squares (NNLS) [10] and Non-negative Matrix Factorization (NMF) [3]–[6], [11]. In practice, non-negativity can be enforced in several manners, most notably with optimization schemes such as active set algorithms [10], iterative update rules that maintain non-negativity [3], [4], [11], and constrained alternating least squared [11], [12] exploiting proximal operators [13], among others. Additionally, part of the inspiration behind non-negativity constraints comes from numerous situations in nature where a process or phenomenon can be explained by the additive mixing of multiple contributing factors [3], [4]. As such, they allow for the development of part-based representations with a wide variety of applications, ranging from learning parts of faces or semantic features of text [3] to music analysis and audio blind source separation of convolutive mixtures [5], [6].

Sparsity constraints, on the other hand, where introduced for dimensionality reduction, and to allow for the development of simpler representations of high-dimensional data [14]. This led to the development of simpler models and representations that are easier to understand. Constraining solutions to be as sparse as possible will nullify all but the strongest parts or components of the solution, thus allowing for these to be given greater relative importance in the final reconstruction (provided that an adequate number of non-zero components is parametrized). Formally, sparsity constraints impose restrictions on the number of non-zero elements of the solution $\mathbf{s}$. In practice, this is generally achieved through the minimization of the ℓ_0-norm $||\mathbf{s}||_0$, i.e. the number of non-zero elements of solution $\mathbf{s}$. As this minimization is an NP-hard, usually intractable problem, the constraint is generally relaxed and the ℓ_1-norm is considered instead. Several sparse representation algorithms exist to compute (or approximate) the solution of this minimization problem, including (but not restricted to) Lasso [14], Orthogonal Matching Pursuit [15] and proximal splitting methods [13].

Interestingly, even though an extensive literature exists on the problem of blind source separation, the very similar problem of observation-based characterization and decomposition of operators (relationships between variables of interest), on the other hand, has not been studied as extensively. To cite some examples, operator decomposition techniques have been used in domains such as fluid dynamics [16]–[18], oceanography and meteorology [19] and image processing [20]. Most of these approaches, however, rely on hypothesis such as classical orthogonality priors and time invariance, and exploit l_2-norm penalizations, which may not always yield the best representations. For instance, in the field of physical oceanography, understanding the relationships and interactions between different geophysical tracers is a major challenge [21]–[26], with approaches ranging from Fourier-based representations [27], [28] to latent class regression model [29]. However, with ocean dynamics depicting strong stochastic variabilities in space and time, these approaches may not be relevant to reveal hidden processes.

Overall, given the need for more complex formulations that can tackle the shortcomings of current models, recent advances in blind source separation applications using sparse and non-negative constraints make them particularly appealing

M. Lopez-Radcenco, R. Fablet and A. Aïssa-El-Bey are with IMT Atlantique, Lab-STICC, UBL, 29238 Brest, France.

This work was supported by ANR (grant ANR-13-MONU-0014), Labex Cominlabs project SEACS and OSTST project MANATEE.

to address the observation-based characterization and decomposition of operators. This paper addresses these issues and develops mathematically-sound and computationally-efficient schemes. Our main contributions are three-fold:

- A least-square formulation in the observation space under non-negativity constraints associated with different estimation algorithms;
- A reformulation of the considered non-negative issue as a dictionary leaning problem to gain modeling flexibility, including the ability to consider alternative priors, such as sparsity;
- The experimental evaluation of the proposed numerical schemes, which point out the relevance of the dictionary learning framework.

The remainder of the paper is organized as follows. Section II formally introduces the problem of observation-based operator decomposition and related work. Section III presents the proposed models and algorithms. We also include a brief analysis on computational complexity. In Section IV-B, we study and compare the different algorithms' performance in terms of parameter recovery and observation reconstruction for multiple cases, including ideal noiseless settings, cases involving a variable number of decomposition modes and configurations considering noisy observations and parameters. Finally, Section V presents our concluding remarks and future work perspectives.

II. Problem statement and Related work

The problem of separating different contributions associated with multiple sources or processes using only an observation dataset (without any a priori knowledge of the hidden processes behind the dataset generation), formally known as blind source separation [1], [2], is a classical problem that has been extensively studied in signal and image processing. In the last few years, considerable advancements have been reported with the development of non-negativity [3]–[6] and sparsity [7]–[9] based models. In a general way, we aim at decomposing a signal or image as the sum of K components:

$$\mathbf{y} = \sum_{k=1}^{K} \alpha_k \mathbf{s}_k + \boldsymbol{\omega} \tag{1}$$

where coefficient α_k quantifies the contribution of component $\mathbf{s}_k$, which corresponds to the k-th reference signal or image and $\boldsymbol{\omega}$ is a white Gaussian noise process that models the estimation residual. Depending on the specific characteristics of the dataset or application considered, model (1) may become intractable, which justifies the introduction of additional constraints in order to improve model identifiability.

The very similar problem of observation-based characterization and decomposition of operators (relationships between variables of interest), on the other hand, has not been studied as extensively. Formally, the general and unconstrained observation-based decomposition of operators amounts to considering operators which relate variables of interest $\mathbf{x}$ and $\mathbf{y}$, and state a general decomposition according to K modes as the superposition of K responses to input variable $\mathbf{x}$. For a dataset $\{\mathbf{x}_n, \mathbf{y}_n\}$, this may be given by:

$$\mathbf{y}_n = \sum_{k=1}^{K} \alpha_{nk} f_k(\mathbf{x}_n) + \boldsymbol{\omega}_n \tag{2}$$

where $\mathbf{x}_n \in \mathbb{R}^J$, $\mathbf{y}_n \in \mathbb{R}^I$, $\alpha_{nk} \in \mathbb{R}$ are mixing coefficients that model the contribution of each mode to the reconstruction of $\mathbf{y}_n$ given $\mathbf{x}_n$, $f_k : \mathbb{R}^J \to \mathbb{R}^I$ is a linear or non-linear function associated with mode k, and $\boldsymbol{\omega}_n \in \mathbb{R}^I$ is a noise process, usually a Gaussian noise. In this paper, we will focus on linear, constrained versions of model (2).

Fluid dynamics and dynamical system analysis are among the scientific fields where such decomposition models are of key interest. Dynamical Mode Decomposition (DMD) [16] is among the most popular approaches for the decomposition of operators governing dynamical systems into physically-relevant modes. DMD relies on orthogonality priors through an SVD [30] based eigendecomposition of a finite-dimensional approximation of the Koopman operator[1] [31]. Meanwhile, the Koopman operator relies on strong hypothesis, such as space-time stationnarity, which may not be fulfilled especially, and the accuracy of the finite-dimensional approximation of the (infinite-dimensional) Koopman operator needed to ensure the feasibility of DMD. To improve the interpretability and representation power of DMD, a few studies have investigated extensions with joint sparsity and non-negativity constraints [17] and Bayesian priors [18]. Beyond the DMD, orthogonality-based decomposition approaches [19], [32] are widely used for the analysis and characterization of geophysical systems. The assumption that the underlying modes may be separated according to orthogonality constraints is, in general, not supported by theoretical evidence. As stated in [33], it should be noted that, in spite of their computational efficiency, there are no guarantees that an orthogonal decomposition will yield physically-meaningful individual dynamical modes or even modes that relate to independent processes or individual kinematic degrees of freedom. Moreover, the extracted modes will most probably not be statistically independent, and might be strongly influenced by the modal non-locality needed to ensure that variance is maximized globally. In the image processing field, super-resolution methods based on linear patch-based operators have recently investigated operator decomposition issues as a means to reduce memory complexity [20]. These approaches exploit an ensemble of locally-estimated linear regressors to perform super-resolution. To reduce memory complexity, this ensemble is decomposed onto a set of basis regressors and representation coefficients, so that only basis regressors and coefficients are stored. Stored parameters are then used for the re-computation of original regressors at run-time, thus greatly reducing memory requirements (at the expense of some extra computational complexity). These methods, however, rely on Tikhonov regularization (also known as ridge regression), i.e. an l_2-norm penalization, for the decomposition

[1] The Koopman operator associated with a given non-linear finite-dimensional dynamical system is a linearization of its associated operator obtained by projecting it onto an appropriate infinite-dimensional space.

of operators, whereas other decomposition constraints might be more suitable.

III. Non-negative decomposition of Operators

We describe in this section the proposed formulation and algorithms for the non-negative decomposition of operators. We first introduce the considered formulation (Section III-A) and associated alternating least squares based algorithms for the estimation of model parameters (Section III-B). We then introduce a dictionary-learning based formulation of the considered problem (Section III-C). We further detail model training and application issues (Section III-D) and analyze the computational complexity of the considered algorithms (Section III-E).

A. General formulation

Let us consider a multivariate observation dataset $\{\mathbf{x},\mathbf{y}\}_n$, where $\mathbf{x}_n \in \mathbb{R}^J$, $\mathbf{y}_n \in \mathbb{R}^I$ denote the n^{th} observation pair. Variables $\mathbf{x}_n$ and $\mathbf{y}_n$ may, for instance, refer to feature vectors, image patches for different modalities or successive states of a dynamical system, depending on the targeted case-study. We focus on model (2) under the assumption that the potentially non-linear relationship between $\mathbf{x}_n$ and $\mathbf{y}_n$, given by functional response $f_k(\mathbf{x}_n)$, can be locally approximated, with reasonable accuracy, by a linear operator. The idea of exploiting a local linear approximation of non-linear operators directly relates to classical approaches such as local linear embbeding [34]. We consider the decomposition of the approximated linear operator relating variables $\mathbf{x}_n$ and $\mathbf{y}_n$ under non-negativity constraints. As stated in [35], this translates into the following model:

$$\begin{aligned} \mathbf{y}_n &= \sum_{k=1}^{K} \alpha_{nk}\boldsymbol{\beta}_k\mathbf{x}_n + \boldsymbol{\omega}_n \\ \text{Subject to } &\begin{cases} \alpha_{nk} \geq 0, & \forall\, k \in [\![1,K]\!], \forall n \in [\![1,N]\!] \\ ||\boldsymbol{\beta}_k||_F = 1, & \forall\, k \in [\![1,K]\!] \end{cases} \end{aligned} \tag{3}$$

where $\alpha_{nk} \in \mathbb{R}^+$ are non-negative mixing coefficients quantifying the contribution of linear modes k to the reconstruction of $\mathbf{y}_n$ for a given $\mathbf{x}_n$, $\boldsymbol{\beta}_k \in \mathbb{R}^{I\times J}$ is a regression matrix representing mode k, $||\cdot||_F$ is the Frobenius norm and $\boldsymbol{\omega}_n \in \mathbb{R}^I$ is a noise process, typically a centered Gaussian noise with covariance matrix $\boldsymbol{\Sigma}$, representing both model uncertainties and observation errors. N and K denote, respectively, the total number of observations and modes, while $k \in [\![1,K]\!]$ and $n \in [\![1,N]\!]$ indicate, respectively, the current mode and observation.

Here, a non-negativity constraint has been imposed on mixing coefficients α_{nk}, drawing inspiration from the success of non-negative decompositions in applications where naturally occurring positive superposition of parts exists [3], [4]. Additionally, a normalization constraint on modal regression matrices $\boldsymbol{\beta}_k$ has been added to eliminate scaling indetereminancies and improve model identifiability.

It may also be noted that any non-linear decomposition model (2) may be restated as a linear decomposition model (3) according to the vector of regression variables $(f_0(\mathbf{x}_n),..,f_K(\mathbf{x}_n))$. In the subsequent, we assume that candidate non-linear functional responses $(f_K(\mathbf{x}_n))$ are given a priori and we address the estimation of mixing coefficients (α_{nk}) and regression matrices $(\boldsymbol{\beta}_k)$.

B. Alternating least squares based estimation of model parameters

We state the estimation of model parameters for model (3) from a set of observations $\{\mathbf{x},\mathbf{y}\}_n$ as the resolution of the following non-linear, non-convex constrained optimization problem:

$$\forall n, \begin{cases} \left[\hat{\alpha}_{nk},\hat{\boldsymbol{\beta}}_k\right] = \underset{\alpha_{nk},\boldsymbol{\beta}_k}{\operatorname{argmin}} & \sum_{m=1}^{N} W_m^n \left|\left|\mathbf{y}_m - \sum_{k=1}^{K}\alpha_{nk}\boldsymbol{\beta}_k\mathbf{x}_m\right|\right|_{\boldsymbol{\Sigma}}^2 \\ \alpha_{nk} \geq 0, & \forall n \in [\![1,N]\!], \forall k \in [\![1,K]\!] \\ ||\boldsymbol{\beta}_k||_F = 1, & \forall k \in [\![1,K]\!] \end{cases} \tag{4}$$

where $||\cdot||_{\boldsymbol{\Sigma}}$ is a weighted norm according to covariance Σ. We assume that, according to weighing factors W_m^n, multiple observation pairs $(\mathbf{x}_m,\mathbf{y}_m)$ may share relatively similar mixing coefficients $\{\alpha_{nk}\}$. The greater W_m^n, the more similar the expected mixing coefficients $\{\alpha_{nk}\}$ and $\{\alpha_{mk}\}$. Weighing matrix W may encode both space-time smoothness priors, such that observation pairs close in space and/or time are expected to share similar operator decompositions, as well as observation-space similarity priors, for instance that observation pairs with similar regression variables may share similar decompositions. This seems reasonable for many applications where parameters are expected to correlate and vary smoothly in the considered spatio-temporal space. The parameterization of weighing matrix W is expected to be application-dependent and may be related to similar ideas used in covariance-based modeling [36] and non-local schemes [37]. Regarding identifiability issues, if the number of modes K verifies $K > I$ (where I is the dimension of observation vector $\mathbf{y}_n$), the estimation of mixing coefficients α_{nk} becomes intractable from a single observation pair $(\mathbf{x}_n,\mathbf{y}_n)$ [2]. As such, weighing matrix also provides a means to address the estimation of mixing parameters in such situations.

An interesting particular case of model (3), studied in [35], arises when $K \leq I$. In this case, model parameters may be estimated from only observation pair $(\mathbf{x}_n,\mathbf{y}_n)$ for each index n, with relates to the parameterization of weighing matrix W as

$$W_m^n = \begin{cases} 1, & m = n \\ 0, & m \neq n \end{cases} \tag{5}$$

and translates to the following constrained minimization problem:

$$\forall n, \begin{cases} \left[\hat{\alpha}_{nk},\hat{\boldsymbol{\beta}}_k\right] = \underset{\alpha_{nk},\boldsymbol{\beta}_k}{\operatorname{argmin}} & \left|\left|\mathbf{y}_n - \sum_{k=1}^{K}\alpha_{nk}\boldsymbol{\beta}_k\mathbf{x}_n\right|\right|_{\boldsymbol{\Sigma}}^2 \\ \alpha_{nk} \geq 0, & \forall n \in [\![1,N]\!], \forall k \in [\![1,K]\!] \\ ||\boldsymbol{\beta}_k||_F = 1, & \forall k \in [\![1,K]\!] \end{cases} \tag{6}$$

[2]For a fixed set of linear modes β_k, the estimation of mixing coefficients α_{nk} requires solving a linear system involving K unknowns and I equations

Nonetheless, it should be noted that the condition $K \leq I$ does not singlehandedly guarantee that Equation (6) will have a solution. Indeed, for pathological cases where the system's Gramian matrix is not invertible more observations need to be considered to compute a solution (as in Equation (4)). In this respect, a compromise exists between the number of observations considered, which increases model's robustness and numerical stability, and the locality of the model, which increases when fewer (or more local when dealing with space-time variabilities) observations are considered.

Given the non-linear, non-convex nature of constrained minimization problem (4), the joint estimation of model parameter sets α_{nk} and $\boldsymbol{\beta}_k$ is not straightforward. Conveniently, this jointly non-convex minimization problem becomes convex when estimation is performed for one set of parameters while considering the other set of parameters to be fixed. This suggests an alternating minimization approach, which leads to the following updates of model parameter sets α_{nk} and $\boldsymbol{\beta}_k$ being iterated until convergence:

- **$\boldsymbol{\beta}$-step:** Minimization over $\boldsymbol{\beta}_k$ with fixed α_{nk} and externally forced normalization constraints[3]

$$\hat{\boldsymbol{\beta}}_k^{i+1} = \hat{\boldsymbol{\beta}}_k^{i} + \left[\sum_{n=1}^{N} \hat{\alpha}_{nk}^{i} \left(\mathbf{y}_n - \sum_{p=1}^{K} \hat{\alpha}_{np}^{i} \hat{\boldsymbol{\beta}}_p^{i} \mathbf{x}_n\right) \mathbf{x}_n^T\right] \left[\sum_{n=1}^{N} \left(\hat{\alpha}_{nk}^{i}\right)^2 \mathbf{x}_n \mathbf{x}_n^T\right]^{-1} \tag{7}$$

$$\hat{\boldsymbol{\beta}}_k^{i+1} = \frac{\hat{\boldsymbol{\beta}}_k^{i+1}}{\left|\left|\hat{\boldsymbol{\beta}}_k^{i+1}\right|\right|_F}, \quad \forall k \in [\![1, K]\!] \tag{8}$$

- **α-step:** Minimization over α_{nk} with fixed $\boldsymbol{\beta}_k$ and externally forced non-negativity constraints

$$\hat{\alpha}_{nk}^{i+1} = \hat{\alpha}_{nk}^{i} + \frac{\sum_{m=1}^{N} W_m^n \left[\mathbf{x}_m^T \left(\hat{\boldsymbol{\beta}}_k^{i}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}_m - \sum_{p=1}^{K} \hat{\alpha}_{np}^{i} \hat{\boldsymbol{\beta}}_p^{i} \mathbf{x}_m\right)\right]}{\sum_{m=1}^{N} W_m^n \left[\mathbf{x}_m^T \left(\hat{\boldsymbol{\beta}}_k^{i}\right)^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}_k^{i} \mathbf{x}_m\right]} \tag{9}$$

$$\hat{\alpha}_{nk}^{i+1} = \max\left\{0, \hat{\alpha}_{nk}^{i+1}\right\} \tag{10}$$

For the case where a single observation pair suffices to estimated model parameters (Equation (6)), the α-step of the alternating minimization approach reduces to:

$$\hat{\alpha}_{nk}^{i+1} = \hat{\alpha}_{nk}^{i} + \frac{\mathbf{x}_n^T \left(\hat{\boldsymbol{\beta}}_k^{i}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}_n - \sum_{p=1}^{K} \hat{\alpha}_{np}^{i} \hat{\boldsymbol{\beta}}_p^{i} \mathbf{x}_n\right)}{\mathbf{x}_n^T \left(\hat{\boldsymbol{\beta}}_k^{i}\right)^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}_k^{i} \mathbf{x}_n} \tag{11}$$

[3]Since linear modes $\boldsymbol{\beta}_k$ are shared by all observation pairs, one set of regression matrices is estimated using all observation pairs in the training dataset. In this respect, all observation pairs are weighted equally for the estimation of modal regression matrices, under the assumption that they contribute uniformly to the estimation of the globally shared linear modes. Hence, to correctly fit the model, all global linear modes should be adequately sampled so as to be represented equally within the training dataset.

The downside to the simplicity of the alternating minimization approach is that it is prone to numerical issues. As acknowledged in the blind source separation literature [11], the alternating projections on the unconstrained and constrained solution spaces may induce divergent or numerically unstable behaviour. To handle such a problem, the direct minimization introduced in Equation (9) may be softened by considering a gradient descent:

$$\hat{\alpha}_{nk}^{i+1} = \hat{\alpha}_{nk}^{i} + 2\delta \left[\sum_{m=1}^{N} W_m^n \mathbf{x}_m^T \left(\hat{\boldsymbol{\beta}}_k^{i}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}_m - \sum_{p=1}^{K} \hat{\alpha}_{np}^{i} \hat{\boldsymbol{\beta}}_p^{i} \mathbf{x}_m\right)\right] \tag{12}$$

where δ is the used defined gradient descent step.
For the single-observation case (Equation (6)), the α-step reduces to:

$$\hat{\alpha}_{nk}^{i+1} = \hat{\alpha}_{nk}^{i} + 2\delta \left[\mathbf{x}_n^T \left(\hat{\boldsymbol{\beta}}_k^{i}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}_n - \sum_{p=1}^{K} \hat{\alpha}_{np}^{i} \hat{\boldsymbol{\beta}}_p^{i} \mathbf{x}_n\right)\right] \tag{13}$$

This is then combined with a projection onto the constrained non-negative solution space (Equation (10)), which comes down to a gradient based proximal splitting method [13].
Even though less necessary (since the renormalization constraint imposed in Equation (8) comes down to a simple rescaling), the same gradient based reformulation can be used for the estimation of modal linear regression matrices $\boldsymbol{\beta}_k$:

$$\hat{\boldsymbol{\beta}}_k^{i+1} = \hat{\boldsymbol{\beta}}_k^{i} + 2\delta \left[\sum_{n=1}^{N} \hat{\alpha}_{nk}^{i} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}_n - \sum_{p=1}^{K} \hat{\alpha}_{np}^{i} \hat{\boldsymbol{\beta}}_p^{i} \mathbf{x}_n\right) \mathbf{x}_n^T\right] \tag{14}$$

C. Dictionary-based formulation

As detailed below, the considered decomposition issue may be restated as a dictionary learning problem. In (3), linear operator $\sum_{k=1}^{K} \alpha_{nk} \boldsymbol{\beta}_k$ can be regarded as a decomposition of the local linear operator relating variables $\mathbf{y}$ and $\mathbf{x}$ for index n. This local linear operator may be estimated as follows according to a weighted least-square criterion using weighing matrix W:

$$\boldsymbol{\Theta}_n = \left(\sum_{m=1}^{N} W_m^n \mathbf{y}_m \mathbf{x}_m^T\right) \left(\sum_{m=1}^{N} W_m^n \mathbf{x}_m \mathbf{x}_m^T\right)^{-1} \tag{15}$$

where again W_m^n are weighting coefficients that account for the relative contributions of observation pairs $(\mathbf{x}_m, \mathbf{y}_m)$ to the estimation of the linear operator $\boldsymbol{\Theta}_n$ relating observation pair $(\mathbf{x}_n, \mathbf{y}_n)$. This least-square estimate comes to solve independently the least-square criterion for each index n in (3). Here, as in model (3), there is also a compromise between model robustness and computational stability and model locality, ultimately determined by the number of auxiliary observations considered for the estimation of local linear operators $\boldsymbol{\Theta}_n$.
Given local models $\{\boldsymbol{\Theta}_n\}_n$, problem (3) relates to the non-negative decomposition of linear operators $\boldsymbol{\Theta}_n$. It can be

shown that model (3) (which yields optimization problem (4)) can be reformulated as:

$$\boldsymbol{\Theta}_n = \sum_{k=1}^{K} \alpha_{nk}\boldsymbol{\beta}_k + \overbrace{\left(\sum_{m=1}^{N} W_m^n \boldsymbol{\omega}_m \mathbf{x}_m^T\right)\left(\sum_{m=1}^{N} W_m^n \mathbf{x}_m \mathbf{x}_m^T\right)^{-1}}^{\boldsymbol{\Upsilon}_n}$$
$$\text{Subject to } \begin{cases} \alpha_{nk} \geq 0, & \forall k \in [\![1,K]\!], \forall n \in [\![1,N]\!] \\ ||\boldsymbol{\beta}_k||_F = 1, & \forall k \in [\![1,K]\!] \end{cases} \tag{16}$$

The reformulation introduced by the estimation of local linear operators $\boldsymbol{\Theta}_n$ induces an error matrix $\boldsymbol{\Upsilon}_n \in \mathbb{R}^{I\times J}$ that depends directly on observations $\mathbf{x}_m$ and weights W_m^n. However, given the Gaussian nature of the original error term $\boldsymbol{\omega}_m$, the new error matrix $\boldsymbol{\Upsilon}_n$, being a linear combination of Gaussian terms, is a Gaussian matrix. The first and second order moments of the new error matrix elements $[\boldsymbol{\Upsilon}_n]_{ij}$ introduced in Equation (16) are:

$$\mathbb{E}\left([\,\boldsymbol{\Upsilon}_n\,]_{ij}\right) = 0 \tag{17}$$

$$\begin{aligned}[\boldsymbol{\Psi}_n]_{(ij)(i^*j^*)} =& \mathbb{E}\left([\,\boldsymbol{\Upsilon}_n\,]_{ij}[\,\boldsymbol{\Upsilon}_n\,]_{i^*j^*}\right) = \\ & \sum_{m=1}^{N}\left((W_m^n)^2 \mathbf{x}_m^T \left[\left(\sum_{m=1}^{N} W_m^n \mathbf{x}_m \mathbf{x}_m^T\right)^{-1}\right]_{:j} \right. \\ & \left. \left[\left(\sum_{m=1}^{N} W_m^n \mathbf{x}_m \mathbf{x}_m^T\right)^{-1}\right]_{j^*:} \mathbf{x}_m\right)[\boldsymbol{\Sigma}]_{ii^*}\end{aligned} \tag{18}$$

where sub-indexes $[A]_{:j}$ and $[A]_{l:}$ denote, respectively, the j-th column and the l-th line of matrix A. This leads to the conclusion that $\boldsymbol{\Upsilon}_n$ (and thus $\boldsymbol{\Theta}_n$) is heteroscedastic, i.e., its elements present a non-constant variance $\boldsymbol{\Psi}_n$ that depends on the observations $\{\mathbf{x}_n\}$ and weights W_m^n used to estimate the considered linear operator $\boldsymbol{\Theta}_n$. As such, adequately choosing the linear regression weights W_m^n should allow us to better manage the heteroscedastic nature of model (16).
Given local models $\{\boldsymbol{\Theta}_n\}_n$, parameter estimation for model (16) then translates to the following constrained optimization problem:

$$\begin{cases} \left[\hat{\alpha}_{nk}, \hat{\boldsymbol{\beta}}_k\right] = \underset{\alpha_{nk},\boldsymbol{\beta}_k}{\operatorname{argmin}} & \sum_{n=1}^{N}\left(\left|\left|\boldsymbol{\Theta}_n - \sum_{k=1}^{K}\alpha_{nk}\boldsymbol{\beta}_k\right|\right|_{\boldsymbol{\Psi}_n}^2\right) \\ \alpha_{nk} \geq 0, & \forall n \in [\![1,N]\!], \forall k \in [\![1,K]\!] \\ ||\boldsymbol{\beta}_k||_2 = 1, & \forall k \in [\![1,K]\!] \end{cases} \tag{19}$$

The direct minimization of this least-square criterion using ALS or gradient-spilling schemes would require the computation of an error covariance matrix $\boldsymbol{\Psi}_n$ for each local linear operator $\boldsymbol{\Theta}_n$, which implies a considerable increase in the computational complexity. Therefore, we propose an alternative method based on the dictionary-based decomposition of vectorized versions of local linear operators $\boldsymbol{\Theta}_n$ where, for the sake of simplicity, we also consider a simplified homoscedastic covariance structure. More precisely, the constrained minimization problem presented in Equation (19) can be restated as a blind dictionary learning based decomposition. We can consider the set $\{\boldsymbol{\Theta}\}_n$ of all N local linear operators, to which we apply the vectorization operator in order to rewrite Equation (19) as:

$$\begin{cases} \left[\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}\right] = \underset{\boldsymbol{A},\boldsymbol{B}}{\operatorname{argmin}} & ||\boldsymbol{\Phi} - \boldsymbol{B}\boldsymbol{A}||_F^2 \\ \boldsymbol{A}_{kn} \geq 0, & \forall k \in [\![1,K]\!], \forall n \in [\![1,N]\!] \\ |||[\boldsymbol{B}]_{:k}||_2 = 1, & \forall k \in [\![1,K]\!] \end{cases} \tag{20}$$

where matrix $\boldsymbol{\Phi} \in \mathbb{R}^{IJ\times N}$ is obtained by concatenating vectorized operators $\boldsymbol{\theta_n} = vec(\boldsymbol{\Theta_n})$ (i.e. $\boldsymbol{\Phi} = [\boldsymbol{\theta_1}|...|\boldsymbol{\theta_N}]$), columns of matrix $\boldsymbol{A} \in \mathbb{R}^{K\times N}$ contain mixing coefficients α_{nk} quantifying the contribution of each mode k for the reconstruction of vectorized local linear operator $\boldsymbol{\theta_n}$ and columns of $\boldsymbol{B} \in \mathbb{R}^{IJ\times K}$ (noted as $[\boldsymbol{B}]_{:k}$) contain vectorized versions of modal linear regression matrices $\boldsymbol{\beta}_k$, i.e., $[\boldsymbol{B}]_{:k} = vec(\boldsymbol{\beta}_k)$. The estimation of model parameters for model (20) resorts, under this new formulation, to a classical dictionary learning problem coupled with a non-negativity constraint. Dictionary learning is a classical problem in signal processing, for which numerous methods, exploiting different constraints, have been proposed [3], [4], [7], [9], [11]. Here, since we consider a non-negative constraint, we solve minimization (20) using a proximal splitting method [13] to account for the non-negativity of mixing coefficients matrix $\boldsymbol{A}$. It involves the iteration of the following two steps until convergence:

- The least-squares estimation of dictionary matrix $\boldsymbol{B}$ under normalization constraints $|||[\boldsymbol{B}]_{:k}||_2 = 1, \forall k$:

$$\boldsymbol{B}^{i+1} = \boldsymbol{\Phi}\left(\boldsymbol{A}^i\right)^T\left(\boldsymbol{A}^i\left(\boldsymbol{A}^i\right)^T\right)^{-1} \tag{21}$$

$$\left[\boldsymbol{B}^{i+1}\right]_{:k} = \frac{\left[\boldsymbol{B}^{i+1}\right]_{:k}}{|||\left[\boldsymbol{B}^{i+1}\right]_{:k}||_2} \quad \forall k \in [\![1,K]\!] \tag{22}$$

- The estimation of mixing coefficients matrix $\boldsymbol{A}$ using a gradient descent based proximal splitting method [13] to enforce non-negativity:

$$\boldsymbol{A}^{i+1} = \boldsymbol{A}^i - 2\lambda\left(\boldsymbol{B}^i\right)^T\left(\boldsymbol{\Phi} - \boldsymbol{B}^i\boldsymbol{A}^i\right) \tag{23}$$

$$\begin{aligned}\left[\boldsymbol{A}^{i+1}\right]_{kn} = \max\left\{0, \left[\boldsymbol{A}^{i+1}\right]_{kn}\right\}, & \forall k \in [\![1,K]\!] \\ , & \forall n \in [\![1,N]\!]\end{aligned} \tag{24}$$

Alternatively, one may choose a different dictionary-learning technique to enforce a different constraint (e.g. KSVD [7] for sparsity). This gives the dictionary-based formulation increased flexibility and adaptability, since alternative model constraints can thus be introduced seamlessly into model (3).

D. Model training and application

We may distinguish two different situations in terms of model parameter estimation for this dictionary-based formulation:

- **Model training:** Regression matrices $\hat{\boldsymbol{\beta}}_k$ (matrix $\hat{\boldsymbol{B}}$ in formulation (20)) and mixing coefficients $\hat{\alpha}_{nk}$ (matrix

$\hat{\boldsymbol{A}}$ in formulation (20)) are jointly estimated for a set of local linear operators $\boldsymbol{\Theta}_n$ obtained from a training dataset $\{\mathbf{x}, \mathbf{y}\}_n$. Estimated regression matrices $\hat{\boldsymbol{\beta}}_k$ will be considered as the dictionary of regression modes $\boldsymbol{\beta}_k$ when the model is applied to new observations and are thus stored for future use.

- **Model application:** Given a trained dictionary of operators $\{\hat{\boldsymbol{\beta}}\}_k$ (matrix $\hat{\boldsymbol{B}}$ in formulation (20)), mixing coefficients $\hat{\alpha}_{nk}$ (matrix $\hat{\boldsymbol{A}}$ in formulation (20)) are estimated for a new observation dataset $\{\mathbf{x}^*, \mathbf{y}^*\}_n$. Two approaches may be considered. Similarly to the training step, linear operators $\{\boldsymbol{\Theta}^*\}_n$ can be estimated for the new dataset, and mixing coefficients can be computed by projecting these operators onto the previously trained dictionary (using a non-negativity constraint). Alternatively, mixing coefficients can be estimated directly from observations using a least-squares criterion derived from model (3) without the prior estimation of linear operators $\{\boldsymbol{\Theta}^*\}_n$. Both approaches can be implemented using proximal operators, as in the model training step, or classical non-negative least-squares solvers [10]. It should be noted that the estimation of mixing coefficients α^*_{nk} for new observations $(\mathbf{x}^*_n, \mathbf{y}^*_n)$ may exploit only data from the training dataset, which, in the context of dynamical system prediction, provides the algorithm with actual prediction capabilities (since no knowledge of $\mathbf{y}^*_n = \mathbf{x}^*_{n+1}$ is needed for the estimation of mixing coefficients α_{nk}).

E. Computational complexity analysis

Table I presents a summary of complexity of the different algorithms, namely the alternating least squares exploiting a direct minimization (equations (7), (8), (9) and (10)), the alternating least squares exploiting a gradient descent (equations (7), (8), (12) and (10)) and the dictionary-based local linear operator decomposition (equations (21), (22), (23) and (24)), expressed in number of operations. Subsequently, we will refer to these algorithms as ALS-direct, ALS-gradient and LLOD, respectively. From these results, it is clear that differences in computational complexity arise from the different strategies used to approximate the unconstrained solution, as the cost of implementing model constraints is identical for all algorithms. ALS-gradient is more computationally demanding that ALS-direct, which seems in agreement with the more gradual manner in which the solution is approximated. In this respect, the added computational cost comes as a downside of having a more regular, smoother approach. As far as LLOD is concerned, complexity is shifted from the optimization stage to the estimation of local linear operators. Globally, however, LLOD involves a lower computational complexity than both variants of the ALS algorithm.

IV. Experiments

In this section we evaluate the performance of the proposed algorithms to address the general decomposition model (3) under ideal and non-ideal settings. We consider the three algorithms introduced in the previous section, namely ALS-direct, ALS-gradient and LLOD. We report numerical experiments to evaluate the proposed models and algorithms. We exploit synthetic data to perform a quantitative analysis of the estimation performance and a sensitivity analysis w.r.t. key parameters and modeling hypotheses.

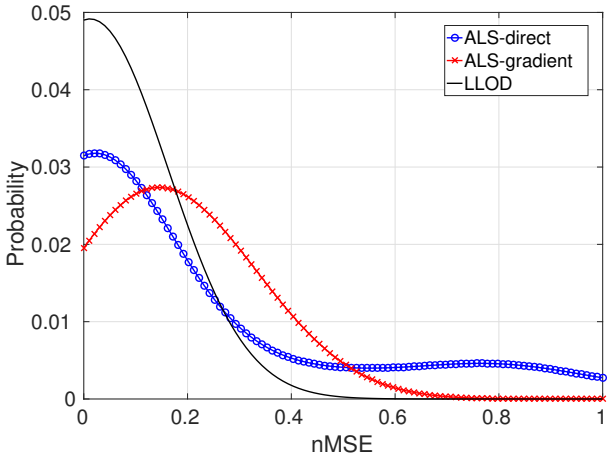

Fig. 1. Probability density function (PDF) for the normalized mean squared estimation error (nMSE) for mixing coefficients α_{nk}. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD). All presented probability distributions where computed using a gaussian kernel.

A. Synthetic dataset generation

We consider synthetic data $(\mathbf{x}_n, \mathbf{y}_n)$ so that we are provided with groundtruthed data. We proceed as follows. Mixing coefficients α_{nk} are simulated by means of a clustering-based approach so that they involve state-dependent variabilities. Elements of linear regression matrices $\boldsymbol{\beta}_k$ are sampled from a normal distribution $\mathcal{N}(0, 1)$, and regression matrices are subsequently normalized. A cluster-based strategy is used to generate observation pairs, so that N_c cluster centroids $\mathbf{x}_c$ are sampled from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \sigma_c^2\mathbf{I})$ and N_x clustered observations are sampled for each cluster from a multivariate normal distribution $\mathcal{N}(\mathbf{x}_c, \sigma_x^2\mathbf{I})$ centered around each cluster centroid. Mixing coefficients α_{nk} are emulated by sampling the same mixing coefficient for all observation pairs in a given cluster from a uniform distribution $\mathcal{U}_{[0,G_a]}$. Corresponding observations $\mathbf{y}_n$ are then generated by applying model (3) to $\mathbf{x}_n$.

B. Estimation performance under ideal settings

We first evaluate estimation performance under ideal noise-free conditions, i.e., when no observation noise is present, which means that noise process $\boldsymbol{\omega}_n$ in Equation (3) represents modeling error only. Moreover, we consider that all observations pairs within the same cluster share exactly the same operator decomposition, in the sense that no parameter noise in either mixing coefficients α_{nk} or modal regression

TABLE I: Computational cost of the different steps of the proposed algorithms, in number of operations.

	ALS-direct	ALS-gradient	LLOD
$\boldsymbol{\Theta}_n$ estimation	-	-	$NM[2J^2+2IJ]+N[J^3+IJ^2]$
$\alpha/\mathbf{A}$-step	$2NK[MKIJ+1]$ $+NK[I^2+IJ+2I+1]$ $+MNK[I^2+2IJ+J+1]$	$2NK[MKIJ+1]$ $+NK[I^2+IJ+2I+1]$ $+1$	$NK[2IJ+2]+NIJ+1$
$\alpha/\mathbf{A}\geq 0$	NK	NK	NK
$\boldsymbol{\beta}/\mathbf{B}$-step	$NK[IJ+I]$ $+N[2J^2+2IJ+I+1]]$ $+J^3+IJ^2+IJ$	$NK[IJ+I]$ $+N[2J^2+2IJ+I+1]]$ $+J^3+IJ^2+IJ$	$NK^2+K^2IJ+KIJ+K^3$
$\boldsymbol{\beta}/\mathbf{B}$ normalization	$K[3IJ+1]$	$K[3IJ+1]$	$K[3IJ+1]$

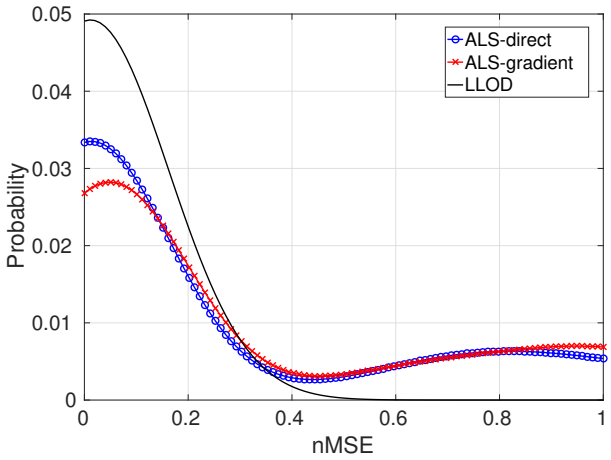

Fig. 2. Probability density function (PDF) for the normalized mean squared estimation error (nMSE) for linear modes $\boldsymbol{\beta}_k$. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD). All presented probability distributions where computed using a Gaussian kernel.

matrices $\boldsymbol{\beta}_k$ is considered. The minimal inter-cluster distance d_{min} verifies $d_{min} > 6\sigma_x$, which ensures a nearest neighbour search will only select points within the same cluster, such that they truly share the same mixing coefficients. All considered algorithms (ALS-direct, ALS-gradient, LLOD) where applied to a dataset generated considering $I = 30$, $J = 2$, $K = 2$, $N_c = 100$, $N_x = 300$, $\sigma_c^2 = 1$ and $\sigma_x^2 = d_{min}^2/360$. $M = 100$ nearest neighbours are used to estimate model parameters for each observation pair $(\mathbf{x}_n, \mathbf{y}_n)$, with uniform weighting $W_m^n = 1/M, \forall n, m$. The experience is repeated 100 times and results are averaged over all runs to ensure statistical significance.

Figure 1 presents the probability density function (PDF) of the normalized mean squared estimation error (nMSE) for mixing coefficients α_{nk}, defined as $nMSE\left(\alpha_{nk}, \hat{\alpha}_{nk}\right) = 1/K \cdot \sum_{k=1}^{K} \left[\sum_{n=1}^{N} \left(\alpha_{nk} - \hat{\alpha}_{nk}\right)^2 / \sum_{n=1}^{N} (\alpha_{nk})^2\right]$, with α_{nk} being the real mixing coefficients and $\hat{\alpha}_{nk}$ being the estimated mixing coefficients. Figure 2 presents similar results for linear modes $\boldsymbol{\beta}_k$. All PDFs where computed from the 100 simulation runs using a non-parametric Gaussian kernel based estimation. The dictionary-based LLOD algorithm yields a better reconstruction performance for both α_{nk} and $\boldsymbol{\beta}_k$, with an error PDF presenting higher values around zero and a rapidly decaying tail for higher error levels. By contrast, the two ALS schemes depict similar patterns for the estimation of regression matrices $\boldsymbol{\beta}_k$, with a secondary mode of the PDF centered around high nMSE values. These patterns indicate that the ALS algorithms do not converge for a significant fraction of cases. For 11% (resp.12%) of the simulations, the nMSE is greater than 0.5 for the ALS-direct (resp. ALS-gradient) scheme, whereas it remains at 0% for the LLOD algorithm. For mixing coefficients α_{nk}, ALS-gradient presents a wider, non-zero-centered mode, which reflects a lower parameter identification performance. As far as the ALS-direct is considered, even though it depicts higher probability levels around zero, its PDF presents, nonetheless, a slowly decaying tail, which reflects a higher instability, with high error values (greater than 0.5) for a significant percentage (approximately 20%) of the simulations. Of the considered algorithms, only the LLOD approach displays consistent and stable performance for the identification of both mixing coefficients α_{nk} and linear modes $\boldsymbol{\beta}_k$.

Regarding convergence properties, we report in Figures 3 and 4 the median nMSE (at convergence) as a function of the iteration number for mixing coefficients α_{nk} and linear modes $\boldsymbol{\beta}_k$. The LLOD approach presents a much slower and smoother convergence than the two ALS schemes for mixing coefficients α_{nk}, while also converging to the a lower nMSE value. Conversely, for linear modes $\boldsymbol{\beta}_k$, convergence is significantly slower for the ALS-gradient algorithm, while both the ALS-direct scheme and the LLOD approach present fast convergence towards low nMSE values. Overall, the ALS-direct scheme depicts a fast convergence (about 10 iterations) for both parameters, but we may underline that the convergence towards the actual parameters is not guaranteed as shown above. Regarding the LLOD approach, convergence is reached in about 10 iterations for linear modes $\boldsymbol{\beta}_k$ and 100

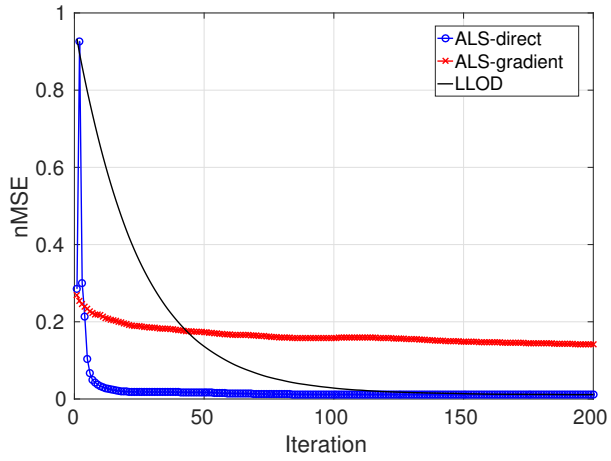

Fig. 3. Normalized mean squared estimation error (nMSE) median evolution for mixing coefficients α_{nk}. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD).

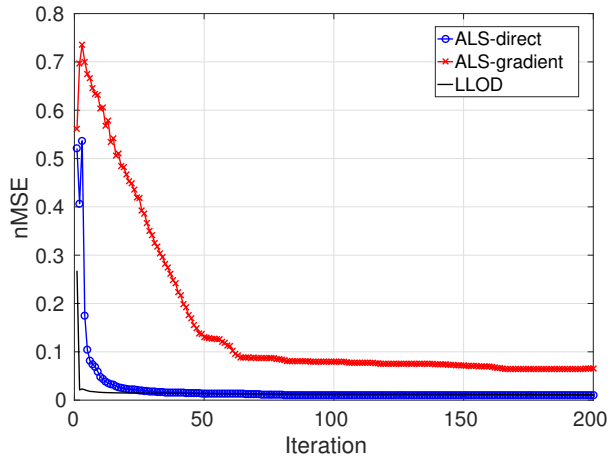

Fig. 4. Normalized mean squared estimation error (nMSE) median evolution for linear modes $\boldsymbol{\beta}_k$. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD).

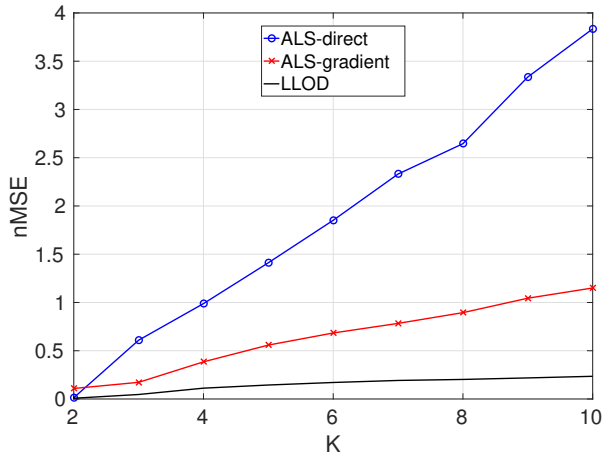

Fig. 5. Normalized mean squared estimation error (nMSE) final median value (at convergence) for mixing coefficients α_{nk} as a function of the number of classes K considered. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD).

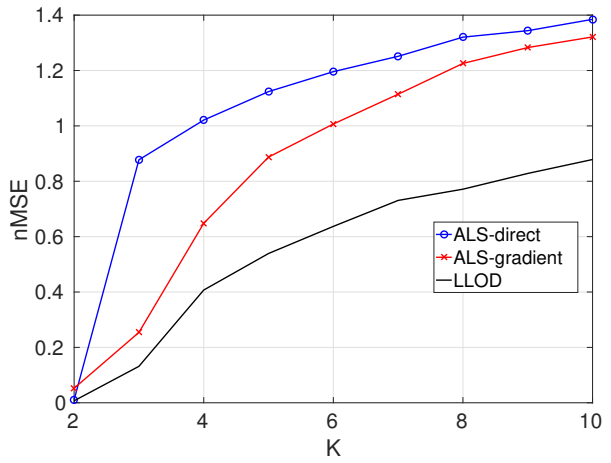

Fig. 6. Normalized mean squared estimation error (nMSE) final median value (at convergence) for linear modes $\boldsymbol{\beta}_k$ as a function of the number of classes K considered. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD).

iterations for mixing coefficients α_{nk}.

A complementary experiment addresses the evaluation of estimation performance with respect to the number of classes K. We vary the number of classes $K = 2, \ldots, 10$ and generate observations using the same procedure as previously. The experience is repeated 100 times for each number of classes K and results are averaged over all runs for each value of K. Figures 5 and 6 present the median nMSE (at convergence) for mixing coefficients α_{nk} and linear modes $\boldsymbol{\beta}_k$ as a function of the number of classes K, for the different algorithms considered. We also depict the median nMSE (at convergence) for the reconstruction of variables $\{\mathbf{y}_n\}$ in Figure 7. Obtained results show that the LLOD outperofrms both variants of the ALS for the recovery of both mixing coefficients α_{nk} and linear modes $\boldsymbol{\beta}_k$. As expected, ALS-direct is the least performant algorithm, which can be explained by a greater numerical instability and a higher estimation variance. Specifically, results show that the high performance degradation for $K > 3$ for ALS-direct is related to the existence of rapid oscillations between multiple local minima at each iteration, most probably due to the instabilities brought about by the alternating projections onto the constrained and unconstrained solution spaces. Overall, parameter recovery performance is degraded as K increases, so that we report

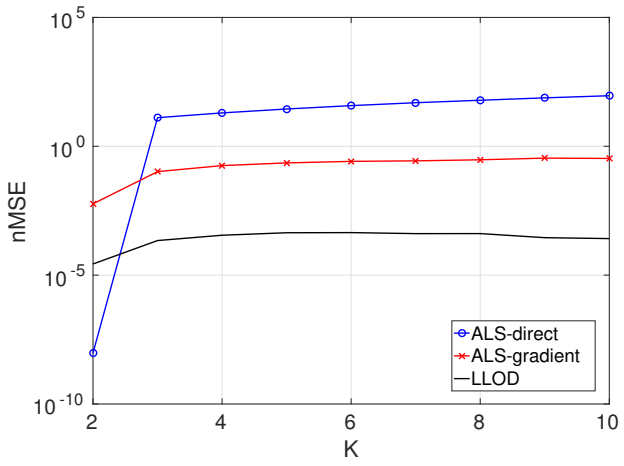

Fig. 7. Normalized mean squared $\mathbf{y}_n$ reconstruction error (nMSE) final median value (at convergence) as a function of the number of classes K considered. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD).

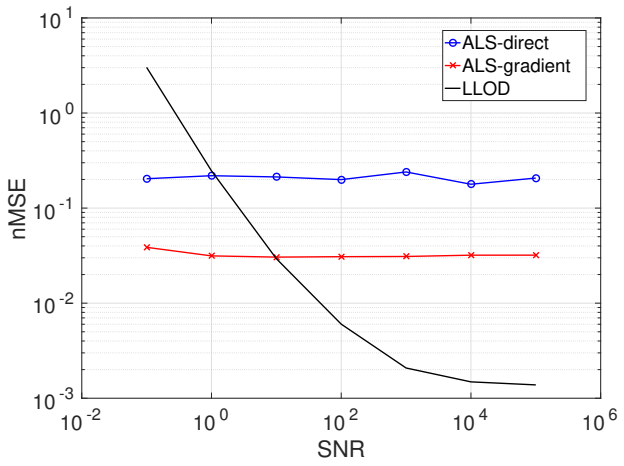

Fig. 8. Normalized mean squared estimation error (nMSE) final median value (at convergence) for mixing coefficients α_{nk} as a function of mixing coefficient SNR when Guassian noise is added to cluster-specific mixing coefficients. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD).

good parameter recovery performance only for a low number of classes ($K < 4$). Reconstruction performance, on the other hand, is weakly affected by the number of classes K, with low nMSE values for LLOD and rather poor nMSE levels both ALS variants (and particularly ALS-direct). These results relate to the identifiability of the model. This identifiability becomes weaker as the number of classes K increases, since so does the number of parameters to be estimated (given by $K(N + IJ)$), while the quantity of available information to estimate these parameters remains constant (since N, I and J, the number and dimensions of observations $\mathbf{x}_n$ and $\mathbf{y}_n$, remain unchanged).

C. Estimation performance with noisy mixing coefficients

We further evaluate the robustness of the proposed algorithms in the case of noisy mixing coefficients, that is to say that for a given observation index n in Equation (4) not all auxiliary observations pairs with index m and non-zero coefficients W_m^n may share exactly the same mixing coefficients α_{nk}. The considered experiment proceeds as follows. A random Gaussian noise is added to the initially cluster-specific mixing coefficients α_{nk} in order to obtain observation-specific coefficients, which will no longer be shared by observations in the same cluster. To prevent the existence of negative mixing coefficients due to the addition of Gaussian noise, the initial cluster-specific mixing coefficients are now sampled from a uniform distribution $\mathcal{U}_{[100G_a, 101G_a]}$.
This simulation setting implies that the $M = 100$ nearest-neighbors of sample n involve varying mixing coefficients, such that model (3) does not hold exactly and is only an approximation. In this respect, parameter similarity for close observations will now depend on the noise variance and, thus, on the signal-to-noise ratio (SNR) between the generated mixing coefficients α_{nk} and the added noise. As noise

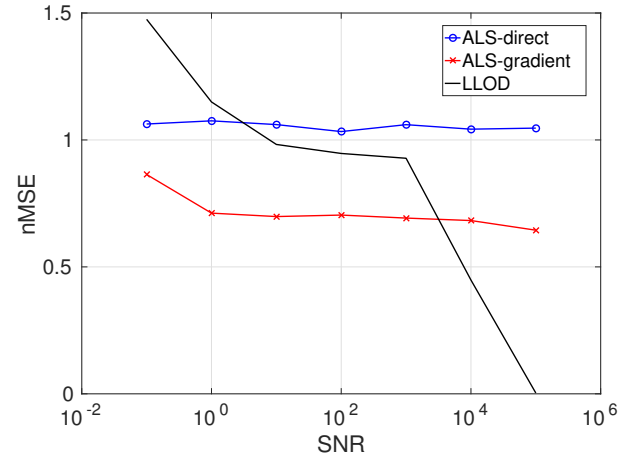

Fig. 9. Normalized mean squared estimation error (nMSE) final median value (at convergence) for linear modes $\boldsymbol{\beta}_k$ as a function of mixing coefficient SNR when Guassian noise is added to cluster-specific mixing coefficients. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD).

variance increases (SNR decreases), the relationship between observation similarity (in terms of distance and of belonging to a given cluster) and parameter similarity becomes weaker. The minimal inter-cluster distance d_{min} verifies $d_{min} > 6\sigma_x$, which ensures a nearest neighbour search will only select points within the same cluster. All considered algorithms (ALS-direct, ALS-gradient, LLOD) where applied to a dataset generated considering $I = 30$, $J = 2$, $K = 2$, $N_c = 100$, $N_x = 300$, $\sigma_c^2 = 1$ and $\sigma_x^2 = d_{min}^2/360$. Again, no observation noise is present, so that noise process $\boldsymbol{\omega}_n$ in Equation (3) represents modeling error only. $M = 100$ nearest neighbours

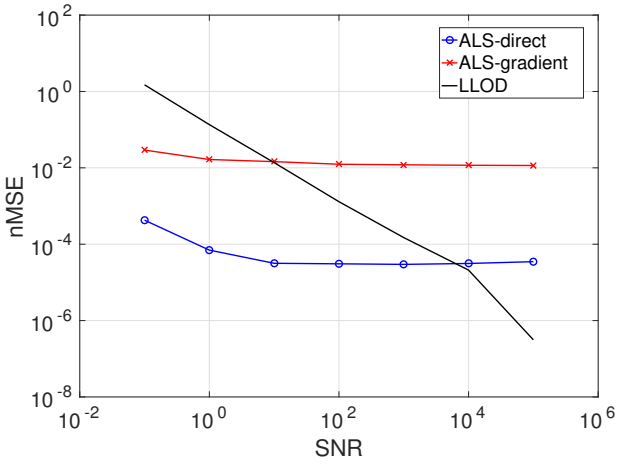

Fig. 10. Normalized mean squared $\mathbf{y}_n$ reconstruction error (nMSE) final median value (at convergence) as a function of mixing coefficient SNR when Guassian noise is added to cluster-specific mixing coefficients. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD).

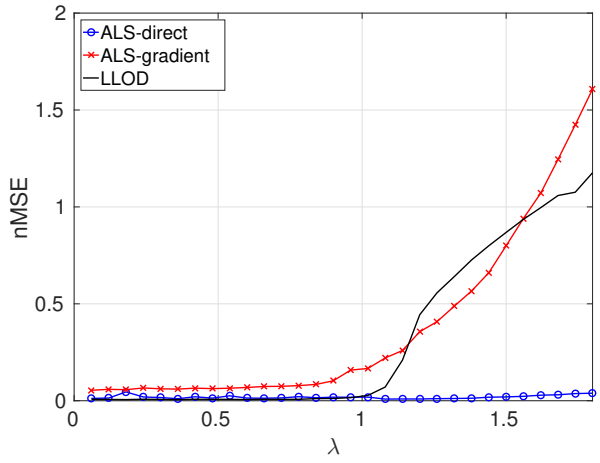

Fig. 11. Normalized mean squared estimation error (nMSE) final median value (at convergence) for mixing coefficients α_{nk} as a function of the ratio between cluster standard deviation σ_x and minimal distance d_{min} (parameter $\lambda = 6\frac{\sigma_x}{d_{min}}$), for initially non-overlapping clusters. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD).

are used to estimate model parameters for each observation pair $(\mathbf{x}_n, \mathbf{y}_n)$, with uniform weighting $W_m^n = 1/M, \forall n, m$. The experience is repeated 100 times and results are averaged over all runs, to ensure statistical significance. Moreover, these 100 simulation runs are repeated considering varying SNR levels: $SNR = \{10^{-1}, 10^0, \ldots, 10^5\}$.

Figures 8 and 9 present the median nMSE (at convergence) for mixing coefficients α_{nk} and linear modes $\boldsymbol{\beta}_k$ as a function of the SNR of mixing coefficients. We also depict the median nMSE (at convergence) for the reconstruction of variables $\{\mathbf{y}_n\}$ in Figure 10. The LLOD approach is clearly the most sensitive to noisy mixing coefficients. Linear modes $\boldsymbol{\beta}_k$ are highly affected even for low noise levels. For SNR values below 10^4, the approach converges to linear modes significantly different from the groundtruth ones. By contrast, the retrieval of mixing coefficients α_{nk} as well as reconstruction performance for variables $\{\mathbf{y}_n\}$ from observations $\{\mathbf{x}_n\}$ seems consistent for SNR levels greater than 10^1. As far as ALS algorithms are concerned, their performance is weakly affected by noisy mixing coefficients α_{nk} as illustrated by Figures 8 and 9. They however lead to poor estimation performance for the identification of linear modes $\boldsymbol{\beta}_k$ even for SNR values greater than 10^4. Overall, these experiments suggest identifiability issues for model (3) for noisy mixing coefficients even at high SNR values. It seems that there may exist a set of estimated linear regression matrices $\hat{\boldsymbol{\beta}}_k$, different from the true modal regression matrices $\boldsymbol{\beta}_k$, that lead to low reconstruction errors (typically nMSE values below 0.01). This implies that the proposed algorithms will be suitable for reconstruction applications, but will also suffer from non-unique solutions for the identification of regression modes $\{\boldsymbol{\beta}_k\}$. Nonetheless, it is worth noting that the non-uniqueness of the solution will not necessarily prevent the algorithms to be considered for identification/segmentation applications using mixing coefficients $\{\alpha_{nk}\}$ as illustrated in [35], [38]. From a computational point of view, one may investigate additional constraints or priors onto mixing coefficients α_{nk} and/or linear modes $\boldsymbol{\beta}_k$ to overcome such identifiability issues.

We further evaluate the extent to which we may account for other noise configurations, especially when neighbours in the observation space may not share similar mixing patterns. To study such situations, we simulate possibly overlapping clusters. As such, neighboring observation pairs $(\mathbf{x}_m, \mathbf{y}_m)$ and $(\mathbf{x}_n, \mathbf{y}_n)$, which are associated with non-null weighing coefficients W_m^n, may belong to different clusters and have, hence, different mean mixing coefficients α_{mk} and α_{nk}. Numerically, we proceed as follows to simulate such datasets. Initial cluster centroids are sampled from a multivariate Guassian distribution $\mathcal{N}(\mathbf{0}, \sigma_c^2)$. To ensure initial cluster separation, an additional acceptance/rejection sampling strategy is used to reject all cluster centroids that are too close to other centroids, according to a minimal distance d_{min}. Given the Gaussian nature of the centroid sampling distribution, the distance between cluster centroids will follow a Rayleigh distribution with scale parameter σ_c. Taking this into account, the minimal distance is chosen as $d_{min} = \sigma_c/e$, which ensures a relatively uniform spatial distribution of cluster centroids. For each cluster, we sample observation data $\{\mathbf{x}_n\}$ from Gaussian distributions $\mathcal{N}(\mathbf{x}_c, \sigma_x^2)$ with a standard deviation σ_x ranging from $\frac{1}{100}d_{min}$ to $\frac{30}{100}d_{min}$. For a standard deviation of $\frac{1}{100}d_{min}$, the simulation leads to non-overlapping clusters, whereas overlapping starts to occur from standard deviation values of $\frac{1}{6}d_{min}$ and above. We then evaluate estimation performance as a function of parameter $\lambda = 6\frac{\sigma_x}{d_{min}}$. Figures 11 and 12 present the median nMSE (at convergence) for mixing coefficients α_{nk} and linear modes $\boldsymbol{\beta}_k$ as a function of parameter $\lambda = 6\frac{\sigma_x}{d_{min}}$. Obtained results indicate, most notably,

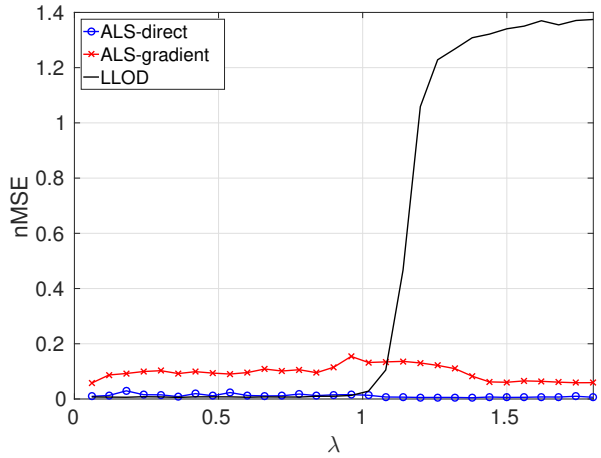

Fig. 12. Normalized mean squared estimation error (nMSE) final median value (at convergence) for linear modes $\boldsymbol{\beta}_k$ as a function of the ratio between cluster standard deviation σ_x and minimal distance d_{min} (parameter $\lambda = 6\frac{\sigma_x}{d_{min}}$), for initially non-overlapping clusters. Results presented for the ALS algorithm using a gradient descent approach (ALS-gradient), the ALS algorithm using a direct minimization (ALS-direct) and the dictionary-based decomposition of local linear operators (LLOD).

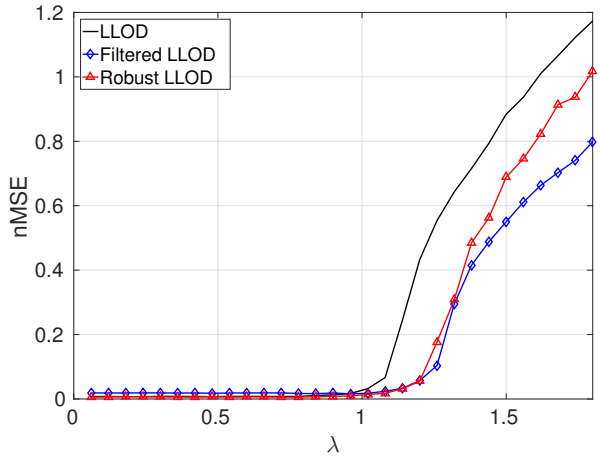

Fig. 13. Normalized mean squared estimation error (nMSE) final median value (at convergence) for mixing coefficients α_{nk} as a function of the ratio between cluster standard deviation σ_x and minimal distance d_{min} (parameter $\lambda = 6\frac{\sigma_x}{d_{min}}$), for initially non-overlapping clusters. Results presented for the original dictionary-based local linear operator decomposition (LLOD) and for two robust variants, namely Filtered LLOD, a filtering of local linear operators with mean value deviation higher than $f_c = \sigma_{\boldsymbol{\Phi}}$ (where $\sigma^2_{\boldsymbol{\Phi}}$ is the local linear operator mean value variance), and Robust LLOD, which involves the iterative re-weighted least squares estimation of local linear operators $\boldsymbol{\Theta}_n$.

that ALS-gradient seems to be unable to correctly recover mixing coefficients α_{nk} as soon as clusters are close enough so that observations from neighbouring clusters start to intervene in the estimation of model parameters, which occurs at around $\sigma_x/d_{min} = 1/6$, i.e. when $\lambda = 6\sigma_x/d_{min} = 1$. ALS-direct, on the other hand, seems more robust to cluster overlapping, with a slightly increasing nMSE as clusters merge. Moreover, LLOD seems to perform worst in the intermediate variance ranges, where parameters for observations near the cluster edge will be computed using wrongly selected neighbours from nearby clusters, while parameters for observations closer to the centroid will be estimated correctly from observations selected from the same cluster. Such behaviour can be also observed for the estimation of linear modes $\boldsymbol{\beta}_k$, whereas ALS-based algorithms seems to remain relatively robust to cluster overlap for the recovery of regression matrices $\boldsymbol{\beta}_k$.

Taking all previous considerations into account, it seems clear that ALS-direct should be used when cluster overlap may exist or when doubts may arise over how many auxiliary observations should be used and whether the chosen number of auxiliary observations may lead to the incorrect selection of nearest neighbours from nearby clusters. LLOD, on the other hand, should be used for cases where cluster overlap and the correct selection of the number of neighbours M are not an issue, since in such cases it will allow for a better model identification performance, both in terms of mixing coefficients α_{nk} and linear modes $\boldsymbol{\beta}_k$.

We specifically investigate robust estimation schemes to improve the performance of LLOD w.r.t. such overlap patterns. When cluster overlapping occurs, local linear operators $\boldsymbol{\Theta}_n$ for points near the clusters' edge are computed using observations from both the current and neighbouring clusters. When compared to local linear operators computed for observations closer to the cluster centroid (which are estimated using only observations from the current cluster), the later local linear operators tend to involve considerably larger values and will thus dominate the dictionary-based decomposition (Equation (20)). To tackle this problem, two different strategies are explored. The first strategy comes to compute the mean $m_{\boldsymbol{\Phi}}$ and standard deviation $\sigma_{\boldsymbol{\Phi}}$ of mean values of estimated local linear operators and filter all observations whose associated local linear operator $\boldsymbol{\Theta}_n$ mean value deviates from $m_{\boldsymbol{\Phi}}$, with cutoff values $m_{\boldsymbol{\Phi}} \pm f_c$, where $f_c = n\sigma_{\boldsymbol{\Phi}}$, for $n = \{1, 2, 3\}$. The second strategy involves the robust estimation of local linear operators $\boldsymbol{\Theta}_n$ using an iterative re-weighted least squares approach (considering $i = 25$ iterations) [39].Figures 13 and 14 present the median nMSE (at convergence) for mixing coefficients α_{nk} and linear modes $\boldsymbol{\beta}_k$ as a function of parameter $\lambda = 6\frac{\sigma_x}{d_{min}}$, for the original LLOD and for the two robust variants considered. For the sake of simplicity, only the most performant filtering strategy, namely that considering $f_c = \sigma_{\boldsymbol{\Phi}}$, is depicted. Reported results suggest that both approaches increase the robustness of LLOD, with best results obtained with the filtering scheme with the lowest cutoff value, closely followed by the robust regression approach, which has the additional advantage of not discarding any observations. These approaches consistently improve the working range of LLOD, which we define as the range of values for λ in which $nMSE < 0.1$. The working range of the original LLOD is $\lambda \in [0, \sim 1.08]$, which corresponds to a maximum overlap (in term of percentage of overlapping points between two clusters) of 0.55%, while the working range of the robust LLOD variants is $\lambda \in [0, \sim 1.26]$, which corresponds to a maximum overlap of 1.73%.

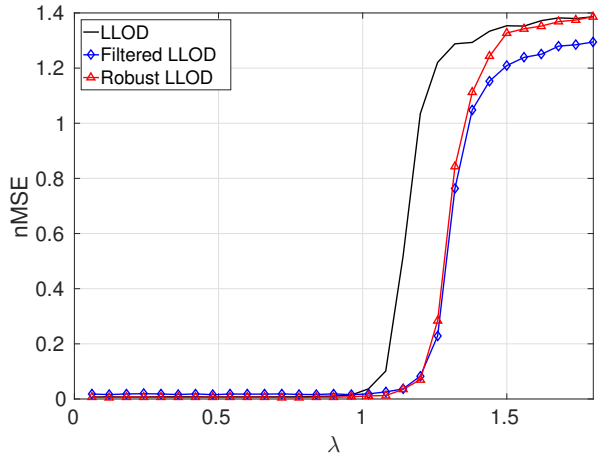

Fig. 14. Normalized mean squared estimation error (nMSE) final median value (at convergence) for linear modes $\boldsymbol{\beta}_k$ as a function of the ratio between cluster standard deviation σ_x and minimal distance d_{min} (parameter $\lambda = 6\frac{\sigma_x}{d_{min}}$), for initially non-overlapping clusters. Results presented for the original dictionary-based local linear operator decomposition (LLOD) and for two robust variants, namely Filtered LLOD, a filtering of local linear operators with mean value deviation higher than $f_c = \sigma_{\boldsymbol{\Phi}}$ (where $\sigma^2_{\boldsymbol{\Phi}}$ is the local linear operator mean value variance), and Robust LLOD, which involves the iterative re-weighted least squares estimation of local linear operators $\boldsymbol{\Theta}_n$.

V. Conclusion

In this paper, we addressed the extension of constrained blind source separation models to the observation-based decomposition of operators. We formally introduced a non-negative additive mixing model for operators, including a dictionary-based reformulation, and derived associated estimation algorithms. The dictionary-based formulation led to a greater modeling flexibility and possible straightforward extensions to sparsity-based priors. We performed numerical experiments to evaluate the estimation performance of the proposed algorithms. Overall, the dictionary-based decomposition of local linear operators seems to provide the best performance in terms of model identification, stability and computational complexity under favorable settings. Alternatively, under non-ideal settings, less stable algorithms, such as the ALS-direct, may nonetheless prove useful for model identification and observation reconstruction. In this respect, reported results suggest the need for additional regularization constraints or priors to tackle identifiability issues for model (3) in non-ideal configurations. Indeed, even though our experiments suggest that the proposed model and algorithms have good reconstruction performance in most settings, which makes them suitable for most reconstruction/prediction issues, model identification appears to be considerably sensitive to non-ideal settings, where the parameter sharing hypothesis is relaxed or where the number or selection of auxiliary observations for parameter estimation induces errors. Results also suggest that model identifiability can be improved by introducing robust estimation approaches for local linear operators and/or additional model constraints. The proposed models and algorithms, however, have been successfully used in both reconstruction/forecasting and segmentation applications [35], [38], [40]. These applications stress the relevance of the proposed non-negative decomposition of operators compared with orthogonality-based or latent class settings, which are considered in most previous works [19], [27]–[29].
Future work will focus on developing strategies for increasing model robustness and algorithm performance, further exploring sparsity and/or other alternative or additional constraints, and identifying and evaluating new possible applications.

References

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation*, Academic Press, Oxford, 2010.

[2] M. Pal, R. Roy, J. Basu, and M.S. Bepari, "Blind source separation: A review and analysis," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*, Nov 2013, pp. 1–5.

[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 2000, NIPS'00, pp. 535–541, MIT Press.

[5] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[6] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 550–563, March 2010.

[7] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[8] K. Stadlthanner, F. J. Theis, C. G. Puntonet, and E. W. Lang, "Extended sparse nonnegative matrix factorization," in *Computational Intelligence and Bioinspired Systems*, J. Cabestany, A. Prieto, and F. Sandoval, Eds., vol. 3512 of *Lecture Notes in Computer Science*, pp. 249–256. Springer Berlin Heidelberg, 2005.

[9] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi Morel, "K-WEB: Nonnegative dictionary learning for sparse image representations," in *IEEE International Conference on Image Processing (ICIP)*, Melbourne, Australia, Sept. 2013.

[10] C. L. Lawson and R. J. Hanson, *Solving least squares problems*, vol. 15, chapter 23, p. 161, SIAM, 1995.

[11] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics Data Analysis*, vol. 52, no. 1, pp. 155 – 173, 2007.

[12] R. Zadeh, H. Li, B. He, M. Lublin, and Y. Perez, "Cme 323: Distributed algorithms and optimization, spring 2015," University Lecture, 2015.

[13] P. L. Combettes and J. C. Pesquet, "Proximal Splitting Methods in Signal Processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, R. S.; Combettes P. L.; Elser V.; Luke D. R.; Wolkowicz H. (Eds.) Bauschke, H. H.; Burachik, Ed., pp. 185–212. Springer, 2011.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[15] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, Nov 1993, pp. 40–44 vol.1.

[16] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of Fluid Mechanics*, vol. 656, pp. 5–28, 2010.

[17] N. Takeishi, Y. Kawahara, and T. Yairi, "Sparse Nonnegative Dynamic Mode Decomposition," in *IEEE International Conference on Image Processing (ICIP)*, Beijing, China, Sept. 2017.

[18] N. Takeishi, Y. Kawahara, Y. Tabei, and T. Yairi, "Bayesian dynamic mode decomposition," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2814–2821.

[19] A. Hannachi, I. T. Jolliffe, and D. B. Stephenson, "Empirical orthogonal functions and related techniques in atmospheric science: A review," *International Journal of Climatology*, vol. 27, no. 9, pp. 1119–1152, 2007.

[20] E. Agustsson, Timofte R., and L. Van Gool, "Regressor Basis Learning for Anchored Super-Resolution," in *IEEE International Conference on Pattern Recognition (ICPR)*, Cancun, Mexico, Dec. 2016.

[21] M. M. Ali, D. Swain, and R. A. Weller, "Estimation of ocean subsurface thermal structure from surface parameters: A neural network approach," *Geophysical Research Letters*, vol. 31, no. 20, pp. n/a–n/a, 2004, L20308.

[22] K. S. Casey and D. Adamec, "Sea surface temperature and sea surface height variability in the north pacific ocean from 1993 to 1999," *Journal of Geophysical Research: Oceans*, vol. 107, no. C8, pp. 14–1–14–12, 2002.

[23] U. Hausmann and A. Czaja, "The observed signature of mesoscale eddies in sea surface temperature and the associated heat transport," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 70, pp. 60 – 72, 2012.

[24] J. Isern-Fontanet, B. Chapron, G. Lapeyre, and P. Klein, "Potential use of microwave sea surface temperatures for the estimation of ocean currents," *Geophysical Research Letters*, vol. 33, no. 24, pp. n/a–n/a, 2006, L24608.

[25] E. W. Leuliette and J. M. Wahr, "Coupled pattern analysis of sea surface temperature and TOPEX/Poseidon sea surface height," *Journal of Physical Oceanography*, vol. 29, no. 4, pp. 599–611, 2016/01/22 1999.

[26] M. Saraceno, C. Provost, and A. R. Piola, "On the relationship between satellite-retrieved surface temperature fronts and chlorophyll a in the western south atlantic," *Journal of Geophysical Research: Oceans*, vol. 110, no. C11, pp. n/a–n/a, 2005, C11016.

[27] J. Isern-Fontanet, M. Shinde, and C. González-Haro, "On the transfer function between surface fields and the geostrophic stream function in the Mediterranean Sea," *Journal of Physical Oceanography*, vol. 44, pp. 1406–1423, 2014.

[28] C. González-Haro and J. Isern-Fontanet, "Global ocean current reconstruction from altimetric and microwave SST measurements," *Journal of Geophysical Research: Oceans*, vol. 119, no. 6, pp. 3378–3391, 2014.

[29] P. Tandeo, B. Chapron, S. Ba, E. Autret, and R. Fablet, "Segmentation of mesoscale ocean surface dynamics using satellite sst and ssh observations," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 52, no. 7, pp. 4227–4235, July 2014.

[30] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numer. Math.*, vol. 14, no. 5, pp. 403–420, Apr. 1970.

[31] B. O. Koopman, "Hamiltonian systems and transformations in hilbert space," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 17, no. 5, pp. 315–318, 1931.

[32] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, vol. 2, no. 11, pp. 559–572, 1901.

[33] A. H. Monahan, J. C. Fyfe, M. H. P. Ambaum, D. B. Stephenson, and G. R. North, "Empirical orthogonal functions: The medium is the message," *Journal of Climate*, vol. 22, no. 24, pp. 6501–6514, 2009.

[34] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[35] M. Lopez-Radcenco, A. Aissa-El-Bey, P. Ailliot, P. Tandeo, and R. Fablet, "Non-negative decomposition of linear relationships: Application to multi-source ocean remote sensing data," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 4179–4183.

[36] A. L. Boulesteix and K. Strimmer, "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data," *Briefings in Bioinformatics*, vol. 8, no. 1, pp. 32–44, 2007.

[37] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, June 2005, vol. 2, pp. 60–65 vol. 2.

[38] M. Lopez-Radcenco, A. Aissa-El-Bey, P. Ailliot, and R Fablet, "Non-negative decomposition of geophysical dynamics," in *ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium, April 2017.

[39] C. Sidney Burrus, "Iterative reweighted least squares," December 2012, OpenStax CNX.

[40] M. Lopez-Radcenco, R. Fablet, A. Aissa-El-Bey, and P. Ailliot, "Locally-adapted convolution-based super-resolution of irregularly-sampled ocean remote sensing data," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 4307–4311.