

Optimisation déterministe et stochastique pour des décompositions matricielles/tensorielles couplées. Application au traitement de signaux multi-dimensionnels et au traitement d'images en grande dimension.

Mots clefs : analyse de données multi-dimensionnelles, algorithmes de factorisation tensorielle, optimisation non-linéaire, optimisation sous contraintes, décompositions (canoniques) polyadiques (CanDe-comp/Parafac), CP, Tucker, parallélisation, décomposition tensorielles/matricielles couplées.

Contexte général

Nous nous intéressons à des problèmes d'analyse de données multi-dimensionnelles [9]. Les blocs de données interprétés de manière tensorielle peuvent être modélisés au moyen de décompositions multi-linéaires (Parafac, Tucker, etc.). Afin d'estimer les facteurs impliqués dans ces décompositions, il est classique de recourir à des algorithmes d'optimisation de critères consistant à mesurer l'erreur de modélisation. Les algorithmes d'optimisation déterministes et stochastiques sous contrainte que nous développons s'appliquent à des problèmes de traitement de signaux multi-dimensionnels ou de traitement d'images en grande dimension. Les applications visées portent principalement sur :

- La spectroscopie de fluorescence (imagerie 3D [5][10])
- La déconvolution multi-dimensionnelle de signaux ou l'estimation de directions d'arrivées utilisant des distributions temps-fréquence spatiales quadratiques pouvant exploiter différentes modalités : temps-fréquence-capteurs, fréquence Doppler-retard-capteurs voire même temps-fréquence-capteurs-polarité (tenseurs d'ordre 4) [13]
- La télédétection et l'imagerie hyperspectrale, les problèmes de fusion d'information et les décompositions NMF [7][8]

La finalité des algorithmes que nous développons est : i) l'identification des variables latentes et la réduction de dimension ii) la séparation (ou démixage) des sources ou composés purs présents au niveau des mélanges étudiés voire leur classification. Ainsi, en spectroscopie de fluorescence, les méthodes que nous développons ont pour but de détecter, séparer voire classifier :

1. des composés de matière organique fluorescents ou des polluants de type HAP [Vu17]
2. des métaux lourds (implication au niveau de l'action "Surveillance" du projet PREVENT porté par l'axe transverse MEDD de l'UTLN ce qui nous a amené à déposer conjointement avec les laboratoires MIO et MAPIEM, une demande ANR actuellement dans la seconde phase d'évaluation (baptisée COSMIQ pour Capteurs fluoRescents innovants pour la meSure de Métaux : outils de surveillance In situ de la Qualité des eaux. Ce projet porte sur l'élaboration de nouveaux capteurs à base de polymères à empreinte dotés d'une sonde fluorescente pour la détection du plomb et du cadmium) ou encore
3. des radionucléides présents dans les échantillons prélevés en un même site de surveillance (ce qui a fait l'objet d'un projet CNRS NEEDS avec l'IRSN).

En contrôlant l'évolution de la composition de l'eau au cours du temps, l'objectif est de parvenir à détecter de manière précoce des changements environnementaux tels que ceux survenant lors d'incidents de pollution et pouvoir ainsi prévenir ou minimiser leurs risques.

Objectifs

Cette thèse aura pour objet des développements d'ordre théorique avec un souci d'application à des données expérimentales. Ces développements porteront sur les décompositions tensorielles (non-négatives) couplées (tenseurs d'ordre deux, trois et/ou quatre) pour le démélange d'images (de spectroscopie de fluorescence, hyperspectrale) ou encore de signaux multi-dimensionnels. Alors que les traitements conjoints des données consistent à décomposer simultanément un grand nombre de blocs de même dimensions, les traitements couplés des données nécessitent la manipulation et la décomposition de blocs (matrices ou tenseurs) de dimensions différentes, correspondant à des grilles d'échantillonnage spatial, spectral (ou toute autre modalité considérée) non superposables et à des rapports signal à bruit éventuellement très différents. Il est alors nécessaire de résoudre un problème de décomposition couplée où le couplage fait non seulement intervenir des opérateurs d'interpolation mais éventuellement aussi d'autres types d'opérateurs (décalage, atténuation, etc.). La plupart des méthodes de la littérature procèdent de manière itératives en alternant d'un problème de décomposition à l'autre [20] - ce qui est très sous-optimal - au lieu de considérer directement le problème couplé [16][15][17]. Ce type de problème peut également se poser en dimension encore plus grande dans tous les problèmes de surveillance par des réseaux de capteurs.

Cette thèse qui s'inscrit dans les thématiques de deux des trois axes transverses de l'UTLN (MEED et INP), est le prolongement naturel des travaux de J.-P. Royer [4] et de X. Vu [2] (tous deux ayant été co-encadrés par N. Thirion-Moreau). Des compétences supplémentaires en algorithmique, optimisation stochastique et parallélisation d'algorithmes seront apportées par C. Prissette. Elle portera sur des problèmes d'optimisation sous contrainte en grande dimension. Nous y traiterons les aspects modélisation, couplage et choix des fonctions de coût à optimiser. La première étape de la thèse consistera à formaliser et à valider un modèle mathématique suffisamment général et flexible pour pouvoir s'appliquer à la grande diversité des configurations réelles que l'on peut rencontrer. Cela implique :

- Le choix de la (les) matrice(s) commune(s) dans les décompositions couplées,
- Les liens entre les autres matrices intervenant dans les décompositions : opérateurs d'interpolation mais éventuellement aussi d'autres types d'opérateurs (décalage, atténuation, etc.)
- La nécessité de prise en compte de la majorité des contraintes du problème traité (parcimonie, non négativité, etc.)

Ensuite, à partir de la modélisation retenue et du simulateur de données réalisé, il s'agira de développer de nouveaux algorithmes permettant d'identifier l'ensemble des paramètres du modèle considéré. Pour cela, une étude théorique sera réalisée en termes d'existence, voire d'unicité, de régularité et de stabilité de la solution. Ensuite, nous passerons à l'analyse numérique qui conditionnera le choix de l'algorithme d'optimisation à mettre en oeuvre et fournira une approximation de la solution. Nous étudierons la robustesse des algorithmes développés vis-à-vis d'erreurs (rang des tenseurs, modèle, bruit additif, non-linéarités, non-stationnarité, etc.) et testerons également l'intérêt dans un tel contexte d'introduire des pénalisations. L'étude de la complexité algorithmique sera réalisée, répondant à notre contrainte de rapidité d'exécution. Dans un dernier temps, les problèmes liés à la grande dimension seront traités. Pour accélérer les temps de traitement, nous envisageons de proposer des versions adaptatives [12] des méthodes développées afin de mieux prendre en compte l'aspect dynamique des phénomènes observés ou parallélisables [19][11] en utilisant MapReduce et l'architecture Hadoop popularisée dans le domaine du Big Data [18]. Enfin des tests sur des jeux de données réelles issus de plusieurs domaines d'application différents seront ensuite réalisés.

Informations pratiques

La thèse se déroulera au sein des laboratoires LIS, UMR CNRS 7020 (Ecole d'ingénieurs SeaTech, Université de Toulon, Toulon, France), dans l'équipe SIIM (Signal IMages) du Pôle Signal Image.

Encadrants :

- Nadège Thirion-Moreau : Professeur des Universités,
Laboratoire LIS, UMR CNRS 7020, Equipe SIIM,
Tél: 04.94.14.24.56, email: thirion@univ-tln.fr
- Cyril Prissette : Maître de Conférences non HDR
Laboratoire LIS, UMR CNRS 7020, Equipe SIIM (Signaux et Images),
Tél: 04.94.14.67.03, email: prissette@univ-tln.fr

Profil du candidat ou de la candidate :

Les candidats doivent être titulaires d'un diplôme de Master ou équivalent (école d'ingénieur).

Ils doivent témoigner de compétences dans au moins l'un des domaines suivants : optimisation, traitement du signal et des images, analyse numérique, programmation, calcul distribué.

La thèse est prévue pour un démarrage en septembre ou octobre 2018. Le financement de la thèse est un financement du Ministère de l'Enseignement Supérieur et de la Recherche (ED 548, Mer et Sciences) s'élevant à un montant net de 1367,80 euros (1684,93 euros bruts) par mois durant 36 mois. Possibilité d'avenant en 2ème année pour des activités complémentaires en enseignement.

Pour tout renseignement complémentaire, contacter les encadrants de thèse prissette@univ-tln.fr et thirion@univ-tln.fr, pour faire acte de candidature, leur transmettre un CV détaillé, une copie du dernier diplôme obtenu et les relevés de notes.

Date limite de candidature : 21 juin 2018

References

- [1] Jean-Philip Royer, "Identification aveugle de mélanges et décomposition canonique de tenseurs : application à l'analyse de l'eau". Thèse de doctorat du pôle de recherche et d'enseignement supérieur (PRES) euro-méditerranéen (thématique : "environnement, développement durable et société") de l'Université de Nice Sophia-Antipolis, soutenue le 4 octobre 2013.
- [2] Thi Thanh Xuan Vu, "Optimisation déterministe et stochastique pour des problèmes de traitement d'images en grande dimension".
Thèse sur sujet pluri-disciplinaire du LABEX Archimède de l'AMU, soutenue le 13 novembre 2017.
- [3] J.-P. Royer, N. Thirion-Moreau, P. Comon "Computing the polyadic decomposition of nonnegative third order tensors", in *Eurasip Signal Processing*, Vol. 91, Issue 9, pp. 2159-2171, September 2011.
- [4] I. Kopriva, J.-P. Royer, N. Thirion-Moreau, P. Comon "Error Analysis of Low-Rank Three-Way Tensor Factorization Approach to Blind Source Separation", accepted in Proc. International Conference on Acoustic Speech and Signal Processing (*ICASSP'2014*), Florence, Italy, 4-9 May 2014

- [5] J.-P. Royer, N. Thirion-Moreau, P. Comon, R. Redon, S. Mounier, “A regularized nonnegative Canonical Polyadic Decomposition algorithm with preprocessing for 3D fluorescence spectroscopy”, in *Journal of Chemometrics*, vol. 29, pp. 253-265, DOI: 10.1002/cem.2709, March 2015.
- [6] S. Maire, C. Prissette, “A Restarted estimation of distribution algorithm for solving sudoku puzzles”, *Monte Carlo methods and applications* 18, no. 2, pp. 147-160, 2012.
- [7] Z. Ben Rabah “Démixage spectral d’images hyperspectrales et fusion possibiliste des connaissances. Application à l’érosion hydrique”, thèse de doctorat de l’ENST Bretagne, septembre 2011.
- [8] O. Echès “Méthodes bayésiennes pour le démixage d’images hyperspectrales”, Thèse de doctorat de l’Université de Toulouse, octobre 2010.
- [9] A. Cichocki, R. Zdunek, A. H. Phan and S. I. Amari “Non negative matrix and tensor factorizations: Application to exploratory multi-way data analysis and blind separation”, Wiley, 2009.
- [10] C. A. Stedmon, S. Markager and R. Bro, “Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy”, *Marine Chemistry*, Volume 82, pp. 239–254, 2003.
- [11] N. Pustelnik, C. Chaux, J.-C. Pesquet, “Parallel ProXimal Algorithm for image restoration using hybrid regularization”, *IEEE Trans. on Image Proc.*, Vol. 20, No. 9, pp. 2450 - 2462, Sep. 2011.
- [12] D. Nion, N. D. Sidiropoulos, “Adaptive algorithms to track the PARAFAC decomposition of a third-order tensor”. *IEEE Transactions on Signal Processing* 57(6), pages 2299-2310, 2009
- [13] A. Belouchrani, M. Amin, N. Thirion-Moreau, Y. Zhang, “Source separation and localization using time-frequency distributions”. *IEEE Signal Processing Magazine*, pp. 97-107, November 2013.
- [14] X. T. Vu, S. Maire, C. Chaux and N. Thirion-Moreau “A new stochastic algorithm to decompose large nonnegative third order tensors”. *IEEE Signal Processing Letters*, vol. 22, issue 10, pp. 1713-1717, April 2015.
- [Vu17] X. T. Vu, C. Chaux, N. Thirion-Moreau, S. Maire and E. M. Carstea “A new penalized nonnegative third order tensor decomposition using a block coordinate proximal gradient approach: application to 3D fluorescence spectroscopy” in *Journal of Chemometrics*, Wiley & Sons, special issue on penalty methods, Vol. 3, CEM 31.1 March 2017, doi:10.1002/cem.2859.
- [15] M. Sørensen and L. De Lathauwer “Coupled Tensor Decompositions for Applications in Array Signal Processing”, 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2013.
- [16] J. E. Cohen, R. Cabral Farias, P. Comon, “Joint tensor compression for coupled CPD”, 24th EUSIPCO Signal Processing Conference, 2016.
- [17] E. Acar, T. G. Kolda, D. M. Dunlavy, “All at once optimization for coupled matrix and tensor factorizations”, in *MLG’11: Proceedings of Mining and Learning with Graphs* (arXiv:1105.3422), May 2011
- [18] A. Beutel, A. Kumar, E. E. Papalexakis, “FlexiFaCT: Scalable Flexible Factorization of Coupled Tensors on Hadoop”, *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 109-117, 2014.
- [19] R. Gemulla, E. Nijkamp, P. J. Haas, and Yannis Sismanis. “Large-scale matrix factorization with distributed stochastic gradient descent”. In *ACM SIGKDD*, pages 69–77, New York, NY, USA, 2011.
- [20] N. Yokoya, T. Yairi, T., A. Iwasaki, “Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion”, *IEEE Transactions on Geoscience and Remote Sensing*, n. 99, pp. 1–10, 2011

Deterministic and stochastic optimization for coupled matrix and tensor decompositions. Application to multi-dimensionnal signal processing and large size inverse problems in image processing.

Keywords: Multi-dimensional data analysis, coupled matrix/tensor factorization algorithms, non-linear optimization, constrained optimization (non-negativity, sparsity), stochastic algorithms, (Canonical) Polyadic decompositions (CanDecomp, Parafac) CP, Tucker, parallelization, distributed computing.

General context

We are interested in multi-dimensional data analysis problems [9]. The data blocks in a tensorial framework are modeled by multi-linear decompositions (Parafac, Tucker). In order to estimate the latent variables (factors) involved in these decompositions, one often resorts to optimization algorithms which measure the approximation error related to the model. The various constrained stochastic or deterministic optimization algorithms that we develop find applications in multi-dimensional signal processing and in large size image processing problems. Of particular interest are

- The fluorescence spectroscopy analysis (3D or 4D images [5][10])
- The multidimensional deconvolution of signals or the estimation of directions of arrivals (DOA) based on spatial quadratic time-frequency distributions [13] in order to take advantage of a higher number of modalities: time-frequency-sensor, Doppler-delay-sensor or even time-frequency-sensor-polarity (4th order tensor)
- The remote sensing and hyperspectral imaging, fusion problems and use of NMF decompositions [7][8]

The aims of the algorithms that we develop are: i) the identification of the latent variables and dimension reduction, ii) the separation (or demixing) of sources or pure components that are present in the studied mixtures and their classification. For example, in fluorescence spectroscopy, the methods that we develop, aim at detecting, separating and classifying i) fluorescent organic matter compounds or pollutants such as PAH [Vu17] or ii) heavy metals (we were involved in the track “surveillance” of a pluri-laboratories project called PREVENT, carried by the transverse axis MEDD of UTLN. It is the reason why we are also involved with laboratories MIO and MAPIEM, in an ANR call (“COSMIQ” project which is about the development of innovating passive fluorescent sensors for the measure of heavy metals for the in situ survey of the quality of waters). It is based on the use of imprinted polymer including a fluorescent probe in order to detect cadmium and lead) or iii) radionuclides present in water samples collected along time in a given site of observation (CNRS NEEDS project with IRSN). By monitoring the water composition, we expect the early detection of environmental changes such as the ones occurring during pollution events and thus, anticipate and minimize the associated risks.

Objectives

This PhD thesis aims at developing new theoretical tools with a view of application to experimental data. These developments will be around (non-negative) coupled tensor decompositions (of order two, three or four) for the demixing of images (fluorescence spectroscopy, hyperspectral imaging) or the processing of multi-dimensionnel signals. Whereas joint decomposition of data sets consists of simultaneously decompose a great number of data blocks of the same dimension, coupled decompositions handle the decomposition of blocks (matrices or tensors) of different sizes, corresponding to different spatial sampling grids, spectral (or

other kind of considered modality) grids non superposable and eventually to really different signal to noise ratios. It is then necessary to solve a problem of coupled decomposition where the coupling can involve an interpolation operator and eventually other kind of operators (such as a shift, attenuation, etc.). Most methods of the literature proceed iteratively, alternating between one decomposition problem to the other [20] - which is highly sub-optimal - instead of considering directly the coupled problem [16][15][17]. This kind of problem also arises with a higher dimension in all the surveillance problems when sensors networks are used.

This PhD thesis fall within the topic of two of the three main research axis of the university of Toulon (MEED and INP). It is in the continuation of J.-P. Royer's [4] and X. Vu's [2] works (both supervised by N. Thirion-Moreau). Additional skills in stochastic optimization, algorithmic and distributed computation are brought by C. Prissette.

The main topic of the PhD thesis is about large size constrained optimisation problems. We will jointly work on the models, the coupling aspects and the choice of the associated cost functions. The first step will consist of choosing and validating a sufficiently general and flexible enough mathematical model to be applicable to the wide variety of real configurations that one may encounter (the raw observations !). This implies:

- Choice of the common matrix (matrices) involved in the coupled decompositions,
- The links between the other matrices involved in the decomposition: interpolation operator or eventually other operators (shift, attenuation, etc.)
- The necessity to take into account most of the constraints that can be encountered (sparsity, non negativity, etc.)

Then, from the chosen model and the realized simulator, we will have to develop new algorithms able to identify all the parameters of the considered model. To that aim, a theoretical study will be conducted in terms of existence, uniqueness, regularity and stability of the solution. Then, we will focus on the numerical analysis in order to decide which optimisation algorithm will be used to obtain a good approximation of the solution.

The robustness of the derived algorithms with respect to errors (related to the model, to noise, to non-linearity and non-stationarity, etc.) will be deeply investigated. We will also study the interest to introduce regularization in such a context. The study of the algorithmic complexity will be performed due to our constraint of speed in the execution. Finally, concerning the problems due to the large size aspects, we will either consider adaptative versions of the considered methods [12] to better take into account the dynamical aspect of the considered phenomena or distributed computing [19][11] using MapReduce and the Hadoop framework very popular in the Big Data community [18]. Finally tests will be laid on several real data sets coming from different application fields.

Practical informations

The PhD will take place in the LIS laboratory, UMR CNRS 7020 (SeaTech School of Engineering, University of Toulon in the South of France, Toulon, France).

The advisors are:

- Nadège Thirion-Moreau : Professor,
Laboratoire LIS, UMR CNRS 7020, Equipe SIIM,
Tél: 04.94.14.24.56, email: thirion@univ-tln.fr
- Cyril Prissette : Maître de Conférences,
Laboratoire LIS, UMR CNRS 7020, Equipe SIIM,
Tél: 04.94.14.67.03, email: prissette@univ-tln.fr

Candidate profile: Applicants should be at graduate level (Master degree or equivalent engineering degree).

They should have skills in at least one of the following topics: optimization, signal and image processing, scientific computing and computer science.

The PhD will start on September or October 2018. The PhD grant is from the French Minister of Higher Education and Research (ED 548, Sea and Sciences). The net income is fixed at 1367,80 euros per month for 36 months. During the second year, possibility of an amendment to apply for teaching activities. For further information or to apply, please contact the PhD advisors prissette@univ-tln.fr and thirion@univ-tln.fr or send them a detailed CV, a copy of the last obtained diploma and a copy of the marks.

Application deadline: june 21 2018