Université d'Angers–University of Wrocław–Politechnika Varsovie – Institut de Cancérologie de l'Ouest, Angers-Nantes

Projet PhD 2020
Développement d'un modèle SLOPE graphique coloré pour l'analyse de données massives des cancers du sein

Directeur de la thèse : PIOTR GRACZYK (LAREMA Mathématiques, Université d'Angers)
Co-encadrants :
Agnès BASSEVILLE (ICO Angers),
Malgorzata BOGDAN (Université Wroclaw)
Bartosz KOLODZIEJEK (Polytechnique Varsovie)

Contact: Piotr.Graczyk@univ-angers.fr

Description du sujet de thèse

Contexte

En santé comme dans beaucoup de domaines, l'augmentation de la production de données combinée à l'amélioration des performances en informatique a permis la création de banques de données massives nommées big data. L'analyse de ces données massives n'est pas possible par les tests statistiques classiques, trop simples pour déchiffrer ces données multivariées, et nécessite l'utilisation d'algorithmes plus complexes basés sur l'apprentissage statistique (machine learning). Dans le cas d'apprentissage supervisé, les algorithmes sont entrainés sur un nombre d'observation suffisamment important pour lequel les variables et le phénotype associé sont connus. Une fois le modèle établi par entrainement/apprentissage, il peut être appliqué à des données pour lesquels les mêmes variables sont connues, afin de prédire un phénotype encore inconnu, le but étant de minimiser l'erreur de prédiction. En santé, la tâche est particulièrement ardue car le nombre de variables (gènes, protéines, etc) dépasse de beaucoup le nombre d'observations (les individus). On parle alors de données de grandes dimensions, qui pour l'instant génèrent des modèles dont les performances de prédiction ne sont pas assez bonnes pour être utilisables en clinique. Dans le but d'améliorer l'analyse de données de grandes dimensions en santé, le laboratoire LAREMA s'est associé avec l'Institut de Cancérologie de l'Ouest (ICO) dans un projet d'analyse de big data dans les cancers du sein par une nouvelle approche combinant un algorithme de machine learning adapté aux grandes dimensions (SLOPE, pour Sorted L-One Penalized Estimator) avec un modèle statistique graphique illustrant les dépendances entre variables.

SLOPE (Bogdan et al., 2015) est une amélioration du LASSO (least absolute shrinkage and selection operator), un algorithme de machine learning utilisant la régression pénalisée pour sélectionner les variables d'intérêt et construire un modèle statistique. SLOPE se base sur une nouvelle méthode de pénalisation qui privilégie une meilleure sélection du modèle (par control du taux des fausses découvertes) plutôt que la minimisation des erreurs de prédiction du modèle.

Les modèles statistiques graphiques (graphical models) (Lauritzen, 1996, Graczyk et al., 2019) sont quant à eux utilisés en machine learning non supervisé afin de comprendre les interactions entre variables. Ces modèles utilisent la matrice inverse de covariance (ou matrice de précision) comme un outil d'apprentissage non supervisé afin détablir les relations entre variables issues de données complexes sous la forme d'un graphique simple non orienté révélant les interactions significatives. Ainsi, le LASSO graphique (développé par Friedman et al., 2008) utilise les méthodes de pénalisation du LASSO appliqué à un modèle graphique pour une analyse non supervisée des variables.

De la même manière, le SLOPE graphique a été récemment développé dans le laboratoire du Pr Bogdan (Sobczyk PhD manuscript, 2019), et les résultats préliminaires suggèrent que le SLOPE graphique serait plus robuste que le LASSO graphique. Par ailleurs, afin d'améliorer les modèles graphiques pour les données de grandes dimensions (entre autre), Hojsgaard et Lauritzen (2008) ont introduit une nouvelle classe de modèles graphiques par ajout de restrictions sur certains coefficients de la matrice de précision. Ces modèles sont représentés par graphes colorés. Les modèles statistiques graphiques colorés ainsi créés ont déjà été validés dans plusieurs domaines statistiques, mais leur utilisation en sélection de modèles n'a pas encore donné de résultat satisfaisant (Gehrmann, 2011, Massam et al., 2018). Ils sont par exemple inadaptés pour le LASSO, car cet algorithme génère une grande sparsité (c'est-à-dire l'existence de nombreux zéros) dans la matrice de précision. A contrario, il a été observé empiriquement que le SLOPE regroupait ensemble les variables de même influence en leur attribuant une valeur identique mais ne générait pas de sparsité. Le développement d'un modèle SLOPE graphique coloré est une voie pour l'instant inexplorée qui semble donc particulièrement adaptée dans un contexte d'analyse de données de santé de grandes dimensions.

Objectifs.

L'objectif de ce projet est de découvrir les variables d'intérêt (nommées biomarqueurs en santé) permettant de prédire la réponse au traitement dans les cancers du sein grâce à une méthode novatrice d'analyse de données massives adaptées aux données de grandes dimensions. Pour cela, nous proposons un projet collaboratif et pluridisciplinaire entre le laboratoire de mathématique LAREMA à Angers (compétence en modèles graphiques), l'Université Wroclaw en Pologne (co-créateur de l'algorithme SLOPE), l'Ecole Polytechnique de Varsovie en Pologne (compétence en modèle graphique coloré) et l'Institut de Cancérologie de l'Ouest (application du modèle en cancérologie). Cette analyse sera basée sur des données issues de biopsies de cancer du sein réalisées chez des patientes avant traitement (qui ont permis de quantifier l'expression des 20000 gènes de la tumeur). Le suivi des patientes après traitement permet de comptabiliser le temps durant lequel la tumeur ne grossit pas (survie sans progression). Cette donnée, également recueillie, sera utilisée comme indicateur de réponse au traitement.

Programme prévisionnel.

Dans une première étape, le doctorant devra s'approprier les méthodes de SLOPE Graphique ainsi que les méthodes de modèles graphiques et de modèles graphiques colorés, en les adaptant aux spécificités des données d'expression génique. Cette première étape se basera sur lanalyse des données issues de la banque de donnée publiques TCGA (The Cancer Genome Atlas).

Dans une deuxième étape, le doctorant mettra en place le SLOPE graphique coloré à partir des mêmes données d'expression géniques issues du TCGA. Il établira également son fondement mathématique par analyse théorique.

Dans une troisième étape, le doctorant appliquera le modèle SLOPE graphique coloré aux données provenant de 82 patientes suivies à l'Institut de Cancérologie de l'Ouest afin de comparer les résultats obtenus et valider les variables importantes (validation externe). La pertinence des variables sélectionnées par le modèle en tant que biomarqueur pourra être évaluée en partenariat avec les cliniciens de l'ICO.

Résultats et valorisation attendus.

Les résultats de ce projet permettront tout d'abord l'apport d'une nouvelle méthode statistique dans l'analyse de données de grandes dimensions.

De plus, l'identification des variables d'intérêt à partir d'un modèle basé sur l'analyse de tumeurs du sein permettra de sélectionner des biomarqueurs candidats qui pourraient être utilisées en clinique. Ces biomarqueurs permettraient à terme de classifier de nouveaux sous-types de cancer afin offrir des traitements adaptés au cas par cas chez les patientes, permettant ainsi de promouvoir une médecine de précision.

Pour finir, ces variables sélectionnées et l'identification du lien existant entre elles pourraient également permettre d'identifier de nouvelles cibles thérapeutiques par analyse du rapport existant entre leur fonction biologique (rôle dans le développement tumoral) et leur implication entre elles et dans la réponse au traitement.

Les résultats de ce projet en big data et de ses applications cliniques seront présentés lors d'évènements destinés à un public non scientifique tels que Ma Thèse en 180s et la Fête de la Science.

Planning prévisionnel de la thèse

1ère année :

- Appropriation des connaissances en génomique et cancérologie et préparation de la base de données (2 mois)
- Appropriation (connaissance théorique et code) des méthodes de SLOPE (4 mois)
- Appropriation des Modèles Statistiques Graphiques (4 mois)
- Appropriation des Modèles Statistiques Graphiques colorés (2 mois)

2ème année :

- Création de la méthode du SLOPE graphique colorés à partir des données du TCGA (6 mois)
- Etablissement de son fondement mathématique (6 mois)

3ème année :

- Validation externe du modèle sur des données de patientes traitées à l'ICO (4 mois)
- Rédaction de publication et réponse aux rapporteurs (4 mois)
- Rédaction de la thèse (3 mois)
- Réponse aux rapporteurs de thèse, préparation de la soutenance (1 mois)

Inscription du travail du doctorant dans la stratégie actuelle de recherche

Ce projet de doctorat s'intègre pleinement dans la politique scientifique actuelle de l'Université d'Angers et du Laboratoire LAREMA pour le développement de projets rattachés à l'analyse des big data, ainsi que dans le développement des collaborations et projets interdisciplinaires et internationaux dans le domaine des big data. L'Université d'Angers a adopté une stratégie globale d'investissement en matière de traitement des données, incluant la recherche, la formation initiale et continue, et les équipements. L'Université d'Angers fait de ce fait partie des lauréats retenus par le ministère de l'enseignement supérieur et de la recherche pour développer un ensemble de formations continues, courtes ou diplômantes, dans le domaine des big data. Depuis 2018, l'UA finance un projet MIR "Big Data et Médecine", dirigé par Dr Graczyk, en collaboration avec Dr Bogdan et son équipe de l'Université de Wroclaw. Cette proposition de doctorat s'inscrit pleinement dans ce projet en fournissant des outils mathématiques pour l'analyse de données massives dans les cancers du sein.

Ce projet de thèse permettra également de renforcer le partenariat avec l'Institut de Cancérologie de l'Ouest qui développe des projets collaboratifs pour l'analyse des données de santé. En particulier, l'ICO a obtenu deux financements de la Commission Européenne pour développer l'analyse des données massives pour améliorer les traitements du cancer du sein (projets Epicure https://projet-epicure.fr/) et PredAlgoBC https://cordis.europa.eu/project/id/841313)

Bibliographie:

- Bogdan M., van den Berg, E., Sabatti C., Su W., and Candes, E. J. SLOPE - adaptive variable selection via convex optimization. Annals of Applied Statistics, 9(3):1103-1140, 2015.

- Lauritzen, S.L. Graphical Models; Clarendon Press: Oxford, UK, 1996

- Graczyk P., Ishi H., and Kolodziejek B. Wishart laws and variance function on homogeneous cones, to appear in Prob. Mathematical Stat., pp. 1-24, 2019

- Friedman, J, Hastie, T, and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432-441, 2008.

- Hojsgaard, S., and Lauritzen, S.L. Graphical Gaussian models with edge and vertex symmetries. J. R. Stat. Soc. Ser. B, 70, 1005-1027, 2008

- Massam H., Li Q., and Gao X. Bayesian precision and covariance matrix estimation for graphical Gaussian models with edge and vertex symmetries, Biometrika, Volume 105, Issue 2, 371-388, 2018

- Gehrmann H. Lattices of graphical Gaussian models with symmetries. Symmetry 3, 653-79, 2011

Université d'Angers–University of Wrocław–Politechnika Varsovie – Institut de Cancérologie de l'Ouest, Angers-Nantes

PhD Project 2020
SHORT VERSION IN ENGLISH
Graphical SLOPE for coloured graphical models with applications in genetics and medecine

Co-Directors of the PhD Thesis :

Piotr Graczyk(Université d'Angers)

Małgorzata Bogdan(University of Wrocław)

Co-Tutors of the PhD Thesis :

Agnès Basseville (Institut de Cancérologie Ouest, Angers-Nantes)

Bartosz Kołodziejek (Polytechnique Warsaw)

Financial support: The PhD fellowship (approx. 1050E/month) will be financed by

- bourse récurrente LAREMA / CNRS
- Université d'Angers
- Erasmus fellowships for short stays in Varsovie and Wroclaw
- Possibility of a thesis in co-tutelle France-Poland. Then a co-tutelle fellowship may be granted.

Contact e-mail : graczyk@univ-angers.fr

Description of the subject:

Estimation of the sparse inverse covariance matrix of multivariate Gaussian variables (precision matrix) has been studied quite actively in recent years. Actually, the inverse covariance matrix provides a practical tool of unsupervised learning, in order to understand statistical relations of variables in complex data in the form of a simple undirected graph, which often reveals meaningful interactions of genes, illness factors etc. This is the object of the **statistical theory of Graphical Models** ([7, 6]).

SLOPE([3]) is a substantial improvement of popular Lasso methods. Very recently, the Graphical Lasso methods of graphical model selection ([4],[10],[12]) have been essentially extended and strengthened by the methods of **Graphical SLOPE** ([2],[11]). Graphical SLOPE proves to have much higher power at the cost of treating few zero entries in precision matrix as non-zeros, i.e. introducing small number of false discoveries. The false discovery rate (FDR) is controlled by Graphical SLOPE in various scenarios.

In order to make Graphical Gaussian Models a viable modelling tool in the modern Big Data Science, i.e. when the number of variables outgrows the number of observations, Højsgaard and Lauritzen([8]) introduced in 2008 model classes which place equality restrictions on precision matrix terms or partial correlations. The models can be represented by vertex and edge **coloured** graphs. This started **the statistical theory of Coloured Graphical Models**. The estimation theory for Coloured Graphical Models is well established ([8]), whereas the Model Selection within the Coloured Graphical Models class is still not satisfactory ([5], [9]).

In particular, no Lasso-type methods exist for Coloured Graphical Models, since Lasso generates a great level of sparsity(0 terms in the precision matrix). Graphical SLOPE, instead, has a strong tendency to average similar elements of the precision matrix and in this way it naturally generates Coloured Graphical Models.

The main objectives of this PhD Project are:

1. to develop Graphical SLOPE methods for Coloured Graphical models, taking into account the specifics of genetics and medicine data

2. to apply Graphical SLOPE methods for Coloured Graphical models for model selection in genetical and medical data

3. to establish statistical guarantees for the Graphical SLOPE for Coloured Graphical models

The applications in genetics/medecine/biology will be done within the SIRIC ILIAD (INCa-DGOS-Inserm 12558) programme, with scientific advice of Agnès Basseville (Institut de Cancérologie Ouest, Angers-Nantes), researcher in genetics/medecine/biology of this programme. The first application could be done to TCGA(The Cancer Genome Atlas) data analysed by other methods in [1].

Prerequsites. Master in Mathematics, Applied Mathematics or Data Sciences.
Teaching duty. None.

References

[1] M. Bogdan, P. Graczyk, F. Panloup, V. Seegers, P. Sobczyk, S. Wilczynski, VARCLUST: MATHEMATICAL BASES AND APPLICATIONS, https://www.overleaf.com/15782965ngbdnqpfyxpd, January 2019.

[2] Bogdan M, Lee S., Sobczyk P., Sparse Inverse Covariance Matrix Estimation with Graphical SLOPE, Technical report, 2018

[3] Bogdan M., van den Berg, E., Sabatti C., Su W. , and Candes, E. J. SLOPE - adaptive variable selection via convex optimization. Annals of Applied Statistics, 9(3):1103-1140, 2015.

[4] Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432-441, 2008.

[5] Gehrmann H. (2011). Lattices of graphical Gaussian models with symmetries. Symmetry 3, 653-79

[6] P. Graczyk, H. Ishi et B. Kołodziejek, Wishart laws and variance function on homogeneous cones, to appear in Prob. Mathematical Stat. (2019), pp. 1-24.

[7] Lauritzen, S.L. Graphical Models; Clarendon Press: Oxford, UK, 1996

[8] Hojsgaard, S.; Lauritzen, S.L. Graphical Gaussian models with edge and vertex symmetries. J. R. Stat. Soc. Ser. B 2008, 70, 1005-1027.

[9] H Massam, Q. Li, X. Gao Bayesian precision and covariance matrix estimation for graphical Gaussian models with edge and vertex symmetries, Biometrika, Volume 105, Issue 2, 1 June 2018, 371-388

[10] Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. Annals of Statistics, 34:1436-1462, 2006.

[11] Sobczyk P., Identifying low-dimensional structures through model selection in high-dimensional data, PhD Thesis 2019.

[12] Yuan, Ming. High dimensional inverse covariance matrix estimation via linear programming. Journal of Machine Learning Research, 11:2261-2286, 2010.