*Université catholique de Louvain – Institut de statistique, biostatistique et sciences actuarielles*

# Young Researchers' Day

## 18 September, 2015

| | | |
|---|---|---|
| $9^{00}$ | **Benjamin Colling** | Goodness-of-fit tests in semiparametric transformation models using the integrated regression function |
| $9^{30}$ | **Baptiste Féraud, Bernadette Govaerts and Manon Martin** | Metabolomics studies in disease diagnostic and prevention. Where and how can statisticians help? |

$10^{15}$ *Coffee Break*

| | | |
|---|---|---|
| *$10^{45}$* | **Mickaël De Backer** | *Copula Quantile Regression with Censored Data* |
| *$11^{15}$* | **Johan Segers, Michal Warchol and Yuwei Zhao** | *Modelling serial extremal dependence with tail processes* |

*At $12^{00}$ the seminar is followed by a sandwich lunch in the cafeteria.*
*Please do not forget to bring your Laptop with you.*

*The Young Researchers' Day is held in room c.115 of the Institut de statistique, biostatistique et sciences actuarielles, Voie du Roman Pays 20, Louvain-la-Neuve.*

## *Goodness-of-fit tests in semiparametric transformation models using the integrated regression function*

Benjamin Colling
*(Benjamin.Colling@uclouvain.be)*

*Consider a semiparametric transformation model of the form $\Lambda_\theta(Y) = m(X) + \epsilon$, where $Y$ is a univariate dependent variable, $X$ is a $d$-dimensional covariate, and $\varepsilon$ is independent of $X$ and has mean zero. We assume that $\{\Lambda_\theta : \theta \in \Theta\}$ is a parametric family of strictly increasing functions, while $m$ is the unknown regression function. We use a profile likelihood estimator for the parameter $\theta$ and a semiparametric local polynomial estimator for $m$. Our goal is to develop a new test for the parametric form of the regression function $m$ and to compare its performance to that developed by Colling and Van Keilegom (2015). The basic idea of the test developed by Colling and Van Keilegom (2015) was to compare the distribution function of $\varepsilon$ estimated in a semiparametric way to the distribution function of $\varepsilon$ estimated under the null hypothesis whereas here we compare the integrated regression function estimated in a semiparametric way to the integrated regression function estimated under the null hypothesis. We consider two different test statistics, a Kolmogorov-Smirnov and a Cramér-von Mises type statistic. We establish the limiting distributions of these two test statistics under the null hypothesis and under a local alternative. We use a bootstrap procedure to approximate the critical values of the different test statistics under the null hypothesis. Finally, a simulation study is carried out to illustrate the performance of our testing procedure, to compare this new test to the previous one and to see under which model conditions which test behaves the best.*

## References

*[1] Bierens, H.J. (1982), Consistent model specification tests.* Journal of Econometrics*, 20, 105–134.*

*[2] Colling, B. and Van Keilegom, I. (2015), Goodness-of-fit tests in semiparametric transformation models.* TEST *(accepted, in press).*

*[3] Escanciano, J.C. (2006), A consistent test for regression models using projections.* Econometric Theory*, 22, 1030–1051.*

*[4] Linton, O., Sperlich, S. and Van Keilegom, I. (2008), Estimation of a semiparametric transformation model.* Annals of Statistics*, 36, 686–718.*

*[5] Stute, W. (1997), Nonparametric model checks for regression.* Annals of Statistics*, 25, 613–641.*

*[6] Van Keilegom, I., González-Manteiga, W. and Sánchez Sellero, C. (2008), Goodness of fit tests in parametric regression based on the estimation of the error distribution.* TEST*, 17, 401–415.*

## *Metabolomics studies in disease diagnostic and prevention. Where and how can statisticians help?*

Baptiste Féraud, Bernadette Govaerts and Manon Martin
*(Baptiste.Féraud@uclouvain.be, Bernadette.Govaerts@uclouvain.be and Manon.Martin@uclouvain.be)*

*Metabolomics studies are increasingly used in a variety of health, pharmaceutical, quality control and food applications with the aim of better understanding the link between the metabolomic profile of a biofluid or tissue with given variables of interest that are indicative of diseases, treatments, exposition to toxins, etc. In this context, 1D and 2D NMR (Nuclear Magnetic Resonance) spectroscopy is widely used, along with Mass spectrometry, to produce complete snapshots (and corresponding data) of the metabolomic fingerprint of bio-samples. These very rich and high-dimensional data must then be pre-processed by finely tuned algorithms and analyzed with advanced multivariate statistical methods in order to extract useful information - or signal - from the raw data to answer questions under investigation. For instance, typical issues in medical projects are related to the research for disease biomarkers and the development of classification models that are able to predict a patient's medical state or condition.*

*More and more biostatisticians collaborate with spectroscopists, biologists and physicians during metabolomic projects and their role is increasingly crucial to promote innovative data pre-processing strategies and analysis tools or solutions. ISBA members are working in such context with medical and spectroscopist teams of the University of Liège and are involved in several medical projects, among which the research of potential biomarkers for an early detection of endometriosis.*

*The talk will follow the steps of a typical metabolomic study and present challenging projects currently conducted by ISBA members:*

- *SOAP-NMR, a R library developed to pre-process 1D $^{1}$H-NMR data and its statistical challenges involved, principally the crucial interest for the optimization of each pre-processing step from row data (FID) to exploitable data for further multivariate statistical analyses.*
- *The pre-processing steps for 2-dimensional COSY (COrrelation SpectroscopY) spectral data.*
- *The concept of Metabolomic Informative Content (MIC) involving different clustering-based measures able to quantify the quality of metabolomic data sets in terms of amount of captured signal (with applications and comparisons between 1D and 2D NMR data).*
- *The analysis of multi-factor metabolomic studies through ASCA and APCA methods for balanced and unbalanced designs.*
- *A review of some classical multivariate approaches to detect potential biomarkers and to build classification models from spectral matrices (PLS-DA, OPLS-DA, ICA).*

- *Some preliminary insights into the use of sparse methods (sparse-PLS and sparse-OPLS) and Random Forests in the context of biomarkers discovery (feature selection).*

## References


*[1] Worley, B. and Powers,R. (2013), Multivariate analysis in metabolomics.* Current Metabolomics*, 1(1): p. 92-107.*

*[2] Feraud B., Govaerts B., Verleysen M., de Tullio P. (2015), Statistical treatment of 2D NMR COSY spectra in metabolomics: data preparation, clustering-based evaluation of the Metabolomic Informative Content and comparison with $^{1}$H-NMR, Metabolomics, in press, DOI: 10.1007/s11306-015-0830-7.*

*[3] Rousseau, R., Feraud, B., Govaerts, B., Verleysen, M. (2013), Combination of Independent Component Analysis and statistical modelling for the identification of metabonomic biomarkers in 1H-NMR spectroscopy, in* Discussion Paper 2013/06*, ISBA - UCL.*

## *Copula Quantile Regression with Censored Data*

Mickaël De Backer
*(mickael.debacker@uclouvain.be)*

*Quantile regression is a common way to investigate the possible relationships between a d-dimensional covariate $\boldsymbol{X}$ and a response variable $T$. Since it was introduced by Koenker and Bassett (1978) as a robust (to outliers) and flexible (to error distribution) method, quantile regression has received notable interest in the literature of theoretical and applied statistics as a very attractive alternative to the classical mean regression method that captures only the central tendency of the data. In survival analysis, the quantile regression approach allows the analyst to estimate the functional dependence between variables for all portions of the conditional distribution of the (possibly) right-censored response variable. In that sense, quantile regression provides a more complete view of relationships between $T$ and $\boldsymbol{X}$ and constitutes an alternative to popular regression techniques like the Cox proportional hazards model or the accelerated failure time model.*

*In this talk, under the usual assumption of conditional independence between the survival time and the censoring time, we consider a new class of estimators that would allow practitioners to analyse in a flexible way medium to high-dimensional censored data using quantile regression. Actually, our methodology is an extension of the recent work of Noh et al. (2013) and Noh et al. (2015) to allow for the presence of censoring. Accordingly, in a similar spirit as for the case without censoring, the methodology makes use of the advantages of copulas in dependence modelling as the main idea consists of expressing the characterization of the quantile regression in terms of a multivariate copula and marginal distributions. Numerical examples will be used to illustrate the validity and performance of our procedure.*

## References

*[1] Noh, H., El Ghouch, A. and Bouezmarni, T. (2013), Copula-Based Regression Estimation and Inference.* Journal of the American Statistical Association, *108, 676–688.*

*[2] Noh, H., El Ghouch, A. and Van Keilegom, I. (2015), Semiparametric Conditional Quantile Estimation through Copula-Based Multivariate Models.* Journal of Business and Economic Statistics, *33(2), 167–178.*

## *Modelling serial extremal dependence with tail processes*

Johan Segers, Michal Warchol and Yuwei Zhao
*(Johan.Segers@uclouvain.be, Michal.Warchol@uclouvain.be and Yuwei.Zhao@uclouvain.be)*

*Classical time series techniques are ill-suited to model serial dependence of extremes. The framework of regular variation allows to model time series extremes for non-Gaussian data. The theory will be illustrated on linear time series models with heavy-tailed innovations. The final part of the lecture will involve a short quiz with the usage of a Shiny application and financial data. Participants are encouraged to bring a laptop with them.*

## References

*[1] Basrak, B. and Segers, J. (2009), Regularly varying multivariate time series.* Stochastic Processes and their applications*, 119, 1055–1080.*

*[2] Davis, R. and Mikosch, T. and Zhao, Y. (2013), Measures of serial extremal dependence and their estimation.* Stochastic Processes and their applications*, 123, 2575–2602.*

*[3] Drees, H. and Segers, J. and Warcholł, M. (2015), Statistics for tail processes of Markov chains.* Extremes*, 18, 369–402.*