



2009 IAP Short Course & Workshop

Modeling Association and Dependence in Complex Data

November 17-20, 2009

Katholieke Universiteit Leuven



Local organizers:

- **E. Lesaffre** (Katholieke Universiteit Leuven & Erasmus Universiteit Rotterdam)
- **G. Molenberghs** (Katholieke Universiteit Leuven & Universiteit Hasselt)
- **G. Verbeke** (Katholieke Universiteit Leuven)

Program

Tuesday & Wednesday, November 17–18
Short course: “Non-parametric Bayesian statistics” T. Hanson & A. Jara Vallejos

Thursday, November 19		Friday, November 20	
8.30 – 9.00	Registration		
9.00 – 9.15	Welcome & Network introduction I. Van Keilegom	9.00 – 10.15	Keynote session: G. Claeskens (part II)
9.15 – 10.15	Keynote session: G. Claeskens (part I)		
10.15 – 11.25	WP2 contributions (4 papers) Chair: I. Van Keilegom	10.15 – 11.25	WP3 contributions (4 papers) Chair: G. Molenberghs
11.25 – 11.50	Coffee	11.25 – 11.50	Coffee
11.50 – 13.15	WP1 contributions (5 papers) Chair: I. Gijbels	11.50 – 13.15	WP4 contributions (5 papers) Chair: G. Verbeke
13.15 – 14.00	Lunch	13.15 – 14.00	Lunch
14.00 – 15.40	WP5 contributions (6 papers) Chair: L. Duchateau	14.00 – 15.40	WP6 contributions (6 papers) Chair: E. Lesaffre
15.40 – 16.00	Coffee	15.40 – 16.00	Coffee
16.00 – 17.15	Keynote session: T. Hanson (part I)	16.00 – 17.00	Keynote session: T. Hanson (part II)
17.30 – 19.00	Administrative meeting with BELSPO (closed meeting)	17.00 – . . .	Closing reception

List of presenters

Pre-conference short course	1
<i>T. Hanson (University of Minnesota, U.S.A.)</i>	
<i>A. Jara Vallejos (Universidad de Concepción, Chile)</i>	1
Keynote speakers	2
<i>G. Claeskens (Katholieke Universiteit Leuven, Belgium)</i>	2
<i>T.E. Hanson (University of Minnesota, U.S.A.)</i>	3
WP1: Multivariate data with qualitative constraints	4
<i>I. Van Keilegom (Université Catholique de Louvain, Belgium)</i>	4
<i>F. Abegaz (University of Pune, India)</i>	5
<i>A. El Ghouch (Université Catholique de Louvain, Belgium)</i>	6
<i>P. Eilers (Erasmus Medical Center, Rotterdam, The Netherlands)</i>	7
<i>I. Gijbels (Katholieke Universiteit Leuven, Belgium)</i>	8
WP2: Temporally and spatially related data	9
<i>L. Slaets (Katholieke Universiteit Leuven, Belgium)</i>	9
<i>M. Bekaert (Ghent University)</i>	10
<i>T. Meinguet (Université Catholique de Louvain, Belgium)</i>	11
<i>T. Lodewyckx (Katholieke Universiteit Leuven, Belgium)</i>	13
WP3: Incomplete data	14
<i>A. Gaddah (Universiteit Hasselt, Belgium)</i>	14
<i>A. Sujica (Université Catholique de Louvain, Belgium)</i>	15
<i>R.M. Daniel (London School of Hygiene and Tropical Medicine, United Kingdom)</i>	16
<i>C. Sotto (Universiteit Hasselt, Belgium)</i>	17
WP4: Data with latent heterogeneity	18
<i>T.F. Wilderjans (Katholieke Universiteit Leuven, Belgium)</i>	18
<i>R. Van Oirbeek (Katholieke Universiteit Leuven, Belgium)</i>	19
<i>E. Vande Gaer (Katholieke Universiteit Leuven, Belgium)</i>	20
<i>K. De Roover (Katholieke Universiteit Leuven, Belgium)</i>	21
<i>S. Frederickx (Katholieke Universiteit Leuven, Belgium)</i>	22

WP5: High-dimensional and compound data	23
<i>J. De Neve (Ghent University, Belgium)</i>	23
<i>Q. Zhu (Universiteit Hasselt, Belgium)</i>	24
<i>R. van den Berg (Katholieke Universiteit Leuven, Belgium)</i>	25
<i>I. Van Mechelen (Katholieke Universiteit Leuven, Belgium)</i>	26
<i>M. Schouteden (Katholieke Universiteit Leuven, Belgium)</i>	27
<i>K. Van Deun (Katholieke Universiteit Leuven, Belgium)</i>	28
WP6: Miscellaneous topics	29
<i>K. Tharmaratnam (Katholieke Universiteit Leuven, Belgium)</i>	29
<i>M. Babanezhad (Golestan University, Iran)</i>	30
<i>L. De Lobel (Ghent University, Belgium)</i>	31
<i>B. Van Rompaye (Ghent University, Belgium)</i>	32
<i>D. Vanpaemel (Katholieke Universiteit Leuven, Belgium)</i>	33

Pre-conference short course

Non-parametric Bayesian statistics

T. Hanson

University of Minnesota, U.S.A.

A. Jara Vallejos

Universidad de Concepción, Chile

Bayesian nonparametric statistics is a relative new area of statistics. The intersection between Bayesian and non-parametric statistics was almost empty until the late sixties where the first advances were made, primarily on the mathematical formulations. It was only in the early nineties with the advent of sampling based methods, in particular Markov chain Monte Carlo methods, that substantial progress has been made in the area. Posterior distributions ranging over functional spaces are highly complex and hence sampling methods play a key role.

This course is designed to provide basic coverage of Bayesian nonparametrics and includes advanced use of the R package DPpackage and BUGS. The course will include theoretical input, but also practical elements and participants will be involved hands-on in the use of DPpackage and BUGS. Emphasis on the course is placed on the probability models for probability distributions and their practical applications for model building. Some knowledge on Bayesian computation is assumed.

Keynote speakers

Tailor-made model selection Goodness-of-fit tests in linear mixed models

G. Claeskens

Katholieke Universiteit Leuven, Belgium

We will discuss some specific model selection and model averaging methods. In particular we will explain the concept of focussed model selection. Sometimes a model is good for estimating a certain parameter, but not for another one. The focussed information criterion is tailored to select the best model for estimating your focus. We explain how to construct and use it in various situations. Model averaging amounts to estimation via a number of possible models and forming a weighted average of the resulting estimators. One example of model averaged estimators are estimators in a selected model. Model averaging clearly shows that ignoring model selection in any further inference in the selected models can lead to dramatic wrong conclusions.

Linear mixed models with both fixed and random effects are most often estimated on the assumption that the random effects have a normal distribution. The goodness-of-fit tests that we study are used to formally test the hypothesis that the random effects and/or the errors are normally distributed. The proposed tests are nonparametric and based on the order selection idea. They are designed to detect virtually any alternative to normality. In case of rejection of the null hypothesis, the nonparametric estimation method that is used in the construction of the test provides an estimator of the alternative distribution.

Bayesian survival analysis: An overview of models and methods

T.E. Hanson

University of Minnesota, U.S.A.

In this talk I will provide an overview of approaches for analyzing time-to-event data using semiparametric and nonparametric models. Popular models including proportional hazards, accelerated failure time, proportional odds, additive hazards, and proportional mean residual life will be discussed, along with common parametric and nonparametric priors on the baseline distribution. Nonparametric priors include the gamma process, beta process, Dirichlet process mixtures, Polya trees, penalized splines and extensions of these. Then I will discuss various model generalizations including time dependent covariates, joint longitudinal and survival modeling, various frailty structures, cure models, and completely nonparametric dependent process approaches. Various models will be fit and illustrated in several software packages and languages including WinBUGS, BayesX, DPpackage, and FORTRAN.

WP1: Multivariate data with qualitative constraints

Univariate frontier estimation in the presence of measurement error

I. Van Keilegom

Université Catholique de Louvain, Belgium

Consider the model $Y = X \cdot Z$, where Y is observable, X is the variable of interest with support $[0, \tau]$ and density f_X and Z is the noise. Suppose that $f_X(\tau) > 0$, and that Z is independent of X and is log-normally distributed with unknown variance σ^2 . This model differs from most models considered in the literature in the sense that the variance σ^2 is here unknown. We show that the model is identifiable, and propose estimators for τ and σ^2 . The asymptotic consistency and the rate of convergence of the estimators is established, and simulations are carried out to verify the performance of the estimators for small samples.

Semiparametric score test for varying Copula parameter in Markov time series

F. Abegaz

University of Pune, India

This talk examines a semiparametric test for checking the constancy of serial dependence via copula models for Markov time series. A semiparametric score test is proposed for testing the constancy of the copula parameter against stochastically varying copula parameter. The asymptotic null distribution of the test is established. A semiparametric bootstrap procedure is employed for the estimation of the variance of the proposed score test. Illustrations are given based on simulated series and historic interest rate data.

Local polynomial quantile regression with parametric features

A. El Ghouch

Université Catholique de Louvain, Belgium

We propose a new approach to conditional quantile function estimation that combines both parametric and nonparametric techniques. At each design point, a global, possibly incorrect, pilot parametric model is locally adjusted through a kernel smoothing fit. The resulting quantile regression estimator behaves like a parametric one when the latter is correct and converges to the nonparametric solution as the parametric start deviates from the true underlying model. We give a Bahadur-type representation of the proposed estimator from which consistency and asymptotic normality are derived under α -mixing assumption. We also propose a practical bandwidth selector based on the plug-in principle and discuss the numerical implementation of the new estimator. Finally, we investigate the performance of the proposed method via simulations and we illustrate the methodology on a data example.

Shape constraints in curve resolution

P. Eilers

Erasmus Medical Center, Rotterdam, The Netherlands

Many (chemical) data sets can be modeled as sums of unimodal positive curves or surfaces. A powerful algorithm for non-parametric smooth unimodal curve modeling, is an essential building block. In this contribution I describe such an algorithm. The key idea is to model the logarithm of a curve, using a penalty that encourages unimodality. In addition an asymmetric penalty can be added to enforce the desired shape, if needed. Real data sets will be used to illustrate applications in one and two dimensions.

Nonparametric estimation of Copulas and conditional copulas

I. Gijbels

Katholieke Universiteit Leuven, Belgium

One way to model a dependence structure between random variables is through the copula function. Association measures such as Kendall's tau or Spearman's rho can be expressed as functionals of the copula. The dependence structure between two variables can be highly influenced by a covariate, and it is of interest to know how this dependence structure changes with the value taken by the covariate. This leads to considering conditional copulas and conditional association measures. In this talk an overview will be given of recently developed nonparametric estimators of (un)conditional copulas and association measures.

WP2: Temporally and spatially related data

A multiresolution approach to time warping achieved by Bayesian prior-posterior transferring

L. Slaets

Katholieke Universiteit Leuven, Belgium

Functional data analysis refers to the statistical methodology designed for data sets involving functions. For many reasons, the rate at which data points are observed over time varies and does not necessarily reflect the underlying biological, physical, or other process governing the data. This is what we refer to as the phase variability in a sample of curves. The procedure known as warping aims at reducing this phase variability, by applying a smooth bijection, the warping function, to the argument of each of the functions. It is important to recognize the identifiability problem between phase and amplitude variability. We propose a natural representation of warping functions in terms of a new type of elementary function named warping component functions, localized in location and scale, which are combined into the warping function by composition. For the estimation a special sequential Bayesian estimation strategy is introduced. The corresponding C++ code makes the MCMC computations sufficiently fast to estimate a semi-parametric amplitude model. The estimated warping component and amplitude coefficients constitute a meaningful dimension reduction for the original curve sample.

References:

Ramsay, J.O. and Silverman, B.W. (2006) Functional Data Analysis. Springer, New York.

Adjusting for time-varying confounding in the subdistribution analysis of a competing risk

M. Bekaert

Ghent University

The intensive care unit (ICU) is a data-rich environment, where information technology may enhance patient care and scientific research by improving access to clinical data. Such automatic surveillance systems are important tools for the construction of high-quality data sets which have a rich collection of day-by-day information about a patient's health status. The analysis of such high dimensional data is statistically challenging, as we will motivate by considering the problem of how to quantify the causal effect of nosocomial infections (i.e., hospital-acquired) on ICU mortality.

Nosocomial infections are highly prevalent in intensive care unit (ICU) patients because of their poor health conditions and, additionally, because of the prevalent use of invasive treatments (e.g., mechanical ventilation) that make it more difficult for the human body to conquer hostile bacteria. They form a major public health problem in the Western world. Despite decades of research in the medical literature, assessment of the attributable mortality due to nosocomial infections in ICU remains controversial, with several studies describing effect estimates ranging from being neutral to extremely risk increasing. Interpretation of study results is further hindered by inappropriate adjustment (a) for censoring of the survival time by discharge from the ICU, and (b) for time-dependent confounders on the causal path from infection to mortality.

In this talk, we address these issues by considering ICU discharge as a competing risk and then inferring the risk of ICU mortality over time that would be observed if nosocomial infections could be prevented for the entire study population. For this purpose we develop marginal structural subdistribution hazard models with accompanying estimation methods. In contrast to subdistribution hazard models with time-varying covariates, the proposed approach (a) can accommodate high-dimensional confounders, (b) avoids regression adjustment for post-infection measurements and thereby so-called collider-stratification bias, and (c) results in a well-defined model for the cumulative incidence function. The methods are used to analyze data from the National Surveillance Study of Nosocomial Infections in ICUs (Belgium) and a large French multi-center ICU database (OUTCOMEREA).

Heavy tailed functional linear processes

T. Meinguet

Université Catholique de Louvain, Belgium

A powerful way to model spatio-temporal phenomena is by means of time series of functional observations. For risk management purposes, the interest is often in the extremes of such processes; examples from the literature include sea levels along dikes [5], windspeeds along the faces of a building [4] and precipitation [3]. In such cases, second moments cannot be assumed to exist, violating the basic assumption in standard functional data analysis based on the sequence of autocovariance operators [2, 8].

While originally defined for univariate functions and random variables, the concept of regular variation has by now been defined and studied in quite abstract settings, including the one of stochastic processes [6, 7]. As for random variables, regular variation provides the mathematical backbone for a coherent theory of extreme values. By considering functional observations as points in a suitable function space, we are led to consider regularly varying time series taking values in Banach spaces.

As in the finite-dimensional case [1], joint regular variation of a stationary time series in a separable Banach space is shown to be equivalent to the existence of a tail process. It admits a familiar-looking decomposition into independent radial and angular components. The radial component is fully determined by the index of regular variation, while the angular component, called spectral process, effectively captures all aspects of extremal dependence: extremal indices, point processes of extremes, etc.

An explicit expression of the spectral process is obtained for linear time series of the type

$$X_t = \sum_{i \geq 0} T_i(Z_{t-i}), \quad t \in \mathbf{Z},$$

where B_1 and B_2 are separable Banach spaces, the innovations $(Z_t)_{t \in \mathbf{Z}}$ are i.i.d. Z in B_1 , the law of Z being regularly varying with index $\alpha > 0$, and the T_i are linear continuous maps between B_1 and B_2 satisfying the requirement $\sum_{i \geq 0} \|T_i\|^p < +\infty$ for some $0 < p < \min(1, \alpha)$. As an application, the spectral process provides the formula

$$\theta = \frac{E \left[\sup_{t \geq 0} \|T_t(\Theta^Z)\|^\alpha \right]}{\sum_{t \geq 0} E \left[\|T_t(\Theta^Z)\|^\alpha \right]}$$

for the extremal index θ of the series $(\|X_t\|)_{t \in \mathbf{Z}}$ under the finite cluster condition, Θ^Z being a random element in B_1 having the spectral measure of Z as law.

References:

- [1] Basrak, B., Segers, J. (2008) *Regularly varying multivariate time series*, Stochastic Processes and their Applications, doi:10.1016/j.spa.2008.05.004.
- [2] Bosq, D. (2000) *Linear Processes in Function Spaces*, Springer (New-York).
- [3] Cooley D., Nychka D., Naveau P. (2007) *Bayesian Spatial Modeling of Extreme Precipitation Return Levels*, Journal of the American Statistical Association 102, 824–840.
- [4] Davis, R.A., Mikosch, T. (2008) *Extreme value theory for space-time processes with heavy-tailed distributions*, Stochastic Processes and their Applications 118, 560–584.
- [5] de Haan, L., Lin, T. (2001) *On convergence toward an extreme value distribution in $C[0,1]$* , Annals of Probability 29, No. 1, pp. 467–483.
- [6] Hult, H., Lindskog, F. (2005) *Extremal behavior of regularly varying stochastic processes*, Stochastic Processes and Their Applications 115, 249–274.
- [7] Hult, H., Lindskog, F. (2006) *Regular variation for measures on metric spaces*, Publications de l'Institut Math. (Beograd) 80, 121–140.
- [8] Pumo, B. (1998) *Prediction of Continuous Time Processes by $C[0,1]$ -valued Autoregressive Process*, Statistical Inference for Stochastic Processes 1, 297–309.

A Bayesian state-space approach to affective dynamics

T. Lodewyckx

Katholieke Universiteit Leuven, Belgium

In the last years, emotion research has been focusing on the conceptualization of emotions as multicomponential, dynamical systems. This development created a new set of challenging research questions, concerning for instance autoregressive dependencies (related to concepts of emotional homeostasis) or cross-lagged relations (related to the mutual influence of emotion components). In a first part, we want to introduce a state-space approach for the dynamical modeling of emotion components. It will be shown how Markov chain Monte Carlo methods are used to estimate the model parameters. Various model extensions are discussed (e.g. external covariates, regime-switching). In a second part, we apply this framework to high resolution psychophysiological and behavioral data obtained during emotionally evocative adolescent-parent interactions and illustrate how it can be used to obtain new insights in the dynamical nature of emotions.

WP3: Incomplete data

A Flexible extension of the Koziol-Green model by a copula function

A. Gaddah

Universiteit Hasselt, Belgium

In many industrial or clinical studies, researchers are interested in the time until an event. However, due to practical constraints, it is often not possible to fully observe the event time of interest (lifetime) and only a lower bound is observable. That is, for the lifetime of each study unit Y_i ($i = 1, 2, \dots, n$), there exists a censoring time variable C_i such that we can only observe $Z_i = \min(Y_i; C_i)$ together with a censoring indicator $\delta_i = \mathbf{1}\{Y_i \leq C_i\}$. To make inference about the lifetime, it is imperative to make a non-verifiable assumption about the relationship between Y_i and C_i (Tsiatis, 1975). In some settings, the censoring time is additionally informative to the lifetime through its distribution function. When the lifetime and censoring time are independent and the observable variables Z_i and δ_i are also independent, the Koziol and Green (1976) estimator is commonly used to model the possible information contained in the distribution of the censoring time. However, there are some situations where Z_i and δ_i are not independent. Using a general copula function to generalize the relationship between the observable lifetime and the censoring indicator, we propose a flexible extension of the Koziol-Green model under random censorship. In this proposition, we also allow for possible dependence between Y_i and C_i by means of an Archimedean copula function. We derive in this extended model, an estimator for the distribution function of the lifetime of interest. In addition, we give some asymptotic and numerical results of the estimator. Afterwards, we introduce a bootstrap testing procedure to check the validity this extended model.

References:

Koziol, J.A., Green, S.B., 1976. A Cramer-von Mises statistic for randomly censored data. *Biometrika* 63, 465-474.

A. Tsiatis, 1976. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America* 72, 20-22.

The Copula-graphic estimator in censored nonparametric location-scale regression models

A. Sujica

Université Catholique de Louvain, Belgium

In survival analysis, we are interested in the distribution function of the lifetime of some event. Due to different practical reasons we only observe a lower bound of the true lifetime. The survival time often depends on some covariate. In this case Van Keilegom and Akritas (1999) proposed a nonparametric location-scale regression model. Under the assumption of independence between the survival time and the censoring time they derived an explicit form for the estimator and its asymptotics. However, often the assumption of independence is not satisfied. In this case, Braekers and Veraverbeke (2005) considered a fixed design regression model where the dependence is described via an Archimedean copula. We are extending the model from Van Keilegom and Akritas by using the idea from Braekers and Veraverbeke, that is by assuming that the relation between the survival time and the censoring time is described via a known Archimedean copula which depends on a covariate.

References:

- Van Keilegom, I. and Akritas, M.G. (1999). Transfer of tail information in censored regression models. *Ann. Statist.*, **27**, 1745-1784.
- Braekers, R. and Veraverbeke, N. (2005). A copula-graphic estimator for the conditional survival function under dependent censoring. *Canad. J. Statist.*, **33**, 429-447.

Doubly robust multiple imputation

R.M. Daniel

London School of Hygiene and Tropical Medicine, United Kingdom

Missing data are common wherever statistical methods are applied in practice. They present a problem in that they require that additional assumptions be made about the mechanism leading to the incompleteness of the data. By incorporating two models for the missing data process, doubly robust (DR) weighting-based methods offer some protection against misspecification bias since inferences are valid when at least one of the two models is correctly specified. The balance between robustness, efficiency and analytical complexity is one which is difficult to strike, resulting in a split between the likelihood and multiple imputation (MI) school on one hand and the weighting and DR school on the other. We propose a new method, doubly robust multiple imputation, combining the convenience of MI with the robustness of the DR approach, thereby constructing approximately DR estimators in settings (such as non-monotone missing at random data) where, hitherto, estimators with this property have not been implemented. We apply the method to data from the RECORD study, a clinical trial comparing anti-glycaemic combination therapies in type II diabetes patients.

MCMC-based estimation methods for continuous longitudinal data with non-random (non)-monotone missingness

C. SOTTO

Universiteit Hasselt, Belgium

Analysis of incomplete longitudinal data requires joint modeling of the longitudinal outcomes (observed and unobserved) and the response indicators. When these two elements are independent, or at least when the non-response depends only on the observed (but not unobserved) outcomes, within a likelihood framework, the missingness is said to be ignorable, obviating the need to formally model the process that drives it. For the non-ignorable or nonrandom case, in which non-response depends on the unobserved outcomes, estimation is less straightforward, because one must work with the observed data likelihood, which involves integration over the missing values, thereby giving rise to computational complexity, especially for high-dimensional missingness. The expectation-maximization (EM) algorithm (Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977, *Journal of the Royal Statistical Society, Series B* 39, 1-38) is popular for handling incomplete data, but the so-called E(xpectation) step can be intractable, for which case the stochastic EM algorithm (Celeux, G. and Diebolt, J., 1985, *Computational Statistics Quarterly* 2, 73-82) provides an alternative, replacing the E- by an S-step, in which the missing data are imputed with plausible values, given the observed values and the current parameter estimates. Maximization of the complete data likelihood then follows using the pseudo-complete data. The method is appealing due to its computational simplicity. In this paper, we apply the SEM algorithm to fit non-random models for continuous longitudinal data with monotone or nonmonotone missingness, using simulated, as well as case study, data. Resulting SEM estimates are compared with their direct likelihood counterparts.

WP4: Data with latent heterogeneity

Methods for a global analysis of coupled data blocks that are subject to heterogeneity in the amount of noise

T.F. Wilderjans

Katholieke Universiteit Leuven, Belgium

In many fields of research, problems often result in the collection of coupled data, which may consist of different N-way N-mode data blocks that have one or more modes in common. In psychology, for example, often different pieces of information are available about the same persons, like, for instance, cognitive performance and physiological measures, implying a set of two coupled two-way data matrices. To get an overall picture of the mechanisms that may underlie the data in each of the data blocks, a global model may be used in which each data block is represented by a multi-way model, with the parameters for each shared mode being the same in all the multi-way models in which that shared mode is included; these common parameters may be estimated based on the information in all data blocks simultaneously. In this presentation, a set of novel global models will be introduced to analyze two-way and three-way coupled data blocks. Furthermore, we will discuss how analyses on the basis of the proposed models can be adjusted in order to deal with coupled data blocks that are subject to different amounts of noise (i.e., noise heterogeneity). The performance will be evaluated by means of extensive simulation studies and by means of applications to empirical real life coupled data.

A set of discrimination indices for frailty survival models

R. Van Oirbeek

Katholieke Universiteit Leuven, Belgium

Discrimination indices quantify how a survival model discriminates between high and low risk groups, thereby measuring the predictive ability of that survival model. Harrell et al. (1982) developed a discrimination index for right-censored proportional hazards (PH) models based on the ROC concepts and Harrells C-index for logistic regression. This measure has been extended to non-proportional hazards (NPH) with time-varying covariate effects and/or time-varying covariates by Antolini et al. (2005).

We suggest a discrimination measure for PH frailty models, which are specific types of NPH models. Our approach is based on Harrells C-index definition, adapted to frailty survival models.

We propose to calculate the discrimination measure in a Bayesian context employing MCMC technology. This approach has the advantage that Harrells C-index can be applied each time the frailty component is sampled in the MCMC simulation. As such an overall- and a within-cluster concordance measure is computed and their posterior uncertainty is established. The properties of the suggested concordance measure are investigated in an extensive simulation study and by making use of a real data set.

Simulations indicated that the correct estimation of the overall concordance measure heavily depends on the censoring percentage of the data set, the cluster size of the individual clusters and the variability of the frailty distribution. When the cluster size is higher than 5 and the censoring percentage lower than 50%, the estimated concordance measure shows a stable behaviour. The within concordance measure showed a good behaviour under every tested condition.

The approach gives insight in the discrimination ability of a frailty model by quantifying the overall- and within-cluster concordance of the survival frailty model.

Extending the CLASSI model for the study of individual differences in sequential processes: from crossed to nested data

E. Vande Gaer

Katholieke Universiteit Leuven, Belgium

In this paper we will focus on the modeling of binary data regarding individual differences in the responses (f.e. behaviors, emotions,) that people display in reaction to specific stimuli (f.e. situations, objects,). Underlying such data, psychologists typically assume a sequential process with two links: stimuli activate specific mediating variables (f.e. appraisals, action tendencies,) (link 1); subsequently, specific patterns of activated mediating variables elicit a particular response (link 2). It is further hypothesized that these two sorts of links may differ across persons. An important challenge then consists of retrieving the place and the nature of the key individual differences in the process under study. To meet this challenge, Ceulemans and Van Mechelen (2008) recently introduced the CLASSI model. However, the CLASSI model requires the persons and stimuli to be fully crossed, implying that for each person information has to be available about the same set of stimuli. This is a major restriction since not all stimuli are equally relevant for every person. To overcome this restraint we propose an extension of the CLASSI model which permits the set of rated stimuli to differ across persons, implying that the stimuli are nested within persons rather than crossed. Like the original CLASSI model for crossed data, the new CLASSI model for nested data (1) reduces the mediating variables and persons to a few types, and (2) defines linking structures between the stimuli and the mediating variable types on the one hand and the appraisal types and response on the other hand, which represent individual differences in these two sorts of links.

References:

E. Ceulemans and I. Van Mechelen. CLASSI: A classification model for the study of sequential processes and individual differences therein. *Psychometrika*, 73:107-124, 2008.

Clusterwise SCA for the analysis of structural differences in multivariate multilevel data

K. De Roover

Katholieke Universiteit Leuven, Belgium

Numerous research questions in educational sciences and psychology concern the structure of a set of variables. In educational sciences, one is for instance interested in the structure of the beliefs, competencies, and expectancies of students (e.g., Vanhoof, Castro Sotos, Onghena, Verschaffel, Van Dooren and Van den Noortgate, 2006). The debate about the structure of emotions (e.g., Kuppens, Ceulemans, Timmerman, Diener and Kim-Prieto, 2006) is an example from psychology: In this debate, one often assumes that emotions can be organized in a low-dimensional space; however, considerable disagreement exists about the number and the nature of the dimensions of this space (e.g., Fontaine, Scherer, Roesch and Ellsworth, 2007).

To study the structure of such a set of variables, a group of persons are scored on these variables. However, one may wonder whether the same structure would have been retrieved if another group of persons had been studied. After all, the structure of the beliefs, competencies, and expectations of students may differ strongly across disciplines. Similarly, the covariation of emotions may vary strongly across cultures (e.g., Eid and Diener, 2001). It goes without saying that to trace such structural differences, one will have to gather data from students from different disciplines, or from inhabitants of different nations. Formally, the resulting data constitute multivariate multilevel data, with the persons being nested within groups. Obviously, the crucial question is how such data have to be analyzed to find out whether and in what way the structure of the variables differs across the groups of persons.

A number of principal component analysis techniques exist to study such structural differences, for instance, simultaneous component analysis (Timmerman and Kiers, 2003). However, these techniques suffer from some important limitations. Therefore, in this presentation, we propose a novel generic modelling strategy, called clusterwise SCA, that solves these limitations by combining clustering and SCA analysis and that encompasses the existing techniques as special cases.

An item mixture model to detect Differential Item Functioning

S. Frederickx

Katholieke Universiteit Leuven, Belgium

In this presentation we present a new methodology for detecting Differential Item Functioning (DIF). We introduce a DIF model that is based on a Rasch model with random item difficulties (besides the common random person abilities). In addition, a mixture model is assumed for the item difficulties such that the items may belong to one of two classes: a DIF or a non-DIF class. The crucial difference between the DIF class and the non-DIF class is that the item difficulties in the DIF class may differ according to the observed person groups while they are equal across the person groups for the items from the non-DIF class. Statistical inference for the item mixture DIF model is carried out in a Bayesian framework. The performance of the item mixture DIF model is evaluated using a simulation study in which it is compared with traditional procedures, like the Likelihood Ratio test, the Mantel-Haenszel procedure and the standardized p-DIF procedure. In this comparison, the item mixture DIF model performs better than the other methods. Finally, the usefulness of the model is also demonstrated on a real life dataset.

WP5: High-dimensional and compound data

Detection of differentially expressed genes in microarrays: A semiparametric unified approach

J. De Neve

Ghent University, Belgium

A general method is proposed for detecting differential genes for Affymetrix GeneChip microarrays. It is a unified approach in the sense that it integrates the three preprocessing steps and the statistical testing methods into one semi-parametric model. An important characteristic is that no stringent assumptions are imposed on the background correction and normalization steps. Instead of focusing on mean differences in gene expression, we formulate the model in terms of the probabilistic index. In particular, probabilities $P(Y_1 < Y_2)$, with Y_i the intensity of a gene in group i ($i = 1, 2$), are modeled in terms of predictor variables. We present some theoretical results, and spike-in studies are considered for comparing the performance of this new method with existing methods.

References:

Wu, Zhijin and Irizarry, Rafael A. (2007) A statistical framework for the analysis of microarray probe-level data. *The Annals of Applied Statistics*, 1 333-357.

Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3.

Tsiatis, Anastasios A. (2006) *Semiparametric Theory and Missing Data* (Springer Series in Statistics). Springer.

An extended model for the enzymatically ^{18}O -labeled mass spectra

Q. Zhu

Universiteit Hasselt, Belgium

To reduce the influence of the between-spectra variability on the results of peptide quantification, one can consider the ^{18}O -labeling approach. The idea is similar to, e.g., two-channel cDNA microarrays. Peptides from two biological samples are analyzed in the same spectrum. To distinguish between the two samples, peptides from one sample are labeled with a stable isotope of oxygen, ^{18}O , while those from the other sample are left unlabeled. As a result, a mass shift of 4 Da of the isotopic distributions of peptides from the labeled sample is induced, which allows to distinguish the two samples and to quantify the relative abundance of the peptides.

It is worth noting, however, that, due to the presence of small quantities of ^{16}O and ^{17}O atoms during the labeling step, the labeled peptide may get various isotopes of oxygen. As a result, not all molecules of the labeled peptide will be shifted by 4 Da in the spectrum. This incomplete labeling may result in the biased estimation of the relative abundance of the peptide in the compared samples.

To address this issue, Valkenborg et al. developed a Markov model, which allows to adjust the analysis of the ^{18}O -labeled spectra for incomplete labeling. The model assumed that the peak intensities, observed in a spectrum, were normally distributed with a constant variance. This assumption is most likely too restrictive from a practical point of view.

To account for the problem, we extend the model, proposed by Valkenborg et al., to include a heteroscedastic normal error. In particular, we use a variance function, which allows the variance of the observed peak intensity to be equal to a power function of the intensity. Such a dependence has been observed in practice. Moreover, we formulate the model within the Frequentist as well as Bayesian frameworks. Both approaches open the possibility to further extend the model by, e.g., the inclusion of random effects that can be used to capture the biological variability of the peptide abundance. We investigate the operational characteristics of the models by applying it to real-life mass-spectrometry datasets and by conducting simulation studies.

References:

Valkenborg, D. and Burzykowski, T. A discrete-time Markov-chain model for the analysis of high-resolution enzymatically ^{18}O -labeled mass spectra. Submitted.

Integrating functional genomics data using maximum likelihood based simultaneous component analysis

R. van den Berg

Katholieke Universiteit Leuven, Belgium

In contemporary biology, complex biological processes are increasingly studied by collecting and analyzing measurements of the same entities that are collected with different analytical platforms. Such data comprise a number of data blocks that are coupled via a common mode. The goal of collecting this type of data is to discover biological mechanisms that underlie the behavior of the variables in the different data blocks. The simultaneous component analysis (SCA) family of data analysis methods is suited for this task. However, a SCA may be hampered by the data blocks being subjected to different amounts of measurement error, or noise. To unveil the true mechanisms underlying the data, it could be fruitful to take noise heterogeneity into consideration in the data analysis. Maximum likelihood based SCA (MxLSCA-P) was developed for this purpose. In a previous simulation study it outperformed normal SCA-P. This previous study, however, did not mimic in many respects typical functional genomics data sets, such as, data blocks coupled via the experimental mode, more variables than experimental units, and medium to high correlations between variables. Here, we present a new simulation study in which the usefulness of MxLSCA-P compared to ordinary SCA-P is evaluated within a typical functional genomics setting. Subsequently, the performance of the two methods is evaluated by analysis of a real life *Escherichia coli* metabolomics data set.

In the simulation study, MxLSCA-P outperforms SCA-P in terms of recovery of the true underlying scores of the common mode and of the true values underlying the data entries. MxLSCA-P further performed especially better when the simulated data blocks were subject to different noise levels. In the analysis of an *E. coli* metabolomics data set, MxLSCA-P provided a slightly better and more consistent interpretation.

MxLSCA-P is a promising addition to the SCA family. The analysis of coupled functional genomics data blocks could benefit from its ability to take different noise levels per data block into consideration and improve the recovery of the true patterns underlying the data. Moreover, the maximum likelihood based approach underlying MxLSCA-P could be extended to custom-made solutions to specific problems encountered.

A generic model for data fusion

I. Van Mechelen

Katholieke Universiteit Leuven, Belgium

In many research contexts, data show up that take the form of multiple linked data blocks. As an example, one may think of several batches of information with regard to a same set of entities as stemming from different sources of information. Data sets that consist of multiple linked data blocks imply novel challenges for the data analyst. More in particular, the data analyst may wish to go for a simultaneous modeling of the different linked blocks, and this for several possible reasons, including: (a) to arrive at more reliable inferences, (b) to grasp in an effective way common as well as distinct pieces of structural information as included in the different data blocks, and (c) to get a better understanding of the linkage relations between the different data blocks. In this paper, we will outline a conceptual framework for the simultaneous modeling of linked data blocks. Subsequently, we will introduce a generic model for this problem, which subsumes a broad range of specific models (existing as well as to be developed) as special cases.

SCA and Rotation to distinguish common and distinctive information in coupled data

M. Schouteden

Katholieke Universiteit Leuven, Belgium

Often data are collected consisting of different blocks that all contain information about the same entities (e.g., items, persons, situations). In order to unveil both information that is common to all data blocks and information that is distinctive for one or a few of them, an integrated analysis of all data blocks jointly may be most useful. An interesting class of methods for such an integrated analysis is the family of methods of simultaneous component analysis. These methods yield dimensions underlying the data that maximally account for the variance in all data blocks. Unfortunately, in results from simultaneous component methods common and distinctive information are mixed up. This paper proposes a novel method to disentangle the two kinds of information, making use of the rotational freedom of the component model. We illustrate the proposed method with gene expression data obtained for the same set of genes in three different populations that were studied with respect to autism spectrum disorder.

Simultaneous component analysis with rotation to common and distinctive components as an alternative to the generalized singular value decomposition

K. Van Deun

Katholieke Universiteit Leuven, Belgium

Currently, integrating information on the same set of biomolecules or conditions that stems from different sources (e.g., different organisms, measurement platforms) is one of the main challenges in the biomedical sciences. Often the aim of such data integration is to find the important biological processes that underly the data and to disentangle therein processes that are common for all data sources and processes that are distinctive for a particular source. The generalized singular value decomposition (GSVD) has been proposed as a data processing method to attain this goal (Alter, Brown, and Botstein, 2003). However, a motivation why the method would approach such data well lacks. As an alternative we propose to use a method that has an optimal approximation property and that is based on rotating simultaneous components to common and distinctive components (the so-called DISCO-SCA method).

Using simulated data, we illustrate that DISCO-SCA gives a better approximation than the GSVD and we show that DISCO-SCA recovers both the common and distinctive components while the GSVD only recovers the distinctive components. We also show an application of DISCO-SCA to metabolomics data for *Escherichia Coli* obtained with two different chemical analytical methods.

References:

Alter, O., Brown, P.O., and Botstein, D. (2003). Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *PNAS*, 100, 3351 - 3356.

WP6: Miscellaneous topics

Model selection strategy for generalized linear models based on robust estimators

K. Tharmaratnam

Katholieke Universiteit Leuven, Belgium

This paper is about robust model selection strategies for regression models. Model selection is a key component in any statistical analysis. There are several strategies for model selection in regression analysis. Akaike's information criterion (AIC) is one of the most popular strategy. We derive a model selection strategy in the style of AIC based on S- and MM-estimators for regression models. Also, we derive AIC for generalized linear models based on quasi-likelihood with M-estimators. We compare different robust AIC methods based on M-, S- and MM- estimators with the classical AIC method. In a simulation study and real data examples we observe that the proposed AIC with S- and MM- estimators selects more appropriate models in case outliers are present.

The association of obesity and gestational weight gain during pregnancy with obstetric outcomes

M. Babanezhad

Golestan University, Iran

It is known that the obesity and gestational weight gain are important factors on the obstetric and neonatal outcomes. This study investigates the effects of obesity and gestational weight gain on the obstetric and neonatal outcomes, in different maternal Body Mass Index (BMI) classes. Specifically, we investigate the association of gestational weight gain and body mass index during pregnancy with obstetric outcomes in a prospective population-based cohort study. The study population consisted of 1, 222 singleton term pregnancies referred to Imam Khomeini Hospital of Sari in North of Iran, in December 20, 2006 through December 20, 2007, where maternal height, maternal age, maternal weight in early pregnancy or in first trimester were available. The women were classified in 3 classes of BMI (Kg/m) in two gestational weight gain classes. We calculated the adjusted odds ratio to estimate the risk for the rate of neonatal morbidity, the rate of cesarean delivery, post-term delivery, birth weight, and low Apgar score (<7 at 5 min).

A family-based association test to detect gene-gene interactions in the presence of linkage

L. De Lobel

Ghent University, Belgium

For many complex diseases, quantitative traits contain more information than dichotomous traits. One of the approaches used to analyse these traits in family-based association studies is the Quantitative Transmission Disequilibrium Test (QTDT). The QTDT is a regression-based approach that models simultaneously linkage and association, resulting in a test for association in the presence of linkage. It splits up the association effects in a between-family and a within-family component to adjust and test for population stratification. Furthermore, a variance components method is included in the model to be able to model linkage.

We extend this approach to detect gene-gene interactions between two unlinked QTLs in the family-based setting while correcting for population stratification. We adjust the definition for the between-family and within-family component and the variance components included in the model. To capture the influence of population stratification on the detection of gene-gene interactions, we calculate the bias of the estimated interaction effect and discuss the influence on type I error rates. We simulate data to investigate the influence of the epistatic model, LD patterns between the markers and the QTLs, family structures and allele frequencies on the power and type I error rates of the approach. Results show that for some of the investigated settings, power gains are obtained in comparison with other techniques (e.g. FBAT-LC). We conclude that our approach shows promising results for studies where too few markers are available to correct for population stratification using standard methods (e.g. EIGENSTRAT). The proposed method is applied to a real-life dataset on hypertension.

References:

G. R. Abecasis, L. R. Cardon, W. O. C. Cookson. (2000), A General Test of Association for Quantitative Traits in Nuclear Families, *American Journal of Human Genetics*, no. (66), 279-292.

Design and testing for clinical trials faced with misclassified causes-of-death

B. Van Rompaye

Ghent University, Belgium

Cause-specific mortality is a common endpoint both for clinical trials and epidemiologic studies. Careful modelling of cause-specific hazards (CSHs) and reconstruction of the cumulative incidence of the event under study can help guide future policy. Failure type is however commonly subject to misclassification, for example in RCTs of cancer screening, epidemiological research based on death certificates, and in developing countries where the death cause is often obtained through proxy interviews, a so called verbal autopsy. Jaffar et al. show how such misclassification can bias effect estimates and substantially reduce power. This widens the gap between bench and bedside, hindering translation of fundamental research into clinical practice. Correction for possible misclassification can therefore increase efficiency in a wide variety of clinical studies.

We derive a modified logrank test accounting for misclassification at given sensitivity and specificity of recorded death cause, under proportional CSH assumptions. The test compares favorably with standard methods in terms of relative efficiency, allowing for a substantial reduction in sample size. Simulation of the Gambia Pneumococcal Vaccine (GPV) Trial relying on verbal autopsy information indicates a possible gain in required sample size of more than 30%. Applying the test to data from this trial we found a strong decrease of the p-value of the treatment effect on the CSH.

To support more general data structures, we apply the same approach to Cox models. This allows asymptotically unbiased estimation of covariate effects and more powerful testing. A discussion of asymptotic and small-sample properties of the covariance matrix estimator is provided. After reviewing the properties of the method we apply it again to the GPV data. The results are compared to classical logrank and Cox-model approaches.

References:

Jaffar S., Leach A., Smith P.G. et al. (2003). Effects of misclassification of causes of death on the power of a trial to assess the efficacy of a pneumococcal conjugate vaccine in The Gambia. *IJE* 32, 430-436

Detecting influential data points for Pareto-type distributions

D. Vanpaemel

Katholieke Universiteit Leuven, Belgium

In extreme value statistics, the Extreme Value Index (EVI), often denoted by γ , is used to characterize the tail behavior of a distribution. This real-valued parameter helps to indicate the size and frequency of certain extreme events under a given probability distribution: the larger γ is, the heavier the tail of the distribution. Distributions for which $\gamma > 0$ are called Pareto-type (or heavy tailed) distributions. The Hill estimator $\hat{\gamma}_{k,n}^H = \frac{1}{k} \sum_{j=1}^k Z_j$, with k a well-chosen tuning parameter and $Z_j := j(\log X_{n-j+1,n} - \log X_{n-j,n})$ the log-spacings of the data, is a well known estimator for γ (Beirlant et al., 2004).

We present a new method for measuring the influence of the individual observations on the estimation of γ , using an empirical influence function (EIF) (Pison and Van Aelst, 2004). The empirical influence function of the Hill estimator uses an estimate of γ , so it is important to insert a robust estimate of γ . Otherwise the estimate itself will be very biased by highly influential data points.

In this presentation, a robust estimator will be introduced based on a robust GLM estimator applied to the log-spacings of the data (Cantoni and Ronchetti, 2001, Beirlant et al., 1999). Based on the asymptotic normality of this estimator, we can also derive cutoff values for automatically detecting highly influential data points.

References:

- J. Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys, Tail Index Estimation and an Exponential Regression Model. *Extremes* 2 (1999), 177-200.
- J. Beirlant, Y. Goegebeur, J. Segers and J. Teugels, *Statistics of Extremes: Theory and Applications* (2005). John Wiley & Sons, New York.
- E. Cantoni, and E. Ronchetti, Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association* 96 (2001), 1022-1030.
- G. Pison and S. Van Aelst, Diagnostic Plots for Robust Multivariate Methods. *Journal of Computational and Graphical Statistics* 13 (2004), 310-329.