Unsupervised Complex Semi-Binary Matrix Factorization for Activation Sequence Recovery of Quasi-Stationary Sources

Romain Delabeye, Martin Ghienne, Olivia Penas, Jean-Luc Dion

Unsupervised Complex Semi-Binary Matrix Factorization for Activation Sequence Recovery of Quasi-Stationary Sources

Romain Delabeye[a,*], Martin Ghienne[a], Olivia Penas[a], Jean-Luc Dion[a]

[a] *Quartz Laboratory, EA7393, ISAE-Supméca, 3 rue Fernand Hainaut, Saint-Ouen, 93400, France*

Abstract

Advocating for a sustainable, resilient and human-centric industry, the three pillars of Industry 5.0 call for an increased understanding of industrial processes and manufacturing systems, as well as their energy sustainability. One of the most fundamental elements of comprehension is knowing when the systems are operated, as this is key to locating energy intensive subsystems and operations. Such knowledge is often lacking in practice. Activation statuses can be recovered from sensor data though. Some non-intrusive sensors (accelerometers, current sensors, etc.) acquire mixed signals containing information about multiple actuators at once. Despite their low cost as regards the fleet of systems they monitor, additional signal processing is required to extract the individual activation sequences. To that end, sparse regression techniques can extract leading dynamics in sequential data. Notorious dictionary learning algorithms have proven effective in this regard. This paper considers different industrial settings in which the identification of binary sub-system activation sequences is sought. In this context, it is assumed that each sensor measures an extensive physical property, source signals are periodic, quasi-stationary and independent, albeit these signals may be correlated and their noise distribution is arbitrary. Existing methods either restrict these assumptions, e.g., by imposing orthogonality or noise characteristics, or lift them using additional assumptions, typically using nonlinear transforms.

This paper addresses these limitations, and introduces the unsupervised

*Corresponding author

complex semi-binary matrix factorization (ℂSBMF) as its main contribution. In particular, we show that the exact recovery of source activation sequences from non-intrusive sensor data is intrinsically tied to the presence of problematic phase shifts, the causes of which are detailed. A greedy algorithm is proposed, iteratively resynchronizing sources to converge towards the maximum decomposition of each operation despite these phase shifts. The ℂSBMF is verified and compared to existing techniques on synthetic use cases, then validated on experimental data with signals of different nature. To that occasion, the CAFFEINE dataset for unsupervised time series multi-label classification is introduced.

Keywords: Underdetermined blind source separation, Semi-binary matrix decomposition, Unsupervised time series multi-label clustering, Energy disaggregation, Sparse dictionary learning, Inverse problems

1. Introduction

The manufacturing industry is inherently energy-intensive, accounting for around 37% of global final energy consumption in 2022 [1]. Amid the recent energy crisis, some factory pilots are looking to increase the efficiency of their machines and processes, aiming for substantial reductions in energy use and associated carbon emissions. Among the many solutions sought to achieve this objective, digital twins stand out, a popular cross-industry concept that has seen a rapid rise in recent years. A digital twin is *"a set of adaptive models that emulate the behavior of a physical system in a virtual system getting real time data to update itself along its life cycle. The digital twin replicates the physical system to predict failures and opportunities for changing, to prescribe real time actions for optimizing and/or mitigating unexpected events observing and evaluating the operating profile system"* [2]. In particular, this enables a number of enhancements, from predicting energy consumption, locating energy drifts induced by faulty components, misuse or environmental changes, to optimizing machine control, component replacement, and process scheduling.

These improvements come at a cost though. A manufacturing plant is characterized by the many cyber-physical systems (CPS) it contains and the similarities they may present. If scalability is sought as regards production, monitoring is no exception. That is, in number of cases, few sensors must monitor a large fleet of machines. Not only are they able to monitor multiple

systems at once, but their implementation does not require direct intervention on these systems' hardware or software. Placing sensors inside a machine, provided that a suitable location can be found without tampering with the production system, often requires to stop the said machine during the intervention. Intrusive sensors may also require rewiring, reprogramming or accessing data from a CPS's dedicated programmable logic controller (PLC). Actuator-specific operating statuses may also be programmed in a low-level language without being returned to the user interface. Overall, the activation sequences are difficult to retrieve. Manually labeling data as a post-processing layer is expensive at plant scale, and so is the development of a physical model for each actuator. As an alternative, with a view to learn energy consumption models, estimate and predict actuator-specific performance indicators, the activation sequences can be extracted from sensor data instead. Hence, in order to unlock the above-mentioned applications, this paper focuses on fully unsupervised non-intrusive load monitoring (NILM), and more specifically on the recovery of actuator activation sequences in sensor data. Moreover, the elements presented here are not limited to current load, but to piecewise mixed quasi-stationary periodic signals in the broadest sense.

Unsupervised NILM, or energy disaggregation, aims to discover the active appliances in energy consumption data [3]. In these methods, the activation sequences are often the result of CPSs entering successive states. Although this problem appears very close to the one at hand, many techniques in this community rely on dedicated features, active and reactive power or peak current to name but a few, and problem-specific methods such as change or event detection [3, 4]. Popular techniques include most notably hidden Markov models (HMM) [5], where the states of a Markov model are not observed directly but are implicitly defined by a probability density function (pdf). Deep learning architectures have also proven successful either using denoising autoencoders in a sequence-to-sequence fashion [6] or convolutional neural networks [7]. Overall, energy disaggregation rather focuses on pattern recognition such as device-specific power distribution or state transitions, putting the emphasis on the process and the machines instead of the underlying actuators. This can be problematic in presence of flexible processes or event-based control, where the process changes and the actuator's activation sequences are not well separated.

A more general approach to this problem is through blind source separation (BSS), the action of retrieving a set of S source signals from M mixed signals. BSS has received great attention over the years in multiple domains,

from audio source separation [8, 9], energy disaggregation [10] to fault detection and diagnosis [11, 12, 13]. The BSS problem is underdetermined if $M < S$. This problem is often dealt with as a matrix factorization which consists in reconstructing data as the composition of two matrices. Two mathematical formulations stand out. On the one hand, clustering can be used to separate operating phases in which sources are mixed, the two matrices then correspond to a *mixing matrix* and *source signals* different in each phase [14, 15, 16]. On the other hand, data can be reconstructed as a sparse *representation* on a *dictionary* [10, 17].

Early Line Spectra Estimators (LSE), inspired by Prony's method in the 1980s, exploit the fact that amplitude and phase estimation becomes least squares solvable when the frequencies are known [18]. This led to the popular subspace methods decomposing discrete data into signal and noise subspaces, among which figure the MUSIC [19] and ESPRIT [20] algorithms, and variants thereof either on-grid [21, 22, 23], or off-grid with a parameterized sparse Fourier representation [24]. Statistical decomposition can also be achieved through underdetermined independent component analysis (ICA) [25, 26], although this technique is limited to sub-Gaussian signals and does not perform well in presence of discrete events. Alternatively, sparse component analysis (SCA) [27, 28] first applies a sparsifying transform on data such as the short time Fourier transform (STFT) or wavelet transform (WT), the rationale being that sources are easier to separate in a lifted space where they exhibit noticeable differences. Spectral decomposition techniques can also construct an orthonormal eigenbasis, onto which the projection of the data results in separated sources. These techniques include singular value decomposition (SVD) [29, 30], difference mode decomposition [31] as well as dynamic mode decomposition (DMD) [32, 33]. In the time-frequency (TF) domain, Nonnegative Matrix Factorization (NMF) and Nonnegative Tensor Factorization (NTF) [13] are a natural choice when working on positive features such as spectrograms or bi-frequency maps [34]. Although effective in isolating sequential dynamic behaviors, subspace methods fall short when the underlying signals do not have orthogonal spectra, as the embedding no longer represents the true sources, but their common characteristics. Moreover, these techniques lack a mechanism to force the representation to lie in a binary space as expected in a multi-label classification problem.

To remedy this limitation, semi-binary NMF forces the representation to be binary [10]. Setting aside partially supervised implementations, popular in energy disaggregation [10], binarity can be enforced through means of reg-

ularization [11] or directly using a coupled factorization method and relaxed alternating least squares (ALS) [35]. The nonnegative requirement imposes the use of nonlinear transforms and hence phase removal.

An alternative consists in learning a shift-invariant dictionary based on the convolution operator [36]. Convolutional sparse coding strategies learn a sparse representation as its convolution with temporal patterns. Traditional techniques include the shift-invariant sparse coding model from Grosse et al. [37] and variants [38]. These models either use prior knowledge on the source signals (pre-computed dictionary with source responses), or learn the dictionary along with the representation [39]. Sub-dictionaries containing time shifted copies of the initial dictionary were recently proposed by Wang et al. [40], although this discrete approximation fails to capture events lying off-grid.

At last, the mentioned techniques require the number of sources to be known. For dimensionality reduction, this number can be approximated [15], yet this often boils down to rank estimation [41]. Noisy data may induce a large Pareto front though. This makes rank estimation highly imprecise. Another possible cause of error in rank estimation is the use of nonlinear transforms. Proposals have been made in previous work to decompose signals with an accurate estimation of the number of sources, either using tracking for non-stationary cases [42] or matrix factorization [43].

To the best of the authors' knowledge, to date there is no matrix factorization method in the literature capable of recovering the exact activation sequences from a mixed signal under the constraints considered in this paper: fully unsupervised underdetermined blind source separation with unknown number of sources, applied to additive, potentially correlated, periodic, ergodic and quasi-stationary source signals with arbitrary noise distribution and no prior knowledge, resulting in a complex dictionary with a binary representation. Our main contribution is twofold, (i) we propose a novel formulation circumventing the pitfalls arising in semi-binary matrix factorization in a complex vector space, and (ii) a greedy algorithm to learn both the dictionary and the sparse representation.

The remainder of this paper is set out as follows. After formally describing the particular underdetermined blind source separation problem this paper is concerned with in Section 2.1, a clustering-based two-step algorithm is proposed. In Section 2.3, we show that the exact recovery of source activation sequences from non-intrusive sensor data is intrinsically tied to the presence of problematic phase shifts, the causes of which are detailed. A so-

lution to the complex semi-binary matrix decomposition problem is found in Section 2.4 by carefully resynchronizing the sources. This method is finally verified on synthetic data and compared to existing methods in Section 3.1. Experimental validation is undertaken in Section 3.2, in which we introduce the CAFFEINE dataset [44]. Limitations and perspectives are discussed in Section 4, before concluding.

2. Methods

2.1. Problem formulation

A production process is a sequence of *operations*. Each of these involves a collection of *actuators*. Operations are therefore successive and cannot overlap. An actuator comprehends all the elements of a connected power chain, actionable simultaneously and controlled as a whole. These actuators produce *source* signals when aggregated by a sensor, resulting in a time series of sequential mixed signals. An actuator is a source s, switched on (1) and off (0) according to its activation sequence $\mathrm{ACT}_s(t)$ over time. Let $\mathcal{C}$ denote the alphabet of all distinct operations in data, each operation involving simultaneous sources. An operation is a group of sources $c \in \mathcal{C}$, later denoted *cluster*. It is attributed an activation status $\mathrm{OPS}_c(t)$.

The short time Fourier transform (STFT) is used in this paper, as it is well suited to the study of piecewise stationary signals. Its definition is recalled for a signal x of finite support, uniformly sampled over time with frequency f_s. The STFT can be viewed as a sliding discrete Fourier transform (DFT) applied to partially overlapping windows with hop size H, using an analysis window w with size W, indexed by discrete time step m, and frequency bin index k associated with discretized pulse $\boldsymbol{\omega}$, with $\forall k \in [\![0, W-1]\!], \omega_k = \frac{2\pi k}{W}$:

$$STFT\{x\}[m,k] = \sum_{n=0}^{W-1} x[n]\mathrm{w}[n-m]e^{-\jmath\frac{2\pi k}{W}n} \tag{1}$$

The time-shift theorem of the DFT is used to shift the signal in time while remaining in the frequency domain. The time-shift operator $\boldsymbol{S}_\Delta$ for a time difference Δ is defined as:

$$\boldsymbol{S}_\Delta = \mathrm{diag}\left(\left(e^{-\jmath\frac{2\pi k}{W}\Delta}\right)_{0\leq k<W}\right) \tag{2}$$

Let $\boldsymbol{X} \in \mathbb{R}^{T'}$ denote a univariate time series representing T' sensor measurements sampled at frequency f_s. $\boldsymbol{X}$ is produced by S actuators sequentially activated, in use in N_{ops} distinct operations, with $N_{ops} \geq S$. From time series $\boldsymbol{X}$, the feature matrix $\boldsymbol{Z} \in \mathbb{C}^{W \times T}$ is computed using the STFT, where W is the number of frequency bins (and window size) and T is the number of feature samples (time windows) in the TF domain, with $T' \geq T$.

The retrieval of the sources' descriptors and activation sequences can be sought as the optimal solution to an underdetermined semi-binary matrix decomposition problem [11]. Here, S sources are mixed over a single channel. Undetermined blind source separation is an inverse problem, ill-posed in that the matrix factorization (regardless the formulation) does not admit a unique stable solution. This issue is often overcome through regularization and constraints on the dictionary, the representation or both. A classic formulation is as a sparse dictionary learning problem [11, 12, 34, 40]:

$$\underset{\substack{\boldsymbol{D} \in \mathcal{D}, \boldsymbol{R} \\ \boldsymbol{R} \geq 0}}{\arg\min} \sum_{m=1}^{T} \left(\Psi(\boldsymbol{Z}_m - \boldsymbol{D}\boldsymbol{R}_m) + \mathcal{R}_{sparse}(\boldsymbol{R}_m) + \mathcal{R}_{binary}(\boldsymbol{R}_m) \right) \tag{3}$$

where $\boldsymbol{D} \in \mathbb{R}^{W \times N}$ and $\boldsymbol{R} \in \mathbb{R}^{N \times T}$ are the *dictionary* and *representation* (multi-labels over time) to be learnt. $\mathcal{R}_{sparse}$ is a sparsity-promoting penalty, typically the Least Absolute Shrinkage and Selection Operator (LASSO). Binary solutions are enforced either using a regularizer $\mathcal{R}_{binary}$ [45] or an alternative to the least-squares functional [11]. $\mathcal{D}$ is a set of constraints on $\boldsymbol{D}$, necessary to prevent the penalties on $\boldsymbol{R}$ from being compensated by larger elements in $\boldsymbol{D}$ due to the coupled functional, as both are jointly optimized. Labels are constrained to positive values for convenience. Indeed, allowing signed labels would result in implicitly defined actuators, i.e., as the presence of a collection of actuators (positive labels) while excluding some others (negative labels).

The signals produced by the sources are assumed to be quasi-stationary and ergodic. Stationarity is a strong assumption, here its weak form is preferred, in which only the mean and the covariance of the process must be time-invariant and finite [46]. In other words, for each operation, the steady state lasts long enough for the transient response to have a negligible impact on the mean and variance of its descriptor.

Input signal is thus piecewise consistent and there exists a *dictionary* $\boldsymbol{C} \in \mathbb{C}^{W \times N}$ of *atoms* $[\boldsymbol{C}_1, ..., \boldsymbol{C}_N]$, also called *centroids* or *descriptors*, prop-

erly describing the stationary state of each operation. In a complex vector space, phase shifts occur and a single operation may be present in more than one configuration, i.e., the sources' time shifts may differ from a realization to another. Hence $N \geq N_{ops}$. A subset $\tilde{\boldsymbol{C}} \in \mathbb{C}^{W \times S}$ of these centroids describes the actuators isolated from all others. It is here assumed that each actuator is seen at least once alone in the dataset.

Similarly, the operation the underlying system is in at each time step is represented by the one-hot encoded label matrix $\breve{\boldsymbol{L}} \in \{0,1\}^{N_{ops} \times T}$, or $\boldsymbol{L} \in \{0,1\}^{N \times T}$ to distinguish all configurations. Let $\|.\|_0$ denote the ℓ_0 norm, $\forall m \in [\![0, T-1]\!]$, $\|\boldsymbol{L}_m\|_0 = 1$. Whereas the activation statuses of a system's actuators over time are gathered in a multi-hot encoded matrix $\tilde{\boldsymbol{L}} \in \{0,1\}^{S \times T}$, with $\|\tilde{\boldsymbol{L}}_m\|_0 \geq 1$. Thus the time series $\boldsymbol{L}^c$ and $\tilde{\boldsymbol{L}}^s$ indicate whether the system is in operation c and uses actuator s at each time step. Matrix superscripts and subscripts represent vectors lying in the row and column spaces respectively.

Here we lay out the assumptions, humorously coined the ten plagues of unsupervised complex semi-binary matrix factorization, which arise from the blind source separation problem and from the application, recovering actuator activation statuses in an industrial environment from non-intrusive sensors and without supervision.

P 1. *Fully unsupervised*

P 2. *Unknown number of sources*

P 3. *Underdetermined ($S < N$)*

P 4. *Periodic source signals*

P 5. *Potentially correlated sources*

P 6. *Quasi-stationary source signals*

P 7. *No prior knowledge on sources*

P 8. *Any noise distribution*

P 9. *Complex dictionary and binary representation*

P 10. *Each sensor measures an extensive property*

The above-described problem uses a dictionary, each atom involving a group of sources. Due to binarity, the decomposition result only makes sense if each atom in the minimal set $\tilde{\boldsymbol{C}}$ represents exactly one source tied to an actuator. In the method proposed in this paper, this property is ensured using time series clustering in conjunction with **P11**.

P 11. *Each source appears alone at least once in data*

The proposed method and associated matrix notations are summarized in Figure 1.

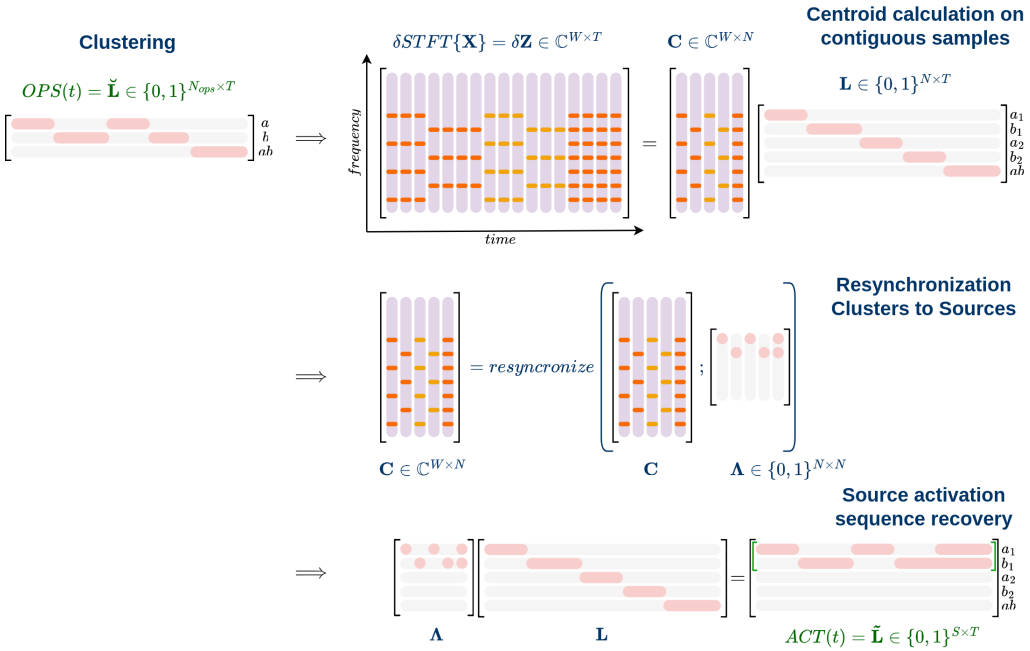

Figure 1: Source activation recovery steps

2.2. Clustering data into successive operations

A first step of the proposed method consists in segmenting the signal into successive operations. Clustering techniques prove useful in addressing this problem, by breaking down data into groups of similar objects [47]. In this context, clustering is applied to multivariate time series, and boils down to grouping timestamps according to a similarity measure between sample vectors. In hard clustering, each data point belongs to a single partition, whereas soft clustering allows for partial membership to different clusters. Time series clustering has been extensively tackled in the literature, and widely applied to structural health monitoring in particular, from fault detection [48, 49] to dynamic load identification [50].

Despite its simplicity, the *k-means* algorithm is considered here as a suitable candidate to segment the data when the number of operations N is known. Otherwise N is estimated from a dendrogram or by maximizing a criterion. Alternatives in which the number of clusters is not required exist

in the literature, yet it is often replaced by other hyperparameters. For instance, clustering techniques such as *DBSCAN* or its variant *OPTICS* [14] may be better suited in this setting.

In the proposed approach, clusters are sought using any relevant feature, obtained through linear or nonlinear transforms, so long as it exhibits piecewise stationarity. This paper focuses on periodic signals, for which phase-invariant spectral descriptors are well suited to clustering tasks (e.g., spectrogram, STFT magnitude, first four statistical moments applied to sliding windows, or any alternative suitable for stationary periodic signals).

Eventually, the one-hot encoded operation labels $\check{\boldsymbol{L}}$ are obtained. The centroids associated with these labels, cluster-wise average features, are of little use though. Indeed, nonlinear transformations were introduced, making the centroids lose their additive property (**P11**). That is, the centroid of a superimposed state must be equal to the sum of the centroids of its states. An operator is hence introduced in Section 2.3 to deduce consistent centroids suitable for the decomposition part.

2.3. Computing consistent complex centroids despite phase shifts

In this section, centroids are computed with a view to later expressing each operation as the sum of other operations. We seek a transform (applied to signal $\boldsymbol{X}$) such that (i) stationarity assumption **P6** is preserved in the feature space, and (ii) the additive property of the signal (**P10**) is kept throughout the transform. To that end, we propose a modified time-shifted STFT operator, denoted $\delta STFT\{.\}[m,k]$. The proposed linear operator leaves the phase unchanged across windows operating on the same signal.

Indeed, when expressing the input signal as a sum of sources, we notice that phase shifts occur from a window to another. Let x be a stationary signal over a single operation containing S simultaneously active harmonic sources. Each source s produces a signal with amplitude A_s, pulse ω_s, reference phase φ_s and phase shift ξ_s with respect to this reference. From this definition, an operation is characterized by source-dependent invariant parameters $\theta = \{[\omega_s, \varphi_s]\}_{s=1}^{S}$ and $\rho = [A_1, ..., A_S]^T$. Because an operation starts as the consequence of an event, the activation or deactivation of a source, entering an operation resets one of the time shifts ξ_s. This means that parameters $\Xi = [\xi_1, ..., \xi_S]^T$ are invariant only throughout one sub-operation, i.e., one realization of this operation. Parameters ρ, θ, Ξ are unknown. As illustrated in Figure 2, every time a source is switched off then on again after a non-integer number of periods, the initial phase of the source signal

rotates. As a result, the k-th component of the DFT of x over a window of the STFT, starting at time step m_0 and ending at time step m_1, indexed by $m \in [\![m_0, m_1]\!]$ has the form:

$$\begin{aligned} DFT_k\{x_{mH:mH+W-1;\rho,\theta,\Xi}\} &= \sum_{s=1}^{S} \sum_{n=0}^{W-1} A_s e^{\jmath\left(\omega_s n + \varphi_s + \xi_s - \omega_k n - \omega_k mH\right)} \\ &= A(\rho) e^{\jmath \boldsymbol{\varphi}(m;\theta,\Xi)} \end{aligned} \tag{4}$$

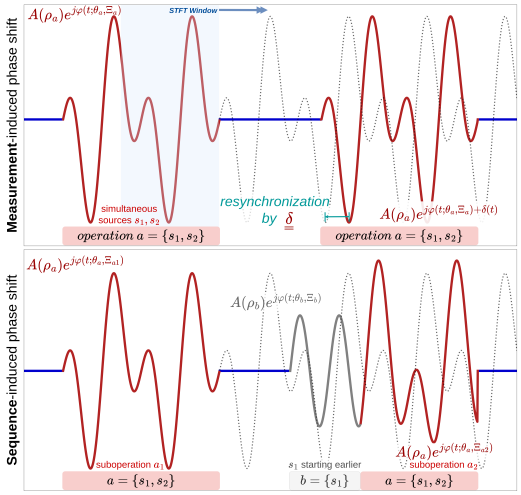

Figure 2: Phase discrepancies translated in the time domain

In order to recover one reference DFT for a sub-operation, we estimate the time lag δ by which the measured DFT should be shifted so the resulting feature remains constant over $[\![m_0, m_1]\!]$. Indeed, the event having led to this sub-operation may have occurred after a non-integer number of hops.

$$\begin{aligned} & \frac{\partial}{\partial m} DFT\{x_{\delta+mH:\delta+mH+W-1}; \rho, \theta, \Xi\} && = 0 \\ \iff & \frac{\partial}{\partial m} \boldsymbol{S}_{\delta+mH} DFT\{x_{0:W-1}; \rho, \theta, \Xi\} && = 0 \\ \iff & -(H + \frac{\partial \delta}{\partial m}) diag(\boldsymbol{\omega}) \boldsymbol{S}_{\delta+mH} DFT\{x_{0:W-1}; \rho, \theta, \Xi\} && = 0 \end{aligned} \tag{5}$$

Then either $A(\rho) = 0$ and shifting is irrelevant (trivial solution), δ is time linear, or the phase cancels out. Solving for δ yields its least squares estimate:

$$\delta = \frac{\boldsymbol{\omega}^T}{\boldsymbol{\omega}^T\boldsymbol{\omega}}\big(\boldsymbol{\varphi}(0;\theta,\Xi) + (2\pi p - mH)\mathbb{1}_W^T\big) \tag{6}$$

where $p \in \mathbb{Z}$ and $\mathbb{1}_W^T$ is the one-vector of size W. As expected, δ depends in turn on θ and Ξ, and $\delta[m] = \delta_0(\theta,\Xi) - mH$. The phase only conveys noise in the low energy regions of the spectrum though. Since the STFT is naturally sparse, the least squares solution is a very poor estimator in this case. Instead, we simply measure the phase $\varphi(m;\theta,\Xi)[K]$ at the maximum of amplitude — excluding the DC component — to estimate $\hat{\delta}[m]$:

$$\hat{\delta}[m] = \frac{\boldsymbol{\varphi}(m;\theta,\Xi)[K]}{\omega_K} \tag{7}$$

Alternatively, $\hat{\delta}[m]$ can be estimated using Equation 6 on the dominant frequencies. Or using an optimal state estimator on the successive realizations of $\delta[m]$ in the sub-operation $[\![m_0, m_1]\!]$, by noticing that $\delta[m]$ has strictly the same linear dynamics (provided that the phase is carefully unwrapped) and statistical properties as $\boldsymbol{\varphi}(m;\theta,\Xi)$.

A collection of centroids $\boldsymbol{C} = [\boldsymbol{C}_1, ..., \boldsymbol{C}_N]$ is obtained by averaging out the samples in every sub-operation with contiguous samples, with $N \geq N_{ops}$. Each sub-operation j holds sequence- and signal-dependent phase differences $\Xi^{(j)}$ with respect to the operation's reference θ. These different configurations for a single operation cannot be easily untangled, as illustrated on Figure 2. Decomposition is thus run on these centroids, and the optimization process proposed in Section 2.4 will naturally recombine these sub-operations using source resynchronization.

Spectral leakage constitutes yet another cause of error in estimating an operation's reference DFT, often due to a non-integer number of periods present in a window. Energy ends up distributed across the spectrum which results in undesired frequency components. Choosing an appropriate window function w alleviates this phenomenon.

A procedure to compute the δSTFT is proposed in Algorithm 1.

Algorithm 1: proposed $\delta STFT\{.\}[m,k]$

Input: Time series $\boldsymbol{X} \in \mathbb{R}^{T'}$; clustering labels $\breve{\boldsymbol{L}} \in \{0,1\}^{N_{ops} \times T}$;
Result: Piecewise constant DFTs over time $\delta\boldsymbol{Z} \in \mathbb{C}^{W \times T}$, and centroids $\boldsymbol{C} \in \mathbb{C}^{W \times N}$;
Step 1: Apply the STFT as $\boldsymbol{Z} \in \mathbb{C}^{W \times T}$, $\boldsymbol{Z} = STFT\{\boldsymbol{X}\}$;
Step 2: Extract features for decomposition
for $0 \leq m < T$ **do**
  Measure the phase at the maximum of magnitude (excluding the DC component)
  $\varphi_{max} = \angle \underset{\boldsymbol{z} \in \{\forall k > 0,\, \boldsymbol{Z}_m^k\}}{\arg\max} |\boldsymbol{z}|$;
  corresponding to the frequency bin with pulse ω_{max} and estimated time shift $\hat{\delta} = \frac{\varphi_{max}}{\omega_{max}} \frac{f_s}{2}$;
  Time shift the phase accordingly $\delta\boldsymbol{Z}_m \leftarrow \boldsymbol{S}_{\hat{\delta}} \boldsymbol{Z}_m$;
end
Lift clustering labels as $\boldsymbol{L}$ to represent only contiguous samples;
Step 3: Compute sub-centroids as $\boldsymbol{C} = \delta\boldsymbol{Z}\boldsymbol{L}^T diag\Big(\big(\frac{1}{\|\boldsymbol{L}^g\|_0}\big)_{1 \leq g \leq N}\Big)$;

Lastly, if present, the vector with the least root mean square (RMS) is removed from $\boldsymbol{C}$ as it relates to the *stand-by* operation. This operation corresponds to background noise or a persistent component detrimental to the decomposition problem (much like the neutral element of a set). Decomposing centroids instead of samples greatly reduces the computational complexity, as the matrix factorization no longer depends on the number of samples but the number of operations. The use of centroids is also more robust to noise.

2.4. Matrix decomposition as a resynchronization problem

In this section, the goal is to retrieve the actuators' activation sequences $\tilde{\boldsymbol{L}} \in \{0,1\}^{S \times T}$, given the elicited centroids $\boldsymbol{C} \in \mathbb{C}^{W \times N}$.

In this paper, we propose a convenient parameterization for both the dictionary and the representation, in which the optimization problem can be effectively regularized. Indeed, by computing the dictionary as a set of centroids from the $\delta STFT$, the atoms are forced to retain physical properties. This dictionary is then parameterized in the time lags $\boldsymbol{\Delta} \in \mathcal{I}_{\boldsymbol{\Delta}}$ required to optimally reconstruct each operation's centroid. Since the sources are periodic, so is the process of resynchronizing each source in a sum. The optimization could hence be carried out on $\cup_{c=1}^{N} [-\frac{\hat{T}^{(c)}}{2}, \frac{\hat{T}^{(c)}}{2}[^N$, where $\frac{1}{\hat{T}^{(c)}}$ is

the estimated fundamental frequency of each atom. Similarly, atoms are expressed as linear combinations of others. That is, the content of each operation is stacked in column form in matrix $\mathbf{\Lambda} \in \mathcal{I}_{\mathbf{\Lambda}}$, where each column is a collection of operations meant to be learnt in place of the sources' activation sequences. $\mathbf{\Lambda}$ thus constitutes a Rosetta Stone, translating each operation as its content in terms of other operations, or directly in terms of the underlying sources when the maximal decomposition is reached. The sought solutions lie in $\{0,1\}^{N\times N}$.

Even under these conditions, decomposition remains challenging. In particular, source resynchronization is well known to be highly non-convex and entails a combinatorial number of spurious minimizers [51]. This phenomenon is clearly illustrated on a synthetic use case in Figure 3. Moreover, the dictionary is redundant down to the time shifts, hence there exists a myriad of global minimizers for the representation as well [40], albeit only a handful are relevant.

We thus propose a novel formulation to overcome the outlined difficulties. Constraints are lifted to begin with. Tikhonov regularization is applied to the time shifts, allowing for an unconstrained optimization on $\mathcal{I}_{\mathbf{\Delta}} = \mathbb{R}^{N\times N}$ directly. This does not prevent the existence of a combinatorial number of local minima though. The representation suffers from many more causes of indetermination. Using the regularization approaches in Equation 8, detailed in Appendix A, the constraint on the representation can be lifted to optimize on $\mathcal{I}_{\mathbf{\Lambda}} = \mathbb{R}^{N\times N}$, instead of $\{0,1\}^{N\times N}$ which is NP-hard.

In the frequency domain, if an operation c with descriptor $\boldsymbol{C}_c$ can be decomposed as a sum of operations $\mathbf{\Lambda}_c$ given time-shifts $\mathbf{\Delta}_c$, then the $\mathbb{C}$SBMF is formulated as an optimization problem:

$$\inf_{\substack{\mathbf{\Delta}\in\mathcal{I}_{\mathbf{\Delta}},\\ \mathbf{\Lambda}\in\mathcal{I}_{\mathbf{\Lambda}},}} \sum_{c=1}^{N} \Bigg(\Big\| \boldsymbol{C}_c - \sum_{i=1}^{N} \boldsymbol{S}_{\Delta_c^i} \boldsymbol{C}_i \Lambda_c^i \Big\|_2^2 + \lambda \mathcal{L}_{col}(\mathbf{\Lambda}_c) + \mathcal{E}\mathcal{T}(\mathbf{\Lambda}_c) + \beta \mathcal{B}_2\big(\mathbf{\Lambda}_c\big) + \Gamma \|\mathbf{\Delta}_c\|_2^2 \Bigg) + L\mathcal{L}_{row}(\mathbf{\Lambda}) \tag{8}$$

$$\mathcal{L}_{col}(\mathbf{\Lambda}_c) = \big\|\mathbf{\Lambda}_c\big\|_p \tag{9}$$

$$\mathcal{T}(\mathbf{\Lambda}_c) = \frac{\big\|\mathbf{\Lambda}_c\big\|_p}{\|\boldsymbol{C}_c\|_2^2} \tag{10}$$

$$\mathcal{L}_{row}(\mathbf{\Lambda}) = \big\|\mathbf{\Lambda}^T\big\|_{2,p} \tag{11}$$

$$\mathcal{B}_2(\mathbf{\Lambda}_c) = \big\|\frac{1}{2}\mathbb{1} - |\mathbf{\Lambda}_c - \frac{1}{2}\mathbb{1}|\big\|_2 \tag{12}$$

where $\mathcal{L}_{col}$, $\mathcal{T}$, $\mathcal{B}_2$, $\|\boldsymbol{\Delta}_c\|_2^2$ and $\mathcal{L}_{row}$ denote the penalties associated with the regularization coefficients λ, $\mathcal{E}$, β, Γ, and L respectively. The least squares functional is denoted $F(\boldsymbol{\Delta}, \boldsymbol{\Lambda})$. $\|.\|_p$ is the ℓ_p norm (for $0 < p \leq 1$), and $\|\boldsymbol{\Lambda}\|_{2,p} = \left(\sum_{i=1}^{N} \|\boldsymbol{\Lambda}_i\|_2^p \right)^{\frac{1}{p}}$ is $\ell_{2,p}$ matrix norm. $\mathbb{1}$ is the one vector.

Sparsity is promoted in two ways, column-wise with the ℓ_p norm to sparsely decompose each atom, and row-wise with an $\ell_{2,p}$ penalization to fight the dictionary's redundancy. The latter is motivated by the fact that the $\ell_{2,p}$ norm is an adequate approximation of the $\ell_{2,0}$ norm which is the exact number of non-empty rows [52]. Here, the number of nonzero rows in $\boldsymbol{\Lambda}$ is the estimated number of sources.

A particularity of the proposed method is that the dictionary was built based on clustering. In compressed sensing, this is the worst choice for a dictionary since the most sparse solution is actually the one involving all atoms. That is, the trivial solution $\boldsymbol{\Lambda} = \boldsymbol{I}$ (identity) corresponds exactly to the clustering result. For instance, $abc = ab + c$ is more sparse than $abc = a + b + c$, yet the latter is sought. The competing objectives $\mathcal{L}_{col}$ and $\mathcal{L}_{row}$ remedy this situation. This calls for a subtle choice for L though, making $\mathcal{L}_{row}$ always greater than $\mathcal{L}_{col}$, and thus prioritizing the estimation of the number of sources.

Another unorthodox regularization term $\mathcal{T}$ is proposed. This term endows the column-wise sparse regularization parameter with a bias decreasing as the squared ℓ_2 norm of a suspected source increases. This penalty is crucial in that it avoids a pitfall arising in complex vector spaces: phase reversal. Indeed, phase resynchronization induces rotations. Hence in a resynchronized sum, vectors can flip and cancel out other components. As an example, $ab = a + b$ could be strictly equivalent (in cardinality and residual on $F(\boldsymbol{\Delta}, \boldsymbol{\Lambda})$) to $a = ab + b$. An arbitrary rule is required to distinguish these minima, since these combinations are algebraically equivalent (all satisfy the triangle inequality). Here, $\mathcal{T}$ is designed so the energy of the sum is higher than that of any of its constituents. The physical interpretation of $\mathcal{T}$ corresponds to the assumption that a collection of systems operated concurrently cannot draw less power than any of the underlying systems operated alone. While unlikely, independent sources may damp each other out and lead to a less energetic sum violating this assumption, as sometimes occurs in vibration mechanics. If this phenomenon is identified, moving $\|\boldsymbol{C}_c\|_2^2$ to the numerator of $\mathcal{T}$ reverses the order.

A binarity penalty $\mathcal{B}_2$ is added, similarly to the one proposed by Darabi

et al. [45]. Sparse regularization leads to a suitable approximation of $\boldsymbol{\Lambda}$, up to a factor since the ℓ_p norm draws the minimum towards zero. Binary regularization rectifies this, as well as any noise-originated discrepancy in the estimation of $\boldsymbol{\Lambda}$.

Overall, given a base dictionary $\boldsymbol{C} \in \mathbb{C}^{W \times N}$ and optimal regularization coefficients, Equation 8 admits non-equivalent minima on $\mathbb{R}^{N \times N} \times \mathbb{R}^{N \times N}$. The optimal representation $\boldsymbol{\Lambda}$ is meaningful in that it corresponds to the maximal binary decomposition of each atom. This claim is supported by the generic example presented in Appendix A, where the regularization mechanisms and their effect on the minima's locations are detailed. Source activation sequences are finally recovered as $\tilde{\boldsymbol{L}} = \boldsymbol{\Lambda}\boldsymbol{L}$.

2.5. Practical implementation

Despite the regularization terms introduced in Equation 8, the cost function is still highly non-convex and entails spurious local minimizers. There are also multiple hyperparameters on which depends the relevance and accuracy of the representation. Hyperparameter optimization has been extensively studied in the literature [53], yet the reliability of these methods remains limited, especially as the number of parameters to tune grows.

Alternate optimization strategies can aim towards the sought minimum [54], separately and gradually optimizing for the time shifts $\boldsymbol{\Delta}$ and the representation $\boldsymbol{\Lambda}$. Convergence cannot be guaranteed though. The effectiveness of such methods is therefore limited, especially as the time shifts of a combination do not inform on those of another combination.

For these reasons, optimizing with respect to both the time lags and the combinations at once may be too big a leap to ensure convergence towards the desired optimum. On an important note, any suboptimal solution to Equation 8 is completely useless for classification, as composite operations could be assigned a label distinct from the sources they contain.

Industrial applications come with a silver lining though. Sensors are often limited to the monitoring of a few systems at a time (which limits the number of sources), and these systems may not use their actuators in all possible configurations (which limits the number of operations). We hence advocate for a greedy algorithm to optimize for $\boldsymbol{\Lambda}$. Indeed, the entire parameter space is known and can be discretely mapped in a tractable way so long as the number of operations is reasonable (application-specific). That is, for each possible decomposition of operation c into a group of operations $\mathcal{G}$ indexed by g with $g_{(2)} = \boldsymbol{\Lambda}_c^{(g)}$ (notation for an integer expressed in base 2), a residual

r_g^c is compared to a threshold τ, an energy bound on additive noise, to accept or reject the decomposition.

Every element of the residual matrix $\boldsymbol{R} \in \mathbb{R}^{N \times 2^N}$ is found as the solution to the resynchronization problem between vectors $\boldsymbol{C}_c$ and $\{\boldsymbol{C}_i\}_{i \in \mathcal{G}}$:

$$r_g^c = \inf_{\{\Delta_c^i\}_{i \in \mathcal{G}}} \left\| \boldsymbol{C}_c - \sum_{i \in \mathcal{G}} \boldsymbol{S}_{\Delta_c^i} \boldsymbol{C}_i \right\|_2^2 \tag{13}$$

The Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [55] is used to compute these residuals. As a result, the residual matrix $\boldsymbol{R}$ illustrated in Figure 3 is obtained, each row indicating the possible combinations in $\{\boldsymbol{\Lambda}_c^{(g)}\}_{g=0}^{2^N}$ for a given operation with centroid $\boldsymbol{C}_c$.

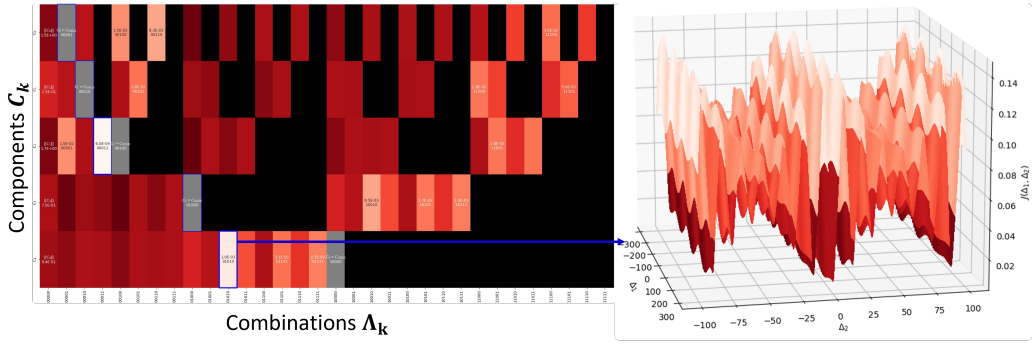

Figure 3: Residual matrix (left), after optimal source resynchronization (right)

The parameter space results from the following heuristic. For each operation $c \in [\![1, N]\!]$ and combination $g \in [\![1, 2^N - 1]\!]$ carrying indices $\mathcal{G}$:

- If $g = 0$, then $r_g^c = \|\boldsymbol{C}_c\|_2^2$ is the squared norm.
- If $g_{(2)} = (2^k)_{(2)}$, then $r_g^c = 0$ (trivial decomposition).
- If $\|\boldsymbol{C}_c\|_2 > \sum_{i \in \mathcal{G}} \|\boldsymbol{C}_i\|_2$, then exclude g (triangle inequality unsatisfied).
- If $\|\boldsymbol{C}_c\|_2^2 < \max_{i \in \mathcal{G}} \|\boldsymbol{C}_i\|_2^2$, then exclude g (energy-based ordering).

In the absence of the sparsity regularizers, multiple admissible combinations may be found. These minimizers bear different residuals due to noise and other sources of uncertainty, albeit centroids are naturally resilient in that respect. For this reason, the decomposition returned by the proposed

technique corresponds to the minimum number of sources to begin with, and only then the lowest residual is sought.

At last, the proposed ℂSBMF algorithm to retrieve the activation sequences is presented in Figure 4.

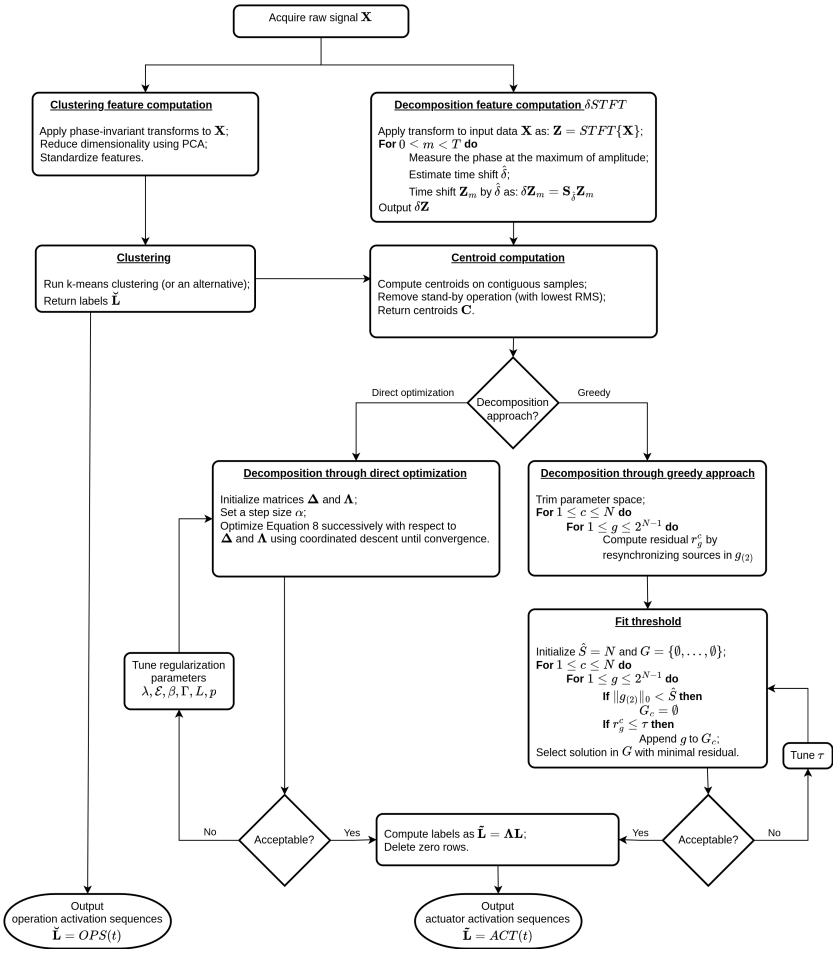

Figure 4: Flowchart of the ℂSBMF for source activation sequence retrieval

2.6. Limitations

There are cases in which machines operate at different regimes. For instance, a motor operating at constant speed with different loads, monitored by an accelerometer, might produce the same signature (same power spectrum), varying only by a scaling factor. Then recombining these sources during post-processing is straightforward. Should the actuator produce different

signatures under these regimes (different speeds in the previous example), recombination is not possible with the proposed technique.

Another practical issue is when distinct actuators produce the exact same signature. If run concurrently, this case is no different from a single device operated at different regimes. An indetermination hence remains between both cases. In the greedy algorithm, multiplicity can be taken into account by expressing the parameter space in base b, with b the maximum multiplicity, instead of base 2. The final complexity of this algorithm is $\mathcal{O}(Nb^N)$, times the optimizer's complexity as regards residual calculation. This remains acceptable for monitoring small dedicated systems.

3. Results

3.1. Numerical experiments

We verify our method against synthetic signals to begin with. A representative scenario was selected here among our numerical experiments. A piecewise stationary univariate signal x is produced as the sequence of all possible sums of sources from an alphabet $\mathcal{C}$, containing a square wave a (frequency $70Hz$, amplitude $1u$, zero-centered), a triangle wave b (frequency $50Hz$, amplitude $1u$, zero-centered) and a sine wave c (frequency $50Hz$, amplitude $2u$, zero-centered). Signal is supplemented with a zero-mean Gaussian noise $w(t)$ with standard deviation $\sigma = 0.1u$.

In order to shed light on practical limitations of existing methods, we compare qualitatively the $\mathbb{C}$SMBF to traditional and state-of-the-art techniques tackling similar problems. The results are presented in Figure 5. The proposed benchmark comprehends semi-binary matrix factorization (SBMF) [35], alpha-stable convolutional sparse coding (αCSC) [38, 39], as well as NMF. Sparse coding steps have been performed using the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [56].

The SBMF is limited to the study of real-valued signals. Less ambiguous than NMF, it captures "the direct sum (as opposed to the average) of community activities" [35]. Albeit similar to the proposed $\mathbb{C}$SBMF, it does not rely on assumption **P11** to build the dictionary and rather uses an SVD-based initialization. It is hence better at capturing intrinsic characteristics, yet concerns remain as to the validity of the result and its interpretation, as will testify the decomposition in Figure 5. Applied to spectrogram data, the SBMF accurately predicted the presence of the sine wave, but difficulties

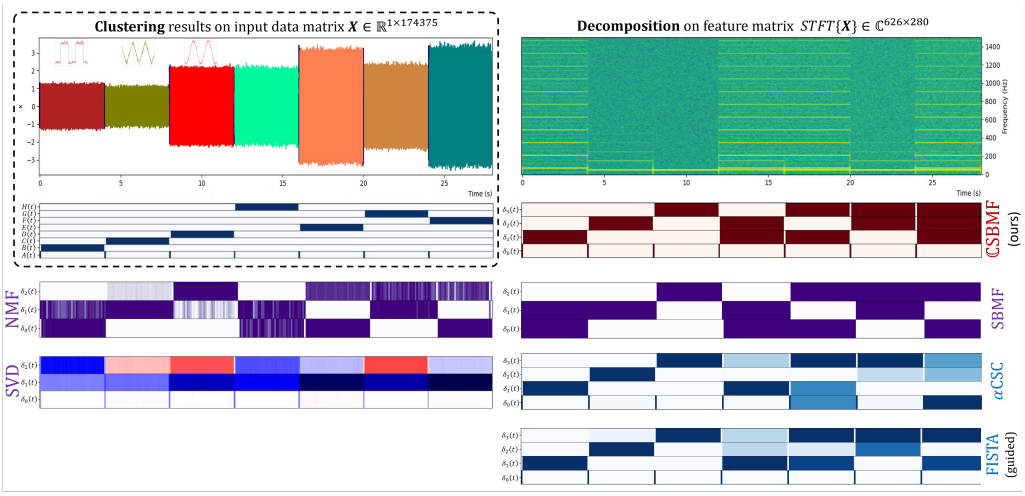

Figure 5: Decomposition of a synthetic signal ($70Hz$ square, $50Hz$ triangle and $50Hz$ sine waves, and their combinations). Six decomposition methods are presented: the ℂSBMF (red), dictionary learning techniques (purple) and sparse coding (blue). Non-binary labels lie in [0, 1] (scale is dictionary-dependent otherwise).

subsist in differentiating the triangle and square waves. We suspect this behavior is caused by noise and spectrogram-induced nonlinearity. To put the emphasis on the nonlinear aspect, the decomposition referred to as *guided FISTA* in our experiment uses the optimal dictionary directly (power spectral density of each wave). Projecting the spectrogram onto it, the triangle wave remains poorly identified.

In our investigations, the αCSC, applied to the time series directly, was found to excel at retrieving temporal patterns. By taking the DFT of these patterns to build the dictionary and projecting the STFT onto them, linearity is preserved. While properly identifying the waves alone, resynchronization is absent from this process, and indeed this method fails to retrieve the combinations.

In comparison to these techniques, the ℂSBMF reliably finds meaningful centroids using clustering, and effectively recovers the activation sequences. In our experiments, source resynchronization allowed to lower the residuals of the desired decompositions by at least two orders of magnitude with respect to their countepart computed using the modulus of the atoms. The performance of the proposed method is tied to the effectiveness of the clustering as well as the averaging process, which is affected by transients, outliers and

noise distributions. We therefore validate the ℂSBMF on real-world signals.

3.2. Experimental results

As a condensed version of an industrial system, we validate the ℂSBMF on the CAFFEINE dataset [44] presented in Figure 6, and more specifically on the current and vibration signals. This use case consists in an automated coffee machine, made of four multiphysical actuators: one *heating coil*, one vibration *pump*, one *infuser* (motor with a worm gear to displace the infuser) and one *grinder* (motor with an epicyclic gearing to grind coffee beans). The relevance of such a system for industrial applications was shown in [57].

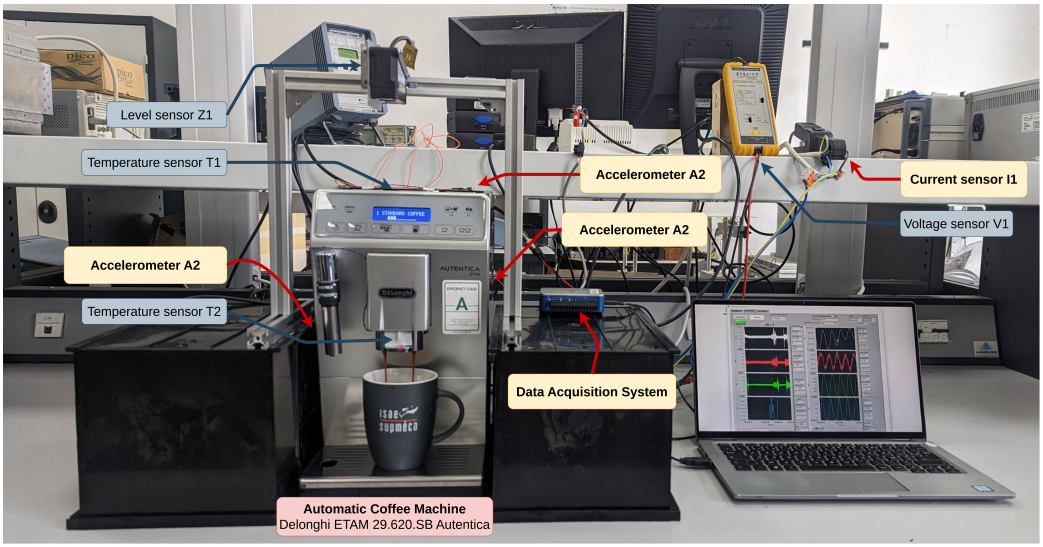

Figure 6: Experimental setup of the CAFFEINE dataset [44] using the NI USB-6003 data acquisition system (16bit, 8channels, 6.25kHz)

The results of the ℂSBMF applied to current sensor data are presented in Figure 7. The residual threshold was set to $\tau = 0.0315$. We observed that this hyperparameter was more sensitive as the sources' scales were unbalanced. Here, the heating coil consumes 65 times more than the infuser, and the infuser's RMS is only twice that of the background noise. On this dataset, the activation sequences of the three actuators are well recovered despite slight nonstationarities and noise. In order to push the boundary of our method, centroids of background noise and low consumption electronics were kept.

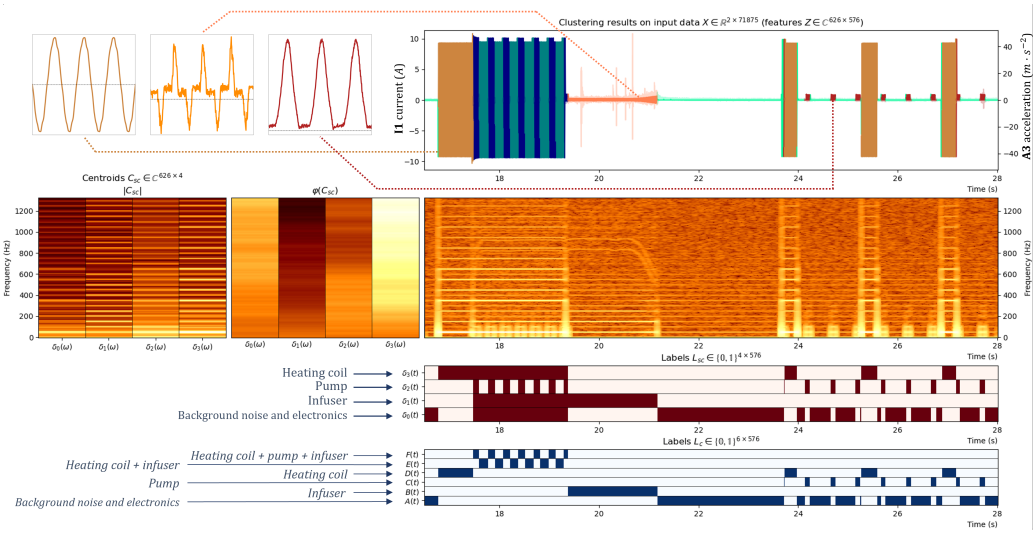

Figure 7: Decomposition of the current signal, trial 42 of the CAFFEINE dataset

In Figure 8, we apply the ℂSBMF on accelerometer data. These components produce extremely noisy (grinder) and partially nonstationary (infuser) signals. This behavior breaks the quasi-stationarity assumption, which translates into inaccurate centroid estimation and leads to incorrect classification. In spite of these challenging conditions, an alternate use for the ℂSBMF is to recombine wrongfully clustered centroids. Indeed, as an intra-cluster variance minimization algorithm, $k-means$ often performs better when overestimating the number of clusters. This isolates outliers. In this case, the outliers and the actuators activated in different operations share similarities, making it possible to regroup them in a meaningful fashion.

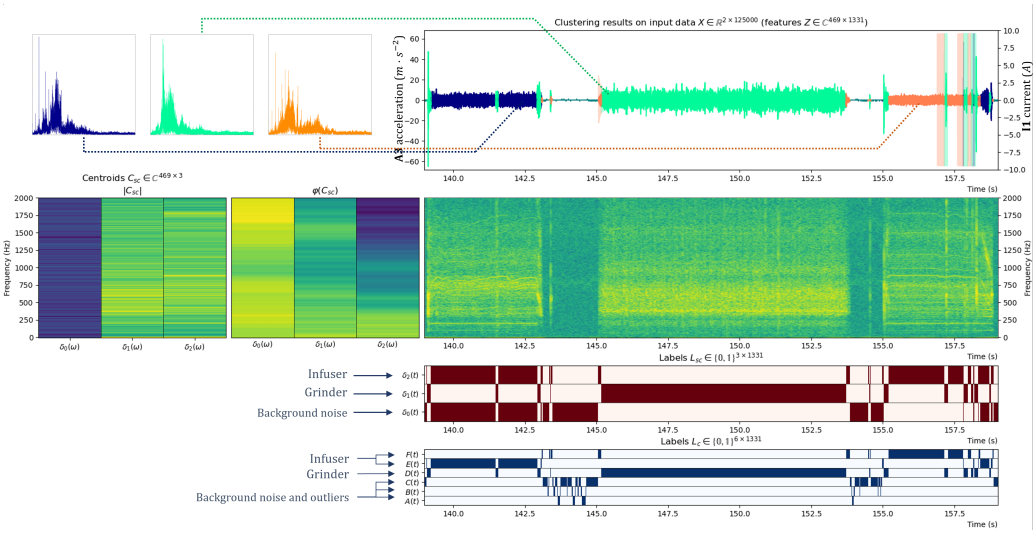

Figure 8: Decomposition of the accelerometer signal, trial 42 of the CAFFEINE dataset

4. Discussion

In this work, we highlight the importance of keeping data's extensive property to recover the activation sequences. Be it through convolutions in the time domain [36, 39], or phase resynchronization in the TF domain as proposed here, we qualify the matrix factorization as *exact* insofar as most of the signal information has been preserved. Although a phase-preserving decomposition is interesting, (i) other sources of uncertainty remain, noise and transients in particular, and (ii) the approach is inherently subject to the curse of dimensionality. Indeed, retaining the properties of the DFT forces to represent data in a high dimensional space, thus limiting the discrimination between centroids or samples. For this reason, using nonlinear transforms, or even simple standardization, to learn a more discriminating manifold would be beneficial [58]. This could also help lifting assumption **P10** to tackle cases where the subsystems are coupled, e.g., in series association.

The ℂSBMF heavily relies on clustering to build a dictionary. Whilst this is a debatable choice for compressed sensing, its relevance to source identification is made clear in this paper. This approach is particularly interesting as regards (i) computational complexity (decomposing centroids instead of samples, $N \ll T$), (ii) resilience to noise thanks to the averaging process, and (iii) estimation of the number of sources. Clustering also constitutes the method's Achilles' heel, as the decomposition is as accurate as the clustering

technique is. Centroids are also subject to transient-originated outliers, and their misestimation is detrimental to the factorization process. Robust kernel smoothing can be used to compute outlier-free centroids though [59].

At last, the ℂSBMF overlooked the case where a system is operated at different regimes. Periodic regularization functions could be considered to extend the representation's domain to $\mathbb{Z}$ ($\mathbb{N}$, ideally) [60], a common practice in quantization neural networks. Should the signature shift however smoothly from a regime to another, matrix decomposition is no longer appropriate. Graph neural networks (GNN) excel at this type of task, therefore making them good candidate architectures to learn an embedding in which sources can be easily separated [61]. Tracking could also be used to that effect, learning trajectories and their principal characteristics instead of centroids [42]. Overall, the rationale is that a single vector may prove insufficient to represent a source in multiple configurations.

5. Conclusions and perspective

In this paper, the ℂSBMF method was proposed to recover source activation sequences in mixed stationary periodic signals. This study highlighted limitations in traditional methods in challenging conditions, where the sources may be correlated and their number difficult to estimate. A formulation to this semi-binary decomposition was proposed as a phase-preserving bi-variate optimization problem. Although direct solving proved tedious and convergence could not be guaranteed, the proposed greedy algorithm stems from a meticulous study of this formulation. A novel operator, coined $\delta STFT$, was introduced in an effort to extract meaningful centroids in the complex plane, thus keeping the Fourier transform's linearity. Additionally, a phase resynchronization mechanism allowed to express centroids with respect to others, and thus find a minimal basis in which data can be reconstructed. Finally, due to spurious minimizers — both in dictionary and representation learning — jeopardizing the optimization process, and building up on the fact that only a tiny proportion of the representation's parameter space is actually relevant, a greedy algorithm was designed.

This work paves the way for interesting prospects. As a trade-off between effective variants of the NMF and the proposed ℂSBMF, an efficient parameter space reduction and search could be proposed by using a phase-invariant space as proxy before resynchronization, without loss of generalization. Future work will also aim for direct solving of Equation 8 using scalable opti-

mization techniques and building up a lower dimensional dictionary. That is, in light of this work, we suspect a more discriminating dictionary could be learnt appropriately. Although the absence of training is interesting, this limits the potential for performance gains. We believe deep-learning-based underdetermined blind source separation techniques will benefit from the findings presented in this paper, by avoiding phase-shift-induced pitfalls in particular.

Number of industrial use cases contain transient, non-stationary and disturbed signals. At present, our method is very sensitive to these non-stationarities, although experimental validation has shown it performed well under mild quasi-stationary conditions in a representative use case. Means to tackle greater levels of non-stationarity and outliers in source signals will hence be investigated.

Acknowledgments

This work was funded by the EnerMan project from the European Union's Horizon 2020 research and innovation programme under grant agreement No 958478. The authors would also like to thank Christophe Ben Brahim for their support in designing the experimental setup.

Appendix A. Detailed study of the ℂSBMF formulation

This section details the mechanisms at play in the ℂSBMF formulation (Equation 8). In particular, we highlight the existence of a global minimum, or at least we show through a generic example that the minimizers are not equivalent to one another under suitable regularization conditions. Inequalities for some of the regularization parameters are provided. This constrains the hyperparameters to tune and guarantees the validity of the expected properties of each penalty.

A simple sequence "$a-b-ab$" is considered as a synthetic use case, where a and b are normalized T-periodic triangle and square waves respectively. Operation ab is the sum of a and b, shifted by $T/4$ and $T/3$ respectively. F denotes the least squares functional of Equation 8 and J is the complete cost function including all regularizers.

Appendix A.1. Tikhonov regularization on the time shifts

Due to periodicity, there is an infinite number of minimizers enabling the resynchronization of each atom i to reconstruct another atom c using time

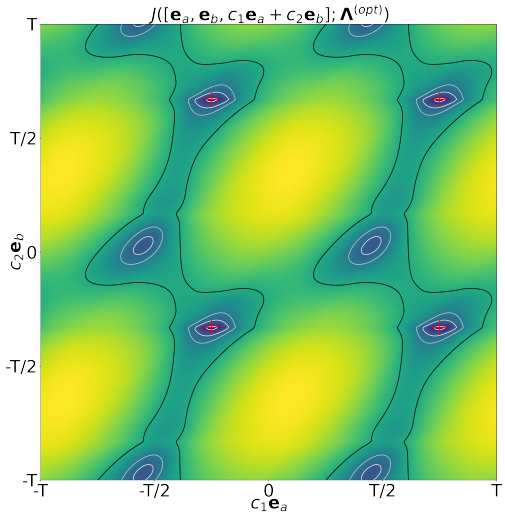

Figure A.9: No regularization

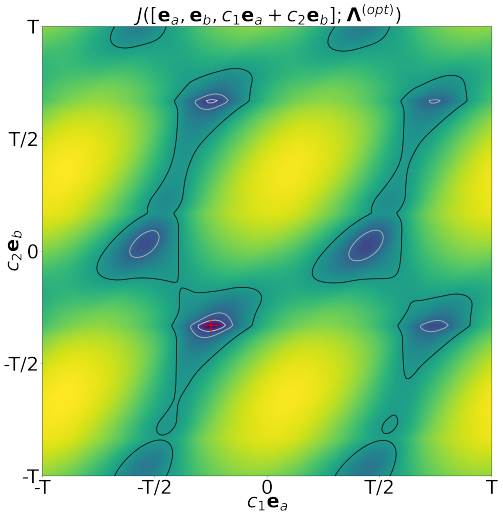

Figure A.10: Tikhonov regularization

Figure A.11: Minimizers of $F(\mathbf{\Delta}_{ab}, \mathbf{\Lambda}_{ab}^{opt})$ (marked in red, with level curves)

shifts $\Delta_c^i \in \mathbb{R}$. Indeed, the time shift operator is periodic, hence so is the functional $F(\mathbf{\Delta}, \mathbf{\Lambda})$. Each time shift is applied to a single atom though, $F(\mathbf{\Delta}, \mathbf{\Lambda})$ is hence periodic of period $\hat{T}^{(i)}$ along each dimension Δ_c^i. Given the desired representation $\mathbf{\Lambda}_{ab}^{(opt)} = [\boldsymbol{e}_a, \boldsymbol{e}_b, \boldsymbol{e}_a + \boldsymbol{e}_b]$, Figure A.11 shows the presence of multiple minimizers (T-periodic), with $\boldsymbol{e}_a$, $\boldsymbol{e}_b$, $\boldsymbol{e}_{ab}$ the canonical vectors. Uniqueness of the solution in $\mathbf{\Delta}$ is obtained through Tikhonov regularization ($\Gamma = 10^{-5}$).

Appendix A.1.1. Regularization of the representation

Three types of regularizations are implemented for (i) sparsity, (ii) binarity, and (iii) consistency as regards the energy profiles of the decompositions. Diverse experiments are conducted to study the effect of these regularizers. The following parameterization is used to represent the cost function in two dimensions: $\mathbf{\Lambda}_{ab} = c_1\boldsymbol{e}_a + c_2\boldsymbol{e}_b$ (Equation A.1), $\mathbf{\Lambda}_{ab} = (\frac{c_1+c_2}{2})\boldsymbol{e}_{ab}$ (Equation A.2), $\mathbf{\Lambda}_b = c_1\boldsymbol{e}_a + c_2\boldsymbol{e}_{ab}$ (Equation A.3) and $\mathbf{\Lambda}_a = c_1\boldsymbol{e}_{ab} + c_2\boldsymbol{e}_b$ (Equation A.4).

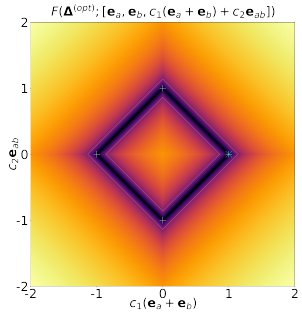

Figure A.12: None

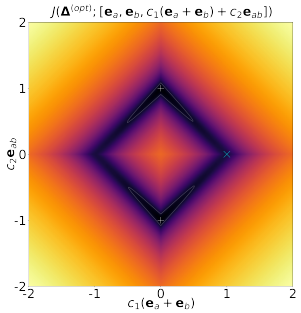

Figure A.13: ℓ_p

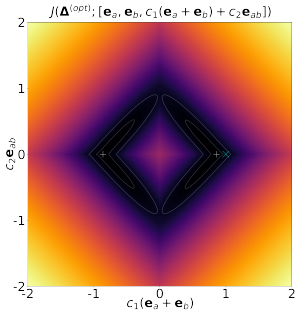

Figure A.14: ℓ_p and $\ell_{2,p}$

Figure A.15: Minimizers of $J_{sparse}(\mathbf{\Delta}_{ab}^{(opt)}, \mathbf{\Lambda}_{ab})$ (marked in white, with level curves)

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 & c_1 \\ 0 & 1 & c_1 \\ 0 & 0 & c_2 \end{bmatrix} \quad (A.1) \qquad \mathbf{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{c_1+c_2}{2} \end{bmatrix} \quad (A.2)$$

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & c_1 & 0 \\ 0 & 0 & 0 \\ 0 & c_2 & 1 \end{bmatrix} \quad (A.3) \qquad \mathbf{\Lambda} = \begin{bmatrix} 0 & 0 & 0 \\ c_2 & 1 & 0 \\ c_1 & 0 & 1 \end{bmatrix} \quad (A.4)$$

From this point on, it is assumed that the optimal time shifts $\mathbf{\Delta}^{opt}$ have been reached.

Appendix A.1.2. Sparsity regularization

The functional $F(\mathbf{\Delta}^{(opt)}, \mathbf{\Lambda})$ includes a number of minima as regards $\mathbf{\Lambda}$. Sparsity regularization usually consists in minimizing the ℓ_0 norm of a vector or an estimator thereof. This produces two noticeable effects: (i) it allows to reconstruct a sample with a minimal number of relevant atoms, (ii) negligible coefficients tend to zero. A difficulty in using a dictionary derived from clustering lies in the fact that the desired decomposition maximizes the number of relevant components instead of minimizing it. For this reason, the trivial solution $\mathbf{\Lambda} = \boldsymbol{I}$ always exists and it is the easiest minimum to find. To remedy this limitation, classic ℓ_p regularization on the columns of $\mathbf{\Lambda}$ is coupled with $\ell_{2,p}$ penalization on $\mathbf{\Lambda}^T$, with $p \leq 1$. The latter estimates the number of nonzero rows in $\mathbf{\Lambda}$ (number of sources). This mechanism is applied to the synthetic use case in Figure A.15, where J_{sparse} denotes the cost function including the functional as well as the column-wise ℓ_p and row-wise $\ell_{2,p}$ sparsity promoting penalties.

As illustrated in Figure A.12, an infinite number of minimizers are con-

nected through valleys as regards the representation in the absence of regularization. That is, the reconstruction is a weighted sum of several admissible combinations. When an element of $\mathbf{\Lambda}$ is negative, the corresponding atom may be flipped as a result of resynchronization (worst case). Figure A.13 shows the effect of column-wise sparsity regularization — pushing $\mathbf{\Lambda}$ towards the identity —, whereas Figure A.14 is the result of both column- and row-wise sparsity penalization, — minimizing the number of nonzero rows while penalizing negligible coefficients —. The scale is lost in the process though. The discrepancy may be observed on Figure A.14 between the minimum marked in white and the expected solution in cyan.

The tradeoff between sparse penalties is found by prioritizing the estimation of the number of sources:

$$\sum_{c=1}^{N} \left(\lambda + \frac{\mathcal{E}}{\|\boldsymbol{C}_c\|_2^2} \right) \|\mathbf{\Lambda}_c\|_p < L \|\mathbf{\Lambda}^T\|_{2,p} \tag{A.5}$$

which, by maximizing the left hand side (maximum decomposition $\mathbb{1}_{N-1}$, given the least energetic centroid $\mu_{min} = \min_c \|\boldsymbol{C}_c\|_2^2$) and minimizing the right hand side (one source), yields:

$$L > (\lambda + \frac{\mathcal{E}}{\mu_{min}}) N \|\mathbb{1}_{N-1}\|_p \tag{A.6}$$

Appendix A.1.3. Binary regularization

Binary regularization as it is referred to in this paper is a special case of quantization. The penalty is zero when $\mathbf{\Lambda}$ is binary. The expected behavior of this regularizer is to displace the desired minimum towards binary locations. This effect appears in Figure A.18.

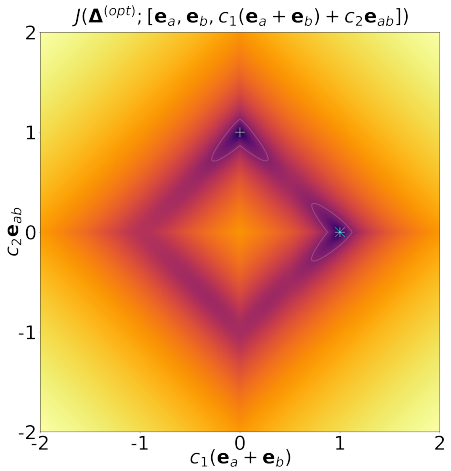

Figure A.16: Binary regularization only

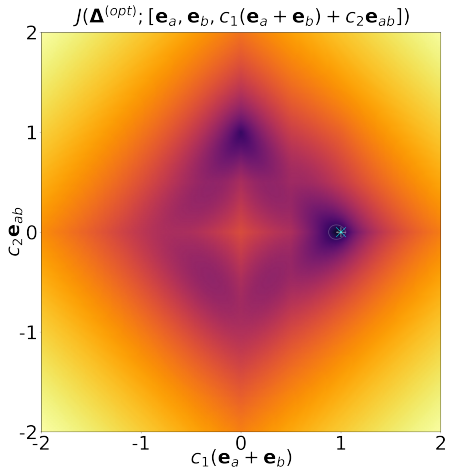

Figure A.17: Binary and sparse penalties

Figure A.18: Minimizers of $J(\mathbf{\Delta}_{ab}^{(opt)}, \mathbf{\Lambda}_{ab})$ (marked in red, with level curves)

Worth noting, the least penalized location is $\mathbf{\Lambda} = 0$. This location corresponds to the sum of the atoms' squared norms (all penalties are null). Hence care must be taken to have a higher cost at the center than at every minimizer's location $\mathbf{\Lambda}^{(opt)}$. This condition binds λ, L, $\mathcal{E}$ and Γ together as $J(\mathbf{\Delta}^{(opt)}, 0) < J(\mathbf{\Delta}^{(opt)}, \mathbf{\Lambda}^{(opt)})$, where J is the complete cost function, in which $\mathcal{B}_2$ can be considered null.

Appendix A.1.4. Regularization for iso-cardinality combinations

At last, differentiating between combinations with the same cardinality involving the same atoms is problem-specific. Geometrically, these solutions are equivalent, albeit the penalty on their respective time shifts may differ. The regularization term $\mathcal{T}$ defined in Equation 10 is proposed to tell these solutions apart, by prioritizing solutions in which the atom to reconstruct has the highest energy level. That is, for two minimizers c and j, if $\|\mathbf{\Lambda}^{(c)}\|_p = \|\mathbf{\Lambda}^{(j)}\|_p$ and $\|\boldsymbol{C}_c\|_2^2 > \|\boldsymbol{C}_j\|_2^2$, then coefficients $\mathcal{E}$ and Γ must be such that $J(\mathbf{\Delta}^{(c)}, \mathbf{\Lambda}^{(c)}) < J(\mathbf{\Delta}^{(j)}, \mathbf{\Lambda}^{(j)})$, where $\mathcal{L}_{col}, \mathcal{L}_{row}$ cancel out and $\mathcal{B}_2, F$ are zero. As shown on A.20, where all regularizers are present but $\mathcal{T}$, without further consideration iso-cardinality combinations correspond to equivalent minima.

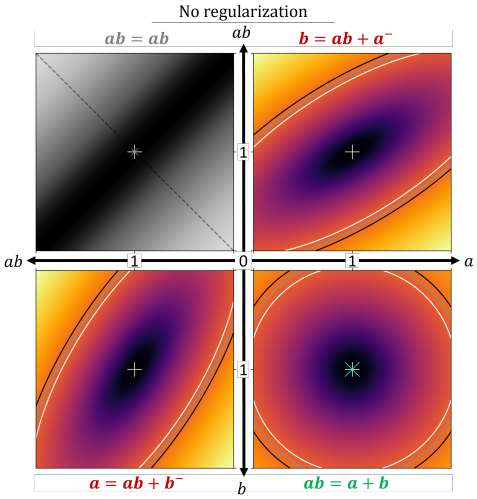

Figure A.19: No regularization

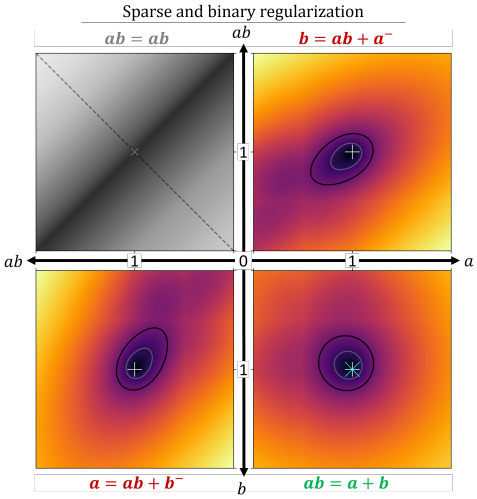

Figure A.20: Binary and sparse only

Figure A.21: Minimizers of $F(\mathbf{\Delta}^{(opt)}, \mathbf{\Lambda})$ (marked in red, with level curves)

Finally, applying all penalties, a single minimum remains, as illustrated in Figure A.22. The values used to regularize the problem in this example are as follows: $\Gamma = 10^{-5}$, $\lambda = 0.0225$, $L = 3.43$, $\beta = 0.7$, $\mathcal{E} = 0.1$ and $p = 0.9$.

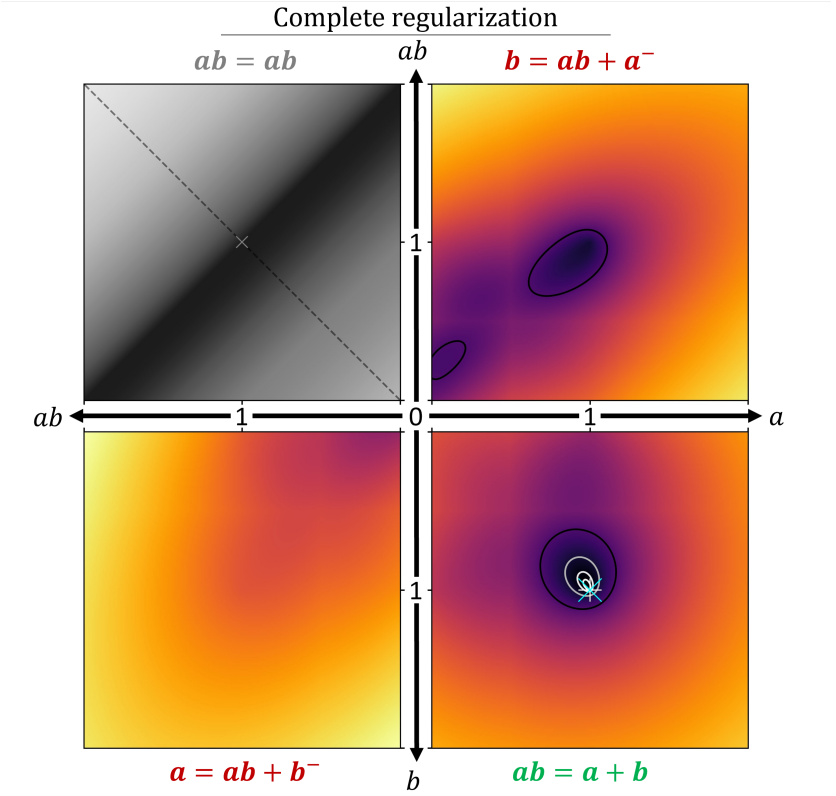

Figure A.22: Complete regularization (Equation 8)

References

[1] IEA, World energy outlook 2022, IEA Paris, France, 2022.
URL `https://www.iea.org/reports/world-energy-outlook-2022`

[2] C. Semeraro, M. Lezoche, H. Panetto, M. Dassisti, Digital twin paradigm: A systematic literature review, Computers in Industry 130 (2021) 103469.

[3] P. A. Schirmer, I. Mporas, Non-intrusive load monitoring: A review, IEEE Transactions on Smart Grid (2022).

[4] A. Faustine, N. H. Mvungi, S. Kaijage, K. Michael, A survey on non-intrusive load monitoring methodies and techniques for energy disaggregation problem, arXiv preprint arXiv:1703.00785 (2017).

[5] H. Lange, M. Bergés, Variational bolt: Approximate learning in factorial hidden markov models with application to energy disaggregation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[6] J. Kelly, W. Knottenbelt, Neural nilm: Deep neural networks applied to energy disaggregation, in: Proceedings of the 2nd ACM international conference on embedded systems for energy-efficient built environments, 2015, pp. 55–64.

[7] C. Zhang, M. Zhong, Z. Wang, N. Goddard, C. Sutton, Sequence-to-point learning with neural networks for non-intrusive load monitoring, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 32, 2018.

[8] R. Hennequin, R. Badeau, B. David, Nmf with time–frequency activations to model nonstationary audio events, IEEE Transactions on Audio, Speech, and Language Processing 19 (4) (2010) 744–753.

[9] R. Wang, Y. Zhang, L. Yu, J. Antoni, Q. Leclère, W. Jiang, A probability model with variational bayesian inference for the complex interference suppression in the acoustic array measurement, Mechanical Systems and Signal Processing 191 (2023) 110181.

[10] M. Matsumoto, Y. Fujimoto, Y. Hayashi, Energy disaggregation based on semi-binary nmf, in: Machine Learning and Data Mining in Pattern Recognition: 12th International Conference, MLDM 2016, New York, NY, USA, July 16-21, 2016, Proceedings, Springer, 2016, pp. 401–414.

[11] J. Wodecki, R. Zdunek, A. Wyłomańska, R. Zimroz, Local fault detection of rolling element bearing components by spectrogram clustering with semi-binary nmf, Diagnostyka 18 (2017).

[12] L. Liang, X. Ding, H. Wen, F. Liu, Impulsive components separation using minimum-determinant kl-divergence nmf of bi-variable map for bearing diagnosis, Mechanical Systems and Signal Processing 175 (2022) 109129.

[13] M. Gabor, R. Zdunek, R. Zimroz, J. Wodecki, A. Wylomanska, Nonnegative tensor factorization for vibration-based local damage detection, Mechanical Systems and Signal Processing 198 (2023) 110430.

[14] Q. Wang, Y. Zhang, S. Yin, Y. Wang, G. Wu, A novel underdetermined blind source separation method based on optics and subspace projection, Symmetry 13 (9) (2021) 1677.

[15] B. Loesch, B. Yang, Source number estimation and clustering for underdetermined blind source separation, in: Proc. IWAENC, 2008.

[16] Y. Xie, K. Xie, Z. Wu, S. Xie, Underdetermined blind source separation of speech mixtures based on k-means clustering, in: 2019 Chinese Control Conference (CCC), IEEE, 2019, pp. 42–46.

[17] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: Proceedings of the 26th annual international conference on machine learning, 2009, pp. 689–696.

[18] P. Stoica, H. Li, J. Li, Amplitude estimation of sinusoidal signals: survey, new results, and an application, IEEE Transactions on Signal Processing 48 (2) (2000) 338–352.

[19] R. Schmidt, Multiple emitter location and signal parameter estimation, IEEE transactions on antennas and propagation 34 (3) (1986) 276–280.

[20] R. Roy, T. Kailath, Esprit-estimation of signal parameters via rotational invariance techniques, IEEE Transactions on acoustics, speech, and signal processing 37 (7) (1989) 984–995.

[21] Z. Liu, F. Duan, G. Niu, D. Ye, J. Feng, Z. Cheng, X. Fu, J. Jiang, J. Zhu, M. Liu, Reconstruction of blade tip-timing signals based on the music algorithm, Mechanical Systems and Signal Processing 163 (2022) 108137.

[22] C. Xu, J. Wang, S. Yin, M. Deng, A focusing music algorithm for baseline-free lamb wave damage localization, Mechanical Systems and Signal Processing 164 (2022) 108242.

[23] S. L. Kiser, M. Rébillat, M. Guskov, N. Ranc, Real-time sinusoidal parameter estimation for damage growth monitoring during ultrasonic very high cycle fatigue tests, Mechanical Systems and Signal Processing 182 (2023) 109544.

[24] M. Lasserre, S. Bidon, O. Besson, F. Le Chevalier, Bayesian sparse fourier representation of off-grid targets with application to experimental radar data, Signal Processing 111 (2015) 261–273.

[25] S. G. Kim, C. D. Yoo, Underdetermined independent component analysis by data generation, in: Independent Component Analysis and Blind Signal Separation: Fifth International Conference, ICA 2004, Granada, Spain, September 22-24, 2004. Proceedings 5, Springer, 2004, pp. 445–452.

[26] Y. Zheng, I. Ng, K. Zhang, On the identifiability of nonlinear ica: Sparsity and beyond, Advances in Neural Information Processing Systems 35 (2022) 16411–16422.

[27] R. Gribonval, S. Lesage, A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges, in: ESANN'06 proceedings-14th European Symposium on Artificial Neural Networks, d-side publi., 2006, pp. 323–330.

[28] Y. Xu, J. M. Brownjohn, D. Hester, Enhanced sparse component analysis for operational modal identification of real-life bridge structures, Mechanical Systems and Signal Processing 116 (2019) 585–605.

[29] X. Zhao, B. Ye, Similarity of signal processing effect between hankel matrix-based svd and wavelet transform and its mechanism analysis, Mechanical Systems and Signal Processing 23 (4) (2009) 1062–1075.

[30] Z. Feng, M. Liang, F. Chu, Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples, Mechanical systems and signal Processing 38 (1) (2013) 165–205.

[31] B. Hou, D. Wang, T. Xia, Z. Peng, K.-L. Tsui, Difference mode decomposition for adaptive signal decomposition, Mechanical Systems and Signal Processing 191 (2023) 110203.

[32] P. J. Schmid, Dynamic mode decomposition of numerical and experimental data, Journal of fluid mechanics 656 (2010) 5–28.

[33] J. L. Proctor, S. L. Brunton, J. N. Kutz, Dynamic mode decomposition with control, SIAM Journal on Applied Dynamical Systems 15 (1) (2016) 142–161.

[34] J. Wodecki, A. Michalak, R. Zimroz, T. Barszcz, A. Wyłomańska, Impulsive source separation using combination of nonnegative matrix factorization of bi-frequency map, spatial denoising and monte carlo simulation, Mechanical Systems and Signal Processing 127 (2019) 89–101.

[35] M. Sørensen, N. D. Sidiropoulos, A. Swami, Overlapping community detection via semi-binary matrix factorization: Identifiability and algorithms, IEEE Transactions on Signal Processing 70 (2022) 4321–4336.

[36] H. Zhou, J. Chen, G. Dong, R. Wang, Detection and diagnosis of bearing faults using shift-invariant dictionary learning and hidden markov model, Mechanical systems and signal processing 72 (2016) 65–79.

[37] R. Grosse, R. Raina, H. Kwong, A. Y. Ng, Shift-invariance sparse coding for audio classification, arXiv preprint arXiv:1206.5241 (2012).

[38] M. Jas, T. Dupré la Tour, U. Simsekli, A. Gramfort, Learning the morphology of brain signals using alpha-stable convolutional sparse coding, Advances in Neural Information Processing Systems 30 (2017).

[39] T. Dupré la Tour, T. Moreau, M. Jas, A. Gramfort, Multivariate convolutional sparse coding for electromagnetic brain signals, Advances in Neural Information Processing Systems 31 (2018).

[40] H. Wang, G. Dong, J. Chen, X. Hu, Z. Zhu, A novel dictionary learning named deep and shared dictionary learning for fault diagnosis, Mechanical Systems and Signal Processing 182 (2023) 109570.

[41] J. Antoni, S. Chauhan, Second order blind source separation techniques (so-bss) and their relation to stochastic subspace identification (ssi) algorithm, in: Structural Dynamics, Volume 3: Proceedings of the 28th IMAC, A Conference on Structural Dynamics, 2010, Springer, 2011, pp. 177–187.

[42] R. Delabeye, M. Ghienne, J.-L. Dion, Sequential harmonic component tracking for underdetermined blind source separation in a multi-target tracking framework (2023).

[43] R. Delabeye, M. Ghienne, A. Kosecki, J.-L. Dion, Unsupervised manufacturing process identification using non-intrusive sensors, in: 25ème Congrès Français de la Mécanique, 2022.

[44] R. Delabeye, A. Baldassarre, C. B. Brahim, A. Kosecki, J.-L. Dion, M. Ghienne, O. Penas, F. Renaud, N. Peyret, Caffeine dataset (Jul. 2023). doi:10.5281/zenodo.8351431.
URL https://doi.org/10.5281/zenodo.8351431

[45] S. Darabi, M. Belbahri, M. Courbariaux, V. P. Nia, Bnn+: Improved binary network training (2018).

[46] M. B. Priestley, Non-linear and non-stationary time series analysis, London: Academic Press (1988).

[47] P. Rai, S. Singh, A survey of clustering techniques, International Journal of Computer Applications 7 (12) (2010) 1–5.

[48] P. Gangsar, R. Tiwari, Signal based condition monitoring techniques for fault detection and diagnosis of induction motors: A state-of-the-art review, Mechanical systems and signal processing 144 (2020) 106908.

[49] A. Rai, S. H. Upadhyay, Bearing performance degradation assessment based on a combination of empirical mode decomposition and k-medoids clustering, Mechanical Systems and Signal Processing 93 (2017) 16–29.

[50] Y. Liu, L. Wang, M. Li, Z. Wu, A distributed dynamic load identification method based on the hierarchical-clustering-oriented radial basis function framework using acceleration signals under convex-fuzzy hybrid uncertainties, Mechanical Systems and Signal Processing 172 (2022) 108935.

[51] A. Gossard, F. de Gournay, P. Weiss, Spurious minimizers in non uniform fourier sampling optimization, Inverse Problems 38 (10) (2022) 105003.

[52] M. Zhao, M. Lin, B. Chiu, Z. Zhang, X.-s. Tang, Trace ratio criterion based discriminative feature selection via l2, p-norm regularization for supervised learning, Neurocomputing 321 (2018) 1–16.

[53] R. Fuentes, R. Nayek, P. Gardner, N. Dervilis, T. Rogers, K. Worden, E. Cross, Equation discovery for nonlinear dynamical systems: A bayesian viewpoint, Mechanical Systems and Signal Processing 154 (2021) 107528.

[54] L. Yu, J. Antoni, H. Zhao, Q. Guo, R. Wang, W. Jiang, The acoustic inverse problem in the framework of alternating direction method of multipliers, Mechanical Systems and Signal Processing 149 (2021) 107220.

[55] R. Fletcher, Practical methods of optimization, John Wiley & Sons, 2013.

[56] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM journal on imaging sciences 2 (1) (2009) 183–202.

[57] R. Delabeye, O. Penas, R. Plateaux, Scalable ontology-based v&v process for heterogeneous systems and applications, in: Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings, 2022, pp. 341–350.

[58] R. Balshaw, P. S. Heyns, D. N. Wilke, S. Schmidt, Importance of temporal preserving latent analysis for latent variable models in fault diagnostics of rotating machinery, Mechanical Systems and Signal Processing 168 (2022) 108663.

[59] P. Humbert, B. Le Bars, L. Minvielle, Robust kernel density estimation with median-of-means principle, in: International Conference on Machine Learning, PMLR, 2022, pp. 9444–9465.

[60] M. Naumov, U. Diril, J. Park, B. Ray, J. Jablonski, A. Tulloch, On periodic functions as regularizers for quantization of neural networks, arXiv preprint arXiv:1811.09862 (2018).

[61] O. Shchur, S. Günnemann, Overlapping community detection with graph neural networks, arXiv preprint arXiv:1909.12201 (2019).