PROGRAMME AND ABSTRACTS

5th CSDA International Conference on

Computational and Financial Econometrics (CFE 2011)

http://www.cfe-csda.org/cfe11

and

4th International Conference of the
ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on

Computing & Statistics (ERCIM 2011)

http://www.cfe-csda.org/ercim11

Senate House, University of London, UK
17-19 December 2011

ERCIM
European Research Consortium
for Informatics and Mathematics

Queen Mary University of London
http://www.qmul.ac.uk

Birkbeck UNIVERSITY OF LONDON
http://www.bbk.ac.uk

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE
http://www.lse.ac.uk

E225: **A goodness-of-fit procedure for semiparametric copula models under random censoring**
Presenter: **Olivier Lopez**, UPMC Paris VI, France
Co-authors: Svetlana Gribkova, Philippe Saint Pierre

A new goodness-of-fit procedure for semiparametric copula models under a bivariate censoring scheme is studied. The work is motivated by a data-set containing lifetimes of married couples who subscribed to a certain type of insurance contract. These bivariate data are subject to bivariate random censoring, with the particularity that the two components of the censoring vector are identical. We provide the asymptotic behavior of our test-statistic, along with a bootstrap procedure that allows us to compute the critical values. The extension of our approach to other kinds of bivariate censoring schemes is also discussed.

E321: **Complete case analysis revisited**
Presenter: **Ursula Mueller**, Texas A&M University, United States of America

A general method is presented for obtaining limiting distributions of complete case statistics for missing data models from those of the corresponding statistics when all data are observed. This provides a convenient tool to obtain the asymptotic behaviour of complete case versions of established methods without (reproducing) lengthy proofs. It is well known that a statistical analysis which ignores cases that are only partially observed does not always perform well and that an approach which imputes missing values often has better properties. However, there are situations where a complete case analysis is appropriate. The methodology is illustrated by analysing three inference procedures for partially linear regression models with responses missing at random. Firstly, asymptotically efficient estimators of the slope parameter are derived. Secondly, an asymptotically distribution free (ADF) test for fitting a normal distribution to the errors is derived. Finally, we obtain an ADF test for linearity, i.e. for testing that the nonparametric component of these models is a constant.

E584: **Flexibly extending the classical Koziol-Green model by a copula function.**
Presenter: **Roel Braekers**, Hasselt University, Belgium
Co-authors: Auguste Gaddah

In survival analysis, we are interested in the time until an event. Due to practical reasons, we do not fully observe this time. A second independent random variable, a censoring time, obscures the observation process and only the smallest of both times is observed together with an indicator variable. The classical approach to estimate the distribution function for the lifetime under informative censoring time is by the Koziol-Green model. In this model, it is assumed that the survival function of the censoring time is a power of the survival function of the time until an event. We generalize this model and assume a parametric function for the relationship between these distribution functions. Hereby we link this assumption to a slide of a parametric copula function between the observed variables. We propose a maximum likelihood estimator for the copula parameter and construct a semi-parametric estimator for the lifetime distribution. As results, we show the consistency and asymptotic normality of the copula parameter. We derive the weak convergence of the semi-parametric distribution estimator. In a simulation study, we investigate the finite sample performance of these estimators and apply this model to a practical data set.

ES51 Room B18 ADVANCES IN SOFTWARE FOR TREE MODELS Chair: Achim Zeileis

E329: **TINT R-package for advanced subgroup analysis**
Presenter: **Elise Dusseldorp**, TNO-Dutch Center for Applied Scientific Research, Netherlands
Co-authors: Iven Van Mechelen

When multiple treatments are available, it is often difficult to decide which treatment is best for an individual given his/her state and characteristics. Empirical evidence is lacking. The method TINT focuses on supplying such empirical evidence by means of advanced subgroup analysis. It answers the question: "Who reacts better to treatment A and who reacts better to treatment B?" TINT results in a binary tree that subdivides the patients into subgroups on the basis of their characteristics; these subgroups are assigned to one of three classes: a first one for which A is better than B, a second one for which B is better than A, and an optional third one for which type of treatment makes no difference. The optimal size of the tree is determined with a bias-corrected bootstrap procedure. The method is appropriate for data from randomized clinical trials with many patient characteristics that could interact with treatment in a complex way. We will show applications of the method using the R-package TINT.

E401: **evtree: Evolutionary learning of globally optimal classification and regression trees in R**
Presenter: **Thomas Grubinger**, Innsbruck Medical University, Austria
Co-authors: Achim Zeileis, Karl-Peter Pfeiffer

Commonly used classification and regression tree methods like the CART algorithm are recursive partitioning methods that build the model in a forward stepwise search. Although this approach is known to be an efficient heuristic, the results of recursive tree methods are only locally optimal, as splits are chosen to maximize homogeneity at the next step only. An alternative way to search over the parameter space of trees is to use global optimization methods like evolutionary algorithms. The evtree package implements an evolutionary algorithm for learning globally optimal classification and regression trees in R. CPU and memory-intensive tasks are fully computed in C++ while the partykit package is leveraged to represent the resulting trees in R, providing unified infrastructure for summaries, visualizations, and predictions. evtree is compared to rpart, the open-source CART implementation, and ctree (conditional inference trees). On several benchmark problems from the UCI machine learning repository, evtree models offered at least similar and most of the time increased predictive accuracy.

E424: **partykit: A toolkit for recursive partytioning**
Presenter: **Achim Zeileis**, Universitat Innsbruck, Austria
Co-authors: Torsten Hothorn

Recursive partitioning methods, or simply "trees", are simple yet powerful methods for capturing regression relationships. Hence, many different algorithms have been suggested in both the statistics and machine learning communities and many standard algorithms are available as R packages, e.g., in rpart, RWeka, party, and many others. However, no common infrastructure is available for representing trees fitted by different packages. Consequently, the capabilities for extraction of information - such as predictions, printed summaries, or visualizations - vary between packages and come with somewhat different user interfaces. Similarly, implementations of new tree models might also require new infrastructure, e.g., for multi-way splits or more complex models in the leafs. To overcome these difficulties, the partykit package offers a unified representation of tree objects along with predict(), print(), and plot() methods. Trees are represented through a new flexible class "party" which can, in principle, capture all trees mentioned above. The package is currently under development at R-Forge and already provides conversion methods for trees of classes rpart, J48, and pmmlTreeModel as well as a re-implementation of conditional inference trees. Furthermore, the new packages evtree and CHAID employ the partykit infrastructure.

E510: **Detecting threshold interactions in supervised classification and regression: STIMA**
Presenter: **Claudio Conversano**, University of Cagliari, Italy

Simultaneous Threshold Interaction Modeling Algorithm (STIMA) has been implemented in the R environment as a tool enabling us to automat-