DOSSIER DE CANDIDATURE
A UNE ALLOCATION DE RECHERCHE
POUR LA RENTREE 2018

Dossier complété et revêtu des signatures à transmettre impérativement pour le :
07 janvier 2018 au plus tard,
A la Direction de la Recherche et Valorisation
secretariat.recherche@univ-littoral.fr

Titre de la thèse :

Méthodes de NMF Bayésiennes informées pour la séparation de sources

Laboratoire d'accueil ULCO : LISIC

Priorité du laboratoire, tous supports de financements confondus : 5/5

Directeur de thèse ULCO : G. ROUSSEL
co-encadrement : Gilles DELMAIRE

Merci de renseigner l'ensemble des demandes de financements envisagées pour ce sujet (NB : Les demandes peuvent porter sur plus de deux cofinanceurs envisagés):

☐ Région 50 % (Dans ce cas, ne pas oublier de remplir également le dossier « Région »)

☐ PMCO 50 %

☐ ULCO 50 %

☐ ULCO 100 %

☐ ADEME 50 %

☐ ADEME 100 %

Merci de nous indiquer si d'autres financements ont été demandés pour ce sujet :

☐ Autre Financeur 50 %, préciser le financeur :

☐ Autre Financeur 100 %, préciser le financeur :

■ **LIBAN – Université Libanaise**

Pour ce dispositif, merci d'indiquer en plus :

- le nom du codirecteur étranger et le laboratoire partenaire

Professeur Antoine ABCHE
Université de Balamand, Tripoli.

- Thématique **1**

(1) La qualité de l'air
(2) Le milieu aquatique
(3) L'obésité, la nutrition et les activités sportives
(4) Les énergies propres et renouvelables
(5) La gestion et le traitement des déchets
(6) L'urbanisme

LIBAN – CNRS Libanais

Pour ce dispositif, merci d'indiquer en plus :

- le nom du codirecteur étranger et le laboratoire partenaire
- Thématique : (1)

(1) La qualité de l'air
(2) Le milieu aquatique
(3) L'obésité, la nutrition et les activités sportives
(4) Les énergies propres et renouvelables
(5) La gestion et le traitement des déchets
(6) L'urbanisme

■ **ARCUS E2D2 :**

Pour ce dispositif, merci d'indiquer en plus :

- le nom du codirecteur étranger et le laboratoire partenaire
- Thématique :

(1) Planifier et habiter la ville durable
(2) Surveillance et gestion durable des infrastructures
(3) Environnement, Atmosphère, Eau
(4) Développement énergique durable

□ **ALGERIE - Université Badji Mokhtar d'Annaba (UBMA)**

Pour ce dispositif, merci d'indiquer en plus :

- le nom du codirecteur étranger et le laboratoire partenaire
- Thématique :

(1) La gestion et le traitement des déchets,
(2) L'aménagement littoral et portuaire,
(3) Le milieu aquatique,
(4) La surveillance et la gestion durable des Infrastructures.

*__LABORATOIRE D'ACCUEIL__

Nom du laboratoire d'accueil : LISIC

Nombre de HDR dans le laboratoire : 15

Nombre de thèses encadrées dans le laboratoire (rentrée 2017) : 24 (dont 8 cotutelles)

Durée moyenne des thèses soutenues dans le laboratoire, sur la période 2013-2017 : 3 ans et 3 mois

ENCADREMENT

Nom, Prénom du directeur de laboratoire : **Christophe Renaud**

Nom, Prénom du directeur de thèse (si différent du directeur de laboratoire) :
Gilles ROUSSEL

Nombre de doctorats en préparation sous la direction du directeur de thèse : **1**

Avis détaillé du directeur de thèse :

Fort d'une expérience importante de développement des méthodes de factorisation pour la séparation de sources (3 thèses soutenues en 3 ans) et de collaborations transversales avec les spécialistes de la chimie de l'environnement, nous cherchons à accroître encore davantage la robustesse des algorithmes, la manière d'introduire l'expertise humaine et l'information disponible dans les problèmes d'inversion des données environnementales. Cette thèse s'inscrit pleinement dans cet objectif en associant le paradigme bayésien pour l'introduction de contraintes expertes souples, mais aussi d'appliquer la problématique de caractérisation des sources de pollution (en particulier en particules) dans une région industrielle du Liban. Pour cela, nous nous appuyons sur des compétences locales grâce à des partenariats déjà établis (soutenance récente d'une thèse en co-tutelle avec l'université de Balamand, collaborations UCEIV et Institut de l'environnement (IOE)).

Signature du directeur de thèse

Avis détaillé du directeur de laboratoire :

Le sujet proposé s'appuie sur une expertise reconnue de l'équipe d'accueil dans le domaine de l'identification de sources de pollution via des techniques de factorisation matricielle NMF. Il vise à accroître la robustesse des méthodes actuellement développées et à appliquer leurs résultats à la fois sur notre territoire et sur un territoire similaire au point de vue industriel au Liban. Les encadrants disposent de collaborations établies avec nos collègues libanais, qui souhaitent poursuivre les travaux sur le sujet de thèse proposé. Des collaborations existent également avec nos collègues de l'UCEIV, qui sont gages de sérieux quant aux données qui peuvent être récoltées et interprétées. J'émets donc un avis très favorable au financement de ce sujet.

Signature du directeur de laboratoire

PROJET DE THESE
Intitulé du projet de thèse :

Méthodes de NMF Bayésiennes informées pour la séparation de sources

Domaine scientifique : Traitement des signaux

Résumé (1/2 page)
Le démélange de sources pour la pollution de l'air peut être formulé comme un problème de NMF en décomposant la matrice d'observation X en le produit de deux matrices non négatives G et F, respectivement la matrice de contributions et de profils. Généralement, les données chimiques sont entâchées d'une part de données aberrantes. En dépit de l'intérêt de la communauté pour les méthodes de NMF, elles souffrent d'un manque de robustesse à un faible nombre de données aberrantes et aux conditions initiales et elles fournissent habituellement de multiples minimas.
En rupture par rapport aux méthodes traditionnelles de factorisations matricielles cette thèse est orientée d'une part vers les NMF informées utilsant les connaissances expertes sous l'angle global des méthodes Bayesiennes, d'autre part en ayant à l'esprit le fait de rendre la NMF robuste même lorsque le nombre d'échantillons est relativement faible. Deux types de connaissances ont été déjà par le passé introduites dans la matrice de profil F. La première hypothèse est la connaissance exacte de certaines composantes de la matrice F tandis que la deuxième information utilise la propriété de somme-à-1 de chaque ligne de la matrice F. Une paramétrisation qui tient compte de ces deux informations a été développée au préalable mais l'objectif ici va consister en la définition d'*a priori* statistiques sur le profil qui permettront de calculer un a posteriori pour les profils ainsi que pour les contribution.

L'application cible consiste à identifier les sources de particules dans l'air dans la région côtière du Nord Beyrouth au Liban. Le challenge essentiel réside dans la pertinence potentielle des méthodes Bayesiennes pour estimer les sources en présence.

Projet de thèse (5 pages maxi.) :
Développer sur cinq pages :

- ***Le sujet de recherche choisi et son contexte scientifique***
- ***L'état du sujet dans le laboratoire et l'équipe d'accueil***
- ***Le programme et l'échéancier de travail***
- ***Les retombées scientifiques et économiques attendues***
- ***Les collaborations prévues et une liste de 10 publications maximum portant directement sur le sujet***

Contexte :

Les méthodes de type Factorisation Matricielles Non négatives (NMF) sont actuellement très en vogue dans la communauté traitement du signal de par la diversité des applications qu'elles peuvent concerner. Ces méthodes consistent à approcher la matrice de données en un produit de deux facteurs de dimensions plus faibles, vérifiant la non-négativité de leurs éléments, appelés respectivement la matrice de sources et la matrice de contribution. Cependant, la plupart de ces méthodes ne présentent pas toutes des garanties satisfaisantes de stabilité et d'interprétation des résultats concernant les différents facteurs. Dans certaines situations, en effet, les lignes de la matrice des sources peuvent être géométriquement proches, ce qui rend les méthodes classiques peu efficaces. Par ailleurs, les données réelles sont entachées de mesures aberrantes.

A cette fin, nous avons proposé, dans des travaux récents, des approches dites informées dans lesquelles les informations des experts du domaines sont considérées comme des contraintes dures du problème de NMF [1-4]. Ces informations peuvent provenir de la connaissance experte de certaines sources contenues dans la matrice des sources (contraintes égalité [1,2] ou de bornitude [3]). Par ailleurs, nous avons récemment développé des méthodes capables de résister à des points aberrants présents dans les données [4]. Extrêmement flexibles, ces approches de NMF informées ont été appliquées à la séparation des polluants atmosphériques par l'équipe du Professeur Courcot (UCEIV-ULCO). En particulier, dans le cadre sa thèse de doctorat, C. Roche a utilisé nos outils pour mesurer l'**impact du trafic maritime** sur la pollution globale affectant la **population** vivant sur la **Côte d'Opale** (Région Nord Pas de Calais, projet ECUME financé par la DREAL).

Nous proposons dans le cadre de cette thèse d'étendre ces notions de factorisations collectives à des situations où la matrice des sources contient uniquement une partie commune, conduisant à considérer des couplages plus flexibles entre les différentes matrices à co-factoriser tout en permettant l'ajout de connaissances expertes telles que proposées dans nos précédentes approches de NMF informée [1-4]. Les approches envisagées utiliseront la statistique Bayesienne [5] afin de mieux cerner les profils a posteriori.

D'un point de vue applicatif, ces méthodes sont appliquées à la pollution de l'air dans l'environnement marin et terrestre. L'enjeu principal consiste en un diagnostic fin de la matrice des sources de pollution. Cette matrice rassemble les signatures chimiques des différentes sources en présence. L'apport de la nouvelle méthodologie permettra de prendre en compte plusieurs sites urbains ou ruraux de manière simultanée en considérant que certaines sources sont communes à l'ensemble des sites.

Du point de vue applicatif, ce projet vise à étudier la contribution des différentes industries sur différents sites de topologie différentes. L'exploitation simultanée des deux sites de mesure (ou plus) nécessitera donc la mise en œuvre de nouvelles méthodes dites de co-factorisation.

Fort de notre expérience dans le Nord-pas de Calais, nous souhaitons étendre notre expertise sur des sites de topologies différentes au Liban, situé au nord de Beyrouth. Ces deux sites pourront être traités ensemble pour séparer l'influence des différentes sources.

Travail proposé :

La thématique générale de cette thèse est le traitement du signal. Le travail du doctorant consistera dans un premier temps à proposer des extensions informées des méthodes avancées de NMF à la co-factorisation partielle. Ce travail pourra être abordé dans un premier temps sous forme d'optimisation déterministe, faisant suite aux travaux menés dans [1,2]. Dans un second temps, il est envisagé d'aborder la factorisation collective sous l'angle Bayesien [5], permettant ainsi de rendre plus flexible la recherche tout en intégrant les informations des experts.

Etat du sujet dans le Laboratoire et l'équipe d'accueil :

Le travail proposé constitue un prolongement scientifique de deux thèses, la première financée par Arcelor Mittal soutenue en 2014 (A. Limem) et d'une 2nde thèse (R. Chreiky, 19 décembre 2017) sur la NMF informée (cotuelle avec l'Université de Balamand). Une troisième thèse utilise des méthods de factorisation pour la calibration aveugle des d'une foule de capteur pour le crowdsensing. Parmi les 10 publications en référence, 5 d'entre elles concernent directement la thématique abordée.

Echéancier de travail :

L'échéancier suivant est proposé :

Octobre 2018-Octobre 2019 :

Etude bibliographique sur les méthodes de factorisation, et d'autre part les méthodes Bayésiennes. Compréhension de la problématique environnementale.

Octobre 2019- Octobre 2020 :

Implémentation des a priori intégrant des contraintes souples. Développement de solutions de NMF Baysésiennes. Intégration des données réelles provenant de nos partenaires.

Octobre 2020- Octobre 2021 :

Comparaison avec des méthodes de référence. Validation

Rédaction de la thèse.

Retombées économiques et scientifiques attendues

L'analyse du district de Koura au Nord Liban représentée sur la carte représenté avec deux sites de mesures aux caractéristiques différentes (l'un urbain, l'autre rural) permettra de mieux comprendre les influences des différentes sources en présence dans le secteur. Ce site comprend en particulier quelques industries lourdes comme les cimenteries et une usine de phosphate. En dehors de ces sources, une influence marine est aussi attendue.

Différentes sources en présence dans la région de Zakroun.

La décomposition prévue permettra de reconstruire les concentrations en les différents sites en fonction des différents émetteurs recensés. Elle permettra donc d'expliquer les origines des pollutions et procurer aux acteurs de la région les éléments de décision permettant de réduire la pollution.

Collaborations prévues :

Nous travaillons actuellement avec l'équipe du Professeur Courcot (UCEIV/ULCO) qui prend en charge la partie analyse chimique des échantillons, exploite et valide les résultats de nos méthodes de traitement.

Dans la continuité de la cotutelle de thèse de Robert Chreiky, soutenue le 19/12/2017, nous renouvelons notre collaboration avec le professeur Antoine Abche (Université de Balamand, Tripoli) qui a une expertise des méthodes Bayesiennes.

Enfin, nous prévoyons une collaboration qui vise à confier une partie des analyses chimiques à l'Institut d'environnement (IOE) de Balamand, dirigé par le Professeur Manal Nader.

Références citées :

[1] A. Limem, G. Delmaire, M. Puigt, G. Roussel, and D. Courcot, *Non-negative matrix factorization under equality constraints–a study of industrial source identification*, Applied Numerical Mathematics (APNUM), Volume 85, pp. 1-15, November 2014.

[2] R. Chreiky, G. Delmaire, M. Puigt, G. Roussel, D. Courcot, A. Abche. Split Gradient Method for Informed Non-negative Matrix Factorization, Proc. of LVA/ICA, pp. 376-383, Liberec, Czech Republic, August 25-28, 2015.

[3] A. Limem, M. Puigt, G. Delmaire, G. Roussel, D. Courcot. Bound constrained weighted NMF for industrial source apportionment. Proc. of MLSP, Reims, France, September 21-24, 2014.

[4] *R. Chreiky, G. Delmaire, C. Dorffer, M. Puigt, G. Roussel, A. Abche*. Robust Informed Split Gradient NMF Using Alpha Beta-Divergence For Source Apportionment. *IEEE International Workshop on Machine Learning for Signal Processing, (MLSP 2016), Vietri Sul Mare, September 13-16, 2016.*

[5] N. Dobigeon, S. Moussaoui, J. Y. Tourneret, C. Carteret, Bayesian separation of spectral sources under non-negativity and full additivity constraints, Signal Processing 89, 2009, pp 2657-2669.

[6] H. Lee and S. Choi. Group nonnegative matrix factorization for EEG classification. Proc. of AISTATS. vol. 5, pp. 320-327. JMLR - Proceedings Track, 2009.

[7] J. Yoo, M. Kim, K. Kang, and S. Choi. Nonnegative matrix partial co-factorization for drum source separation. In: Proc. of ICASSP. pp. 1942-1945, March 2010.

[8] J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. Proc. of SDM. vol. 13, pp. 252-260, 2013.

[9] N. Seichepine, S. Essid, C. Févotte, and O. Cappé. Soft non-negative matrix co-factorization. IEEE Trans. on Sig. Process., 62(22):5940–5949, Nov. 2014.

[10] R. Cabral Farias, J. E. Cohen, C. Jutten, and P. Comon, Joint decompositions with flexible couplings, Proc. of LVA/ICA, Liberec, Czech Republic, Aug. 2015.