Thank you for your participation to this annual doctoral workshop of the Graduate School in Statistics and Actuarial Sciences.
Programme

9h15: Welcome coffee

9h45: Presidential address, ROOM 01
10h - 11h: Invited Session I (Chair: Ph. Lambert, ULg), ROOM 01
  John Mac McDonald (University of London, England)
  Effects of Exposure Misspecification in Log-linear Models for Rates
11h - 12h20: Senior PhD Students invited session (Chair: R. Von Sachs, UCL), ROOM 01
  11h - 11h25:
  Alexandre Petkovic (Université Libre de Bruxelles)
  Modelling and pricing multi-asset derivatives
  11h25 - 11h50:
  Sophie Vanbelle (Université de Liège)
  Agreement between raters and groups of raters
  11h50 - 12h15:
  Céline Le Bailly de Tilleghem (Merck Sharp & Dohme and UCL)
  Taking into account prediction error in the virtual optimization of molecules using QSAR models

12h30 - 14h: Lunch

14h - 15h20: Parallel contributed sessions

Contributed session ACTU (Chair: G. Deelstra, ULB), ROOM S.33
  14h - 14h25:
  Xavier De Scheemaekere (Université Libre de Bruxelles)
  Dynamic risk indifference pricing and backward stochastic differential equations
  14h25 - 14h50:
  Gregory Rayée (Université Libre de Bruxelles)
  Vanna-Volga methods applied to FX derivatives: from theory to market practice
  14h50 - 15h15:
  Julien Hunt (Université Catholique de Louvain)
  Discrete-time semi-Markov regime switching interest rate models

Contributed session STAT (Chair: C. Heuchenne, ULg), ROOM S.42
  14h - 14h25:
  Hiroyuki Taniai (Université Libre de Bruxelles)
  Residual Empirical Process of Exponential ARCH via Quantile Regression
  14h25 - 14h50:
  Christel Ruwet (Université de Liège)
  Influence function of the error rate of classification based on clustering
  14h50 - 15h15:
  Maïk Schwarz (Université Catholique de Louvain)
  Consistent density deconvolution under partially known error distribution
Contributed session BIOSTAT (Chair: B. Govaerts, UCL), ROOM S36
14h - 14h25:
Jean-Marc Freyermuth (Université Catholique de Louvain)
Tree-structured Wavelet Estimation in a Mixed Effects Model for Spectra of Replicated Time Series
14h25 - 14h50:
Aysun Cetinyurek (Université de Liège)
Bayesian Density Estimation for Interval Censored Data with an Extension to COX Proportional Hazards Model
14h50 - 15h15:
Bernard Francq (Université Catholique de Louvain)
Development of statistical tools to test the equivalence between two measurement methods

15h20 - 15h45: Coffee break

15h45 - 16h45: Invited Session II (Chair: G. Haesbroeck, ULg), ROOM 01
Chris Jones (Open University, England)
Distributionology
**Çetinyurek Aysun**  
*Université de Liège*  

*Bayesian density estimation for interval censored data with an extension to Cox proportional hazards model*

There has been increasing interest in statistical analysis of interval-censored time-to-event data. This type of data is quite usual for clinical trials or longitudinal studies especially in practical settings of AIDS and cancer research where the individuals have pre-scheduled visits but the event of interest occurs between the visits. Then, the event times are not known exactly but rather to lie in an interval of time.

Interval-censored data is a natural generalization of right censored time-to-event data. For right censored time-to-event data, an extensive number of statistical techniques are available to tackle most research questions under a variety of assumptions. However, for interval-censored time-to-event data less well developed procedures are available.

We first show how a smooth estimate of density of the time-to-event variable can be obtained from interval-censored data. It combines Bayesian P-splines with a generalization of the composite link model (Lambert and Eilers, 2009). It is extended to regression setting under the hypothesis of Cox proportional hazards (PH) model which is also a generalization of the strategy advocated by Lambert and Eilers (2005) for right censored data.

We illustrate the methodology on a well known real data set from the literature.

*This is joint work with Philippe Lambert, University of Liège.*

**References**


**De Scheemaekere Xavier**  
*Université Libre de Bruxelles*  

*Dynamic risk indifference pricing and backward stochastic differential equations*

In this work, we investigate a dynamic pricing formula in incomplete markets based on the risk indifference principle: we replace the criterion of maximizing utility by minimizing risk exposure because the latter is more often used in practice and because it is a natural extension to the idea of pricing and hedging in complete markets.
Using a dual characterization of dynamic risk measures coming from BMO martingales (see [3]), the risk indifference pricing problem reduces to two (zero-sum) stochastic differential games, which we solve by means of backward stochastic differential equation (BSDE) theory. We find an explicit formula for the dynamic risk indifference price in terms of solutions of BSDEs. We follow the spirit of [4] who study a similar risk indifference pricing problem using PDE techniques. Importantly, our stochastic analysis approach does not impose Markovian assumptions on the coefficients and it encompasses the case of dynamic time consistent risk measures (as, e.g., risk measures coming from g-expectations; see [2]).

The work will also include the extension to the jump diffusion case as well as a Malliavin calculus derivation of a (quasi) hedging formula, as in [1].

References

Francq Bernard
Institut de Statistique, Université Catholique de Louvain

Development of statistical tools to test the equivalence between two measurement methods

The needs of the industries to quickly assess the quality of products and the performance of the manufacturing methods leads to the development and improvement of alternative analytical methods sometimes faster, easier to handle, less expensive or even more accurate than the method corresponding reference. These so-called alternative methods should ideally lead to results comparable to those obtained by a standard method known as a reference.

To test statistically the equivalence between two measurement methods, a certain characteristic of a sample can be measured by the two methods in the experimental domain of interest. Pairs of points (Xi, Yi), representing the measures taken by the reference method and the alternative one can be modelled by a linear regression (a straight line). The estimated parameters are very useful to test the equivalence. Indeed, an intercept significantly different from zero indicates a systematic analytical bias between the two methods of measurement and a slope significantly different from one indicates a proportional bias. The estimated parameters and their confidence intervals are then used to test the equivalence of the two measurement methods. To achieve this correctly, it is essential to take into account the errors in both axes and heteroscedasticity in both axes. Different types of regression exist to handle these cases and a lot of confusion still exists in the literature. We review therefore the equations for estimating the regression by these different techniques and standardize the notations. Then, well focus on the most suitable equivalence techniques tests and we propose...
a new methodology to test the proportional bias and the analytical bias at the same time with a simultaneous confidence interval.

This is joint work with Bernadette Govaerts (Université Catholique de Louvain).

Freyermuth Jean-Marc
Institut de Statistique, Université Catholique de Louvain

Tree-structured Wavelet Estimation in a Mixed Effects Model for Spectra of Replicated Time Series

Wavelet thresholding has been proven to be a powerful technique for nonparametric function estimation when the underlying curve shows inhomogeneous spatial structure. However, in situations of low signal-to-noise ratio, it is possible to improve this denoising technique by imposing a tree-structure on the thresholded wavelet coefficients (Baraniuk, 1999). On the theoretical side, it is known from approximation theory (Cohen et al., 2001) and from studying the mean-squared error for functions in certain smoothness classes (Autin, 2008) that tree-structured wavelet thresholding has some powerful near-optimality properties. In Freyermuth et al. (2009), we apply a simple instance of tree-structured wavelet thresholding in combination with recursive dyadic partitioning to functional mixed effects modelling of log-spectra of stationary time series traces recorded from experimental designs.

Despite its simplicity, our method succeed in using pooled information carried over from all subjects to reach the following objectives:
• improve the denoising of the population spectrum,
• construct a model-based estimator of the variability between subjects,
• predict the subject specific spectra.

This is joint work with Hernando Ombao, Center for Statistical Sciences, University of Brown, USA and Rainer von Sachs, Université Catholique de Louvain.

References
Hunt Julien  
Institut de statistique, Université Catholique de Louvain  
*Discrete-time semi-Markov regime switching interest rate models*

In recent years, there has been a lot of effort put into finding realistic models of the term structure of interest rates. Several such models have been developed in a discrete-time binomial framework namely the Ho and Lee model, the Black-Derman-Toy model and many more. In our talk, we will present a discrete-time "binomial" regime switching model of the term structure. The idea is to model the state of the economy or the so-called regime by an underlying process influencing the term structure. Our choice for the underlying process is a discrete-time semi-Markov process of which Markov processes are a special case. The advantage of semi-Markov processes is to allow for more general duration distributions. Of course, this process acts as an extra source of uncertainty in our model. We will see how this influences the notion of arbitrage and its relation to equivalent martingale measures. We will also discuss in a special case what influence the introduction of the underlying process has on the possible values for the parameters and what influence it has on our "binomial" structure. Finally, time allowing, we will conclude with some ideas about possible future research.

*This is joint work with Pierre Devolder.*

Jones Chris  
The Open University, England  
*Distributionology*

The silly title is my shorthand way of saying that the talk will be about the study of parametric families of distributions and the fitting of them to data. To keep things relatively simple, I will concentrate on four-parameter distributions for univariate continuous data taking values on the whole real line. The four parameters consist of the usual location and scale parameters plus two shape parameters which between them allow for skewness and variation in tailweight. I shall probably have most to say about two particular attractive and potentially useful families of such distributions. One will be one of my discoveries, the "sinh-arcsinh distribution"; the other will be "two-piece distributions", of much longer history, with some emphasis on the skew t case. In addition to looking at various properties of these distributions, I hope to address some simple aspects of statistical inference via maximum likelihood for such models. In particular, tests of normality (and symmetry) based on sinh-arcsinh distributions are powerful quite generally in both heavy- and light-tailed situations. Two-piece distributions, on the other hand, can be made to have helpful properties in terms of parameter orthogonality.

Le Bailly de Tilleghem Céline  
Institut de statistique, Université Catholique de Louvain  
*Taking into account prediction error in the virtual optimization of molecules using QSAR models*

The motivation problem of the presentation is the virtual optimization of molecules using QSAR (Quantitative Structure-Activity Relationship) models. In this context, QSAR models
are used to predict druggability properties (ADME, toxicity,…) as a function of chemical and physical descriptors of molecules of interest. This presentation proposes an integrated multicriteria optimization methodology to select a subset of equivalent optimal compounds in a collection of nonsynthesized molecules.

First the OECD principles on how to efficiently collect data and fit QSAR models having good predictive power are reviewed. Those QSAR models can be used to predict the optimized properties for new compounds.

Then a desirability index is proposed to combine the predicted properties into a value between zero and one. Such index is often used in the context of experimental design to solve multiresponse optimization problem. It provides a ranking of possible compounds in the library and the optimal one can be found by an adequate optimization algorithm. But, as model predictions are soiled with error, so is the desirability index and the optimal molecule found. In practice, in the related literature of design of experiments, this error is neglected.

To take into account the propagation of the models error on the estimated or predicted desirability index, confidence and prediction intervals are constructed. The stochastic character of the index leads also to an uncertainty on the optimum and a methodology is proposed to build an equivalence zone containing no significantly different optimal molecules.

This methodology is illustrated on the optimization of a virtual combinatorial library in the context of the development of a molecule in the pharmaceutical industry.

This is a joint work with and B. Govaerts, UCL, and B. Beck and B. Boulanger, Eli Lilly & Company, Belgium.

References
- Govaerts, B. and Le Bailly de Tilleghem, C. (2005), Uncertainty Propagation in Multiresponse Optimization using a Desirability Index, DP0532, Institut de Statistique, Université catholique de Louvain, Louvain-la-Neuve, Belgium

McDonald John W. "Mae"
Centre for Longitudinal Studies, Institute of Education, University of London
Effects of Exposure Misspecification in Log-linear Models for Rates

Log-linear models are used in demography, actuarial science and epidemiology to model rates cross-classified by explanatory factors. A rate is defined as the ratio of the number of events of interest to the exposure. For example, for mortality rates, the exposure is total person-years
at risk. The rates are not modelled directly, but a table of counts of events of interest and a table of exposures are required.

Often, the true exposure is unknown and an estimate of exposure is used. The problem we study is the effects of misspecification of the exposures on inferences drawn from a log-linear model for rates. Our literature review suggests that the effects of this type of misspecification have not been investigated. In particular, we investigate how this type of rates model misspecification affects parameter estimates, estimated standard errors, confidence intervals, test statistics, etc.

Petkovic Alexandre
ECARES, Université Libre de Bruxelles
*Modelling and pricing multi-asset derivatives*

This presentation will be composed of two parts.

First, we will discuss the pricing and hedging of arithmetic Asian basket spread options of the European type within the standard Black and Scholes framework. More specifically we will discuss approximations obtained through moment matching and comonotonic approximations.

Then, we will discuss the construction of multivariate Levy processes and multivariate stochastic volatility Levy processes. We will see how such processes could be used to price Asian basket spread options.

Rayée Grégory
Université Libre de Bruxelles
*Vanna-Volga methods applied to FX derivatives: from theory to market practice*

We study Vanna-Volga methods which are used to price first generation exotic options in the Foreign Exchange market. They are based on a rescaling of the correction to the Black-Scholes price through the so-called 'probability of survival' and the 'expected first exit time'. Since the method relies heavily on the appropriate treatment of market data we also provide a summary of the relevant conventions. We offer a justification of the core technique for the case of vanilla options and show how to adapt it to the pricing of exotic options. Our results are compared to a large collection of indicative market prices and to more advanced models. Finally we propose a simple calibration method based on one-touch prices that allows the Vanna-Volga results to be in line with our pool of market data.

*This is joint work with F. Bossens, N. S. Skantzoz and G. Deelstra.*

References
Ruwet Christel  
University of Liège  

Influence function of the error rate of classification based on clustering

Cluster analysis may be performed when one wishes to group similar objects into a given number of clusters. Several algorithms are available in order to construct these clusters. In this talk, focus will be on two particular cases of the generalized k-means algorithm: the classical k-means procedure as well as the k-medoids algorithm, while the data of interest are assumed to come from an underlying population consisting of a mixture of two groups. Among the outputs of these clustering techniques, a classification rule is provided in order to classify the objects into one of the clusters. When classification is the main objective of the statistical analysis, performance is often measured by means of an error rate.

In case there are some outliers in the data, the classification rule may be corrupted. The error rate may then be contaminated. To measure the robustness of classification based on clustering, influence functions have been computed. Similar results as those derived by Croux et al (2008) and Croux et al (2008) in discriminant analysis were observed. More specifically, under optimality (which happens when the model distribution is \(F_N = 0.5 \text{N}(\mu_1, \sigma) + 0.5 \text{N}(\mu_2, \sigma)\), Qiu and Tamhane 2007), the contaminated error rate can never be smaller than the optimal value, resulting in a first order influence function identically equal to 0. Second order influence functions need then to be computed. When the optimality does not hold, the first order influence function of the theoretical error rate does not vanish anymore and shows that contamination may improve the error rate achieved under the non-optimal model.

The first and, when required, second order influence functions of the error rate are useful in their own right to compare the robustness of the 2-means and 2-medoids classification procedures. They have also other applications. For example, they may be used to derive diagnostic tools in order to detect observations having an unduly large influence on the error rate. Also, under optimality, the second order influence function of the theoretical error rate can yield asymptotic relative classification efficiencies.

This is joint work with Gentiane Haesbroeck, University of Liège.

References:
- Croux C., Filzmoser P. and JoosSENS K. (2008), Classification efficiencies for robust linear discriminant analysis, Statistica Sinica 18, pp. 581-599.
Schwarz Maik  
Université catholique de Louvain  
*Consistent density deconvolution under partially known error distribution*

We are interested in estimating the density $f^X$ of a real-valued random variable $X$ based on an i.i.d. sample from $Y = X + \varepsilon$, where $\varepsilon$ is an independent additive error. In the literature, the density of the noise is usually supposed to be fully known. In contrast to this, we assume that $\varepsilon$ is normally distributed, but with unknown variance $\sigma^2 > 0$. First, we show that $(f^X, \sigma^2)$ can be identified from the observations when $f^X$ vanishes on a set of positive Lebesgue measure. As opposed to standard procedures, this identification condition is not based on properties of the densities’ characteristic functions.

Deconvolving a density is well-known not to be continuous with respect to the $L^2$-norms, that is why it is called an ill-posed inverse problem. However, we show that deconvolution becomes continuous if we choose the topology of weak convergence for the deconvolution density and an appropriate topology for the observations’ density. As a consequence, a minimum distance estimator of $f^X$ will be weakly consistent only imposing the identification condition, that is, without further conditions on the densities’ characteristic functions. The result remains true even if we do not require the involved probability distributions to have densities.

*This is joint work with Sébastien Van Bellegem, Toulouse School of Economics, France.*

**References**
- Schwarz and Van Bellegem (2009) Consistent density deconvolution under partially known error distribution, Discussion paper.

Tanai Hiroyuki  
ECARES, Université Libre de Bruxelles  
*Residual Empirical Process of Exponential ARCH via Quantile Regression*

This talk discusses a tool to study the Residual Empirical Process (REP) of several ARCH models, including Exponential ARCH, in a comprehensive way. Its essential part will be derived from the asymptotic behavior of Quantile Regression (QR) estimators. As the QR estimators, together with the explanatory variables, represent the conditional sample quantiles, their asymptotic representation in probability provides a behavior of the conditional sample quantiles around the true conditional quantile. Also, the parameter of the QR model, QR coefficients, say, has some explicit relation with each ARCH models of our concern. Then, after rewriting the Regression Quantile Process (Reg.QP) in terms of the ARCH parameter, its unconditional behavior will be studied by a sort of the tower property of conditional expectations. The transformation of the Residual Quantile Process (Resid.QP) into the REP is done by the functional delta method. By this approach, we may clarify why the REP of ARCH models, unlike that of ARMA models, contains estimator-depending terms, since QR models themselves are defined identically for the AR model and the pure ARCH model.
Vanbelle Sophie  
Medical Informatics and Biostatistics, University of Liège  

Agreement between raters and groups of raters

Agreement between raters on a categorical scale is a frequently encountered problem in practice. We first review the kappa-like family of agreement indexes between two raters. An interpretation of the linear weighting scheme when using the weighted kappa coefficient is proposed (Vanbelle and Albert, 2009a). We then extend the theory to the case of an isolated rater and a group of raters (Vanbelle and Albert, 2009b) and to the case of two groups of raters (Vanbelle and Albert, 2009c). Finally, we turn to the problem of comparing several agreement indexes and present an extension of a bootstrap method permitting the comparison of two dependent agreement indexes to several agreement indexes (Vanbelle and Albert, 2009d). The research was motivated by the script concordance test (SCT) used in medical education to evaluate the clinical reasoning of students when confronted to a group of experts (Vanbelle et al., 2007; Collard et al., 2009).

References