Master Theses

Students can choose a topic from the list below, find a related topic in consultation with a promotor or propose a topic to one of the teachers in the program.

The list of thesis subjects may help you to identify promotor(s) that may be interested in your topic. We suggest that you contact and talk to professors to identify an interesting thesis topic.

The deadline for the submission of a title and the name of the promoter is 1 April 2017.

Master thesis subjects 2016-2017

- A Statistical Analysis of the Microbiome Data of the American Gut Project
- A Study on the Impact of Volume Variability in Digital PCR
- A Web Tool for Digital PCR Data Analysis
- Analysis of University Rankings
- Discovering Relationships in Climate-vegetation Dynamics using Dynamic Feature Selection Techniques
- Effect of Early Grade Retention on Math Development
- Effect of Weight Choice in "Data-driven Hypothesis Weighting" for RNA-Seq Studies
- Estimating the Biodiversity of Aquatic Microbial Ecosystems
- Identifying Synergy and Redundancy in Neuroimaging Data
- Large-scale/High-dimension Prediction Problems with Multiple Independent Datasets
- Machine Learning for the Modeling of Climate-vegetation Dynamics in a Non-linear Granger Causality Framework
- Microbiome
- Multivariate Statistics and decoding for EEG Data
- Network Structure of Multivariate Time Series of Brain Recordings
- Rank Tests under the Probabilistic Index Model Framework
- Regression Trees for Longitudinal Data
- Robust Estimation in Probabilistic Index Models
- Simulation-guided Group-sequential Design to Evaluate a Candidate Vaccine for HIV Prevention
- Zero-Inflated Generalized Linear Mixed Models for Count Data

Title: A statistical analysis of the microbiome data of the American Gut Project

Promoter : Prof. dr. Olivier Thas (Olivier.Thas@UGent.be)

Tutor : Stijn Hawinkel

Target group : students with interest in genomics

Summary :

The healthy human body is inhabited by billions of bacteria, viruses and fungi on all of its outer and inner surfaces, such as oral cavity, skin and gut. Thereby each body site has its own unique community of micro-organisms adapted to its environmental conditions. The ensemble of these communities of non-human beings living on our bodies is called the human microbiome. The composition of the microbiome varies with subject, body site and other factors such as antibiotics use.

The composition of the gut microbiome is characterized by sequencing the 16S rRNA marker gene in stool samples, after which these sequences are mapped to known databases. This way count tables are constructed that show the bacterial composition of every sample. A recent data collection effort called the American Gut project culminated in a dataset of over 2000 sequenced fecal samples, with a total of 1800 species detected. Along with the stool composition of the patient, over 200 baseline covariates are recorded as well. Many methods for testing for association between covariates and bacterial species abundance can accomodate for additional covariates, but little is known about their effect on the resulting model and their association with the species abundances.

The goal of the thesis is to gain insight in the role of clinical baseline covariates with respect to bacterial abundances and estimated nuisance parameters and to develop a model selection procedure for selecting important covariates.

References:

-McDonald, D., Birmingham, A., & Knight, R. (2015). Context and the human microbiome. *Microbiome*, *3*, 52. http://doi.org/10.1186/s40168-015-0117-2

Title: A study on the impact of volume variability in digital PCR

Promotor: Prof. dr. Olivier Thas (Olivier.Thas@UGent.be)

Tutor: Matthijs Vynck

Target group: students with interest in genomics, R programming

Several variables impacting the accuracy and precision of digital PCR experiments have been identified (Jacobs et al., 2014)). One of these variables is the volume of the partitions: this is typically assumed to be a constant, leading to an underestimation (bias) of the concentration. Especially now platforms with higher dynamic ranges (and smaller partitions) are being marketed, ignoring this variability may have large unwanted effects. Recently, a method to account for volume variability has been suggested, relying on modelling the concentration as a Poisson-Gamma distribution as opposed to a Poisson distribution (Huggett et al., 2015)). This method has not been thoroughly investigated.

The objective of this master's thesis is to set up a simulation study to study the effects of volume variability under a range of assumptions, based on a realistic set of parameters as obtained from literature.

References:

Huggett JF, Cowen S, Foy CA. Considerations for digital PCR as an accurate molecular diagnostic tool. Clin Chem. 2015 Jan;61(1):79-88. doi: 10.1373/clinchem.2014.221366.

Jacobs BK, Goetghebeur E, Clement L. Impact of variance components on reliability of absolute quantification using digital PCR. BMC Bioinformatics. 2014 Aug 22;15:283. doi: 10.1186/1471-2105-15-283.

Title: A web tool for digital PCR data analysis

Promotor: Prof. dr. Olivier Thas (Olivier.Thas@UGent.be)

Tutor: Matthijs Vynck

Target group: students with interest in genomics, R programming

Recently, we have proposed several methods to improve the data analysis pipeline of digital PCR experiments (Trypsteen et al., 2015), Vynck et al., 2016)). R code and data analysis tutorials on how to apply these methods are available. Several other groups have suggested alternative data analysis procedures (Strain et al., 2013); Dreo et al. ,2014).

Unfortunately, the use of the methodology outlined in these papers is hampered by a lack of experience with R in the digital PCR community. Indeed, many biomedical researchers prefer an easy to use graphical interface. To facilitate the use of the developed methodology, two standalone Shiny apps for applying the methodology have been developed. However, these apps have not been integrated, reducing the workflow efficiency. Moreover, not all developed methodology has been implemented into these graphical interfaces.

The purpose of this master's thesis is to
- implement several statistical methodologies for the analysis of digital PCR data in R,
- combine the workflow into one application,
- extend the functionality of the currently available graphical interfaces.

References:

Dreo T, Pirc M, Ramšak Ž, Pavšič J, Milavec M, Zel J, Gruden K. Optimising droplet digital PCR analysis approaches for detection and quantification of bacteria: a case study of fire blight and potato brown rot. Anal Bioanal Chem. 2014 Oct;406(26):6513-28. doi: 10.1007/s00216-014-8084-1.

Strain MC, Lada SM, Luong T, Rought SE, Gianella S, Terry VH, Spina CA, Woelk CH, Richman DD. Highly precise measurement of HIV DNA by droplet digital PCR. PLoS One. 2013;8(4):e55943. doi: 10.1371/journal.pone.0055943.

Trypsteen W, Vynck M, De Neve J, Bonczkowski P, Kiselinova M, Malatinkova E, Vervisch K, Thas O, Vandekerckhove L, De Spiegelaere W. ddpcRquant: threshold determination for single channel droplet digital PCR experiments. Anal Bioanal Chem. 2015 Jul;407(19):5827-34. doi: 10.1007/s00216-015-8773-4.

Vynck M, Vandesompele J, Nijs N, Menten B, De Ganck A, Thas O. Flexible analysis of digital PCR experiments using generalized linear mixed models. Biomol Detect Quantif. 2016 Jun 24;9:1-13. doi: 10.1016/j.bdq.2016.06.001.

Title: Analysis of university rankings

Promotor: Jan De Neve and Christophe Ley

According to the most recent Times Higher Education (THE) University ranking, Ghent University is ranked 118 worldwide, making it the second best Belgian university after the KU Leuven. Besides the THE, there exist various other university rankings such as the Academic Ranking of World Universities (Shanghai ranking) where Ghent University is ranked 71 worldwide, making it the top-ranked Belgian university. The aims of the present thesis are three-fold: 1) understanding the mechanisms behind the current university rankings, 2) comparing the various rankings and, as ultimate goal, 3) proposing a new ranking based on probabilistic index models (Thas et al., 2012).

References:

Thas, O., De Neve, J., Clement, L., & Ottoy, J. P. (2012). Probabilistic index models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(4), 623-671.

Title: Discovering relationships in climate-vegetation dynamics using dynamic feature selection techniques

Promoter : Prof. dr. Willem Waegeman, Prof. Dr. Olivier Thas

Tutor : Christina Papagiannopoulou

Target group : students with interest in predictive modelling

Summary :

Earth observation satellite data provide a wealth of information about the dynamics of our planet in recent decades. Composite global records of important environmental and climatic variables now span up to 30 years, enabling the study of climate–vegetation interactions over multi-decadal scales. These records have the form of multivariate time series with different spatial and temporal resolutions.

Climatic conditions are known to be key drivers of ecosystem dynamics, which are sensitive to temperature, availability of water and the solar irradiance. In the other direction, vegetation has a major influence on climate systems on a global scale. Specifically, we are interested in extreme events such as droughts, heatwaves or extreme precipitation and how they are related to vegetation. Statistical methods and machine learning techniques can be applied on these data in order to help in discovering correlations and forming models that may predict future situations. An additional challenge is the efficient handling of the massive datasets that are used to study climate extremes.

The aim of this thesis is to discover hidden correlations between the different climate drivers and vegetation, and more specifically to reveal the effects that these drivers may have on vegetation in each area of the world. Based on machine learning techniques, and taking into account past observations, we can model patterns that may exist in the data. These patterns can dynamically change in time due to climate change, showing that drivers which influence the vegetation of a region in the past are different in comparison with the ones in the present. To this end, dynamic feature selection techniques will have to be implemented.

Effect of early grade retention on math development

Promotor : prof. Stijn Vansteelandt (Stijn.Vansteelandt@ugent.be)

Target group : students with interest in educational data, consulting

Summary :

One of the main objectives of many empirical studies in the social and behavioral sciences is to assess the causal effect of a treatment or intervention on the occurrence of a certain event. The randomized controlled trial is generally considered as the gold standard to evaluate such causal effects. However, because of ethical or practical reasons, social scientists are often bound to the use of non-experimental, observational designs. When the treatment and control group are different with regard to variables that are related to the outcome, this may raise the problem of confounding. A variety of statistical techniques, such as regression, matching and subclassification, is now available and routinely used to adjust for confounding due to measured variables. However, these techniques are not appropriate for dealing with time-varying confounding, which arises in situations where the treatment or intervention can be received at multiple time points. In this thesis, this problem of time-varying confounding will be addressed in the analysis of a Flemish study of grade retention effects on mathematics development throughout primary school.

For this thesis, you will be expected to communicate with subject-matter experts on a regular basis to make yourself acquainted with the scientific question and data, and to report the results. Being a fluent communicator, as well as taking one of the courses on Analysis of Longitudinal and Clustered Data, Causality and Missing Data, is an assett for this project.

Title: Effect of weight choice in "data-driven hypothesis weighting" for RNA-Seq studies

Promoter : Prof. dr. Olivier Thas (Olivier.Thas@UGent.be)

Tutor : Alemu Takele

Target group: students with interest in genomics

Summary:

Hypothesis weighting is a powerful approach for improving the power of data analyses that employ multiple testing. However, in general it is not evident how to choose the weights. In data-driven hypothesis weighting, different functions of covariates can be used as weights for each hypothesis.

The goal of this thesis is to explore the effect of weight choice in data-driven hypothesis weighting on the results of differential expression testing for RNA-seq studies (effect on the power and the false discovery rate of the tests).

References:

- Ignatiadis, N., Klaus, B., Zaugg, J. B., & Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*.

Title: Effect of weight choice in "data-driven hypothesis weighting" for RNA-Seq studies

Promoter : Prof. dr. Olivier Thas (Olivier.Thas@UGent.be)

Tutor : Alemu Takele

Target group: students with interest in genomics

Summary:

Hypothesis weighting is a powerful approach for improving the power of data analyses that employ multiple testing. However, in general it is not evident how to choose the weights. In data-driven hypothesis weighting, different functions of covariates can be used as weights for each hypothesis.

The goal of this thesis is to explore the effect of weight choice in data-driven hypothesis weighting on the results of differential expression testing for RNA-seq studies (effect on the power and the false discovery rate of the tests).

References:

- Ignatiadis, N., Klaus, B., Zaugg, J. B., & Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*.

Title: Estimating the biodiversity of aquatic microbial ecosystems

Promoter : Prof. dr. Willem Waegeman (Willem.Waegeman@UGent.be)
Copromoter: Prof. dr. Olivier Thas (Olivier.Thas@UGent.be)

Tutor : Peter Rubbens (Peter.Rubbens@UGent.be)
Tutor: Ruben Props (Ruben.Props@UGent.be)

Target group : students with interest in machine learning and applications in synthetic ecology

Summary :

Researchers have been showing great interest in studying microbial communities present in drinking water. This stems from the fact that information about these communities is related to global properties such as drinking water quality. Recently, the use of *flow cytometry* is becoming more and more popular to perform this analysis [1]. This is because flow cytometry is able to analyze thousands of single cells in only seconds of time. Of course, this gives rise to a large amount of data which needs to be interpreted accordingly.

Numerous techniques exist in the field of immunology to analyze flow cytometric data [2]. However, a recent study has shown that these methods do not perform optimally when studying samples coming from microbial communities in an aquatic context [3]. The aim of this thesis is therefore to perform a comparative study of state-of-the-art methods applied to aquatic microbial ecosystems. The target of the study is to achieve a proper estimate of the *biodiversity*. The thesis therefore comprises both a literature study and a thorough computational and statistical analysis. Depending on the interest of the student, the focus of the thesis can be shifted towards one or the other.

References:
[1]: Monitoring microbiological changes in drinking water systems using a fast and reproducible flow cytometric method. Prest, E.I., Hammes, F., Kötzsch, S., Van Loosdrecht, M.C.M., Vrouwenvelder, J.S.; *Water Research* (2013) doi:10.1016/j.watres.2013.07.051

[2]: Critical assessment of automated flow cytometry data analysis techniques. Aghaeepour, N., Finak, G., The FlowCAP Consortium, The DREAM Consortium, Hoos, H., Mosmann T.R., Brinkman, R., Gottardo, R., Scheuermann, R.H.; *Nature Methods* (2013). doi:10.1038/nmeth.2365

[3: Scalable clustering algorithms for continuous environmental flow cytometry. Hyrkas, J., Clayton, S., Ribalet, F., Halperin, D., Armbrust, E., Howe, B.; *Bioinformatics* (2015) doi:10.1093/bioinformatics/btv594.

Identifying synergy and redundancy in neuroimaging data

Promotor : prof. Daniele Marinazzo (daniele.marinazzo@ugent.be)

Target group : students with interest in how the brain works, graph theory and time series analysis

Summary :

Information theoretic treatment of groups of correlated degrees of freedom can reveal their functional roles as memory structures or information processing units. Furthermore by looking at the common amount of information shared in a group of variables we can tell whether they are mutually redundant or synergetic. The application of these insights to identify functional connectivity structure is a promising line of research. Another topic of general interest is the understanding of couplings between dynamical systems and their parts. Transfer entropy and Granger causality are popular approaches used to distinguish effectively driving and responding elements and to detect asymmetry in the interaction of subsystems. These two methods can be unified under some conditions, opening new computational and methodological perspectives. Several techniques can evidence sets of variables which provide information for the future state of the target. This information can be synergetic or redundant, with important implication on our understanding of the functioning of the dynamical system under analysis.
The thesis work will be dedicated to develop and statistically validate both exact and approximated algorithms to identify redundancy and synergy in neuroimaging data sets, both functional magnetic resonance imaging (fMRI) and electroencephalography (EEG). Analyses will be carried out by default in MATLAB and R, but students with an expertise in other languages (python, c etc) are most welcome (indeed they would bring added value to the lab).

References:

Synergetic and redundant information flow detected by unnormalized Granger causality: application to resting state fMRI
S. Stramaglia, L. Angelini, G. Wu, J. Cortes, L. Faes, D. Marinazzo
IEEE Transactions on Biomedical Engineering, PP (99), in press (2016)
http://ieeexplore.ieee.org/document/7462237/

Expanding the transfer entropy to identify information circuits in complex systems
S. Stramaglia, G. Wu, M. Pellicoro, D. Marinazzo
Physical Review E, 86, 066211 (2012)

Large-scale/high-dimension prediction problems with multiple independent datasets

October 5, 2016

Promoter: Prof. dr. Olivier Thas (Olivier.Thas@UGent.be)

Tutor: Chamberlain Mbah

Target group: students with interest in genomics/ analysis of high dimensional data.

Summary:

The success of radiation therapy is based on maximizing tumour control and minimising damage of healthy tissues around the tumour. Optimising radiation therapy depends on how well normal tissue damage can be quantified and controlled. Dose parameters are obvious predictors for normal tissue reactions, but genetic information can be useful. With an increasing number of randomised and observational studies of radiotherapy patient cohorts, numerous predictors and prediction models for normal tissue damage have been proposed. However, it is challenging to validate these predictors and models across multiple cohorts.

The aim of this thesis is to combine clinical, treatment, dosimetric, and genetic information of patients from multiple cohorts to build a single prediction model that predicts normal tissue damage. The effect of different predictors in prediction, will vary across cohorts; referred to as cohort heterogeneity. A universally stable model with cohort independent effects will be developed and applied to data from multiple European breast cancer patient cohorts.

Title: Machine learning for the modeling of climate-vegetation dynamics in a non-linear Granger causality framework.

Promoter : Prof. dr. Willem Waegeman, Prof. Dr. Ir. Olivier Thas

Tutor : Stijn Decubber

Target group : students with interest in machine learning and high performance computing on large datasets.

Summary :

Earth observation satellite data provide a wealth of information about the dynamics of our planet in recent decades. Composite global records of important environmental and climatic variables now span up to 30 years, enabling the study of climate–vegetation interactions over multi-decadal scales. These records have the form of multivariate time series with different spatial and temporal resolutions.

Climatic conditions are known to be key drivers of ecosystem dynamics, which are sensitive to temperature, availability of water and the solar irradiance. In the other direction, vegetation has a major influence on climate systems on a global scale. Specifically, we are interested in extreme events such as droughts, heatwaves or extreme precipitation and how they relate to vegetation.

Granger causality is a predictive notion of causality used to understand relations between time series data. Traditionally, Granger modelling has been applied to multivariate time series in a linear setting, usually combined with statistical tests to quantify predictive causality in a sound way.

The goal of this thesis is to use machine learning techniques to investigate which climate variables drive changes in vegetation. Because non-linear methods such as random forests have been shown to perform best at capturing patterns in our dataset, a challenge at the interplay of statistics and machine learning arises to cast the predictive performance of machine learning models in a statistical Granger causality framework.

An additional challenge is the size of the datasets that are used. The modelling experiments will be performed on the high performance computing infrastructure of the university. Furthermore, because of the nature and the size of the data, there is an opportunity to go into more advanced methods such as neural networks or deep learning.

Promoter : Prof. dr. Olivier Thas (Olivier.Thas@UGent.be)

Tutor : Stijn Hawinkel

Target group : students with interest in genomics

Summary :

The healthy human body is inhabited by billions of bacteria, viruses and fungi on all of its outer and inner surfaces, such as oral cavity, skin and gut. Thereby each body site has its own unique community of micro-organisms adapted to its environmental conditions. The ensemble of these communities of non-human beings living on our bodies is called the human microbiome. The composition of the microbiome varies with subject, body site and other factors such as antibiotics use.

The composition of the gut microbiome is characterized by sequencing the 16S rRNA marker gene in stool samples, after which these sequences are mapped to known databases. This way count tables are constructed that show the bacterial composition of every sample. A recent data collection effort called the American Gut project culminated in a dataset of over 2000 sequenced fecal samples, with a total of 1800 species detected. Along with the stool composition of the patient, over 200 baseline covariates are recorded as well. Many methods for testing for association between covariates and bacterial species abundance can accomodate for additional covariates, but little is known about their effect on the resulting model and their association with the species abundances. The goal of the thesis is to gain insight in the role of clinical baseline covariates with respect to bacterial abundances and estimated nuisance parameters and to develop a model selection procedure for selecting important covariates.

References:

-McDonald, D., Birmingham, A., & Knight, R. (2015). Context and the human microbiome. *Microbiome, 3*, 52. http://doi.org/10.1186/s40168-015-0117-2

Multivariate statistics and decoding for EEG data

Promotor : prof. Daniele Marinazzo (daniele.marinazzo@ugent.be)

Target group : students with interest in how the brain works, machine learning and statistics

Summary :

Magnetic- and electric-evoked brain responses have traditionally been analyzed by comparing the peaks or mean amplitudes of signals from selected channels and averaged across trials. More recently, tools have been developed to investigate single trial response variability (e.g., EEGLAB) and to test differences between averaged evoked responses over the entire scalp and time dimensions (e.g., SPM, Fieldtrip). LIMO EEG is a Matlab toolbox (EEGLAB compatible) to analyse evoked responses over all space and time dimensions, while accounting for single trial variability using a simple hierarchical linear modelling of the data. In addition, LIMO EEG provides robust parametric tests, therefore providing a new and complementary tool in the analysis of neural evoked responses.

This toolbox is now expanding, improving and merging with other EEG data toolboxes. The student will start working on a MANOVA, then implementing a linear and quadratic discriminant analysis, which utilizes the same matrices, but combines the info in a different way.
A further step forward could be the development of a decoding function.

References:

LIMO EEG: A Toolbox for Hierarchical LInear MOdeling of ElectroEncephaloGraphic Data
Cyril R. Pernet, Nicolas Chauveau, Carl Gaspar and Guillaume A. Rousselet
Computational Intelligence and Neuroscience, Volume 2011 (2011), Article ID 831409,
http://dx.doi.org/10.1155/2011/831409

Network structure of multivariate time series of brain recordings

Promotor : prof. Daniele Marinazzo (daniele.marinazzo@ugent.be)

Target group : students with interest in how the brain works, graph theory and time series analysis

Summary :

Our understanding of a variety of phenomena in physics, biology and economics crucially depends on the analysis of multivariate time series. While a wide range of tools and techniques for time series analysis already exist, the increasing availability of massive data structures calls for new approaches for multidimensional signal processing. The student will use a non-parametric method to analyze multivariate time series, based on the mapping of a multidimensional time series into a multilayer network, which allows to extract information on a high dimensional dynamical system through the analysis of the structure of the associated multiplex network. Simple structural descriptors of the associated multiplex network may allow to extract and quantify nontrivial properties of a complex dynamical system such as the brain.
The multivariate time series which will be object of the analysis will come from both functional magnetic resonance imaging (fMRI) and electroencephalographic (EEG) recordings, which measure brain activity at different spatial and temporal scales.
Analyses will be carried out by default in MATLAB and R, but students with an expertise in other languages (python, c etc) are most welcome (indeed they would bring added value to the lab).

References:

- **Network structure of multivariate time series.** Lacasa L, Nicosia V, Latora V
Sci Rep 2015 5: 15508. doi: 10.1038/srep15508
http://www.ncbi.nlm.nih.gov/pubmed/23422254

Title: Rank tests under the Probabilistic Index Model framework

Promoter : Prof. dr. Olivier Thas (Olivier.Thas@UGent.be) and Prof. dr. Jan de Neve (Jan.DeNeve@UGent.be)

Tutor : Gustavo Amorim

Target group : Students interested in statistical computing and regression models.

Summary :

As parametric tests require sometimes strong underlying assumptions and are often greatly affected by outliers or extreme observations, nonparametric rank tests, which are distribution-free and robust under significantly less assumptions, should be generally preferable. This is however not the case. Rank tests are usually constructed for specific designs and the lack of software that readily implement rank tests for more complicated scenarios limits their widespread use.

To overcome this issue, De Neve and Thas (2015) showed that the recently introduced probabilistic index model (PIM, Thas et al., 2012) can be used to generate several classical and new rank tests. PIMs model the probability $P(Y_1 \leq Y_2)$ in terms of a set of regressors X and, as shown by the authors, for certain choices of X and of the regression parameter, different rank tests such as the Wilcoxon-Mann-Whitney, Kruskal-Wallis and Friedman rank tests, to name a few, arise naturally from a PIM.

In this thesis the student is expected to implement most classical rank tests discussed in De Neve and Thas (2015), under the PIM framework, in the software R. Several tests need to be applied to example data sets so as to demonstrate the use of the new R functions and the added value of the PIM. These results will be later added to the PIM package, currently available on CRAN.

References:

- Thas, O., Neve, J. D., Clement, L., and Ottoy, J. P. (2012). Probabilistic index models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(4), 623-671.
- De Neve, J and Thas, O. (2015). A regression framework for rank tests based on the probabilistic index model. Journal of the American Statistical Association, 110(511), 1276-1283.

Regression Trees for Longitudinal Data

Promotor : prof. Tom Loeys (tom.loeys@ugent.be)

Target group : students with interest in longitudinal data analysis and data mining

Summary :

When a longitudinal change is studied we often find that changes over time are heterogeneous (in terms of time and/or covariates' effect). A traditional linear mixed effect model assuming common parametric form for covariates and time may not be applicable to the entire population. This is usually the case in studies when there are many possible predictors influencing the response trajectory. In such cases, a group-averaged trajectory can mask important subgroup differences.
The aim of this thesis to explore techniques that can identify and characterize longitudinally homogeneous subgroups based on the combination of (baseline) covariates in a parsimonious way. The potential of tree methods such as CART (classification and regression trees) is assessed and may be illustrated using real data.

References:

- McArdle, J. & Ritschard, G. (2014) Contemporary issues in exploratory data mining in the behavioral sciences. Routledge

Title: Robust estimation in Probabilistic index models

Promoter : Prof. dr. Olivier Thas (Olivier.Thas@UGent.be)

Tutor : Gustavo Amorim

Target group : Students interested in statistical computing, mathematical statistics and regression models.

Summary :

Probabilistic index model (PIM) is a flexible class of semi-parametric models that generate many classical and new rank tests. It relates the probability $P(Y_1 \leq Y_2)$, which arise naturally for instance in engineering or in clinical trials, to a set of predictors through some link function *g(.)*. This link function is assumed to be known or correctly specified, otherwise PIM-based estimators for the regression parameter may be inconsistent. As *g(.)* is often unknown, the applicability of PIMs are reduced to few cases.

To solve this issue, we expect to borrow results from semiparametric single index models and derive a PIM-based estimator that uses a nonparametric link function. This would allow us to construct diagnostics tools for checking whether particular choice of *g(.)* is acceptable or not as well as providing a more adequate fit to the data when the link function is hard to be specified.

In this thesis, the student will be part of a research team where s/he is expected to implement all aspects of the estimating procedure in R. Therefore, interest in mathematical statistics and in statistical computing, more specifically proficiency in R, are required.

References:

- Thas, O., Neve, J. D., Clement, L., and Ottoy, J. P. (2012). Probabilistic index models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(4), 623-671.
- De Neve J, Thas O. (2015) A regression framework for rank tests based on the probabilistic index model. Journal of the American Statistical Association. 110(511):1276-1283.
- Weisberg, S. and Welsh, A. H. (1994) Adapting for the Missing Link. Ann. Statist. 22(4), 1674-1700.
- Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics, 58(1), 71-120.

Simulation-guided group-sequential design to evaluate a candidate vaccine for HIV prevention

Promotor : Dr. An Vandebosch (avandebo@its.jnj.com)
Prof. Stijn Vansteelandt (stijn.vansteelandt@ugent.be)

Summary :

The aim of this project is to explore via simulations several design aspects for a phase 3 trial of a candidate vaccine to prevent HIV infections. In this study, healthy subjects will be randomly assigned to receive a candidate vaccine or appropriate control vaccination. After vaccination, subjects will be followed regularly to assess whether or not they were infected with HIV. The primary objective is to compare time to HIV infection between randomized groups.
In first instance, the required number events, number of subjects and study power will be determined under various incidence, including and effect size scenario's. These can be extended to take more complex scenarios into account, such as site heterogeneity and time-dependent vaccine effects.
Secondly, the opportunity will be explored to build in interim evaluations to monitor for harm (an unanticipated increased HIV risk), futility (no difference between randomized groups) or early evidence for efficacy. Via simulations, the student will evaluate the probability to stop early under various scenarios, potential of an incorrect decision.

In the first part, a literature evaluation will be conducted to understand the necessary background on vaccine studies, study design as well as group-sequential studies. In the next steps, a step-by-step exploration will be done via simulations on the various aspects of trial design.

Requirements: Epi&RCT, survival analysis

Zero-Inflated Generalized Linear Mixed Models for Count Data

Promotor : prof. Tom Loeys (tom.loeys@ugent.be)

Target group : students with interest in categorical and longitudinal data analysis

Summary :

It is not uncommon in psychological research that the outcome of interest is counting the occurrence of a behavioral event. In relational psychology for example, the number of post-breakup unwanted pursuit behaviors (UPB) of harassment in former partners may be a measurement of interest. Such count data are typically very skewed and exhibit a lot of zero count observations. Zero-inflated count models or hurdle models can be used to model associations between such count data and predictors of interest. When the count outcomes are repeatedly measured over time, one needs to take into account the correlation as well, hereby relying on the generalized linear mixed modeling (GLMM) framework for example. The purpose of this thesis is to explore the implementation of zero-inflated models within the GLMM-framework. The different techniques may be applied to longitudinal data on UPB in ex-partners.

References:

- Loeys, T., Moerkerke, B., De Smet, O., & Buysse, A. (2012) The analysis of zero-inflated count data: beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology*, 65(1), 163-180
- Buu, A., Li, R., Tan, X., & Zucker, R. (2012) Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine*, 31 (29), 4074-4086.