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 which allows jointly estimating the model parameters and its structure. 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], [6] 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 a identifiable low-rank matrix factorization involving a Vandermonde matrix [7], 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 measured signal 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 [8]. 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. 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 [9], [8].

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 [10] or the alternating least squares method with enhanced line search (ALS-ELS) [11]. 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 [12]; 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 [13], 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 computation as the minimization of a (least-squares) fitting

P. Comon is with Univ. Grenoble Alpes, CNRS, Gipsa-Lab, F-38000 Grenoble, France (pierre.comon@gipsa-lab.fr).

J. H. de M. Goulart, 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 ({goulart, marinho, cabral, zarzoso}@i3s.unice.fr).

Manuscript received XXX XX, 20XX; revised XXX XX, 20XX.

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), which is simpler than the algorithm of [13] and provably convergent. We also devise a variant termed constrained AGL (CAGL) to deal with linear (subspace) constraints over block matrices. For this purpose, alternating projections are performed at each iteration to ensure that the block matrices have low rank and belong to the specified subspace. This strategy is well-known as Cadzow's algorithm (CA) [14], and allows in particular imposing a low-rank Hankel structure.

As an application of CAGL, we consider the problem of extracting the atrial activity (AA) in electrocardiograms (ECGs) of atrial fibrillation (AF). 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 the one we propose in this paper.

Other BSS techniques such as PCA and ICA have been applied to noninvasive AA extraction [15]–[16] 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 [3], [17], [4], [5], [6]. 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 [3], [17], [4], [5], [6], but differs from [3], [17], [5], [6], since we do not impose the structural parameters of the model. It also differs from [4], 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 the following: in Section II we present the problem of source separation of sums of complex exponentials using BTD with Hankel structure. Section III formulates regularized BTD approximation problem and the algorithms AGL and CAGL. A numerical evaluation of AGL algorithm is shown in Section IV. The application of CAGL approach to AA extraction is then presented in Section V, where semi-synthetic ECG datasets and real ECG datasets are considered. Finally, conclusions are given in Section VI.

Notation: Tensors are denoted in uppercase bold script letters $\boldsymbol{\mathcal{X}}$, matrices in uppercase bold letters $\boldsymbol{X}$, vectors in lowercase bold letters $\mathbf{x}$ and scalars in lowercase letters x. 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 $\|\mathbf{X}\|_{2,1} = \sum_{r=1}^{R} \|\mathbf{x}_r\|_2$, respectively. The symbols $\otimes$, $\boxtimes$ and $\odot$ are used for tensor, Kronecker, and Khatri-Rao (column-wise Kronecker) product, respectively. Block Khatri-Rao product is denoted by $\odot_L$, for a product $\boldsymbol{Z} = \boldsymbol{X} \odot_L Y$ with $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_R]$ and $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_{LR}]$, the ith block of $\boldsymbol{Z}$ is given by $\mathbf{x}_i \boxtimes \left[\mathbf{y}_{(i-1)L+1} \cdots \mathbf{y}_{(i-1)L+L}\right]$. $\operatorname{Diag}\mathbf{x}$ denotes a diagonal tensor or matrix with entries given by $\mathbf{x}$. 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 [8] 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

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

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 [18]

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

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 [19]. It implies that a "simple" signal of the form (1) 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 real-valued signals, usual conjugacy conditions must hold in (1). Correspondingly, complex-valued columns of $\mathbf{V}_s$ as well as coefficients α_ℓ must arise in complex conjugate pairs.

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

$$s_r(n) = \sum_{\ell=1}^{L_r} \alpha_\ell^{(r)} \exp\left(\zeta_\ell^{(r)} n\right), \quad L_r < M, \tag{2}$$

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 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

$$\mathbf{Y} = \mathbf{X}\mathbf{S}^{\mathsf{T}}, \tag{3}$$

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

$$\begin{aligned} \mathcal{Y} &= \textstyle\sum_{k=1}^{K} \mathbf{Y}_k \otimes \mathbf{e}_k = \sum_{k=1}^{K} \left(\sum_{r=1}^{R} x_{k,r} \mathbf{H}_r\right) \otimes \mathbf{e}_k \\ &= \textstyle\sum_{r=1}^{R} \mathbf{H}_r \otimes \left(\sum_{k=1}^{K} x_{k,r} \mathbf{e}_k\right) \\ &= \textstyle\sum_{r=1}^{R} \mathbf{H}_r \otimes \mathbf{x}_r, \end{aligned} \tag{4}$$

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 (4).

It turns out that the tensor decomposition (4), known as block-term decomposition (BTD) and introduced by [9], is essentially[1] unique under relatively mild assumptions. Its uniqueness properties have been first studied in [9], and further results were given in [8]. In particular, Theorem 2.4 of [8] 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 (4) is essentially unique. (It is thus necessary that $L_r < M$.)

C. Approximate block term decomposition

In practice, the data matrix $\mathbf{Y}$ is only approximately given by (3), due to noise and imperfect modeling. Hence, one can only approximate tensor $\mathcal{Y}$ by a low-rank BTD model of the form in (4).

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 (4) and considering a third-order tensor $\mathcal{Y} \in \mathbb{C}^{I \times J \times K}$ (in model (4) 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. Mostly often, a least-squares criterion is adopted (as in, *e.g.*, [11]), leading to

$$\min_{(\mathbf{A},\mathbf{B},\mathbf{X}) \in \mathcal{S}} f(\mathbf{A},\mathbf{B},\mathbf{X}) \triangleq \left\| \mathcal{Y} - \textstyle\sum_{r=1}^{R} \left(\mathbf{A}_r \mathbf{B}_r^{\mathsf{T}}\right) \otimes \mathbf{x}_r \right\|_F^2 \tag{5}$$

with $\mathcal{S} \triangleq \mathbb{C}^{I \times \sum_{r=1}^{R} L_r} \times \mathbb{C}^{J \times \sum_{r=1}^{R} L_r} \times \mathbb{C}^{K \times R}$, 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 (5), and also the imposition of a Hankel constraint over the blocks in (4).

1) Existence: Problem (5) 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 [11]. As recently shown in [12], 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.

3) Linear constraints over block matrices: In the special case of interest (4), $\mathbf{H}_r$ must belong to the subspace of $M \times M$ Hankel matrices, $\mathcal{H}$. Although the slices $\mathbf{Y}_k$ are Hankel by construction, there is no reason why a solution $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{X}})$ of (11) or of (5) should satisfy $\hat{\mathbf{A}}_r \hat{\mathbf{B}}_r^{\mathsf{T}} \in \mathcal{H}$. 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 (1).

To date, this issue has been circumvented by projecting the unconstrained estimated block matrices onto $\mathcal{H}$ [1], [5]:

$$\hat{\mathbf{H}}_r = \mathcal{P}_{\mathcal{H}}(\hat{\mathbf{A}}_r \hat{\mathbf{B}}_r^{\mathsf{T}}), \quad r = 1, \ldots, R, \tag{6}$$

[1]Note that (4) 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 $\alpha \neq 0$.

where $\mathcal{P}_{\mathcal{H}}$ is the orthogonal projector onto the Hankel subspace $\mathcal{H}$. However, there are two problems with this approach:

(i) It can happen in practice that $\|\mathcal{P}_{\mathcal{H}^\perp}(\hat{\mathbf{A}}_r\hat{\mathbf{B}}_r^\mathsf{T})\|_F^2 \not\ll \|\hat{\mathbf{H}}_r\|_F^2$ for one or more indices r, where $\mathcal{H}^\perp$ denotes the orthogonal complement of $\mathcal{H}$. In this case, it is also hard to interpret the results, because a significant portion of the "energy" of the rth block is discarded.

(ii) $\hat{\mathbf{H}}_r$ has full rank in general, and thus the simplicity constraint (small number of poles) is not satisfied anymore, not even approximately if $\|\mathcal{P}_{\mathcal{H}^\perp}(\hat{\mathbf{A}}_r\hat{\mathbf{B}}_r^\mathsf{T})\|_F$ is large.

Instead of performing a single projection as in (6), one can employ Cadzow's Algorithm (CA) [14], performing a sequence of alternating projections onto $\mathcal{H}$ and the (nonconvex) set $\mathcal{L}_r \triangleq \{\mathbf{M} \in \mathbb{C}^{M\times M} : \operatorname{rank}\mathbf{M} \leq \hat{L}_r\}$, where $\hat{L}_r \triangleq \operatorname{rank}\hat{\mathbf{A}}_r\hat{\mathbf{B}}_r^\mathsf{T}$. It is cheap to apply $\mathcal{P}_{\mathcal{H}}$: it suffices to compute the means of antidiagonals of its argument. Computing $\mathcal{P}_{\mathcal{L}_r}$ is also easy but more costly, since it requires truncating a singular value decomposition (SVD) according to the rank bound. This strategy addresses issue (ii) above, but issue (i) still remains.

Finally, note that the above discussion applies almost entirely to arbitrary constraints of the form $\mathbf{H}_r \in \mathcal{U}$ where $\mathcal{U}$ is a subspace of $\mathbb{C}$, except for the fact that the projector $\mathcal{P}_{\mathcal{U}}$ may be not as cheap to apply as $\mathcal{P}_{\mathcal{H}}$. Though CA is known to converge to generally suboptimal solutions [21], it is very simple to implement and to adapt to arbitrary constraints, whilst often producing good solutions in practice.

III. Alternating group lasso algorithm for BTD

In the following, we derive a provably convergent algorithm for computing an (unconstrained) approximate BTD of a given tensor. 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, one can include penalization terms promoting low-rank blocks and controlling the number of blocks in the formulation, as in

$$\min_{(\mathbf{A},\mathbf{B},\mathbf{X})\,\in\,\mathcal{S}} F(\mathbf{A},\mathbf{B},\mathbf{X}) \triangleq f(\mathbf{A},\mathbf{B},\mathbf{X}) + \gamma\, g(\mathbf{A},\mathbf{B},\mathbf{X}), \quad (7)$$

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 (5), $\gamma > 0$ is a regularization parameter and g is a regularization function of the form

$$g(\mathbf{A},\mathbf{B},\mathbf{X}) \triangleq \|\mathbf{A}\|_{2,1} + \|\mathbf{B}\|_{2,1} + \|\mathbf{X}\|_{2,1}. \quad (8)$$

Adding a mixed $\ell_{2,1}$-norm regularization term is a well-known strategy for inducing *group sparsity* of its argument's columns. 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 [22].

Hence, for sufficiently high γ, minimizers of (7) 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 (5), 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 coercive (due to g) and continuous.

B. Algorithm for unconstrained blocks

To tackle the nonconvex and nonsmooth problem (7), 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 $\hat{\mathbf{A}}^{(t)}$, $\hat{\mathbf{B}}^{(t)}$ and $\hat{\mathbf{X}}^{(t)}$ denote the estimates obtained at iteration t. Fixing $\mathbf{B} = \hat{\mathbf{B}}^{(t)}$ and $\mathbf{X} = \hat{\mathbf{X}}^{(t)}$ in (7), the subproblem in $\mathbf{A}$ of iteration $t+1$ becomes a standard group lasso problem

$$\begin{aligned} &\min\nolimits_{\mathbf{A}\in\mathbb{C}^{I\times LR}}\ F_{\mathbf{A}}^{(t)}(\mathbf{A}), \quad (9)\\ &\text{with}\quad F_{\mathbf{A}}^{(t)}(\mathbf{A}) \triangleq \tfrac{1}{2}\|\mathcal{Y} - \mathcal{W}_{\mathbf{A}}^{(t)}(\mathbf{A})\|_F^2 + \gamma\,\|\mathbf{A}\|_{2,1}, \end{aligned}$$

where $\mathcal{W}_{\mathbf{A}}^{(t)}$ is a linear map depending on $\hat{\mathbf{B}}^{(t)}$ and $\hat{\mathbf{X}}^{(t)}$. The groups are disjoint and correspond to the columns of $\mathbf{A}$. Existing group lasso algorithms such as that in [23] can be readily invoked[2] to solve (9), which is convex. Similar subproblems can be derived for $\mathbf{B}$ and $\mathbf{X}$.

Now, despite being convex, subproblem (9) may not be strictly convex, because $\mathcal{W}_{\mathbf{A}}^{(t)}$ may not be injective at some iterations. It can thus fail to have a unique solution, and convergence to a stationary point cannot be guaranteed [24]. One can remedy this shortcoming by adding a proximal term, as in $\min_{\mathbf{A}\in\mathbb{C}^{I\times LR}} F_{\mathbf{A}}^{(t)}(\mathbf{A}) + \frac{\tau}{2}\|\mathbf{A} - \hat{\mathbf{A}}^{(t)}\|_F^2$, with $\tau > 0$. Putting $\mathbf{a} \triangleq \operatorname{vec}(\mathbf{A})$ 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.* [8]), this is equivalent to

$$\min_{\mathbf{a}\in\mathbb{C}^{ILR}} \frac{1}{2}\left\|\begin{bmatrix}\mathbf{y}_1\\ \sqrt{\tau}\,\hat{\mathbf{a}}^{(t)}\end{bmatrix} - \begin{bmatrix}\mathbf{W}_{\mathbf{A}}^{(t)}\\ \sqrt{\tau}\mathbf{I}\end{bmatrix}\mathbf{a}\right\|_2^2 + \gamma\sum_{r=1}^{R}\sum_{l=1}^{L}\|\mathbf{a}_{r,l}\|_2, \quad (10)$$

where $\mathbf{W}_{\mathbf{A}}^{(t)} \triangleq \left(\hat{\mathbf{X}}^{(t)} \odot_L \hat{\mathbf{B}}^{(t)}\right) \boxtimes \mathbf{I}_I$ 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, (10) is strictly convex, since the Hessian of the least-squares term is positive definite. Analogous strictly convex subproblems can also be derived for $\mathbf{B}$ and $\mathbf{X}$, with

$$\begin{aligned} \mathbf{W}_{\mathbf{B}}^{(t)} &\triangleq \left(\hat{\mathbf{X}}^{(t)} \odot_L \hat{\mathbf{A}}^{(t+1)}\right) \boxtimes \mathbf{I}_J,\\ \mathbf{W}_{\mathbf{X}}^{(t)} &\triangleq \left[\left(\hat{\mathbf{B}}^{(t+1)} \odot \hat{\mathbf{A}}^{(t+1)}\right)\operatorname{Diag}(\mathbf{1}_L,\ldots,\mathbf{1}_L)\right] \boxtimes \mathbf{I}_K. \end{aligned}$$

The 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 [11]. As we explain in Appendix A, AGL's formulation satisfies the conditions of [24, Theorem 2], and hence its iterates converge to a stationary point of problem (7).

It should be noted that AGL is a general approach to compute an unconstrained approximate low-rank BTD and is

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

TABLE I: Pseudocode for constrained AGL algorithm.

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 (10) to obtain $\mathbf{A}^{(t)}$ from $\mathbf{A}^{(t-1)}$, $\mathbf{B}^{(t-1)}$ and $\mathbf{X}^{(t-1)}$
4: **for** $r = 1, \ldots, R$ **do**
5: $L_r^{(t)} \leftarrow$ number of nonzero columns in $\mathbf{A}_r^{(t)}$
6: $(\mathbf{A}_r^{(t)}, \mathbf{B}_r^{(t-1)}) \leftarrow \texttt{cadzow}(\mathbf{A}_r^{(t)}(\mathbf{B}_r^{(t-1)})^\mathsf{T}, L_r^{(t)})$
7: $(\mathbf{A}_r^{(t)}, \mathbf{B}_r^{(t-1)}) \leftarrow ([\mathbf{A}_r^{(t)}\ \mathbf{0}_{I \times L - L_r^{(t)}}], [\mathbf{B}_r^{(t-1)}\ \mathbf{0}_{I \times L - L_r^{(t)}}])$
8: Solve group lasso subproblem in $\mathbf{B}$ analogous to (10) to obtain $\mathbf{B}^{(t)}$ from $\mathbf{A}^{(t)}$, $\mathbf{B}^{(t-1)}$ and $\mathbf{X}^{(t-1)}$
9: **for** $r = 1, \ldots, R$ **do**
10: $L_r^{(t)} \leftarrow$ number of nonzero columns in $\mathbf{B}_r^{(t)}$
11: $(\mathbf{A}_r^{(t)}, \mathbf{B}_r^{(t)}) \leftarrow \texttt{cadzow}(\mathbf{A}_r^{(t)}(\mathbf{B}_r^{(t)})^\mathsf{T}, L_r^{(t)})$
12: $(\mathbf{A}_r^{(t)}, \mathbf{B}_r^{(t)}) \leftarrow ([\mathbf{A}_r^{(t)}\ \mathbf{0}_{I \times L - L_r^{(t)}}], [\mathbf{B}_r^{(t)}\ \mathbf{0}_{I \times L - L_r^{(t)}}])$
13: Solve group lasso subproblem in $\mathbf{X}$ analogous to (10) to obtain $\mathbf{X}^{(t)}$ from $\mathbf{A}^{(t)}$, $\mathbf{B}^{(t)}$ and $\mathbf{X}^{(t-1)}$
14: $t \leftarrow t + 1$

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], [25] and [26].

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

In order to address issues (i) and (ii) stated in Section II-C3, we propose a variant of AGL, termed constrained AGL (CAGL), to address the problem

$$\min_{(\mathbf{A},\mathbf{B},\mathbf{X}) \in \mathcal{S}} F(\mathbf{A}, \mathbf{B}, \mathbf{X}) \quad \text{subj. to} \quad \forall r,\ \mathbf{A}_r \mathbf{B}_r^\mathsf{T} \in \mathcal{H}. \quad (11)$$

In CAGL, the constraint in (11) is enforced by applying CA 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.

Note that each application of CA re-estimates both $\mathbf{A}_r$ and $\mathbf{B}_r$, which are updated as the factors of the last SVD computed by CA. The reason for this joint update is that, for a fixed rank L' and a fixed $J \times L'$ matrix $\mathbf{V}_r$, it can happen that the only solution $I \times L'$ matrix $\mathbf{U}_r$ for $\mathbf{U}_r \mathbf{V}_r^\mathsf{T} \in \mathcal{H}$ is $\mathbf{U}_r = \mathbf{0}$. Another observation is that the rank estimate $\hat{L}_r$ given to CA is a conservative one, since we only take into account the columns of one factor. 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 CA, as done in lines 7 and 12 of Table I.

As we will see in Section V, in practice CAGL yields meaningful solutions to the source separation problem and seems to be robust with respect to its initialization, in spite of the suboptimality of CA. The price to pay for constraining the block matrices is a heavier computational load, mainly due to the computation of the SVDs required by CA. Note also that any other structured low-rank approximation algorithm, such as those in [27], can be used in place of CA.

IV. Numerical evaluation on random block term decomposition models

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

A. Unconstrained BTD

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 $\max_{i,j} |\mathbf{x}_i^\mathsf{T} \mathbf{x}_i| < 0.9$ 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 achieve SNR $\triangleq \|\mathcal{Y}_0\|_F^2\, \sigma_\mathcal{N}^{-2} \|\mathcal{N}\|_F^{-2} = 20$ dB.

We set $I = J = 8$ and $K = R = 3$. The block ranks are $L_1 = 4$, $L_2 = 3$ and $L_3 = 2$. For each realization, AGL is applied three times with $L = 4$ using the following procedure:

- starting from a random initial point, the algorithm is run with $\gamma = \gamma_0$, producing an initial solution;
- for $p = 1, \ldots, P-1$, it 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 [28] and produces a sequence of candidate solutions. Of all P candidate solutions, we keep the best one according to the normalized mean squared error (NMSE) over the blocks:

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

For comparison, Tensorlab's Gauss-Newton BTD algorithm (BTD-NLS) [10] is also used. The same rank L is used for all three blocks, since true ranks (or at least adequate ones) are typically unknown in practice. Ten random initializations are used, the first three being the same as used by AGL. Fig. 1(a) shows the empirical cumulative distribution function (ECDF) of NMSE obtained by selecting the best solution among the first N_i initializations. For BTD-NLS, we let N_i vary from 1 to 10; for AGL, it varies from 1 to 3. Clearly, AGL outperforms BTD-NLS by a significant margin. Furthermore, its performance seems much less sensitive with respect to initialization. Table II displays the proportion of realizations for which each block rank is estimated by AGL at a given value, for $N_\text{i} = 2$. It shows that AGL is able to correctly estimate each block's rank most of the time, though there is some non-negligible chance of overestimation.

Now, even if the true ranks are given as input to BTD-NLS (which is an unrealistic assumption), it is still outperformed by AGL with $N_\text{i} \in \{2, 3\}$, as seen in Fig. 1(b). The average computing times and standard deviations (in seconds) for AGL with $N_\text{i} = 1$ and $N_\text{i} = 2$ are ($\mu = 25.56$, $\sigma = 9.70$) and ($\mu = 50.75$, $\sigma = 14.65$), respectively; for BTD-NLS with ranks $(4, 4, 4)$, they are ($\mu = 71.24$, $\sigma = 32.38$), and when the true ranks $(4, 3, 2)$ are used, ($\mu = 70.13$, $\sigma = 34.78$).

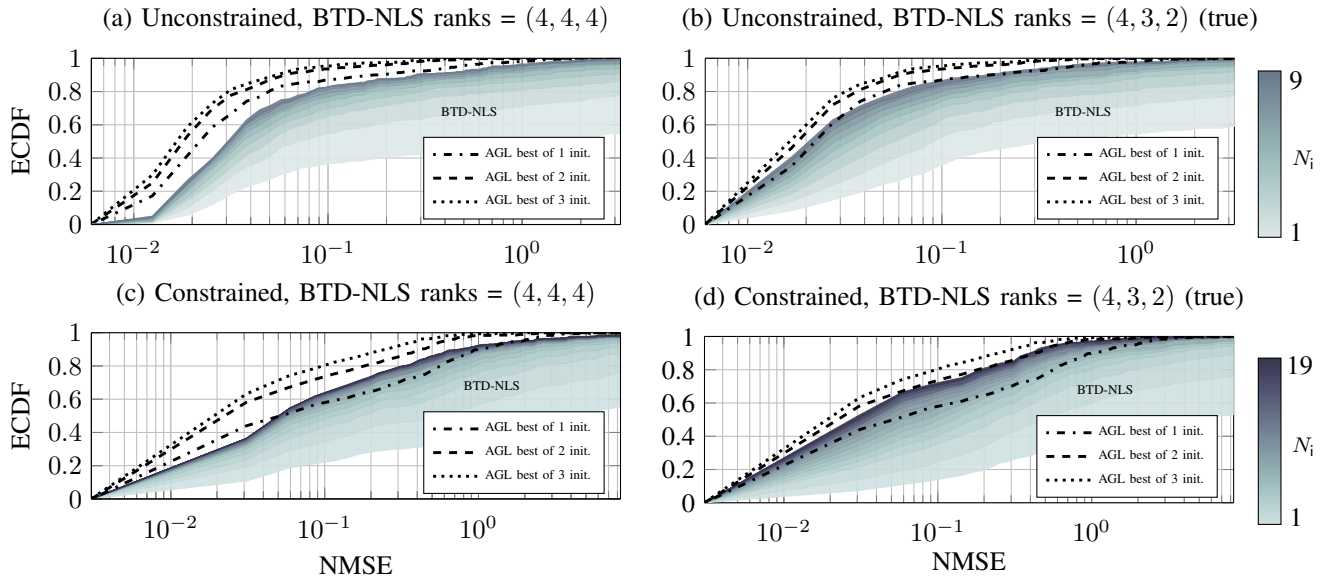

Fig. 1: Empirical CDFs of NMSE (over estimated blocks) attained by AGL and BTD-NLS for 500 realizations of a noisy random BTD model, in the constrained and unconstrained scenarios. BTD-NLS' results are portrayed by the shaded regions: each such region indicates the area between the curve obtained by keeping the best result among N_{i} initializations and that obtained by keeping the best result among $N_{\mathrm{i}} + 1$ initializations.

TABLE II: Proportion of block rank values estimated by AGL in both scenarios of Section IV, when the best solution obtained with two initializations ($N_{\mathrm{i}} = 2$) is kept.

	Unconstrained			Constrained		
	$L_1 = 4$	$L_2 = 3$	$L_1 = 2$	$L_1 = 4$	$L_2 = 3$	$L_1 = 2$
$\hat{L}_i = 1$	00.0%	00.0%	00.4%	00.0%	00.4%	01.6%
$\hat{L}_i = 2$	00.2%	00.6%	65.4%	01.2%	04.6%	52.2%
$\hat{L}_i = 3$	02.2%	69.2%	22.2%	06.6%	53.8%	12.8%
$\hat{L}_i = 4$	97.6%	30.2%	12.0%	92.2%	41.2%	33.2%

B. Constrained BTD

A similar evaluation is performed in this scenario, also with 500 realizations and SNR= 20 dB, but now the block matrices are random low-rank Hankel matrices generated using CA. Moreover, to enforce the constraint after convergence, CA is also applied to the outcomes of BTD-NLS. Setting all ranks to $L = 4$ in BTD-NLS leads to the results shown in Fig. 1(c). The performance of both CAGL and BTD-NLS are worse in comparison to the previous scenario, with CAGL still ahead. Yet, it appears to be more sensitive with respect to initialization now. Also, block rank estimation is less accurate, as shown in Table II. If the true ranks are given to BTD-NLS, then CAGL with $N_{\mathrm{i}} = 2$ still produces slightly better results than BTD-NLS with $N_{\mathrm{i}} = 20$, as seen in Fig. 1(d). The average computing times and standard deviations (in seconds) for CAGL ($N_{\mathrm{i}} = 2$) are ($\mu = 199.46$, $\sigma = 145.25$); for BTD-NLS ($N_{\mathrm{i}} = 20$) with ranks $(4, 4, 4)$, they are ($\mu = 207.97$, $\sigma = 96.07$), and when the true ranks $(4, 3, 2)$ are used, ($\mu = 218.56$, $\sigma = 87.39$). Hence, overall CAGL outperforms BTD-NLS also in this scenario.

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 (3), 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 [16].

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 (4), 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 (1) 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 (1) 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 [29]

TABLE III: Parameters of the synthetic AA signal model (12).

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

that mimics the sawtooth pattern (a typical characteristic of the f waves) is used. This model is given by

$$s(n) = -\sum_{p=1}^{P} a_p(n) \sin\left(p\,\theta(n)\right) \tag{12}$$

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.

1) One AA source: We first consider a scenario with one AA source $s(n)$, which is generated using (12) with the parameters of Model 1 given in Table III. This signal is shown in Fig. 2(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 [3]. This ECG, that belongs to the database of [30], 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

$$\mathbf{Y} = \mathbf{V} + \alpha\, \mathbf{x}\mathbf{s}^{\mathsf{T}} + \mathbf{N} \in \mathbb{R}^{12 \times N}, \tag{13}$$

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. 2(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 $[8 \times 10^{-4}, 0.33 \times 10^{-2}]$ and keeping the last solution. For γ_0, we start the algorithm with $R = 6$ random blocks of rank $L = 40$. Among the estimated sources, the AA source is chosen as that which maximizes (in absolute value) the correlation coefficient ρ with respect to the ground truth $s(n)$.

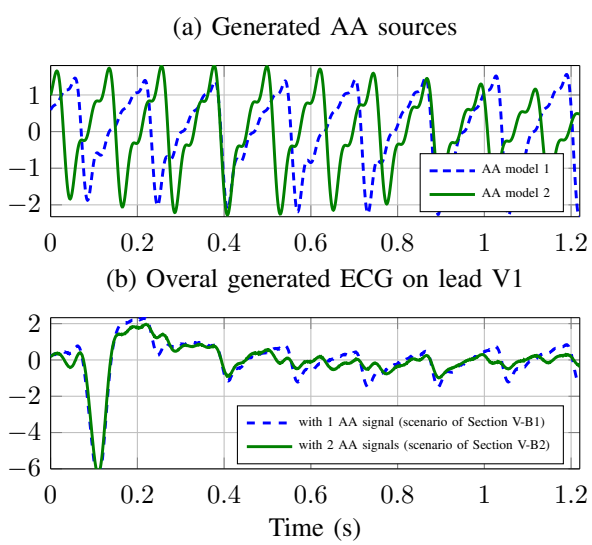

Fig. 2: Examples of generated semi-synthetic models: (a) AA sources following model (12) with the parameters shown in Table III; (b) overall synthesized ECG signals on lead V1.

BTD-NLS 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. By contrast, the results obtained with BTD-NLS 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. 3(a). For AGL, the rank chosen for the AA signal block is almost always 10 as seen in Fig. 3(a).

Though the most adequate choice for BTD-NLS 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. 4(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-NLS 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-NLS achieves a poor ρ for every R. By contrast, ρ is very likely to be quite close to 1 for CAGL.

In conclusion, BTD-NLS only produces good results with a proper combination of R, L and the 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. 2(a).

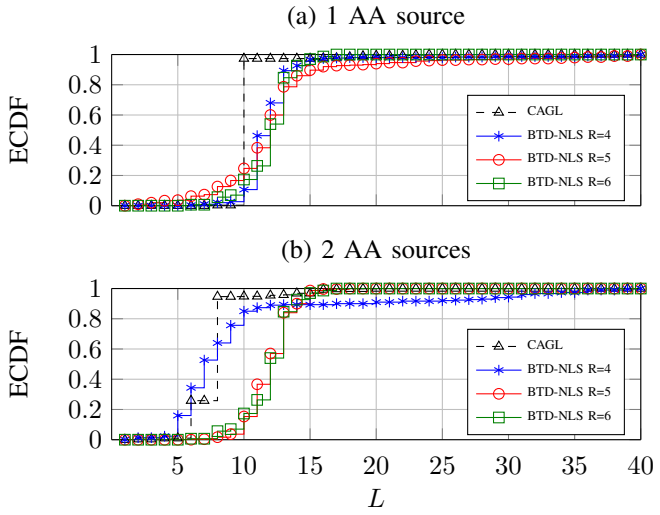

Fig. 3: Empirical distribution of rank chosen by CAGL for the AA source and of rank L yielding the best AA extraction for BTD-NLS with different numbers of blocks R.

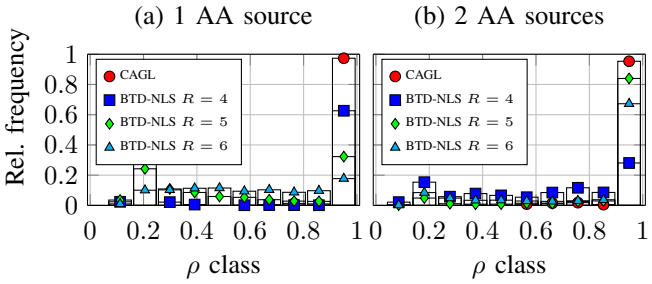

Fig. 4: Histogram of computed correlation coefficient ρ between true and estimated AA sources.

Accordingly, the model now reads

$$\mathbf{Y} = \mathbf{V} + \mathbf{X}\mathbf{S}^{\mathsf{T}} + \mathbf{N} \in \mathbb{R}^{12\times N}, \tag{14}$$

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. 2(b) displays one example of the overall generated ECG signal (solid curve).

CAGL and BTL-NLS 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. 3(b) and Fig. 4(b). The conclusions are similar to the previous scenario, with three main differences: (i) BTD-NLS now performs best with $R = 5$, as expected; (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 with exactly the same procedure as in the previous scenario. By contrast, BTD-NLS 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 two patients suffering from persistent AF. These recordings belong to a database provided by the Cardiology Department of Princess Grace Hospital 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 are about 1.17 and 1.40 seconds, for Patient 1 and 2, respectively.

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 [31]:

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

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 [31]. 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 [32], 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 [5].

Among six blocks estimated by CAGL for Patient 1, one (having rank $L_r = 8$) is identified as a potential AA source, with DF = 6.44 Hz, SC = 77.07% and $\hat{\kappa} = 155.65$. The other five estimated blocks have ranks 4, 9, 10, 18 and 21. The corresponding AA estimate on lead V1 is shown by Fig. 5(a), while Fig. 5(b) shows the overall estimated signal on that lead along with the measured signal. It also shows the estimated

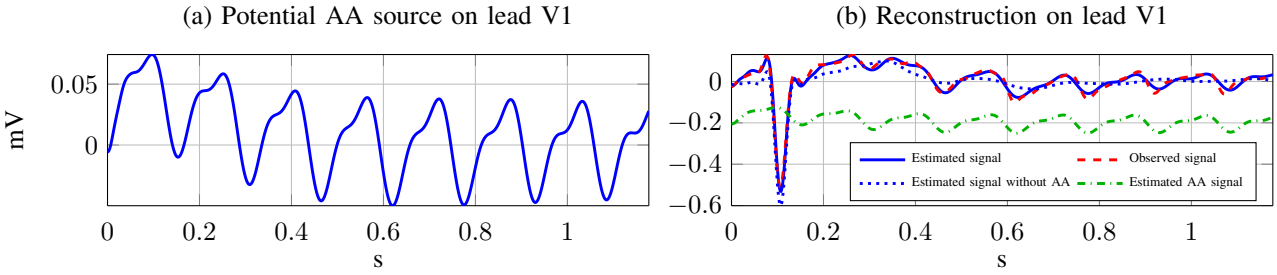

Fig. 5: Results produced by CAGL with real-world ECG data from Patient 1. The AA signal in (b) is vertically shifted by -0.2 mV for ease of visualization.

AA signal (which is vertically displaced by -0.2 mV for an easier visualization) and the estimated overall lead output after exclusion of the AA signal. From the observed AA pattern in Fig. 5(a), the high value of SC and the relatively high value of $\hat{\kappa}$, it is clear that an effective AA extraction is performed.

For Patient 2, two blocks, having ranks 36 and 29, were identified as potential AA sources: they have typical f-wave features, as can be seen in Fig. 6(a), and we have measured DF = 6.2 Hz, SC = 57.68% and $\hat{\kappa} = 107.62$ for AA source 1, and DF = 6.2 Hz, SC = 81.74% and $\hat{\kappa} = 166.50$ for AA source 2. The other sources have ranks 37 and 39. SC and $\hat{\kappa}$ values are not so high due to the residual of the T-wave (ventricular repolarization), which can be seen around 0.3 seconds of Fig. 6(a). The estimated overall signal (with and without the AA contribution) on lead V1 is displayed in Fig. 6(b), along with the estimated AA signal on that lead (which is the sum of the two signals shown in Fig. 6(a)) and the overall observed signal. Again, the AA signal is vertically shifted by -0.2 mV for clarity. It is seen that an effective extraction is achieved. Figs. 6(c)–(e) show the potential AA source contributions on leads III, V2 and V3 as well. It can be seen 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 lead III. By contrast, AA source 1 gives significant contributions to all plotted 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 (normalized) columns of $\mathbf{X}$ is around -0.19, 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. 6(a), further supports the hypothesis that the associated activities may arise from different areas of atrial tissue. 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 during catheter ablation interventions.

A final important observation is that in both examples the sum of the block ranks exceeds the dimension of the Hankel matrices: for Patient 1, $\sum_r L_r = 70 > M = 63$; for Patient 2, $\sum_r L_r = 141 > M = 74$. This showcases the benefit of using a tensor method, since a matrix technique could not possibly identify all the poles constituting each model.

VI. Conclusion

We have proposed a convergent 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. 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 standard least-squares implementation of BTD.

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 one of the 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 properly choose 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 more statistically significant results.

Appendix A
Convergence of (unconstrained) AGL

AGL's formulation satisfies the conditions of Theorem 2 of [24] for the following reasons:

- since the directional derivative of g (see (8)) exists at every point $(\mathbf{A}, \mathbf{B}, \mathbf{X}) \in \mathcal{S}$ and AGL's subproblems satisfy

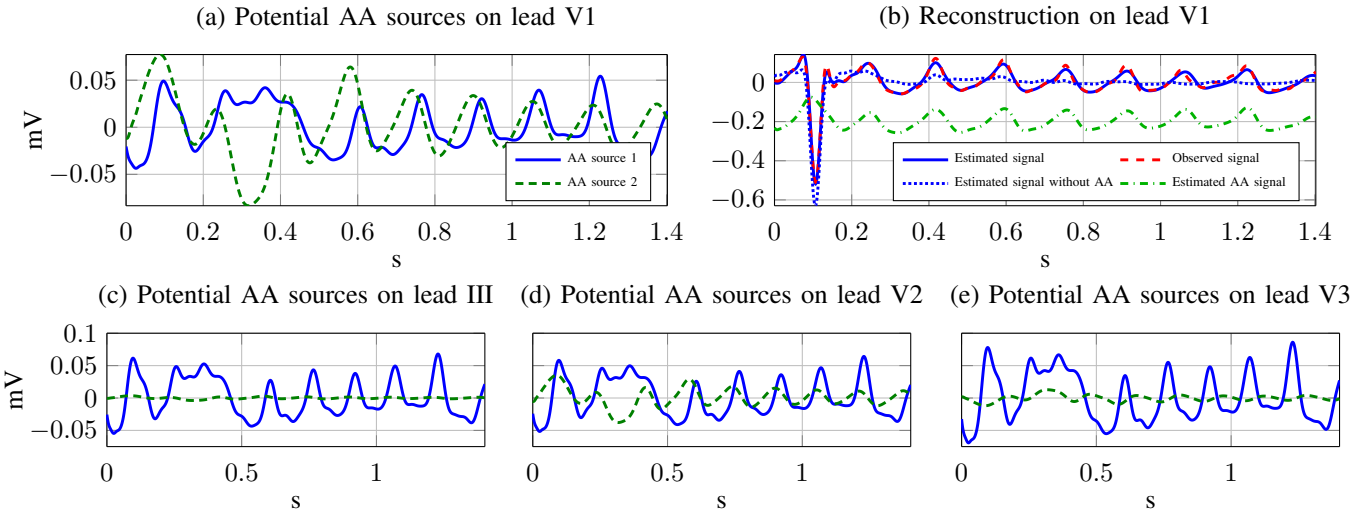

Fig. 6: Results produced by CAGL with real-world ECG data from Patient 2. The legend in (a) applies also to (c), (d) and (e). The AA signal in (b) is vertically shifted by -0.2 mV for ease of visualization.

$F(\mathbf{A} \,|\, \hat{\mathbf{B}}^{(t)}, \hat{\mathbf{X}}^{(t)}) + \frac{\tau}{2}\|\mathbf{A} - \mathbf{A}^{(t)}\|_F^2 \geq F(\mathbf{A} \,|\, \hat{\mathbf{B}}^{(t)}, \hat{\mathbf{X}}^{(t)})$ with equality at $\mathbf{A} = \mathbf{A}^{(t)}$ and likewise for $\mathbf{B}$ and $\mathbf{X}$, Assumption 2 of [24] holds (see [24, Proposition 2]);

- $F(\mathbf{A}, \mathbf{B}, \mathbf{X})$ is continuous and coercive, and thus its sublevel sets $\mathcal{L}_c = \{(\mathbf{A}, \mathbf{B}, \mathbf{X}) \in \mathcal{S} \,:\, F(\mathbf{A}, \mathbf{B}, \mathbf{X}) \leq c\}$, with $c \in \mathbb{R}$, are compact;
- AGL's subproblems are strictly convex, thus having unique minimizers;
- as $f(\mathbf{A}, \mathbf{B}, \mathbf{X}) < \infty$ and f is Gâteaux-differentiable on all $(\mathbf{A}, \mathbf{B}, \mathbf{X}) \in \mathcal{S}$, [33, Lemma 3.1] implies F is regular at every point $(\mathbf{A}, \mathbf{B}, \mathbf{X}) \in \mathcal{S}$.

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. V. Zarzoso is a member of the *Institut Universitaire de France*.

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, G. Favier, J. C. M. Mota, A. L. De Almeida, and V. Zarzoso, "A tensor decomposition approach to noninvasive atrial activity extraction in atrial fibrillation ECG," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*. Nice, France: IEEE, 2015, pp. 2576–2580.

[3] 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. (EBMC)*. Milan, Italy: IEEE, aug 2015, pp. 6987–6990.

[4] 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.

[5] 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.

[6] ——, "Block term decomposition analysis in long segments of atrial fibrillation ECGs," in *Proc. XXXVI Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT-2018)*, Campina Grande, Brazil, sep 2018.

[7] 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.

[8] 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.

[9] ——, "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.

[10] 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.

[11] 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.

[12] 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, Jan. 2019.

[13] 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.

[14] J. A. Cadzow, "Signal enhancement—a composite property mapping algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 1, pp. 49–62, 1988.

[15] 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, jul 2004.

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

[17] 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.

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

[19] L. L. Scharf and C. Demeure, *Statistical signal processing: detection,*

estimation, and time series analysis. Addison-Wesley Reading, MA, 1991, vol. 63.

[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] M. T. Chu, R. E. Funderlic, and R. J. Plemmons, "Structured low rank approximation," *Linear algebra and its applications*, vol. 366, pp. 157–172, 2003.

[22] 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.

[23] J. Liu and J. Ye, "Moreau-yosida regularization for grouped tree structure learning," in *Proc. Adv. Neural Inf. Process. Systems*, 2010, pp. 1459–1467.

[24] 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.

[25] 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.

[26] 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.

[27] I. Markovsky and K. Usevich, "Software for weighted structured low-rank approximation," *J. Comput. Appl. Math.*, vol. 256, pp. 278–292, 2014.

[28] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[29] 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.

[30] 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.

[31] 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, Feb. 2005.

[32] 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.

[33] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, 2001.