# MAP573 Group projects

October 2019

## Introduction

The present document aims at introducing the group projects for the course MAP573. The projects are conducted by groups of **5 students** (2 groups of 6) and result in a report and an oral presentation lasting 20 minutes. The oral presentation will be on Friday, the 6th of December and the report is required for the 20th of December so you will have time to improve your work, add analyses, etc. There may surely have mid-term lightning talks so that each group can present their work and the other groups can benefit from other topics. The day with all the defenses is mandatory for all students.

Some projects are suggested in the following list. You can also work on your own proposed project. Be aware that, in the latter case, the subject must beforehand be validated by the course coordinators. The projects may be related with statistics or data analysis. They may also be purely computational (time and space costs, algorithm's optimization) or methodological. Projects may also be deepenings of concepts that you have seen in class.

Once a subject chosen, you must **fill the following link with your names (first and last names), group number and subject**.

`https://docs.google.com/spreadsheets/d/1-dZGJF_93SRqHSndjDf1eXVEG6BuH5EQdv8Lv5KvwTo/edit?usp=sharing`.

One project can be taken by two groups, as some projects may easily be split.

## Causal inference to assess the effect of a treatment on survival (Julie Josse - Imke Mayer)

Causal inference aims at establishing causal relationships from observational data. It is an active field in statistics and have proved to give promising results in econometrics (policy evaluation), healthcare as well as marketing (to assess advertising campaigns) ... The main pitfall is that data was not collected following an experiment outline, which would have enabled to capture and control all sources of variability.

The two projects propsed are related with the larger Traumabase project[1]. One important question, among others, is to assess the effect of an intervention such as the administration of one or multiple treatments on survival by adjusting for confounding variables. Such an adjustment process is essential since the given treatment is more often prescribed to sick patients; whose survival rates are obviously lower than the ones of healthy patients. The study relies on data collected by the AP-HP (Paris hospitals). A confidentiality agreement is to be signed by the students to ensure privacy guarantees.

### Project 1: Matching

Matching techniques pair treated and non-treated patients to assess the effects of treatment on similar individuals. In this project, students will first thoroughly list and describe main matching techniques and existing software. Matching will then be implemented and evaluated on the AP-HP data set to assess the effect of tranexamic acid on survival for head trauma patients. Results obtained in this project can potentially be compared to the soon released results of a large clinical trial studying the same medical question.

#### References

- D. Rubin. Matching to remove bias in observational studies. Biometrics, 29:159–183, 1973.
- P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70:41–55, 1983.
- Iacus, Stefano M., Gary King, and Giuseppe Porro. "Causal inference without balance checking: Coarsened exact matching." Political analysis 20.1 (2012): 1-24.
- Introduction to causal inference (online course on coursera) https://www.coursera.org/learn/crash-course-in-causality
- Clinical randomization of an antifibrinolytic in significant head injury (CRASH 3 trial) https://crash3.lshtm.ac.uk/

[1] http://www.traumabase.eu/fr_FR

### Type of tools

Matching methods are often based on nearest neighbor type approaches but recent papers are based on optimization, optimal transport, etc.

## Project 2: Heterogeneous Treatment Effect HTE

The students will first compile a list of HTE techniques (random forests, with the package grf for causal forest, BART, TMLE, etc) and apply them on the AP-HP Traumabase dataset to assess the effect of tranexamic acid administration on patients' survival. Optionally, in addition to the HTE estimation, the students will have the possibility to analyze (average) treatment effects on subgroups of patients defined by a recent classification based on a brain lesion criterion validated by a team of medical experts. The estimated treatment effects will be compared to the results obtained with HTE estimation methods. The results of this study will potentially be compared to the soon released results of a large clinical trial studying the same medical question.

### References

- Susan Athey and Stefan Wager. Estimating Treatment Effects with Causal Forests: An Application. Observational Studies, 5, 2019.
- Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. Statistical Science, 34(1), pp.43-68, 2019.
- Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., & Gallego, B. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. Statistics in medicine, 37(23), 3309-3324, 2018.
- Introduction to causal inference (online course on coursera) `https://www.coursera.org/learn/crash-course-in-causality`
- Clinical randomization of an antifibrinolytic in significant head injury (CRASH 3 trial) `https://crash3.lshtm.ac.uk/`

## Project 3: Taskview causal inference

In R, there are task views `https://cran.r-project.org/web/views/` which aim at listing all the available packages on a subject and explaining their main functionality and scope. Such a task view does not exist for causal inference yet and would be very useful for future users. The aim of the project is twofold: (1) to provide a comprehensive overview of causal inference and (2) to assess the performance and versatility of the different packages on data sets (either simulated and benchmark data sets taken from the standard causal inference literature). Students can have a look at the yearly data challenge to investigate behavior of methods (`https://sites.google.com/view/ACIC2019DataChallenge/home`). We expect an output which include a Rmd file which can be used as a pipeline for users who want to compare methods (importation of the data, visualisation of the data, comparison of different causal inference methods, etc. ) If the project is well conducted, it will be published online.

**References**

- Two examples of attempts compiling a selection of software/packages allowing for causal analyses in R: `https://shiny.sund.ku.dk/zms499/` or in different programming languages: `https://github.com/rguo12/awesome-causality-algorithms`
- Introduction to causal inference (online course on coursera) `https://www.coursera.org/learn/crash-course-in-causality`

## Project 4: Assessing initial brain perfusion status with transcranial Doppler in patients with trauma brain injury: what contribution to predict outcome ? With Sophie Hamada, medical doctor and Olivier Auliard, Chief Data Scientist Capgemini

See document DTC traumabase version 12-11-2018.

## Project 5: Impact of the choice of the definition of a hemorrhagic shock on patient characteristics, consumption of care, patient's outcomes and methodological implications: a systematic review and registry based study, With Sophie Hamada, medical doctor and Olivier Auliard, Chief Data Scientist Capgemini

See document Project Summary Traumabase HS definitions

## Project 6: Other dimensionality reduction methods - Julie Josse

We have seen dimensionality reduction with PCA which is dedicated to quantitative data and linear relationships. Other methods are dedicated to other nature of variables or to nonlinear relationships. The aim of this project is to learn a or several new method and to apply it on real data.

### Methods

- Multiple Correspondence Analysis (method to analyse categorical variables such as survey data), compositional data
- Nonlinear PCA, , t-SNE (t-distributed Stochastic Neighbor Embedding), Uniform Manifold Approximation and Projection (UMAP)
- CCA/MFA/Tuckey/Parafac (methods to analyse groups of variables)

### References

- http://factominer.free.fr/index_fr.html
- Multiple Factor Analysis by Example Using R, Pages.
- https://github.com/lmcinnes/umap
- http://www.econ.upf.edu/~michael/IFCS/

### Type of tools

A lot of Singular Value Decomposition

## Project 7: Time series (with missing values) - Julie Josse, Elise Dumas

Time series are essential data. In this project, we will first familiarize ourselves with the main methods and packages (forecast, prophet) for visualizing and analyzing time series and then focus on methods for managing missing data. Recent work based on matrix completion techniques may be of interest. Possibility to work with EDF. Otherwise, a datachallenge of this type can be considered: https://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting

### References

Papers related to matrix completion and missing values. https://hal.archives-ouvertes.fr/hal-02068455/file/Matrix_Factorization_for_Multivariate_Time_Series_Analysis.pdf

https://arxiv.org/abs/1802.09064

Classical lecture on time series can be found here: https://sites.google.com/site/pierrejacob/teaching

Many resources are available on the web for time series with R, the taskview https://cran.r-project.org/web/views/TimeSeries.html can help student navigating between all the ressources.

## Project 8: Work on R-miss-tatic - Elise Dumas, Thomas Schmitt, Imke Mayer

"R-miss-tastic", `https://rmisstastic.netlify.com/` is a platform which aims to provide an overview of standard missing values problems, methods, how to handle them in analyses, and relevant implementations of methodologies. The objective is not only to collect, but also to comprehensively organize materials, to create standard analysis workflows, and to unify the community.

The aim of this project is to learn more about missing values methods and contribute to this platform especially by organizing data challenges. Possible directions of work:

- initiate first steps for the data challenge,
- extend existing or propose new workflows (in R or python) for handling missing values in different scenarios/under different frameworks,
- propose new features for the platform,
- propose solutions for automation of platform updates (new packages, bibliography, events, etc.)

### References

- R-miss-tastic: a unified platform for missing values methods and workflows, `https://arxiv.org/abs/1908.04822`

## Project 9: Unsupervised learning (dimension reduction, clustering) for count data: application in ecology / genomics – Julien Chiquet, Julie Josse

Many application domains, such as ecology or genomics, have to deal with multivariate non-Gaussian observations. A typical example is the joint observation of the respective abundances of a set of species in a series of sites aiming to understand the covariations between these species. Another example is the joint observation of gene counts observed in various subpopulation of cell (e.g., tumour cells). The Gaussian multivariate setting provides a canonical way to model such data but does not apply in general.

This project will explore dimension reduction and clustering techniques to discover patterns and/or grouping when the data table are counts, and thus non-Gaussian. The discovered structured will be put in perspective with the knowledge available in the field of application (ecology, genomics).

**Data / web resources**

- Ecological data set: bird species abundance, in collaboration with ecologists, See ProjetOiseaux3A.doc.
- Genomics data set: single-cell transcriptomics (`https://github.com/LuyiTian/sc_mixology`)

**Some suggestions**

Depending on the students' interest in data analysis or code development, one or the other of the following avenues may be investigated:

- Apply transformations to fulfil the Gaussian assumption and apply classical – yet advanced – clustering / dimension reduction methods (PCA, Gaussian mixture models, k-means, Hierarchical clustering, and regularized versions).
- Couple dimension reduction or spectral method with clustering approaches
- Generalize some dedicated count models (e.g. the Poisson lognormal model) by developing a clustering and/or a Zero-inflated version.

## References

[1] Aitchison, J., and C.H. Ho. 1989. "The Multivariate Poisson-Log Normal Distribution." Biometrika 76 (4). Oxford University Press: 643–53.

[2] J. Chiquet, M. Mariadassou and S. Robin: Variational inference for probabilistic Poisson PCA, the Annals of Applied Statistics, 12: 2674–2698, 2018.

[3] J. Chiquet, M. Mariadassou: PLNmodels: Poisson lognormal models. `https://github.com/jchiquet/PLNmodels`

## Project 10: Network Data Analysis – Julien Chiquet

The objective of this project is to analyse some social network data sets gathered on the internet with various graph clustering methods. In particular, the students will investigate variants of the stochastic block model that handle valued graph, external covariates, dynamic data or include degree-based correction in the inference. An interesting question is to study the robustness of the clustering depending on the method/variant used, or depending on the way the network is (sub)-sampled.

### Web resources

Find some network data of your choice on the internet (with less than 500/1000 nodes for your convenience!). The more exciting approach is to crawl some (social) network data with ‘R‘, for instance your own Tweeter network of followers, or LinkedIn / Facebook / Google scholar equivalent. You may also find some network data on the following web pages

- 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`

### R packages

APIs for tweeter, facebook or Google scholar are interfaced to ‘R‘ with (for instance!) one of the following packages: **graphTweets**, **scholar**, **Rfacebook**.

The following ‘R‘ packages fit the Stochastic Block Model and some of its most useful extensions: **blockmodels**, **dynsbm**, **randnet**

The following tools might be useful for graph manipulations and to build fancy representations of your results: **igraph**, **ggraph**, **tidygraph**

### References

Mariadassou et al (2010) present 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 models are implemented in the package **blockmodels**. Use it to analysis some weighted network data and/or binary network with covariates. Karrer and Newman (2010) introduce a degree-corrected version of the SBM, when the wished structure is not directly related to the degree distribution. The package **randnet** implements this method. Miele and Matias (2017) introduce an extension of the Stochastic Block Model where memberships may vary across time in order to analyse an ecological network gathered in time. The corresponding model is implemented in the package **dynsbm**. Use it to analysis some time-varying network data.

[1] M. Mariadassou, S. Robin and C. 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`

[2] B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks, Physical Review E, 83(1):016107, 2011. `https://arxiv.org/abs/1008.3926`

[3] Vincent Miele 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`

## Project 11: Learn how to code in R – Julien Chiquet & Élise Dumas

For those who want to learn more about R software and study its full potential (data treatment, visualization with ggplot2, interactive graphics, Shiny web application). The objective of this project is to learn how to program in R and to give an overview of what can be done with R.

To this end, the students will

- choose a standard data set adapted to unsupervised learning
- choose and extension of one of the dimension reduction or clustering method studied during the course for which a reference algorithm is available (ask if you need idea!)
- perform all the pre-treament, data wrangling and plotting with the **tidyverse** suite of packages
- implement their own version of the chosen algorithm in `R`, by means of some advanced computational tools (e.g., by interfacing R to C++ with **Rcpp/RcppArmadillo**, to TensorFlow with **RTensorFlow**, or with other advanced optimization library).

## Project 12: Similar patients detection in the context of Traumabase – Olivier Auliard (Capgemini) & Tobias Gauss (Traumabase)

Handling severe trauma patients is a major concern for public health. To help M.Ds to take the best decision in the shortest amount of time, the Trauma Insights tool aims at giving information on past similar cases, giving new levers to practitioners to capitalize on their peer's work. To build this tool, you will have to explore the Traumabase dataset and build a practical dashboard using R or Python that, given a patient, will provide similar cases from the Traumabase and display relevant clinical information using similarity calculation methods and clustering, as well as dimensionality reduction and missing value completion. You could also find some inspiration from recommender system. The purpose of this tool will be to collect pre-hospital data from a new patient and to see if there were similar cases in order to give insights on potential outcomes associated with therapeutic strategies for complicated cases.

# References

[1] DeepReco: Deep Learning Based Health Recommender System Using Collaborative Filtering https://www.mdpi.com/2079-3197/7/2/25/htm

[2] A k-mean clustering algorithm for mixed numeric and categorical data https://www.sciencedirect.com/science/article/pii/S0169023X0700050X

[3] Sparse Nonnegative Matrix Factorization for Clustering https://smartech.gatech.edu/handle/1853/20058

[4] Dimensionality reduction of unsupervised data https://ieeexplore.ieee.org/abstract/document/632300