B-Stat News

The Newsletter
of the

Belgian Statistical Society
Belgische Vereniging voor Statistiek
Société Belge de Statistique

Number 47 – September 2009

Editeurs responsables/Verantwoordelijke uitgevers :
Sophie Vanbelle, Service de Biostatistique, Université de Liège
CHU Sart Tilman, 4000 Liège
Herbert Thijs, Center for Statistics, Hasselt University
Agoralaan - building D, 3590 Diepenbeek

THE COUNCIL OF THE SOCIETY

President

Pr. Paul Janssen, Universiteit Hasselt

Vice-President

Pr. Marcel Rémon, Facultés universitaires ND de la Paix, Namur

Secretary of the Society

Pr. Gentiane Haesbroeck, Université de Liège.

Treasurer

Pr. Marc Aerts, Universiteit Hasselt

Administrators

Pr. Thomas Bruss, Université Libre de Bruxelles
Dr. Guy Brys, Directorate General, Statistics Belgium
Dr. Filip De Ridder, Janssen Pharmaceutica
Pr. Uwe Einmahl, Vrije Universiteit Brussel
Pr. Peter Goos, Universeit Antwerpen
Pr. Christian Ritter, Université Catholique de Louvain
Pr. Martina Vandebroek, Katholiek Universiteit Leuven
Pr. Dirk Van Den Poel, Universiteit Gent
Pr. Ingrid Van Keilegom, Université Catholique de Louvain
Pr. Kristel Van Steen, Université de Liège

Website of the Society

www.sbs-bvs.be

B-Stat News editor

Sophie Vanbelle : sophie.vanbelle@ulg.ac.be
Herbert Thijs: herbert.thijs@uhasselt.be

Webmaster

Laurence Seidel : laurence.seidel@ulg.ac.be

TABLE OF CONTENTS

BELGIAN STATISTICIANS

with Léopold Simar

Can you explain how you became interested in statistics was born during your engineer studies?

This is a complicated and somewhat funny story. There was a course of statistics and probability when I was in second year of civil engineer at the university and I had problems with this course. As a matter of fact, I failed the exam in June because the professor did not let me ask clarifications about his question. I did not really understand the question and I ask: "Please, do you mean that or that?" and he replied: "Usually, it is me who asks the questions not the students, take the door out" and I received a zero. All my other exams were good on average; therefore I had only to pass again this exam in September. Thus, I spent time during my holidays on reading books on statistics and probability. When I went back to the exam in September, the professor asked me several questions but I replied in another way he taught us. Then he told me: "Well, what you are saying is correct but this is different from what I taught you, where did you learn that?" So, I said I red this book and this other book. The professor was fine and honest and I received a good mark. But most importantly, I found much more interesting what I red in probability and statistics during my holidays, than what this professor taught us during the year. Therefore, the next year, I continued to attend some courses in statistics but now taught by José Paris. Of course, José Paris knew something about statistics (which I think was not really the case of the first professor I had) and my passion for statistics continued to increase. My first professor was probably a mathematician regarding probability and statistics as an exercise of mathematics and nothing more. That was my starting point. From this time, I was always fascinated by this idea of trying to learn about complicate phenomena and by observing them. This remained the main theme of my studies in statistics.

You are a leading scientist in research on frontier estimation. Can you explain what is frontier estimation and give some applications?

Well, yes indeed. It is rather easy. In fact, if you think about economy or management, you can regard a production process as follows; a unit uses factors of production called "inputs" that produce goods called "outputs". There are many inputs and outputs and you want to compare whether some firms produce more efficiently than other ones in the same economic activity. The problem can be represented in a multidimensional space, the space of the inputs and the outputs which in practice may be high dimensional. You want then to determine the upper boundary of what is technically possible (the maximum achievable level of outputs for a given level of inputs). Mathematically, this is a surface and economically this represents the "benchmark" production frontier. Once this surface is determined, you can measure the efficiency of a firm by measuring its distance to the point where it should be for being on the frontier. This distance allows comparing efficiency of firms being on different levels of inputs and outputs. It's easier in two dimensions of course, with one input and one output. You have a curve describing the frontier and the firms you observe are below this curve. It is different from regression. In regression you estimate an average curve, here you are estimating an upper boundary curve. The difficulty, and so the interest for me, is when you are in a high dimensional space. Of course, the main interest is when you take a nonparametric approach, avoiding some restrictive parametric assumptions both on the shape of the frontier and on the stochastics behind the production process. There are many applications in business and management but also in public services; we have applications for the efficiency analysis of the post offices, for hospitals, for universities (they received money and they produce students, researchers, papers, etc…), banks, trains networks, airlines, air controllers, etc. In fact this can be applied in all fields of human activities where the objective is to reach a maximum level of outputs given some inputs. This is quite interesting.

Can you said a few word about your motivation to create the Institute of Statistics at UCL in 1992 ?

After my PhD, I went for a post-doc at Cornell University in the United States and when I was there, I was asked to start a position as junior lecturer at the Facultés universitaires Saint-Louis (FUSL) in Brussels, but I kept during that period a research position at CORE (Center for Operational Research in Econometrics, UCL), where I prepared my PhD. So, I was doing my teaching at FUSL, my research at CORE, but very quickly I was asked to take some administrative duties at Saint-Louis (I was the dean of the faculty of economy for 12 years). I was very happy during these years to be able to keep my research activities at CORE. After a couple of years, the

University of Louvain asked me also to teach some courses in statistics. Then, after something like 15 years of this "steady state", the Rector Pierre Macq of the UCL asked me to make an audit of the situation of statistics in the university and to propose solutions to try to improve the situation. The idea was that being outside the university -but partially inside because I was still working at CORE and I was teaching one or two courses at UCL- I had some chance of success... So, I received a mission of two years for preparing something. After one year, I proposed to the rector the project of creating an Institute of Statistics. The rector, supported my project at one condition: to leave Saint-Louis, become a full professor at UCL and accept to chair this new Institute for a couple of years. I did it in fact for twelve years. The motivation behind this project was that there were a lot of good statisticians in the university of Louvain-la-Neuve, but they were spread all over the university. There were researchers in statistics of high quality in medicine, in agronomy, in economy, in mathematics, in sciences, etc… but there was no name and no visible place covering their activities. So, it was not well known by the international scientific community that there was an intense research activity in statistics in Louvain, except of course at the individual level. The idea was thus to create a single entity that would bring together people coming on one side, from the more mathematical and theoretical world of statistics, and on the other side from the more applied world (agronomy, psychology, medicine, economy, social sciences, etc…) The basic project was to try to group all these people together in one single entity with the missions of doing research of high level in statistics, organizing the teaching in statistics throughout the university, advising new PhDs in statistics and providing consulting help in and outside the university. In summary, these are the missions of any department of the university. Of course, for me, it was crucial that this department, this Institute, would not be located inside one single faculty, but should remain "interfacultary" because it would cover fields coming from mathematics to medicine, passing through economy, social science and psychology. It would have been very bad for our image to receive a label of being in one particular faculty. My fear was that having the label of being mathematicians, then the economists, the social scientists, the psychologists or the people from medicine would not join us. On the other hand, having the label of being from psychology, social sciences or economy, then people coming from sciences and mathematics would not be ready to consider us as being "serious". By the "Organic law" of our university, every department had to depend from one and only one faculty, but the rector agreed to add to this rule the following statement: "except for the Institute of Statistics, which does not depend of any faculty". We are indeed in Belgium... The Institute was thus directly depending on the rector until today. This year, our university has new projects, reforms and new rules but I hope this will not

change the main reasons of the success we had for almost the past twenty years.

What was the situation at the creation compared to now?
At the start, we were only something like five persons full time, mainly coming from mathematics, economics and one from medicine. Then, the university opened a position for me because I was not full professor here and some other new positions because I was requesting in my project more positions to achieve a critical size. So, we were very quickly 6, 7, 8 and today we are 10 full-time academics here at the Institute. We have now also agreements and contacts with the Institute of actuarial sciences. The organization of the teaching is also important, which means the organization of the exercises linked to the courses of statistics of the university. It was also in my project that we had to organize all of this. Before the Institute, there were some isolated teaching assistants in medicine, in psychology, etc… not speaking the same statistical language and this was not optimal. We were able to convince the deans to put most of them together in our Institute and today I think we have something like 13 or maybe more full-time teaching assistants. We have also about 15 researchers sponsored by research projects (like IAP, ARC, etc...). So today, we are not far from 50 persons, including 3 computer persons (experts in computers and/or in statistical computing). Of course we haven't forgotten the administrative staff. At the start, I was able to get one-half person and today I think there is something like 4 full-time persons, including one director of the administration. So, it was a quite an amazing experience because we started with a very small budget, and with a very small number of persons. I was patient and year after year, with the support of the rector, we were able to increase the number of positions. In the 90's, the universities, like today, were not rich enough to create new positions from nothing, so it was not easy. In fact, all these positions were obtained by transforming positions from the other faculties. The deans agreed because I was able to convince them that this was a good plan for them... So, I have to thank them today for their confidence... But it was a lot of work! It is amazing to see year after year how we started with just one corridor in the building of CORE and now we are a whole building with 50 persons working there.

Do you have time to continue your personal research with all this work?
Yes, that's a good question. I was indeed able to keep a good level in my research with a reasonable number of publications. This means indeed a lot of work but research is not really working. For me, research is fun and I like fun, as maybe you know. Research is part of this fun so it's probably true

that during those years, I invested a lot of my time in my life trying to keep a good balance between work and family, both including fun. I think I was able to do it but it was not always easy. It was important for me, first because my main motivation for doing the job of professor in a university is doing research. Secondly, if you are not a good researcher with international recognition, people working in the university will not listen to you. And remember, I had to convince a lot of people (deans, etc.) to trust my project. That's maybe a bad thing, but that's the reality in the university, people may have some power if they are good researchers (it is a necessary but not sufficient condition, of course, some diplomacy is also probably useful...).

Can you tell us the best and worst souvenir during your carreer?
I don't know if I have a worst souvenir. At some moments in my career, there were some fights for such or such idea with some colleagues. It was not always easy because some people wanted to attack me, but it is maybe not a so bad souvenir because finally I won. Maybe the worst part of my job, that I really hated, was the exams. I remember one year, a long time ago, maybe in the 90's, I had for the month of June to enter in the system 1500 marks before the deliberation. I was of course helped for the corrections of the written exams by teaching assistants, but handling these copies and going to all the deliberations was awful. I asked the rector at that time to give me less teaching but my teaching load has always been very intense. My average load was well above 200 hours per year. Today, I have no more teaching here and I will no more teach in this university. I really love to teach but the idea of not having to correct and score exams is the most interesting thing of my life. You ask me what is the best souvenir of my career; well… this is maybe the one that has still to come. Although I have already so many... For instance, when you have a paper that is accepted for publication, that's a real achievement. A good paper accepted in a very high level journal, this is really fun. I am still very happy about that, even today. Otherwise of course, a very good souvenir is the day when the academic council of this university decided to support my project of creating the Institute of Statistics. That was a big moment for me, but there are so many others that I don't remember which was the best one.

Do you have a particular souvenir of meeting with another famous statistician?
Yes of course, I had a lot of contacts with many visitors at our Institute. For instance, a researcher like Peter Hall is really impressive, I had the chance to meet him in LLN and I am very proud of having produced a couple of papers with him. Luc Devroye is an other big person, Luc is even a Belgian guy and he is Honorary Member of the BSS. For those who know him (and

me) you can imagine the fun we had in testing some good Belgian beers at several places over the world... Both received the honors of Doctor Honoris Causa of the University of Louvain. It is impressive to work with people like them. They are of such high level, remain just simple natural persons and they don't consider themselves as being really big stars, although they really are. It is always a pleasure for me to meet them and I am happy to continue to meet people like these two "bigs".

You will now retire. Do you have any particular plan ?
Yes, my plan is very easy. It is to continue as before except that I will not teach anymore and remember, it is not teaching that is a load for me, but the exams. I will certainly continue to make research, writing papers, going to conferences, going for time to time where people invite me to do research, etc . So, freedom now is almost complete for me.

Do you want to say some word to young statisticians ?
Yes, probably for young statisticians willing to start a scientific or academic career. For me, it is the best thing that you could do. First, the discipline is quite interesting, quite fascinating because you can do many different things depending on what you like. You can do more theoretical or more applied work, as you will. You can do applications in medicine, economy, etc... It is very broad. You are not prisoner of a particular field of applications. That's the first thing. The second thing is that academic life is probably the best life over the world because you have the freedom, in a sense. You don't have complete freedom but you have the freedom of choosing the problem you want to work on. Of course it is not easy every day, you have teaching load, administrative meetings, you have to prepare research projects, to find money to make your research but that's interesting, and if you succeed, you have to prepare activities reports... At the end, this game is somewhat exciting... I would be today at the beginning of my professional life, I will not hesitate to do exactly the same things and maybe that is a sign that it is not so terrible to be a professor at the university.

Do you want to add something?
I am very happy... And to add a final word, I should even admit that I was a little bit moved when Paul Janssen announced at the conference we had in May, that the Belgian statistical society would make me the honor, because it is really a honor for me, to be considered as being an Honorary Member.

MATHEMATICS AND COMMUNICATION

Interview of F. Thomas Bruss

How do you consider the communication in mathematics?

In our department of mathematics, the communication between colleagues is in general very good. The interdisciplinary communication should be better.

A great asset for our department - and I think, for the entire field of mathematics in Belgium - is the UREM team of the ULB. This group is very active and provides an excellent service, as many people know. Obviously, their effort focuses by definition on the important contact with students and teachers of secondary schools.

Generally, I think, mathematicians should try to communicate more with the outside world. There is a tendency to focus on just what we do and not to look for communication and interaction. However, we should not criticize this lack because this is also a comprehensive tendency caused by the nature of our domain. Mathematics itself can be very difficult, and communication of Mathematics can be already difficult on a medium or even elementary level.

What are your personal experiences concerning the ways of communication in mathematics ?

My personal experiences in scientific communication are good. I wrote my first paper addressing the general public in 1998. It appeared in the DMV-Mitteilungen of the German Mathematician's Association. Since 2000, I have published general-public papers with a certain regularity, always trying to choose very visible journals.

For such papers, there are two sides to the coin. On one side, the gratification or recognition within our discipline is modest because these papers are typically not referenced in the « *Mathematical Reviews* » or in any other classicial indicator of scientific activity in our domain. In a certain sense, this lack of recognition is not justified. According to my experience it

can be more demanding and take more effort to publish in « *Pour la science* » or « *Spektrum der Wissenschaft* » (National editions of the Scientific American), say, than in « *Annals of Probability* », say, which is a real top journal in my domain. The other side of the coin is however really positive. The reactions of the readership are more numerous than what we receive as reaction to a publication in a scientific journal in a classical sense. Suddenly, there is a laboratory here, or a group of researchers there, who are interested in our work. Some readers will phone and come to see you.. In other words, we can receive interest and attention from outside, a certain notoriety.

And the conferences ?
Scientific conferences are for the mathematicians, as in all other scientific disciplines, necessary to have an efficient interaction. It is true that interaction still remains difficult between mathematicians during conferences because of the intrinsic difficulty of mathematics in several domains at the research level. The advantage of conferences is nevertheless that we can talk with colleagues which is better than just e-mails. In summary, going to conferences (not too many of them) and having an active participation, is in my point of view, indispensable.

What ways of communication do you use most ?
For my research these are the classical ways, as the majority of my colleagues would use, i.e. giving talks and publishing in scientific journals. Then, and I think this is your true question, I also publish papers for the general public: in national editions of *Scientific American (German, French, Chinese),* in newspaper like "Die Welt" or "Le Soir", in the Newsletter of the European Mathematical Society, in several Bulletins of national mathematical societies, and others. It is not rare that editors simply ask me the permission to re-publish a paper in translation, and this is easiest for me if I do not speak the language. Then those do everything! This explains why I have such publications also in Chinese, Japanese, Spanish, Polish, and Arabic. Also to mention my participation at two article competitions in scientific communication (European Mathematical Society), a lecture at the Royal Academy, another one at the Collège Belgique, some interviews in Belgian and German newspaper and on radio (BBC, RTBF, Radio Contact+ and others), and two round-table-debates on television (BEL-RTL and RTBF). Finally, there were invitations as expert in probability … and some expertise requests.

What are your experiences in the communication between mathematicians and other scientists or industries?
It happened twice that companies (one chemical, one pharmaceutical) contacted me, surprisingly, after articles of which I first thought that they have little to do with their problems. But they were both right, and this was fun. Recently one of them approached me again, and perhaps this will be more fun. I also had discussions with a political party, a couple of engineers, even with a German professor doing research on control of rocket orbits (of which I know nothing as you can guess)… but in particular with several physicians. Hence , as I said, the second side of the coin can be really rewarding. But it is also time consuming and cuts in your time for real research.

Importance of that kind of communication ?
Yes, this type of communication is, to my point of view, very important. In particular for the motivation of mathematics and for its image in public. We should be open to the outside world.

Possible improvements ?
Indeed, it should be possible and rewarding to improve communication. There was, by the way, a meeting of a commission of the European Mathematical Society in May in Brussels, just discussing this objective.

People have to communicate with other scientifics in all domains of mathematics, even in pure mathematics. For the latter, I understand this can be very difficult.

Are there communication differences following the different domains in mathematics?
There are good and bad communicators everywhere. In some domains of Pure Mathematics, communication is, I agree, and will stay probably really difficult. But even there one should be able to improve communication.

What are your personal experiences in communication with the large audiences ?
Do you mean in public speeches?

I gave commencement speeches in Berlin and Wiesbaden, and others in Brussels, Maria Laach, Leiden, … This asks for a lot of preparation. I find it not easy at all to speak about Mathematics to people with no interest in

Mathematics or a low level of mathematical education. But sometimes it's really these people you would like to win, in particular if their position and/or power of decision is in contrast with their ability in rational reasoning. And it is worth to try.

Is the communication between mathematicians and the large public sufficient ?
No, as I said I think it is not, but this is not or not only the fault of mathematicians. The pressure to do so many things does not allow us to devote enough time to it,, neither to research nor to communication of research. Decision makers should hold this in mind.

Is it always possible for you to disseminate your researches?
No. I can do this only for a part of it. In my subject of probability I also encounter problems where my interest is just "pure". Then, there are technical details in the demonstrations taking often a long time. Here I see usually nothing worth communicating to the public. This is normal, this is the problem of mathematics, as every mathematician knows it. After all, "pure" or "less pure", "applied" or less "applied", a mathematician is a mathematician. We work with the same deltas and epsilons. And this is exactly what all mathematicians should do when we have to.

Interviewed by
Carine Bartholmé

BELGIAN RESEARCH IN BAYSIAN STATISTICS REWARDED

Although it is still a topic of marginal (but rising) interest in our Society, Belgian research in Bayesian biostatistics was at the forefront at the last edition of the Joint Statitical Meetings (Washington D.C., 2009).

Indeed, the four Savage Award finalists, including two former PhD students from UCLouvain and KULeuven, were invited for a dedicated session chaired by Prof. J.G. Ibrahim (Univ. of North Carolina) and organized by Prof. M. Vannucci (Rice University) for the ASA section on Bayesian Statistical Science.

This Award, named in honor of Leonard J. Savage, is bestowed each year to two outstanding doctoral dissertations in Bayesian econometrics and statistics, one each in "Theory and Methods" and one in "Applied Methodology". The theses submitted for the Award were first read and scored by three members of a Jury. The finalists, who obtained the Savage Travel Award, had their thesis ranked in the first four positions. Then, a Jury of six members made a detailed reading of the theses to select the two Savage Awards.

Lorenzo Trippa (L. Bocconi Univ., Milano) obtained the Savage Award in "Theory and Methods" and Alejandro Jara (associate professor at Univ. de Concepción, Chile) was preferred to Donatello Telesca (Univ. of Washington, Seattle) and Astrid Jullion (Biostatistician at UCB Pharma, Belgium) for the Savage Award in "Applied Methodology".

Alejandro Jara did his PhD thesis (entitled "Bayesian semiparametric methods for the analysis of complex data") at KULeuven under the supervision of Prof. E. Lesaffre, I. Gijbels and G. Verbeke. Astrid Jullion did her thesis (entitled "Adaptive Bayesian P-splines models for fitting time-activity curves and estimating associated clinical measures in Positron Emission Tomography and Pharmacokinetic studies") at UCLouvain under the supervision of Prof. P. Lambert.

After Prof. Luc Bauwens (CORE, UCLouvain) in 1984, this is the second time in 30 editions that the Savage Award honours a research in Bayesian statistics made in our country.

Philippe Lambert

17TH ANNUAL MEETING OF THE BELGIAN STATISTICAL SOCIETY

Lommel, (14) 15-16 October 2009

The preliminary program of the next meeting of the Belgian Statistical Society (BSS 2009) is now available on the conference web-site: http://www.sbs-bvs.be/

Wednesday 14th October (Young Statisticians Afternoon)

13h30 Registration and welcome

14h00 Contributed talks young statisticians

- Gordon Gudendorf (U.C.Louvain) *Nonparametric estimation of multivariate extreme value copulas*
- Koen Mahieu (K.U.Leuven) *Maxbias Curves for Multivariate Regression Estimators*
- Girma Minalu (U.Hasselt*) Estimating the impact of school closure on social mixing behavior and the transmission of close contact infections in eight European countries*

15h00 Presentations Quetelet price winners

- Bernard Francq (U.C. Louvain) *TBA*
- Dan Lin (U. Hasselt) *Testing Procedures on Comparisons of Several Treatments with one Control in a Microarray Setting*
- Andrea Kvitkovičová (Charles University in Prague) *Changepoint fractional polynomials in modelling the seroprevalence data*
- Elke Moons (U. Hasselt) *TBA*

- Veronique Storme (U.Gent) *Integration of Transcriptomics and Metabolomics for understanding Lignin Biosynthesis*

16h00 Break

16h20 Contributed talks young statisticians

- Jonathan Jaeger (U.C.Louvain) *Functional estimation in systems defined by differential equation using Bayesian smoothing methods*

- Christel Ruwet (U.Liège) *Detection of influential observations on the error rate based on the generalized k-means clustering procedure*

17h00 Quiz

18h00 Dinner

Thursday, 15th October

8h30 Registration

9h30 Welcome and Opening of the meeting.

9h40 Invited talk by Prof. Anthony Davison (EPFL, Lausanne)
Modelling spatial extreme values.

10h30 Break and Posters

10h50 Contributed talks

Quetelet

- Jan de Neve (U.Gent) *A semiparametric unified approach for the detection of differential gene expression in microarrays.*

- Emanuele Del Fava (U.Hasselt) *The co-infection between HCV and HIV: A joint modelling approach.*

- Tatsiana Khamiakova (U.Hasselt) *Gene filtering of microarray experiments with a complex design.*

- Forcheh Chiara Anyiawung (K.U.L.) *A novel approach for exploring functional relationships in gene expression data.*

Methodology

- Herbert Thijs (U.Hasselt) *The future of missing data.*

- Roula Tsonaka (K.U.L.) *A goodness-of-fit test for the random effects distribution in mixed models.*

- Olga Reznikova (U.C.L.) *On the estimation of the dynamic conditional correlation models.*

- Teh Amouh (F.U.N.D.P.Namur) *Stratified aggregation rule for multiple classifier systems.*

12h10 Lunch and Posters

14h00 Quetelet invited talk by Prof. Dimitris Rizopoulos (Erasmus University Rotterdam)
Challenges in joint modelling of longitudinal and time-to-event data.

14h50 Contributed talks

Biostatistics

- Réjane Rousseau (U.C.L.) *Combination of Independent Component Analysis and statistical modelling for the identification of metabonomic biomarkers in 1H-NMR spectroscopy.*

- Qi Zhu (U.Hasselt) *A Bayesian approach to modeling of the enzymatically 18O-labeled mass spectra.*

- Lixin Zhang (U.Liège) *Revisiting the multivariate coefficient of variation for comparing electrophoretic techniques in External Quality Assessment (EQA) schemes.*

Methodology

- Ilaria Prosdocimi (K.U.L.) *Analyzing Italian induced abortion data with extended generalized additive models*.

- Géraldine Laurent (U.Liège) *Estimation of the error distribution in right censored and selection biased regression models.*

- Jan Johannes (U.C.L.) *Global and local minimax-optimal estimation in functional linear model.*

15h50 Break and Posters

16h20 Contributed talks

Biostatistics

- Geert Molenberghs (U.Hasselt) *On the identifiability of the incomplete-data models and random-effects.*

- Teshome Birhanu (U.Hasselt) *Pseudo-likelihood estimation for incomplete data.*

Methodology

- Michiel Debruyne (U.Antwerpen) *An outlier map for support vector machine classification.*

- Arnout Van Messem (V.U.B.) *Consistency and robustness properties of SVMs for heavy-tailed distributions.*

17h00 Invited talk by Prof. Marc Hallin (U.L.B.)
Multivariate quantiles and statistical depth: from L1 optimization to halfspace contours.

17h50 Break

18h00 General assembly of the Belgian Statistical Society

19h00 Dinner

Friday, 16th October

9h00 IAP invited talk by Prof. Alois Kneip (University of Bonn)
Functional linear regression.

9h50 Contributed talks

Biostatistics

- Aysun C¸ etiny¨urek Yavuz (U.Liège) *A flexible Cox proportional hazards model for interval-censored data using Bayesian P-splines with an extension to frailty model*.

- Bart Van Rompaye (U.Gent) *Analyzing competing risks survival data faced with misclassified causes-of-death.*

- Jan Ramon (K.U.L.) *Handling missing values and censored data in PCA of pharmacological matrices.*

B-ENBIS

- Vishva M. Danthurebandara (K.U.L.) *Effect of choice complexity on design efficiency in conjoint choice experiments*.

- Bagus Sartono (U.Antwerpen) *A new method for the orthogonal blocking of strength-three orthogonal arrays.*

- Christian Ritter (U.C.L.) *Introducing R2wd and R2PPT, two R packages to create Word documents and Powerpoint presentations from R.*

10h50 Break and Posters

11h20 Contributed talks

Quetelet

- Kristof De Beuf (U.Gent) *Analysis of tiling array data using wavelet-based functional mixed models.*

- Amparo Yovanna Castro (U.Hasselt) *The estimation of the force of infection for HCV among injecting drug users (IDU's) using interval censored data.*

Methodology

- Tim Verdonck (U.Antwerpen) *A deterministic algorithm for the MCD estimator.*

- Stephan Van der Veeken (K.U.L) *Robust transformation to symmetry: a maximum trimmed likelihood approach.*

12h00 Biostatistics invited talk by Prof. Marta Fiocco (Leiden University Medical Center)
A new multivariate gamma distribution and its implication for the Poisson correlated gamma frailty model.

12h50 Closing

13h00 Lunch

Posters

Abdelkamel Alj (U.L.B.) *The exact quasi-likelihood of time dependent VARMA Models*

Tom Cattaert (U. Liège) *Selecting multiple epistatic models using MB-MDR*

Lieven De Clercq (K.U.L.) *TBA*

Auguste Gaddah (U. Hasselt) *A flexible Koziol-Green model*

Arthur Gitome (U. Hasselt) *Resampling based multiple testing in microarray experiments: permutations or bootstrap?*

Nele Goeyvaerts (U. Hasselt) *Statistical Inference on Parvovirus B19 immunology using data on social contacts and serological status*

Maarten Jansen (K.U.L.) *Multiscale kernel smoothing using a lifting scheme*

Christophe Ley (U.L.B.) *On the singularity of Fisher information in skew-symmetric density estimation*

Harrison Machiaria (U. Antwerpen) *Equivalent-estimation versus D-optimal second-order split-plot designs*

Setia Pramana (U. Hasselt) *Order restricted semi-parametric inference for testing dose-response relationship in pre-clinical microarray experiments*

Maik Schwarz (U.C.L.) *Consistent density deconvolution under partially known error distribution*

Christelle Senterre (U.L.B.) *The contribution of biostatistics to the epidemiologic study of injuries*

Leen Slaets (K.U.L.) *Clustering based on multiresolution time warping*

Kukatharmini Tharmaratnam (KUL) *A model selection strategy for regression models based on robust estimators*

Pushpike Thilakarathne (KUL) *The use of mixed models to identify differentially expressed genes when a single replicate per biological condition is present*

Catherine Timmermans (U.C.L.) *Measuring dissimilarities between time series using their unbalanced Haar wavelets expansion*

Dina Vanpaemel (K.U.L.) *Detecting influential data points in extreme value statistics*

GENERAL ASSEMBLY OF THE SBS-BVS

Centerparcs De Vossemeren (Lommel), 15 October 2009

The SBS-BVS annual General Assembly will take place at the Centerparcs De Vossemeren (Lommel), on Thursday 15 October 2009 at 18:00 hours. The election of six Board members is one of the points on the agenda.

The General Assembly is organised during the annual meeting of the Society, but you are all cordially invited to attend.

CONFERENCE IN HONOR OF MARC HALLIN

NONPARAMETRIC STATISTICS AND TIME SERIES MODELS

Bruxelles, 27-28 November 2009

This conference is organized on the occasion of the retirement of Marc Hallin, Professor of Statistics at ULB. The organizers want to honor his very exceptional career, exceptional for both his scientific achievements and his implication in numerous national and international academic institutions (FNRS, Académie Royale de Belgique, Société Belge de Statistique, Société Française de Statistique, etc.). Beyond his former students and colleagues, the conference will bring together fifteen Belgian and foreign speakers, who count among his main co-authors and collaborators (see the list below). The program is structured around the two major research topics of Marc Hallin, namely nonparametric statistics and the analysis of time series models.

Invited speakers:

- Jan BEIRLANT (Katholieke Universiteit Leuven)
- Denis BOSQ (Université Pierre-et-Marie Curie (Paris 6))
- Christophe CROUX (Katholieke Universiteit Leuven)
- Paul DEHEUVELS (Université Pierre-et-Marie Curie (Paris 6))
- Holger DETTE (Ruhr-Universität Bochum)
- Jean-Marie DUFOUR (McGill University)
- Jana JURECKOVA (Charles University in Prague 9)
- Hira KOUL (Michigan State University)
- Marco LIPPI (Université La Sapienza in Roma)
- Zudi LU (University of Adelaide)
- Hannu OJA (University of Tampere)
- Madan PURI (University of Texas at Arlington)
- Ivan MIZERA (University of Alberta)
- Noël VERAVERBEKE (Hasselt University)
- Bas WERKER (Tilburg University)

More details on the program can be found on the web site
http://mh2009.ulb.ac.be/

FORTHCOMING STATISTICAL EVENTS

October 1, 6, 8, 2009 – Louvain-la-Neuve, Belgium, *short course: Statistical analysis of network data*,
More information:
http://www.graduatecollegescience.be/edt_objectifs.php?lang=en&ecole=STATS

October (14) 15-16, 2009 – Lommel, Belgium, *17th annual meeting of the Belgian Statistical society*,
More information: http://www.sbs-bvs.be/

November 12 2009 – Dijon, France, *workshop on "pronostic modelling"*,
More information: catherine.quantin@chu-dijon.fr

November 19-20, 2009 – Leuven, Belgium, *Modeling association and dependence in complex data*,
More information: http://med.kuleuven.be/biostat/IAP2009/index.html

November 27-28, 2009 – Bruxelles, Belgium, *Conference in honor of Marc Hallin: nonparametric statistics and time series models*
More information: http://mh2009.ulb.ac.be/

December 16-17, 2009 – Berlin, Germany, *Workshop on "demographic risk"*
More information: http://sfb649.wiwi.hu-berlin.de/demrisk/

February 4-5, 2010 – Bruxelles, Belgium, *Actuarial and financial mathematics conference 2010 interplay*,
More information: http://www.afmathconf.ugent.be/

RECENT PHD THESES

University of Hasselt (UHasselt)

Jacobs Tom. *Non-linear mixed-effects modeling for complex biopharmaceutical data* (2009) – Promotor: Pr. G. Molenberghs.

A lot of the attention on statistical models used to be focused on (generalized) linear models. However, several phenomena in nature exhibit nonlinearity. One well-known example is the plasma concentration-time profile after the administration of a drug product. With the increase in computer capacity and the availability of software such as nonmem, SAS, WinBugs, and monolix, fitting the nonlinear mixed effects models required in pharmacokinetics became feasible. As the plasma concentration-time profile is the driving force behind the efficacy and safety of a drug product, a good understanding and an accurate estimation of the models is required.

Controlled-release formulations pose additional difficulties: not only the metabolization and the elimination determine the model fit, but also the accuracy and precision of the in-vivo release mechanism ought to be investigated. However, the in-vivo release is not observed. Therefore, one combines a multiple of responses and models to overcome this. Gillespie and Veng-Pederson (1985) demonstrated that the controlled-release plasma concentration-time profile can be modeled as the convolution product of the immediate release plasma concentration-time profile and the in-vivo release of the drug product. As one is not interested in the elimination of the controlled-release formulation, this is estimated from the immediate-release plasma concentration-time profile. A hypothetical relation between the in-vitro and in-vivo release is imposed (O'Hara et al. 2001) and the adequacy of the model fit for the controlled-release plasma concentration-time profile is assessed. Such a model is referred to as an in-vitro – in-vivo correlation model (IVIVC). These models play a crucial role in the pharmaceutical industry: once an IVIVC is established, it allows assessing the impact of changing to the manufacturing procedures, batch differences, etc. based on an in-vitro test. IVIVC models also enhance formulation development.

In the dissertation, a model was presented that copes with heterogeneous formulation, i.e., with a dual release mechanism. The methodology was also extended to a one-stage procedure to allow the exchange of information

between the different sub-models of the IVIVC, coping with the in-vitro release time profile, and the immediate-release and controlled-release plasma concentration-time profile. Further, the existing model diagnostics (%PE) imposed by the authorities (FDA 2003) were criticized and local influence was introduced to detect potential outlying plasma concentration-time profiles. Further, the range of applications was extended: the IVIVC model was combined with a PK/PD model to link the in-vitro dissolution properties of a drug formulation to the in-vivo receptor binding. As such, changes in the in-vitro dissolution profile were translated into changes in the clinical effects of the controlled-release formulation. This allows determining clinically significant dissolution specifications.

In the same vein, the existing bioequivalence methodology, which determines whether two tablets or capsules yield the same drug exposure, was modified to incorporate the therapeutic window of the drug product; the newly proposed bioequivalence acceptance ranges are more conservative when minor changes in exposure lead to major clinical effects, whereas more liberal acceptance ranges are imposed when minor clinical changes are observed for large changes in exposure.

U.S. Food and Drug Administration, Center for Drug Evaluation and Research; ``Guidance for Industry: Bioavailability and Bioequivalence Studies for Orally Administered Drug Products --- General Considerations"; 2003.

Gillespie, W.R., Veng-Pedersen, P. (1985). Gastro-intestinal bioavailability: determination of in vivo release profiles of solid dosage forms by deconvolution. *Biopharmaceutics and Drug Disposition*, **6**, 351–355.

O'Hara, T., Hayes, S., Davis, J., Devane, J., Smart, T., and Dunne, A. (2001). in vivo—in vitro correlation (IVIVC) modelling incorporating a convolution step. *Journal of Pharmacokinetics and Pharmacodynamics*, **28**, 277--298.

Laenen Annouschka. *Psychometric Validation of Continuous Rating Scales from Complex Data* (2008) – Promotor: Pr. G. Molenberghs.

Psychometric Validation of Continuous Rating Scales from Complex Data
In clinical trials, the response of interest is sometimes difficult to measure. This happens in psychopharmacological trials, where interest lies in measuring traits like anxiety or depression, but also in other fields, when measuring concepts like quality of life or pain. Rating scales are generally used to measure this type of latent variables. Such scales then need to be *valid* and *reliable*. A scale is valid if it actually measures the latent variable that we are aiming at, and it is reliable if the measurement error is limited. However, the latter is not a fixed scale characteristic but it is population dependent. The same scale applied to two different populations can result in different reliabilities. In our work we have focussed on the question how we can evaluate the reliability of the outcome measurements within a clinical study.

The classical psychometric approach for evaluating reliability is based on a very restrictive modelling framework that is unlikely to hold for longitudinal clinical trial data. We therefore base our approach on the flexible family of linear mixed models that can handle the typical characteristics of repeated measurements. The first step was then to extend the concept of reliability as defined in the classical test theory to this more general setting. For doing this we used an axiomatic approach and proposed four properties that any measure for reliability should fulfill. These properties state that any measure for reliability (1) should always lie between zero and one, (2) should be zero only if there is only measurement error, and (3) should be one only if there is no measurement error at all, and finally (4) should be equal to the classical definition whenever the conditions of the classical theory are satisfied.

The above definition led to a whole family of reliability measures of which all members are weighted sums of the same basic elements: the roots of the equation $q(\lambda) = |\boldsymbol{\Sigma}\text{-}\lambda \boldsymbol{V}| = 0$; where $\boldsymbol{\Sigma}$ expresses the error variability and $\boldsymbol{V}$ the total variability. Two measures, R_T and R_Λ, were of particular interest and have been further scrutinized. In a longitudinal context, the R_T coefficient expresses the average reliability over the different measurement occasions. Having a single measure has the advantage of facilitating interpretation and is very useful whenever two scales should be compared on their reliability. On the other hand, it is possible to obtain R_T values per time point, which can be important when one is interested in the evolution of reliability over the course of the study.

The R_Λ coefficient, even though structurally similar to R_T , bears a totally different interpretation. This measure expresses the reliability of the longitudinal sequence as a whole. It captures not the average reliability per time point, but the reliability of the information that is available when considering the repeated measures jointly. As a consequence, R_Λ will always increase when the number of measurements increases. Relevantly, this implies that we can always obtain a pre-specified level of reliability if the patient is followed long enough. Indeed, even if we only have to our disposal a scale that gives rise to a relatively large amount of measurement error, we can still increase the reliability of our conclusions by repeating the measurement over time.

The previous developments were based upon a longitudinal framework. However, in psychometric research much interest has gone to the study of reliability in the context of cross-sectional, multivariate measurement. We have illustrated that the same measures as proposed in the longitudinal context also apply when studying reliability in a multivariate setting. The R_T coefficient then expresses the average reliability per item whereas the R_Λ coefficient refers to the reliability of the information available in the entire scale.

Vangeneugden Tony. *Applying Psychometric Validation Methodology to Longitudinal Clinical Trial Data* (2008) – Promotor: Pr. G. Molenberghs.

Before questionnaires or measurement scales are used in clinical trials, the psychometric properties must be assed to validate the measurement scale. Specifically, the reliability and the validity are evaluated. Reliability consists in determining the extent the measurement is free from random error. This can be performed through analyzing internal consistency and reproducibility of the questionnaire The calculation of the intraclass correlation coefficient (ICC) is one of the most commonly used methods. The validity of a questionnaire is defined as the degree which the questionnaire measures to what it purports to measure. This can be performed through the analysis of content, construct and criterion validity. This psychometric validation is done in a separate and often in a rather small sample of stable subjects. However it is important to note that validity and reliability of a scale are not unique aspects of a scale but relative to the population in which the scale is used.

In this thesis we showed that these psychometric validation techniques can also be applied to longitudinal data collected in clinical trials. The goal is not the replace upfront psychometric validation, but rather to offer methodology to evaluate validity and reliability in the specific trial population at hand. This work provided a flexible framework to evaluate the actual performance of the scale in terms of reliability or validity. More specifically, to evaluate reliability, we used the Linear Mixed Model framework to develop a general formula to derive test re-test reliability in case of interval scaled data. The same frame work was then used to extend reliability to generalizability testing. The purpose of this concept is to evaluate which factors influence reliability. Data from 5 clinical studies in schizophrenia were used to study test-retest reliability of the total Positive and Negative Syndrome Scale (PANSS). The same data was also used to evaluate generalizability of the total PANSS versus country and the baseline total negative PANSS subscore. Analysis showed the reliability of the total PANSS was not impacted by so much by the former but more by the latter factor.

This framework was then extended to binary case by means of the General Linear Mixed Model (GLMM) framework to derive approximate formulae for the ICC of reliability. These derivations allowed to derive reliability and generalizability of Clinical Global Impression (CGI) response captured in 4 clinical trials in schizophrenia. Additionally, the special case of count data was addressed. A closed form was derived to calculate the ICC of reliability and applied to number of seizures captured in a clinical study in epilepsy.

Finally, the GLMM framework was used to investigate correlation between joint longitudinal sequences of different measures. Similar a criterion validity, we evaluated the correlation between the PANSS total score and CGI response. Analyses showed that the correlation between the total PANSS and CGI response was as high as 0.75.

Van Sanden Suzy. *Statistical Methods for Microarray-based Analysis of Gene-expression, Classification, and Biomarker Validation* (2008) – Promotors: Pr. T. Burzykowski and Pr. Z. Shkedy.

One of the most important research questions that scientists still face today, is how do living organisms function, down to the cellular level. To find the answer, a substantial amount of research has been devoted to the study of the basic building block of life, DNA. However, the identification and localization of all of the approximately 20,000-25,000 genes in human DNA, a task which was completed in 2003, led to other questions: what is their function and how is their expression regulated? Gene expression is the process, by which the information carried by a gene is transformed into a protein. This process can be influenced by a number of internal (e.g., a disease) or external (e.g., the environment) factors.

Addressing these questions is, however, complicated by the fact that the genome of every living organism consists of a substantial number of genes. It is nearly impossible to examine expression levels gene by gene. The demand for techniques that would allow to simultaneously monitor a large numbers of genes was met by the development of DNA microarrays.

These technological advances in the field of genomics have brought about a new statistical research area, the analysis of data from high-throughput screening experiments. The main issue in such studies is the interest in a large number of parameters, while confronted with a small biological sample size. In addition, there is inherent "noise" in microarray data. The process of obtaining the gene-expression measurements is one of many stages. Each of these stages is vulnerable to the inclusion of unwanted systematic and random effects, possibly leading to bias in the results. Hence, there is a need for proper statistical procedures to deal with the above mentioned issues in the design and analysis of the experiments.

In the dissertation, we focus on data normalization, discovery of differentially expressed genes, and class prediction. A new transformation is proposed to pre-process data from different slides and to prepare them for further analysis. Several gene selection and classification methods are compared, with respect to their ability to classify samples to predefined groups, in an extensive simulation study. Furthermore, new modelling techniques are explored to discover differentially expressed genes. And in the final part, some methods are presented for biomarker detection.

University of Leuven (KULeuven)

Consentino Fabrizio. *Modelling the missingness: estimation, testing and model selection* (2009) – Promotor: Pr. G. Claeskens.

The term `missing data' indicates the presence of missing observations in the data set of interest, which could be both in the response and the explanatory variables. When data are collected, in any possible scientific field, the occurrence of missing observations is quite probable; this presence has an impact on analyzing the data, since the standard statistical methods could fail in order to obtain reliable results. For this reason dealing with missing data has had an increasing impact in statistical analysis, leading, in the last years, to new techniques able to overcome these statistical problems, such as for estimation, model selection, etc.

The primary aim of the work presented in this dissertation is to develop new tools in order to analyze data sets with missing observations and to draw reliable conclusions from them. Nowadays many studies in different scientific fields have to deal with large amounts of information; these datasets include a large number of variables. Hence choosing the most relevant variables for the analysis is fundamental in order to obtain valid results. However problems could rise when missing data are present potentially leading to a failure of the procedures.

Since model selection could fail in the presence of missing data, a new variant of the classical Akaike information criterion (AIC) has been proposed. The new AIC is able to perform model selection when missing observations are present in the explanatory variables. This is done by exploiting the use of the EM algorithm and by modelling the covariates with missing observations. When the covariates contain missing observations, they are treated as random variables; hence modelling them has an impact on the model selection criteria with missing observations.

Using different distributions for this purpose could lead to better results. We propose a distribution selection criterion when either a normal or a t-distribution is chosen for modelling the missing covariates. In particular this is performed using a non iterative method, which is feasible when a logistic regression model is considered. Furthermore we consider estimation of parameters in these models.

Performing hypotheses tests is fundamental in statistics. The goal of the presented research in Chapter 4 is to develop nonparametric tests that are applicable to data sets with missing observations. The studied order selection tests do not require the specification of a particular parametric alternative hypothesis. The main advantage of these tests is that by considering orthogonal series expansions the tests have non-trivial power against a wide range of alternatives. For the construction of the tests a model selection criterion is used to select amongst the models in the series expansion. Our construction uses likelihood ratio tests based on multiple imputations in order to perform the order selection test in missing data situations.

Rizopoulos Dimitris. *Joint modelling of longitudinal and survival data* (2008) – Promotor: Pr. G. Verbeke

Many longitudinal studies collect information on outcomes such as time to infection or death, as well as covariates that vary with time. These covariates are usually measured intermittently, often at different times for each participant, and with substantial error. In such studies the prognostic value of these time-dependent covariates and/or the covariate process itself may be of interest, since it sheds light on the natural history of the disease. This type of studies have lead to a new and active area of biostatistical research that deals with the joint modelling of longitudinal and event time data.

In this thesis we investigate a number of issues in the joint modelling area. In particular, we theoretically investigate the effect of misspecifying the random effects distribution in parameter estimators and standard errors, especially as the number of repeated measurements per subject increases. Further, we propose an alternative parameterization for joint models using copulas for the random effects, and discuss sensitivity analysis issues. On the computational part of the thesis, a new type of Laplace approximation is developed that can efficiently handle multidimensional random effects vectors in joint models. Finally, we postulate a flexible model for the event outcome using B-splines under which the estimation of standard errors is facilitated.

Tsonaka Roula. *Models for handling coarsening and non-monotone missingness in clinical trials* (2008) – Promotors: Pr. G. Verbeke, Pr. E. Lesaffre and Pr. M. Hubert

In clinical trials often data are either missing or inaccurately recorded. This is the so called phenomenon of data coarsening that can occur in the form of missingness, censoring and grouping or rounding. For instance, consider a longitudinal study where a variable of interest (e.g., blood pressure) is repeatedly recorded in time on the same subjects. Even though data collection is often scheduled at pre-specified points in time, not all measurements are obtained. For various reasons the study participants may fail to appear at the study centers leading thus to incomplete response profiles. Moreover, in a study where the time to a particular event (e.g., death) is of interest, the true event times may not be recorded for all individuals. The end of the study may be reached or the subjects drop out from the study before experiencing the event. In this case the variable of interest is imposed to censoring. Furthermore, in survey studies the recorded variables are often prone to misreporting and grouping or rounding (e.g., age reporting for infants is typically done in weeks or months, for adolescents and adults it is truncated to the next lower year, etc.). Analysing such coarse responses as if they were complete can lead to incorrect inferences. Therefore the reasons of data coarsening need to be carefully considered and in some settings sophisticated methods of analysis are required.

In this thesis we have developed methods for handling data coarsening in clinical trials that occurs in the form of missingness and grouping or rounding. Specifically, in Chapter 1 we explain how coarseness arises in clinical trials and discuss the implications in analysis when it is not properly addressed. In Chapters 2 and 3 we concentrate on the analysis of non-monotone missing profiles, namely incomplete response profiles that arise when the subjects miss intermittently some of the scheduled visits. In particular, we consider the Shared Parameter Model (SPM) framework, in which a latent process (e.g., the true health status) known as random effects is assumed to affect both the longitudinal responses and the mechanism that produces the missing data. Two important issues are addressed; the choice of the random effects distribution and the interpretation of the model parameters. Regarding the random effects distribution, parametric assumptions are typically considered that can be unrealistic and affect the validity of inference. Therefore, in this thesis we leave this distribution completely unspecified. For the estimation of this model a semi-parametric maximum likelihood method is used. Regarding parameter interpretation, the use of random effects induces a conditional on the random effects interpretation, which may not always be desirable (e.g., when the population

treatment effect is required). Therefore a reparameterization is applied on the longitudinal model to allow for model parameters with a population averaged interpretation. This gives rise to the marginalized semi-parametric SPM presented in Chapter 3 for the analysis of incomplete longitudinal ordinal responses.

In Chapters 4 and 5 we deal with the analysis of grouped data, namely, responses that are a coarsened version of a latent variable of primary interest. In particular, the analysis of quality of life indexes is the main topic of this part of the thesis. Such indexes often follow non-standard distributions, namely J- or U-shaped, precluding classical parametric statistical approaches for analysis. Therefore, a new parametric approach, based on the logistic transformation, is proposed for analysing index data that can capture various shapes of distribution while allowing for covariate adjustments. The proposed model and the derived formulas for power and sample size calculations presented in Chapter 5 are applied to Barthel index evaluations of patients with an acute ischemic stroke.

Finally, in Chapter 6 we conclude with a discussion of the main findings of this doctoral research and refer to topics for further research.

Vermeulen Bart. *Design issues in conjoint analysis for market and non-market valuation* (2009) – Promotors: Pr. M. Vandebroek and Pr. P. Goos (UA)

The stated preferences of individuals for different products reveal a wealth of information to researchers and practitioners in market and non-market valuation. By exploring the preferences of individuals, one can gain insight in the economic values of the features of a product. In market valuation, the focus is on measuring the impact of new features on the individuals' purchases and on predicting market shares for new products. In non-market valuation, public goods are evaluated for which there exists no market and valuation is even more complex.

To explore the preferences of individuals, conjoint experiments have become a popular tool. In these experiments, a respondent is confronted with a number of hypothetical commodities or alternatives each described by its features or attributes. The respondent is requested to choose his or her preferred alternative, to rank or rate the alternatives or to choose the options he/she likes or dislikes most. This task is then repeated for a specified number of sets of hypothetical profiles. After analyzing these conjoint data by a random utility model, the resulting part-worth or utility coefficient estimates reflect the importance of the attributes for the individuals.

One of the key challenges in implementing a conjoint experiment is the statistical design of it. This involves the combination of attributes and their levels and the allocation of the resulting alternatives to the choice sets. Because the number of combinations of attributes and their levels can be huge even with relatively simple commodities, some theory is required to construct the most appropriate design. Depending on the objective of the analyst, an optimality criterion is used to do so. Often, the researcher is interested in maximizing the information on the utility coefficients. However, a researcher might put first another goal: e.g. measuring the individuals' willingness-to-pay to obtain a particular feature of a commodity. It is obvious that other optimality criteria will be required in each of these cases.

In this work, we discuss various topics related to the design of conjoint experiments. In the first chapter, we deal with conjoint choice experiments including a no-choice option which mimics an individual's market behaviour more realistically. We examine whether including this no-choice option in the design phase of the experiment improves the accuracy of the estimated utility coefficients and predictions of the individuals' choices.

In chapter 2 and 3, we consider *rank-order conjoint experiments*, in which the respondent is asked to rank all or a number of the alternatives in each choice set, and *best-worst choice experiment*, in which the respondent indicates the most and least preferred alternative in each choice set. We propose an optimality criterion for both types of experiments to develop optimal designs leading to precise estimates of the utility coefficients. We examine whether considerable improvements in terms of estimation and prediction accuracy are obtained by using the resulting tailor-made designs compared to benchmark designs which are often used in practice. Moreover, we measure the additional information of indicating a second choice in each choice set, i.e. choosing the second best or the least preferred option.

However, a conjoint experiment is not always focused on precise estimates of the utility coefficients. Measuring individuals' willingness-to-pay (WTP) is a frequently stated goal in non-market valuation. Obtaining accurate estimates of the willingness-to-pay necessitates an other design criterion than the one yielding precise estimates of the utility coefficients. Chapter 4 and 5 provide several design criteria to obtain precise willingness-to-pay estimates. We compare the designs constructed using the proposed criteria in terms of the precision of the WTP estimates.

Yu Jie.*Optimal design methodology for choice experiments in the presence of model uncertainty and consumer heterogeneity* (2009) – Promotors: Pr. M. Vandebroek and Pr. P. Goos (UA)

Conjoint analysis is by far the most preferred technique for exploring consumers' preferences for different features of an individual product or service. One of the most important strengths of this technique is the ability to develop market simulation models that can predict how consumers would react to product changes. This assists companies in determining what features a new product should have and how it should be priced.

In conjoint experiments, any product or service (called profile or alternative) is described in terms of a number of attributes. For example, a television may have attributes of screen size, screen format, brand, price, and so on. Each attribute consists of a number of levels. For instance, levels for screen format may be CRT, LCD, or Plasma. Respondents or test persons would be shown a set of products created from a combination of levels from all or some of the attributes and asked to rank or rate these products or choose their preferred one from this set of products. The last type of experiment is called the choice-based or conjoint choice or discrete choice experiment. It is popular because it imitates consumer behavior in real-life. It is a more realistic exercise for individuals to indicate which product they would purchase rather than rating or ranking since making choices is what they actually do in the marketplace. To date, discrete choice experiments have been extensively applied in many of the social sciences and applied sciences including marketing, product management, operations research, econometrics, transportation, environmental and health economics.

In a typical choice experiment, respondents need to indicate their preferred profile from each of several choice sets presented to them. As the number of combinations of attributes and levels increases, the number of potential profiles increases exponentially. This brings us to the question of what profiles should be used in the experiment and how to group them into choice sets such that the experiment can provide maximum information on the parameter estimates? The main focus of this thesis is to construct efficient experimental choice designs.

University of Liège (ULg)

Vanbelle Sophie. *Agreement between raters and groups of raters* (2009) – Promotors: Pr. A. Albert and Pr. G. Haesbroeck.

Agreement between raters on a categorical scale is a situation often encountered in practice. For example, one may want to test for agreement between two psychiatrists on the assessment of depression in out patients. We also may want to test for the equality of such obtained agreements with two different methods. The first part of this work concerns with the first situation and the second part with the second one.

In the first part of the work, the kappa-like family of agreement indexes is described in various situations: when agreement is searched between two singles raters or more, between an isolated rater and a group of raters and between two groups of raters. To quantify the agreement between two single raters, Cohen's kappa coefficient and intraclass kappa coefficient can be used for binary and nominal scales while weighted kappa coefficient concerns with ordinal scales. An interpretation of the quadratic and the linear weighting scheme is given. When agreement is searched between several raters, agreements indexes corresponding to Cohen's kappa and intraclass kappa coefficients are exposed. Then, the kappa-like family of agreement coefficients is extended to the case of an isolated rater and a group of raters and to the case of two groups of raters. The agreement coefficients are derived on a population model and reduce to the well-known Cohen's coefficient in case of two single raters. The proposed agreement indexes are also compared to existing methods, the consensus method and Schouten's agreement index. The superiority of the new methods on these methods is shown.

In the second part of the work, methods to compare several agreement indexes are presented. Firstly, the method proposed by Fleiss (1981) to compare several independent agreement indexes is developed. Then, a bootstrap method is presented to compare two dependent agreement indexes. This method was extended by to compare more than two dependent agreement indexes. The methods cited above can be applied to compare all kind of agreement indexes introduced in the first part of the work. Finally, regression methods to test the effect of continuous and categorical covariates on the agreement between two or several raters are presented. This includes the weighted least square method, allowing only for categorical covariates and a regression method based on two sets of generalized estimating equations. The latter method was developed for the intraclass kappa coefficient, Cohen's kappa coefficient and the weighted

kappa coefficient. Finally, an heuristic method, limited to the case of independent observations is presented which turns out to be equivalent to the generalized estimating equation approach. The regression methods have still to be adapted to allow for agreement indexes between a single and a group of raters and between two groups of raters.

University of Louvain-la-neuve (UCL)

Motta Giovanni. *Evolutionary factor analysis* (2009) – Promotors: R. von Sachs and C. Hafner

Linear factor models have attracted considerable interest over recent years especially in the econometrics literature. The intuitively appealing idea to explain a panel of economic variables by a few common factors is one of the reasons for their popularity. From a statistical viewpoint, the need to reduce the cross-section dimension to a much smaller factor space dimension is obvious considering the large data sets available in economics and finance.

One of the characteristics of the traditional factor model is that the process is stationary in the time dimension. This appears restrictive, given the fact that over long time periods it is unlikely that e.g. factor loadings remain constant. For example, in the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965), typical empirical results show that factor loadings are time-varying, which in the CAPM is caused by time-varying second moments.

In this thesis we generalize the tools of factor analysis for the study of stochastic processes whose behavior evolves over time. In particular, we introduce a new class of factor models with loadings that are allowed to be smooth functions of time. To estimate the resulting non-stationary factor model we generalize the properties of the principal components technique to the time-varying framework. We mainly consider separately two classes of Evolutionary Factor Models: Evolutionary Static Factor Models (Chapter 2) and Evolutionary Dynamic Factor Models (Chapter 3).

In Chapter 2 we propose a new approximate factor model where the common components are static but non-stationary. The non-stationarity is introduced by the time-varying factor loadings, that are estimated by the eigenvectors of a non-parametrically estimated covariance matrix. Under simultaneous asymptotics (cross-section and time dimension go to infinity simultaneously), we give conditions for consistency of our estimators of the time varying covariance matrix, the loadings and the factors. This paper generalizes to the locally stationary case the results given by Bai (2003) in the stationary framework. A simulation study illustrates the performance of these estimators. The estimators proposed in Chapter 2 are based on a nonparametric estimator of the covariance matrix whose entries are computed with the same smoothing parameter. This approach has the advantage of guaranteeing a positive definite estimator but it does not adapt

to the different degree of smoothness of the different entries of the covariance matrix.

In Chapter 5 we give an additional theoretical result which explains how to construct a positive definite estimate of the covariance matrix while while permitting different smoothing parameters. This estimator is based on the Cholesky decomposition of a pre-estimator of the covariance matrix.

In Chapter 3 we introduce the dynamics in our modeling. This model generalizes the dynamic (but stationary) factor model of Forni et al. (2000), as well as the non-stationary (but static) factor model of Chapter 2. In the stationary (dynamic) case, Forni et al. (2000) show that the common components are estimated by the eigenvectors of a consistent estimator of the spectral density matrix, which is a matrix depending only on the frequency. In the evolutionary framework the dynamics of the model is explained by a time-varying spectral density matrix. This operator is a function of time as well as of the frequency. In this chapter we show that the common components of a locally stationary dynamic factor model can be estimated consistently by the eigenvectors of a consistent estimator of the time-varying spectral density matrix.

In Chapter 4 we apply our theoretical results to real data and compare the performance of our approach with that based on standard techniques. Chapter 6 concludes and mentions the main questions for future research.

Teodorescu Bianca. *General conditional linear models with time-dependent coefficients under censoring and truncation* (2008) – Promotor: Pr. Ingrid Van Keilegom

In survival analysis interest often lies in the relationship between the survival function and a certain number of covariates. It usually happens that for some individuals we cannot observe the event of interest, due to the presence of right censoring and/or left truncation. A typical example is given by a retrospective medical study, in which one is interested in the time interval between birth and death due to a certain disease. Patients who die of the disease at early age will rarely have entered the study before death and are therefore left truncated. On the other hand, for patients who are alive at the end of the study, only a lower bound of the true survival time is known and these patients are hence right censored.

In the case of censored and/or truncated responses, lots of models exist in the literature that describe the relationship between the survival function and the covariates (proportional hazards model or Cox model, log-logistic model, accelerated failure time model, additive risks model, etc.). In these models, the regression coefficients are usually supposed to be constant over time. In practice, the structure of the data might however be more complex, and it might therefore be better to consider coefficients that can vary over time. In the previous examples, certain covariates (e.g. age at diagnosis, type of surgery, extension of tumor, etc.) can have a relatively high impact on early age survival, but a lower influence at higher age. This motivated a number of authors to extend the Cox model to allow for time-dependent coefficients or consider other type of time-dependent coefficients models like the additive hazards model. In practice it is of great use to have at hand a method to check the validity of the above mentioned models.

First we consider a very general model, which includes as special cases the above mentioned models (Cox model, additive model, log-logistic model, linear transformation models, etc.) with time-dependent coefficients and study the parameter estimation by means of a least squares approach. The response is allowed to be subject to right censoring and/or left truncation.

Secondly we propose an omnibus goodness-of-fit test that will test if the general time-dependent model considered above fits the data. A bootstrap version, to approximate the critical values of the test is also proposed.

In this dissertation, for each proposed method, the finite sample performance is evaluated in a simulation study and then applied to a real data set.

We would like to publish in this *Newsletter* any statistical matter such as :

- information about universities, institutes (1 to 3 pages);
- lists of recent publications and technical reports;
- abstracts of recent PhD theses;
- news of members;
- forthcoming statistical events and announcements;
- short papers about teaching methods in statistics, statistics in the industry, official statistics, etc.

Suggestions are welcome: please, contact us.

Suitable information for the next issue, prepared as (**LA)TEX or WORD FILES**, should reach the editors of the Newsletter **BEFORE December 31, 2009**, preferable by e-mail to:

sophie.vanbelle@ulg.ac.be or herbert.thijs@uhasselt.be

Any change of job, address, phone number, ... ?

Please notify the Secretary of the Society:

Gentiane Haesbroeck:
Université de Liège
Institut de mathématique
Sart Tilman B37
B-4000 Liège
g.haesbroeck@ulg.ac.be