An Effective Iterative Solution for Independent Vector Analysis with Convergence Guarantees

Clément Cosserat, Ben Gabrielson, Emilie Chouzenoux, Jean-Christophe Pesquet, Jean-Christophe Pesquet, Tülay Adali

An Effective Iterative Solution for Independent Vector Analysis with Convergence Guarantees

Clément Cosserat, *Student Member* IEEE, Ben Gabrielson, *Student Member*, IEEE, Emilie Chouzenoux, *Senior Member*, IEEE, Jean-Christophe Pesquet, *Fellow*, IEEE, and Tülay Adali, *Fellow* IEEE

***Abstract*—Independent vector analysis (IVA) is an attractive solution to address the problem of joint blind source separation (JBSS), that is, the simultaneous extraction of latent sources from several datasets implicitly sharing some information. Among IVA approaches, we focus here on the celebrated IVA-G model, that describes observed data through the mixing of independent Gaussian source vectors across the datasets. IVA-G algorithms usually seek the values of demixing matrices that maximize the joint likelihood of the datasets, estimating the sources using these demixing matrices. Instead, we write the likelihood of the data with respect to both the demixing matrices and the precision matrices of the source estimate. This allows us to formulate a cost function whose mathematical properties enable the use of a proximal alternating algorithm based on closed form operators with provable convergence to a critical point. After establishing the convergence properties of the new algorithm, we illustrate its desirable performance in separating sources with covariance structures that represent varying degrees of difficulty for JBSS.**

***Index Terms*—IVA, PALM Algorithm, Maximum-Likelihood, Blind Source Separation, Proximal methods**

I. Introduction

THE *blind source separation* problem (BSS) aims at factorizing a data matrix as a product of a mixing matrix and a source data matrix. BSS is thus a data-driven method to extract latent features from a dataset, which have a broad variety of uses. It offers a wide range of applications in signal processing and engineering including neuroimaging data analysis [1], communications [2], [3], remote sensing [4], to name a few [5]. The latent sources are interpreted as physical quantities of interest that cannot be measured directly, e.g. brain connectivity networks [1], [6], [7], or can be used as features for further tasks such as classification [8]. The BSS problem generalizes to the *joint blind source separation* problem (JBSS) when multiple datasets are analyzed jointly to benefit from their shared information. JBSS is necessary to fully analyze datasets that share similarities. This case presents itself when the datasets contain measures of a same phenomenon for different subjects [6], [9], different measurement methods [10], or more generally for various modalities [11]. Allowing the interaction between the datasets leads JBSS to achieve more accurate separation than multiple separate BSS in general [6], [12].

One way to address BSS is to model the rows of the source dataset as samples of mutually independent random variables or random process called sources and the datasets as mixtures of independent sources [13]. This approach is known as *independent component analysis* (ICA) [14], which has been one of the most popular ways to achieve ICA due to its uniqueness guarantees under very general conditions. ICA can be generalized to *independent vector analysis* (IVA) [6], [12], [15], [16], [17] to address the JBSS problem. In this case, sources are considered as random vectors (or random process vectors), called *source component vectors* (SCVs). Each entry of a SCV accounts for one source in a given dataset. In IVA, now the SCVs are assumed to be independent rather than univariate sources as in ICA. Within each SCV, the statistical dependence (correlation for IVA-G) across the datasets are taken into account. IVA methods gather a family of algorithms that aim at recovering the SCVs and mixing coefficients that generated the observed data. The JBSS problem can be addressed with many other methods such as groupICA [7], [16], or joint ICA [18], but IVA is more powerful as it offers a greater flexibility and helps preserve the individual variability represented by each dataset, e.g., in multisubject analyses [19], [20]. Moreover, it enables a common factorization of the datasets without needing to realign the sources $a\ posteriori$, which can be costly, thus alleviating the inherent permutation ambiguity across the datasets.

IVA is typically formulated using a *maximum of likelihood* (ML) estimator, which means that the estimation is done by solving an optimization problem where the cost-function is derived from the log-likelihood of the data. In [12], we see IVA-G presented as a minimization of the mutual information of the SCVs to maximize their independence, those two formulations are equivalent when the number of samples tends to infinity [6].

In this paper, we focus on the case of non-degenerate, centered Gaussian SCVs, placing ourselves in the so-called IVA-G framework, as presented in [12]. This model is convenient, as the SCVs are fully described by their covariance matrix (or equivalently by their precision matrix). Besides, since IVA-G algorithms only make use of *second order statistics* (SOS) of the data, this makes them the simplest and most efficient among the IVA algorithms. In the Gaussian case, having the demixing matrices gives a natural estimate of the covariance matrices of the SCVs through their empirical covariance, this is why the cost function in [12] only depends on the demixing matrices. In this work however, we choose to write the ML such that it explicitly takes the SCVs *probability density function* (pdf) as an input variable, to benefit from analytical properties of the resulting cost.

So far, the IVA-G algorithms are based on standard optimization methods, like Newton's method or gradient descent, without demonstration of explicit convergence guarantees. In this article, we show that our proposed cost function to jointly search for the Gaussian SCVs and demixing matrices generalizes the cost function in [12], in the sense that replacing the precision matrices of the SCVs by the inverse of their empirical covariance matrices for the given demixing matrices, we recover a cost function that only depends on those demixing matrices and that is equal to the one in [12] up to a constant. The choice of a cost function that jointly

acts on the sought demixing and covariance matrices offers an attractive structure enabling the explicit incorporation of priors and constraints, and the use of mathematically sound minimization algorithms. Hence, we design a *Proximal Alternating Linearized Minimization* algorithm (**PALM**), dedicated to IVA-G problem, and show that it converges to a critical point under some mild assumptions. The resulting scheme, denoted **PALM-IVA-G**, is also shown to be fast, and to establish a reliable estimation performance in practice, and at least as good as the state-of-the-art methods for IVA-G.

Our contributions can be summarized as follows. First, we provide a novel variational formulation for IVA-G, introducing a new cost function and establishing its mathematical properties. Second, building upon these properties, we design the **PALM-IVA-G** algorithm, to solve the resulting minimization problem, and show its convergence under mild assumptions. Third, we numerically illustrate the desirable performance of our approach, in comparison with state-of-the-art approaches, on a number of scenarios that cover various degrees of dependence across the datasets.

The paper is organized as follows. Section II introduces the JBSS problem, the IVA-G framework to address it, and presents our cost function based on maximum likelihood estimation. Section III then focuses on our proposition of an original iterative algorithm to solve the IVA-G problem. Convergence results are presented in Section IV. Numerical experiments assessing the validity and good performance of our approach are presented in Section V. Section VI concludes the paper.

II. The IVA-G problem

A. Notation

Throughout the paper, we use bold upper case symbols for matrices, bold lower case symbols for column or row vectors, calligraphic upper case symbol for tensors of order three or more, and regular lower case symbols for scalars. Italic font is used for deterministic quantities while regular one is used for random quantities. For instance, let $\mathbf{u} = [\mathrm{u}_1, \dots, \mathrm{u}_N]^\top$ be an $\mathbb{R}^N$-valued random vector, for which we draw V realisations that we stack, columnwise, in a matrix $\boldsymbol{U} \in \mathbb{R}^{N\times V}$. The latter can also be written as $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_N]^\top$ with, for every $n \in \{1, \dots, N\}$, $\boldsymbol{u}_n \in \mathbb{R}^V$ a column vector of V realizations of the scalar random variable u_n.

The subset of non-singular (resp. symmetric) matrices of $\mathbb{R}^{N\times N}$ is denoted $\mathcal{GL}_N(\mathbb{R})$ (resp. $\mathcal{S}_N$). $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. We also denote $\mathcal{S}_N^+$ the set of positive semi-definite matrices of size N, and $\mathcal{S}_N^{++} = \mathcal{S}_N^+ \bigcap \mathcal{GL}_N(\mathbb{R})$ the set of *positive definite* (PD) matrices. $\|\cdot\|$ denotes the Frobenius norm for any vector, matrix, and tensor of order three or more.

For any matrix $\boldsymbol{A}$, we denote by $\|\boldsymbol{A}\|_{\mathrm{S}}$ its spectral norm, equal to the largest singular value of $\boldsymbol{A}$. For every $n \in \{1, \dots, N\}$, $\boldsymbol{a}_n^\top$ denotes the n-th row of $\boldsymbol{A}$. If $\boldsymbol{A}$ is a square matrix, we note $\boldsymbol{\sigma}_{\boldsymbol{A}} = (\sigma_{\boldsymbol{A},l})_{1\le l\le N} \in \mathbb{R}^N$ the vector of its singular values, $\mathrm{tr}(\boldsymbol{A})$ its trace, and $\det(\boldsymbol{A})$ its determinant. $\mathrm{diag}(\boldsymbol{A}) \in \mathbb{R}^N$ is the vector whose entries are the diagonal coefficients of $\boldsymbol{A}$, whereas for $\boldsymbol{a} \in \mathbb{R}^N$, $\mathrm{Diag}(\boldsymbol{a})$ (with an upper case) is the diagonal matrix whose coefficients are the components of $\boldsymbol{a}$. We also note $\mathrm{Diag}(\boldsymbol{A})$ the diagonal matrix whose diagonal coefficients are the same as those of $\boldsymbol{A}$. Finally, for any $(N, M) \in (\mathbb{N}\backslash\{0\})^2$, we consider the matricial scalar product defined as $(\forall(\boldsymbol{A}, \boldsymbol{B}) \in (\mathbb{R}^{N\times M})^2)\ \langle \boldsymbol{A} \mid \boldsymbol{B}\rangle = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B})$.

B. JBSS problem

Let K, N, V be positive integers, we consider K datasets denoted $(\boldsymbol{X}^{[k]})_{1\le k\le K}$ where $\forall k \in \{1, \dots, K\}, \boldsymbol{X}^{[k]} \in \mathbb{R}^{N\times V}$ is a real-valued matrix. For instance, in an fMRI analysis problem, for $n \in \{1, \dots, M\}$, and $k \in \{1, \dots, K\}$, the n-th row of $\boldsymbol{X}^{[k]}$, $(\boldsymbol{x}_n^{[k]})^\top$ could model the blood-oxygen-level-dependent (BOLD) contrast in the V voxels of a 3D model of the brain, measured at acquisition time n for the k-th subject, within a cohort of K patients [1], [7]. We assume that the observed datasets are obtained by a linear mixing of latent source datasets, i.e, $(\forall k \in \{1, \dots, K\})$:

$$\boldsymbol{X}^{[k]} = \boldsymbol{A}^{[k]}\boldsymbol{S}^{[k]} \in \mathbb{R}^{N\times V}, \tag{1}$$

where $\boldsymbol{A}^{[k]} = \left(a_{m,n}^{[k]}\right)_{1\le m\le N, 1\le n\le N} \in \mathbb{R}^{N\times N}$ is a square non-singular mixing matrix and $\boldsymbol{S}^{[k]} \in \mathbb{R}^{N\times V}$ is a latent matrix whose coefficients are typically interpreted as hidden features of the phenomenon described by $\boldsymbol{X}^{[k]}$. The JBSS problem consists in estimating simultaneously for all $k \in \{1, \dots, K\}$, both $\boldsymbol{A}^{[k]}$ and $\boldsymbol{S}^{[k]}$ from $\boldsymbol{X}^{[k]}$. We estimate the inverse of the mixing matrices by the so-called demixing matrices $(\boldsymbol{W}^{[k]})_{1\le k\le K}$, and deduce estimates $(\boldsymbol{Y}^{[k]})_{1\le k\le K}$ for the source datasets $(\boldsymbol{S}^{[k]})_{1\le k\le K}$ by calculating

$$(\forall k \in \{1, \dots, K\}) \quad \boldsymbol{Y}^{[k]} = \boldsymbol{W}^{[k]}\boldsymbol{X}^{[k]}. \tag{2}$$

In a nutshell, using tensor notations, JBSS amounts to providing an estimate $\mathcal{Y} = [\boldsymbol{Y}^{[1]}, \dots, \boldsymbol{Y}^{[K]}]$ of the source tensor $\mathcal{S} = [\boldsymbol{S}^{[1]}, \dots, \boldsymbol{S}^{[K]}]$ via the demixing tensor $\mathcal{W} = [\boldsymbol{W}^{[1]}, \dots, \boldsymbol{W}^{[K]}]$, which is an estimate of the slice-wise inverse of the mixing tensor $\mathcal{A} = [\boldsymbol{A}^{[1]}, \dots, \boldsymbol{A}^{[K]}]$, given the datasets $\mathcal{X} = [\boldsymbol{X}^{[1]}, \dots, \boldsymbol{X}^{[K]}]$. Here, $\mathcal{S}, \mathcal{X}$ and $\mathcal{Y} \in \mathbb{R}^{N\times V\times K}$ and $\mathcal{A}$ and $\mathcal{W} \in \mathbb{R}^{N\times N\times K}$.

C. IVA model

Assuming that the datasets share some information which can be leveraged to separate the sources more accurately than if dealt with individually, IVA [17] models this interdependence of the $\boldsymbol{X}^{[k]}$ through statistical links between the latent datasets. More precisely, the rows of the latent datasets with the same index are assumed to be correlated, while being independent from rows with different indices. To formalize this, IVA models the columns of the $\boldsymbol{X}^{[k]}$ (resp. $\boldsymbol{S}^{[k]}$) as independent samples from $\mathbb{R}^N$-valued random vectors $\mathbf{x}^{[k]} = [\mathrm{x}_1^{[k]}, \dots, \mathrm{x}_N^{[k]}]^\top$ (resp. $\mathbf{s}^{[k]} = [\mathrm{s}_1^{[k]}, \dots, \mathrm{s}_N^{[k]}]^\top$). We can thus rewrite the model in (1) using random vector notations: $(\forall k \in \{1, \dots, K\}), \mathbf{x}^{[k]} = \boldsymbol{A}^{[k]}\mathbf{s}^{[k]}$. Regrouping the components of the $\mathbf{s}^{[k]}$ with corresponding indices, we obtain N all $\mathbb{R}^K$-valued random vectors, $\mathbf{s}_n = [\mathrm{s}_n^{[1]}, \dots, \mathrm{s}_n^{[K]}]^\top$ for $n \in \{1, \dots, N\}$ called source component vectors, where each entry of a SCV accounts for the corresponding dataset. The goal of IVA is to make the SCVs as independent as possible.

IVA aims at recovering the sources by building demixing matrices $(\boldsymbol{W}^{[k]})_{1\le k\le K}$. At the same times, it builds estimated SCVs $(\mathbf{y}_n)_{1\le n\le N}$ whose distributions have probability density functions respectively denoted $(p_{\mathbf{y}_n})_{1\le n\le N}$ and that we suppose mutually independent. Similarly as for SCVs,

we denote $(\forall n \in \{1,\dots,N\}), \mathbf{y}_n = [\mathrm{y}_n^{[1]},\dots,\mathrm{y}_n^{[K]}]^\top$ the estimated SCVs, and we reorganize the components to define $(\forall k \in \{1,\dots,K\}), \mathbf{y}^{[k]} = [\mathrm{y}_1^{[k]},\dots,\mathrm{y}_N^{[k]}]^\top$. Then, we see with (2) that for all $k \in \{1,\dots,K\}$, $\mathbf{x}^{[k]}$ is estimated by $(\boldsymbol{W}^{[k]})^{-1}\mathbf{y}^{[k]}$, whose probability distribution is entirely determined by the demixing matrices and the $(p_{\mathbf{y}_n})_{1\le n\le N}$. Said otherwise, $\mathcal{W}$ and $(p_{\mathbf{y}_n})_{1\le n\le N}$ yield an estimated generative model for our observed datasets.

D. IVA-G cost function

In the following, for every $n \in \{1,\dots,N\}$, we denote by $\boldsymbol{Y}_n \in \mathbb{R}^{K\times V}$ the matrix obtained by stacking vertically the n-th rows of $\boldsymbol{Y}^{[k]}$ for $k \in \{1,\dots,K\}$. Therefore, using (2), we have

$$(\forall n \in \{1,\dots,N\}) \quad \boldsymbol{Y}_n = \boldsymbol{W}_n \boldsymbol{X}, \tag{3}$$

where

$$\boldsymbol{W}_n = \begin{pmatrix} \boldsymbol{w}_n^{[1]\top} & 0 & \dots & 0 \\ 0 & \boldsymbol{w}_n^{[2]\top} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{w}_n^{[K]\top} \end{pmatrix} \in \mathbb{R}^{K\times KN} \tag{4}$$

and

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}^{[1]} \\ \vdots \\ \boldsymbol{X}^{[K]} \end{pmatrix} \in \mathbb{R}^{KN\times V}. \tag{5}$$

The objective of IVA is to determine $\mathcal{W}$ and $(p_{\mathbf{y}_n})_{1\le n\le N}$ that maximize the log-likelihood of $\boldsymbol{X}$ in our estimated generative model, given by [15]

$$\sum_{n=1}^{N} \log p_{\mathbf{y}_n}(\boldsymbol{Y}_n) + V \sum_{k=1}^{K} \log|\det \boldsymbol{W}^{[k]}|.$$

In the Gaussian case, considered here, we model $(\mathbf{y}_n)_{1\le n\le N}$ as centered non-degenerate Gaussian vectors, whose pdf is thus entirely determined by their covariance matrices, or equivalently, their precision matrices that we denote $(\boldsymbol{C}_n)_{1\le n\le N}$. As done previously, to simplify our notation, we introduce the tensor $\mathcal{C} = [\boldsymbol{C}_1,\dots,\boldsymbol{C}_N] \in (\mathcal{S}_K^{++})^N$ that gathers the estimated precision matrices of all the SCVs. Moreover, we assume sample independence, so the pdf of $\boldsymbol{Y}_n$ is the product of the pdfs of its columns. Under this Gaussian modeling, we have,

$$(\forall \boldsymbol{y} \in \mathbb{R}^K)$$
$$\log p_{\mathbf{y}_n}(\boldsymbol{y}) = \frac{1}{2}\log\det \boldsymbol{C}_n - \frac{K}{2}\log 2\pi - \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{C}_n \boldsymbol{y}. \tag{6}$$

Hence, the problem becomes equivalent to minimize, with respect to $\mathcal{W},\mathcal{C}$, the cost function $J_{\mathrm{IVA-G}}(\mathcal{W},\mathcal{C})$ defined as

$$\begin{aligned} &J_{\mathrm{IVA-G}}(\mathcal{W},\mathcal{C}) + \frac{1}{2}\sum_{n=1}^{N}\log\det \boldsymbol{C}_n + \sum_{k=1}^{K}\log|\det \boldsymbol{W}^{[k]}| \\ &= \frac{1}{2}\sum_{n=1}^{N}\frac{1}{V}\operatorname{tr}(\boldsymbol{C}_n \boldsymbol{Y}_n \boldsymbol{Y}_n^\top) \\ &= \frac{1}{2}\sum_{n=1}^{N}\operatorname{tr}(\boldsymbol{C}_n \boldsymbol{W}_n \widehat{\boldsymbol{R}}_{\mathbf{x}} \boldsymbol{W}_n^\top), \end{aligned} \tag{7}$$

where $\widehat{\boldsymbol{R}}_{\mathbf{x}} = \frac{1}{V}\boldsymbol{X}\boldsymbol{X}^\top$ is the empirical covariance matrix of $\hat{\mathbf{x}}$.

Note that the domain of function (7) can be extended to $\mathbb{R}^{N\times N\times K} \times \mathbb{R}^{K\times K\times N}$ by setting $J_{\mathrm{IVA-G}}(\mathcal{W},\mathcal{C}) = +\infty$ if $\boldsymbol{W}^{[k]}$ is singular for some $k \in \{1,\dots,K\}$ or if $\boldsymbol{C}_n$ is not symmetric positive definite for some $n \in \{1,\dots,N\}$.

Remark 1 *As we will show in Section IV, the non-singularity of $\widehat{\boldsymbol{R}}_{\mathbf{x}}$ is a sufficient condition to the lower-boundedness of (7), and as such, to the well-posedness of our optimization problem. In practice, for a large number of samples — that is $V > KN$ — drawn from a continuous probability distribution, the empirical covariance matrix $\widehat{\boldsymbol{R}}_{\mathbf{x}}$ is non-singular almost surely. Hence, we will suppose that this condition holds in the remainder of this article.*

E. IVA-G-V and IVA-G-N approaches

In [12], the authors proposed two algorithms, called **IVA-G-V** and **IVA-G-N**, to solve the IVA-G estimation problem. To do so, they reformulate the problem as the minimization of the following cost function [12]:

$$\widetilde{J}_{\mathrm{IVA-G}}(\mathcal{W}) = \frac{1}{2}\sum_{n=1}^{N}\log\det(\boldsymbol{W}_n \widehat{\boldsymbol{R}}_{\mathbf{x}} \boldsymbol{W}_n^\top) - \sum_{k=1}^{K}\log|\det \boldsymbol{W}^{[k]}|. \tag{8}$$

These approaches implicitly estimate the covariance matrices of the sources by $\boldsymbol{C}_n^{-1} = \boldsymbol{W}_n \widehat{\boldsymbol{R}}_{\mathbf{x}} \boldsymbol{W}_n^\top$, that is, the empirical covariance matrices of the $(\mathbf{y}_n)_{1\le n\le N}$. The minimization of (8) is usually performed using either a gradient-based, or a Newton-based solver, leading to **IVA-G-V** and **IVA-G-N** schemes, respectively. To the best of our knowledge, these algorithms do not benefit from strong guarantees in terms of convergence of the iterates.

It is easy to show that finding $(\mathcal{W},\mathcal{C})$ that minimizes the proposed objective function (7) is actually equivalent to finding $\mathcal{W}$ that minimizes the cost (8) used in state-of-the-art IVA-G approaches, as stated in the following known result.

Theorem 1 (Equivalence between $J_{\mathrm{IVA-G}}$ and $\widetilde{J}_{\mathrm{IVA-G}}$)
For a given $\mathcal{W} \in \mathbb{R}^{N\times N\times K}$, if all the $\boldsymbol{W}^{[k]}$ are non-singular, then

(i) *$J_{\mathrm{IVA-G}}(\mathcal{W},.)$ is minimized over $\mathbb{R}^{K\times K\times N}$ at $\widehat{\mathcal{C}}(\mathcal{W}) = [\widehat{\boldsymbol{C}}_1(\boldsymbol{W}_1),\dots,\widehat{\boldsymbol{C}}_N(\boldsymbol{W}_N)]$ where*

$$(\forall n \in \{1,\dots,N\})\ \widehat{\boldsymbol{C}}_n(\boldsymbol{W}_n) = (\boldsymbol{W}_n \widehat{\boldsymbol{R}}_{\mathbf{x}} \boldsymbol{W}_n^\top)^{-1}. \tag{9}$$

(ii) *We have, for every $\mathcal{W} \in \mathbb{R}^{N\times N\times K}$:*

$$\min_{\mathcal{C}\in(\mathcal{S}_K^{++})^N} J_{\mathrm{IVA-G}}(\mathcal{W},\mathcal{C}) = \widetilde{J}_{\mathrm{IVA-G}}(\mathcal{W}) + \frac{KN}{2}. \tag{10}$$

Theorem 1 shows that $\widehat{\mathcal{W}}$ is a minimizer for $\widetilde{J}_{\mathrm{IVA-G}}$ if and only if $(\widehat{\mathcal{W}},\widehat{\mathcal{C}}(\widehat{\mathcal{W}}))$ is a minimizer for $J_{\mathrm{IVA-G}}$, hence the equivalence. The advantage of relying on the proposed cost function that takes $\mathcal{C}$ as an input block variable is that it offers a more structured form, allowing an efficient use of a block alternating minimization scheme, and benefiting from sounded convergence, as we will show in Section III.

F. Ambiguities in IVA-G model

There are two ambiguities in the IVA-G generative model, that correspond to information on the parameters that cannot be deduced from the observed data. The first one is the *permutation ambiguity*. New mixing matrices can be obtained by renumbering of the SCVs, that is, by defining $\widetilde{\boldsymbol{A}}^{[k]} = \boldsymbol{A}^{[k]}\boldsymbol{P}$ with $\boldsymbol{P}$ a permutation matrix. This change however does not modify the value of the cost function (7). The second one is the *scaling ambiguity*. For every $k \in \{1,\dots,K\}$ and $n \in \{1,\dots,N\}$, and for any $\alpha_n^{[k]} \in \mathbb{R}\setminus\{0\}$, we can replace $\mathbf{s}_n^{[k]}$ with $\widehat{\mathbf{s}}_n^{[k]} = \alpha_n^{[k]}\mathbf{s}_n^{[k]}$ and $a_{m,n}^{[k]}$ with $\hat{a}_{m,n}^{[k]} = (\alpha_n^{[k]})^{-1}a_{m,n}^{[k]}$ for every $m \in \{1,\dots,N\}$. Those transformations let the random vectors $\mathbf{x}_n^{[k]}$ unchanged, and consequently, they do not affect the likelihood expression. Hence, the demixing matrices can only estimate the inverse of the ground truth mixing matrices, up to the permutation and scaling ambiguity. The latter ambiguity moreover raises the problem that a minimizing sequence of $J_{\mathrm{IVA-G}}$ is not necessarily bounded. This motivates our proposition for a regularized version for the cost.

G. Proposed regularized cost function

Let us note, for all $\mathcal{C} \in (\mathcal{S}_K^{++})^N$ and $n \in \{1,\dots,N\}$, $(c_{n,k,k})_{1\le k\le K} = \operatorname{diag}(\boldsymbol{C}_n)$. Due to the scaling ambiguity raised above, for any $(\mathcal{W},\mathcal{C})$, it is possible to rescale the coefficients to define $(\mathcal{W}',\mathcal{C}')$ such that $J_{\mathrm{IVA-G}}(\mathcal{W}',\mathcal{C}') = J_{\mathrm{IVA-G}}(\mathcal{W},\mathcal{C})$ and that

$$(\forall n \in \{1,\dots,N\})(\forall k \in \{1,\dots,K\}) \quad c'_{n,k,k} = 1. \tag{11}$$

To do this, one can set, for each (k,n), $\alpha_n^{[k]} = \frac{1}{\sqrt{c_{n,k,k}}}$. As a consequence, if a minimizer of $J_{\mathrm{IVA-G}}$ exists, then there exists at least another minimizer, satisfying $c_{n,k,k} = 1$ for all (k,n). To mitigate this ambiguity, we thus propose to add a quadratic penalty to the cost function to control the distance to 1 of the diagonal coefficients of the precision matrices.[1]

Remark 2 *Let us notice that a minimizer of $J_{\mathrm{IVA-G}}$, such that $c_{n,k,k} = 1$ has no reason to be qualitatively better than any other minimizer of $J_{\mathrm{IVA-G}}$. The proposed regularization aims at reducing the number of distinct minima, and at ensuring some mathematical properties of the cost we will leverage to prove the convergence of our optimization algorithms. It is still possible, once convergence is reached, to rescale the sources a posteriori.*

In addition to the scaling penalty term, we also propose to add an extra term to the cost function, to constrain the singular values of the recovered precision matrices to be positive by a minimum margin. This term aims at avoiding numerical issues that could arise at the boundary of the logarithm determinant definition domain, without damaging the quality of the results. In practice, the constraint is simply imposed, by adding an indicator function $\iota_{[\epsilon,+\infty)^K}$, equal to 0 for non-negative entries, $+\infty$ otherwise. Our objective is to impose that the components of the vector of singular values $\boldsymbol{\sigma}_{\boldsymbol{C}_n} \in \mathbb{R}^K$ of matrix $\boldsymbol{C}_n$ are above a certain $\epsilon > 0$, for every n. We thus obtain the final form for our proposed (regularized) cost function, denoted by $J^{\mathrm{Reg}}_{\mathrm{IVA-G}}$:

$$\begin{aligned}&(\forall(\mathcal{W},\mathcal{C}) \in (\mathcal{GL}_N(\mathbb{R})^K \times (\mathcal{S}_K^{++})^N))\\ &J^{\mathrm{Reg}}_{\mathrm{IVA-G}}(\mathcal{W},\mathcal{C}) = J_{\mathrm{IVA-G}}(\mathcal{W},\mathcal{C}) + \frac{\alpha}{2}\sum_{n=1}^{N}\|\operatorname{diag}(\boldsymbol{C}_n) - \mathbf{1}_K\|^2\\ &\qquad + \sum_{n=1}^{N}\iota_{[\epsilon,+\infty)^K}(\boldsymbol{\sigma}_{\boldsymbol{C}_n}),\end{aligned} \tag{12}$$

$\alpha > 0$ is an hyper-parameter that controls the strength of the introduced regularization.

Our IVA-G method then aims to minimize (12). This is a challenging non-convex and non-smooth problem. The next section is dedicated to discuss further the properties of (12), and to design an efficient optimization algorithm to find a critical point of it.

III. Proposed minimization algorithm

A. Mathematical properties of $J^{\mathrm{Reg}}_{\mathrm{IVA-G}}$

In order to design an appropriate minimization algorithm for (12), let us examine the structure and properties of the cost function $J^{\mathrm{Reg}}_{\mathrm{IVA-G}}$. First, let us remark that minimizing (12) on $\mathcal{GL}_N(\mathbb{R})^K \times (\mathcal{S}_K^{++})^N$ is equivalent to

$$\underset{(\mathcal{W},\mathcal{C})\in\mathbb{R}^{N\times N\times K}\times\mathbb{R}^{K\times K\times N}}{\text{minimize}}\; h(\mathcal{W},\mathcal{C}) + f(\mathcal{W}) + g(\mathcal{C}), \tag{13}$$

with

$$2h(\mathcal{W},\mathcal{C}) = \sum_{n=1}^{N}\operatorname{tr}(\boldsymbol{C}_n\boldsymbol{W}_n\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_n^\top) + \alpha\|\operatorname{diag}(\boldsymbol{C}_n) - \mathbf{1}_K\|^2, \tag{14}$$

$$f(\mathcal{W}) = \begin{cases} -\sum_{k=1}^{K}\log|\det\boldsymbol{W}^{[k]}| \\ \quad \text{if } (\forall k\in\{1,\dots,K\})\,\boldsymbol{W}^{[k]} \in \mathcal{GL}_N(\mathbb{R}) \\ +\infty \text{ otherwise,}\end{cases} \tag{15}$$

$$g(\mathcal{C}) = \begin{cases} -\frac{1}{2}\sum_{n=1}^{N}\log\det\boldsymbol{C}_n \\ \quad \text{if } (\forall n\in\{1,\dots,N\})\,\boldsymbol{C}_n - \epsilon\boldsymbol{I}_K \in \mathcal{S}_K^+ \\ +\infty \text{ otherwise.}\end{cases} \tag{16}$$

As we already mentioned, $\epsilon > 0$ serves as obtaining better regularity for the cost function (with closed definition domain). It is typically taken very small (e.g., $\epsilon = 10^{-12}$ in our experiments) to ensure the problem is essentially equivalent to a search for a maximum of the likelihood. A similar strategy was adopted in [21]. Formulation (13) has the advantage of isolating a differentiable term h acting on both set of variables $(\mathcal{W},\mathcal{C})$, and two non differentiable terms, f and g, acting separately on $\mathcal{W}$ or $\mathcal{C}$.

a) Function h: Let us first study function h acting jointly on $\mathcal{W}$ and $\mathcal{C}$ variables. We state the following lemmas, regarding the expression and smoothness properties, of its partial gradients, with respect to the first or second entry, the other being fixed.

Lemma 1 *The partial gradient of h with respect to variable $\mathcal{W}$, evaluated at $(\mathcal{W},\mathcal{C}) \in \mathbb{R}^{N\times N\times K}\times\mathbb{R}^{K\times K\times N}$, reads:*

$$\nabla_{\mathcal{W}}h(\mathcal{W},\mathcal{C}) = (\nabla_{\boldsymbol{W}^{[k]}}h(\mathcal{W},\mathcal{C}))_{1\le k\le K} \in \mathbb{R}^{N\times N\times K}, \tag{17}$$

[1] In **IVA-G-V** and **IVA-G-N**, the scale ambiguity is managed by an ad-hoc renormalizing of the rows of the demixing matrices after each iteration of the minimization solver.

where, for all $k \in \{1,\dots,K\}$,

$$\nabla_{\boldsymbol{W}^{[k]}} h(\mathcal{W},\mathcal{C}) = \Big(\frac{\partial h(\mathcal{W},\mathcal{C})}{\partial w_{n,m}^{[k]}}\Big)_{1\le n,m\le N}$$
$$= \begin{pmatrix} [\boldsymbol{C}_1\boldsymbol{W}_1\widehat{\boldsymbol{R}}_{\mathbf{x}}]_{k,(k-1)N+1} & \cdots & [\boldsymbol{C}_1\boldsymbol{W}_1\widehat{\boldsymbol{R}}_{\mathbf{x}}]_{k,(k-1)N+N} \\ \vdots & & \vdots \\ [\boldsymbol{C}_N\boldsymbol{W}_N\widehat{\boldsymbol{R}}_{\mathbf{x}}]_{k,(k-1)N+1} & \cdots & [\boldsymbol{C}_N\boldsymbol{W}_N\widehat{\boldsymbol{R}}_{\mathbf{x}}]_{k,(k-1)N+N} \end{pmatrix}.$$

Moreover, for every $\mathcal{C} \in \mathbb{R}^{K\times K\times N}$, $\nabla_{\mathcal{W}} h(\cdot,\mathcal{C})$ *is Lipschitz continuous, with modulus*

$$L_{\mathcal{W}}(\mathcal{C}) = \rho_{\mathcal{C}}\varrho_{\widehat{\boldsymbol{R}}_{\mathbf{x}}}, \tag{18}$$

where $\widehat{\boldsymbol{R}}_{\mathbf{x}}^{[k]}$ *is the matrix obtained by extraction of the columns* $kN+1$ *to* $(k+1)N$ *of* $\widehat{\boldsymbol{R}}_{\mathbf{x}}$,

$$\varrho_{\widehat{\boldsymbol{R}}_{\mathbf{x}}} = \max_{1\le k\le K} \|\widehat{\boldsymbol{R}}_{\mathbf{x}}^{[k]}\|_{\mathrm{S}} \tag{19}$$

and

$$(\forall \mathcal{C} \in \mathbb{R}^{K\times K\times N}) \quad \rho_{\mathcal{C}} = \max_{1\le n\le N} \|\boldsymbol{C}_n\|_{\mathrm{S}}. \tag{20}$$

Proof: See Appendix A. ∎

Lemma 2 *The partial gradient of* h *with respect to* $\mathcal{C}$*, evaluated at* $(\mathcal{W},\mathcal{C}) \in \mathbb{R}^{N\times N\times K} \times \mathbb{R}^{K\times K\times N}$*, reads:*

$$\nabla_{\mathcal{C}} h(\mathcal{W},\mathcal{C}) = (\nabla_{\boldsymbol{C}_n} h(\mathcal{W},\mathcal{C}))_{1\le n\le N} \in \mathbb{R}^{K\times K\times N}, \tag{21}$$

where, for all $n \in \{1,\dots,N\}$,

$$\nabla_{\boldsymbol{C}_n} h(\mathcal{W},\mathcal{C}) = \frac{1}{2}\boldsymbol{W}_n\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_n^{\top} + \alpha(\mathrm{Diag}(\boldsymbol{C}_n) - \boldsymbol{I}_K).$$

Moreover, for every $\mathcal{W} \in \mathbb{R}^{N\times N\times K}$, $\nabla_{\mathcal{C}} h(\mathcal{W},\cdot)$ *is Lipschitz continuous with constant modulus*

$$L_{\mathcal{C}} = \alpha. \tag{22}$$

Proof: See Appendix B ∎

According to Lemmas 1 and 2, both partial derivatives of h are well defined and continuous with respect to $\mathcal{W}$ and $\mathcal{C}$, which shows in particular that h is a C^1 function.

b) Functions f and g: Functions f and g are proper (i.e., finite-valued at least at one point), and lower-semicontinuous. Furthermore, function g is convex (see, for e.g., [22, Example 24.66], for a proof), while function f is not. Both functions f and g are not differentiable but still, it is possible to manipulate them efficiently, for minimization purpose, through their proximity operator [23].[2] The following lemmas provide the expression for these operators, that will then be perused in our algorithm. The proofs for these lemmas mainly rely on the fact that both f and g are so-called spectral functions [24], depending only on the spectral values of their inputs. Note that, despite the non-convexity of f, its proximity operator is still uniquely defined, as shown in the proof for Lemma 3.

Lemma 3 *Let* $\mathcal{W}' \in \mathbb{R}^{N\times N\times K}$*, and some* $c > 0$*. We define the proximity operator of* f *at* $\mathcal{W}'$ *as*

$$prox_{cf}(\mathcal{W}') = \underset{\mathcal{W}\in\mathbb{R}^{N\times N\times K}}{\mathrm{argmin}} \frac{1}{2}\|\mathcal{W}-\mathcal{W}'\|^2 + cf(\mathcal{W})$$
$$= (\boldsymbol{U}_{\boldsymbol{W}'^{[k]}}\,\mathrm{Diag}(\boldsymbol{\sigma}_{\boldsymbol{W}^{[k]}})\boldsymbol{V}_{\boldsymbol{W}'^{[k]}})_{1\le k\le K}, \tag{23}$$

[2] See also https://proximity-operator.net/

where $(\forall k \in \{1,\dots,K\})$ $\boldsymbol{W}'^{[k]} = \boldsymbol{U}_{\boldsymbol{W}'^{[k]}}\,\mathrm{Diag}(\boldsymbol{\sigma}_{\boldsymbol{W}'^{[k]}})\boldsymbol{V}_{\boldsymbol{W}'^{[k]}}$ *is the singular value decomposition of* $\boldsymbol{W}'^{[k]}$*, and*

$$\boldsymbol{\sigma}_{\boldsymbol{W}^{[k]}} = \frac{\boldsymbol{\sigma}_{\boldsymbol{W}'^{[k]}} + \sqrt{\boldsymbol{\sigma}^2_{\boldsymbol{W}'^{[k]}} + 4c}}{2}. \tag{24}$$

Proof: See Appendix C. ∎

Lemma 4 *Let* $\mathcal{C}' \in \mathbb{R}^{K\times K\times N}$*, and some* $c > 0$*. The proximity operator of* g *at* $\mathcal{C}'$ *is given by*

$$prox_{cg}(\mathcal{C}') = \underset{\mathcal{W}\in\mathbb{R}^{K\times K\times N}}{\mathrm{argmin}} \frac{1}{2}\|\mathcal{C}-\mathcal{C}'\|^2 + cg(\mathcal{C})$$
$$= (\boldsymbol{U}_{\boldsymbol{C}'_n}\,\mathrm{Diag}(\boldsymbol{\sigma}_{\boldsymbol{C}_n})\boldsymbol{V}_{\boldsymbol{C}'_n})_{1\le n\le N} \tag{25}$$

where $(\forall n \in \{1,\dots,N\}), \boldsymbol{C}'_n = \boldsymbol{U}_{\boldsymbol{C}'_n}\,\mathrm{Diag}(\boldsymbol{\sigma}_{\boldsymbol{C}'_n})\boldsymbol{V}_{\boldsymbol{C}'_n}$ *is the singular value decomposition of* $\boldsymbol{C}'_n$*, and*

$$\boldsymbol{\sigma}_{\boldsymbol{C}_n} = \max\left(\epsilon, \frac{\boldsymbol{\sigma}_{\boldsymbol{C}'_n} + \sqrt{(\boldsymbol{\sigma}_{\boldsymbol{C}'_n})^2 + 2c}}{2}\right). \tag{26}$$

Proof: See Appendix D. ∎

*B. Proposed **PALM-IVA-G** algorithm*

As shown in the previous subsection, the minimization of (12) amounts to solve Problem (13), that has a structure particularly well suited to block alternating minimization. More precisely, we have shown that the cost function includes (partially) Lipschitz differentiable terms acting on both $(\mathcal{W},\mathcal{C})$ variables (Lemmas 1 and 2), as well as two terms acting separately on these variables. Despite being non differentiable, these terms are proper, lower-semicontinuous, and with a tractable proximity operator (Lemmas 3 and 4). These results pave the way for applying a block alternating proximal gradient algorithm, as studied for instance in [25], [26], [27]. Here, we opted for **PALM** introduced in [28], because of its powerful convergence results. We adapted here **PALM** mechanism to the minimization of the cost (12) and thus designed **PALM-IVA-G** presented in Alg. 1. In **PALM** initial study, the convergence was shown in the case of sequential block updates. Here, we instead opted for a more versatile update scheme that follows the so-called *essentially cyclic rule* [25], allowing each block to be updated more than once, per main iteration. This assumption hence gives more flexibility to our algorithm, and it is straightforward to adapt the proof given in [28] to this case, using a similar technique to [25].

At each step $i \in \mathbb{N}$ of **PALM-IVA-G** main loop, we update $\mathcal{W}$ (resp. $\mathcal{C}$) a number $n_{\mathcal{W}}(i) \le \overline{n}_{\mathcal{W}}$ (resp. $n_{\mathcal{C}}(i) \le \overline{n}_{\mathcal{C}}$) of times, with $\overline{n}_{\mathcal{W}}$ and $\overline{n}_{\mathcal{C}}$ positive integers. The updates include gradient, and proximal steps, as follows. First, gradient steps on h with respect to the active block, $\mathcal{W}$ or $\mathcal{C}$, with positive stepsizes $c_{\mathcal{W}}^{(i)}$ or $c_{\mathcal{C}}$, respectively, are conducted. Then, proximal steps on f or g, are run, using the same stepsizes. Inner and outer loops are controlled by a maximum number of iterations, and furthermore include early stopping tests, based on comparing the following quantities to the (small) precision parameter $\delta > 0$:

$$(\forall (\mathcal{W},\mathcal{W}') \in (\mathbb{R}^{N\times N\times K})^2)$$
$$\theta_{\mathcal{W}}(\mathcal{W},\mathcal{W}') = \max_{\substack{1\le n\le N\\ 1\le k\le K}} \frac{\|\boldsymbol{w}_n'^{[k]} - \boldsymbol{w}_n^{[k]}\|^2}{2N}, \tag{27}$$

Algorithm 1 PALM-IVA-G

Require: Empirical covariance $\widehat{\boldsymbol{R}}_{\mathbf{x}}$, initial tensors $(\mathcal{W}^{(0)}, \mathcal{C}^{(0)}) \in \mathcal{GL}_N(\mathbb{R})^K \times (\epsilon \boldsymbol{I}_K + \mathcal{S}_K^+)^N$, penalty weight $\alpha > 0$, stepsizes $\gamma_{\mathcal{C}} \in (0,2), \gamma_{\mathcal{W}} \in (0,1)$, maximal inner/outer loops $\overline{n}_{\mathcal{W}} \in \mathbb{N} \setminus \{0\}, \overline{n}_{\mathcal{C}} \in \mathbb{N} \setminus \{0\}, \overline{N} \in \mathbb{N} \setminus \{0\}$, precision $\delta > 0$.

Ensure: $(\mathcal{W}_{\text{out}}, \mathcal{C}_{\text{out}})$

▷ *Initialization* ◁
$\mathcal{W}^{(0,0)} \leftarrow \mathcal{W}^{(0)}$
$\mathcal{C}^{(0,0)} \leftarrow \mathcal{C}^{(0)}$
$c_{\mathcal{C}} \leftarrow \frac{\gamma_{\mathcal{C}}}{\alpha}$
$i \leftarrow 0$
$\theta_{\text{ext}}^{(0)} \leftarrow +\infty$
▷ *Start Main Loop* ◁
while $\theta_{\text{ext}}^{(i)} > \delta$ **or** $i < \overline{N}$ **do**
  ▷ *Update* $\mathcal{W}$ ◁
  $c_{\mathcal{W}}^{(i)} \leftarrow \frac{\gamma_{\mathcal{W}}}{L_{\mathcal{W}}(\mathcal{C}^{(i)})}$ using (18)
  $j \leftarrow 0$
  $\theta_{\text{int}}^{(0)} \leftarrow +\infty$
  while $\theta_{\text{int}}^{(j)} > \delta$ **or** $j < \overline{n}_{\mathcal{W}}$ **do**
    $\mathcal{W}^{(i,j+1)} \leftarrow \text{prox}_{c_{\mathcal{W}}^{(i)} f}(\mathcal{W}^{(i,j)} - c_{\mathcal{W}}^{(i)} \nabla_{\mathcal{W}} h(\mathcal{W}^{(i,j)}, \mathcal{C}^{(i)}))$
      using (17) and (23)
    $\theta_{\text{int}}^{(j)} \leftarrow \theta_{\mathcal{W}}(\mathcal{W}^{(i,j+1)}, \mathcal{W}^{(i,j)})$
    $j \leftarrow j+1$
  $\mathcal{W}^{(i+1,0)} \leftarrow \mathcal{W}^{(i,j)}$
  $\mathcal{W}^{(i+1)} \leftarrow \mathcal{W}^{(i+1,0)}$
  ▷ *Update* $\mathcal{C}$ ◁
  $j \leftarrow 0$
  $\theta_{\text{int}}^{(0)} \leftarrow +\infty$
  while $\theta_{\text{int}}^{(j)} > \delta$ **or** $j < \overline{n}_{\mathcal{C}}$ **do**
    $\mathcal{C}^{(i,j+1)} \leftarrow \text{prox}_{c_{\mathcal{C}} g}(\mathcal{C}^{(i,j)} - c_{\mathcal{C}} \nabla_{\mathcal{C}} h(\mathcal{C}^{(i,j)}, \mathcal{W}^{(i+1)}))$
      using (21) and (25)
    $\theta_{\text{int}}^{(j)} \leftarrow \theta_{\mathcal{C}}(\mathcal{C}^{(i,j+1)}, \mathcal{C}^{(i,j)})$
    $j \leftarrow j+1$
  $\mathcal{C}^{(i+1,0)} \leftarrow \mathcal{C}^{(i,j)}$
  $\mathcal{C}^{(i+1)} \leftarrow \mathcal{C}^{(i+1,0)}$
  ▷ *Evaluate Stopping Criteria* ◁
  $\theta_{\mathcal{W}}^{(i+1)} \leftarrow \theta_{\mathcal{W}}(\mathcal{W}^{(i+1)}, \mathcal{W}^{(i)})$ using (27)
  $\theta_{\mathcal{C}}^{(i+1)} \leftarrow \theta_{\mathcal{C}}(\mathcal{C}^{(i+1)}, \mathcal{C}^{(i)})$ using (28)
  $\theta_{\text{ext}}^{(i+1)} \leftarrow \max(\theta_{\mathcal{W}}^{(i+1)}, \theta_{\mathcal{C}}^{(i+1)})$
  $i \leftarrow i+1$
$(\mathcal{W}_{\text{out}}, \mathcal{C}_{\text{out}}) \leftarrow (\mathcal{W}^{(i)}, \mathcal{C}^{(i)})$
return $(\mathcal{W}_{\text{out}}, \mathcal{C}_{\text{out}})$

$$(\forall (\mathcal{C}, \mathcal{C}') \in (\mathbb{R}^{K\times K\times N})^2) \quad \theta_{\mathcal{C}}(\mathcal{C}, \mathcal{C}') = \max_{\substack{1\le n\le N\\ 1\le k\le K}} \frac{||\boldsymbol{c}'_{n,k} - \boldsymbol{c}_{n,k}||^2}{2K}, \tag{28}$$

where, for $n \in \{1, \dots, N\}$ and $k \in \{1, \dots, K\}$, the notation $\boldsymbol{c}_{n,k}$ refers to the k-th row of the matrix $\boldsymbol{C}_n$.

Let us now move to Section IV, with the aim to establish the convergence of **PALM-IVA-G** iterates $(W^{(i)}, \mathcal{C}^{(i)})_{i\in\mathbb{N}}$. Practical settings for **PALM-IVA-G** hyper-parameters will be discussed in Section V.

IV. Convergence result

Let us state our convergence theorem for the proposed **PALM-IVA-G** algorithm.

Theorem 2 (Convergence of PALM-IVA-G) *Assuming the setting of Algorithm 1, the infinite sequence $(\mathcal{W}^{(i)}, \mathcal{C}^{(i)})_{i\in\mathbb{N}}$ converges to a critical point $(\mathcal{W}^*, \mathcal{C}^*)$ of function $J_{\text{IVA-G}}^{Reg}$ given in (12).*

Here, a critical point is defined as in [28, Rem.1 (iv)], i.e., $0 \in \partial J_{\text{IVA-G}}^{\text{Reg}}(\mathcal{W}^*, \mathcal{C}^*)$, where ∂ denotes the limiting subdifferential operator. The proof of the above result relies on [28, Theorem 1] in the case of a cyclic update of the blocks. The latter states the convergence of a generic form of **PALM** method under several assumptions regarding the properties of the cost function, and provided the sequence is bounded. It is hence sufficient to show that such conditions hold in our case, to prove Theorem 2. In the previous section, we have already seen that f and g are proper and lower semi-continuous functions such that the proximal operators prox_{cf} and prox_{cg} are defined for all $c > 0$, at any point $\mathcal{W} \in \mathcal{GL}_N(\mathbb{R})^K$ and $\mathcal{C} \in (\mathcal{S}_K^{++})^N$, respectively. We also outlined that h is a C^1 function and for every given $\mathcal{W}' \in \mathbb{R}^{N\times N\times K}$ (resp. $\mathcal{C}' \in \mathbb{R}^{K\times K\times N}$), the function $\mathcal{C} \mapsto h(\mathcal{W}', \mathcal{C})$ (resp. $\mathcal{W} \mapsto h(\mathcal{W}, \mathcal{C}')$) is $C^{1,1}_{L_{\mathcal{C}}(\mathcal{W}')}$, i.e. its partial gradient $\mathcal{C} \mapsto \nabla_{\mathcal{C}} h(\mathcal{W}', \mathcal{C})$ is globally Lipschitz with modulus $L_{\mathcal{C}}(\mathcal{W}')$ (resp. $C^{1,1}_{L_{\mathcal{W}}(\mathcal{C}')}$). We note that those first properties were necessary to well-defining the algorithm itself.

Using our notation, to complete the proof, we need to demonstrate each item of the following Assumption 1 is satisfied.

Assumption 1

1) $\inf_{\mathbb{R}^{N\times N\times K}\times\mathbb{R}^{K\times K\times N}} J_{\text{IVA-G}}^{Reg} > -\infty$.
2) *∇h is Lipschitz continuous on bounded subsets of $\mathbb{R}^{N\times N\times K} \times \mathbb{R}^{K\times K\times N}$.*
3) *$(\mathcal{W}^{(i)}, \mathcal{C}^{(i)})_{i\in\mathbb{N}}$ is bounded.*
4) *There exists $(\lambda_{\mathcal{W}}^+, \lambda_{\mathcal{W}}^-, \lambda_{\mathcal{C}}^+, \lambda_{\mathcal{C}}^-) > 0$ such that:*

$$(\forall i \in \mathbb{N})\ \lambda_{\mathcal{W}}^- \le L_{\mathcal{W}}(\mathcal{C}^{(i)}) \le \lambda_{\mathcal{W}}^+,$$

and

$$(\forall i \in \mathbb{N})\ \lambda_{\mathcal{C}}^- \le L_{\mathcal{C}}(\mathcal{W}^{(i)}) \le \lambda_{\mathcal{C}}^+.$$

5) *$J_{\text{IVA-G}}^{Reg}$ is a KL function.*

Indeed, following the steps of [28], the update scheme of the algorithms makes the sequence of the costs decreasing, and using the first item of the above assumption, this sequence has a finite limit. Then, items 2) to 4) enable to prove that the limit set of $(\mathcal{W}^{(i)}, \mathcal{C}^{(i)})_{i\in\mathbb{N}}$ is nonempty, compact, and contains only critical points of $J_{\text{IVA-G}}^{\text{Reg}}$. Finally, the last item is used to prove that $(\mathcal{W}^{(i)}, \mathcal{C}^{(i)})_{i\in\mathbb{N}}$ is a Cauchy sequence, hence convergent. The verification of Assumption 1 is technically involved and is therefore provided in Appendix E.

V. Experimental results

We now present a set of experiments, to assess the quantitative, qualitative, and computational performance of **PALM-IVA-G** on tasks of independent Gaussian sources separation.

A. Experimental protocol

1) Qualitative evaluation: In all the experiments, the proposed method **PALM-IVA-G**, as well as the benchmarks, are evaluated by means of the so-called jISI (joint Intersymbol Interference) score, also used in [12]. This score is an extension of the ISI score that was introduced in [29] to assess the performance of ICA methods. jISI score measures the correspondance between demixing matrices $\mathcal{W} = [\boldsymbol{W}^{[1]}, \dots, \boldsymbol{W}^{[K]}]$ and the ground truth mixing matrices $\mathcal{A} = [\boldsymbol{A}^{[1]}, \dots, \boldsymbol{A}^{[K]}]$, up to a common permutation and a rescaling of their rows. Let us note

$$(\forall k \in \{1, \dots, K\}) \quad \mathbf{G}^{[k]} = \boldsymbol{W}^{[k]} \boldsymbol{A}^{[k]} \tag{29}$$

and, for every $(m, n) \in \{1, \dots, N\}^2$, the mean $\overline{g}_{n,m} = \sum_{k=1}^{K} |g_{n,m}^{[k]}|$ of the (n, m)-th entry of tensor $\mathcal{G} = [\mathbf{G}^{[1]}, \dots, \mathbf{G}^{[K]}]$. The jISI score of the pair $(\mathcal{W}, \mathcal{A})$ is

$$\text{jISI}(\mathcal{W}, \mathcal{A}) = \frac{1}{2N(N-1)} \left[\sum_{n=1}^{N} \left(\sum_{m=1}^{N} \frac{\overline{g}_{n,m}}{\max_p \overline{g}_{n,p}} - 1 \right) + \sum_{m=1}^{N} \left(\sum_{n=1}^{N} \frac{\overline{g}_{n,m}}{\max_p \overline{g}_{p,m}} - 1 \right) \right]. \tag{30}$$

As defined in (30), the jISI score is a real number between 0 and 1, with jISI score equals 0 if and only if $\overline{\mathbf{G}} = (\overline{g}_{n,m})_{1 \leq n,m \leq N}$ has exactly one positive coefficient by row and by column, that is when it is a (possibly permuted) diagonal matrix. This happens when the $(\boldsymbol{W}^{[k]} \boldsymbol{A}^{[k]})_{1 \leq k \leq K}$ are simultaneously permuted diagonal matrices, which is the best situation one can expect from a source separation step. Hence, the smallest the jISI score, the better quality for the source separation.

2) Benchmark methods: Our algorithm **PALM-IVA-G** is compared against two state of the art algorithms for independant Gaussian source separation, namely **IVA-G-V** and **IVA-G-N**, both introduced in [12]. Those two algorithms perform the minimization of the cost function $\widetilde{J}_{\text{IVA}-\text{G}}$ (8) (which, as we remind, only depends on variable $\mathcal{W}$) using, respectively, a gradient descent and a Newton's method. Both implement a backtracking linesearch, and a normalization of the rows of the demixing matrix at each iteration. Hyper-parameters are set as recommended in [12]. Let us emphasize that both of these algorithms are empirical, and, to our knowledge, do not benefit from established convergence guarantees, though in practice we did not observe any failure.

For every experiment, and each algorithm, we obtain an estimate for which we record the jISI score reached at convergence (i.e., when stopping criterion is reached), and the computational time, in seconds, denoted $\mathtt{T}$, to reach this convergence point. All algorithms are implemented in Python 3.10.12 and run on a Dell Precision 5820 Workstation with 11th Gen Intel(R) Core(TM) i9-10900X at 3.00GHz, equipped with 32Go Ram.

3) Synthetic dataset generation: For each experimental setup, we define a generative model defined by the ground-truth variables $(\mathcal{A}, \boldsymbol{\Sigma})$, and use this model to generate source data $\mathcal{S}$ of length $V = 10000$, and then mix it into observed data $\mathcal{X}$. The goal is to recover estimates of $([(\boldsymbol{A}^{[1]})^{-1}, \dots, (\boldsymbol{A}^{[K]})^{-1}], [\boldsymbol{\Sigma}_1^{-1}, \dots, \boldsymbol{\Sigma}_N^{-1}])$ up to the scaling and permutation ambiguities. All trials are repeated over 100 independent runs, and we compute the mean μ_{jISI} (resp. $\mu_{\mathtt{T}}$) and standard-deviation σ_{jISI} (resp. $\sigma_{\mathtt{T}}$) values for the jISI score (resp. computational time $\mathtt{T}$). Most experiments are conducted for various values for the dimensions (K, N), specified in each test case. Depending on the nature of the phenomenon modeled, the covariance and the mixing matrices may have various properties, leading us to define several sets of experiments, detailed hereafter.

We aim at exploring the impact of the overall level of correlation across the datasets (given by the extra-diagonal coefficients of the SCVs covariance matrices), and the variability of those correlations. Our generative model used to simulate SCVs depends on two parameters $\boldsymbol{\rho} = [\rho_1, \dots, \rho_N]^\top \in [0, 1]^N$ and $\lambda \in [0, 1]$. Given those, we compute a ground truth tensor $\boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_N]$ with, for every n,

$$\boldsymbol{\Sigma}_n = \rho_n \mathbf{1}\mathbf{1}^\top + \frac{\lambda}{R} \boldsymbol{Q}\boldsymbol{Q}^\top + \eta_n \boldsymbol{I}_K \tag{31}$$

with

$$\eta_n = 1 - \rho_n - \lambda \in [0, 1], \tag{32}$$

matrix $\boldsymbol{Q}$ randomly drawn in $\mathbb{R}^{K \times R}$ with elements $q_{i,j} \sim \mathcal{N}(0, 1)$ mutually independent, and $R \in \mathbb{N} \setminus \{0\}$ a predefined rank value.

Then, we have, for every n, and every entry (i, j),

$$\begin{cases} \mathbb{E}\{(\boldsymbol{\Sigma}_n)_{i,j}\} = \rho_n + (\lambda + \eta_n)\delta_{i,j} \\ \text{Var}\{(\boldsymbol{\Sigma}_n)_{i,j}\} = \frac{\lambda^2}{R}(1 + \delta_{i,j}). \end{cases} \tag{33}$$

This means that $\boldsymbol{\rho}$ controls the average correlation across the datasets while λ controls the variability between the correlations. The third term ensures that the covariance matrices we use are positive definite. In our experiments, we opt for four scenarios, corresponding to low/high variability, and low/high correlation, as defined below:

- Case A: low correlation, low variability. $\lambda = 0.04$ and, $\boldsymbol{\rho}$ regularly sampled in $[0.2, 0.3]^N$,
- Case B: low correlation, high variability. $\lambda = 0.25$ and, $\boldsymbol{\rho}$ regularly sampled in $[0.2, 0.3]^N$,
- Case C: high correlation, low variability. $\lambda = 0.04$ and, $\boldsymbol{\rho}$ regularly sampled in $[0.6, 0.7]^N$,
- Case D: high correlation, high variability. $\lambda = 0.25$ and, $\boldsymbol{\rho}$ regularly sampled in $[0.6, 0.7]^N$,

The sources are expected to be easier to separate in case D, while case A is more challenging, since increasing the correlation or the variability of the sources decreases the Cramer-Rao Lower-Bound of the ML estimator and thus tends to improve the separation [15]. Each of the four cases is investigated, for various sizes $K \in \{5, 10, 20\}$ and $N \in \{10, 20\}$, and we set $R = K + 10$. In all experiments, the ground truth tensor $\mathcal{A}$ is simulated by drawing its entries independently from a zero-mean Gaussian distribution with a standard deviation of 1.

For each simulated pair $(\mathcal{A}, \boldsymbol{\Sigma})$, we build the corresponding input data $\mathcal{X}$, following the mixing model (1). Note that, as recommended in [12], we systematically performed a whitening of the data, before applying the algorithms. This step amounts to multiplying, for every $k \in \{1, \dots, K\}$, the latent sources by full-rank matrices $(\boldsymbol{B}^{[k]})_{1 \leq k \leq K}$, and solving the demixing problem on $\boldsymbol{X}'^{[k]} = \boldsymbol{B}^{[k]} \boldsymbol{X}^{[k]} = \boldsymbol{B}^{[k]} \boldsymbol{A}^{[k]} \boldsymbol{S}^{[k]}$. The whitening matrices are set to decrease the spectral norm of $\widehat{\boldsymbol{R}}_{\mathbf{x}}$, with the aim to improve the conditioning of the loss function, and hence accelerate the empirical convergence of the methods.

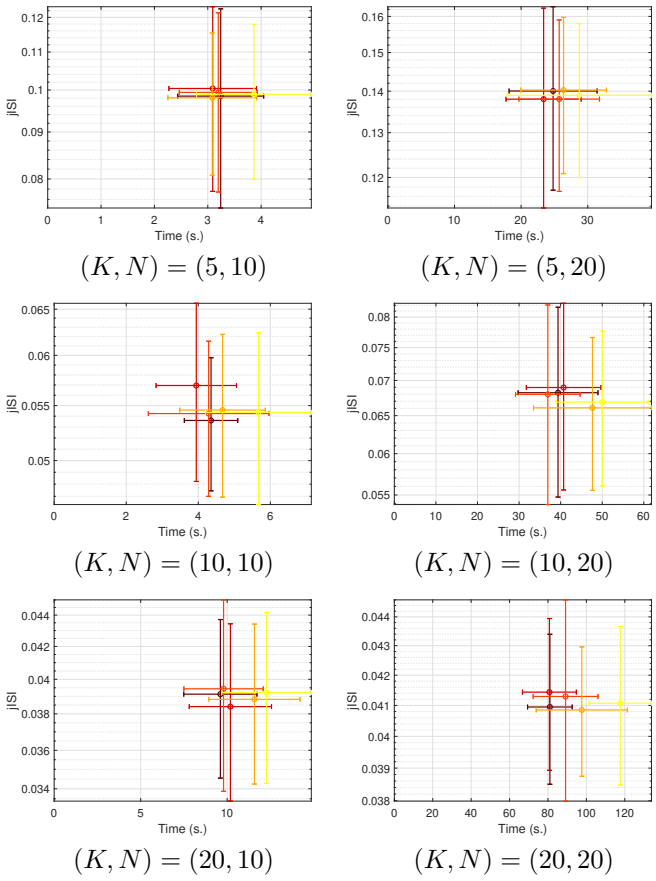


Fig. 1: jISI scores vs computation times (mean $\pm$ standard deviation) using **PALM-IVA-G** for Case A over 20 runs, with α taking values in $\{0.1, 0.5, 1, 5, 10\}$ (the brighter color, the higher α). The jISI score varies little with α, the latter impacting mostly the time, with good compromise at $\alpha = 1$ (dark orange).

4) Algorithmic settings: The implementation of the proposed **PALM-IVA-G** algorithm requires the setting of (i) the stepsizes, (ii) the regularization weight, and (iii) the stopping conditions for internal and external loops.

We set the stepsizes to constant values $(\gamma_{\mathcal{C}}, \gamma_{\mathcal{W}}) = (1.99, 0.99)$, i.e. as large as possible to meet the convergence theorem assumptions. The penalty parameter α appears to have little influence over the performance of **PALM-IVA-G**, in terms of jISI metric and computational time, as long as it is chosen in a reasonable range. This can be observed empirically in Fig. 1 summarizing results for Case A and various values of α. Indeed, except for $\alpha = 10$ which seems to yield systematically slower computations, we cannot observe that a value of α gives consistently better results than the others. Similar behaviors were obtained for Cases B to D. For the experiments, we thus retain $\alpha = 1$, as it achieves a good compromise in terms of time complexity.

PALM-IVA-G algorithm, as well as its competitors **IVA-G-V** and **IVA-G-N** are run until a certain stopping criterion is reached, with a maximum number of $\overline{N} = 20000$ iterations. The precision threshold is set to $\delta = 10^{-10}$, and the maximum number of iterations within the internal loops of **PALM-IVA-G** are $\bar{n}_{\mathcal{C}} = 1$ and $\bar{n}_{\mathcal{W}} = 15$. For **IVA-G-V** (resp. **IVA-G-N**), we monitor only the value of (27) between consecutive iterates, stopping once lower than $\delta = 10^{-6}$ (resp. $\delta = 10^{-7}$). Note that the values of $(\bar{n}_{\mathcal{C}}, \bar{n}_{\mathcal{W}}, \delta)$ have been empirically set to reach the best trade-off between jISI score and computational time, ensuring fair comparisons of the methods.

The validity of our settings can be assessed visually on Fig. 2, showing the cost function evolution using our **PALM-IVA-G** algorithm in a representative example.

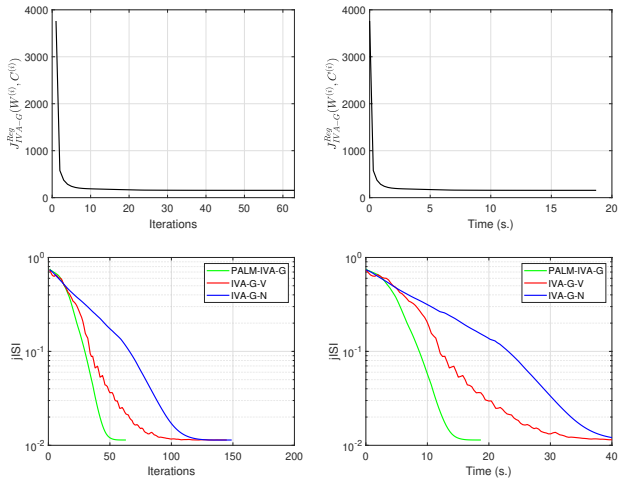


Fig. 2: Top: Empirical convergence of **PALM-IVA-G**, on a synthetic example from Case D, with $(K, N) = (20, 20)$, $\lambda = 0.25$ and $\boldsymbol{\rho} \in [0.2, 0.3]^\top$. Cost function across iterations (left) and time in second (right). Bottom: On the same example, evolution of jISI score along iterations (left) and time in seconds (right), for the compared methods.

B. Results

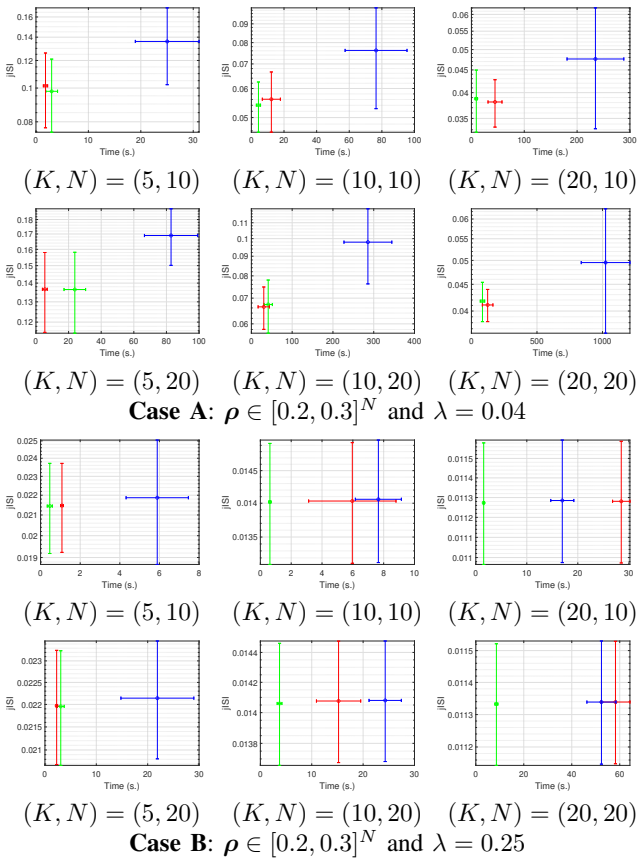
We now present the results of the experiments, and comment on the strengths and weaknesses of our algorithm in comparison with the benchmarks. The results are illustrated in Figs. 3 and 4, and summarized in Tab. I. On Fig. 3 and 4, we display, for each of the three algorithms, a cross centered at $(\mu_{\mathrm{T}}, \mu_{\mathrm{jISI}})$, and spread out by $\pm\sigma_{\mathrm{T}}$ (horizontal) and $\pm\sigma_{\mathrm{jISI}}$ (vertical) axis. The best results thus correspond to a cross located on the bottom left side of the figure (i.e., low jISI score reached in a minimal time). In Tab. I, we highlight in bold font (resp. italic bold font) the best results in terms of μ_{jISI} (resp. μ_{T}), considering of similar quality jISI scores with less than 10^{-4} difference (resp. computation times with less that 10^{-2} difference).

All algorithms achieve what appears to be optimal separation in easy cases (B and D). In difficult cases, it is generally either **IVA-G-V** or **PALM-IVA-G** that have the best jISI, with an advantage for **PALM-IVA-G** in small dimensions and an advantage for **IVA-G-V** in larger dimensions, but in all cases performance remains close. **IVA-G-N** is generally not as good, but manages to keep its performance close to that of the other algorithms, except for Case A, where it gives jISI values 20 to 50 percent higher than the other algorithms on average.

Computations of **PALM-IVA-G** are tractable, and stay under two minutes. The running time is sensitive to the number of sources, and seems to grow linearly with the number of datasets within the tested dimensions. For $N = 10$, our algorithm takes less than 15 seconds to run in average in all

cases. Meanwhile, **IVA-G-V** is less sensitive to N and also manages to separate the sources in two minutes in average at most for the dimensions in the experiment. However, **IVA-G-N** is the slowest of all three algorithms, it takes several tens of seconds in small dimensions for Cases A and C, and its computational cost becomes prohibitive in larger dimensions, taking an average time of fifteen minutes to run in Case A for $K = 20$ and $N = 20$. Visually, it leads to blue crosses often positioned on the right side of the plots in Figs. 3 and 4.

In contrast to gradient descent, iterations using Newton's method or proximal gradient are more informed and are expected to find the local minimum more efficiently. On the other hand, these methods are more costly, since they involve Hessian inversions for **IVA-G-N**, and SVDs for our **PALM-IVA-G**, and such cost is not necessarily offset by a gain in number of iterations. Besides, the update scheme of **PALM-IVA-G** implies many more updates of the block $\mathcal{W}$ than the block $\mathcal{C}$, so most of the matrices for which we compute the SVD are of size $N \times N$ rather than $K \times K$, this is why the computation time of **PALM-IVA-G** increases faster with respect to N than K.



$(K, N) = (5, 10)$ $(K, N) = (10, 10)$ $(K, N) = (20, 10)$

$(K, N) = (5, 20)$ $(K, N) = (10, 20)$ $(K, N) = (20, 20)$

Case A: $\boldsymbol{\rho} \in [0.2, 0.3]^N$ and $\lambda = 0.04$

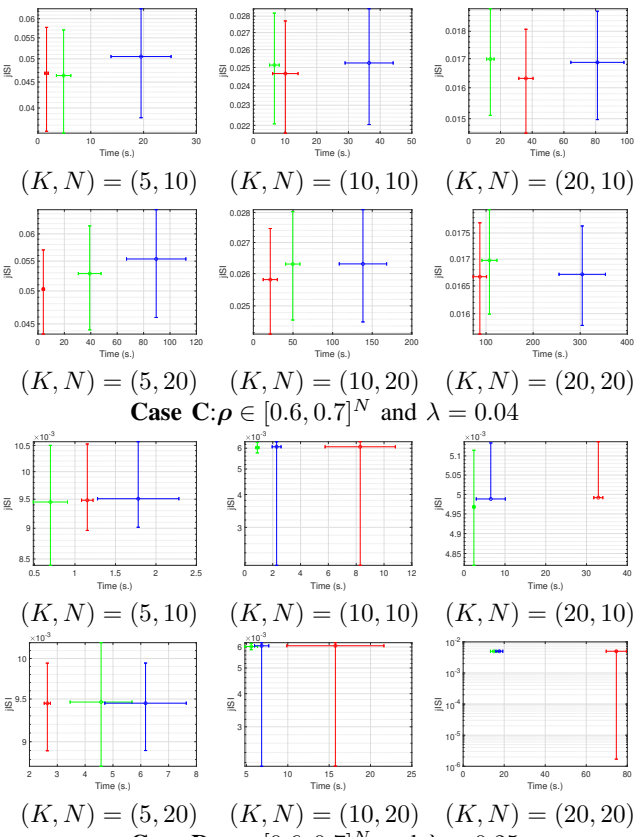
$(K, N) = (5, 10)$ $(K, N) = (10, 10)$ $(K, N) = (20, 10)$

$(K, N) = (5, 20)$ $(K, N) = (10, 20)$ $(K, N) = (20, 20)$

Case B: $\boldsymbol{\rho} \in [0.2, 0.3]^N$ and $\lambda = 0.25$

Fig. 3: jISI score vs computational time in seconds (mean $\pm$ standard deviation), for **PALM-IVA-G** (green), **IVA-G-V** (red) and **IVA-G-N** (blue), for Case A and Case B.

In conclusion, in addition to its established convergence guarantees, **PALM-IVA-G** appears to be competitive with the state of the art IVA-G algorithm, consistently achieving good jISI scores and taking reasonable time to run, especially when the number of sources is not too high.



$(K, N) = (5, 10)$ $(K, N) = (10, 10)$ $(K, N) = (20, 10)$

$(K, N) = (5, 20)$ $(K, N) = (10, 20)$ $(K, N) = (20, 20)$

Case C:$\boldsymbol{\rho} \in [0.6, 0.7]^N$ and $\lambda = 0.04$

$(K, N) = (5, 10)$ $(K, N) = (10, 10)$ $(K, N) = (20, 10)$

$(K, N) = (5, 20)$ $(K, N) = (10, 20)$ $(K, N) = (20, 20)$

Case D:$\boldsymbol{\rho} \in [0.6, 0.7]^N$ and $\lambda = 0.25$

Fig. 4: jISI score vs computational time in seconds (mean $\pm$ standard deviation), for **PALM-IVA-G** (green), **IVA-G-V** (red) and **IVA-G-N** (blue), for Cases C and D.

VI. Conclusion

In this article, we addressed the problem of joint blind source separation, through the IVA-G formulation. First, we derived a cost function parameterised by the demixing matrices and the precision matrices of the sources, to solve IVA-G based on the maximum likelihood estimator. We then introduced an additional term to fix the scaling ambiguity, hence forcing the precision matrices of output solution to have ones on its diagonal, completing the design of our cost function.

Then, we studied the different terms of this non-convex and non-smooth cost, in particular, we gave explicit formulas for the partial gradients of the smooth term with respect to both blocks of variable and for the proximity operators of the separable terms. Based on these results, we proposed an algorithm adapted from the **PALM** optimizer, and proved that the sequence of iterates converges globally to a critical point. We compared our method to two state-of-the-art IVA-G algorithms and showed that our method is competitive in terms of jISI score and computation time, especially for a moderate number of sources.

As a future work, those encouraging results on synthetic datasets should be confirmed on real data, for instance from fMRI or MEEG. Depending on the application considered, we could leverage model knowledge in the form of new regularization terms in the cost function. The proposed proximal alternating algorithm is versatile and would be easily adaptable to a large class of penalties and constraints.

TABLE I: Averaged jISI scores, μ_{jISI}, and averaged computational times, μ_{T}, in seconds, for **PALM-IVA-G**, **IVA-G-V** and **IVA-G-N** (from top to bottom). Best (i.e., lowest) jISI results (resp. lowest times) are highlighted in bold (resp. italic bold).

			$K=5$		$K=10$		$K=20$	
			$N=10$	$N=20$	$N=10$	$N=20$	$N=10$	$N=20$
PALM-IVA-G	Case A	μ_{jISI}	**9.79E-02**	**1.36E-01**	**5.40E-02**	6.74E-02	3.88E-02	4.18E-02
		μ_{T}	3.0	23.8	***4.3***	41.2	***9.2***	***84.2***
	Case B	μ_{jISI}	**2.14E-02**	**2.20E-02**	**1.40E-02**	**1.41E-02**	**1.13E-02**	**1.13E-02**
		μ_{T}	***0.5***	3.1	***0.6***	***3.7***	***1.5***	***8.6***
	Case C	μ_{jISI}	**4.63E-02**	5.29E-02	2.51E-02	2.63E-02	1.70E-02	1.70E-02
		μ_{T}	4.9	39.2	***6.6***	49.7	***13.6***	107.2
	Case D	μ_{jISI}	**9.45E-03**	**9.46E-03**	**6.03E-03**	**6.03E-03**	**4.97E-03**	**4.98E-03**
		μ_{T}	***0.7***	4.6	***0.9***	***5.6***	***2.4***	***14.9***
IVA-G-V	Case A	μ_{jISI}	1.01E-01	1.37E-01	5.61E-02	**6.64E-02**	**3.81E-02**	**4.11E-02**
		μ_{T}	***1.8***	***5.4***	12.2	***30.4***	44.6	123.7
	Case B	μ_{jISI}	**2.15E-02**	**2.20E-02**	**1.40E-02**	**1.41E-02**	**1.13E-02**	**1.13E-02**
		μ_{T}	1.1	***2.3***	6.0	15.2	28.5	58.2
	Case C	μ_{jISI}	4.68E-02	**5.03E-02**	**2.47E-02**	**2.58E-02**	**1.63E-02**	**1.67E-02**
		μ_{T}	***1.6***	***4.1***	10.1	***21.3***	36.0	***86.8***
	Case D	μ_{jISI}	**9.48E-03**	**9.45E-03**	**6.08E-03**	**6.07E-03**	**4.99E-03**	**5.00E-03**
		μ_{T}	1.2	***2.6***	8.3	15.8	32.9	74.7
IVA-G-N	Case A	μ_{jISI}	1.36E-01	1.69E-01	7.62E-02	9.79E-02	4.76E-02	4.95E-02
		μ_{T}	25.0	82.8	76.4	285.7	234.7	1023.5
	Case B	μ_{jISI}	2.19E-02	2.21E-02	**1.41E-02**	**1.41E-02**	**1.13E-02**	**1.13E-02**
		μ_{T}	5.9	21.9	7.7	24.3	17.0	52.4
	Case C	μ_{jISI}	5.05E-02	5.54E-02	2.52E-02	2.63E-02	1.69E-02	**1.67E-02**
		μ_{T}	19.5	89.5	36.5	138.3	81.1	304.3
	Case D	μ_{jISI}	**9.51E-03**	**9.45E-03**	**6.08E-03**	**6.06E-03**	**4.99E-03**	**5.00E-03**
		μ_{T}	1.8	6.2	2.3	6.8	6.5	17.6

APPENDIX A
PROOF OF LEMMA 1

Let $(\mathcal{W},\mathcal{C}) \in \mathbb{R}^{N\times N\times K}\times\mathbb{R}^{K\times K\times N}$. The partial derivative of h in (14), with respect to the (n,m,k)-th entry of tensor $\mathcal{W}$ reads

$$\frac{\partial h(\mathcal{W},\mathcal{C})}{\partial w_{n,m}^{[k]}} = \frac{1}{2}\sum_{n'=1}^{N}\frac{\partial \operatorname{tr}(\boldsymbol{C}_{n'}\boldsymbol{W}_{n'}\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_{n'}^{\top})}{\partial w_{n,m}^{[k]}}. \tag{34}$$

For $n' \neq n$, the terms of the above sum are null, hence

$$\frac{\partial h(\mathcal{W},\mathcal{C})}{\partial w_{n,m}^{[k]}} = \frac{1}{2}\frac{\partial \operatorname{tr}(\boldsymbol{C}_{n}\boldsymbol{W}_{n}\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_{n}^{\top})}{\partial w_{n,m}^{[k]}}. \tag{35}$$

Then, by applying the formula of the derivative of a matricial scalar product [30],

$$\begin{aligned}
\frac{\partial h(\mathcal{W},\mathcal{C})}{\partial w_{n,m}^{[k]}} &= \frac{1}{2}\frac{\partial\langle \boldsymbol{C}_{n}\boldsymbol{W}_{n}\widehat{\boldsymbol{R}}_{\mathbf{x}}|\boldsymbol{W}_{n}\rangle}{\partial w_{n,m}^{[k]}}\\
&= \frac{1}{2}\Big(\langle\frac{\partial(\boldsymbol{C}_{n}\boldsymbol{W}_{n}\widehat{\boldsymbol{R}}_{\mathbf{x}})}{\partial w_{n,m}^{[k]}}|\boldsymbol{W}_{n}\rangle + \langle\boldsymbol{C}_{n}\boldsymbol{W}_{n}\widehat{\boldsymbol{R}}_{\mathbf{x}}|\frac{\partial \boldsymbol{W}_{n}}{\partial w_{n,m}^{[k]}}\rangle\Big)\\
&= \frac{1}{2}\Big(\langle\boldsymbol{C}_{n}\frac{\partial\boldsymbol{W}_{n}}{\partial w_{n,m}^{[k]}}\widehat{\boldsymbol{R}}_{\mathbf{x}}|\boldsymbol{W}_{n}\rangle + \langle\boldsymbol{C}_{n}\boldsymbol{W}_{n}\widehat{\boldsymbol{R}}_{\mathbf{x}}|\frac{\partial \boldsymbol{W}_{n}}{\partial w_{n,m}^{[k]}}\rangle\Big)\\
&= \operatorname{tr}\Big(\frac{\partial\boldsymbol{W}_{n}}{\partial w_{n,m}^{[k]}}\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_{n}^{\top}\boldsymbol{C}_{n}\Big).
\end{aligned} \tag{36}$$

Hereabove, $\frac{\partial \boldsymbol{W}_{n}}{\partial w_{n,m}^{[k]}}$ is a matrix of dimension $K\times KN$ whose elements are equal to 0, except one, at row index k and column index $(k-1)N+m$, which is equal to 1. We deduce

$$\frac{\partial h(\mathcal{W},\mathcal{C})}{\partial w_{n,m}^{[k]}} = [\boldsymbol{C}_{n}\boldsymbol{W}_{n}\widehat{\boldsymbol{R}}_{\mathbf{x}}]_{k,(k-1)N+m}. \tag{37}$$

The above expression can be reexpressed in a more concise matrix form which gives (17), and proves the first part of the lemma.

Let $\mathcal{C}\in\mathbb{R}^{K\times K\times N}$. From (17), we can see that $\nabla_{\mathcal{W}}h(.,\mathcal{C})$ is linear, and thus Lipschitz continuous. Moreover, for every $\mathcal{W}\in\mathbb{R}^{N\times N\times K}$,

$$\begin{aligned}
\Big\|\frac{\partial h(\mathcal{W},\mathcal{C})}{\partial \mathcal{W}}\Big\|^2 &= \sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{m=1}^{N}([\boldsymbol{C}_{n}\boldsymbol{W}_{n}\widehat{\boldsymbol{R}}_{\mathbf{x}}]_{k,(k-1)N+m})^2\\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{m=1}^{N}([\boldsymbol{C}_{n}\boldsymbol{W}_{n}\widehat{\boldsymbol{R}}_{\mathbf{x}}]_{k,(k-1)N+m})^2\Big)\\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\|[\boldsymbol{C}_{n}\boldsymbol{W}_{n}\widehat{\boldsymbol{R}}_{\mathbf{x}}]_{k,(k-1)N+1,\ldots,N}\|^2\\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\|\boldsymbol{c}_{n,k}\boldsymbol{W}_{n}\widehat{\boldsymbol{R}}_{\mathbf{x}}^{[k]}\|^2.
\end{aligned} \tag{38}$$

Using the identity $\|\boldsymbol{A}\boldsymbol{B}\|\leq\|\boldsymbol{A}\|_{\text{S}}\|\boldsymbol{B}\|$, we have, for all $k\in\{1,\ldots,K\}, n\in\{1,\ldots,N\}$,

$$\|\boldsymbol{c}_{n,k}\boldsymbol{W}_{n}\widehat{\boldsymbol{R}}_{\mathbf{x}}^{[k]}\|^2 \leq \|\widehat{\boldsymbol{R}}_{\mathbf{x}}^{[k]}\|_{\text{S}}^2\|\boldsymbol{c}_{n,k}\boldsymbol{W}_{n}\|^2 \leq \varrho_{\widehat{\boldsymbol{R}}_{\mathbf{x}}}^2\|\boldsymbol{c}_{n,k}\boldsymbol{W}_{n}\|^2. \tag{39}$$

$\boldsymbol{C}_{n}\boldsymbol{W}_{n}$ is a $K\times KN$ matrix whose k-th line is equal to $\boldsymbol{c}_{n,k}\boldsymbol{W}_{n}$, hence,

$$\sum_{k=1}^{K}\|\boldsymbol{c}_{n,k}\boldsymbol{W}_{n}\|^2 = \|\boldsymbol{C}_{n}\boldsymbol{W}_{n}\|^2 \leq \|\boldsymbol{C}_{n}\|_{\text{S}}^2\|\boldsymbol{W}_{n}\|^2. \tag{40}$$

Overall,

$$\begin{aligned}
\Big\|\frac{\partial h(\mathcal{W},\mathcal{C})}{\partial \mathcal{W}}\Big\|^2 &\leq \sum_{n=1}^{N}\varrho_{\widehat{\boldsymbol{R}}_{\mathbf{x}}}^2\sum_{k=1}^{K}\|\boldsymbol{c}_{n,k}\boldsymbol{W}_{n}\|^2\\
&\leq \varrho_{\widehat{\boldsymbol{R}}_{\mathbf{x}}}^2\sum_{n=1}^{N}\rho_{\mathcal{C}}^2\|\boldsymbol{W}_{n}\|^2\\
&= \rho_{\mathcal{C}}^2\varrho_{\widehat{\boldsymbol{R}}_{\mathbf{x}}}^2\|\mathcal{W}\|^2.
\end{aligned} \tag{41}$$

This concludes the second part of the Lemma, i.e. $\mathcal{W}\rightarrow h(\mathcal{W},\mathcal{C})$ is Lipschitz differentiable with constant (18).

APPENDIX B
PROOF OF LEMMA 2

For all $\mathcal{W}\in\mathbb{R}^{N\times N\times K}$, $h(\mathcal{W},.)$ in (14) is quadratic, which yields the expression (21) in a straightforward manner, and concludes the first part of the proof. Let $\mathcal{W}'\in\mathbb{R}^{N\times N\times K}$. According to (21), it can be easily checked that $\nabla_{\mathcal{C}}h(\mathcal{W}',.)$ is Lipschitz with moduli α, hence $\mathcal{C}\rightarrow h(\mathcal{W}',\mathcal{C})$ is Lipschitz differentiable with constant (22), which ends the proof.

APPENDIX C
PROOF OF LEMMA 3

Function f in (15), reads, equivalently, for all $\mathcal{W}\in\mathbb{R}^{N\times N\times K}$, $f(\mathcal{W})=\sum_{k=1}^{K}\hat{f}(\boldsymbol{W}^{[k]})$, where, for all $\boldsymbol{M}\in\mathbb{R}^{N\times N}$, with singular values $\boldsymbol{\sigma}_{\boldsymbol{M}}=(\sigma_{\boldsymbol{M},\ell})_{1\leq\ell\leq N}$, $\hat{f}(\boldsymbol{M})=-\log|\det\boldsymbol{M}|=\phi_{\hat{f}}(\boldsymbol{\sigma}_{\boldsymbol{M}})$, with

$$\begin{aligned}
\phi_{\hat{f}}\colon \mathbb{R}^{N} &\mapsto (-\infty,+\infty]\\
\boldsymbol{\sigma}=(\sigma_1,\ldots,\sigma_N) &\mapsto \begin{cases}+\infty \text{ if } (\exists\ell\in\{1,\ldots,N\})\ \sigma_\ell\leq 0\\ -\sum_{l=1}^{N}\log\sigma_\ell \text{ otherwise.}\end{cases}
\end{aligned} \tag{42}$$

Let $c > 0$ and $\mathcal{W}' \in \mathbb{R}^{N\times N\times K}$. Then,

$$cf(\mathcal{W})+\frac{1}{2}\|\mathcal{W}-\mathcal{W}'\|^2 = \sum_{k=1}^{K} c\hat{f}(\boldsymbol{W}^{[k]})+\frac{1}{2}\|\boldsymbol{W}^{[k]}-\boldsymbol{W}^{'[k]}\|^2.$$

Hence, we can minimize the sum by minimizing all its terms independently. In other words, $\mathcal{W} \in \operatorname{prox}_{cf}(\mathcal{W}')$, if and only if, for every $k \in \{1,\dots,K\}$,

$$\boldsymbol{W}^{[k]} \in \operatorname{prox}_{c\hat{f}}(\boldsymbol{W}^{'[k]}).$$

It remains to derive an explicit form for $\operatorname{prox}_{c\hat{f}}$. To do so, let us use the following lemma whose proof is given in [31] (see also [24], [22] for similar results on the proximity operators of spectral functions).

Lemma 5 *Let $N \geq 1$ and $\phi : \mathbb{R}^N \mapsto (-\infty,+\infty]$ be a function whose domain is not empty and is contained in $[0,+\infty)^N$, such that ϕ is invariant by any permutation of its arguments and that the proximity operator of ϕ is defined. Let $\Phi : \mathbb{R}^{N\times N} \mapsto (-\infty,+\infty], \boldsymbol{M} \mapsto \phi(\boldsymbol{\sigma}_{\boldsymbol{M}})$ where $\boldsymbol{\sigma}_{\boldsymbol{M}}$ is a vector containing the N singular values of $\mathbf{M}$ in any order. Then, the proximity operator of Φ is defined, and for any $\boldsymbol{M}' \in \mathbb{R}^{N\times N}$, whose singular value decomposition reads $\boldsymbol{M}' = \boldsymbol{U}_{\boldsymbol{M}'}\operatorname{Diag}(\boldsymbol{\sigma}_{\boldsymbol{M}'})\boldsymbol{V}_{\boldsymbol{M}'}^{\top}$, if $\boldsymbol{\sigma}_{\boldsymbol{M}} \in \operatorname{prox}_{\phi}(\boldsymbol{\sigma}_{\boldsymbol{M}'})$, then we have $\boldsymbol{U}_{\boldsymbol{M}'}\operatorname{Diag}(\boldsymbol{\sigma}_{\boldsymbol{M}})\boldsymbol{V}_{\boldsymbol{M}'}^{\top} \in \operatorname{prox}_{\Phi}(\boldsymbol{M}')$.*

Function (42) is invariant by permutation of its arguments and verifies $\varnothing \neq \operatorname{dom}\phi_{\hat{f}} \subset [0,+\infty)^N$. Moreover, for any $\boldsymbol{\sigma}' = (\sigma'_1,\dots,\sigma'_N) \in \mathbb{R}^N$,

$$\begin{aligned}\operatorname{prox}_{c\phi_{\hat{f}}}(\boldsymbol{\sigma}') &= \operatorname*{argmin}_{\boldsymbol{\sigma}\in\mathbb{R}^N} cf(\boldsymbol{\sigma}) + \frac{1}{2}\|\boldsymbol{\sigma}-\boldsymbol{\sigma}'\|^2 \\ &= \operatorname*{argmin}_{\boldsymbol{\sigma}\in(0,+\infty)^N} \sum_{\ell=1}^{N}\frac{1}{2}(\sigma_\ell-\sigma'_\ell)^2 - c\log\sigma_\ell.\end{aligned} \tag{43}$$

The above prox calculation requires to minimize a sum of functions, each term acting on a different variable, which means we can minimize each term independently. Let $l \in \{1,\dots,N\}$, then

$$\begin{aligned}&\forall\sigma_\ell > 0 : \frac{d\left(\frac{1}{2}(\sigma_\ell-\sigma'_\ell)^2 - c\log\sigma_\ell\right)}{d\sigma_\ell} = \sigma_\ell - \sigma'_\ell - c\sigma_\ell^{-1} = 0\\ &\iff \sigma_\ell^2 - \sigma'_\ell\sigma_\ell - c = 0\\ &\iff \sigma_\ell \in \left\{\frac{\sigma'_\ell - \sqrt{(\sigma'_\ell)^2+4c}}{2}, \frac{\sigma'_\ell + \sqrt{(\sigma'_\ell)^2+4c}}{2}\right\}.\end{aligned}$$

As $\sigma_\ell \mapsto \frac{1}{2}(\sigma_\ell-\sigma'_\ell)^2 - c\log\sigma_\ell$ diverges toward $+\infty$ in 0^+ and $+\infty$ and is C^1 on $(0,+\infty)$, it must have a minimum where its derivative is equal to 0. The only point where it happens is $\sigma_\ell = \frac{\sigma'_\ell+\sqrt{(\sigma'_\ell)^2+4c}}{2}$, so we can conclude that this point is a global minimum.

Finally, $\operatorname{prox}_{c\phi_{\hat{f}}}$ is uniquely defined for any $\boldsymbol{\sigma}' \in \mathbb{R}^N$ and has the explicit form:

$$\boldsymbol{\sigma} = \operatorname{prox}_{c\phi_{\hat{f}}}(\boldsymbol{\sigma}') = \frac{\boldsymbol{\sigma}' + \sqrt{(\boldsymbol{\sigma}')^2+4c}}{2}$$

where the square and square root operations are applied component-wise. Applying Lemma 5, with $\boldsymbol{M} = \boldsymbol{W}^{[k]}$, $\boldsymbol{M}' = \boldsymbol{W}^{'[k]}$, for every $k \in \{1,\dots,K\}$, and $\Phi = \widehat{f}$, allows to conclude the proof.

APPENDIX D
PROOF OF LEMMA 4

For every $\mathcal{C} \in \mathbb{R}^{K\times K\times N}$, we can rewrite function (16), as $g(\mathcal{C}) = \sum_{n=1}^{N}\widehat{g}(\boldsymbol{C}_n)$ where, for every $\boldsymbol{M} \in \mathbb{R}^{K\times K}$, with singulare values $\boldsymbol{\sigma}_{\boldsymbol{M}} = (\sigma_{\boldsymbol{M},\ell})_{1\leq\ell\leq K}$, $\widehat{g}(\boldsymbol{M}) = \phi_{\hat{g}}(\boldsymbol{\sigma}_{\boldsymbol{M}})$, with

$$\phi_{\hat{g}} : \mathbb{R}^K \to (-\infty,+\infty]$$
$$\boldsymbol{\sigma} = (\sigma_1,\dots,\sigma_K) \mapsto \begin{cases}+\infty \text{ if } (\exists\ell\in\{1,\dots,K\})\ \sigma_\ell \leq 0\\ -\sum_{l=1}^{K}\log\sigma_\ell \text{ otherwise.}\end{cases}$$

Moreover, for any $\boldsymbol{\sigma}' \in \mathbb{R}^K$,

$$\begin{aligned}\operatorname{prox}_{c\phi_{\hat{g}}}(\boldsymbol{\sigma}') &= \operatorname*{argmin}_{\boldsymbol{\sigma}\in[\epsilon,+\infty)} \frac{1}{2}\sum_{\ell=1}^{K}(\sigma_\ell-\sigma'_\ell)^2 - c\log\sigma_\ell\\ &= \max\left(\epsilon, \frac{\boldsymbol{\sigma}' + \sqrt{(\boldsymbol{\sigma}')^2+2c}}{2}\right).\end{aligned}$$

Hence the result, applying Lemma 5 for every $n \in \{1,\dots,N\}$, with $\Phi = \hat{g}$, and $(\boldsymbol{M},\boldsymbol{M}') = (\boldsymbol{C}_n,\boldsymbol{C}'_n)$.

APPENDIX E
CONVERGENCE PROOF

a) Lower-boundedness of the cost function: Using Theorem 1,

$$(\forall(\mathcal{W},\mathcal{C}) \in \mathbb{R}^{N\times N\times K}\times\mathbb{R}^{K\times K\times N})\quad J_{\text{IVA}-\text{G}}(\mathcal{W},\mathcal{C}) \geq \widetilde{J}_{\text{IVA}-\text{G}}(\mathcal{W}) + \frac{KN}{2}.$$

Let us note σ^- the smallest eigenvalue of $\widehat{\boldsymbol{R}}_{\mathbf{x}}$, then $\widehat{\boldsymbol{R}}_{\mathbf{x}} - \sigma^-\boldsymbol{I}_{KN}$ is a symmetric matrix whose eigenvalue are non-negative, i.e., $\widehat{\boldsymbol{R}}_{\mathbf{x}} - \sigma^-\boldsymbol{I}_{KN}$ is symmetric positive. Therefore, it is also the case of $\boldsymbol{W}_n(\widehat{\boldsymbol{R}}_{\mathbf{x}} - \sigma^-\boldsymbol{I}_{KN})\boldsymbol{W}_n^{\top}$ for any $n \in \{1,\dots,N\}$, and

$$\boldsymbol{W}_n\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_n^{\top} \succeq \sigma^-\boldsymbol{W}_n\boldsymbol{W}_n^{\top}, \tag{44}$$

where $\succeq$ denotes the Loewner order relationship defined on $(\mathcal{S}_N)^2$ as $(\forall(\boldsymbol{A},\boldsymbol{B}) \in (\mathcal{S}_N)^2)\ \boldsymbol{A}\succeq\boldsymbol{B} \iff \boldsymbol{A}-\boldsymbol{B} \in \mathcal{S}_N^+$.

Yet, $\boldsymbol{W}_n\boldsymbol{W}_n^{\top}$ is a diagonal matrix whose k-th coefficient is $\|\boldsymbol{w}_n^{[k]}\|^2$. Hence $\det(\boldsymbol{W}_n\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_n^{\top}) \geq (\sigma^-)^K\det(\boldsymbol{W}_n\boldsymbol{W}_n^{\top})$ and

$$\frac{1}{2}\log\det(\boldsymbol{W}_n\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_n^{\top}) \geq \frac{1}{2}K\log(\sigma^-) + \frac{1}{2}\sum_{k=1}^{K}\log\|\boldsymbol{w}_n^{[k]}\|^2. \tag{45}$$

Besides, by Hadamard inequality applied to the lines of $\boldsymbol{W}^{[k]}$ for any $k \in \{1,\dots,K\}$, $|\det\boldsymbol{W}^{[k]}|^2 \leq \prod_{n=1}^{N}\|\boldsymbol{w}_n^{[k]}\|^2$, hence,

$$-\log|\det\boldsymbol{W}^{[k]}| \geq -\frac{1}{2}\sum_{n=1}^{N}\log\|\boldsymbol{w}_n^{[k]}\|^2. \tag{46}$$

By summing (45) for $n \in \{1,\dots,N\}$, and (46) for $k \in \{1,\dots,K\}$, it comes:

$$(\forall\mathcal{W} \in \mathbb{R}^{N\times N\times K})\ \widetilde{J}_{\text{IVA}-\text{G}}(\mathcal{W}) \geq \frac{KN}{2}(1+\log(\sigma^-)). \tag{47}$$

We can conclude that $J_{\text{IVA}-\text{G}}$ is lower bounded on $\mathbb{R}^{N\times N\times K}\times\mathbb{R}^{K\times K\times N}$. Finally, since the quadratic regularization term is positive, $J_{\text{IVA}-\text{G}}^{\text{Reg}}$ is lower bounded too, which ends the proof.

b) Lipschitz continuity of the gradient: The expressions of the partial gradients of h given in Lemmas 1 and 2 show clearly that ∇h is a C^1 function on $\mathbb{R}^{N\times N\times K}\times\mathbb{R}^{K\times K\times N}$. Consequently, the mean value theorem can be applied to show that it has Lipschitz continuity on any bounded domain.

c) Boundedness of the sequence:

• Boundedness of $(\mathcal{C}^{(i)})_{i\in\mathbb{N}}$

First, we notice that (20) defines a norm on $\mathbb{R}^{K\times K\times N}$. For all $(i,j)\in\mathbb{N}^2$ such that $\mathcal{C}^{(i,j)}$ belongs to the sequance generated by **PALM-IVA-G**, let us note:

$$\begin{aligned}\overline{\boldsymbol{C}}_n^{(i,j)} &= \boldsymbol{C}_n^{(i,j)} - c_{\mathcal{C}}(\alpha(\operatorname{Diag}(\boldsymbol{C}_n^{(i,j)}) - \boldsymbol{I}_K)\\ &\quad - \frac{1}{2}\boldsymbol{W}_n^{(i+1)}\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_n^{(i+1)\top}).\end{aligned}$$

$$\begin{aligned}\|\boldsymbol{C}_n^{(i,j+1)}\|_{\mathrm{S}} &= \max\Big(\epsilon, \frac{\|\overline{\boldsymbol{C}}_n^{(i,j)}\|_{\mathrm{S}} + \sqrt{\|\overline{\boldsymbol{C}}_n^{(i,j)}\|_{\mathrm{S}}^2 + 2c_{\mathcal{C}}}}{2}\Big)\\ &\leq \|\overline{\boldsymbol{C}}_n^{(i,j)}\|_{\mathrm{S}} + \sqrt{\frac{c_{\mathcal{C}}}{2}},\end{aligned} \tag{48}$$

assuming that $\epsilon < \sqrt{\frac{c_{\mathcal{C}}}{2}}$, which is verified in our experimental settings.

As $\boldsymbol{W}_n^{(i+1)}\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_n^{(i+1)\top}$ is symmetric positive,

$$\begin{aligned}\|\overline{\boldsymbol{C}_n}^{(i,j)}\|_{\mathrm{S}} &\leq \|\boldsymbol{C}_n^{(i,j)} - \gamma_{\mathcal{C}}(\operatorname{Diag}(\boldsymbol{C}_n^{(i,j)}) - \boldsymbol{I}_K)\|_{\mathrm{S}}\\ &= \|\boldsymbol{C}_n^{(i,j)} - \gamma_{\mathcal{C}}\operatorname{Diag}(\boldsymbol{C}_n^{(i,j)})\|_{\mathrm{S}} + \gamma_{\mathcal{C}}.\end{aligned} \tag{49}$$

Let $\boldsymbol{u}\in\mathbb{R}^K$ be a unit norm vector. Using the inequality for matrices of $\mathcal{S}_K^+, \forall(k,l)\in\{1,\dots,K\}\,|c_{n,k,l}^{(i,j)}| \leq \sqrt{c_{n,k,k}^{(i,j)}}\sqrt{c_{n,l,l}^{(i,j)}}$, it comes

$$\begin{aligned}&\boldsymbol{u}^\top\boldsymbol{C}_n^{(i,j)}\boldsymbol{u}\\ &= \sum_{1\leq k,l\leq K} c_{n,k,l}^{(i,j)}u_k u_l \leq \sum_{1\leq k,l\leq K}|c_{n,k,l}^{(i,j)}||u_k||u_l|\\ &\leq \sum_{1\leq k,l\leq K}\sqrt{c_{n,k,k}^{(i,j)}}|u_k|\sqrt{c_{n,l,l}^{(i,j)}}|u_l|\\ &\leq \sum_{1\leq k,l\leq K}\frac{c_{n,k,k}^{(i,j)}u_k^2}{2} + \frac{c_{n,l,l}^{(i,j)}u_l^2}{2} = K\boldsymbol{u}^\top\operatorname{Diag}(\boldsymbol{C}_n^{(i,j)})\boldsymbol{u}.\end{aligned}$$

It follows that, for all $\boldsymbol{u}\in\mathbb{R}^K$ such that $\|\boldsymbol{u}\| = 1$, $\boldsymbol{u}^\top\Big(\boldsymbol{C}_n^{(i,j)} - \gamma_{\mathcal{C}}\operatorname{Diag}(\boldsymbol{C}_n^{(i,j)})\Big)\boldsymbol{u} \leq (1-\frac{\gamma_{\mathcal{C}}}{K})\boldsymbol{u}^\top\boldsymbol{C}_n^{(i,j)}\boldsymbol{u} \leq (1-\frac{\gamma_{\mathcal{C}}}{K})\|\boldsymbol{C}_n^{(i,j)}\|_{\mathrm{S}}$. Combining this result with (48) and (49), we deduce that $\|\boldsymbol{C}_n^{(i,j+1)}\|_{\mathrm{S}} \leq (1-\frac{\gamma_{\mathcal{C}}}{K})\|\boldsymbol{C}_n^{(i,j)}\|_{\mathrm{S}} + \gamma_{\mathcal{C}} + \sqrt{\frac{c_{\mathcal{C}}}{2}}$, then we can prove by recurrence that for all (i,j), $\|\boldsymbol{C}_n^{(i,j)}\|_{\mathrm{S}} \leq \max\big(\|\boldsymbol{C}_n^{(0)}\|_{\mathrm{S}}, K(1+\sqrt{\frac{1}{2\alpha\gamma_{\mathcal{C}}}})\big)$. Consequently, if we define

$$\bar{\rho} = \max\big(\|\rho_{C^{(0)}}\|_{\mathrm{S}}, K(1+\sqrt{\frac{1}{2\alpha\gamma_{\mathcal{C}}}})\big), \tag{50}$$

$\bar{\rho}$ is an upper bound for $(\rho_{\mathcal{C}^{(i)}})_{i\in\mathbb{N}}$, which proves that $(\mathcal{C}^{(i)})_{i\in\mathbb{N}}$ is bounded.

• Boundedness of $(\mathcal{W}^{(i)})_{i\in\mathbb{N}}$

Again, we fix $i\in\mathbb{N}$, the proximal operator used in Algorithm 1 ensures that $\mathcal{C}^{(i)} \in (\epsilon\boldsymbol{I}_K + \mathcal{S}_K^+)^N$. Thus, $(\forall n\in\{1,\dots,N\})\ \operatorname{tr}(\boldsymbol{C}_n^{(i)}\boldsymbol{W}_n^{(i)}\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_n^{(i)\top}) \geq \epsilon\operatorname{tr}(\boldsymbol{W}_n^{(i)}\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_n^{(i)\top}) \geq \epsilon\sigma^-\|\boldsymbol{W}_n^{(i)}\|^2$ using (44). By summing these inequalities,

$$\frac{1}{2}\sum_{n=1}^N\operatorname{tr}(\boldsymbol{C}_n^{(i)}\boldsymbol{W}_n^{(i)}\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_n^{(i)\top}) \geq \frac{\epsilon\sigma^-}{2}\|\mathcal{W}^{(i)}\|^2. \tag{51}$$

Besides, let us consider the concave inequality on the logarithm function at point $\frac{2}{\epsilon\sigma^-} > 0$,

$$(\forall x\in(0,+\infty)) \quad \log x \leq \log\frac{2}{\epsilon\sigma^-} + \frac{\epsilon\sigma^-}{2}x - 1. \tag{52}$$

Taking $k\in\{1,\dots,K\}$ and summing (52) applied on $\|\boldsymbol{w}_n^{(i)[k]}\|$ for $n\in\{1,\dots,N\}$, we obtain

$$\sum_{n=1}^N\log\|\boldsymbol{w}_n^{(i)[k]}\| \leq N(\log\frac{2}{\epsilon\sigma^-} - 1) + \frac{\epsilon\sigma^-}{2}\|\boldsymbol{W}^{(i)[k]}\|^2. \tag{53}$$

And combining this inequality with (45) gives

$$-\log|\det\boldsymbol{W}^{(i)[k]}| \geq -\frac{\epsilon\sigma^-}{4}\|\boldsymbol{W}^{(i)[k]}\|^2 - \frac{N}{2}(\log\frac{2}{\epsilon\sigma^-} - 1).$$

Finally, summing for $k\in\{1,\dots,K\}$,

$$-\sum_{k=1}^K\log|\det\boldsymbol{W}^{(i)[k]}| \geq -\frac{\epsilon\sigma^-}{4}\|\mathcal{W}^{(i)}\|^2 - \frac{KN}{2}(\log\frac{2}{\epsilon\sigma^-} - 1). \tag{54}$$

Summing (54) and (51) yields

$$\begin{aligned}&\frac{\epsilon\sigma^-}{4}\|\mathcal{W}^{(i)}\|^2 - \frac{KN}{2}(\log\frac{2}{\epsilon\sigma^-} - 1)\\ &\leq \frac{1}{2}\sum_{n=1}^N\operatorname{tr}(\boldsymbol{C}_n^{(i)}\boldsymbol{W}_n^{(i)}\widehat{\boldsymbol{R}}_{\mathbf{x}}\boldsymbol{W}_n^{(i)\top}) - \sum_{k=1}^K\log|\det\boldsymbol{W}^{(i)[k]}|\\ &= J_{\mathrm{IVA-G}}(\mathcal{W}^{(i)},\mathcal{C}^{(i)}) + \frac{1}{2}\sum_{n=1}^N\log\det\boldsymbol{C}_n^{(i)}\\ &\leq J_{\mathrm{IVA-G}}(\mathcal{W}^{(i)},\mathcal{C}^{(i)}) + \frac{KN}{2}\log\bar{\rho}.\end{aligned}$$

Following the proof of [28, Lemma 2], it is sufficient that the first four items of Assumption 1 hold to obtain that $J_{\mathrm{IVA-G}}^{\mathrm{Reg}}(\mathcal{W}^{(i)},\mathcal{C}^{(i)})$ is a non-increasing sequence. We thus have

$$\|\mathcal{W}^{(i)}\|^2 \leq \frac{4}{\epsilon\sigma^-}\Big(J_{\mathrm{IVA-G}}^{\mathrm{Reg}}(\mathcal{W}^{(0)},\mathcal{C}^{(0)}) + \frac{KN}{2}\log\bar{\rho}(\frac{2}{\epsilon\sigma^-} - 1)\Big). \tag{55}$$

This shows that $(\mathcal{W}^{(i)})_{i\in\mathbb{N}}$ is bounded too.

d) Bounds for the corrected Lipschitz moduli: Using (18), we can set $\lambda_{\mathcal{W}}^- = \epsilon\varrho_{\widehat{\boldsymbol{R}}_{\mathbf{x}}}$ and $\lambda_{\mathcal{W}}^+ = \bar{\rho}\varrho_{\widehat{\boldsymbol{R}}_{\mathbf{x}}}$. We can also set $\lambda_{\mathcal{C}}^- = \lambda_{\mathcal{C}}^+ = \alpha$.

e) KL property: KL property is a key tool from functional analysis to demonstrate convergence of iterates in the non-convex setting [28]. It is sufficient here to apply the result from [32, Section 4.3], which states that a function verifies KL on the domain of its subdifferential, if it is lower semi-continuous, proper, and definable in an o-minimal structure.

We rely on the o-minimal structure $\mathfrak{S}(\mathbb{R}_{\mathrm{an,exp}})$, defined in [33, Section 2, Example (6)], which contains $\log, \exp$ and all semi-algebraic functions. The properties in [33, Section 5] can then be used to show that $\mathfrak{S}(\mathbb{R}_{\mathrm{an,exp}})$ contains our cost function $J_{\mathrm{IVA-G}}^{\mathrm{Reg}}$, after identifying $\mathbb{R}^{N\times N\times K}\times\mathbb{R}^{K\times K\times N}$ with $\mathbb{R}^{KN(K+N)}$.

ACKNOWLEDGMENT

The work of E.C. and C.C. is funded by the European Research Council Starting Grant MAJORIS ERC-2019-STG-850925.

REFERENCES

[1] V. Calhoun and T. Adali, "Unmixing fMRI with independent component analysis," *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 2, pp. 79–90, Mar. 2006.

[2] Z. Luo, C. Li, and L. Zhu, "A comprehensive survey on blind source separation for wireless adaptive processing: Principles, perspectives, challenges and new research directions," *IEEE Access*, vol. 6, pp. 66 685–66 708, 2018.

[3] M. Castella, J.-C. Pesquet, and A. Petropulu, "A family of frequency- and time-domain contrasts for blind separation of convolutive mixtures of temporally dependent signals," *IEEE Transactions on Signal Processing*, vol. 53, no. 1, pp. 107–120, 2005.

[4] A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Transactions on Image Processing*, vol. 11, no. 3, pp. 293–305, Mar. 2002, conference Name: IEEE Transactions on Image Processing.

[5] I. Santamaria, "Handbook of blind source separation: independent component analysis and applications (Comon, P. and Jutten, C. ; 2010 [Book Review])," *IEEE Signal Processing Magazine*, vol. 30, no. 2, pp. 133–134, 2013.

[6] T. Adalı, M. Anderson, and G.-S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Proc. Mag.*, vol. 31, no. 3, pp. 18–33, May 2014.

[7] V. D. Calhoun and T. Adali, "Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery," *IEEE Reviews in Biomedical Engineering*, vol. 5, pp. 60–73, 2012.

[8] R. García-Bermúdez, F. R. Ruiz, J. G. Peñalver, O. V. Cansino, L. V. Pérez, C. Torres, and R. Becerra-García, "Evaluation of electro-oculography data for ataxia sca-2 classification: A blind source separation approach," in *2010 10th International Conference on Intelligent Systems Design and Applications*, 2010, pp. 237–241.

[9] B. Gabrielson, M. A. B. S. Akhonda, S. Bhinge, J. Brooks, Q. Long, and T. Adali, "Joint-IVA for identification of discriminating features in EEG: Application to a driving study," *Biomedical Signal Processing and Control*, vol. 61, p. 101948, Aug. 2020.

[10] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[11] Z. Boukouvalas, M. Puerto, D. C. Elton, P. W. Chung, and M. D. Fuge, "Independent vector analysis for molecular data fusion: Application to property prediction and knowledge discovery of energetic materials," in *2020 28th European Signal Processing Conference (EUSIPCO)*. Amsterdam, Netherlands: IEEE, Jan. 2021, pp. 1030–1034.

[12] M. Anderson, X.-L. Li, and T. Adalı, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Trans. Signal Processing*, vol. 60, no. 4, pp. 2049–2055, April 2012.

[13] J.-C. Pesquet and E. Moreau, "Cumulant-based independence measures for linear mixtures," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1947–1956, 2001.

[14] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[15] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adali, "Independent vector analysis: Identification conditions and performance bounds," *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4399–4410, 2014.

[16] T. Adali, M. Akhonda, and V. D. Calhoun, "ICA and IVA for data fusion: An overview and a new approach based on disjoint subspaces," *IEEE sensors letters*, vol. 3, no. 1, p. 7100404, Jan. 2019.

[17] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: Definition and algorithms," in *Proceedings of the 40th Asilomar Conference on Signals, Systems and Computers (ASILOMAR 2006)*, 2006, pp. 1393–1396.

[18] V. D. Calhoun and T. Adali, "Feature-based fusion of medical imaging data," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 5, pp. 711–720, 2009.

[19] A. Michael, M. Anderson, R. Miller, T. Adali, and V. Calhoun, "Preserving subject variability in group fMRI analysis: Performance evaluation of GICA vs. IVA," *Frontiers in systems neuroscience*, vol. 8, p. 106, 06 2014.

[20] J. Laney, K. P. Westlake, S. Ma, E. Woytowicz, V. D. Calhoun, and T. Adalı, "Capturing subject variability in fMRI data: A graph-theoretical analysis of GICA vs. IVA," *Journal of neuroscience methods*, vol. 247, pp. 32–40, 2015.

[21] E. Chouzenoux, T. T.-K. Lau, C. Lefort, and J.-C. Pesquet, "Optimal multivariate Gaussian fitting with applications to PSF modeling in two-photon microscopy imaging," *Journal of Mathematical Imaging and Vision*, vol. 61, no. 7, pp. 1037–1050, Sept. 2019.

[22] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed. New York: Springer, 2017.

[23] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Bauschke, H. Burachik, R. Combettes, P. Elser, V. Luke, D. Wolkowicz, and H. (Eds.), Eds. Springer, 2011, pp. 185–212.

[24] A. Benfenati, E. Chouzenoux, and J.-C. Pesquet, "Proximal approaches for matrix optimization problems: Application to robust precision matrix estimation." *Signal Processing*, no. 69, p. 107417, 2020.

[25] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "A block coordinate variable metric forward–backward algorithm," *Journal of Global Optimization*, vol. 66, no. 3, pp. 457–485, 2016.

[26] L. T. K. Hien, D. Phan, and N. Gillis, "An inertial block majorization minimization framework for nonsmooth nonconvex optimization," *Journal of Machine Learning Research*, vol. 24, pp. 1–41, 2023.

[27] E. Chouzenoux, M.-C. Corbineau, J.-C. Pesquet, and G. Scrivanti, "A variational approach for joint image recovery and feature extraction based on spatially varying generalised Gaussian models," *Journal of Mathematical Imaging and Vision*, pp. 1–22, 2024.

[28] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1, pp. 459–494, 2014.

[29] S. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems*, vol. 8. MIT Press, 1995.

[30] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Oct. 2008, version 20081110.

[31] J.-C. Pesquet, "Prox of a function of the singular values of a matrix," internal report, CVN, CentraleSupélec, 2022.

[32] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.

[33] L. van den Dries and C. Miller, "Geometric categories and o-minimal structures," *Duke Mathematical Journal*, vol. 84, no. 2, Aug. 1996.