TELECOM Paris

IP PARIS

INSTITUT POLYTECHNIQUE DE PARIS

PHD THESIS PROJECT

New approaches for non-linear blind source separation, with application to remote sensing data

1. Context

Blind source separation (BSS) [2] is a powerful machine learning paradigm with a wide range of applications such as remote sensing [14] and biomedical imaging [15]. Generally speaking, BSS aims at decomposing a given data set into unknown elementary signals to be recovered, generally referred to as the *sources*.

Because it is simple and easily interpretable, many works [2] have focused on the *linear* mixing model (LMM), which assumes that the ith data set sample $\bar{\mathbf{x}}_i \in \mathbb{R}^m$ ($i \in [\![n]\!] = \{1, 2, \dots, n\}$) has been physically generated by a linear process

$$\bar{\mathbf{x}}_i = \sum_{k=1}^{r} h_{ki}\mathbf{w}_k + \mathbf{n}_i,$$

where $\mathbf{w}_k \in \mathbb{R}^m$ is the kth source ($k \in [\![r]\!]$) and h_{ki} is the associated mixing coefficient in the ith (mixed) observation. The vector $\mathbf{n}_i$ accounts for any additive noise. The goal of BSS is to recover, from the sole knowledge of the $\bar{\mathbf{x}}_i$ ($i \in [\![n]\!]$), both the sources $\mathbf{w}_k$ ($k \in [\![r]\!]$) and the h_{ki} coefficients ($k \in [\![r]\!], i \in [\![n]\!]$) that generated the data.

As an infinity of spurious solutions exists, BSS is an ill-posed problem [2], calling for additional constraints on $\mathbf{w}_k$ and h_{ki} to reduce the number of spurious solutions. The focus will be here on nonnegativity constraints, akin to nonnegative matrix factorization (NMF) [9]. More specifically, this PhD will focus on the subclass of near-separable NMF [1], for which the problem can be solved in a polynomial time with weak indeterminacies. This subclass corresponds to data sets in which each source appears purely in at least one data sample. Identifying the sources then boils down to finding among the $\bar{\mathbf{x}}_i$ ($i \in [\![n]\!]$) which ones are approximately equal to some sources $\mathbf{w}_k$ ($k \in [\![r]\!]$). Building on this, several provably robust algorithms have been proposed, that is algorithms which are proved to recover the sources having generated the data set [9].

Application : hyperspectral unmixing A typical application of BSS to remote sensing is the hyperspectral (HS) unmixing problem. HS images can be seen as a generalization of RGB images, since they measure the energy in a high number m of wavelength bands ($m = 3$ for RGB images). Therefore, a whole spectrum $\mathbf{x}_i$ is acquired for each of the n pixels. Nevertheless, the spatial resolution of HS images is usually low (n small). Thus, when HS imaging systems are embedded in a plane or a satellite in order to perform earth monitoring, each pixel usually corresponds to several square meters on the floor, making that several materials (*e.g.* water, earth, stone...) of the scene are usually present in the pixels [10]. Consequently, the spectrum $\mathbf{x}_i$ measured in each pixel does not correspond to the spectra $\mathbf{w}_k$ of the single materials, but rather to a mixture of them, calling for BSS techniques to unmix them ; see Fig. 1 for a concrete example. Note that in the HS context, the near-separable assumption amounts to assume that for each material, there is at least a pixel in which this material appears purely.

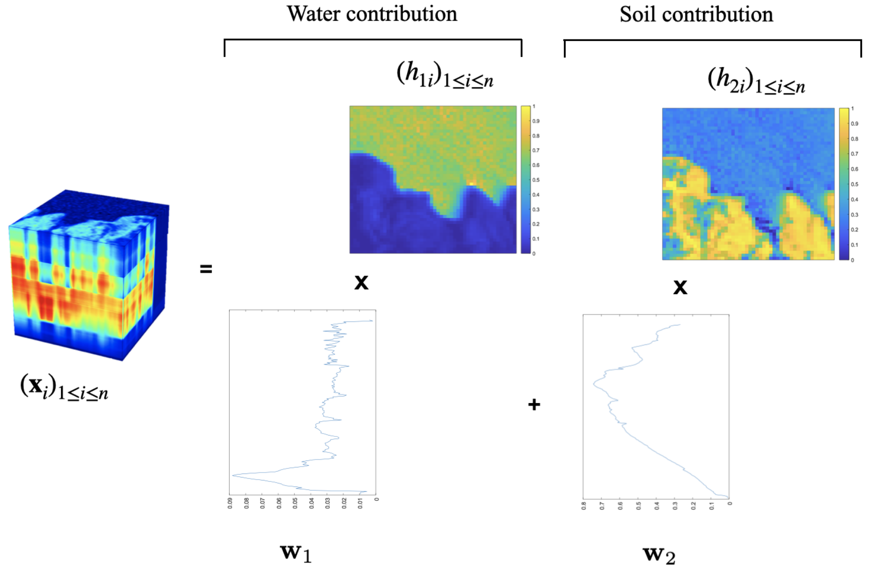
Water contribution
Soil contribution
$(h_{1i})_{1\leq i\leq n}$
$(h_{2i})_{1\leq i\leq n}$
=
x
x
+
$(\mathbf{x}_i)_{1\leq i\leq n}$
$\mathbf{w}_1$
$\mathbf{w}_2$

Figure 1. Illustration of BSS for HS unmxing. The data set, acquired over a coastal scene, is constituted of two materials : water and soil. The spectrum measured at each pixel of the cube is here assumed to be a linear mixture of the soil and water spectra. BSS aims at finding, in a blind fashion, the spectrum of each material $\mathbf{w}_k$ and the concentration h_{1i} of each material in each pixel $i \in [\![n]\!]$.

2. General objective of this PhD

In various applications such as HS unmixing, the LMM is unfortunately only a first-order approximation of non-linear mixing processes [4]. Therefore, linear-quadratic (LQ) models sometimes better account for the physical mixing processes by including termwise products of the sources [4], which writes as

$$\bar{\mathbf{x}}_i = \sum_{k=1}^{r} h_{ki}\mathbf{w}_k + \sum_{p=1}^{r}\sum_{l=p}^{r} \beta_{ipl}(\mathbf{w}_p \odot \mathbf{w}_l) + \mathbf{n}_i. \tag{1}$$

In (1), the linear contribution associated to LMM is complemented by a set of second-order interactions $\mathbf{w}_p \odot \mathbf{w}_l$ between the sources, where $\odot$ denotes the Hadamard product and β_{ipl} is the amount of the interaction $\mathbf{w}_p \odot \mathbf{w}_l$ within the ith observation. Of particular applicative interest is the Nascimento [4] model, consisting of (1) with the two additional constraints :

$$\begin{aligned} &\sum_{k=1}^{r} h_{ki} + \sum_{p=1}^{r}\sum_{l=p}^{r} \beta_{ipl} = 1 \ \text{ for } i \in [\![m]\!], \quad \text{(sum-to-one constraint)} \\ &\mathbf{w}_k \geq 0, h_{ki} \geq 0, \beta_{ipl} \geq 0 \text{ for } k,p,l \in [\![r]\!], i \in [\![n]\!] \quad \text{(nonnegativity constraints)}. \end{aligned} \tag{2}$$

Unfortunately, contrary to linear BSS, much less work has been devoted to the development of provably robust LQ near-separable algorithms. In fact, we are only aware of our recent work [12], where we unfortunately also showed that the conditions ensuring the proposed SNPALQ algorithm robustness were quite restrictive in practice. **The goal of this PhD is therefore to investigate several promising research paths for LQ near-separable provably-robust BSS**.

3. Detailed content of the PhD

To tackle the LQ near-separable BSS problem, some research directions are presented into the three following sections : section 3.1 proposes to focus on the Nascimento mixing model, both by extending the provably robust method we proposed in [13] and introducing new ones. Section 3.2 rather delves on studying different mixing models than the Nascimento one. Lastly, Section 3.3 aims at applying the developed methods to important real-life applications, such as hyperspectral unmixing.

As a side remark, the proposed research directions are mostly independent, making that any difficulty to accomplish one of them should not have high repercussions on the other ones.

3.1. New developments for tackling the Nascimento model

When dealing with the LMM, three main families of provably-robust near-separable NMF algorithms exist [9] : (i) greedy algorithms, (ii) brute-force algorithms, and (iii) optimization-based approaches (self-dictionary). So far, only greedy (i) and brute-force (ii) algorithms have been extended to near-separable LQ Nascimento mixtures, and these extensions have limitations. The objective of the first part of this PhD is to bypass these issues through two complementary approaches :

— *Improving existing greedy LQ-NMF methods (i) using robust-to-outliers approaches* : a first possibility is to improve SNPALQ, the LQ greedy algorithm we proposed in [13]. Specifically, when the recovery conditions of SNPALQ algorithm are not met, some observations can erroneously be extracted and be spuriously considered as sources by the algorithm. To reduce the number of such spurious sources, an option could be to break the fully sequential (or greedy) scheme of SNPALQ by enabling the algorithm to discard some erroneously-extracted observations, in a similar way to what is done in [8] in the context of the LMM. On the other hand, the work of [8] rises open questions (such as how to fully take into account the nonnegativity constraints), which might first be tackled before extending the algorithm to LQ mixtures.

— *Tackling near-separable LQ BSS through optimization based approaches (iii) :* another option is to extend the algorithms of family (iii) to LQ mixtures. The advantage is twofold : first, compared to family (i), the conditions guarantees could be made milder than with the current SNPALQ. Second, LMM algorithms of family (iii) jointly consider the observations in the data set, in contrast to brute-force algorithms (ii). Therefore, they usually lead to better practical results. Consequently, it can be conjectured that extending family (iii) to LQ mixtures would outperform the only current LQ brute-force algorithm (iii) [12].

3.2. Proposing new provably robust LQ algorithms for other LQ models

The above part dealt with the Nacimento model, consisting in mixtures of the form (1) with constraints (2). In contrast, this part aims at tackling different LQ models, which, depending on the application at hand, might be more relevant than the Nascimento one [3]. Although this PhD will not be restricted to it, we here take as an example the Fan model [7], which has been well used in the context of hyperspectral unmixing [4]. The mixing coefficients are then supposed to have the specific form $\beta_{ipl} = h_{pi}h_{pl}$, leading to the following mixing model :

$$\begin{gathered}\bar{\mathbf{x}}_i = \sum_{k=1}^{r} h_{ki}\mathbf{w}_k + \sum_{p=1}^{r}\sum_{l=p}^{r} h_{pi}h_{pl}(\mathbf{w}_p \odot \mathbf{w}_l) + \mathbf{n}_i, \text{ for } i \in [\![1,n]\!], \\ \text{with } \mathbf{w}_k \geq 0,\ h_{ki} \geq 0 \text{ for } k \in [\![r]\!],\ i \in [\![n]\!] \text{ and } \sum_{k=1}^{r} h_{ki} = 1 \text{ for } i \in [\![n]\!].\end{gathered} \tag{3}$$

The rationale is that if a source is absent in a given observation, it cannot (quadratically) interact with the other ones. To the best of our knowledge, no provably robust LQ near separable BSS algorithm exists for the Fan model. The second part of the PhD objectives is thus to fill this gap. For instance, a first step could be to adapt SNPALQ to the constraint $\beta_{ipl} = h_{pi}h_{pl}$. Nevertheless, attention must be paid to the fact that the geometric interpretations of the Naciemento and Fan models are different (see Figure 2), which might jeopardize SNPALQ separation quality and call for the development of another algorithm.

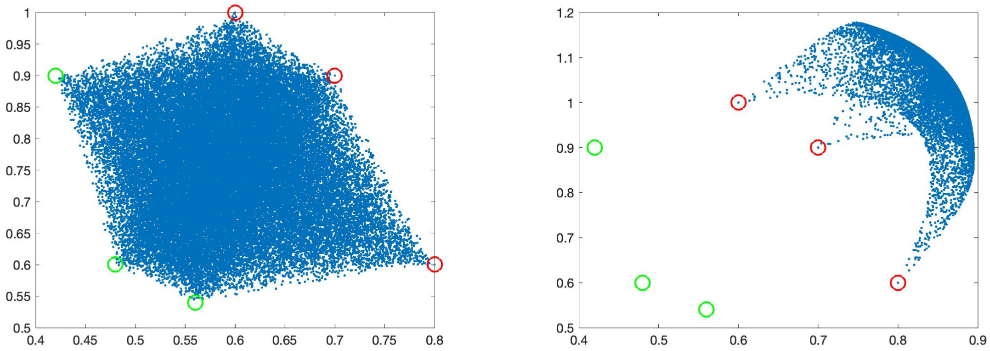

Figure 2. Scatter plot of two LQ near-separable mixtures following (left) the Nascimento and (right) the Fan models. Here, $m = 2$, $r = 3$, $n = 50000$. Each observation $\mathbf{x}_i$, $(i \in [\![n]\!]]$ is represented as a blue point. The red circles correspond to the sources, and the green ones to their second-order products. As can be seen, the geometric interpretations of the two mixing models are quite different. In particular, with the Nascimento model, the sources correspond to the vertices of a polytope comprehending all the observations. This geometric interpretation, which is at the basis of SNPALQ, is not anymore true for complicated Fan mixings, calling for a new algorithm.

3.3. Application to hyperspectral remote sensing imaging

The developed algorithms should in particular improve HS data unmixing. As such, they will be applied on real data sets ; for instance, on the ones acquired by the AVIRIS, APEX or HyMap sensors [1]. Specifically, it has been well established [14] that LQ mixing models (1) generally outperform linear models on data sets captured on 3-dimensional scenes, in which non-linearities are introduced by light multiple scatterings. In this PhD, the two following main HS unmixing applications could be investigated, among others :

— Urban environment monitoring [14, 11], in which multiple scattering stems from the presence of high buildings. The interest is here to better control the increasingly fast urban environment development. Better city planning enables in turn the improvement of human being life quality, as well as sustainable development, through for instance the control of the evolution of vegetated areas in urban environments [16].

— Vegetation monitoring [7, 3], in which multiple light scattering can be induced, among others, by the tree layered structure. Vegetation monitoring has a broad interest : non-linear HS was used in [17] to determine, from the WorldView-2 satellite data, the evolution of vegetation health in the Antartic region, which is related to the impact of global warming. As another example, HS unmixing is used in [6] for change detection in forests induced by fires.

Beyond applying the developed methods on the above applications, LQ models can also be used for defense (see [5] in which coastal monitoring is performed) and also appear in different contexts than 3-dimensional scenes (see for instance intimate mixtures [4] in rocks constituted of closely distributed different minerals). Therefore, new fields of application of the proposed methods could be explored. Lastly, working on real data sets might also rise interesting methodological questions, as the LQ mixing model could be perturbed by further non-linearities, such as spectral variabilities.

4. Working context

This work is intended to be done in collaboration between several members of different institutions and countries :

— *From Télécom Paris (France) :* The project will be conducted under the supervision of Christophe Kervazo (christophe.kervazo@telecom-paris.fr) and Florence Tupin, within the

[1] https ://aviris.jpl.nasa.gov/index.html ; https ://earth.esa.int/web/eoportal/airborne-sensors/apex ; https ://airbornescience.nasa.gov/instrument/HyMap

IMAGES group, at the Telecom Paris engineering school. The IMAGES group has proposed several major contributions in the context of remote sensing and currently aims at developing a research axis around the topic of hyperspectral imaging.

- *From University of Mons (Belgium) :* The work is expected to be performed in collaboration with Nicolas Gillis, who is a reknown expert in the field of near-separable NMF.

5. Candidate

The candidate should have a Master 2 degree (or equivalent) and an excellent academic curriculum. He/she should have a good knowledge in signal/image processing and mathematics (especially, linear algebra). Knowledge in convex optimization is a plus. Ideally, Matlab programming language should be mastered.

The candidate will acquire an expertise in signal processing (in particular, of multi-valued data), which is valuable in many fields : remote sensing, astrophysics, text-mining...

Contact must be taken with Christophe Kervazo (christophe.kervazo@telecom-paris.fr) before April 20th, 2021.

6. Supervisors

- Christophe Kervazo received the Supélec (France) engineering degree in 2015, and the master of science in Electrical and Computer Engineering from Georgia Institute of Technology (USA) in 2016. From 2016 to 2019, he was PhD student in the CosmoStat group at CEA Saclay (France), where he worked on the optimization framework for sparse blind source separation, as well as non-linear component separation. In Mons (Belgium), he then worked as a post-doctoral researcher, under the supervision of N. Gillis, on the extension of Nonnegative Matrix Factorization to Linear-Quadratic mixture unmixing. He is currently an Assistant Professor (maître de conférences) at Télécom Paris (France), in the IMAGES group. More details can be found at `https://sites.google.com/view/christophekervazo/`.
- Florence Tupin research work is dedicated to the development of remote sensing methods for image analysis, processing and interpretation. She has in particular worked on SAR images. She published more than fifty articles in the IEEE Transactions on Geoscience and Remote Sensing (TGRS) and IEEE Transactions on Image Processing journals, and more than fifty conference articles. She co-received the best article prize of IEEE TGRS in 2016, the best student article at ICIP in 2010. The PhD candidates she supervised have obtained several PhD prizes. More details can be found at `https://perso.telecom-paristech.fr/tupin/`.

References

[1] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization—provably. *SIAM Journal on Computing*, 45(4) :1582–1611, 2016.

[2] P. Comon and C. Jutten. *Handbook of Blind Source Separation : Independent component analysis and applications*. Academic Press, 2010.

[3] N. Dobigeon, L. Tits, B. Somers, Y. Altmann, and P. Coppin. A comparison of nonlinear mixing models for vegetated areas using simulated and real hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6) :1869–1878, 2014.

[4] N. Dobigeon, J.-Y. Tourneret, C. Richard, J. C. M. Bermudez, S. McLaughlin, and A. O. Hero. Nonlinear unmixing of hyperspectral images : Models and algorithms. *IEEE Signal Processing Magazine*, 31(1) :82–94, 2014.

[5] O. Eches and M. Guillaume. A bilinear–bilinear nonnegative matrix factorization method for hyperspectral unmixing. *IEEE Geoscience and Remote Sensing Letters*, 11(4) :778–782, 2013.

[6] A. Ertürk, M.-D. Iordache, and A. Plaza. Sparse unmixing-based change detection for multitemporal hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(2) :708–719, 2015.

[7] W. Fan, B. Hu, J. Miller, and M. Li. Comparative study between a new nonlinear model and common linear model for analysing laboratory simulated-forest hyperspectral data. *International Journal of Remote Sensing*, 30(11) :2951–2962, 2009.

[8] N. Gillis. Successive projection algorithm robust to outliers. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 331–335. IEEE, 2019.

[9] N. Gillis. *Nonnegative Matrix Factorization.* SIAM, 2020.

[10] U. Heiden, S. Roessner, K. Segl, and H. Kaufmann. Analysis of spectral signatures of urban surfaces for their identification using hyperspectral hymap data. In *IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (Cat. No. 01EX482)*, pages 173–177. IEEE, 2001.

[11] P. Huard. Study of non-linear mixing in hyperspectral imagery—a first attempt in the laboratory. In *2011 3rd Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, pages 1–4. IEEE, 2011.

[12] C. Kervazo, N. Gillis, and N. Dobigeon. Provably robust blind source separation of linear-quadratic near-separable mixtures. *arXiv preprint arXiv :2011.11966*, 2020.

[13] C. Kervazo, N. Gillis, and N. Dobigeon. Successive nonnegative projection algorithm for linear quadratic mixtures. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1951–1955. IEEE, 2021.

[14] I. Meganem, P. Déliot, X. Briottet, Y. Deville, and S. Hosseini. Linear–quadratic mixing model for reflectances in urban environments. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1) :544–558, 2013.

[15] J. J. Rieta, F. Castells, C. Sánchez, V. Zarzoso, and J. Millet. Atrial activity extraction for atrial fibrillation analysis using blind source separation. *IEEE Transactions on Biomedical Engineering*, 51(7) :1176–1186, 2004.

[16] C. Song. Spectral mixture analysis for subpixel vegetation fractions in the urban environment : How to incorporate endmember variability ? *Remote sensing of environment*, 95(2) :248–263, 2005.

[17] X. Sun, W. Wu, X. Li, X. Xu, and J. Li. Vegetation abundance and health mapping over southwestern antarctica based on worldview-2 data and a modified spectral mixture analysis. *Remote Sensing*, 13(2) :166, 2021.