Alternating group lasso for block-term tensor decomposition with application to ECG source separation

José Henrique de Morais Goulart, Pedro Marinho R. de Oliveira, Rodrigo Cabral Farias, Vicente Zarzoso, Pierre Comon

Alternating Group Lasso for Block-Term Tensor Decomposition and Application to ECG Source Separation

J. H. de M. Goulart, P. M. R. de Oliveira, R. C. Farias, V. Zarzoso, *Senior Member, IEEE,* and P. Comon, *Fellow, IEEE*

***Abstract*—In some applications, blind source separation can be performed by computing an approximate block-term tensor decomposition (BTD), under much milder constraints than matrix-based techniques. However, choosing the BTD model structure (*i.e.*, the number of blocks and their ranks) is a difficult problem, and the standard least-squares formulation can be ill-posed. This paper proposes an alternating group lasso algorithm to compute approximate low-rank BTDs. It solves, in a provably convergent manner, a well-posed mixed-norm regularized tensor approximation problem that allows estimating the model parameters and its structure jointly. A variant is also put forward for dealing with linearly constrained blocks, motivated by the problem of blind separation of sums of complex exponentials, which can be cast as a low-rank Hankel-structured block-term tensor approximation problem. An experimental comparison with a standard nonlinear least-squares algorithm on synthetic tensor data indicates that the proposed algorithm is much more robust with respect to initialization. We also apply the constrained variant to the extraction of atrial activity from semi-synthetic and real-world electrocardiogram recordings during atrial fibrillation episodes. Our results show its ability to consistently select an adequate structure and to extract multiple signals which can be physiologically interpreted as atrial fibrillation patterns.**

***Index Terms*—Tensors, block term decomposition, group lasso, structured low-rank approximation, atrial fibrillation.**

I. Introduction

RECORDED signals in biomedical applications, such as electroencephalography [1] and electrocardiography [2], [3], [4], [5] can be modeled as instantaneous unknown linear mixtures of R sources. To separate them, the unknown sources are assumed to be statistically independent and orthogonal, in independent component analysis (ICA) and principal component analysis (PCA), respectively. Although this renders the underlying blind source separation (BSS) well-posed, such stringent assumptions may lack physiological grounds, hindering results interpretation. A less constraining approach is to assume that sources can be approximated by sums of complex exponentials (SCE), which can model narrowband and transient signals that can be linked to specific clinical conditions. This leads to an identifiable low-rank matrix factorization involving a Vandermonde matrix [6], but the total number of exponentials that can be identified by this matrix approach is bounded by the number of sensors.

Alternatively, by forming a matrix with shifted versions of each observed signal (*e.g.*, each channel output) and by stacking these matrices in a third-order tensor, a process called "Hankelization," BSS of SCE signals can be cast as an approximate block-term tensor decomposition (BTD) problem with Hankel-structured blocks [7]. Each SCE source contributes in the decomposition with a structured term given by a tensor product of a rank-L_r Hankel matrix containing its samples with a vector containing its spatial signature (*i.e.*, its weights for each channel), where L_r is the number of complex exponentials (poles) in the SCE source. The main benefit of this approach is that identifiability of the decomposition is guaranteed under mild conditions which do not involve stringent constraints such as orthogonality or independence and can hold even when the sum of the block ranks exceeds the number of sensors and the dimensions of the Hankel matrices as well [8], [7].

When implementing this approach, however, two problems arise. The first one refers to the computation of the approximate BTD. Some methods have been made available for this task, such as the nonlinear least-squares methods implemented in Tensorlab [9] or the alternating least squares method with enhanced line search (ALS-ELS) [10]. However, the performance of such techniques depends considerably on the choice of structural parameters (R and L_r), which is difficult to make in practice. Furthermore, they usually exhibit strong sensitivity with respect to the initialization and address an optimization problem which may lack a global minimizer [11]; this can lead to the estimation of almost collinear blocks with no physical interpretation, a phenomenon often termed model degeneracy. The second problem refers to the low-rank Hankel structure constraint, which is hard to enforce for real-valued data. To our knowledge, no existing BTD algorithm can reliably impose Hankel constraints in this case.

In [12], a functional promoting group sparsity of the decomposition factor columns was minimized to estimate appropriate structural parameters of an (unconstrained) BTD model, but not the model itself. Moreover, the authors only considered the case where all L_r are equal. In this work, we show that (i) the same principle can be used to *jointly* estimate the model structure and its parameters, and that (ii) the general case with different block ranks can also be addressed.

To achieve this goal, we formulate the approximate BTD

J. H. de M. Goulart and P. Comon are with Univ. Grenoble Alpes, CNRS, Gipsa-Lab, F-38000 Grenoble, France ({henrique.goulart, pierre.comon}@gipsa-lab.fr).

P. M. R. de Oliveira, R. C. Farias and V. Zarzoso are with the Univ. Côte d'Azur, CNRS, I3S Laboratory, 06900 Sophia-Antipolis, France ({marinho, cabral, zarzoso}@i3s.unice.fr).

computation as the minimization of a (least-squares) fitting term plus a regularization term given by the sum of the $\ell_{2,1}$-norms of the matrices containing the model parameters. The latter enforces (column-wise) group sparsity of the factor matrices, thus penalizing models with high R and L_r. In this way, we are able to find an approximate BTD without assumptions on R and L_r, effectively achieving a trade-off between data fitting and model complexity. Furthermore, the regularization renders the BTD approximation problem well-posed.

To solve this problem, we propose an algorithm called *alternating group lasso* (AGL) that is provably convergent and simpler than the algorithm of [12], as the latter sequentially employs three different ADMM (alternating directiom method of multipliers) schemes to first estimate the model structure and then its parameters. We also devise a variant termed constrained AGL (CAGL) to deal with linear (subspace) constraints over block matrices. For this purpose, a structured low-rank approximation method is applied on the block matrices at each iteration to ensure that they have low rank and belong to the specified subspace.

As an application of CAGL, we consider the problem of extracting the atrial activity (AA) in electrocardiograms (ECGs) with atrial fibrillation (AF) pattenrs. AF is the most common sustained cardiac arrhythmia encountered in clinical practice, a major public health and economical concern. The importance given to this challenging cardiac condition has increased in the past few years, since its mechanisms are not completely understood. Accurate analysis of the fibrillatory waves (f-waves) is then necessary for better understanding the arrhythmia. Noninvasive AA extraction from ECG recordings is, therefore, a key problem that motivates the development of signal processing techniques such as that proposed in this paper.

Other BSS techniques such as PCA and ICA have been applied to noninvasive AA extraction [13]–[14] and provided satisfactory results. However, as previously stated, results are difficult to interpret due to the imposed constraints. Motivated by identifiability and interpretability requirements, BTD modeling using "Hankelization" for AA extraction has been recently proposed and studied in [2], [15], [3], [4], [5]. Experimental results in synthetic and real AF ECG data showed that BTD can outperform the matrix-based techniques for AA extraction in short and long segments of AF ECG recordings. The application in this work follows the same line of [2], [15], [3], [4], [5], but differs from [2], [15], [4], [5], since we do not impose the structural parameters of the model. It also differs from [3], where structural parameters are chosen fitting different BTD models and selecting the best one with respect to an information-theoretical criterion.

The outline of the paper is as follows: in Section II we present the problem of SCE source separation using the BTD with Hankel structure. Section III formulates a regularized BTD approximation problem and the AGL and CAGL algorithms. A numerical evaluation of AGL is shown in Section IV. The application of CAGL to AA extraction is then presented in Section V, where both semi-synthetic and real-world ECG datasets are considered. Finally, concluding remarks are made in Section VI.

Notation: Tensors are denoted in uppercase bold script letters $\boldsymbol{\mathcal{X}}$, matrices in uppercase bold letters $\mathbf{X}$, vectors in lowercase bold letters $\mathbf{x}$ and scalars in lowercase letters x. The notations $\|\cdot\|_F$ and $\|\cdot\|_{2,1}$ stand for the Frobenius norm and the matrix mixed $\ell_{2,1}$-norm, respectively. The latter is defined as the sum of the norms of its argument's columns, as in

$$\|\mathbf{X}\|_{2,1} = \textstyle\sum_{r=1}^{R} \|\mathbf{x}_r\|_2. \tag{1}$$

The symbols $\otimes$, $\boxtimes$ and $\odot$ are used for outer (tensor), Kronecker, and Khatri-Rao (column-wise Kronecker) product, respectively. The block Khatri-Rao product $\odot_L$ is such that, given $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_R\end{bmatrix} \in \mathbb{C}^{K\times R}$ and $\mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 & \cdots & \mathbf{y}_{LR}\end{bmatrix} \in \mathbb{C}^{M\times LR}$, it yields $\mathbf{Z} = \mathbf{X} \odot_L \mathbf{Y} = \begin{bmatrix}\mathbf{Z}_1 & \cdots & \mathbf{Z}_R\end{bmatrix} \in \mathbb{C}^{KM\times LR}$ where $\mathbf{Z}_r = \mathbf{x}_r \boxtimes \begin{bmatrix}\mathbf{y}_{(r-1)L+1} \cdots \mathbf{y}_{(r-1)L+L}\end{bmatrix} \in \mathbb{C}^{KM\times L}$. $\mathrm{Diag}(\mathbf{x})$ denotes a diagonal with the components of $\mathbf{x}$ on its diagonal, and the $\mathrm{vec}(\cdot)$ operator maps a matrix $\begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_R\end{bmatrix} \in \mathbb{C}^{K\times R}$ into the vector $\begin{bmatrix}\mathbf{x}_1^\mathsf{T} & \ldots & \mathbf{x}_R^\mathsf{T}\end{bmatrix}^\mathsf{T} \in \mathbb{C}^{RK}$. The superscript $(\cdot)^\mathsf{T}$ denotes matrix transpose, and a hat $(\hat{\cdot})$ denotes an estimate.

II. Tensor-based separation of sums of complex exponentials

This section reviews the method proposed in [7] for separation of SCE sources by means of a block-term tensor decomposition, under the assumption that each source has a small number of poles. We also discuss existence and uniqueness of solutions to the approximate decomposition problem which is addressed in practice.

A. Low-rank Hankel source model

A celebrated result in signal processing states that if a discrete-time signal is a linear combination of L damped complex exponentials, say

$$s(n) = \textstyle\sum_{\ell=1}^{L} \alpha_\ell \exp(\zeta_\ell\, n), \quad n = 0, \ldots, N-1, \tag{2}$$

where $\alpha_\ell, \zeta_\ell \in \mathbb{C}$, then the $M \times M$ Hankel matrix

$$\mathbf{H}_s \triangleq \begin{bmatrix}\mathbf{s}(0) & \mathbf{s}(1) & \ldots & \mathbf{s}(M-1)\end{bmatrix},$$

with $\mathbf{s}(n) \triangleq \begin{bmatrix}s(n) & s(n+1) & \ldots & s(n+M-1)\end{bmatrix}^\mathsf{T}$ and $N = 2M-1$, has rank at most $\min\{L, M\}$. In fact, this property follows immediately from the decomposition [16]

$$\mathbf{H}_s = \mathbf{V}_s \,\mathrm{Diag}(\alpha_1, \ldots, \alpha_L)\mathbf{V}_s^\mathsf{T}, \tag{3}$$

where $\mathbf{V}_s$ is the Vandermonde matrix

$$\mathbf{V}_s \triangleq \begin{bmatrix} 1 & \ldots & 1 \\ \exp(\zeta_1) & \ldots & \exp(\zeta_L) \\ \vdots & & \vdots \\ \exp(\zeta_1(M-1)) & \ldots & \exp(\zeta_L(M-1)) \end{bmatrix} \in \mathbb{C}^{M\times L},$$

and is at the heart of classical modal analysis methods [17]. It implies that a "simple" signal of the form (2) can be mapped into a low-rank Hankel matrix, where simple here means being constituted by a small number L of exponentials. We will see next how signal separation can be performed by relying on this relation.

Remark 1. For a real-valued signal $s(n)$, it is well known that the complex-valued coefficients α_ℓ and poles ζ_ℓ with nonzero imaginary part must arise in complex conjugate pairs. The same requirement thus applies to the columns of $\mathbf{V}_s$ and diagonal elements of $\mathrm{Diag}(\alpha_1, \ldots, \alpha_L)$.

B. Separation of linear mixture via block term decomposition

Consider now a linear instantaneous mixture $y(n) = \sum_{r=1}^{R} x_r\, s_r(n)$, with

$$s_r(n) = \textstyle\sum_{\ell=1}^{L_r} \alpha_\ell^{(r)} \exp\left(\zeta_\ell^{(r)} n\right), \quad L_r < M, \tag{4}$$

and assume one wants to recover the signals $s_r(n)$ from knowledge of y (and of the above model) only. By linearity of the Hankel map discussed above, we have $y \mapsto \mathbf{H}_y = \sum_{r=1}^{R} x_r\, \mathbf{H}_{s_r}$, so that $\operatorname{rank} \mathbf{H}_y \leq \sum_{r=1}^{R} L_r$. Without further information, though, this linear combination of matrices is not of much help for separation.

The situation changes upon introduction of spatial diversity, meaning we now observe $y(k,n) = \sum_{r=1}^{R} x_{k,r}\, s_r(n)$ for $k = 1, \ldots, K$. In matrix notation, we have

$$\mathbf{Y} = \mathbf{X}\mathbf{S}^{\mathsf{T}}, \tag{5}$$

where $\mathbf{S} = (s_{n,r}) = (s_r(n-1))$ is an $N \times R$ matrix containing the source signals and $\mathbf{X} = (x_{k,r})$ is a $K \times R$ mixture matrix specifying how the sources are combined to yield the channels' outputs. Each such output $y_k(n) = y(k,n)$ for a fixed k (*i.e.*, each row of $\mathbf{Y}$) can be mapped into an $M \times M$ Hankel matrix as before, say $y_k \mapsto \mathbf{Y}_k$. Hence, $\mathbf{Y}_k = \sum_{r=1}^{R} x_{k,r} \mathbf{H}_r$, where $\mathbf{H}_r$ is the rank-L_r Hankel matrix associated with s_r. The matrices $\mathbf{Y}_k$ can be viewed as slices of an $M \times M \times K$ tensor $\mathcal{Y}$ satisfying

$$\begin{aligned} \mathcal{Y} &= \textstyle\sum_{k=1}^{K} \mathbf{Y}_k \otimes \mathbf{e}_k = \sum_{k=1}^{K} \left(\sum_{r=1}^{R} x_{k,r} \mathbf{H}_r\right) \otimes \mathbf{e}_k \\ &= \textstyle\sum_{r=1}^{R} \mathbf{H}_r \otimes \left(\sum_{k=1}^{K} x_{k,r} \mathbf{e}_k\right) \\ &= \textstyle\sum_{r=1}^{R} \mathbf{H}_r \otimes \mathbf{x}_r, \end{aligned} \tag{6}$$

where $\mathbf{e}_k$ is the kth canonical basis vector of $\mathbb{C}^K$, $\mathbf{x}_r$ is the rth column of $\mathbf{X}$ and $\otimes$ is the tensor product. The data tensor thus consists of a sum of blocks, each one given by the tensor product of a low-rank matrix and a vector. We refer to the parameters $(R, \{L_r\}_{r=1}^{R})$ as structural parameters or simply the structure of the model in (6).

It turns out that the tensor decomposition (6), known as block-term decomposition (BTD) and introduced by [8], is essentially[1] unique under relatively mild assumptions. Its uniqueness properties have been first studied in [8], and further results were given in [7]. In particular, Theorem 2.4 of [7] states that if $\mathbf{X}$ has full column rank and $\operatorname{rank} \sum_{r=1}^{R} w_r \mathbf{H}_r > \max_r \operatorname{rank} w_r \mathbf{H}_r$ for all $\mathbf{w} = \begin{bmatrix} w_1 & \ldots & w_R \end{bmatrix}^{\mathsf{T}}$ having at least two nonzero components, then the BTD in (6) is essentially unique. (It is thus necessary that $L_r < M$.) Other uniqueness results have been derived in [18], [19], including some that apply to coupled BTDs.

[1]Note that (6) can only be unique modulo a permutation of the summands and a joint rescaling of the components of each summand as in $(\mathbf{H}_r, \mathbf{x}_r) \mapsto (\alpha\, \mathbf{H}_r, (1/\alpha)\, \mathbf{x}_r)$ for some $\alpha \neq 0$.

C. Approximate block term decomposition

In practice, the data matrix $\mathbf{Y}$ is only approximately given by (5), due to noise and imperfect modeling. Hence, one can only approximate tensor $\mathcal{Y}$ by a low-rank BTD model of the form in (6).

Since the approximate BTD problem is important in its own right, we hereby discuss it from a general perspective, momentarily leaving aside the Hankel constraints in (6) and considering a third-order tensor $\mathcal{Y} \in \mathbb{C}^{I \times J \times K}$ (in model (6) we had $I = J = M$). Typically, an approximate BTD is computed by minimizing a measure of distance between the data tensor and a model of fixed structure with respect to the model components. In most cases, a least-squares criterion is adopted (as in, *e.g.*, [10]), leading to

$$\min_{(\mathbf{A},\mathbf{B},\mathbf{X}) \in \mathcal{S}} f(\mathbf{A}, \mathbf{B}, \mathbf{X}), \tag{7}$$

$$\text{with } f(\mathbf{A}, \mathbf{B}, \mathbf{X}) \triangleq \left\| \mathcal{Y} - \textstyle\sum_{r=1}^{R} \left(\mathbf{A}_r \mathbf{B}_r^{\mathsf{T}}\right) \otimes \mathbf{x}_r \right\|_F^2$$

and $\mathcal{S} \triangleq \mathbb{C}^{I \times \sum_{r=1}^{R} L_r} \times \mathbb{C}^{J \times \sum_{r=1}^{R} L_r} \times \mathbb{C}^{K \times R}$, where $\mathbf{A}_r$ contains the columns of indices $1 + \sum_{m=1}^{r-1} L_m$ to $\sum_{m=1}^{r} L_m$ of $\mathbf{A}$, and likewise for $\mathbf{B}_r$. Observe that the factorization $\mathbf{A}_r \mathbf{B}_r^{\mathsf{T}}$ is employed to bound each block rank as $\operatorname{rank} \mathbf{H}_r \leq L_r$.

We discuss next the existence and uniqueness of solutions to (7).

1) Existence: Problem (7) may lack a global minimizer, because the set of tensors having a given BTD structure is not necessarily closed. An example of this phenomenon has been known since the introduction of the BTD [10]. As recently shown in [11], spaces of real-valued tensors can contain sets with nonempty interior whose elements do not admit a best approximate BTD having a given structure. This fact has practical consequences, since it implies that a random tensor drawn from an absolutely continuous distribution has nonzero probability of falling into such a set. For complex-valued tensors, [20] shows that this issue only affects tensors from sets of zero volume, and thus is of lesser practical concern.

2) Uniqueness: In fact, the results of [20] not only imply that a closest tensor having a specified BTD structure exists for almost all complex-valued tensors, but also that it is unique (see [20, Corollary 7.4]). However, there is a subtlety: this simply means that for a random complex-valued tensor $\mathcal{Y}$ the problem

$$\min\nolimits_{\hat{\mathcal{Y}} \in \mathcal{B}_{L_1,\ldots,L_R}} \|\mathcal{Y} - \hat{\mathcal{Y}}\|_F^2$$

has a unique solution almost surely, where $\mathcal{B}_{L_1,\ldots,L_R}$ is the set of all complex-valued tensors which can be written in the form $\sum_{r=1}^{R} \mathbf{H}_r \otimes \mathbf{x}_r$ with $\operatorname{rank} \mathbf{H}_r \leq L_r$. This does *not* imply that the *BTD components* themselves are unique, which requires additional conditions over these components, such as those stated in Section II-B.

For real-valued tensors, though, no analogue of the above mentioned result is known.

D. Shortcomings of the standard least-squares approach

The definition of the BTD does not impose restrictions regarding the ranks of the blocks, which can be very different from one another. This flexibility is a very attractive feature of

the BTD, since in applications it is often the case that the objects being represented by the blocks have different properties and thus require different ranks for their accurate modeling. This occurs, for instance, in the application considered in this paper, where accurate approximation of a ventricular signal typically needs more poles, and therefore higher rank, than an atrial signal.

However, leveraging such a flexibility in practice is a challenging task that currently existing approaches are not able to carry out effectively. First, because it requires adequate estimates for the block ranks. Secondly, even assuming such ranks are known or available, the standard approach of solving (7) with fixed block ranks often performs poorly, as the algorithm can converge to a local minimum where the block ranks have been "inverted." For instance, in a source separation scenario, this approach has no way of controlling which rank will be assigned to each type of source. While the use of an algebraic initialization scheme (see, *e.g.*, [8, Theorem 4.1]) can reduce this effect, it cannot prevent it completely, as our simulations in Section IV-A1 will show.

This issue is often circumvented in practice by simply assigning equal ranks to all blocks in the decomposition. Yet, we argue that this is not a sensible approach, for two reasons:

(i) Setting all ranks equal implies overestimation of the ranks of one or more blocks, which increases estimation variance and slows down convergence.

(ii) If one wants to employ an algebraic initialization, then a necessary condition is that $\sum_{r=1}^{R} L_r \leq \min\{I, J\}$, and thus if $L_r = L$ for all r then one must choose $L \leq \frac{\min\{I,J\}}{R}$. In other words, the range of possible ranks is restricted, and the restriction becomes more severe as the number of blocks grows.

In addition to the above arguments, the ill-posedness of (7) for some tensors $\mathcal{Y}$ (a phenomenon that can happen with positive probability in the real-valued case [11]) also has visible impact on the performance of algorithms. This will also be made clear by the results of Section IV-A1, where the estimates produced by a BTD algorithm that tackles (7) often diverge.

E. Structured low-rank approximation

In the special case of interest (6), $\mathbf{H}_r$ must belong to the subspace of $M \times M$ Hankel matrices, $\mathcal{H}$. Although the slices $\mathbf{Y}_k$ are Hankel by construction, in practice a solution $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{X}})$ of (13) or (7) may not satisfy $\hat{\mathbf{A}}_r \hat{\mathbf{B}}_r^{\mathsf{T}} \in \mathcal{H}$, due to noise, to a possibly inadequate choice of block ranks and to modeling imperfections. In fact, even if the sum $\sum_{r=1}^{R} \hat{x}_{k,r} \hat{\mathbf{A}}_r \hat{\mathbf{B}}_r^{\mathsf{T}}$ is Hankel, the matrices $\hat{\mathbf{A}}_r \hat{\mathbf{B}}_r^{\mathsf{T}}$ do not need to be (though the opposite is certainly true). In other words, the solution may lack temporal structure, not being interpretable as a mixture of sources of the form (2).

There are basically two ways of dealing with this issue:

1) *Projected unconstrained solution*: Ignore the Hankel structure to compute an unconstrained approximate low-rank BTD and then project each estimated block $\hat{\mathbf{H}}_r$ onto the set $\mathcal{H}_{L_r}$ of Hankel matrices having rank up to L_r.

2) *Constrained solution*: Use a constrained optimization algorithm for imposing the Hankel structure while computing the solution of (6).

More generally, a similar choice is faced in related problems such as harmonic retrieval [21] or the so-called shape-from-moments problem [22], which can be cast as the computation of a low-rank approximation of a Hankel data matrix $\mathbf{Y}$. In practice, one either ignores the Hankel structure or solves a structured low-rank approximation problem of the form

$$\min_{\mathbf{H} \in \mathcal{H}_L} \|\mathbf{Y} - \mathbf{H}\|_F, \tag{8}$$

which can be addressed by structured low-rank approximation methods such as those thoroughly discussed in [23].

In principle, one would expect that more accurate results are attained when the Hankel structure is imposed, for this introduces an additional prior in the optimization process. However, this is not necessarily true: for instance, the experiments in [22] concerning the shape-from-moments problem show that, even though a constrained algorithm can outperform an unconstrained one in a low-noise regime, the same does not happen at a higher level of noise. This probably happens because the local optimization algorithms used to solve the nonconvex problem (8) may attain a suboptimal local minimum. Furthermore, these constrained optimization algorithms are typically more computationally demanding than unconstrained ones.[2]

Due to these difficulties, in practice unconstrained algorithms are often employed, as seen in [21], [22]. In the particular case of our BTD problem, to date only the projected unconstrained approach has been followed, as in [1], [4]. However, in spite of the above mentioned difficulties, the use of a constrained approach can lead to more accurate solutions, as our computer experiments will show in Section V.

Finally, note that the above discussion also applies to arbitrary constraints of the form $\mathbf{H}_r \in \mathcal{U}$ where $\mathcal{U}$ is any[3] subspace of $\mathbb{C}$.

III. Alternating group lasso algorithm for BTD

In the following, aiming to overcome the difficulties discussed in Section II-D, we derive a provably convergent algorithm for computing an (unconstrained) approximate BTD of a given tensor that does not impose hard rank constraints, allowing instead the joint estimation of block ranks and parameters. Subsequently, we show how linear (subspace) constraints can be imposed upon the block matrices, Hankel structure being a special case of these constraints.

A. Problem formulation

Instead of determining the BTD structure beforehand, we propose to include penalization terms promoting low-rank

[2]This higher cost can often be alleviated by initializing a constrained optimization algorithm with an unconstrained solution.

[3]In some special cases, such as when $\mathcal{U}$ is the space of circulant matrices, the structured low-rank matrix problem has a closed-form solution.

blocks and controlling the number of blocks in the formulation, as in

$$\min_{(\mathbf{A},\mathbf{B},\mathbf{X})\in\mathcal{S}} F(\mathbf{A},\mathbf{B},\mathbf{X}) \quad \text{with } F(\mathbf{A},\mathbf{B},\mathbf{X}) \triangleq f(\mathbf{A},\mathbf{B},\mathbf{X}) + \gamma\, g(\mathbf{A},\mathbf{B},\mathbf{X}), \tag{9}$$

where $\mathcal{S} \triangleq \mathbb{C}^{I\times LR}\times\mathbb{C}^{J\times LR}\times\mathbb{C}^{K\times R}$, f is the same as in (7), $\gamma > 0$ is a regularization parameter and g is a regularization function of the form

$$g(\mathbf{A},\mathbf{B},\mathbf{X}) \triangleq \|\mathbf{A}\|_{2,1} + \|\mathbf{B}\|_{2,1} + \|\mathbf{X}\|_{2,1}, \tag{10}$$

where the norm $\|\cdot\|_{2,1}$ is defined by (1).

Adding a mixed $\ell_{2,1}$-norm regularization term is a well-known strategy for inducing *group sparsity* of its argument's columns, in the sense that the number of nonzero columns of the penalized matrix will be adjusted to the data according to the weight of the regularization term and to their contribution to the data fit term. This is essentially a generalization of the lasso (least absolute shrinkage and selector operator) estimator principle, called group lasso, and is owed to geometric properties of the $\ell_{2,1}$-norm [24].

Hence, for sufficiently high γ, minimizers of (9) will be formed by $\mathbf{A}$ and $\mathbf{B}$ displaying some columns made entirely of zeros, effectively yielding a BTD of low-rank blocks. The same applies to $\mathbf{X}$, possibly reducing the number of blocks. This allows much more flexibility compared to (7), since now the number of degrees of freedom of the model can adapt to the data $\mathcal{Y}$. Moreover, at least one solution is *guaranteed* to exist: because F is nonnegative, coercive[4] (due to g) and continuous, existence follows from the extreme value theorem.

Note that, in favor of simplicity, we choose to include a regularization term in which all three factors are penalized in the same way, even though $\mathbf{X}$ plays a different role than $\mathbf{A}$ and $\mathbf{B}$ in the model. Alternatively, one could use one constant (say, γ_1) to penalize $\mathbf{A}$ and $\mathbf{B}$ and another (γ_2) to penalize $\mathbf{X}$, provided that the scale of each $\mathbf{A}_r\mathbf{B}_r{}^{\mathsf{T}}$ or of x_r is fixed to eliminate the scaling ambiguity, otherwise these different relative weights can always be compensated for by rescaling the factors. Yet, in practice this would require tuning two regularization constants instead of one, while our results show that the proposed approach is already quite effective for estimating BTD models in a range of situations.

B. Algorithm for unconstrained blocks

To tackle the nonconvex and nonsmooth problem (9), we employ a block coordinate descent (BCD) approach. This widespread technique consists in partitioning the set of optimization variables and then sequentially solving subproblems in each subset of variables (with the others fixed) until all subsets are updated, completing one iteration of the algorithm.

Here, we have a natural partition into three blocks: $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{X}$. Let $\hat{\mathbf{A}}^{(t-1)}$, $\hat{\mathbf{B}}^{(t-1)}$ and $\hat{\mathbf{X}}^{(t-1)}$ denote the estimates obtained at iteration $t-1$. Fixing $\mathbf{B} = \hat{\mathbf{B}}^{(t-1)}$ and $\mathbf{X} = \hat{\mathbf{X}}^{(t-1)}$ in (9), the subproblem in $\mathbf{A}$ of iteration t becomes a standard group lasso problem

$$\min_{\mathbf{A}\in\mathbb{C}^{I\times LR}} \frac{1}{2}\|\mathcal{Y} - \mathscr{W}_{\mathbf{A}}^{(t)}(\mathbf{A})\|_F^2 + \gamma\,\|\mathbf{A}\|_{2,1}, \tag{11}$$

where $\mathscr{W}_{\mathbf{A}}^{(t)}$ is a linear map depending on $\hat{\mathbf{B}}^{(t-1)}$ and $\hat{\mathbf{X}}^{(t-1)}$. The groups are disjoint and correspond to the columns of $\mathbf{A}$. Similar subproblems can be derived for $\mathbf{B}$ and $\mathbf{X}$.

Now, despite being convex, subproblem (11) may not be strictly convex, because $\mathscr{W}_{\mathbf{A}}^{(t)}$ may not be injective at some iterations. It can thus fail to have a unique solution, in which case the convergence of the BCD algorithm to stationary points of (9) cannot be guaranteed (see [25] and references therein). One can remedy this shortcoming by adding a proximal term, as in

$$\min_{\mathbf{A}\in\mathbb{C}^{I\times LR}} \frac{1}{2}\|\mathcal{Y} - \mathscr{W}_{\mathbf{A}}^{(t)}(\mathbf{A})\|_F^2 + \gamma\,\|\mathbf{A}\|_{2,1} + \tfrac{\tau}{2}\|\mathbf{A} - \hat{\mathbf{A}}^{(t-1)}\|_F^2,$$

with $\tau > 0$. Setting $\mathbf{a} \triangleq \operatorname{vec}(\mathbf{A})$ and $\mathbf{y}_1 \triangleq \operatorname{vec}(\mathbf{Y}_{\langle 1\rangle})$, where $\mathbf{Y}_{\langle 1\rangle}$ indicates the mode-1 matrix unfolding of $\mathcal{Y}$ (see *e.g.* [7]), this is equivalent to

$$\min_{\mathbf{a}\in\mathbb{C}^{ILR}} \frac{1}{2}\left\|\begin{bmatrix}\mathbf{y}_1\\ \sqrt{\tau}\,\hat{\mathbf{a}}^{(t-1)}\end{bmatrix} - \begin{bmatrix}\mathbf{W}_{\mathbf{A}}^{(t)}\\ \sqrt{\tau}\,\mathbf{I}\end{bmatrix}\mathbf{a}\right\|_2^2 + \gamma\sum_{r=1}^{R}\sum_{l=1}^{L}\|\mathbf{a}_{r,l}\|_2, \tag{12}$$

where $\mathbf{W}_{\mathbf{A}}^{(t)} \triangleq \left(\hat{\mathbf{X}}^{(t-1)} \odot_L \hat{\mathbf{B}}^{(t-1)}\right) \boxtimes \mathbf{I}_I$ and $\mathbf{a}_{r,l}$ holds components $(((r-1)L+l)-1)I+1$ to $((r-1)L+l)I$ of $\mathbf{a}$. By construction, (12) is strictly convex, since the Hessian $\mathbf{W}_{\mathbf{A}}^{(t)^{\mathsf{T}}}\mathbf{W}_{\mathbf{A}}^{(t)} + \tau\mathbf{I}$ of the least-squares term is positive definite for $\tau > 0$. Analogous strictly convex subproblems can also be derived for $\mathbf{B}$ and $\mathbf{X}$, with

$$\mathbf{W}_{\mathbf{B}}^{(t)} \triangleq \left(\hat{\mathbf{X}}^{(t-1)} \odot_L \hat{\mathbf{A}}^{(t)}\right) \boxtimes \mathbf{I}_J,$$
$$\mathbf{W}_{\mathbf{X}}^{(t)} \triangleq \left[\left(\hat{\mathbf{B}}^{(t)} \odot \hat{\mathbf{A}}^{(t)}\right)(\mathbf{I}_R \boxtimes \mathbf{1}_L)\right] \boxtimes \mathbf{I}_K.$$

Though no analytical solution exists for these subproblems, they can be solved by employing existing (iterative) group lasso algorithms such as that of [26].[5]

The proposed AGL algorithm solves them in alternating fashion, cycling through updates of $\hat{\mathbf{A}}^{(t)}$, $\hat{\mathbf{B}}^{(t)}$ and $\hat{\mathbf{X}}^{(t)}$, in that order, at each iteration t. It can be seen as a regularized version of the alternating least squares scheme proposed in [10]. As we explain in Appendix A, by [25, Theorem 2] all limit points of the sequence of iterates produced by AGL are stationary points of problem (9).

Table I contains a pseudocode for the constrained version of AGL, which we present next. The unconstrained version follows the same scheme, but does not involve lines 5 to 8 (which impose the constraints).

It should be noted that AGL is a general approach to compute an unconstrained approximate low-rank BTD and is not specifically tailored to the application we consider later in the paper. Hence, it could also be applied, for example, to the problems presented in [1], [27] and [28].

[4]Recall that a function $F(\mathbf{A},\mathbf{B},\mathbf{X})$ is said to be coercive if and only if $F(x)\to\infty$ whenever $\|\mathbf{A}\|_F\to\infty$ *or* $\|\mathbf{B}\|_F\to\infty$ *or* $\|\mathbf{X}\|_F\to\infty$.

[5]This algorithm was, however, formulated for the real-valued setting only.

TABLE I: Pseudocode for constrained AGL algorithm. For the unconstrained AGL algorithm, lines 5–8 must be ommited.

Inputs:	Data tensor $\mathcal{Y}$, penalty parameter γ, proximal term weight τ, initial point $(\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{X}^{(0)})$
Outputs:	Approximate BTD factors $(\mathbf{A}, \mathbf{B}, \mathbf{X})$

1: $t \leftarrow 1$
2: **while** stopping criteria not met **do**
3: Solve group lasso subproblem (12) to obtain $\mathbf{A}^{(t)}$ from $\mathbf{A}^{(t-1)}$, $\mathbf{B}^{(t-1)}$ and $\mathbf{X}^{(t-1)}$
4: Solve group lasso subproblem in $\mathbf{B}$ analogous to (12) to obtain $\mathbf{B}^{(t)}$ from $\mathbf{A}^{(t)}$, $\mathbf{B}^{(t-1)}$ and $\mathbf{X}^{(t-1)}$
5: **for** $r = 1, \ldots, R$ **do**
6: $L_r^{(t)} \leftarrow \operatorname{rank}\left(\mathbf{A}_r^{(t)}(\mathbf{B}^{(t)})_r^{\mathsf{T}}\right)$
7: $(\mathbf{A}_r^{(t)}, \mathbf{B}_r^{(t)}) \leftarrow \texttt{slra}(\mathbf{A}_r^{(t)}(\mathbf{B}_r^{(t)})^{\mathsf{T}}, L_r^{(t)})$
8: $(\mathbf{A}_r^{(t)}, \mathbf{B}_r^{(t)}) \leftarrow ([\mathbf{A}_r^{(t)} \ \mathbf{0}_{I \times L - L_r^{(t)}}], [\mathbf{B}_r^{(t)} \ \mathbf{0}_{I \times L - L_r^{(t)}}])$
9: Solve group lasso subproblem in $\mathbf{X}$ analogous to (12) to obtain $\mathbf{X}^{(t)}$ from $\mathbf{A}^{(t)}$, $\mathbf{B}^{(t)}$ and $\mathbf{X}^{(t-1)}$
10: $t \leftarrow t + 1$

C. Handling linear constraints in $\mathbf{H}_r$

In order to address the issue discussed in Section II-E, we propose a variant of AGL, termed constrained AGL (CAGL), to address the problem

$$\min_{(\mathbf{A},\mathbf{B},\mathbf{X}) \in \mathcal{S}} F(\mathbf{A}, \mathbf{B}, \mathbf{X}) \quad \text{subj. to} \quad \forall r,\ \mathbf{A}_r \mathbf{B}_r^{\mathsf{T}} \in \mathcal{U}, \quad (13)$$

with $\mathcal{U}$ denoting a given subspace of $I \times J$ matrices. In CAGL, the constraint in (13) is enforced by applying a structured low-rank approximation (SLRA) algorithm during the iterations, after having estimated $\hat{\mathbf{A}}^{(t)}$ and $\hat{\mathbf{B}}^{(t)}$. For clarity, the algorithm is summarized in Table I, where notation was simplified by dropping the symbol $\hat{(\cdot)}$ from the computed estimates and `slra` denotes the SLRA algorithm, which takes a matrix $\hat{\mathbf{H}}_r$ and a maximum rank L_r as inputs and produces factors $\hat{\mathbf{A}}_r$ and $\hat{\mathbf{B}}_r$ having L_r columns and satisfying $\hat{\mathbf{H}}_r \approx \hat{\mathbf{A}}_r \hat{\mathbf{B}}_r^{\mathsf{T}} \in \mathcal{U}$.

Note that each application of `slra` re-estimates both $\mathbf{A}_r$ and $\mathbf{B}_r$. Also, in practice the computation of $L_r^{(t)}$ at line 6 can be performed by counting how many of the columns of $\mathbf{A}_r$ and $\mathbf{B}_r$ are simultaneously nonzero. Finally, it should be noted that, since $\hat{\mathbf{A}}_r^{(t)}$ and $\hat{\mathbf{B}}_r^{(t)}$ always have L columns, zeros must be added in those blocks after applying `slra`, as done in line 8 of Table I.

As we will see in Section V, in practice CAGL is able to yield meaningful solutions to the source separation problem. However, constraining the block matrices brings a heavier computational load, due to use of the structured low-rank approximation algorithm, which is typically iterative. Furthermore, the arguments which allow showing the convergence of AGL no longer apply here.

Remark 2. In the particular case of a Hankel constraint, another way of imposing it is by estimating for each block matrix a low-rank decomposition of the form (3), where the Vandermonde factors are parameterized by their poles ζ_ℓ. In this case, a sparsity prior over the coefficients α_ℓ can be used to promote low-rank blocks. However, since ζ_ℓ and α_ℓ are generally complex-valued, conjugacy constraints would be needed to obtain real-valued block matrices, which requires fixing the (maximum) number of complex-conjugate pole pairs and of single real-valued poles of each block *a priori*. This loss of generality can be mitigated by using a high value of L for the starting rank (thus including "enough" conjugate pole pairs and single poles), but at the expense of increased computing cost.

IV. Numerical evaluation on random block term decomposition models

We now evaluate AGL and CAGL in the approximate computation of synthetic BTD models, with and without Hankel constraint.

A. Unconstrained BTD

1) Scenario 1: We generate 500 joint realizations of $(\mathbf{A}, \mathbf{B}, \mathbf{X}, \mathcal{N})$ by drawing the (real-valued) elements of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{X}$ and $\mathcal{N}$ in an independent and identically distributed (i.i.d.) fashion from the standard normal distribution. $\mathbf{X}$ is then normalized column-wise and the condition $\max_{i,j} |\mathbf{x}_i^{\mathsf{T}} \mathbf{x}_j| < 0.9$ is imposed (by drawing $\mathbf{X}$ multiple times until it is met). This prevents nearly collinear spatial signatures. Next, we construct the noisy model $\mathcal{Y} = \mathcal{Y}_0 + \sigma_{\mathcal{N}} \mathcal{N}$, where $\mathcal{Y}_0 = \sum_{r=1}^{R} (\mathbf{A}_r \mathbf{B}_r^{\mathsf{T}}) \otimes \mathbf{x}_r$ and $\sigma_{\mathcal{N}}$ is the standard deviation of the noise, which is adjusted to obtain a signal-to-noise ratio (SNR) of 20 dB, defined as $\text{SNR} \triangleq \|\mathcal{Y}_0\|_F^2 \, \sigma_{\mathcal{N}}^{-2} \|\mathcal{N}\|_F^{-2}$.

We set $I = J = 18$, $K = 4$ and $R = 3$. The block ranks are $(L_1, L_2, L_3) = (6, 5, 4)$. Since $L_1 + L_2 + L_3 = 15 < 18 = \min\{I, J\}$, one can compute an approximate initial solution by using an algebraic method. In this simulation, we employ the algebraic method implemented by Tensorlab[6], which is based on [8, Theorem 4.1]. (Note that the procedure based on simultaneous diagonalization proposed in [19] does not apply, because we consider blocks with different ranks.) Given an input tensor and the chosen block ranks $\hat{L}_r$, this implementation first essentially computes a canonical polyadic decomposition with rank $\sum_{r=1}^{R} \hat{L}_r$ and then attempts to cluster the columns of the third factor (which corresponds to $\mathbf{X}$) by using a k-means algorithm, in order to form blocks having the specified ranks.

For each realization, AGL is applied with $(R, L) = (3, 6)$, and also with $(R, L) = (5, 8)$, in both cases using the following procedure:

- an initial solution is generated by employing the above described algebraic method with rank estimates $\hat{L}_1 = \hat{L}_2 = \hat{L}_3 = L$;
- AGL is run from this initial solution and using $\gamma = \gamma_0$;
- for $p = 1, \ldots, P-1$, AGL is run with $\gamma = \gamma_p = (p+1)\gamma_0$ using the solution obtained for γ_{p-1} as the initial point.

This γ-sweeping procedure is inspired by solution-path techniques used in the statistics community [29]. We then keep the last generated solution as the final estimate. The value of γ_0 was tuned empirically, by setting it to a small constant (usually from 10^{-4} to 10^{-1}) times $\|\mathcal{Y}\|_F$. We set the proximal term

[6]This algebraic method is implemented in Tensorlab by the function `lll_gevd`.

weight as[7] $\tau = 10^{-3}$. For comparison, Tensorlab's Gauss-Newton BTD algorithm[8] (BTD-GN) [9] is used with the same rank $L = 6$ for all three blocks, being also initialized by the same algebraic solution used for AGL.

The performance criterion used in our evaluation is the normalized mean squared error (NMSE) over the blocks:

$$\text{NMSE}(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{X}}) \triangleq \frac{1}{R}\sum_{r=1}^{R} \frac{\|(\mathbf{A}_r\mathbf{B}_r^\mathsf{T}) \otimes \mathbf{x}_r - (\hat{\mathbf{A}}_r\hat{\mathbf{B}}_r^\mathsf{T}) \otimes \hat{\mathbf{x}}_r\|_F^2}{\|(\mathbf{A}_r\mathbf{B}_r^\mathsf{T}) \otimes \mathbf{x}_r\|_F^2}.$$

To compute this quantity, one needs to match the estimated blocks to the ground truth, which can be done using the classical Hungarian algorithm for assignment problems [30]. (This means that, when AGL estimates more than 3 blocks, only those that minimize the NMSE are kept.) The empirical cumulative distribution function (ECDF) of the measured NMSE for each algorithm is shown in Fig. 1. It shows a clear superiority of the proposed approach: for instance, the NMSE attained by AGL is smaller than 0.01 for about 83% realizations with $(R, L) = (3, 6)$ and about 80% realizations with $(R, L) = (5, 8)$ while the same is true for only about 58.8% realizations in the case of BTD-GN. This is due to the following reasons:

- For many realizations, BTD-GN seems to run into problems related to the non-existence of a best approximation, as the estimated blocks grow unbounded in norm along the iterations, while nearly canceling each other out to yield a reasonable approximation of the input data tensor (but not of the desired low-rank blocks). For instance, for 81 realizations (16.2%), the final NMSE over the blocks exceeds 100. This may be caused by the lack of a global solution for the noisy input tensor (as discussed in Section II-C1), or even by the algorithm entering a region of the parameter space associated to tensors which are topologically close to the set of tensors that do not have a best approximation (well-posed instances in the vicinity of ill-posed ones are typically ill-conditioned). In other words, ill-posedness of tensor decomposition problems can cause difficulties even when a global solution exists for the tensor being decomposed, a fact that has already been observed in the literature [31].
- AGL is able in this experiment to detect the true model structure with a 97.6% success rate when $(R, L) = (3, 6)$ and a 87.6% success rate when $(R, L) = (5, 8)$, and thus quite often effectively finds the number of relevant variables, whereas BTD-GN is estimating an over-parameterized model, and hence its results are subject to a higher variance.
- Because of the added noise and of the mismatch between the true ranks and the ranks specified to the algorithms, the algebraic initialization procedure described above fails in most realizations (84.6%), in which case a random initialization with the chosen ranks is generated instead by Tensorlab's routine. As it turns out, AGL is much less sensitive with respect to the choice of the initial solution, as will be discussed in the next simulation scenario.

[7]This choice is not very consequential, provided that τ is (numerically) bounded away from zero, and not too high to avoid slowing down convergence.

[8]The Tensorlab function used in our experiments is `lll_nls`.

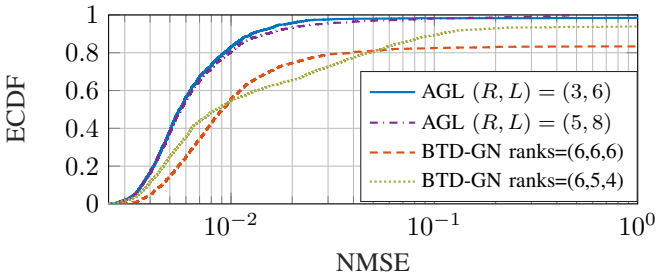

Fig. 1: Empirical CDFs of NMSE over estimated blocks attained by AGL and BTD-GN for 500 realizations of a noisy random BTD model (SNR = 20 dB). The block ranks used in BTD-GN are indicated in the legend, as well as the values of (R, L) used by AGL. The true ranks are $(L_1, L_2, L_3) = (6, 5, 4)$.

For the sake of comparison, we have also run BTD-GN with the true block ranks (which is unlikely to be viable in practice, since the block ranks are typically unknown). In that case, the algebraic initialization procedure only fails at about 25.4% realizations. Even so, BTD-GN is still outperformed by AGL for both initial guesses $(R, L) = (3, 6)$ and $(R, L) = (5, 8)$, as shown in Fig. 1. This is because there is a high probability of the rank inversion phenomenon discussed in Section II-D: in this experiment, this happened for about 51% of the realizations, considerably degrading the block approximation errors.

2) Scenario 2: We now consider a second scenario, where noisy BTD are generated following the same procedure but now with $I = J = 16$ and $K = R = 3$. The block ranks are $L_1 = 8$, $L_2 = 6$ and $L_3 = 4$. Since here $L_1 + L_2 + L_3 = 18 > 16 = \min\{I, J\}$, one cannot employ an algebraic solution to initialize the algorithm. Hence, in the absence of a better initialization strategy, we are forced to employ an arbitrarily chosen (in practice, a random) initial solution. Under these circumstances, the approximate BTD problem becomes particularly challenging because it is non-convex. Moreover, as we shall see in Section V-C, the case where $\sum_{r=1}^{R} L_r > \min\{I, J\}$ is relevant in the application of BTD to source separation.

In this scenario, AGL is applied with $L = 8$ and follows the same procedure as described above, but now a random initial solution is used to start the γ-sweeping procedure that produces a sequence of P solutions. Among these solutions, we keep the best one according to the normalized mean squared error (NMSE) over the blocks. For comparison, the BTD-GN algorithm is run with the same rank L for all three blocks and also with the true ranks L_r.

To study sensitivity with respect to initialization, we run each algorithm starting at several random initial solutions: 12 initial points for AGL and 30 initial points for BTD-GN. (The first 12 initial points used for both algorithms are the same, and when the true ranks are used in BTD-GN, we keep only the first L_r columns of matrices $\hat{\mathbf{A}}_r^{(0)}$ and $\hat{\mathbf{B}}_r^{(0)}$ used in AGL.)

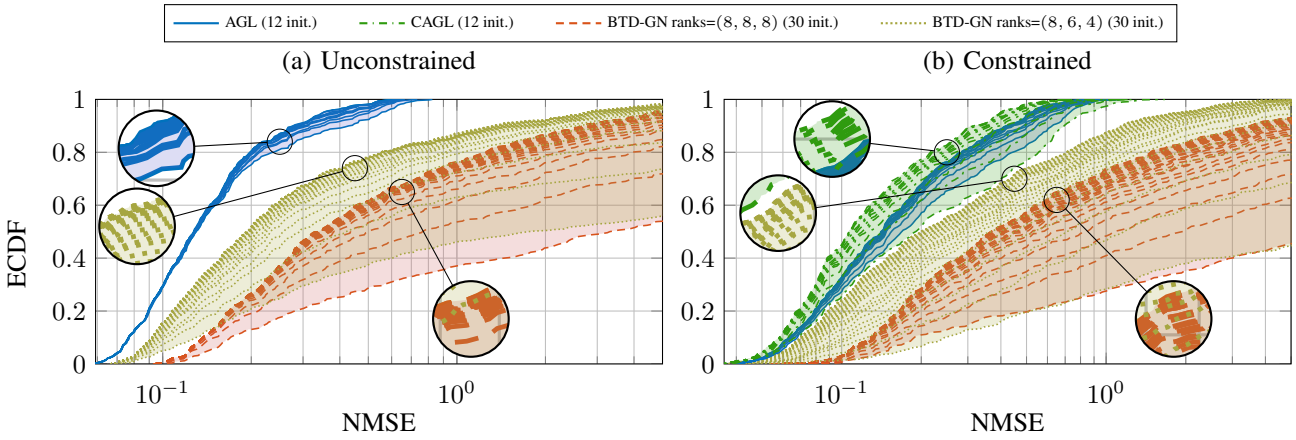

Fig. 2: Empirical CDFs of NMSE over estimated blocks attained by AGL and BTD-GN for 500 realizations of a noisy random BTD model (SNR = 10 dB), in the unconstrained and (Hankel-)constrained scenarios. The block ranks used in BTD-GN are indicated in the legend, while AGL and CAGL start with all ranks set to $L = 8$. The true ranks are $(L_1, L_2, L_3) = (8, 6, 4)$. BTD-GN and (C)AGL are run with 30 and 12 random initializations, respectively. The n_ith curve (from bottom to top) of each algorithm is the ECDF obtained by keeping the best solution among those given by the first n_i initializations. In the constrained scenario, all algorithms are followed by Cadzow's algorithm to enforce the low-rank Hankel constraint.

Fig. 2(a) shows the results for SNR = 10 dB. The n_ith curve (from bottom to top) of each algorithm is the ECDF of the NMSE obtained by keeping the best solution among those given by the first n_i initializations. Clearly, AGL outperforms BTD-GN by a significant margin, even when the latter is given the true ranks (which is an unrealistic condition in practice). Indeed, a single run of AGL suffices to produce statistically better results than 30 runs of BTD-GN. The magnified portions of the plot show that the curves start to "accumulate" when n_i attains a certain value, suggesting that little or no improvement can be achieved by further increasing it. AGL is also visibly much more robust with respect to its initialization, since its curves are much less spread apart. The average computing time and standard deviation for a single run of each algorithm are: $\mu = 4.91$ s and $\sigma = 1.06$ s for AGL (taking into account the execution of AGL for the whole range of values of γ_p, as described above); $\mu = 4.44$ s and $\sigma = 3.89$ s for BTD-GN with ranks $(8, 8, 8)$ and $\mu = 4.88$ s and $\sigma = 4.04$ s for BTD-GN with ranks $(8, 6, 4)$.

Fig. 3(a)–(c) shows the proportion of block ranks estimated by AGL. All block ranks are well estimated most of the time, but there is a significant chance of underestimation of L_1. This behavior probably comes from the choice of the γ parameter range, whose values may be slightly higher than necessary.

For a more comprehensive comparison, we repeat the experiment for other SNR values and then compute the median NMSE over the blocks attained by each algorithm. The results are shown in Fig. 4, where now the n_ith curve from top to bottom of each algorithm is computed by keeping the best result among the first n_i initializations. One can see that AGL is preferable over this entire SNR range, as the median NMSE of BTD-GN approaches that of AGL only for higher SNR values and when the algorithm is run several times (with the known ranks), thus incurring a high computing cost. Finally, the fact that the 12 curves of AGL are indistinguishable

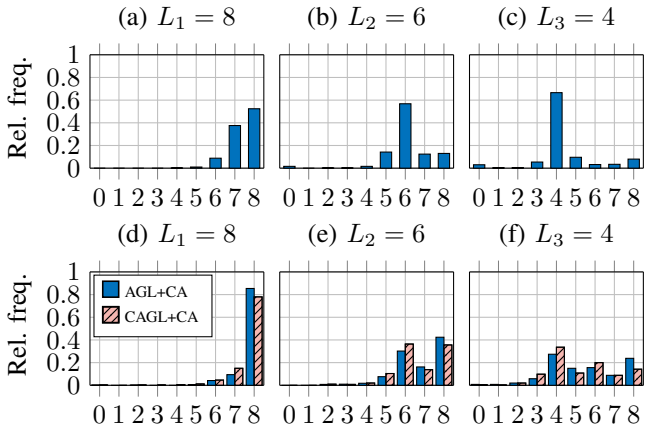

Fig. 3: Proportion of block ranks estimated by AGL and CAGL with SNR=10dB: (a)–(c) refer to the unconstrained scenario of Section IV-A2 and (d)–(f) refer to the constrained scenario of Section IV-B. In the constrained scenario, the rank estimates refer to the best obtained result (with respect to NMSE over blocks) among the first $n_i = 5$ runs.

showcases its robustness regarding the choice of initialization.

The results obtained with BTD-GN highlights the difficulties that it faces in practice: frequent stop due to stagnation in a region of very slow convergence or failure to converge within the maximum allowed number of iterations (which is 1500, the same used for AGL). The employed multi-initialization scheme mitigates these problems, but only to some extent. As in the previous scenario, these difficulties are also possibly related to the ill-posedness of problem (7).

B. Constrained BTD

A similar evaluation was performed in the case where block matrices have Hankel structure. Apart from CAGL

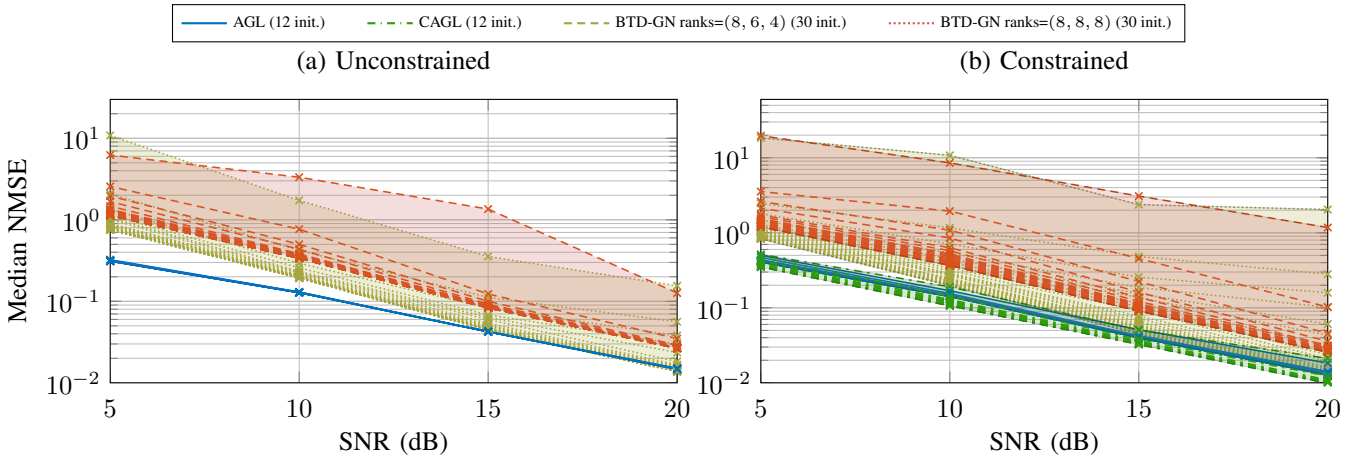

Fig. 4: Median NMSE over estimated blocks attained by AGL and BTD-GN for 500 realizations of a noisy random BTD model, in the unconstrained and (Hankel-)constrained scenarios. The true ranks are $(L_1, L_2, L_3) = (8, 6, 4)$. BTD-GN and (C)AGL are run with 30 and 12 random initializations, respectively. The n_ith curve (from top to bottom) of each algorithm is obtained by keeping the best solution among those given by the first n_i initializations. In the constrained scenario, all algorithms are followed by Cadzow's algorithm to enforce the low-rank Hankel constraint.

TABLE II: Pseudocode of Cadzow's algorithm [32] for a Hankel constraint.

Inputs: Data matrix $\mathbf{Y}$, target rank L, tolerance ϵ_{CA} and maximum number of iterations T_{CA}
Outputs: Low-rank Hankel approximation $\mathbf{H}$ of $\mathbf{Y}$

Initialization: $t \leftarrow 1$, $\mathbf{H}^{(0)} \leftarrow \mathbf{Y}$
for $t = 1, 2, \ldots$ **do**
  $\mathbf{H}^{(t)} \leftarrow \mathcal{P}_{\mathcal{H}}(\mathbf{H}^{(t-1)})$
  Compute the SVD: $\mathbf{H}^{(t)} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}$
  Truncate the computed SVD at rank L: $\mathbf{H}^{(t)} \leftarrow \mathbf{U}_L\mathbf{\Sigma}_L\mathbf{V}_L^{\mathsf{T}}$
  if $\|\mathbf{H}^{(t)} - \mathbf{H}^{(t-1)}\|_F < \epsilon_{\text{CA}} \|\mathbf{H}^{(t-1)}\|_F$ or $t = T_{\text{CA}}$ **then**
    break for loop and output $\mathbf{H} = \mathbf{H}^{(t)}$

and BTD-GN, this evaluation also includes AGL in order to compare the performances of the constrained and unconstrained approaches. For the sake of simplicity, the structured low-rank approximation method used in CAGL is Cadzow's algorithm (CA) [32], which consists in performing alternating projections onto the Hankel subspace $\mathcal{H}$ and onto the set of matrices with rank bounded by a prescribed value. Although the approximations provided by CA are not (locally) optimal in general [33], they are often satisfying in practice. Furthermore, CA is very simple to implement.

For clarity, a description of the employed CA is given in Table II. The projection onto $\mathcal{H}$, denoted by $\mathcal{P}_{\mathcal{H}}$, is cheap to compute by averaging anti-diagonals of $\mathbf{H}^{(t)}$. By the Eckart–Young theorem, the projection of $\mathbf{H}^{(t)}$ onto $\mathcal{L}$ can be obtained via its truncated singular value decomposition (SVD), which is easy but costly to compute. Hence, to avoid an excessive cost per iteration of CAGL, we set the stopping criteria (see Table II) to $\epsilon_{\text{CA}} = 10^{-3}$ and $T_{\text{CA}} = 10$, implying that the constraint is approximately satisfied along the iterations.

In this scenario, we also generate 500 realizations of noisy constrained BTD models for each value of SNR, using the same dimensions, number of blocks and block ranks as in Section IV-A. To this end, the block matrices are random low-rank Hankel matrices generated using CA with $\epsilon_{\text{CA}} = 2.22 \times 10^{-16}$ (machine precision) and $T_{\text{CA}} = 1000$. The algorithms are run as in the previous section, but now we apply CA (with $\epsilon_{\text{CA}} = 10^{-10}$ and $T_{\text{CA}} = 1000$) to their outputs for enforcing the Hankel constraint.

The results obtained for SNR = 10 dB are depicted in Fig. 2(b) and show that, statistically, both AGL and CAGL produce more accurate results than BTD-GN. Though the performance of multi-initialization BTD-GN is now close to that of single-initialization CAGL, this only applies when the former is given the true block ranks. The curves also show that a single run of CAGL performs worse than AGL, but the opposite is true when more runs are performed. Furthermore, CAGL is more sensitive with respect to its initialization, which is expected since the constrained problem is harder to solve. The block rank estimation is less accurate in this scenario than in the unconstrained case, as shown in Fig. 3(d)–(f): while L_1 is mostly well estimated, overestimation occurs for the other blocks, with CAGL producing slightly better estimates.

Evidently, CAGL is also considerably more costly than AGL: the average and standard deviation of computing times (for a single run) are $\mu = 8.96$ s and $\sigma = 2.51$ s for AGL; $\mu = 33.49$ s and $\sigma = 21.60$ s for CAGL; $\mu = 6.71$ s and $\sigma = 5.39$ s for BTD-GN with ranks $(8, 8, 8)$ and $\mu = 7.10$ s and $\sigma = 5.09$ s for BTD-GN with ranks $(8, 6, 4)$. Nonetheless, CAGL is able to attain a superior statistical performance when the multi-initialization scheme is used as the accumulation of the ECDF curves suggest that AGL's results cannot be further enhanced by more runs. Therefore, CAGL is a more costly but more accurate alternative than the other methods in this scenario.

Finally, the curves of Fig. 4(b) show that the above conclusions hold also for other values of SNR, though the advantage of CAGL and AGL with respect to BTD-GN decreases as the SNR grows.

TABLE III: Parameters of the synthetic AA signal model (14).

Model	P	a	Δa	f_a	F_s	f_0	Δf	F_f
1	5	150	50	0.08	1000	6	0.2	0.10
2	3	60	18	0.50	1000	8	0.3	0.23

V. Experimental results with ECG data

A. Tensor representation of ECG signals

ECG produces a time plot that represents the heart's electrical activity recorded from electrodes placed on the body surface. The ECG data matrix with K leads and N samples can be modeled as (5), where $\mathbf{X} \in \mathbb{R}^{K\times R}$ is the mixing matrix, that models the propagation of the cardiac electrical sources from the heart to the body surface, $\mathbf{S} \in \mathbb{R}^{N\times R}$ is the source matrix that contains the atrial, ventricular, and possibly disturbance sources, and R is the number of sources [14].

AA extraction in AF ECG recordings can be viewed as a BSS problem where the goal is to estimate the matrices $\mathbf{X}$ and $\mathbf{S}$ only from the observed data matrix $\mathbf{Y}$. The tensor built from the ECG data matrix as described in Section II-B then satisfies (6), where $\mathbf{H}_r \in \mathbb{R}^{M\times M}$ is a Hankel matrix built from the rth column of $\mathbf{S}$, and thus contains samples of the rth ECG source. Vector $\mathbf{x}_r$, which is the rth column of $\mathbf{X}$, quantifies the contribution of this source to each electrode's output, and so can be thought of as its spatial signature.

Due to the quasi-periodic nature of AF, atrial sources can be represented by the SCE model (2) with a small number of exponentials. Ventricular sources, in their turn, are typically composed by a few transient components, and thus can also be well modeled by (2) with small L. Hence, these signals can be mapped into low-rank Hankel matrices, as discussed in Section II-B.

B. Semi-synthetic AF data

The usefulness of CAGL for ECG source separation is now assessed by resorting to a semi-synthetic AF data model. To simulate the AA signal during AF, the model proposed in [34] that mimics the sawtooth pattern (a typical characteristic of the f waves) is used. This model is given by

$$s(n) = -\sum_{p=1}^{P} a_p(n)\sin\left(p\,\theta(n)\right) \tag{14}$$

with modulated amplitude and phase respectively given by

$$a_p(n) = \tfrac{2}{p\pi}\left[a + \Delta a \sin\left(2\pi\tfrac{f_a}{F_s}n\right)\right]$$

and

$$\theta(n) = 2\pi\tfrac{f_0}{F_s}n + \left(\tfrac{\Delta f}{F_f}\right)\sin\left(2\pi\tfrac{F_f}{F_s}n\right),$$

where a is the sawtooth amplitude, Δa is the modulation peak amplitude, f_a is the amplitude modulation frequency, F_s is the sampling frequency, f_0 is the frequency value in which $\theta(n)$ varies sinusoidally, Δf is the maximum frequency deviation and F_f is the modulation frequency.

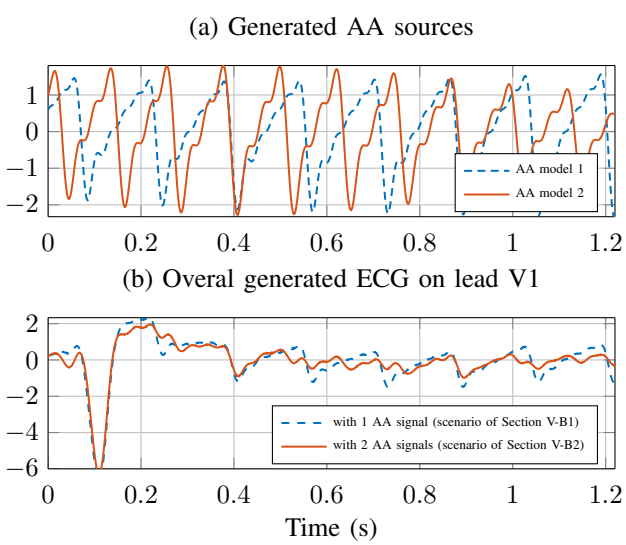

Fig. 5: Examples of generated semi-synthetic models: (a) AA sources following model (14) with the parameters shown in Table III; (b) overall synthesized ECG signals on lead V1.

1) One AA source: We first consider a scenario with one AA source $s(n)$, which is generated using (14) with the parameters of Model 1 given in Table III. This signal is shown in Fig. 5(a). A random spatial signature $\mathbf{x} \in \mathbb{R}^{12}$ over all 12 ECG leads is generated for this source, having standard normal i.i.d. components. The ventricular activity (VA) source is taken from a real 12-lead ECG of a healthy person, after P wave suppression as in [2]. This ECG, that belongs to the database of [35], is acquired at a sampling rate of 1 kHz and is preprocessed by a zero-phase forward-backward type-II Chebyshev bandpass filter with cutoff frequencies of 0.5 and 30 Hz, in order to suppress high-frequency noise and baseline wandering. Additive white Gaussian noise (AWGN) with variance σ^2 was added, yielding the overall model

$$\mathbf{Y} = \mathbf{V} + \alpha\,\mathbf{x}\mathbf{s}^{\mathsf{T}} + \mathbf{N} \in \mathbb{R}^{12\times N}, \tag{15}$$

where $\mathbf{V}$ holds the normalized VA signal, $\mathbf{s} \in \mathbb{R}^N$ holds the AA signal samples, $\mathbf{N}$ contains the AWGN samples and $\alpha = 2$ is a scaling factor chosen to obtain an average atrial-to-ventricular power ratio consistent with clinical observations. A window of about 1.2 seconds is used, yielding 1221 samples. An example of the overall generated signal is shown in Fig. 5(b) (dashed curve). A direct Hankelization of this matrix yields a tensor of dimensions $611 \times 611 \times 12$, whose approximate BTD demands a large computing time. To reduce it, we downsample the signals by a factor of 10 before computing the decomposition. The resulting tensor $\mathcal{Y}$ has dimensions $62 \times 62 \times 12$.

CAGL is run with the same γ-sweeping procedure used in Section IV, but now with γ taking 30 equispaced values in the interval $[8\times10^{-4}, 0.33\times10^{-2}]$ and keeping the last solution. For γ_0, we start the algorithm from random blocks, with the initial guess of $R = 6$ and $L = 40$. Among the estimated sources, the AA source is chosen as that which maximizes (in

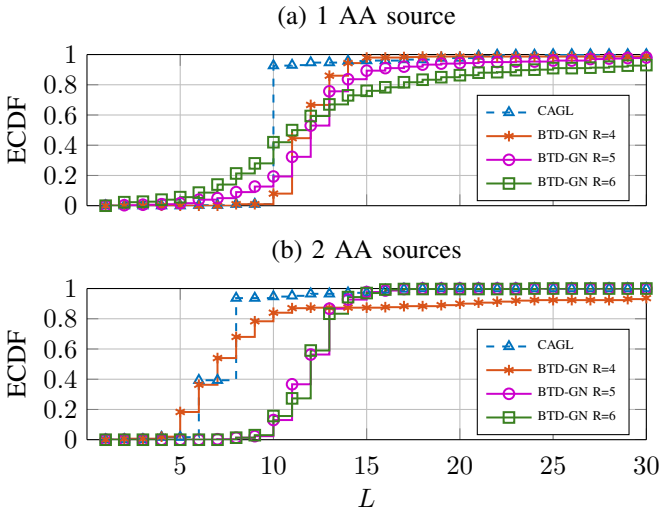

Fig. 6: Empirical distribution of rank chosen by CAGL for the AA source and of rank L yielding the best AA extraction for BTD-GN with different numbers of blocks R. The curves are shown only up to $L = 30$ for ease of visualization.

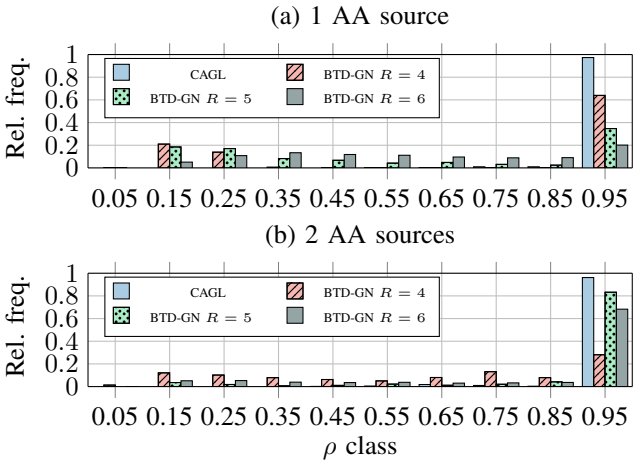

Fig. 7: Histogram of computed correlation coefficient ρ (in absolute value) between true and estimated AA sources.

absolute value) the correlation coefficient ρ with respect to the ground truth $s(n)$.

BTD-GN is run to estimate R blocks of fixed rank L, for all combinations $(R, L) \in \{4, 5, 6\} \times \{1, \ldots, 40\}$. For each rank L, the initial point is generated by filling the L columns of $\mathbf{A}_r$ and $\mathbf{B}_r$ with the first L columns used to initialize these variables in CAGL.

This procedure is repeated 30 times for each of 10 different realizations of $(\mathbf{x}, \mathbf{N})$. We found that the results of CAGL are remarkably consistent, producing very similar estimates regardless of the chosen initial point. In particular, for 95% of the runs, the final number of blocks is 4. By contrast, the results obtained with BTD-GN are much more sensitive in this respect. Specifically, the value of L that yields the best performance for a given run is not the same across different runs, as shown by the ECDF of Fig. 6(a). For CAGL, the rank chosen for the AA signal block is almost always 10 as seen in Fig. 6(a).

Though the most adequate choice for BTD-GN seems to be $R = 4$ and $L \in \{10, \ldots, 15\}$, its performance is highly variable for this range of L. This is seen in Fig. 7(a), which displays the histogram of the correlation coefficient ρ (in absolute value) between the estimated AA source and the ground truth. We have included all results produced by BTD-GN for $L \in \{10, \ldots, 15\}$, and all results produced by CAGL. It can be seen that the choice of R significantly affects performance, and a large proportion of results given by BTD-GN achieves a poor ρ for every R. By contrast, ρ is very likely to be quite close to 1 for CAGL.

In conclusion, BTD-GN only produces good results with a proper combination of R, L and initial point. By contrast, CAGL only requires choosing a reasonable range for γ and behaves much more robustly with regard to initialization.

2) Two AA sources: In this scenario, the VA source is still the same as in the previous scenario, but now two AA sources are generated, each one using the parameters of one row of Table III. These AA signals are shown in Fig. 5(a). Accordingly, the model now reads

$$\mathbf{Y} = \mathbf{V} + \mathbf{X}\mathbf{S}^{\mathsf{T}} + \mathbf{N} \in \mathbb{R}^{12 \times N}, \tag{16}$$

where each column of $\mathbf{S} \in \mathbb{R}^{N \times 2}$ contains one of the AA signals and those of $\mathbf{X} \in \mathbb{R}^{12 \times 2}$ hold their respective spatial signatures. Signals are again downsampled by a factor of 10, yielding data tensors with the same dimensions as before. Fig. 5(b) displays one example of the overall generated ECG signal (solid curve).

CAGL and BTD-GN are run following the same procedure as in the case with one AA source. However, here the estimated AA sources are extracted by choosing first the block with the highest correlation coefficient (in absolute value) with one of the reference AA sources, and secondly the block maximizing the correlation with the remaining AA source.

The corresponding results are shown in Fig. 6(b) and Fig. 7(b). The conclusions are similar to the previous scenario, with three main differences: (i) BTD-GN now performs best with $R = 5$ instead of $R = 4$, which is expected since two AA sources are now present; (ii) its performance is closer to that of CAGL for this choice of R; (iii) two choices of rank were most often made by CAGL (rather than one), namely $L_r = 6$ and $L_r = 8$.

Note that CAGL is run following exactly the same procedure as in the previous scenario. By contrast, BTD-GN now yields best results with a different choice of R. This highlights the fact that CAGL can effectively adapt to a given dataset, typically behaving more stably than the usual approach across different circumstances.

C. Real AF data

To further study the usefulness of CAGL in the target application, we perform experiments with real-world standard 12-lead ECG recordings from five patients suffering from persistent AF. These recordings belong to a database provided by the Cardiology Department of Princess Grace Hospital

TABLE IV: Block ranks of the ECG sources extracted by CAGL and characteristics of the potential AA sources.

Patient	Non-AA source ranks	AA source	SC (%)	DF (Hz)	$\hat{\kappa}$	AA source rank
1	3, 9, 10, 16, 18	1	80.23	6.44	163.69	8
2	28, 32	1	59.36	6.20	116.77	29
		2	82.05	6.20	163.77	12
3	21, 27, 29	1	66.13	6.20	132.29	20
4	8, 16, 20	1	74.51	5.96	196.16	10
5	32, 38, 39	1	91.70	5.72	348.42	10
		2	78.63	5.01	166.95	21

Center, in Monaco. They are acquired at a 977 Hz sampling rate and are preprocessed by a zero-phase forward-backward type-II Chebyshev bandpass filter with cutoff frequencies of 0.5 and 40 Hz, in order to suppress high-frequency noise and baseline wandering. For each patient, the segment with the largest TQ interval recording is chosen for the experiment. The recordings lengths range from about 1.08 to 1.40 seconds.

First, we downsample all signals by a factor of 10. This decreases computing cost, with practically negligible information loss. After normalization of each signal tensor, CAGL is applied using the same γ-sweeping strategy as in Section V-B. However, here we choose the final solution by inspection of the separated signals.

To assess AA estimation, we employ two commonly used performance parameters. The first one is spectral concentration (SC), *i.e.*, the relative amount of energy around the dominant frequency (DF), computed as in [36]:

$$\mathrm{SC} = \left(\sum_{f_i=0.82f_p}^{1.17f_p} P_{AA}(f_i)\right)\left(\sum_{f_i=0}^{F_s/2} P_{AA}(f_i)\right)^{-1},$$

where f_p is the value of the DF, defined as $\arg\max_{f_i} P_{AA}(f_i)$, F_s is the sampling frequency, f_i is the discrete frequency and P_{AA} is the power spectrum of the AA signal computed using Welch's method as in [36]. An AA signal during AF typically should have a DF between 3 and 9 Hz with high SC. The second parameter is kurtosis of the signal in the frequency domain, acquired by a 4096-point FFT. As in [37], we use a sample-based estimate $\hat{\kappa}$ of the general expression of kurtosis valid for non-circular complex data, given by

$$\kappa = \frac{E[|S_r(k)|^4] - 2E[|S_r(k)|^2]^2 - |E[S_r(k)^2]|^2}{E[|S_r(k)|^2]^2}$$

where $S_r(k)$ denotes the FFT of the rth source. As kurtosis measures peakedness and sparsity of a distribution, it naturally provides a quantitative measure of harmonicity of the signal when computed in the frequency domain. A high kurtosis is thus suggestive of a harmonic signal like AA during AF [4].

Table IV displays a quantitative assessment of the potential AA sources extracted for each patient, while Fig. 8 shows the estimated overall and AA signals for each patient along with the observed signals. In all experiments, we have chosen the initial guess $(R, L) = (6, 40)$ for the model structure, which was sufficient to produce satisfying results. Note that, for Patients 2 and 5, two blocks were identified as potential AA sources; accordingly, the estimated AA signals in Fig. 8(b) and Fig. 8(e) are given by the linear combinations of these blocks multiplied by their corresponding spatial weights for lead V1.

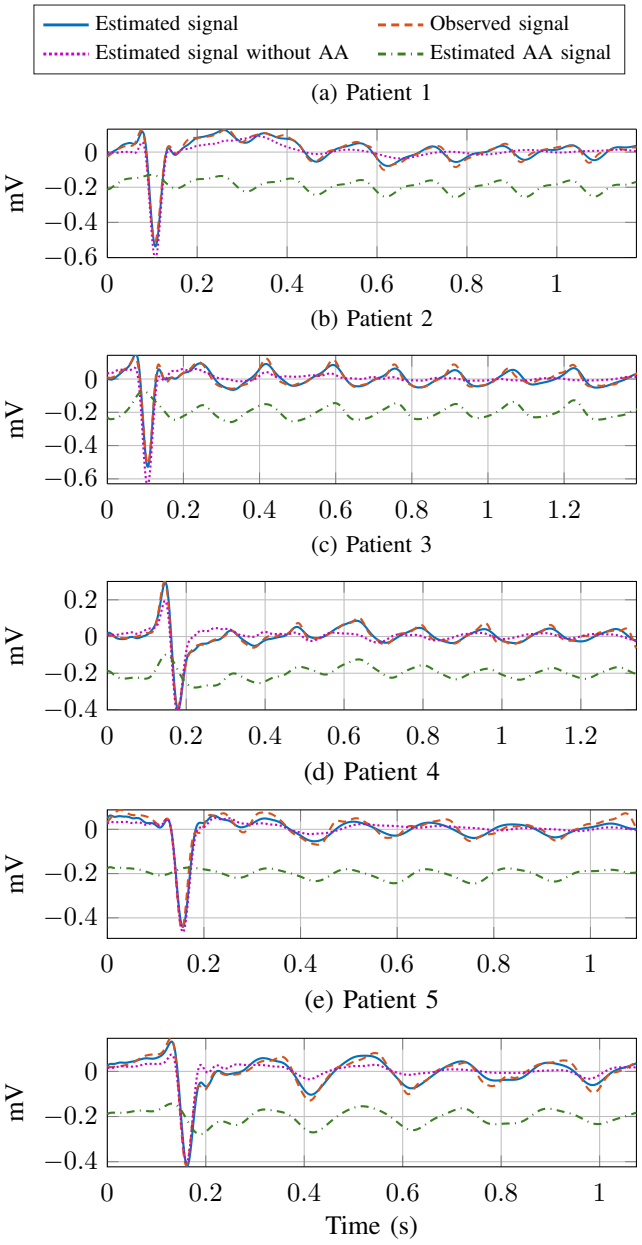

Fig. 8: Results produced by CAGL with real-world ECG data: observed and estimated signals at lead V1. The estimated AA signals are vertically shifted by -0.2 mV for ease of visualization.

These distinct potential AA sources for Patients 2 and 5 are shown in Fig. 9, along with their respective spatial weights (*i.e.*, their associated columns in $\mathbf{X}$). The shown signals are normalized as $[\max_n |s_r(n)|]^{-1} s_r(n)$, and the corresponding spatial weights are then rescaled so as to absorb the factor $\max_n |s_r(n)|$. Based on the results of Table IV and Fig. 8, it is seen that a satisfying extraction is achieved, as the potential AA sources have typical f-wave features, relatively high SC and physiologically plausible DF between 5 and 6.5 Hz.

Regarding Patient 2, the values of SC and $\hat{\kappa}$ for its AA source 1 are not so high due to the residual of the T-wave (ventricular repolarization), which can be seen around 0.3 seconds

of Fig. 9(a). Furthermore, Fig. 9(b) shows that the contribution of AA source 2 decays along the path V1–V2–V3, which suggests that this source may reflect electrical activity mainly occurring in the right atrium, and is almost null on other leads. By contrast, AA source 1 gives significant contributions to most leads, which suggests that the associated electrical propagation pattern may be harbored in a region including both atria. In fact, these sources have very different spatial signatures: the cosine of the angle between their respective columns of $\mathbf{X}$ is around -0.13, and so they are far from being collinear. Furthermore, despite having the same DF, their observed power spectra are considerably different. The lack of temporal synchronization between the estimated atrial sources, as manifested by the time lag between the maxima of the two signals plotted in Fig. 9(a), further supports the hypothesis that the associated activities may arise from different areas of atrial tissue.

Patient 5 provides another example in which two potential AA sources with considerably different spatial signatures are extracted: their respective columns in $\mathbf{X}$ form an angle whose cosine is about 0.09. Moreover, the dominant frequencies computed for these sources also differ significantly: 5.72 for AA source 1 and 5.01 for AA source 2. Here, it is AA source 2 that exhibits significant contributions to most leads, while AA source 1 manifests itself mostly on lead V1 and thus may correspond to more localized electrical activity taking place in the right atrium.

While the possibility of extracting more than one atrial source presents great interest for the noninvasive analysis of AF, a thorough validation of this result would be required by means of ground truth data such as a full electroanatomical mapping of the atria performed, for instance, during catheter ablation interventions. This validation, however, is out of the scope of this paper.

A final important observation is that for all but one patient, the sum of the estimated block ranks exceeds the dimension of the Hankel matrices: for Patient 1, $\sum_r L_r = 64 > M = 63$; for Patient 2, $\sum_r L_r = 101 > M = 74$; for Patient 3, $\sum_r L_r = 97 > M = 71$; and for Patient 5, $\sum_r L_r = 140 > M = 58$. This corroborates the practical importance of the case where $\sum_r L_r$ exceeds all tensor dimensions. Furthermore, for all patients the sum of block ranks far exceeds the number of leads, *i.e.*, $\sum_r L_r > K$, which showcases the benefit of using a tensor method, since a matrix decomposition approach could not possibly identify all the poles constituting each model.

VI. Conclusion

We have proposed an alternating optimization algorithm for a well-posed penalized formulation of the approximate BTD problem that jointly estimates the model structure and its parameters. The resulting subproblems can be solved by existing group lasso methods, and every limit point of the iterates produced by the alternating algorithm is a stationary point of the penalized criterion. Moreover, linear (subspace) structure can be imposed on the block matrices by using a structured low-rank approximation method, though a study of the convergence is still needed in this case. Experimental results with random tensors show that our approach is much more robust with respect to initialization than the conventional least-squares approach (without regularization).

To illustrate its practical usefulness, our algorithm has been applied to extract atrial activity signals from ECG recordings of atrial fibrillation episodes. In this problem, Hankel constraints must be imposed on the block matrices. Our results with semi-synthetic and real-world data highlight the ability of this approach to consistently perform an effective separation without the need of choosing structural parameters *a priori*. A particularly interesting result is the occurrence in (two of the five) real-world examples of two distinct sources with f-wave characteristics and very different spatial signatures. Further study is needed to give this result a more thorough physiological interpretation.

As future developments on approximate BTD computation, we can mention studies on how to automatically tune γ in practice and on the convergence of CAGL using a locally optimal structured low-rank approximation method to impose the linear constraints. Regarding the AA extraction in AF ECGs, future work would aim at analyzing the occurrence of multiple atrial sources, as well as performing the experiments in a large database of AF patients in order to provide statistically more significant results. Finally, the classification of subjects on the basis of their extracted ECG signals can also be envisaged.

Appendix A
Convergence of (unconstrained) AGL

By Theorem 2 of [25], every limit point of the sequence of iterates generated by AGL is a stationary point of (9) when the following conditions are met:

1) The overall cost function F must have compact sublevel sets, which holds since $F(\mathbf{A}, \mathbf{B}, \mathbf{X})$ is continuous and coercive.
2) The directional derivative of F along $(\Delta_\mathbf{A}, \Delta_\mathbf{B}, \Delta_\mathbf{X})$ must be positive whenever the directional derivatives along $(\Delta_\mathbf{A}, 0, 0)$, $(0, \Delta_\mathbf{B}, 0)$ and $(0, 0, \Delta_\mathbf{X})$ are positive, which is true because g is separable and f is differentiable on $\mathcal{S}$ (see [37, Lemma 3.1]).
3) The cost functions of the subproblems in $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{X}$ must be quasi-convex and have a unique solution, which clearly holds since each subproblem is strictly convex due to the proximal term.
4) The subproblems in $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{X}$ must satisfy the conditions stated by [25, Assumption 2], which as shown by [25, Proposition 2] are met when the following hold:
   - the cost function of the subproblem in $\mathbf{A}$ must be continuous and its smooth part must satisfy $f(\mathbf{A}, \mathbf{B}^{(t-1)}, \mathbf{X}^{(t-1)}) + \frac{\tau}{2}\|\mathbf{A} - \mathbf{A}^{(t-1)}\|_F^2 \geq f(\mathbf{A}, \mathbf{B}^{(t-1)}, \mathbf{X}^{(t-1)})$ with equality at $\mathbf{A} = \mathbf{A}^{(t-1)}$, which is clearly true; analogous requirements for $\mathbf{B}$ and $\mathbf{X}$ are also met.
   - the directional derivative of $g(\mathbf{A}, \mathbf{B}, \mathbf{X}) = \|\mathbf{A}\|_{2,1} + \|\mathbf{B}\|_{2,1} + \|\mathbf{X}\|_{2,1}$ must exist at every point $(\mathbf{A}, \mathbf{B}, \mathbf{X}) \in \mathcal{S}$, which also holds.

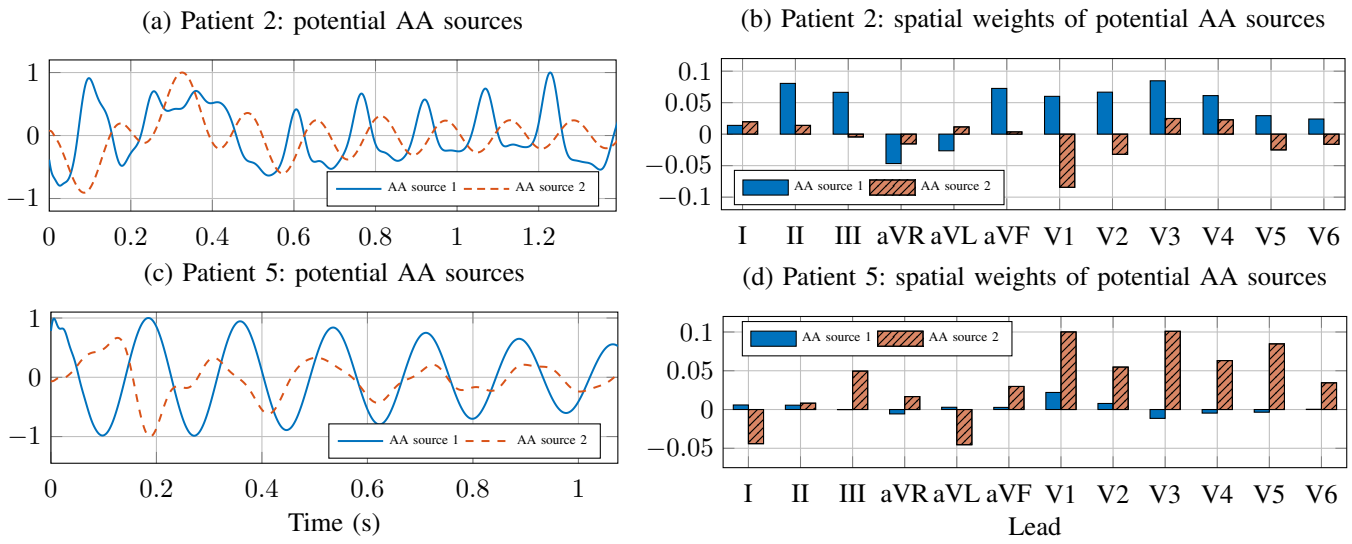

Fig. 9: Results produced by CAGL with ECG data from Patients 2 and 5.

ACKNOWLEDGMENT

The authors would like to thank O. Meste for helpful discussions on the application. J. H. de M. Goulart and P. Comon were supported by the European Research Council under the European Programme FP7/2007-2013, Grant AdG-2013-320594 "DECODA." P. M. R. de Oliveira is funded by a PhD scholarship from the IT Doctoral School (ED STIC) of the Université Côte d'Azur.

REFERENCES

[1] B. Hunyadi, D. Camps, L. Sorber, W. Van Paesschen, M. De Vos, S. Van Huffel, and L. De Lathauwer, "Block term decomposition for modelling epileptic seizures," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, p. 139, 2014.

[2] L. N. Ribeiro, A. R. Hidalgo-Muñoz, and V. Zarzoso, "Atrial signal extraction in atrial fibrillation electrocardiograms using a tensor decomposition approach," in *Proc. IEEE Eng. Med. Biol. Soc. Conf. (EMBC)*. Milan, Italy: IEEE, aug 2015, pp. 6987–6990.

[3] V. Zarzoso, "Parameter estimation in block term decomposition for noninvasive atrial fibrillation analysis," in *Proc. IEEE 7th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*. Curaçao, Dutch Antilles: IEEE, dec 2017.

[4] P. M. R. de Oliveira and V. Zarzoso, "Source analysis and selection using block term decomposition in atrial fibrillation," in *Proc. 14th Int. Conf. , LVA/ICA 2018*. Guildford, UK: Springer, jul 2018, pp. 46–56.

[5] ——, "Block term decomposition of ECG recordings for atrial fibrillation analysis: Temporal and inter-patient variability," *J. Commun. and Inf. Syst.*, vol. 34, no. 1, pp. 111–119, 2019.

[6] I. Markovsky, O. Debals, and L. D. Lathauwer, "Sum-of-exponentials modeling and common dynamics estimation using tensorlab," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 14 150 – 14 155, 2017, 20th IFAC World Congress.

[7] L. De Lathauwer, "Blind separation of exponential polynomials and the decomposition of a tensor in rank-$(L_r, L_r, 1)$ terms," *SIAM J. Matrix Anal. Appl.*, vol. 32, no. 4, pp. 1451–1474, 2011.

[8] ——, "Decompositions of a higher-order tensor in block terms–Part II: Definitions and uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1033–1066, 2008.

[9] N. Vervliet, O. Debals, L. Sorber, M. V. Barel, and L. De Lathauwer, "Tensorlab 3.0," Mar. 2016, available online. URL: http://www.tensorlab.net.

[10] L. De Lathauwer and D. Nion, "Decompositions of a higher-order tensor in block terms–Part III: Alternating least squares algorithms," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1067–1083, 2008.

[11] J. H. de M. Goulart and P. Comon, "On the minimal ranks of matrix pencils and the existence of a best approximate block-term tensor decomposition," *Linear Algebra Appl.*, vol. 561, pp. 161–186, 2019.

[12] X. Han, L. Albera, A. Kachenoura, H. Shu, and L. Senhadji, "Block term decomposition with rank estimation using group sparsity," in *Proc. IEEE 7th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*. Curaçao, Dutch Antilles: IEEE, dec 2017.

[13] J. J. Rieta, F. Castells, C. Sánchez, V. Zarzoso, and J. Millet, "Atrial activity extraction for atrial fibrillation analysis using blind source separation," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1176–1186, 2004.

[14] V. Zarzoso, "Extraction of ECG characteristics using source separation techniques: exploiting statistical independence and beyond," in *Adv. Biosignal Process.* Springer, 2009, pp. 15–47.

[15] L. N. Ribeiro, A. L. De Almeida, and V. Zarzoso, "Enhanced block term decomposition for atrial activity extraction in atrial fibrillation ECG," in *Proc. Sensor Array Multichannel Signal Process. Workshop (SAM), 2016 IEEE*. Rio de Janeiro, Brazil: IEEE, jul 2016, pp. 1–5.

[16] D. L. Boley, F. T. Luk, and D. Vandevoorde, "Vandermonde factorization of a Hankel matrix," in *Proc. Workshop Scientific Comput.*, Hong Kong, 1997.

[17] L. L. Scharf and C. Demeure, *Statistical signal processing: detection, estimation, and time series analysis*. Addison-Wesley Reading, MA, 1991, vol. 63.

[18] M. Sørensen and L. D. De Lathauwer, "Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank-$(l_{r,n}, l_{r,n}, 1)$ terms—Part I: Uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 36, no. 2, pp. 496–522, 2015.

[19] M. Sørensen, I. Domanov, and L. De Lathauwer, "Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank-$(l_{r,n}, l_{r,n}, 1)$ terms—Part II: Algorithms," *SIAM J. Matrix Anal. Appl.*, vol. 36, no. 3, pp. 1015–1045, 2015.

[20] Y. Qi, M. Michałek, and L.-H. Lim, "Complex tensors almost always have best low-rank approximations," *arXiv preprint arXiv:1711.11269v2*, sep 2018.

[21] S.-Y. Kung, K. S. Arun, and D. V. B. Rao, "State-space and singular-value decomposition-based approximation methods for the harmonic retrieval problem," *J. Optical Soc. of America*, vol. 73, no. 12, pp. 1799–1811, 1983.

[22] M. Elad, P. Milanfar, and G. H. Golub, "Shape from moments—An estimation theory perspective," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1814–1829, 2004.

[23] I. Markovsky, *Low-rank approximation*, ser. Communications and Control Engineering. Springer, 2012.

[24] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Stat. Soc.: Series B (Stat. Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[25] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence

analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.
[26] J. Liu and J. Ye, "Moreau-Yosida regularization for grouped tree structure learning," in *Proc. Adv. Neural Inf. Process. Systems*, 2010, pp. 1459–1467.
[27] L. De Lathauwer and A. de Baynast, "Blind deconvolution of DS-CDMA signals by means of decomposition in rank-$(1, l, l)$ terms," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1562–1571, 2008.
[28] J. Spiegelberg, J. Rusz, and K. Pelckmans, "Tensor decompositions for the analysis of atomic resolution electron energy loss spectra," *Ultramicroscopy*, vol. 175, pp. 36–45, 2017.
[29] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.
[30] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logist. Q.*, vol. 2, no. 1-2, pp. 83–97, 1955.
[31] P. Paatero, "Construction and analysis of degenerate PARAFAC models," *J. Chemom.*, vol. 14, no. 3, pp. 285–299, 2000.
[32] J. A. Cadzow, "Signal enhancement—a composite property mapping algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 1, pp. 49–62, 1988.
[33] M. T. Chu, R. E. Funderlic, and R. J. Plemmons, "Structured low rank approximation," *Linear Algebra Appl.*, vol. 366, pp. 157–172, 2003.
[34] M. Stridh and L. Sornmo, "Spatiotemporal QRST cancellation techniques for analysis of atrial fibrillation," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 1, pp. 105–111, 2001.
[35] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
[36] F. Castells, J. J. Rieta, J. Millet, and V. Zarzoso, "Spatiotemporal blind source separation approach to atrial activity estimation in atrial tachyarrhythmias," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 2, pp. 258–267, 2005.
[37] V. Zarzoso and P. Comon, "Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size," *IEEE Trans. Neural Netw.*, vol. 21, no. 2, pp. 248–261, 2010.