Titre : Méthodes étendues de factorisation informée de matrices ou tenseurs (semi-)non-négatifs pour l'analyse de données incomplètes et de grande dimension. Application au traitement de données issues du *mobile crowdsensing*.

Financement prévu : Ministère
Cofinancement éventuel : Région/ADEME
(Co)-Directeur de thèse : Gilles ROUSSEL
E-mail : gilles.roussel@univ-littoral.fr
Encadrant : Matthieu PUIGT
E-mail : matthieu.puigt@univ-littoral.fr

Laboratoire : LISIC (Laboratoire d'Informatique Signal et Image de la Côte d'Opale, EA 4491)
Equipe : SPeciFI

I. Introduction :

Nous avons récemment considéré le traitement de signaux issus du *mobile crowdsensing*, c'est à dire des données datées et géolicalisées recueillies par une foule d'appareils mobiles (de type smartphones). Ce type d'application, en lien avec la **mesure ubiquitaire**, l'**internet des objets** et les données massives dans les **villes intelligentes**, présente des problématiques d'intérêt pour les communautés signal et informatique. Nous travaillons actuellement sur un projet de recherche citoyenne en partenariat avec l'Inria Lille – Nord Europe et les associations ATMO HdF et BES. L'objectif de ce projet est la **surveillance de la qualité de l'air par une foule de capteurs mobiles** fabriqués et portés par des citoyens volontaires. L'une des problématiques clés d'un tel type d'acquisition provient du fait que les **capteurs ne peuvent pas être étalonnés en laboratoire** avant et pendant leur déploiement. Nos travaux actuels proposent de traiter ce problème d'étalonnage sous forme d'une factorisation matricielle à données manquantes.

Les **approches de factorisation matricielle** connaissent depuis une quinzaine d'années un énorme intérêt de la part de la communauté scientifique. En effet, de **nombreux problèmes** tels que la localisation et/ou la séparation aveugle de sources, l'apprentissage de dictionnaire, l'acquisition comprimée, la complétion de matrices, l'analyse archétypale [1] peuvent s'écrire sous la forme d'une factorisation matricielle, pour laquelle diverses contraintes propres au problème considéré sont prises en compte. Compte tenu de la diversité de ces problèmes et de leur intérêt applicatif, des travaux significatifs ont été proposés à la fois en **mathématiques appliquées**, en **informatique** et en **traitement du signal et des images**.

Parmi les outils les plus développés, les approches de factorisation en matrices non-négatives (NMF pour *Non-negative Matrix Factorization* en anglais) connaissent un vif intérêt depuis le célèbre article de Lee et Seung paru dans Nature [2] et une multitude d'approches permettant de traiter ce problème ont été proposées. Les principaux axes de recherche concernent l'ajout de contraintes (de type parcimonie, somme-à-un, orthogonalité) [3], l'utilisation de fonctions de coûts robustes [4], l'accélération des calculs par des méthodes d'optimisation rapide ou par distribution des calculs [5], ou encore l'étude des conditions garantissant l'unicité et/ou l'exactitude de la factorisation [6]. Ces approches ont de plus été étendues au cas de factorisation en matrices semi-non-négatives (Semi-NMF) [7] et en tenseurs (semi-)non-négatifs [5].

Comme expliqué précédemment, nous travaillons actuellemnet sur un projet de traitement de données issues du *mobile crowdsensing*. Comme expliqué précédemment, les **capteurs ne peuvent pas être étalonnés en laboratoire** avant et pendant leur déploiement. Revisitant l'idée de rendez-vous de capteurs [8], qui est une hypothèse standard en étalonnage de capteurs mobiles, nous avons proposé des **approches d'étalonnage aveugle de capteurs et d'estimation des grandeurs physiques mesurées** qui s'expriment sous forme d'une factorisation matricielle informée [9-11]. Les informations incorporées dans la factorisation consistent notamment en une structure particulière d'un facteur matriciel (due à la nature du problème) et à des a priori de parcimonie des grandeurs physiques mesurées fournissant une paramétrisation spécifique du second facteur.

II. Proposition de sujet de thèse :

Dans le cadre du projet OSCAR, les approches développées jusqu'alors supposent que la masse de données à traiter est de taille « raisonnable ». Or, nous savons que, dans la mouvance de l'**internet des objets** et des *Big Data* dans lequel ce projet s'insère, nos approches atteindront leurs limites calculatoires et ne pourront fournir en un temps raisonnable des résultats.
Il est donc nécessaire d'étendre ces méthodes à un formalisme permettant de gérer des matrices de très grande taille. Dans le cadre de cette thèse, nous chercherons à étendre les approches informées [9-11] pour traiter des problèmes de grande taille, par exemple via des méthodes d'optimisation rapides, distribuées et/ou randomisées.

Ces approches rapides et informées pourront être par ailleurs étendues au cas où les données collectées par *mobile crodwsendsing* seront stockées sous forme de tenseur à données manquantes.

Mots clés : factorisation en matrices (semi-)non-négatives ; factorisation tensorielle (semi-)non-négative ; données incomplètes ; données de grande dimension ; factorisation informée unique et/ou exacte ; big data ; réseau mobile de capteurs ; internet des objets ; troisième révolution industrielle ; *mobile crowdsensing* ; surveillance environnementale.

III. Références :

[1] I. Carron, The advanced matrix factorization jungle *[Internet]*, 2011. Available from: https://sites.google.com/site/igorcarron2/matrixfactorizations
[2] D. D. Lee, H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature* 401 (6755), 788-791, 1999.
[3] Y. X. Wang and Y. J. Zhang, "Nonnegative Matrix Factorization: A Comprehensive Review," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336-1353, June 2013.
[4] Cichocki, A., Lee, H., Kim, Y. D., & Choi, S. (2008). Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, *29*(9), 1433-1440.
[5] G. Zhou, A. Cichocki, Q. Zhao and S. Xie, "Nonnegative Matrix and Tensor Factorizations : An algorithmic perspective," in *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 54-65, May 2014.
[6] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, M. Zhu, "A practical algorithm for topic modeling with provable guarantees," in *Proc. of ICML*, 2013.
[7] C Ding, T Li, MI Jordan, Convex and semi-nonnegative matrix factorizations, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32, 45-55, 2010
[8] O. Saukh, D. Hasenfratz, C. Walser, L. Thiele, "On Rendezvous in Mobile Sensing Networks," in *Proc. of RealWSN*, 2013.
[9] C. Dorffer, M. Puigt, G. Delmaire, G. Roussel, "Blind Calibration of Mobile Sensors Using Informed Nonnegative Matrix Factorization," in *Proc. of LVA/ICA 2015*, vol. LNCS 9237, pp. 497-505, 2015.
[10] C. Dorffer, M. Puigt, G. Delmaire, G. Roussel, "Blind Mobile Sensor Calibration Using an Informed Nonnegative Matrix Factorization With a Relaxed Rendezvous Model," in *Proc. of ICASSP*, pp. 2941-2945, 2016.
[11] C. Dorffer, M. Puigt, G. Delmaire, G. Roussel, "Nonlinear Mobile Sensor Calibration Using Informed Semi-Nonnegative Matrix Factorization With a Vandermonde Factor," in *Proc. of IEEE SAM*, 2016.

IV. Publications de l'équipe en lien avec le sujet :

Outre les références [9-11], l'équipe a récémment publié en lien avec le sujet de thèse :

- A. Limem, G. Delmaire, M. Puigt, G. Roussel, 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.
- C. Dorffer, M. Puigt, G. Delmaire, G. Roussel, "Fast nonnegative matrix factorization and completion using Nesterov iterations," to appear in *Proc. of LVA/ICA*, 2017.