

PROGRAMME AND ABSTRACTS

6th CSDA International Conference on
Computational and Financial Econometrics (CFE 2012)

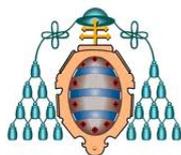
<http://www.cfe-csda.org/cfe12>

and

5th International Conference of the
ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on
Computing & Statistics (ERCIM 2012)

<http://www.cfe-csda.org/ercim12>

Conference Center “Ciudad de Oviedo”, Spain
1-3 December 2012



UNIVERSIDAD DE OVIEDO

<http://www.uniovi.es>



<http://www.qmul.ac.uk>

Monday 03.12.2012

17:10 - 18:10

Parallel Session S – ERCIM

ES83 Room 13 HEALTH DATA AND MEDICAL STATISTICS**Chair: Peter Congdon****E768: Statistical estimation problems in causal inference: Equating propensity scores***Presenter:* **Priyantha Wijayatunga**, Umea University, Sweden

In the statistical literature, particularly in medical domains, causal inference problems for observational data became popular with the emergence of such models as probabilistic graphical models and potential outcome based methods. However researchers using these two techniques are opposing to each other. The equivalence of the two methods is shown along with their need of addressing the same estimation problems. In the potential outcome framework attention is paid to the estimation of the so-called propensity scores and an efficient stratification of data sample using them for estimation of treatment effects. This is usually done by equating some of them together, so that each newly formed stratum has almost the same covariate distribution for both treated and non-treated groups where the covariates are the observed confounders of the treatment and outcome variables. Since in medical domains relatively large databases of observational data are common, using probabilistic arguments it is shown how one could equate propensity scores to preserve the requirement. These score ranges depend on marginal probabilities of covariates too. This raises questions on the estimation of propensity scores using discriminative methods such as logistic regression. It is argued that generative estimation methods of probabilities are needed to preserve required balance within strata.

E928: The effect of aggregation on disease mapping*Presenter:* **Caroline Jeffery**, Liverpool School of Tropical Medicine, United Kingdom*Co-authors:* Al Ozonoff, Marcello Pagano

In public health, studying the relationship between an individual's location and the acquisition of disease can serve to prevent further spread of disease or to guide (re)allocation of resources to improve access to care. In health data, spatial information on cases is either available in point form (e.g. longitude and latitude) or aggregated by an administrative region. Statistical methods developed for spatial data can accommodate either form of data, but not always both. In the case of disease mapping, point data or centroids of aggregated regions can serve as spatial location and produce a smoothed map of estimated risk. However the quality of the mapping is affected by how coarse is the resolution of the spatial information. Previous literature has shown that for cluster-detection methods, power tends to decrease as the spatial information on cases becomes coarser. We study the effect of aggregation on a disease risk mapping method, when the method is used to locate an increase in occurrence of cases in one subregion of the study area. Our simulations in the unit disk show that the accuracy of the mapping diminishes as the resolution of spatial information gets coarser.

E983: Exploring the correlation structure of inter-nucleotide DNA distances*Presenter:* **Sonia Gouveia**, IEETA/UA, Portugal*Co-authors:* Vera Afreixo, Manuel Scotto, Paulo Ferreira

DNA is a long sequence of repeated symbols called nucleotides (A, C, G and T), from which four series of inter-nucleotide distances (IND) are obtained from the consecutive distances between equal nucleotides. Previously, the distributions of IND values were shown to exhibit significant differences with the counterpart geometric reference distributions, which would be obtained if nucleotides were placed randomly and independently. In this work, the goal is to explore the possibility that these differences are due to the IND autocorrelation structure. A simulation study was designed from real data `gij33286443jrefjNM-032427.1` gene of *Homo sapiens* (<ftp://ftp.ncbi.nih.gov/genomes/>) processed in blocks of symbols. Each block was transformed into a binary sequence (0 or 1, according to the nucleotide) and its autocorrelation function was estimated. Afterwards, the binary sequence was scrambled to obtain a random Bernoulli variable with the same success probability and no autocorrelation. This random variable was then colored (with proper filtering/quantization), mimicking the original autocorrelation structure up to an order p . The optimal order was chosen as that leading to non-significant differences between original and colored IND distributions, assessed by chi-square testing adjusted for sample size. This study focuses on the interpretation of the optimal order changes along the gene sequence.

ES90 Room 5 DESIGNS AND SURVEYS**Chair: Kalliopi Mylona****E765: Survey estimates by calibration on dual frames***Presenter:* **Maria del Mar Rueda**, Universidad de Granada, Spain*Co-authors:* Antonio Arcos, Maria Giovanna Ranalli, Annalisa Teodoro

Survey statisticians make use of the available auxiliary information to improve estimates. One important example is given by calibration estimation, that seeks for new weights that are as close as possible to the basic design weights and that, at the same time, match benchmark constraints on available auxiliary information. Recently, multiple frame surveys have gained much attention and became largely used by statistical agencies and private organizations to decrease sampling costs or to reduce frame undercoverage errors that could occur with the use of only a single sampling frame. Much attention has been devoted to the introduction of different ways of combining estimates coming from the different frames. We will extend the calibration paradigm to the estimation of the total of a variable of interest in dual frame surveys as a general tool to include auxiliary information, also available at different levels. In fact, calibration allows us to handle different types of auxiliary information and can be shown to encompass as a special case the raking ratio method and the pseudo empirical maximum likelihood approach.

E650: Optimal experimental designs for conjoint analysis: Estimation of utility functions*Presenter:* **Jose M. Vidal-Sanz**, Universidad Carlos III de Madrid, Spain*Co-authors:* Mercedes Esteban-Bravo, Agata Leszkiewicz

In conjoint analysis consumers utility functions over multiattributed stimuli are estimated using experimental data. The quality of these estimations heavily depends on the alternatives presented in the experiment. An efficient selection of the experiment design matrix allows more information to be elicited about consumer preferences from a small number of questions, thus reducing experimental cost and respondent's fatigue. Kiefer's methodology considers approximate optimal design selecting the same combination of stimuli more than once. In the context of conjoint analysis, replications do not make sense for individual respondents. We present a general approach to compute optimal designs for conjoint experiments in a variety of scenarios and methodologies: continuous, discrete and mixed attributes types, customer panels with random effects, and quantile regression models. We do not compute good designs, but the best ones according to the size (determinant or trace) of the information matrix of the associated estimators without repeating profiles as in Kiefer's methodology. We use efficient optimization algorithms to achieve our goal.

E985: D-optimal two-level factorial designs for logistic regression models with bounded design regions*Presenter:* **Roberto Dorta-Guerra**, Universidad de La Laguna, Spain*Co-authors:* Enrique Gonzalez-Davila, Josep Ginebra

Under first order normal linear models, the amount of information gathered through two-level factorial experiments, as measured through the determinant of their information matrix, does neither depend on where the experiment is centered, nor on how it is oriented relative to the contour lines of the surface, and balanced allocations are always more informative than unbalanced ones with the same number of runs. Thus, when planning