# 26th Annual Meeting of the Royal Statistical Society of Belgium

Ovifat, 17-19 October, 2018

**ROYAL STATISTICAL SOCIETY OF BELGIUM**

The support of the following sponsors is gratefully acknowledged








Avec le soutien de la
Fédération Wallonie-Bruxelles





Ecole Doctorale Thématique en
Statistique et Sciences Actuarielles



# Contents

**Practical details and map** **9**

**Committees** **11**

**Invited speakers** **13**

**Conference schedule** **15**

**Abstracts of posters (Oct 17-18)** **19**

A parametric methodology to estimate the variance matrix of classical measurement error (*Elif Akça* and Ingrid Van Keilegom*) . . . 19

Assessing cure status prediction from survival data using ROC curves (*Maïlis Amico* and Ingrid Van Keilegom*) . . . 19

Extremal dependence in a Hüsler-Reiss Markov tree (*Stefka Asenova*, Johan Segers and Gildas Mazo*) . . . 20

Testing uniformity on high-dimensional spheres against even rotationally symmetric alternatives (*Christine Cutting*, Davy Paindaveine and Thomas Verdebout*) . . . 21

Statistical approaches to estimate economic performance for global business (*Antonio Frenda*) . . . 21

Confidence curves post-AIC selection (*Gerda Claeskens and Andrea Cristina Garcia Angulo**) . . . 22

Randomization inference with general interference and censoring (*Mohammad Ali, John D. Clemens, Michael E. Emch, Michael G. Hudgens and Wen Wei Loh**) . . . 22

Comparison of model regularization methods for identifying groups of predictive variables (*Bernadette Govaerts, Rebecca Marion* and Rainer von Sachs*) . . . 23

Uncertainty quantification in sunspot counts (*Véronique Delouille, Laure Lefèvre, Sophie Mathieu* and Rainer von Sachs*) . . . 24

Semiparametric estimation in AFT mixture cure models (*Motahareh Parsa* and Ingrid Van Keilegom*) . . . 24

Preliminary test estimation in ULAN models (*Davy Paindaveine, Joséa Rondrotiana Rasoafaraniaina* and Thomas Verdebout*) . . . 25

Identification of Wiener-Hammerstein system with random forest (*Kurt Barbe and Md Abu Hanif Shaikh**) . . . 25

Life and health actuarial pricing: a biostatistics approach (*Michel Denuit, Catherine Legrand and Antoine Soetewey**) . . . 26

Weight choices for the penalized composite and model-averaged estimator in quantile regression (*Daumantas Bloznelis, Gerda Claeskens and Jing Zhou**) . . . 27

**Abstracts of talks: Oct 17** **29**

Enhancing your research with R (*Klaus Nordhausen*) . . . . . 29

Improving interim decisions in randomized trials by exploiting information on short-term outcomes and prognostic baseline covariates (*An Vandebosch, Kelly Van Lancker* and Stijn Vansteelandt*) . . . . . 29

Comparison of optimisation bias in unstructured and structured sparse variable selection (*Bastien Marquis* and Maarten Jansen*) . . . . . 30

How (not) to publish statistical research - a personal collection of someone from the middle ages (*Roland Fried*) . . . . . 30

**Abstracts of talks: Oct 18** **31**

NPMLE methods for mixture models (*Roger Koenker* and Jiaying Gu*) . . . . . 31

Climate event attribution using multivariate peaks-over-thresholds modelling (*Anna Kiriliouk* and Philippe Naveau*) . . . . . 31

Community based grouping for undirected graphical models (*Gerda Claeskens and Eugen Pircalabelu**) . . . . . 32

Comparison of different software implementations for spatial disease mapping (*Christel Faes, Thomas Neyens and Maren Vranckx**) . . . . . 32

Estimation of controlled direct effects in longitudinal mediation analyses with latent variables in randomised studies (*Tom Loeys, Wen Wei Loh*, Beatrijs Moerkerke and Stijn Vansteelandt*) . . . . . 33

On scientifically relevant mediation-style questions, and approaches to answering them, from large multi-omic datasets (*Rhian Daniel*) . . . . . 34

Multivariate tail quantile contours based on optimal transport (*Cees de Valk and Johan Segers**) . . . . . 34

Flexible distributional modelling for parametric survival analysis (*Chris Jones*) . . . . 35

Choose or not to choose a prior. That's the question! (*Fatemeh Ghaderinezhad* and Christophe Ley*) . . . . . 35

Bridging the gap between Bayesian P-splines and Laplace's method for inference in generalized additive models (*Oswaldo Gressani* and Philippe Lambert*) . . . . . 36

The social relations model for count data: to Bayes or not to Bayes (*Tom Loeys and Justine Loncke**) . . . . . 36

Modeling longitudinal dyadic data in the SEM framework (*Fien Gistelinck* and Tom Loeys*) . . . . . 37

Prediction of singular VARs and application to generalized dynamic factor models (*Gilles Nisol* and Siegfried Hörmann*) . . . . . 38

Time series models with time-dependent coefficients: asymptotic results (*Rajae Azrak and Guy Mélard**) . . . . . 38

Building a dynamic risk prediction model for cardiovascular disease (*Jessica Barrett*, Michael Sweeting, Ellie Paige, David Stevens and Angela Wood*) . . . . . 39

On standardising quality of care indicators based on summary statistics (*Marion Louvet* and Els Goetghebeur*) . . . . . 39

Generalized pairwise comparison methods to analyze (non)-hierarchical composite endpoints (*Johan Verbeeck*) . . . . . 40

**Abstracts of talks: Oct 19** **41**

Goodness-of-fit tests in proportional hazards models with random effects (*Wenceslao González-Manteiga, María Dolores Martínez-Miranda* and Ingrid Van Keilegom*) 41

Lorenz regression (*Cédric Heuchenne and Alexandre Jacquemain**) . . . . . 41

Smooth time-dependent ROC curve for right censored survival data (*Kassu Mehari Beyene*) . . . . . 42

Stable IPW estimation for longitudinal studies (*Vahe Avagyan* and Stijn Vansteelandt*) 42

Power in high-dimensional testing problems (*Anders Bredahl Kock and David Preiner-storfer**) . . . . . 43

High-dimensional doubly robust tests for regression parameters (*Vahe Avagyan, Oliver Dukes* and Stijn Vansteelandt*) . . . . . 43

Testing for hidden periodicities in functional time series (*Clément Cerovecki, Vaidotas Characiejus*, and Siegfried Hörmann*) . . . . . 44

First aid after model selection (*Gerda Claeskens*) . . . . . 44

On the causal effect of gender: beyond the (many!) anecdotes, a statistician's view on evidence of gender bias (*Els Goetghebeur*) . . . . . 45

James-Stein estimators in factor analysis (*Elissa Burghgraeve*, Jan De Neve and Yves Rosseel*) . . . . . 45

Testing for principal component directions under weak identifiability (*Davy Paindaveine, Julien Remy* and Thomas Verdebout*) . . . . . 46

Distribution comparison tests based on self-similar transport of measure (*Gilles Mordant*) 46

On the impact of residential history in the spatial analysis of diseases with long latency period: a study of Mesothelioma in Belgium (*Christel Faes, Kristiaan Nackaerts, Benoit Nemery, Thomas Neyens and Oana Petrof**) . . . . . 47

Detection of high dimensional intestinal microbiota as biomarker for immunological response: a Bayesian variable selection approach (*Olajumoke Evangelina Owokotomo*, Rudradev Sengupta, Luc Bijnens, Ziv Shkedy and Adetayo Kasim*) . . . . . 47

"cyanoFilter", an automated framework for identifying picocyanobacteria populations obtained via flow cytometry (*Marc Aerts, Frederik De Laender, Thomas Neyens, Olusoji Oluwafemi* and Jurg Spaak*) . . . . . 48

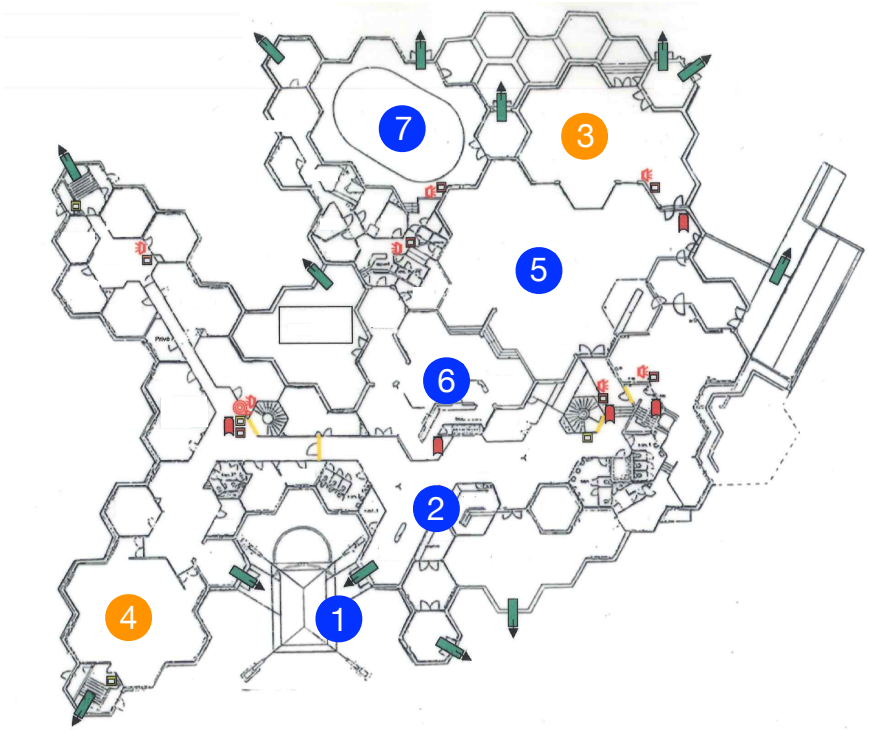
The genesis and use of time-varying frailty models for representing heterogeneities in the transmission of infectious diseases (*Steffen Unkel*, Steven Abrams, Andreas Wienke and Niel Hens*) . . . . . 49

**Presenter Index** **51**

# Practical details and map

Plenary sessions will take place in the room Adrien de Nassau (ADN), whereas parallel sessions will be held in the room ADN and in the room Apollinaire (please see the map below).

You can connect to the Internet by using the WiFi network *Domaine des Hautes Fagnes* (no password is needed).




1 Entrance
2 Reception
3 Room ADN
4 Room Apollinaire
5 Restaurant
6 Bar
7 Swimming pool

# Committees

**Scientific committee**

Davy Paindaveine (ULB, Chair)
Gentiane Haesbroeck (ULiège)
Niel Hens (UHasselt and UAntwerp)
Catherine Legrand (UCLouvain)
Beatrijs Moekerke (UGent)
Marcel Rémon (UNamur)
Germain Van Bever (UNamur)
Ingrid Van Keilegom (KULeuven)
Thomas Verdebout (ULB)

**Organizing committee**

Davy Paindaveine (ULB, Chair)
Nancy De Munck (ULB)
Gentiane Haesbroeck (ULiège)
Niel Hens (UHasselt and UAntwerp)
Pierre Jeurissen (ULB)
Gilles Nisol (ULB)
Julien Remy (ULB)
Germain Van Bever (UNamur)
Thomas Verdebout (ULB)
Catherine Vermandele (ULB)

# Invited speakers

**October 17 (PhD day)**

Roland Fried (TU Dortmund University)
Klaus Nordhausen (Vienna University of Technology)

**October 18-19**

Jessica Barrett (University of Cambridge)
Gerda Claeskens (KULeuven)
Rhian Daniel (Cardiff University)
Els Goetghebeur (UGent)
Chris Jones (The Open University)
Roger Koenker (University College London)
María Dolores Martínez-Miranda (University of Granada)
Johan Segers (UCLouvain)
Steffen Unkel (University of Göttingen)

# Conference schedule

| October 17 | | |
|---|---|---|
| 12:30-13:50 | Registration & sandwich buffet | |
| 13:50-14:25 | **Klaus Nordhausen** (Chair: Davy Paindaveine)<br>*Enhancing your research with R* | ADN |
| 14:25-15:00 | **Kelly Van Lancker** (UGent) (Chair: An Vandebosch)<br>*Improving interim decisions in randomized trials by exploiting information on short-term outcomes and prognostic baseline covariates*<br>Discussant: Elisabeth Coart (IDDI) | ADN |
| 15:00-15:30 | Poster storm (Chair: Thomas Verdebout) | |
| 15:30-16:15 | Poster session & Coffee break | |
| 16:15-16:50 | **Bastien Marquis** (ULB) (Chair: Eugen Pircalabelu)<br>*Comparison of optimisation bias in unstructured and structured sparse variable selection*<br>Discussant: Rainer von Sachs (UCLouvain) | ADN |
| 16:50-17:25 | **Roland Fried** (Chair: Marcel Rémon)<br>*How (not) to publish statistical research – a personal collection of someone from the middle ages* | ADN |
| 17:30-18:30 | **Job fair** | ADN |
| 18:30-19:30 | Reception | |
| 19:30- | Dinner | |

| October 18 | | |
|---|---|---|
| 8:45-9:15 | Registration | |
| 9:15-9:20 | Welcome word | |
| 9:20-10:00 | **Roger Koenker** (Chair: Davy Paindaveine)<br>*NPMLE methods for mixture models* | |
| 10:00-10:40 | Parallel session 1, room ADN<br>(Chair: Germain Van Bever) | Parallel session 2, room Apollinaire<br>(Chair: Ariel Alonso Abad) |
| | **Anna Kiriliouk** (UNamur)<br>*Climate event attribution using multivariate peaks-over-thresholds modelling* | **Maren Vranckx** (UHasselt)<br>*Comparison of different software implementations for spatial disease mapping* |
| | **Eugen Pircalabelu** (UCLouvain)<br>*Community based grouping for undirected graphical models* | **Wen Wei Loh** (UGent)<br>*Estimation of controlled direct effects in longitudinal mediation analyses with latent variables in randomised studies* |
| 10:40-11:00 | Coffee break | |
| 11:00-11:40 | **Rhian Daniel** (Chair: Catherine Legrand), Quetelet society invited talk<br>*On scientifically relevant mediation-style questions, and approaches to answering them from large multi-omic datasets* | |
| 11:40-12:20 | **Johan Segers** (Chair: Marc Hallin)<br>*Multivariate tail quantile contours based on optimal transport* | |
| 12:20-13:40 | Lunch | |
| 13:40-14:20 | **Chris Jones** (Chair: Gentiane Haesbroeck)<br>*Flexible distributional modelling for parametric survival analysis* | |
| 14:20-15:20 | Parallel session 3, room ADN<br>(Chair: Yvik Swan) | Parallel session 4, room Apollinaire<br>(Chair: Jan De Neve) |
| | **Fatemeh Ghaderinezhad** (UGent)<br>*Choose or not to choose a prior. That's the question!* | **Fien Gistelinck** (UGent)<br>*Modeling longitudinal dyadic data in the SEM framework* |
| | **Oswaldo Gressani** (UCLouvain)<br>*Bridging the gap between Bayesian P-splines and Laplace's method for inference in generalized additive models* | **Gilles Nisol** (ULB)<br>*Prediction of singular VARs and application to generalized dynamic factor models* |
| | **Justine Loncke** (UGent)<br>*The social relations model for count data: to Bayes or not to Bayes* | **Guy Mélard** (ULB)<br>*Time series models with time-dependent coefficients: asymptotic results* |
| 15:20-16:20 | Poster session & Coffee break | |
| 16:20-17:00 | **Jessica Barrett** (Chair: Beatrijs Moerkerke), Quetelet society invited talk<br>*Building a dynamic risk prediction model for cardiovascular disease* | |
| 17:00-17:50 | Quetelet prize presentations (Chair: Beatrijs Moerkerke)<br>**Marion Louvel** (UGent). *On standardising quality of care indicators based on summary statistics*<br>**Johan Verbeeck** (UHasselt). *Generalized pairwise comparison methods to analyze (non)-hierarchical composite endpoints* | |
| 18:00-19:30 | General assembly | |
| 19:30- | Reception and conference dinner | |

| October 19 | | |
|---|---|---|
| 8:45-9:00 | Registration | |
| 9:00-9:40 | **María Dolores Martínez-Miranda** (Chair: Ingrid Van Keilegom)<br>*Goodness-of-fit tests in proportional hazards models with random effects* | |
| 9:40-10:40 | Parallel session 5, room ADN<br>(Chair: Anna Kiriliouk) | Parallel session 6, room Apollinaire<br>(Chair: Mia Hubert) |
| | **Alexandre Jacquemain** (UCLouvain)<br>*Lorenz regression* | **David Preinerstorfer** (ULB)<br>*Power in high-dimensional testing problems* |
| | **Kassu Mehari Beyene** (UCLouvain)<br>*Smooth time-dependent ROC curve for right censored survival data* | **Oliver Dukes** (UGent)<br>*High-dimensional doubly robust tests for regression parameters* |
| | **Vahe Avagyan** (UGent)<br>*Stable IPW estimation for longitudinal studies* | **Vaidotas Characiejus** (ULB)<br>*Testing for hidden periodicities in functional time series* |
| 10:40-11:00 | Coffee break | |
| 11:00-11:40 | **Gerda Claeskens** (Chair: Thomas Verdebout)<br>*First aid after model selection* | |
| 11:40-12:20 | **Els Goetghebeur** (Chair: Thomas Bruss) *On the causal effect of gender: beyond the (many!) anecdotes, a statistician's view on evidence of gender bias* | |
| 12:20-13:30 | Lunch | |
| 13:30-14:30 | Parallel session 7, room ADN<br>(Chair: Jan Beirlant) | Parallel session 8, room Apollinaire<br>(Chair: Thomas Neyens) |
| | **Elissa Burghgraeve** (UGent)<br>*James-Stein estimators in factor analysis* | **Oana Petrof** (UHasselt)<br>*On the impact of residential history in the spatial analysis of diseases with long latency period: a study of Mesothelioma in Belgium* |
| | **Julien Remy** (ULB)<br>*Testing for principal component directions under weak identifiability* | **Olajumoke Owokotomo** (UHasselt)<br>*Detection of high dimensional intestinal microbiota as biomarker for immunological response: a Bayesian variable selection approach* |
| | **Gilles Mordant** (UCLouvain)<br>*Distribution comparison tests based on self-similar transport of measure* | **Olusoji Oluwafemi** (UNamur)<br>*"cyanoFilter", an automated framework for identifying picocyanobacteria populations obtained via flow cytometry* |
| 14:35-15:15 | **Steffen Unkel** (Chair: Niel Hens)<br>*The genesis and use of time-varying frailty models for representing heterogeneities in the transmission of infectious diseases* | |
| 15:15 | Closing | |

# Abstracts of posters (Oct 17-18)

### A parametric methodology to estimate the variance matrix of classical measurement error

Elif Akça* and Ingrid Van Keilegom
KULeuven

Measurement error is a ubiquitously faced challenge and in practice, it is possible that more than one covariate is measured with an error. When measurements are taken by the same device, the errors within each covariate are correlated. In this work, we present a novel approach to estimate the variance-covariance matrix of classical additive errors in the absence of validation data or auxiliary variables when two covariates are not measured precisely. We show that the elements of the variance-covariance matrix are identifiable. To investigate the performance of the proposed technique, a diverse set of simulation studies is conducted and the entire course is demonstrated on a real dataset.

Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition. Chapman & Hall, Boca Raton.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. Boca Raton: CRC Press.

---

### Assessing cure status prediction from survival data using ROC curves

Maïlis Amico* and Ingrid Van Keilegom
KULeuven

Survival analysis relies on the hypothesis that, if the follow-up will be long enough, the event of interest will eventually be observed for all observations. This assumption, however, is often not realistic. In fact, in some situations a fraction of the subjects may never experience the event of interest. The survival data then contain a fraction of 'cured' or long-term survivors, usually associated with infinite survival times. A common approach to model this type of data consists in using cure models. Two types of information can therefore be obtained: the survival at a given time and the cure status, both possibly modelled as a function of the covariates. The cure status is often of interest for medical practitioners, and one is usually interested in predicting it based on markers. ROC curves are one way to evaluate these predicting performances. However, the 'classical' ROC curve method is not appropriate since the cure status is partially unobserved due to the presence of censoring. In this research, we propose a ROC curve estimator aiming to evaluate the accuracy of the cure status prediction from survival data. This estimator decomposes

sensitivity and specificity based on the definition of conditional probability and estimates these two quantities by means of weighted empirical distribution functions. Based on simulations, we demonstrate the good performance of our proposal and compare it with the 'classical' nonparametric ROC curve estimator that would be obtained if the cure status was fully observed. We finally illustrate the methodology on a cancer dataset.

---

## Extremal dependence in a Hüsler-Reiss Markov tree

Stefka Asenova[1*], Johan Segers[1] and Gildas Mazo[2]
[1]UCLouvain; [2]MaIAGE, INRA, Université Paris-Saclay

Tree graphical models are suitable for analyzing data where variables are naturally connected through a tree graph. Such variables can be found in studying quantities measured on a river: water level, water flow, pollution concentration in the water or the soil nearby the river, etc. In such a model, the joint distribution is completely determined by the tree structure and the bivariate distributions along the edges of the tree. A regularly varying Markov tree (Segers and Mazo, 2018) is a tree graphical model the joint distribution of which is assumed to be regularly varying. We study models for extremal dependence within tree graphical models, propose estimators of the model parameters, and assess their properties through Monte Carlo simulations. Our focus is on two models in particular, a Hüsler–Reiss and a max-linear tree model. The Hüsler–Reiss model is based on the assumption that the pairwise copula is in the max-domain of attraction of the bivariate one-parameter Hüsler–Reiss copula. We compare three different estimators: the continuously updating weighted least squares estimator of Einmahl, Kiriliouk and Segers (2016), an estimator based on Engelke et al. (2015) and one based on covariance selection models and maximum likelihood estimation (Lauritzen, 1996). The max-linear tree statistical model arises when a max-linear relation is assumed for each pair of adjacent nodes.

Segers, J., and Mazo, G. (2018). Regularly varying Markov trees.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, Oxford.

Engelke, S., Malinowski, A., Kabluchko, Z., and Schlather, M. (2015). Estimation of Hüsler-Reiss distributions and Brown-Resnick processes. *Journal of the Royal Statistical Society Series B*, 77, 239–265.

Einmahl, J., Kiriliouk, A., and Segers, J. (2016). A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes*, 22, 205-233.

---

## Testing uniformity on high-dimensional spheres against even rotationally symmetric alternatives

Christine Cutting*, Davy Paindaveine and Thomas Verdebout
ULB

We are interested in testing uniformity on high-dimensional unit spheres against even rotationally symmetric alternatives (distributions invariant under rotations $O$ such that $O\theta = \theta$ for some direction $\theta$ and symmetrically distributed on either sides of the hyperplane orthogonal to $\theta$). When $\theta$ is specified, the model is Locally Asymptotically Normal (LAN). When it is not specified, there is no $\sqrt{n}$-consistent estimator of $\theta$ in the general case, because there are two very different possibilities: bipolar distributions and girdle distributions. Both cases are investigated in the low-dimensional and high-dimensional settings.

---

## Statistical approaches to estimate economic performance for global business

Antonio Frenda
Sapienza University of Rome, Pegaso Telematic University

In the near future, the statistical estimation of the value added of leading business groups in Europe and of large complex units could become a source for the preliminary estimates of GDP at a European level, or be used for the further development of existing indicators of European growth. However, it is often challenging to produce such data, because in standard accounting formats the distinction between national and foreign activities is not always requested. The following case studies explain how it is possible to solve some of the problems that arise when trying to calculate group accounts that are useful for establishing aggregate statistical indicators, starting from the accounts of the individual companies and branches. The chosen method depends on the availability of data concerning foreign production (in particular for enterprises involved in the construction field and those operating in Internet), accounting criteria used in certain countries, and vertical integration: when some of these are unavailable, it may result in a particular method being chosen over another. As highlighted in the European System of Accounts in Eurostat (2010), the centre of predominant economic interest of an enterprise indicates that a location exists where this unit engages in economic activities and transactions on a significant scale within a country's economic territory. Some statistical, fiscal and administrative sources are outlined that can be used to sketch the domestic economic performance of the main enterprises carrying out activities abroad. The paper provides robust statistical methods regarding the utilisation of such sources.

---

## Confidence curves post-AIC selection

Gerda Claeskens and Andrea Cristina Garcia Angulo*
KULeuven

Post-selection inference is a recent methodology that incorporates the extra variability added by model selection to perform valid inference. Recent work in inference after model selection has focused on producing valid p-values and confidence intervals for the parameters of interest. Charkhi and Claeskens (2018) proposed an asymptotically valid post-selection interval when Akaike's information criterion AIC is used for model selection. They showed that the limiting density function of the parameter estimator conditional on AIC-selection is a truncated normal of which the domain is given by the selection region from which an approximate pivotal quantity for inference is obtained. We propose to use this result and the concept of confidence distributions (Schweder and Hjort, 2016) to produce post-selection confidence distributions and post-selection confidence curves in order to show the effect of model selection at all possible levels of confidence. We use the cumulative distribution function of the obtained pivot to produce the post-AIC selection confidence distribution. As a result, the post-AIC selection confidence curves are much wider than the classic naïve confidence curves with drastic effects even at the lowest levels of confidence reaffirming that ignoring model selection leads to invalid inference.

Charkhi, A., and Claeskens, G. (2018). Asymptotic post-selection inference for Akaike's information criterion. Available at https://lirias.kuleuven.be/bitstream/123456789/616160/1/KBI 1804:pdf

Schweder, T., and Hjort, N. L. (2016). *Confidence, Likelihood, Probability.* New York: Cambridge University Press.

---

## Randomization inference with general interference and censoring

Mohammad Ali[1], John D. Clemens[2], Michael E. Emch[3], Michael G. Hudgens[3] and Wen Wei Loh[4*]

[1]Johns Hopkins University, Baltimore; [2]University of California, Los Angeles; [3]University of North Carolina, Chapel Hill; [4]UGent

Interference occurs between individuals when the treatment (or exposure) of one individual affects the outcome of another individual. Previous work on causal inference methods in the presence of interference has focused on the setting where a priori it is assumed there is 'partial interference', in the sense that individuals can be partitioned into groups wherein there is no interference between individuals in different groups. Bowers, Fredrickson, and Panagopoulos (2012) and Bowers, Fredrickson, and Aronow (2016) consider randomization-based inferential methods that allow for more general interference structures in the context of randomized experiments. In this paper, extensions of Bowers et al. are considered, including allowing for failure time outcomes subject to right censoring. Permitting right censored outcomes is challenging because standard randomization-based tests of the null hypothesis of no treatment effect assume that whether an individual is censored does not depend on treatment. The proposed extension of Bowers et al. to allow for censoring entails adapting the method of Wang, Lagakos, and Gray (2010) for two sample survival comparisons in the presence of unequal censoring. The methods

are examined via simulation studies and utilized to assess the effects of cholera vaccination in an individually-randomized trial of 73,000 children and women in Matlab, Bangladesh.

Bowers, J., Fredrickson, M. M., and Aronow, P. M. (2016). Research note: A more powerful test statistic for reasoning about interference between units. *Political Analysis*, 24, 395–403.

Bowers, J., Fredrickson, M. M., and Panagopoulos, C. (2012). Reasoning about interference between units: A general framework. *Political Analysis*, 21, 97–124.

Wang, R., Lagakos, S. W., and Gray, R. J. (2010). Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring. *Biostatistics*, 11, 676–692.

---

## Comparison of model regularization methods for identifying groups of predictive variables

Bernadette Govaerts, Rebecca Marion* and Rainer von Sachs

UCLouvain

Model regularization methods that perform embedded variable selection during the model estimation process have great potential for increasing model interpretability. In cases where the predictor variables form groups (i.e. variables within a group are highly correlated but variables from different groups are not correlated), the selection of important variables using model regularization is more challenging, especially if these groups are not known a priori. Several methods proposed in the literature (e.g. Cluster Elastic Net, Pairwise Absolute Clustering and Sparsity, Sparse Laplacian Shrinkage, Simultaneous Supervised Clustering and Feature Selection, Bayesian Sparse Group Selection, etc.) are able to learn variable groups from the data and select important groups during model estimation. However, the empirical performance of these methods has not been assessed in situations where variables from different groups are at least somewhat correlated. This poster presents the results of simulation studies assessing the prediction and variable selection performance of these methods in the presence of correlated variable groups, identifying factors that most weaken their performance in this context.

Sharma, D., Bondell, H. D., and Zhang, H. (2013). Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics*, 22, 319–340.

Shen, X., Huang, H.-C., and Pan, W. (2012). Simultaneous supervised clustering and feature selection over a graph. *Biometrika*, 99, 899–914.

Witten, D. M., Shojaie, A., and Zhang, F. (2014). The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics*, 56, 112–122.

---

## Uncertainty quantification in sunspot counts

Véronique Delouille[1], Laure Lefèvre[1], Sophie Mathieu[2*] and Rainer von Sachs[2]
[1]Royal Observatory of Belgium; [2]UCLouvain

Sunspots (SS) are dark spots appearing in groups on the solar surface as a manifestation of solar magnetism. While the time series of SS counts acts as a benchmark in a large variety of physical sciences, as of today it lacks proper uncertainty quantification and modeling. This paper details the first comprehensive noise model of the SS counts in a multiplicative framework. We estimate the various error terms using either mixture or Hurdle models combined with overdispersed distributions. Key results are an estimation of short-term error distribution, and an estimation of long-term drift specific to each observatory. The former allows detecting daily outliers, while the latter brings a key element to monitor the stability over time of SS counts recorded by a given observatory.

Cameron, A. C., and Trivedi, P. K. (2013). *Regression Analysis of Count Data*. Cambridge University Press, 2nd edition.

Dudok de Wit, T., Lefèvre, L., and Clette, F. (2016). Uncertainties in the sunspot numbers: estimation and omplications. *Solar Physics*, 291, 2709–2731.

Taylor, J., and Verbyla, A. (2004). Joint modelling of location and scale parameters of the $t$ distribution. *Statistical Modelling*, 4, 91–112.

---

## Semiparametric estimation in AFT mixture cure models

Motahareh Parsa* and Ingrid Van Keilegom
KULeuven

In survival analysis one models the time it takes for events to occur. Assuming a certain proportion of patients to be cured has led to define the mixture cure models. The accelerated failure time (AFT) mixture cure model is one of the popular models in practice. In this article a semiparametric procedure is provided to estimate an AFT cure model while the right censoring is taken into account. For the estimation of the distribution and density function of the error term, we use the Kaplan-Meier estimator and its corresponding kernel density estimator. The method is compared with some other results existing in the literature, in which smooth seminonparametric (SNP) approach is used to maximize the likelihood of the AFT mixture cure model. A simulation study is conducted to investigate the benefits of the suggested methodology. Also, an asymptotic study of the proposed estimators is provided.

Van Keilegom, I., and Veraverbeke, N. (1996). Uniform strong convergence results for the conditional Kaplan-Meier estimator and its qualities. *Communincations in Statistics - Theory and Methods*, 25, 2251-2256.

Van Keilegom, I., and Veraverbeke, N. (1997). Estimation and bootstrap with censored data in fixed design nonparametric regrassion. *Annals of the Institute of Statistical Mathematics*, 49, 467-491.

Zhang, M., and Davidian, M. (2008). Smooth semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics*, 64, 567-576.

---

## Preliminary test estimation in ULAN models

Davy Paindaveine, Joséa Rondrotiana Rasoafaraniaina* and Thomas Verdebout
ULB

We consider the problem of estimating a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ when it is suspected that $\boldsymbol{\theta}$ belongs to a subset $\boldsymbol{\Theta}_0$ of the parameter space $\boldsymbol{\Theta}$. A classical solution in this framework is the *preliminary test estimation*, introduced in Bancroft (1944), that estimates $\boldsymbol{\theta}$ by a constrained estimate or an unconstrained according to the decision of a preliminary test for the null hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. In this work, we investigate the asymptotic properties of such preliminary test estimators under the assumption that the underlying model is uniformly locally and asymptotically normal (ULAN). More precisely, our asymptotic investigation considers three cases: (i) the case where $\boldsymbol{\theta}$ is fixed and does not belong to $\boldsymbol{\Theta}_0$, (ii) the case where it is fixed and belongs to $\boldsymbol{\Theta}_0$, and (iii) an intermediate case where $\boldsymbol{\theta} = \boldsymbol{\theta}_n$ is associated with a sequence of hypotheses that is contiguous to the null hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. In particular, in this challenging case (iii), we use the resulting asymptotic distribution of the preliminary test estimator to compare its performances to competing estimators in terms of MSE. Finally, we compare our results to those obtained by Saleh (2006) in the context of linear regression.

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of signicance. *The Annals of Mathematical Statistics*, 15, 190–204.

Paindaveine, D., Rasoafaraniaina, J. R., and Verdebout, T. (2017) Preliminary test estimation for multi-sample principal components. *Econometrics & Statistics*, 2, 106-116.

Saleh, A. M. E. (2006). *Theory of preliminary test and Stein-type estimation with applications.* John Wiley and Sons.

---

## Identification of Wiener-Hammerstein system with random forest

Kurt Barbe and Md Abu Hanif Shaikh*
VUB

The Wiener-Hammerstein (W-H) system is a popular and easy to understand class of Volterra nonlinear dynamical system. It consists of a static nonlinearity positioned between two dynamical systems. The main identification challenge resides in separating two linear filters as rational form of poles and zeros. According to previous studies, an initial guess of the separated dynamics is made by browsing through all possible partitions of pole-zero. Then based on the root mean square error (RMSE), good partitions are selected and later all parameters are optimized mutually to determine the best model. As RMSE before and after optimization behaves erratically, this study proposes the use of the Spearman correlation to select good models for optimizing. The proposed technique avoids estimation of local nonlinearity, avoid partitioning and optimize only a

single model. Thus a cosmic speed-up in processing time is achieved without any prior knowledge about model configuration. Through the output is not optimal for RF yet proposed identification of W-H system from SYSID'09 Benchmark data gives a better result than Spearman correlation based technique.

Sjoberg, J., and Schoukens, J. (2012). Initializing Wiener-Hammerstein models based on partitioning of the best linear approximation. *Automatica*, 48, 353–359.

Shaikh, M.A.H. , and Barbe, K. (2018). Spearman correlation for initial estimation of Wiener-Hammerstein system. Proc. of IEEE I2MTC.

---

## Life and health actuarial pricing: a biostatistics approach

Michel Denuit, Catherine Legrand and Antoine Soetewey*

UCLouvain

It is generally thought that patients having suffered from a cancer have a lower probability of survival compared to healthy people. Due to this aggravated risk and the relatively small number of patients wishing to take out insurance coverage in case of death, the insurance industry is reluctant to grant such a guarantee. However, survival and life expectancy of cancer patients have been increasing over the last decades and we can reasonably assume that it will keep increasing in the future thanks to medical and technological progress. In regard to this, France passed a law referred as "the right to forget", that is, the right for a person subscribing to a contract not to declare a previous cancer after a period of 10 years after the end of the therapeutic protocol (Sapin and Touraine, 2017). This period being reduced to 5 years if the person is a minor. But some questions remain: the thresholds of 10 and 5 years are arbitrary and do not reflect survival of the persons having suffered from a cancer. There remains some ambiguity about what is considered as treatment, so what marks the end of a therapeutic protocol and in the end when the patient will start to benefit from this right? Finally, this right is very binary and not flexible at all. The aim of the project is twofold: (i) To develop a method to adequately estimate the threshold after which cancer patients can be considered as cured, and (ii) to find a proper way to adapt the actuarial pricing of life insurance products to each category of risk, disease, person, etc. The goal is also to demonstrate that for some types of cancer, the survivors actually have a chance of survival comparable to that of the general population, or pose a moderately increased risk and could therefore be covered in the event of death. This involves measuring and quantifying the potential excess mortality so that the premiums claimed reflect the risk in terms of financial services.

---

## Weight choices for the penalized composite and model-averaged estimator in quantile regression

Daumantas Bloznelis[1], Gerda Claeskens[2] and Jing Zhou[2*]

[1]Inland Norway University of Applied Sciences, Elverum, Norway; [1]KULeuven

Composite estimation requires a specification of a set of weights to average loss functions; model averaging uses weights to average estimators. We investigate the weight choice for high-dimensional quantile regression models. Due to the regularization, there are several approaches to derive the expressions of the so-called optimal weights for both estimators. While assuming perfect selection, Bradic et al. (2011) derived the optimal weights that achieve the lower bound of the estimator's variance for the penalized composite estimator. Under the same assumption, we derived the same type of lower bound of the variance for the model-averaged estimator in Bloznelis et al. (2017). To relax the assumption that the active set is perfectly selected, we consider the robust approximate message passing algorithm (RAMP) in Bradic (2016); the RAMP algorithm generates solutions to the $L_1$-penalized estimators and does not require the loss functions to be differentiable. For the penalized composite estimator, we fit the weighted loss functions into the framework of the RAMP algorithm in Zhou and Claeskens (2018). For the model averaged estimator, we investigate the correlation structure between multiple paralleled RAMP iterations, and derive the lower bound of the estimator's mean squared error. We compare the performances of the two types of the optimal weights, as well as equal weights, for different sparsity and quantile levels for a group of error distributions.

Bloznelis, D., Claeskens, G., and Zhou, J. (2017). Composite versus model-averaged quantile regression. Technical report.

Bradic, J. (2016). Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing. *Electronic Journal of Statistics*, 10, 3894–3944.

Zhou, J., and Claeskens, G. (2018). Optimal weights for the composite and model-averaged quantile regression. Technical report.

# Abstracts of talks: Oct 17

### Enhancing your research with R

Oct 17
13:50–14:25
ADN

Klaus Nordhausen
Vienna University of Technology

R can be considered the lingua franca of statistics. Ideas are often explored and communicated via R code and R packages. Therefore it is essential for statisticians to be able to produce efficient and shareable code. In this talk we will present some strategies on how to write efficient code and how to debug and profile it. Some tricks to speed it up are also discussed and we will show that it is not difficult to create an R package.

---

### Improving interim decisions in randomized trials by exploiting information on short-term outcomes and prognostic baseline covariates

Oct 17
14:25–15:00
ADN

An Vandebosch[1], Kelly Van Lancker[2*] and Stijn Vansteelandt[2,3]
[1]Janssen Pharmaceutica; [2]UGent; [3]University and London School of Hygiene and Tropical Medicine

Interim analyses are routinely used to monitor accumulating data in clinical trials. A problem in such analyses is that all patients may have been enrolled by the time a sufficient number of patients have their primary endpoint available. When the objective of the interim analysis is to stop the trial when treatment is futile, it must ideally be conducted prior to enrollment completion. To remedy this problem, we propose an interim decision procedure which exploits the information contained in baseline covariates and short-term outcomes that are predictive of the final outcome. We show that the proposed procedure leads to a gain in efficiency, an increased power and a reduced sample size, without compromising the Type I error rate of the procedure, even when the used prediction models are misspecified. In particular, implementing our proposal in the conditional power approach allows earlier stopping for true futility whilst controlling the probability for incorrectly stopping. This has the consequence of reducing the number of recruited patients in case of stopping for futility, such that fewer patients get the futile regimen. We support the proposal by simulation studies based on data from a real clinical trial.

---

Oct 17
16:15–16:50
ADN

## Comparison of optimisation bias in unstructured and structured sparse variable selection

Bastien Marquis* and Maarten Jansen
ULB

In sparse high-dimensional data, the selection of a model can lead to an overestimation of the number of nonzero variables. Indeed, the use of an $l_1$ norm constraint while minimising the sum of squared residuals tempers the effects of false positives, thus they are more likely to be included in the model. On the other hand, an $l_0$ regularisation is a non-convex problem and finding its solution is a combinatorial challenge which becomes unfeasible for more than 50 variables. To overcome this situation, one can come up with a selection via an $l_1$ penalisation but estimate its coefficients without shrinkage. This leads to an additional bias in the optimisation of an information criterion over the model size. Used as a stopping rule, this IC must be modified to take into account the deviation of the estimation with and without shrinkage. By looking into the difference between the Prediction Error and the expected Mallows's Cp, previous work analysed a correction for the optimisation bias and an expression can be found for a signal-plus-noise model given some assumptions. A focus on structured models, in particular grouped variables, shows similar results, though the bias is noticeably reduced.

---

Oct 17
16:50–17:25
ADN

## How (not) to publish statistical research - a personal collection of someone from the middle ages

Roland Fried
TU Dortmund University

Writing and publishing good papers (whatever "good" means in this context) is an art which requires not only the capacity of doing high-level research, but among others also language skills, organising ability, and experience. Starting from the problem of identifying an interesting research problem up to how to monitor the review process of your paper, there are many relevant issues which need to be taken into account, including how to organize your paper, motivate readers, select a journal and convince reviewers. This talk tries to give some guidance based on the presenter's own experience collected during 25 years of research, 20 years of refereeing papers and 10 years of editorial service, combined with a small literature review. The outcome of these efforts, though certainly being far from perfect, might hopefully be worth being considered and serve as a starting point.

---

# Abstracts of talks: Oct 18

### NPMLE methods for mixture models

Oct 18
9:20–10:00
ADN

Roger Koenker[1*] and Jiaying Gu[2]
[1]University College London; [2]University of Toronto

Unobserved heterogeneity is a pervasive feature of many statistical applications often effectively modeled by parametric mixtures. However, the nonparametric maximum likelihood methods proposed by Robbins (1951) and Kiefer and Wolfowitz (1956) offer a more flexible approach and are now efficiently computable with modern convex optimization techniques. This approach will be illustrated with applications to Gaussian mixtures for longitudinal data, Weibull mixtures for survival data and discrete mixtures for binary response.

---

### Climate event attribution using multivariate peaks-over-thresholds modelling

Oct 18
10:00–10:20
ADN

Anna Kiriliouk[1*] and Philippe Naveau[2]
[1]UNamur; [2]Laboratoire des Sciences du Climat et l'Environnement CNRS

Quantifying the human influence on climate change and identifying potential causes is a highly relevant research area which is often referred to as detection and attribution. A common approach is to compare the probability of an extreme event in the factual world to the probability of an extreme event in a counterfactual world, i.e., a world that might have been if no humans would have existed. The event probabilities can be calculated using large scale climate model runs that simulate the evolution of the climate with and without anthropogenic forcings. The Fraction of Attributable Risk (FAR) is defined as the relative ratio of event probabilities in the factual and in the counterfactual world. Estimating the FAR will allow us to quantify the extent to which human activities have increased the risk of occurrence of an extreme event. While this subject has been gaining attention, no multivariate extreme-value theory has been used up to date. We propose a semi-parametric model for the FAR based on the multivariate generalized Pareto distribution, i.e., the asymptotic distribution of suitably normalized exceedances over a high threshold. Contrary to current methods, our approach allows us to take the spatial dependence into account. The model is then used to quantify the increased risk of an extreme rainfall event over a large spatial region in Europe.

---

Oct 18
10:20–10:40
ADN

## Community based grouping for undirected graphical models

Gerda Claeskens[1] and Eugen Pircalabelu[2*]
[1]KULeuven; [2]UCLouvain

A new strategy of probabilistic graphical modeling is developed that draws parallels from social network analysis. Probabilistic graphical modeling summarizes the information coming from multivariate data in a graphical format where nodes, corresponding to random variables, are linked by edges that indicate dependence relations between the nodes. The purpose is to estimate the structure of the graph (which nodes connect to which other nodes) when data at the nodes are available. On the opposite side of the spectrum, social network analysis considers the graph as the observed data. Given thus the graph where connections between nodes are observed rather than estimated, social network analysis estimates models that represent well an underlying mechanism which has generated the observed graph. We propose a new method that exploits the strong points of each framework as it estimates jointly an undirected graph and communities of homogenous nodes, such that the structure of the communities is taken into account when estimating the graph and conversely, the structure of the graph is accounted for when estimating homogeneous communities of nodes. The procedure uses a joint group graphical lasso approach with community detection-based grouping, such that some groups of edges co-occur in the estimated graph. The grouping structure is unknown and is estimated based on community detection algorithms. Theoretical derivations regarding graph convergence and sparsistency, as well as accuracy of community recovery are included, while the method's empirical performance is illustrated in an fMRI context, as well as with simulated examples.

---

Oct 18
10:00–10:20
Apollinaire

## Comparison of different software implementations for spatial disease mapping

Christel Faes, Thomas Neyens and Maren Vranckx*
UHasselt

Disease mapping is a scientific field that aims to understand and predict disease risk at specific locations of a geographic area of interest. Many software implementations to model risk distributions exist, with the most commonly used methods using Bayesian estimation techniques. Many of these applications differ, to varying degrees, in the underlying methodology and it remains unclear if the results that they produce are alike. This study provides an in-depth comparison between analysis results, coming from CARBayes, R2OpenBUGS, R2BayesX, INLA and rstan. We investigate different parameterizations of conditional autoregressive (CAR) models for spatially discrete count data. These models are typically used to estimate the relative disease risks, while correcting for spatial association. Data about diabetics in children and young adults in Belgian are used in a case study, while a simulation is undertaken to assess software performance in different settings. This investigation shows that CAR models may be estimated readily using a variety of implementations, but it also highlights important differences in estimation results and indicates that a number of software implementations are less optimal.

---

## Estimation of controlled direct effects in longitudinal mediation analyses with latent variables in randomised studies

Oct 18
10:20–10:40
Apollinaire

Tom Loeys, Wen Wei Loh*, Beatrijs Moerkerke and Stijn Vansteelandt

UGent

In a randomised study with longitudinal data on a mediator and outcome, the direct effect of the treatment or exposure on the outcome at a particular time includes all paths that avoid any instances of the mediator. Estimation of each direct effect thus requires adjusting for observed confounders between the outcome at that time and all preceding instances of the mediator. But when the confounders are affected by treatment, standard regression adjustment is prone to possibly severe bias. In this paper, we propose a G-estimation method (Robins, 1999) for unbiased estimation of the controlled direct effect of treatment on the outcome at each time in the presence of a time-varying mediator. We adapt existing methods for time-varying treatments (Vansteelandt and Sjolander, 2016) that only require correctly specifying either the mediator model or the outcome model. Further, we extend previous work by Moerkerke, Loeys, and Vansteelandt (2015) where under a certain class of linear models, unbiased estimators of the controlled direct effect can be obtained within the structural equation modelling (SEM) framework by carefully combining the estimated path coefficients for the constituent paths. The setting where the mediator or outcome, or both, are latent but are measured by manifest items at each time is also considered. We develop a novel G-estimation method that entails first estimating unbiased surrogates of the latent variables, then correcting for biases due to measurement error. The method builds on the G-estimation method proposed by Loeys et al. (2014) for a single time point.

Loeys, T., Moerkerke, B., Raes, A., Rosseel, Y., and Vansteelandt, S. (2014). Estimation of controlled direct effects in the presence of exposure-induced confounding and latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 21, 396–407.

Moerkerke, B., Loeys, T., and Vansteelandt, S. (2015). Structural equation modeling versus marginal structural modeling for assessing mediation in the presence of posttreatment confounding. *Psychological methods*, 20, 204.

Robins, J.M. (1999). Testing and estimation of direct e$_{ects}$ by reparameterizing directed acyclic graphs with structural nested models. *Computation, causation, and discovery*, 349–405.

Vansteelandt, S., and Sjolander, A. (2016). Revisiting G-estimation of the e$_e$*ct* of a time-varying exposure subject to time-varying confounding. *Epidemiologic Methods*, 5, 37–56.

Oct 18
11:00–11:40
ADN

## On scientifically relevant mediation-style questions, and approaches to answering them, from large multi-omic datasets

Rhian Daniel
Cardiff University

It is increasingly popular to perform traditional mediation analysis (eg using products of coefficients) in multi-omic datasets with a very large number of mediators. For example, a linear regression model with LDL-cholesterol as the outcome, and a particular SNP together with one of c.4,000 proteins as two predictors will be fitted, followed by a second linear regression model with the protein as the outcome, and the SNP as the predictor. The product of the coefficient of the SNP in the second model and the coefficient of the protein in the first model is taken to represent the part of the effect of the SNP on LDL-cholesterol mediated by that protein. This is then repeated in turn for each of the c.4,000 proteins.

In this talk I will start by reasoning why one might embark on such an analysis, what scientific question might it be roughly addressing, and whether an alternative formulation of this question, which more closely aligns with the scientific objectives, exists. I will then discuss the potential pitfalls of relying on a simple analysis strategy such as the above, and suggest an alternative, justifying its performance using a simulation study and illustrating its use with an analysis of multi-omic data from the UCLEB consortium.

---

Oct 18
11:40–12:20
ADN

## Multivariate tail quantile contours based on optimal transport

Cees de Valk[1] and Johan Segers[2*]
[1]KNMI; [2]UCLouvain

The theory of optimal transport originated in engineering and economics, answering questions of how to rearrange the distribution of resources such as to minimize the cost of the transport plan. A well-known application in statistics is the Wasserstein distance between two distributions, quantifying the best coupling between two random vectors with finite second moments. A new idea was proposed by Chernozhukov, Galichon, Hallin and Henry (2017) in their Monge–Kantorovich quantile contours of a multivariate distribution, defined via transformations of a uniform reference measure through gradients of convex functions, the multivariate equivalent of non-decreasing functions on the real line. We propose a variation of their definition that is better suited to heavy-tailed distributions since it does not require moment constraints. Tail quantile contours of a multivariate regularly varying distribution turn out to converge to a limiting shape as the tail probability tends to zero, a shape that can be described via a transformation of a compact convex body. This information empowers the construction of nonparametric estimators of such tail quantile contours.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge-Kantorovich depth, quantiles, ranks and signs. *Annals of Statistics*, 45, 223–256.

---

## Flexible distributional modelling for parametric survival analysis

Oct 18
13:40–14:20
ADN

Chris Jones
The Open University, U.K.

I'll describe a general, flexible yet parsimonious, parametric survival modelling framework based on an (Adapted) "Power Generalized Weibull" ((A)PGW) distribution, encompassing key hazard function shapes (constant, monotone, bathtub, upside-down-bathtub and no others) and several popular survival distributions (including log-logistic, Weibull, Gompertz). This generality is achieved using four basic distributional parameters, two scale-type and two shape. Incorporating covariate dependence, the scale-type regression components correspond to accelerated failure time and proportional hazards models. In general, the use of multi-parameter regression, in which more than one distributional parameter depends on covariates, is advocated. There is a cute frailty relationship between a PGW distribution with one value of its distribution-choice parameter and a PGW distribution with a smaller value of that parameter. This relationship is exploited to propose a bivariate shared frailty model with PGW marginal distributions: these marginals are linked by the BB9 or "power variance function" copula. This choice of copula is, therefore, natural in the current context. I'll then adapt the bivariate PGW distribution, in turn, to accommodate APGW marginals. In this talk, I intend to concentrate more on ideas and properties rather than on implementation and application. It is based on joint work with Kevin Burke and Angela Noufaily, our papers adding much practical detail and applications. For example, in the univariate case, while many choices are available, it is suggested that covariates are introduced through just one of the scale parameters in combination with a "power" shape parameter, while the other, distribution-choice, shape parameter remains covariate-independent.

---

## Choose or not to choose a prior. That's the question!

Oct 18
14:20–14:40
ADN

Fatemeh Ghaderinezhad* and Christophe Ley
UGent

In Bayesian statistics, one combines the prior information coming from last experiences with the data distribution to set up a full probability model including observable and unobservable quantities consistent with knowledge about the scientific problem and data collection process. The question is as more and more data are collected, how we can quantify the impact of choosing different priors. Stein's method is an instrument which provides lower and upper bounds on the Wasserstein distance between two posteriors based on two different priors (even improper priors) at fixed sample size.

Ley, C., Reinert, G., and Swan, Y. (2017). Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. *The Annals of Applied Probability* 27, 216–241.

Ghaderinezhad, F., and Ley, C. (2018). A general measure of the impact of priors in Bayesian statistics via Stein's Method. Submitted.

---

Oct 18
14:40–15:00
ADN

## Bridging the gap between Bayesian P-splines and Laplace's method for inference in generalized additive models

Oswaldo Gressani[1*] and Philippe Lambert[2]
[1]UCLouvain; [2]ULiège

Ever since the dawn of statistical science, regression analysis has played a central role in the literature, giving birth to models that grew in profusion to accommodate all sorts of relationships between variables of interest. Generalized additive models (GAMs) are a well-established statistical tool for modeling complex nonlinear relationships between a response belonging to an exponential family and a set of covariates. Recently, an approximate Bayesian approach (Rue et al., 2009) termed Integrated Nested Laplace Approximations (INLA) has emerged in the literature and has been largely acclaimed for its computational efficiency to obtain posterior marginals in structured additive regression models with a Gaussian latent field. We develop a novel inferential methodology for GAMs characterized by a flexible estimation of smooth functions with Bayesian P-splines (Lang and Brezger, 2004) and a rapid approximation of joint posterior distributions of latent model variables with Laplace's method. The gradient and Hessian of the hyperparameters are analytically available and a moment-matching technique allows to capture possible asymmetries in the posterior distribution of the penalty parameter vector. The suggested methodology is an extension of the Laplace-P-spline model proposed in Gressani and Lambert (2018) which has proved to work well in a particular class of survival models, largely outperforming the computational speed of Markov chain Monte Carlo methods. A simulation study is implemented as a performance measure with encouraging results.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society, Series B*, 71, 319–392,

Lang, S., and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13, 183–212.

Gressani, O., and Lambert, P. (2018). Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. *Computational Statistics and Data Analysis*, 124, 151–167.

---

Oct 18
15:00–15:20
ADN

## The social relations model for count data: to Bayes or not to Bayes

Tom Loeys and Justine Loncke*
UGent

The family social relations model (SRM) is widely used to identify the sources of variance in interpersonal dispositions in families. Traditionally, it makes use of dyadic measurements that are obtained according to a round-robin design, where each family member rates the other family members on a specific interpersonal disposition. In this study, we employ the blocked design which is restricted to merely intergenerational dyadic measurements (e.g. parent-child). The parameters of the SRM can be estimated by defining it in a multilevel framework and applying a Bayesian method using Gibbs sampling. Such Bayesian approach has already been

proposed for continuous outcomes for the SRM without family roles. Here, we aim to extend that approach to the SRM with family roles and to accommodate it to a count outcome variable. However, this approach might result in biased parameters of the variances for such a small group size in com bination with a small sample size. Alternatively, the parameters of the model can also be estimated using the traditional SEM-framework by approaching the SRM-analysis as a confirmatory factor analysis (CFA) with count indicators. We perform a simulation study where the performance in the Bayesian framework is compared to the performance of CFA with count factor indicators in the SEM-framework

---

## Modeling longitudinal dyadic data in the SEM framework

Oct 18
14:20–14:40
Apollinaire

Fien Gistelinck* and Tom Loeys

UGent

In dyadic research, people are often interested in estimating the effect of one's own (i.e., actor effect) and one's partner (i.e., partner effect) predictor variable on an outcome variable. Due to the nonindependence between the two dyad members, statistical models such as the actor-partner interdependence model (APIM) have been developed to analyze the outcomes of the two dyad members simultaneously.
When dyads are measured over time, not only the scores of the members are nested within a dyad, but the scores of a dyad member at different time points are also correlated. One way to incorporate both interdependencies is to extend the APIM to the longitudinal case: the longitudinal APIM (L-APIM). This model both accounts for between-dyad variation at level-2 (the so-called G-side), and for nonindependence of level-1 residuals in each dyad (the so-called R-side). The latter can take complex forms such as "UN@AR(1)", which allows for correlation within a dyad at a specific time point and a first-order autoregressive process for the measurements over time within each dyad member.
While the implementation of such complex covariance structure is available in few multilevel modeling software (e.g., SAS), we show how it can be implemented in structural equation modeling (SEM) software, such as the R-package "Lavaan". Given the complexity of the code to model such advanced covariance structures in SEM, a Shiny-application was developed to enable applied dyadic researchers to fit the L-APIM on their longitudinal dyadic data within the SEM framework.

---

Oct 18
14:40–15:00
Apollinaire

## Prediction of singular VARs and application to generalized dynamic factor models

Gilles Nisol[1]* and Siegfried Hörmann[2]
[1]ULBruxelles; [2]Graz University of Technology

Vector autoregressive processes (VARs) with innovations having a singular covariance matrix (in short singular VARs) appear naturally in the context of dynamic factor models. The Yule-Walker estimator of such a VAR is problematic, because the solution of the corresponding equation system tends to be numerically rather unstable. For example, if we overestimate the order of the VAR, then the singularity of the innovations renders the Yule-Walker equation system singular as well. Moreover, even with correctly selected order, the Yule-Walker system tends be close to singular in finite sample. In this paper we are going to show that this has a severe impact on predictions. While the asymptotic rate of the mean square prediction error (MSPE) can be just like in the regular (non-singular) case, the finite sample behaviour is suffering. This effect will turn out to be particularly dramatic in context of dynamic factor models, where we do not directly observe the so-called common components which we aim to predict. Then, when the data are sampled with some additional error, the MSPE often gets severely inflated. We will explain the reason for this phenomenon and show how to overcome the problem. Our numerical results underline that it is very important to adapt prediction algorithms accordingly.

---

Oct 18
15:00–15:20
Apollinaire

## Time series models with time-dependent coefficients: asymptotic results

Rajae Azrak[1] and Guy Mélard[2]*
[1]Université Mohammed V – Rabat; [2]ULB

This paper is a follow-up to Azrak and Mélard (2006) and Alj et al. (2017) where we have studied ARMA and VARMA models with time-dependent coefficients, or tdARMA and tdVARMA models. In addition, the innovation scatter can also be time-dependent. The model coefficients and the innovation scatter are supposed to be deterministic functions of time $t$, and possibly of the time series length $n$, and of a small number of parameters. A typical change with respect to classical models would be to replace these coefficients by slowly varying functions of $t/n$. However, the coefficients (and scatter) are not necessarily smooth functions of time, like sudden breaks in a coefficient or periodic heteroscedasticity. Some assumptions are required on the model to obtain a consistent QML estimator for the parameters that is also asymptotically normal. Besides parameter estimation, the asymptotic covariance matrix of the estimator is required with a procedure to estimate it. Contrarily to Azrak and Mélard (2006) where the model does not depend on $n$, limit theorems for sequences are replaced by similar theorems for triangular arrays. With respect to Azrak and Mélard (2006), we provide new theoretical results: (i) a fundamental theorem for the asymptotic theory based on Lehmann and Casella (1998); (ii) a lemma for reducing the moment assumption from 8 to 4; (iii) two theorems to establish convergence for the matrices in the asymptotic covariance matrix of the estimator; (iv) two practical methods to evaluate these matrices. We apply these results to tdVMA models with multivariate Laplace or Student innovations and compare the standard errors to those deduced from the theory.

Alj, A., Azrak, R., Ley, C. and Mélard, G. (2017). Asymptotic properties of QML estimators for VARMA models with time-dependent coefficients. *Scandinavian Journal of Statistics*, 44, 617–635.

Azrak, R., and Mélard, G. (2006). Asymptotic properties of quasi-likelihood estimators for ARMA models with time-dependent coefficients. *Statistical Inference for Stochastic Processes*, 9, 279–330.

Lehmann, E. L., and Casella, G. (1998). *Theory of Point Estimation*. Springer Verlag, New York.

---

## Building a dynamic risk prediction model for cardiovascular disease

Oct 18
16:20–17:00
ADN

Jessica Barrett[1*], Michael Sweeting[2], Ellie Paige[3], David Stevens[1] and Angela Wood[1]
[1]University of Cambridge; [2]University of Leicester; [3]Australian National University

The aim of a risk prediction model is to accurately predict the probability of some event occurring within a pre-specified time window for a new individual, i.e. an individual whose data does not contribute to the model fit. A dynamic risk prediction model allows risk predictions to be updated over time in response to new information becoming available. Possible methods for dynamic risk prediction include (i) using the last-observation-carried forward (LOCF) of each risk factor as a time-varying covariate in a time-to-event model, (ii) landmarking, where a discrete set of landmark times is specified at which risk predictions are to be made and survival is modelled from the landmark time only for individuals still at risk using either (a) the LOCF of each risk factor as a covariate or (b) the current value of each risk factor as estimated by longitudinal modelling of past risk factor measurements, and (iv) joint modelling, where repeated risk factor measurements and the time to event are modelled simultaneously. E.g. for cardiovascular disease (CVD), the 10-year risk of a CVD event is typically used to make clinical decisions about whether to prescribe lipid-lowering medication. Time-varying CVD risk factors, such as blood pressure, cholesterol and smoking status, may be monitored over time and used to dynamically update CVD risk predictions. We consider dynamic risk prediction for CVD in a number of different data scenarios, starting from a single cohort study of around 20,000 individuals to electronic health record data comprising around 2 million individuals.

---

## On standardising quality of care indicators based on summary statistics

Oct 18
17:10–17:30
ADN

Marion Louvel* and Els Goetghebeur
UGent

Twice a year, the Flemish government gathers measures of quality indicators from its care centers as part of the VIP-WZC (Flemish Indicator Project for Care Centers). Currently the VIP breaks its results down per center characteristics. However, it does not yet adjust for resident-level characteristics due to privacy protection restrictions. It is therefore possible that resident characteristics, such as their health status, interfere with the indicators.

The present study adapts a Firth-corrected model fitting approach to work from center averaged resident summary statistics, in order to model binary outcomes of care center residents whose data must be handled confidentially. The consequences of allowing standardised quality of care indicator to condition on resident characteristics are investigated, and the change in bias of the resulting standardised risk estimate is quantified. For the Firth-corrected regression

technique to be available in the context of the VIP, this study also assesses the feasibility of fitting this model based solely on summary statistics aggregated at the care center level. A set of summary statistics enabling the manual construction of the matrices required in the fitting algorithm is identified, and convergence rate of the fitting algorithm is analysed.

This allows us to confirm the worthiness of including the available resident information in the statistical model and the superiority of the Firth-corrected method with regards to both bias reduction and overcoming issues of separation in the data.

---

Oct 18
17:30–17:50
ADN

## Generalized pairwise comparison methods to analyze (non)-hierarchical composite endpoints

Johan Verbeeck
UHasselt

In the analysis of composite endpoints in a clinical trial, time to first event analysis techniques such as the logrank test and Cox proportional hazard test, do not take into account the multiplicity, importance and the severity of events in the composite endpoint. Several Generalized Pairwise Comparison (GPC) analysis methods have been described recently that do allow to take these aspects into account. These methods have the additional benefit that all types of outcomes can be included, such as longitudinal quantitative outcomes, to evaluate the full treatment effect. Four of the generalized pairwise comparison methods, the Finkelstein-Schoenfeld, adapted Buyse, unmatched Pocock and adapted O'Brien test are compared to each other and to the logrank test by means of simulations of the TAVR UNLOAD trial (NCT02661451), a cardiovascular trial that includes in the composite endpoint a quality of life measure. These simulations show that hierarchical generalized pairwise comparison methods perform very similarly and are better powered to detect a treatment difference only when non time-to-event outcomes are added to the composite endpoint.

Buyse, M. (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*, 29, 3245–3257.

Pocock, S. J., Ariti, C. A., Collier, T. J., and Wang, D. (2012). The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal*, 33, 176–182.

Ramchandani, R., Schoenfeld, D., and Finkelstein, D. M. (2016). Global rank tests for multiple, possibly censored, outcomes. *Biometrics*, 72, 926–935.

---

# Abstracts of talks: Oct 19

### Goodness-of-fit tests in proportional hazards models with random effects

Oct 19
9:00–9:40
ADN

Wenceslao González-Manteiga[1], María Dolores Martínez-Miranda[2*] and Ingrid Van Keilegom[3]
[1]University of Santiago de Compostela; [2]University of Granada; [3]KULeuven

In survival analysis the Cox hazard model is without any doubt the most used and well known semiparametric model to study the relationship between a survival time and a set of covariates. The popularity of this model can, among others, be explained by its easy interpretation and the fact that the nonparametric baseline hazard function cancels out in the likelihood, making the estimation of the parametric components as simple as in a purely parametric model. An appealing extension of the Cox model consists of adding random effects. This provides a powerful tool in a wide variety of applications, where the data have a natural clustered structure. The model can reflect then the fact that some of the regression parameters are cluster-dependent and they may be treated as random.

The assumption of linear covariate effects in the Cox model with random effects is quite strong in practice. Nevertheless, linearity is often assumed without any formal verification. This paper deals with testing the functional form of the covariate effects in a Cox model with random effects. The estimation of the model under the null (parametric covariate effect) and the alternative (nonparametric effect) is performed using full marginal likelihood. Under the alternative, the nonparametric effects are estimated using orthogonal expansions. The test statistic is the likelihood ratio statistic, and its distribution is approximated using bootstrap. The performance of the testing procedure is evaluated through simulations. The method is also applied on real survival data.

---

### Lorenz regression

Oct 19
9:40–10:00
ADN

Cédric Heuchenne[1] and Alexandre Jacquemain[2*]
[1]ULiège; [2]UCLouvain

We lay out a regression procedure based on the Lorenz curve, a tool famous for its use in economics to describe income inequalities. The methodology is semiparametric in the sense that a monotone link function is assumed between the dependent variable and a linear index involving explanatory variables. Yet, no further assumption is made on the exact nature of that link. In that sense, our framework runs in the same lines as the single-index model proposed by Ichimura (1993). In applications, we argue that our methodology selects the covariates weights which reproduce as much as possible the observed inequalities in the dependent variable. If all covariates are continuous, we show that the obtained estimator is a special case of the monotone

rank estimator proposed by Cavanagh and Sherman (1998). In presence of discrete covariates, a continuity correction is developed. We present a goodness-of-fit measure, which refers to the proportion of explained inequality. Since the maximization program appears hard to tackle with usual methods, we present a genetic algorithm to solve it. Finally, we assess the performance of our estimator compared to Ichimura's nonlinear least-squares through a series of Monte-Carlo simulations.

---

Oct 19
10:00–10:20
ADN

### Smooth time-dependent ROC curve for right censored survival data

Kassu Mehari Beyene
UCLouvain

The receiver operating characteristic curve (ROC) and its corresponding area under the curve (AUC) are the most commonly used methods to assess the predictive ability (or classification accuracy) of prognostic or diagnostic tools. Several approaches have been proposed to estimate the time-dependent ROC and AUC in case of time-to-event data with censoring. Except one recent approach, the existing methods provide ROC estimators that are not smooth where, by definition, the ROC curve is smooth. In this article we propose a new time-dependent ROC curve and AUC estimators based on the smoothed empirical distribution function. We compare the performance of our estimator with some existing methods using a simulation study. Furthermore, we illustrate its usefulness using a real data example. An R package, under development, makes the methodology easily applicable.

---

Oct 19
10:20–10:40
ADN

### Stable IPW estimation for longitudinal studies

Vahe Avagyan[1*] and Stijn Vansteelandt[1,2]
[1]UGent; [2]University and London School of Hygiene and Tropical Medicine

In this paper, we consider estimation of the average effect of time-varying dichotomous exposure on outcome using Inverse Probability Weighting (IPW) under the assumption that there is no unmeasured confounding of the exposure - outcome association at each time point. Despite the popularity of IPW for this problem, the performance of IPW estimation is often poor due to instability of the estimated weights. We develop an estimating equation-based strategy for the nuisance parameters indexing the weights at each time point, aimed at preventing highly volatile weights and ensuring the stability of IPW estimation. Extensive simulation studies and a real data analysis demonstrate adequate performance of the proposed approach compared with the traditional maximum likelihood estimator and the Covariate Balancing Propensity Score estimator.

Imai, K., and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B*, 76, 243–263.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110, 910–922.

---

## Power in high-dimensional testing problems

Oct 19
9:40–10:00
Apollinaire

Anders Bredahl Kock[1] and David Preinerstorfer[2*]
[1]University of Oxford; [2]ULB

Fan et al. (2015) recently introduced a remarkable method for increasing asymptotic power of tests in high-dimensional testing problems. If applicable to a given test, their power enhancement principle leads to an improved test that has the same asymptotic size, uniformly non-inferior asymptotic power, and is consistent against a strictly broader range of alternatives than the initially given test. We study under which conditions this method can be applied and show the following: In asymptotic regimes where the dimensionality of the parameter space is fixed as sample size increases, there often exist tests that cannot be further improved by the power enhancement principle. When the dimensionality can increase with sample size, however, there typically is a range of "slowly" diverging rates for which every test with asymptotic size smaller than one can be improved with the power enhancement principle. While the latter statement in general does not extend to all rates at which the dimensionality increases with sample size, we give sufficient conditions under which this is the case.

Fan, J., Liao, Y. and Yao, J. (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica*, 83, 1497–1541.

---

## High-dimensional doubly robust tests for regression parameters

Oct 19
10:00–10:20
Apollinaire

Vahe Avagyan, Oliver Dukes* and Stijn Vansteelandt
UGent

After variable selection, standard inferential procedures for regression parameters may not be *uniformly valid*; there is no finite sample size at which a standard test is guaranteed to attain its nominal size (within pre-specified error margins). This problem is exacerbated in high-dimensional settings, where variable selection becomes unavoidable. This has prompted a flurry of activity in developing uniformly valid hypothesis tests for a low-dimensional regression parameter (e.g. the causal effect of an exposure $A$ on an outcome $Y$) in high-dimensional models. So far there has been limited focus on model misspecification, although this is inevitable in high-dimensional settings. We propose tests of the null that are uniformly valid under sparsity conditions weaker than those typically invoked in the literature, assuming working models for the exposure and outcome are both correctly specified. When one of the models is misspecified, by amending the procedure for estimating the nuisance parameters, our tests continue to be valid; hence they are then *doubly robust*. Our proposals are straightforward to implement using existing software for penalized maximum likelihood estimation and do not require sample-splitting.

---

Oct 19
10:20–10:40
Apollinaire

## Testing for hidden periodicities in functional time series

Clément Cerovecki[1], Vaidotas Characiejus[1*], and Siegfried Hörmann[2]
[1]ULB; [2]Graz University of Technology

We propose several procedures to test for the presence of periodicities in functional time series when the length of the period is unknown. The tests are based on the asymptotic distribution of the maximum over all Fourier frequencies of the Hilbert-Schmidt norm of the periodogram operator of independent and identically distributed random elements with values in a real separable Hilbert space. Our approach is based on a projection onto a finite dimensional subspace spanned by a finite number of principal components. When the number of principal components is fixed, we show that the maximum converges in distribution to the standard Gumbel distribution as the sample size increases. Our result generalises the result of Davis and Mikosch (1999). Under stronger assumptions, we show that the same limit holds even if we let the number of principal components grow to infinity as the sample size increases. This allows us to establish the limit in distribution of the maximum over all Fourier frequencies of the Hilbert-Schmidt norm of the periodogram operator of independent and identically distributed random elements with values in a real separable Hilbert space. We use our asymptotic results to propose several tests for hidden periodicities in functional time series and illustrate their performance using a small simulation study.

Davis, R.A., and Mikosch, T. (1999). The maximum of the periodogram of a non-Gaussian sequence. *The Annals of Probability*, 27, 522–536.

---

Oct 19
11:00–11:40
ADN

## First aid after model selection

Gerda Claeskens
KULeuven

By selecting variables or models via information criteria or other formal methods a single selected model comes out as the winner. Often, this winning model is treated as if it was known all the way that precisely this model would get selected. There is, however, randomness involved with selection, with a different data sample another model could have been selected. For the popular method of the Akaike information criterion (AIC), the asymptotic distribution of parameter estimators after model selection is studied. We exploit the overselection property of this criterion in the construction of a selection region, and obtain the asymptotic distribution of parameter estimators and linear combinations thereof in the selected model. The proposed method does not require the true model to be in the model set. We investigate the method in linear and generalized linear models. Most part of this is joint work with A. Charkhi.

---

## On the causal effect of gender: beyond the (many!) anecdotes, a statistician's view on evidence of gender bias

Oct 19
11:40–12:20
ADN

Els Goetghebeur
UGent

The most important questions in society cannot be answered by data alone. Without reliable statistical evidence however, our understanding is doomed to stay with prejudice. When asking about causal effects rather than mere associations, the challenge grows. This happens as causal inference requires a framework beyond distributional assumptions on observed variables for interpretable data analysis. Today, a growing causal inference community has embraced a common formalism for this and several Nobel prizes honored major methodological advances in this area. Still, estimating causal effects of factors, such as gender and race, meets with controversy when there are no randomized experiments to emulate.

Here, we briefly review evidence presented by reliable sources, including a report of the American National Academies on Science and Engineering, and of the League of European Research Universities. We are concerned with hiring practices, advances through the academic ranks, peer review of grant applications and submitted publications, perceived research quality,... We also look at the Harvard hosted test for implicit bias, compulsory for academic staff in many academic institutions, also in Europe.

We go on to examine more closely the methods and conclusions of some well chosen blinded randomized experiments and observational studies. And then we ask our audience to draw their own conclusion, with suitable uncertainty bounds, on whether and how gender bias is driving academia here and now. For those of us who carry important responsibility in this realm – and who doesn't? — the next question is: what can or should be done?

---

## James-Stein estimators in factor analysis

Oct 19
13:30–13:50
ADN

Elissa Burghgraeve*, Jan De Neve and Yves Rosseel
UGent

In the social and behavioral sciences, latent variables are ubiquitous. These latent variables often represent hypothetical constructs (quality of life, motivation, ability) that can not be measured directly. Instead one measures indicators (i.e. observed variables) related to the latent variable of interest. For example, standard indicators of the quality of life include wealth, employment, physical and mental health, education,... The predominant statistical model to relate the latent variables to their indicators, is the factor analysis model. In this conventional method however, the relationship between an indicator and the latent variable is typically considered to be linear, an assumption that might not be fulfilled. We therefore present a semiparametric estimation procedure where we only assume a normal error term. This approach consists of using the James-Stein estimator, commonly known in measurement error theory. We apply this technique in our setting of estimating parameters in factor analysis models and extend it to deal with a possible nonlinear relation.

Whittemore, A.S. (1989). Errors-in-variables regression using Stein estimates. *The American Statistician*, 43, 226–228.

---

Oct 19
13:50–14:10
ADN

## Testing for principal component directions under weak identifiability

Davy Paindaveine, Julien Remy* and Thomas Verdebout
ULB

We consider the problem of testing, on the basis of a $p$-variate Gaussian random sample, the null hypothesis $\mathcal{H}_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$ against the alternative $\mathcal{H}_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_1^0$, where $\boldsymbol{\theta}_1$ is the "first" eigenvector of the underlying covariance matrix and $\boldsymbol{\theta}_1^0$ is a fixed unit $p$-vector. In the classical setup where eigenvalues $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_p$ are fixed, the Anderson (1963) likelihood ratio test (LRT) and the Hallin, Paindaveine and Verdebout (2010) Le Cam optimal test for this problem are asymptotically equivalent under the null, hence also under sequences of contiguous alternatives. We show that this equivalence does not survive asymptotic scenarios where $\lambda_{n1} - \lambda_{n2} = o(r_n)$ with $r_n = O(1/\sqrt{n})$. For such scenarios, the Le Cam optimal test still asymptotically meets the nominal level constraint, whereas the LRT severely overrejects the null hypothesis. Consequently, the former test should be favored over the latter one whenever the two largest sample eigenvalues are close to each other. By relying on the Le Cam theory of asymptotic experiments, we study the non-null and optimality properties of the Le Cam optimal test in the aforementioned asymptotic scenarios and show that the null robustness of this test is not obtained at the expense of power. Our asymptotic investigation is extensive in the sense that it allows $r_n$ to converge to zero at an arbitrary rate. While we restrict to single-spiked spectra of the form $\lambda_{n1} > \lambda_{n2} = \ldots = \lambda_{np}$ to make our results as striking as possible, we extend our results to the more general elliptical case. Finally, we present an illustrative real data example.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34, 122–148.

Hallin, M., Paindaveine, D. and Verdebout, T. (2010). Optimal rank-based testing for principal components. *Annals of Statistics*, 38, 3245–3299.

---

Oct 19
14:10–14:30
ADN

## Distribution comparison tests based on self-similar transport of measure

Gilles Mordant
UCLouvain

In statistics, and especially in nonparametric statistics, having natural order is crucial. It is however well known that ranks do not canonically exist in $\mathbb{R}^d$. In this talk, we show a new method relying on measure transportation onto one dimension using self-similar objects helping to easily perform multivariate distribution comparison tests. We show that our testing procedures are computationally effective and have nice connections with existing tests. In addition to finite sample performance and properties, we provide simulation results. Next, we consider an extension to the study of independence tests. This is based on a distance measure which involves Sklar's theorem and the general idea developed for comparison tests.

---

## On the impact of residential history in the spatial analysis of diseases with long latency period: a study of Mesothelioma in Belgium

Oct 19
13:30–13:50
Apollinaire

Christel Faes[1], Kristiaan Nackaerts[2], Benoit Nemery[2], Thomas Neyens[1] and Oana Petrof[1*]
[1]UHasselt; [2]KULeuven

Mesothelioma is a rare cancer caused by exposure to asbestos. Belgium has had a long history of asbestos production, resulting in one of the highest mesothelioma mortality rates worldwide. While the production of asbestos has stopped completely, the long latency period of mesothelioma, which can fluctuate between 20 to 40 years after exposure, makes that an increased number of mesothelioma patients is currently being observed. While interest is in the geographical distribution of the mesothelioma risk, this is disturbed by the latency period. Many people have changed several residential locations throughout their lifetime, and consequently, the location where the patients develop the disease is often far from the location where they were exposed. Using the residential history of patients, we propose the use of a convolution multiple membership model, which includes both a spatial conditional autoregressive and an un structur ed random effect. Pancreatic patients are used as control population, reflecting the population at risk for mesothelioma cancer. Results show the impact of the residential mobility on the geographical risk estimation, as well as the importance of acknowledging for the latency period of a disease.

Goldstein, H. (2011). *Multilevel Statistical Models*. Fourth Edition. John Wiley & Sons.

Neyens, T., Faes, C., and Molenberghs, G. (2012). A generalized Poisson-gamma model for spatially overdispersed data. *Spatial and spatio-temporal epidemiology*, 3, 185–194.

Van den Borre, L., and Deboosere, P. (2014). Asbestos in Belgium: an underestimated health risk. The evolution of mesothelioma mortality rates (1969-2009). *International Journal of Occupational and Environmental Health*, 20, 134–140.

---

## Detection of high dimensional intestinal microbiota as biomarker for immunological response: a Bayesian variable selection approach

Oct 19
13:50–14:10
Apollinaire

Olajumoke Evangelina Owokotomo[1*], Rudradev Sengupta[2], Luc Bijnens[2], Ziv Shkedy[1] and Adetayo Kasim[3]
[1]UHasselt; [2]Janssen Pharmaceuticals campanies of Johnson and Johnson; [3]Durham University

One of the most important risk factors for immune diseases and disorders is the alteration of the body microbiome. Which is due to antibiotics, genetic and environmental factors affecting the interactions between microbiome and the immune system of the host. Due to the population in the composition of the microbiome, it has been established and therefore important to use this bacterial to ascertain normal functioning of the human body.

Current methods for the detection of high dimensional biomarkers for clinical endpoints do not take into account the uncertainty about the endpoint and the model that is used to detect the biomarker. In this paper, we propose a new modelling approach for the detection of high dimensional biomarkers using Bayesian variable selection and Joint modelling techniques. We formulate a hierarchical joint path analysis model for the biomarker and the clinical outcome, taking into account possible treatment effects on both variables. The proposed path analysis

model allows us to estimate the posterior probability that the intestinal microbiota variable can be used as a biomarker for the clinical endpoint.

The proposed method is applied to the TransPAT study; it consists of 15 C57BL/6 germ free mice. The dataset contains information about the immune response (IgA) and the intestinal microbiome. The microbiome data contains information about the Operational Taxonomic Units (OTUs) and their abundance, information about the samples and also taxonomy information. In addition to model uncertainty, the hierarchical path analysis model allows to include random effects to account for possible within- subject variability. The results showed that the effect of the intestinal microbiota variable as a biomarker for IgA increases over time.

George, E. I., and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American statistical Association*, 88, 881–889.

Rivera-Amill, V. (2014). The human microbiome and the immune system: An ever evolving understanding. *Journal of Clinical & Cellular Immunology*, 5(e114).

Ruiz, V., Battaglia, T., Kurtz, Z.D., Bijnens, L., Ou, A., Engstrand, I., Zheng, X. H., Iizumi, T., Mullins, B. J., Muller, C. L., Cadwell, K., Bonneau, R., Perez-Perez, G. I., and Blaser, M. J. (2017). A single early-in-life macrolide course has lasting effects on murine microbial network topology and immunity. *Nature Communications*, 8, 1.

---

Oct 19
14:10–14:30
Apollinaire

## "cyanoFilter", an automated framework for identifying picocyanobacteria populations obtained via flow cytometry

Marc Aerts[1], Frederik De Laender[2], Thomas Neyens[1], Olusoji Oluwafemi[2*] and Jurg Spaak[2]
[1]UHasselt; [2]UNamur

Flow cytometry is a well-known technique for identifying cell populations. It is largely applied in biomedical and medical sciences for cell sorting, counting, biomarker detections and protein engineering. Cyanobacteria are a bacteria phylum, believed to contribute more than 50% of atmospheric oxygen via photosynthesis and are found almost everywhere.

While there are existing studies showing the applications of flow cytometry to identify picocyanobacteria populations present in a water sample, there has been no standard approach to identifying the different outcomes of these experiments. More often than not, scientists and ecologists resolve to rely on expert knowledge and opinion to identify cell population of interest, a process called manual gating. This makes reproducibility of population identification challenging if not an impossible task. To this effect, we propose a set of steps to follow to filter out desired cell populations and other possible outcomes from data obtained via flow cytometry.

Central to the identification of debris is a gating strategy based on identifying the number of peaks in a two-dimensional kernel density and estimation of a saddle point between these peaks, an idea already implemented in the flowDensity package. We also followed a similar technique in identifying margin events, using a one-dimensional kernel density. These steps are presently being developed into an R package (cyanoFilter), which will enable reproducibility of results as well as enable easy distribution and usage.

Malek, M., Taghiyar, M.J., Chong, L., Finak, G., Gottardo, R., and Brinkman, R. R. (2015). flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics*, 31, 606–-607.

## The genesis and use of time-varying frailty models for representing heterogeneities in the transmission of infectious diseases

Oct 19
14:35–15:15
ADN

Steffen Unkel[1*], Steven Abrams[2], Andreas Wienke[3] and Niel Hens[2,4]

[1]University Medical Center Göttingen; [2]Hasselt University; [3]Martin Luther University Halle-Wittenberg; [4]University of Antwerp

Frailty models are commonly used for representing and making inference on individual heterogeneities relevant to the transmission of infectious diseases, including heterogeneities that evolve over time. This talk takes as its focus time-varying frailty models for the infection-specific hazard rate, concentrating on the genesis of such models and how they can be derived from mechanisms underlying the endemic equilibrium equation for the infection of interest. We show that time-varying frailty models are a natural choice for capturing individual heterogeneities as they follow naturally from the underlying biology of the infections. Multivariate frailty models with shared or correlated frailties are particularly useful for inducing association of infection times within individuals as well as variability among individuals. For time-varying shared and correlated frailty models, we discuss issues of identifiability and methods of estimation. Illustrations with real data from serological surveys are provided.

# Presenter Index


Akça
    Elif, 19
Amico
    Maïlis, 19
Asenova
    Stefka, 20
Avagyan
    Vahe, 42

Barrett
    Jessica, 39
Beyene
    Kassu Mehari, 42
Burghgraeve
    Elissa, 45

Characiejus
    Vaidotas, 44
Claeskens
    Gerda, 44
Cutting
    Christine, 21

Daniel
    Rhian, 34
Dukes
    Oliver, 43

Frenda
    Antonio, 21
Fried
    Roland, 30

Garcia Angulo
    Andrea Cristina, 22
Ghaderinezhad
    Fatemeh, 35
Gistelinck
    Fien, 37
Goetghebeur
    Els, 45
Gressani
    Oswaldo, 36

Jacquemain
    Alexandre, 41
Jones
    Chris, 35

Kiriliouk
    Anna, 31
Koenker
    Roger, 31

Loh
    Wen Wei, 22, 33
Loncke
    Justine, 36
Louvel
    Marion, 39

Marion
    Rebecca, 23
Marquis
    Bastien, 30
Martínez-Miranda
    María Dolores, 41
Mathieu
    Sophie, 24
Melard
    Guy, 38
Mordant
    Gilles, 46

Nisol
    Gilles, 38

Nordhausen
Klaus, 29

Oluwafemi
Olusoji, 48
Owokotomo
Olajumoke Evangelina, 47

Parsa
Motahareh, 24
Petrof
Oana, 47
Pircalabelu
Eugen, 32
Preinerstorfer
David, 43

Rasoafaraniaina
Joséa Rondrotiana, 25
Remy
Julien, 46

Segers
Johan, 34
Shaikh
Md Abu Hanif, 25
Soetewey
Antoine, 26

Unkel
Steffen, 49

Van Lancker
Kelly, 29
Verbeeck
Johan, 40
Vranckx
Maren, 32

Zhou
Jing, 27