Université d'Angers – University of Wrocław – Polytechnique Warsaw – Institut de Cancérologie Ouest, Angers-Nantes

PhD Project 2019
Graphical SLOPE for coloured graphical models with applications in genetics and medicine

Co-Directors of the PhD Thesis: Małgorzata Bogdan(University of Wrocław),
Piotr Graczyk(Université d'Angers)
Co-Tutors of the PhD Thesis : Bartosz Kołodziejek (Polytechnique Warsaw)
Valérie Seegers (Institut de Cancérologie Ouest, Angers-Nantes)

Financial support: The PhD fellowship (approx. 1400EUR/month) will be co-financed by
• Project SIRIC ILIAD (INCa-DGOS-Inserm 12558) Nantes-Angers, 2018-2022.
(SIRIC=Sites de Recherche Intégrée sur le Cancer, ILIAD=Imaging and Longitudinal Investigations to Ameliorate Decision making in multiple myeloma and breast cancer. A grant directed by ICO=Institut Cancerologie Ouest-France, with 19 research institutes of Région Pays de la Loire)
• Université d'Angers
• Erasmus doctoral fellowships

Contact e-mail : Malgorzata.Bogdan@pwr.edu.pl, graczyk@univ-angers.fr

Description of the subject:

Estimation of the covariance matrix of multivariate Gaussian variables has been studied quite actively in recent years. Actually, the inverse covariance matrix (precision matrix) provides a practical tool for unsupervised learning. In order to understand statistical relations of variables in complex data one can visualize their interactions in the form of a simple undirected graph. The goal is to unravel 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 is controlled by Graphical SLOPE in various scenarios.

In order to make Graphical Gaussian Models a viable modeling 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 set equality restrictions on certain entries of covariance matrix or precision matrix. Such models can be represented by **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 (many zero entries in the precision matrix). Graphical SLOPE, instead, has a strong tendency to average similar elements of the precision matrix and in this way Coloured Graphical Models are very natural in such context.

The main objectives of this PhD Project are:

1. to establish statistical guarantees for the Graphical SLOPE,
2. to develop Graphical SLOPE methods for Coloured Graphical models, taking into account the specifics of genetics and medicine data,
3. to apply Graphical SLOPE methods for Coloured Graphical models for model selection in genetical and medical data.

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

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

References

[1] Bogdan, M., Graczyk, P., Panloup, F., Seegers, V., Sobczyk, P., Wilczynski, S. *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, *Ann. Appl. Stat.*, 9(3):1103-1140, 2015.

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

[5] Gehrmann, H. Lattices of graphical Gaussian models with symmetries. *Symmetry* 3:653-679, 2011.

[6] Graczyk, P., Ishi, H., and Kołodziejek, B. Wishart laws and variance function on homogeneous cones, to appear in *Probab. Math. Stat.*, pp. 1-24, 2019.

[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*, 70:1005-1027, 2008.

[9] Massam, H., Li, Q., 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.

[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, M. High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, 11:2261-2286, 2010.