Joint decomposition of multichannel signals with multimarginal optimal transport

Jean Mallejac, Jérôme Idier, Charles Soussen, Xiangyi Li

Joint decomposition of multichannel signals with multimarginal optimal transport

Jean Malléjac[1], Jérôme Idier[2], Charles Soussen[1], Xiangyi Li[3]

[1]*Université Paris-Saclay, CNRS, CentraleSupélec, L2S,* F-91192 Gif-sur-Yvette, France
[2]*Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004,* F-44000 Nantes, France
[3]*Aix Marseille Université, CNRS, Centrale Marseille, Institut Fresnel,* F-13013 Marseille, France
jean.mallejac@centralesupelec.fr, jerome.idier@ls2n.fr, charles.soussen@centralesupelec.fr, xiangyi.li@fresnel.fr

***Abstract*—The joint decomposition of a sequence of signals is addressed, where each signal is coded within a predefined dictionary. We propose a convex variational formulation involving a regularization term based on optimal transport. The latter is designed so as to promote proximity between neighboring channels. Two kinds of regularization criteria are designed, corresponding to distinct cross-channel information on the transport of dictionary atoms from one channel to another. The resulting optimization problems are reformulated as quadratic programs and solved using a proximal algorithm. The proposed approach is analyzed for a multichannel deconvolution problem.**

***Index Terms*—Inverse problem, multichannel decomposition, optimal transport, proximal algorithm.**

I. Introduction

The recording of multichannel signals, corresponding to sequences of signals evolving in time or space occurs in many application fields. Such data can be found in, *e.g.*, spectroscopy, seismology, and hyperspectral imaging [1], [2]. In this paper, channels refer to sensors capturing a physical process at different times or spatial locations. For instance, in hyperspectral imaging, each channel corresponds to a space location (pixel). The related signal depends on the wavelength, and can be analyzed using sparse decomposition in a dictionary containing all possible spectral features [2].

Hereafter, the dictionary is assumed to be known and designed from expert knowledge. When dealing with monovariate signals, sparse decomposition can be carried out using ℓ_0 or ℓ_1 minimization. For strongly correlated dictionaries, the sparse decomposition problem becomes ill-posed, and extra regularization information is needed to improve accuracy.

In the multichannel setting, one aims at jointly decomposing the set of observed signals. Multichannel decomposition can be addressed by leveraging the correlation between neighboring channels. Variational techniques consist of minimizing a penalized least-squares cost function. Several convex penalties can be used, corresponding to various priors on the trajectories of active atoms across channels. Using mixed ℓ_1-ℓ_2 norms is classical to promote simultaneous atom decomposition [3]. However, simultaneous sparsity appears to be restrictive. Extended group sparsity techniques have been proposed to deal with atom decomposition that evolves from one channel to another. The latter is based on the design of groups of neighboring atoms [4]. A limitation of this approach is that the group sparsity structure should be empirically defined.

Optimal Transport (OT) has emerged as a powerful tool in machine learning [5], where one needs to quantify the divergence between probability measures. In inverse problems, a recent trend is to use OT metrics to design non-parametric convex penalties [6], [7] that appear to be more flexible than mixed norm criteria. Apart from flexibility, the attractiveness of OT divergences comes from the fact that their computation amounts to solving a linear program, and fast solvers are available. OT techniques have been explored for multi-task regression in neuroimaging [8] and online signal tracking [9].

We propose an offline approach to decompose slowly evolving signals in a dictionary, with prior knowledge on inter-channel correlation. We design two regularizers based on *(i)* optimal transport between two consecutive channels, and *(ii)* multimarginal OT (MMOT) to deal with an extended channel neighborhood. Note that the MMOT concept was first exploited in [10] to deal with sensor fusion issues.

The proposed method is presented in Sect. II. The related optimization problems are solved using a first order proximal algorithm [11] in Sect. III. Then, the proposed algorithms are assessed on numerical simulations in Sect. IV.

II. Proposed method

We consider a multichannel signal seen as a sequence of monovariate signals $(\boldsymbol{y}_n)_{n\in\mathcal{E}_N}$ with N the number of channels and $\mathcal{E}_N$ the set of integers $[\![1, N]\!]$. Each signal $\boldsymbol{y}_n \in \mathbb{R}^K$ in the sequence is represented in a known dictionary $A \in \mathbb{R}^{K\times M}$, which is a collection of M elementary signals called *atoms*. Gaussian i.i.d noise, denoted $\boldsymbol{n}_n \in \mathbb{R}^K$, is added such that:

$$\forall n \in \mathcal{E}_N, \quad \boldsymbol{y}_n = A\boldsymbol{x}_n + \boldsymbol{n}_n.$$

The inverse problem aims to retrieve the decomposition weights $(\boldsymbol{x}_n)_{n\in\mathcal{E}_N}$ from observations $(\boldsymbol{y}_n)_{n\in\mathcal{E}_N}$ and dictionary A. In this paper, we assume that $\boldsymbol{x}_n$ are non-negative signals that satisfy the sum-to-one assumption:

$$\forall n \in \mathcal{E}_N, \quad \boldsymbol{x}_n \in \mathbb{R}^M_+, \quad \mathbf{1}^\top \boldsymbol{x}_n = 1, \tag{1}$$

with $\mathbf{1}$ the column vector of ones of size M. Such an assumption is rather restrictive, but it is a common one in hyperspectral unmixing [2]. We further assume a slow

evolution of the sequence of signals $\boldsymbol{x}_n$ with respect to the channel index n. This assumption holds when two consecutive channels induce gradual (continuous) changes of the data acquisition with respect to time or space.

Hereafter, $\boldsymbol{x}_n$ will be interpreted as a discrete probability measure. Furthermore, the slow evolution of $\boldsymbol{x}_n$ will be formulated using optimal transport metrics.

A. First-order method

The slow evolution of the sequence $(\boldsymbol{x}_n)_{n\in\mathcal{E}_N}$ induces a correlation between $\boldsymbol{x}_n$ and its neighbors $\boldsymbol{x}_{n-1}$, $\boldsymbol{x}_{n+1}$. We take advantage of this correlation by considering a regularization between $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n+1}$ in a joint framework:

$$\min_{\substack{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\in\mathbb{R}_+^M\\ \forall n,\ \mathbf{1}^\top\boldsymbol{x}_n=1}} \sum_{n=1}^{N} \|\boldsymbol{y}_n - A\boldsymbol{x}_n\|_2^2 + \omega \sum_{n=1}^{N-1} \mathcal{OT}_1(\boldsymbol{x}_n, \boldsymbol{x}_{n+1}). \quad (2)$$

The OT metric quantifies the degree of proximity between measures $\boldsymbol{x}_n$. Generally speaking, OT aims to find the joint probability distribution P_n between two probability measures μ_1 and μ_2 that minimizes a cost function such as:

$$\mathcal{OT}_1(\mu_1,\mu_2) = \min_{P\in\mathbb{R}_+^{M\times M}} \langle P, C\rangle_F, \quad (3a)$$

$$\textbf{s.t.} \quad P\mathbf{1} = \mu_1, \quad P^\top\mathbf{1} = \mu_2, \quad (3b)$$

with $\langle\cdot,\cdot\rangle_F$ the Frobenius inner product [5]. Hereafter, we set:

$$\forall i,j\in\mathcal{E}_M, \quad C_{ij} = (i-j)^2, \quad (4)$$

thus $\mathcal{OT}_1$ identifies with the squared 2-Wasserstein distance. By plugging (3) within (2), we get the following quadratic (convex) formulation:

$$\min_{\substack{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\in\mathbb{R}_+^M\\ P_1,\dots,P_{N-1}\in\mathbb{R}_+^{M\times M}}} \sum_{n=1}^{N} \|\boldsymbol{y}_n - A\boldsymbol{x}_n\|_2^2 + \omega \sum_{n=1}^{N-1} \langle P_n, C\rangle_F, \quad (5a)$$

$$\textbf{s.t.} \quad \forall n\in\mathcal{E}_N, \quad \mathbf{1}^\top\boldsymbol{x}_n = 1, \quad (5b)$$

$$\forall n\in\mathcal{E}_{N-1}, \quad P_n\bar{\times}_2\mathbf{1} = \boldsymbol{x}_n, \quad P_n\bar{\times}_1\mathbf{1} = \boldsymbol{x}_{n+1}. \quad (5c)$$

Note that constraints (3b) and (5c) are identical up to a notation change. Here, we introduce the k-mode vector product $\bar{\times}_k$ to marginalize out the k-th dimension, that is, across the rows and columns for $k=1$ and 2, respectively.

The effect of the $\mathcal{OT}_1$ penalty is illustrated in Fig. 1. Here, A is a convolutive dictionary with Gaussian shaped atoms and $\boldsymbol{x}_n$ are sparse sequences. $A\boldsymbol{x}_n$ identifies with a convolution product $h * \boldsymbol{x}_n$ where h refers to the Gaussian impulse response. One can notice that $\mathcal{OT}_1(\boldsymbol{x}_n, \boldsymbol{x}_{n+1})$ is strongly related to the horizontal shifts $|i-j|$ between the corresponding spikes within $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n+1}$, at respective positions i and j [5]. According to (4), C_{ij} increases with $|i-j|$, therefore $\mathcal{OT}_1(\boldsymbol{x}_n, \boldsymbol{x}_{n+1})$ increases with the speed of displacement of spikes. Therefore, when ω is large, the $\mathcal{OT}_1$ penalty tends to prevent strong mass displacements from $\boldsymbol{x}_n$ to $\boldsymbol{x}_{n+1}$.

The above method will be referred to as first-order since it prevents lateral displacements of the spike locations. Indeed, the method penalizes the (discrete) first-order derivatives (over channels) of the spike location trajectories.

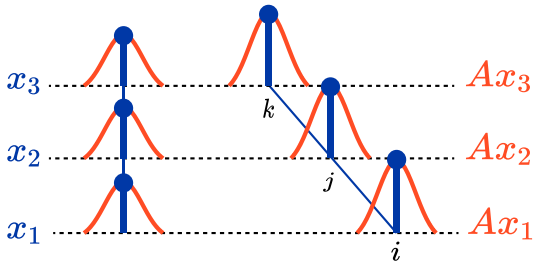

Fig. 1. As spikes move between $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ (on the right), $\mathcal{OT}_1(\boldsymbol{x}_1, \boldsymbol{x}_2) \neq 0$ and $\mathcal{OT}_1(\boldsymbol{x}_2, \boldsymbol{x}_3) \neq 0$. However, $\mathcal{OT}_2(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = 0$ as they are aligned (locally uniform displacement speed).

B. Second-order method

In order to make the OT-based method more flexible and to allow arbitrary (unknown) displacement speed between the spikes, we propose to replace $\mathcal{OT}_1$ by a second-order penalty $\mathcal{OT}_2$. The latter considers an extended neighborhood defined by the triplet $\boldsymbol{x}_n, \boldsymbol{x}_{n+1}$ and $\boldsymbol{x}_{n+2}$, favoring uniform displacements of the 3D spikes within the local time interval $[\![n, n+2]\!]$. The resulting cost function reads:

$$\min_{\substack{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\in\mathbb{R}_+^M\\ \forall n,\ \mathbf{1}^\top\boldsymbol{x}_n=1}} \sum_{n=1}^{N} \|\boldsymbol{y}_n - A\boldsymbol{x}_n\|_2^2 + \omega \sum_{n=1}^{N-2} \mathcal{OT}_2(\boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}) \quad (6)$$

$\mathcal{OT}_2$ is a MMOT functional [5, Chap. 10] aiming to retrieve the tridimensional joint probability distribution related to $\boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}$, seen as marginal distributions. In what follows, we will denote by $\mathbb{P}$ the 3D tensor that minimizes the multimarginal cost function related to three probability measures μ_1, μ_2, μ_3:

$$\mathcal{OT}_2(\mu_1,\mu_2,\mu_3) = \min_{\mathbb{P}\in\mathbb{R}_+^{M\times M\times M}} \langle \mathbb{P}, \mathbb{C}\rangle_F, \quad (7a)$$

$$\textbf{s.t.} \quad \mathbb{P}\bar{\times}_{2,3}(\mathbf{1}\mathbf{1}^\top) = \mu_1, \quad (7b)$$

$$\mathbb{P}\bar{\times}_{1,3}(\mathbf{1}\mathbf{1}^\top) = \mu_2, \quad (7c)$$

$$\mathbb{P}\bar{\times}_{1,2}(\mathbf{1}\mathbf{1}^\top) = \mu_3. \quad (7d)$$

The (k,ℓ)-mode vector product, denoted $\bar{\times}_{k,\ell}$, is used to marginalize out both dimensions k and ℓ simultaneously. We propose to define the 3D cost tensor $\mathbb{C}$ as:

$$\forall i,j,k\in\mathcal{E}_M, \quad \mathbb{C}_{i,j,k} = (2j-i-k)^2. \quad (8)$$

The rationale behind this choice is that i, j and k correspond to the spike locations for measures μ_1, μ_2, μ_3. One can easily see that $\mathbb{C}_{ijk} = 0$ iff $j = \frac{i+k}{2}$, so $\mathbb{C}_{i,j,k} = 0$ whenever a spike trajectory is linear, regardless of the speed of the trajectory.

We refer to problem (6) a second-order version of (2) since the cost matrix $\mathbb{C}_{i,j,k}$ promotes linear trajectories of the spike locations from one channel to another. In other words, the discrete second-order derivatives (across channels) of the spike location trajectories are small. Fig. 1 illustrates the fact that $\mathcal{OT}_1$ penalizes the horizontal displacement of probability

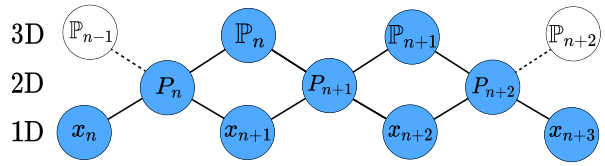

Fig. 2. $\mathbb{P}_n$ is the 3D joint distribution between probability measures $\boldsymbol{x}_n, \boldsymbol{x}_{n+1}$ and $\boldsymbol{x}_{n+2}$. The 2D joint distribution P_n is obtained from $\mathbb{P}_n$ by marginalizing out the third dimension. It is also obtained from $\mathbb{P}_{n-1}$ by marginalizing out the first dimension.

measures displacement whereas $\mathcal{OT}_2$ penalizes local transport that is nonlinear.

Similar to Sect. II-A, the minimization problem (7) related to the computation of $\mathcal{OT}_2$ is incorporated in the target problem (6). The resulting minimization problem (9) is obtained. It involves a set of 1D variables $\boldsymbol{x}_n$, 2D latent variables P_n representing joint distributions between two consecutive measures, and 3D variables $\mathbb{P}_n$ coding for joint distributions between three consecutive measures:

$$\min_{\substack{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n \in \mathbb{R}_+^M \\ P_1,\dots,P_{N-1} \in \mathbb{R}_+^{M\times M} \\ \mathbb{P}_1,\dots,\mathbb{P}_{N-2} \in \mathbb{R}_+^{M\times M\times M}}} \sum_{n=1}^{N} \|\boldsymbol{y}_n - A\boldsymbol{x}_n\|_2^2 + \omega \sum_{n=1}^{N-2} \langle \mathbb{P}_n, \mathbb{C} \rangle_F, \quad (9a)$$

$$\textbf{s.t.} \quad \forall n \in \mathcal{E}_N, \quad \mathbf{1}^\top \boldsymbol{x}_n = 1, \quad (9b)$$

$$\forall n \in \mathcal{E}_{N-1}, \quad P_n \bar{\times}_2 \mathbf{1} = \boldsymbol{x}_n, \quad P_n \bar{\times}_1 \mathbf{1} = \boldsymbol{x}_{n+1}, \quad (9c)$$

$$\forall n \in \mathcal{E}_{N-2}, \quad \mathbb{P}_n \bar{\times}_3 \mathbf{1} = P_n, \quad \mathbb{P}_n \bar{\times}_1 \mathbf{1} = P_{n+1}. \quad (9d)$$

The three-stage structure linking 1D, 2D and 3D variables is illustrated on Fig. 2. The constraint (9c) is applied to enforce that the marginal distributions of 2D probability measures P_n identify with the 1D measures $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n+1}$. Similarly, (9d) is a consistency constraint that forces the marginal distribution of the 3D probability measure $\mathbb{P}_n$ to identify with the 2D joint distributions P_n and P_{n+1}. One can further notice that (9b)-(9c) identify with (5b)-(5c).

C. Fast approximate implementation

The first and second-order methods defined above rely on convex optimization problems. However, these problems are obviously high-dimensional. Here, we propose a slight adaptation aiming to reduce the number of variables involved in the optimization problems. In the proposed adaptation, a maximal support proximity constraint is applied, impeding mass transfer over a certain horizontal range. We denote by $\mathcal{B}_d = \{(i,j) \in [\![1,M]\!]^2, |i-j| \leq d\}$ a 2D band where $d \in \mathbb{N}$ is a user predefined parameter. In order to restrict all admissible spike displacements to belong to $\mathcal{B}_d$, we set the cost matrix to with:

$$C^\infty_{i,j} = \begin{cases} C_{i,j}, & \text{if } (i,j) \in \mathcal{B}_d \\ \infty, & \text{otherwise} \end{cases}$$

Proposition 1 *Setting $C = C^\infty$, the penalty term in problem* (3) *is finite iff $\forall (i,j) \notin \mathcal{B}_d$, $P_{i,j} = 0$. Thus, the penalized-OT problem* (3) *only depends on $2d+1$ diagonals of P. This amounts to $O(dM)$ scalar variables instead of $O(M^2)$.*

The same idea applies for $\mathcal{OT}_2$, with a super-band tensor

$$\mathbb{C}^\infty_{i,j,k} = \begin{cases} \mathbb{C}_{i,j,k}, & \text{if } (i,j,k) \in \mathcal{SB}_d \\ \infty, & \text{otherwise} \end{cases}$$

where $\mathcal{SB}_d = \{(i,j,k) \in [\![1,M]\!]^3, |2j-i-k| \leq d\}$.

Proposition 2 *Setting $\mathbb{C} = \mathbb{C}^\infty$, the OT-penalty term in problem* (7) *is finite iff $\forall (i,j,k) \notin \mathcal{SB}_d$, $\mathbb{P}_{i,j,k} = 0$. Thus, the OT-penalized problem is restricted to a limited number of variables, corresponding to the super-band $\mathcal{SB}_d$ of $\mathbb{P}$. It involves solving $O(dM^2)$ variables instead of $O(M^3)$.*

Hereafter, we will solve problems (5) and (9) using cost matrix C^∞ and cost tensor $\mathbb{C}^\infty$, respectively. The algorithms are described though in the general case where the entries of C and $\mathbb{C}$ are arbitrary (*i.e.*, finite or infinite valued).

III. Primal dual implementation

Both optimization problems (5) and (9) are convex quadratic problems, because the related cost function is a separable sum of quadratic functionals depending on $\boldsymbol{x}_n$ and P_n (respectively, $\mathbb{P}_n$), and their constraints are linear.

Hereafter, we solve both problems using a proximal first-order scheme. The Condat-Vu algorithm is a primal-dual algorithm dedicated to the following type of problems:

$$\min_{\boldsymbol{z}} \quad h(\boldsymbol{z}) + f(\boldsymbol{z}) + \sum_{i=0}^{I-1} g_i(L_i \boldsymbol{z}), \quad (10)$$

where h is convex and differentiable, f and g_i are convex, and L_i are linear operators.

A. Implementation of the first-order method

Let us denote by $\boldsymbol{z} \in \mathbb{R}_+^Z$ an unfolded variable gathering of all variables $\boldsymbol{x}_n$ and P_n involved in the target problem (5):

$$\boldsymbol{z} = \{(\boldsymbol{x}_n, n \in \mathcal{E}_N); (P_n, n \in \mathcal{E}_{N-1})\}.$$

Proposition 3 *Problem* (5) *can be reformulated in the Condat-Vu framework with $I = 2$ and:*

- *h is the quadratic cost function appearing in* (5a)*,*
- *$f = \iota_{\mathbb{R}_+^Z}$ for non-negativity constraints in* (5)*, where ι refers to the indicator function,*
- *$g_1 = \iota_{\{\mathbf{1}\}}$ and L_1 is the linear operator appearing in* (5b)*, which returns $\mathbf{1}^\top \boldsymbol{x}_n$ for all n,*
- *$g_0 = \iota_{\{\mathbf{0}\}}$ and L_0 is the linear operator induced by* (5c)*.*

The Condat-Vu algorithm (in the simplest form, that is, without overrelaxation) is sketched in Algorithm 1 where $\boldsymbol{u}_i^{(k)}$ refer to dual variables at iteration k. It involves three hyperparameters τ and σ_0, σ_1 related to the step sizes in the primal and dual domains, respectively. When dealing with problem (5), the proximal operators appearing in the Condat-Vu algorithm ($\text{prox}_{\tau f}$, $\text{prox}_{\sigma_i g_i^*}$) have simple closed form expressions, leading to Algorithm 1 where $(\boldsymbol{z})_+ := \max(\boldsymbol{z}, \mathbf{0})$.

The primal and dual gradient steps τ, σ_0 and σ_1 should verify the condition expressed in [11, Th. 9.7] to ensure the

Algorithm 1 Condat-Vu (unrelaxed) implementation of (5).

for $k = 1, 2, \dots$ **do**

$$\boldsymbol{z}^{(k+1)} = \left(\boldsymbol{z}^{(k)} - \tau \nabla h(\boldsymbol{z}^{(k)}) - \tau \sum_{i=0}^{1} L_i^\top \boldsymbol{u}_i^{(k)}\right)_+$$

$$\boldsymbol{u}_0^{(k+1)} = \boldsymbol{u}_0^{(k)} + \sigma_0 L_0 \left(2\boldsymbol{z}^{(k+1)} - \boldsymbol{z}^{(k)}\right)$$

$$\boldsymbol{u}_1^{(k+1)} = \boldsymbol{u}_1^{(k)} + \sigma_1 \left(L_1 \left(2\boldsymbol{z}^{(k+1)} - \boldsymbol{z}^{(k)}\right) - \mathbf{1}\right)$$

end for

Algorithm 2 Condat-Vu (unrelaxed) implementation of (9). We set $\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_2 = \mathbf{0}$, $\boldsymbol{\alpha}_1 = \mathbf{1}$.

for $k = 1, 2, \dots$ **do**

$$\boldsymbol{z}^{(k+1)} = \left(\boldsymbol{z}^{(k)} - \tau \nabla h(\boldsymbol{z}^{(k)}) - \tau \sum_{i=0}^{2} L_i^\top \boldsymbol{u}_i^{(k)}\right)_+$$

for $i = [\![0, 2]\!]$ **do**

$$\boldsymbol{u}_i^{(k+1)} = \boldsymbol{u}_i^{(k)} + \sigma_i \left(L_i \left(2\boldsymbol{z}^{(k+1)} - \boldsymbol{z}^{(k)}\right) - \boldsymbol{\alpha}_i\right)$$

end for

end for

convergence of iterates in the Condat-Vu algorithm. Let us now specify this condition in the context of Problem (5).

Proposition 4 *The conditions for convergence [11, Th. 9.7] are met for Algorithm 1 as soon as:*

$$\tau \leq \left(2\left\|A^\top A\right\| + \sigma_1 M + 2\sigma_0(M+1)\right)^{-1}.$$

where $\|\cdot\|$ is the spectral norm of a matrix.

SKETCH OF PROOF: The condition in [11, Th. 9.7] reads:

$$\tau \leq \left\|Q + \sum_{i=0}^{I-1} \sigma_i L_i^\top L_i\right\|^{-1},$$

with Q the Hessian of h. Using the triangle inequality, we get a slightly more restrictive condition:

$$\tau \leq \left(\|Q\| + \sum_{i=0}^{I-1} \sigma_i \left\|L_i^\top L_i\right\|\right)^{-1}.$$

From (5a), we get $\|Q\| = 2\left\|A^\top A\right\|$. Moreover, one can show that $\left\|L_1^\top L_1\right\| = M$ and $\left\|L_0^\top L_0\right\| \leq 2(M+1)$.

A complete proof can be found in [12].

B. Implementation of the second-order method

Similar to Sect. III-A, we denote by $\boldsymbol{z} \in \mathbb{R}_+^Z$ a column vector gathering of all variables $\boldsymbol{x}_n, P_n$ and $\mathbb{P}_n$:

$$\boldsymbol{z} = \left\{(\boldsymbol{x}_n, n \in \mathcal{E}_N); (P_n, n \in \mathcal{E}_{N-1}); (\mathbb{P}_n, n \in \mathcal{E}_{N-2})\right\}$$

Proposition 5 *Problem* (9) *can be reformulated in the Condat-Vu framework with $I = 3$ with:*

- *h the quadratic criterion appearing in* (9a)*,*
- *f, g_0, g_1 and L_0, L_1 defined as in Proposition 3,*
- *$g_2 = g_0 = \iota_{\{\mathbf{0}\}}$ and L_2 the linear operator induced by constraints* (9d)*.*

Regarding the setting of hyperparameters τ and σ_i, let us apply Theorem 9.7 from [11] to Problem (9).

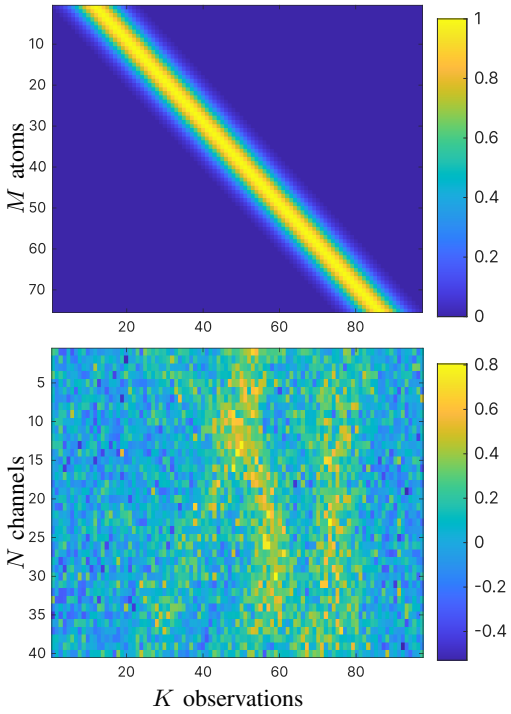

Fig. 3. Gaussian dictionary A (top). Noisy signals $\boldsymbol{y}_n$ at 0 dB (bottom).

Proposition 6 *The conditions for convergence [11, Th. 9.7] are met for Algorithm 2 as soon as:*

$$\tau \leq \left(2\left\|A^\top A\right\| + \sigma_1 M + 2(M+1)(\sigma_0 + \sigma_2)\right)^{-1}.$$

SKETCH OF PROOF: The complete proof is an extension of that of Proposition 3 (see [12]). The spectral norms $\left\|L_1^\top L_1\right\|$ and $\|Q\|$ remain unchanged. Furthermore, one can show that $\left\|L_2^\top L_2\right\|$ can be upper bounded by $2(M+1)$, akin to $\left\|L_0^\top L_0\right\|$.

IV. NUMERICAL RESULTS

Our methods are tested on a sparse deconvolution problem. Each observed signal $\boldsymbol{y}_n$ reads as a noisy convolution product $h * \boldsymbol{x}_n + \boldsymbol{n}_n$ where h is a Gaussian shaped impulse response and $\boldsymbol{x}_n$ is a sparse signal, see Fig. 1. As shown in Fig. 3, the sequence $\boldsymbol{y}_n$ slowly evolves across channels. It is noticeable that the dictionary columns are strongly correlated, which makes the decomposition problem difficult, especially when the noise level is high and when the trajectories are overlapping. Hereafter, the simulation problem has $K = 97$ observations, $N = 40$ channels, and the convolutive dictionary is composed of $M = 75$ shifted atoms.

We compare both first and second order methods (with the fast implementation of Sect. II-C, with $d = 5$) with Basis Pursuit Denoising (BPDN) and Group-Norm minimization (GN) [13]. BPDN boils down to monovariate sparse decomposition, that is, each channel is treated independently. In the GN approach, each group is defined by gathering a given dictionary atom for two consecutive channels n and $n+1$. The results of Fig. 4 were obtained for the dataset of Fig. 3. GN tends to force locally constant, discontinuous

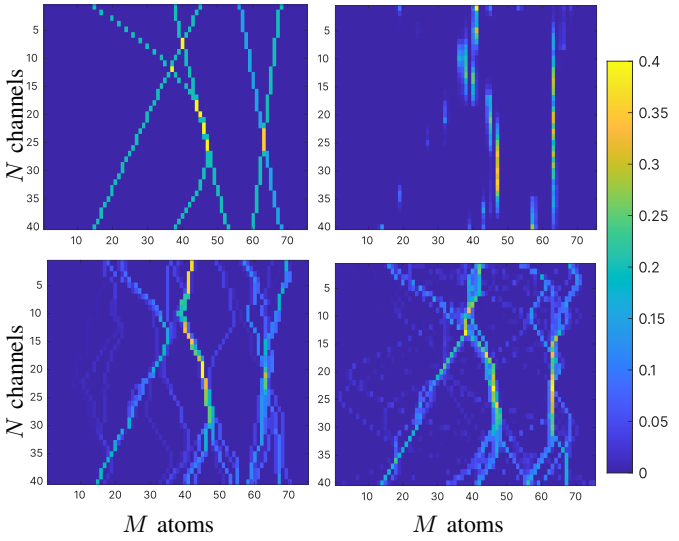

Fig. 4. Ground truth $\boldsymbol{x}_n^*$ (top left) and recovered signals $\hat{\boldsymbol{x}}_n$ for problem of Fig. 3: group norm (top right), $\mathcal{OT}_1$ (bottom left), $\mathcal{OT}_2$ (bottom right).

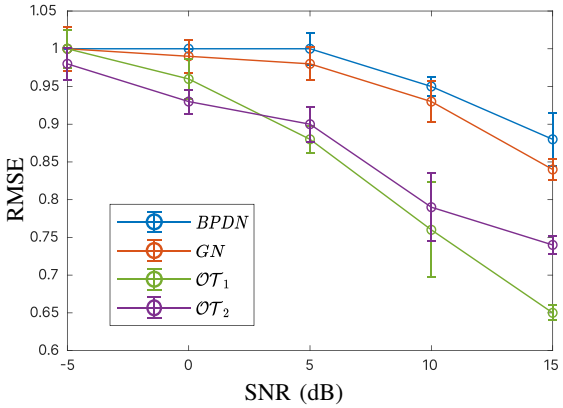

Fig. 5. Root Mean Square Error (RMSE) on $\boldsymbol{x}_n$ versus SNR.

trajectories of atoms. On the contrary, the $\mathcal{OT}_1$ based method succeeds to recover continuous trajectories. One can further notice that the $\mathcal{OT}_2$ method has an improved capacity to follow up overlapping trajectories. The BPDN method (not shown here) gives poor results at this noise level. In Fig. 5, the previous experiment is repeated for various noise levels and noise realizations. The RMSE comparisons confirm the potential of the proposed approach.

It is noticeable that the number of variables of optimization problems (5) and (9) is strongly reduced in the fast implementation setting (Sect. II-C, $d = 5$). When $M = 500$, $K \approx M$, and $N = 40$, the number variables is 2.10^5 and 2.10^7, respectively as compared to 10^7 and 5.10^9 for the original (high dimensional) problems.

In principle, a substantially smaller problem should be much faster to solve. However, we have only observed small gains in terms of speed for the Condat-Vu algorithm, whereas we have obtained the expected one using a totally different solution based on CPLEX to solve (5) or (9), once put in the generic format of quadratic programming problems. Moreover, we have also observed that our CPLEX implementation was far more efficient than our Matlab implementation of Condat-Vu. Indeed, we suspect that the efficiency of the latter may be strongly impacted by calls to a specialized tensor toolbox.

V. Conclusion

We introduced two kinds of regularization schemes based on OT for multichannel signal decomposition. As expected, the first order approach promotes locally constant trajectories of selected dictionary atoms, while the second order approach is more flexible and recovers oblique trajectories. On the sparse deconvolution problem, both methods outperform BPDN for monovariate decomposition and the group-norm approach for multivariate decomposition. Moreover, the fast implementation greatly reduces the number of variables involved in the optimization problems.

The sum-to-one assumption (1) might be unrealistic in many scenarios. A main perspective is to relax this assumption, leading us to investigate the so-called *unbalanced* OT metrics, see, *e.g.*, [9]. Other perspectives include acceleration of the optimization scheme using Sinkhorn's algorithm. This approach is expected to improve efficiency with the disadvantage to yield less sparse decomposition results.

References

[1] D. Pereg, I. Cohen, and A. A. Vassiliou, "Multichannel sparse spike inversion", *Journal of Geophysics and Engineering*, vol. 14, no. 5, pp. 1290–1299, 09 2017.

[2] J. M. Bioucas-Dias, A. J. Plaza, N. Dobigeon, M. Parente, Q. Du, P. D. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches", *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, Apr. 2012.

[3] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties", *Found. Trends Mach. Learn.*, vol. 4, no. 1, pp. 1–106, Jan. 2012.

[4] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social sparsity! Neighborhood systems enrich structured shrinkage operators", *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 53–78, May 2013.

[5] G. Peyré and M. Cuturi, "Computational optimal transport", *Found. Trends Mach. Learn.*, vol. 51, no. 1, pp. 1–44, 2019.

[6] J. Karlsson and A. Ringh, "Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport", *SIAM J. Imaging Sciences*, vol. 10, no. 4, pp. 1935–1962, 2017.

[7] R. Flamary, C. Févotte, N. Courty, and V. Emiya, "Optimal spectral transportation with application to music transcription", *NeurIPS*, vol. 29, 2016.

[8] H. Janati, T. Bazeille, B. Thirion, M. Cuturi, and A. Gramfort, "Multi-subject MEG/EEG source imaging with sparse multi-task regression", *Neuroimage*, vol. 220, pp. 116847, 2020.

[9] J. Lee, N. P. Bertrand, and C. J. Rozell, "Unbalanced optimal transport regularization for imaging problems", *IEEE Trans. Comput. Imaging*, vol. 6, pp. 1219–1232, 2020.

[10] F. Elvander, I. Haasler, A. Jakobsson, and J. Karlsson, "Multi-marginal optimal transport using partial information with applications in robust localization and sensor fusion", *Signal Process.*, vol. 171, 2020.

[11] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi, "Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists", *SIAM Rev.*, vol. 65, no. 2, pp. 375–435, 2023.

[12] J. Malléjac, J. Idier, C. Soussen, and X. Li, "Proximal optimization of OT-regularized cost functions for multichannel signal recovery", Research report, https://hal.science/hal-04488937, Mar. 2024.

[13] L. Yuan, J. Liu, and J. Ye, "Efficient methods for overlapping group lasso", *NeurIPS*, vol. 24, 2011.