Projects

MSc in Statistics for Smart Data – Introduction to graph analysis and modeling

Julien Chiquet, November the 28, 2017

Instruction

In each project, the trinome should read the journal paper and write a report on the general motivations, goals and methods used. You do not need to have a deep understanding of the technical aspects.

Then, you must illustrate these methods on a real world data set picked up on the internet. For each project, their exist an `R` package so you will not need to write much code[1]. The second part of your report will present the results of your analyses.

Your oral presentation will present your report, that should not exceed 10 pages long (but *well written*).

Send me the report back at julien.chiquet@inra.fr before the 17th of December at midnight.

Web resources

Network data (project 1 and 2).

Pick up some network data (with less than 500/1000 nodes for your convenience!) to illustrate the method that you study, for instance in

- Network repository http://networkrepository.com/
- General network data: http://www-personal.umich.edu/˜mejn/netdata/
- Ecological network database: http://networkrepository.com/eco.php
- SNAP database: https://snap.stanford.edu/data/index.html
- . . . *feel free to use your own network data.*

Gaussian multivariate data (project 3).

Find some Gaussian multivariate data (less that 100/200 variables for your convenience!)

- Protein expression data set: https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression
- you can use the `Ecoli.expr` dataset seen during the practical.
- Any expression data sets from http://biogps.org/dataset/
- . . . *feel free to use your own data.*

[1] expect for making some fancy figures!

Project 1 - extension of the SBM: beyond binary edges

This paper presents several extensions of the Stochastic Block Model where edges are weighted with various distributions (Poisson and Gaussian for instance). It also shows how one can include external knowledge on top of the network structure, by means of covariates on the nodes of the graph. All the corresponding model are implemented in the package **blockmodels**. Use it to analysis some weighted network data or binary network with covariates.

Journal paper. Mahendra Mariadassou, Stéphane Robin and Corinne Vacher. *Uncovering latent structure in valued graphs: a variational approach.* The Annals of Applied Statistics (2010): 715-742. https://arxiv.org/pdf/1011.1813.pdf

R package. https://CRAN.R-project.org/package=blockmodels, see also https://arxiv.org/abs/1602.07587

Project 2 - extension of the SBM: dynamic SBM

The first paper presents an extension of the Stochastic Block Model where memberships may vary across time. It is suitable to network data gathered in time. The second one is probably less techical and presents an application in ecology. The corresponding model are implemented in the package **dynsbm**. Use it to analysis some time-varying network data.

Journal papers. Catherine Matias and Vincent Miele. *Statistical clustering of temporal networks through a dynamic stochastic block model.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79.4 (2017): 1119-1141. https://arxiv.org/pdf/1506.07464

Miele, Vincent, and Catherine Matias. *Revealing the hidden structure of dynamic ecological networks.* Royal Society Open Science 4.6 (2017): 170251. http://rsos.royalsocietypublishing.org/content/4/6/170251

R package. https://CRAN.R-project.org/package=dynsbm

Project 3 - stability selection in sparse Gaussian graphical models

The stability selection is a general resampling scheme used to assess the robustness of any variable selection procedure. The paper presents this approach in case of sparse linear regression and sparse Gaussian graphical models. Stability selection is implemented in the package **stabs**. Use it to reconstruct a network where you only keep the most stable edges.

Journal paper. Nicolai Meinshausen and Peter Bühlmann. *Stability selection.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72.4 (2010): 417-473. https://arxiv.org/pdf/0809.2932

R package. https://CRAN.R-project.org/package=stabs