# PROGRAMME AND ABSTRACTS

9th International Conference on

## Computational and Financial Econometrics (CFE 2015)

http://www.cfenetwork.org/CFE2015

and

8th International Conference of the
ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on

## Computational and Methodological Statistics (CMStatistics 2015)

http://www.cmstatistics.org/CMStatistics2015

Senate House & Birkbeck University of London, UK
12 – 14 December 2015








**ISBN 978-9963-2227-0-4**

Technical Editors: Angela Blanco-Fernandez and Gil Gonzalez-Rodriguez.

## Contents

EO1330: **Spatial clustering of time-series via mixtures of autoregressive models and Markov random fields for image analysis**
*Presenter:* **Hien Nguyen**, University of Queensland, Australia
*Co-authors:* Geoffrey McLachlan, Jeremy Ullmann, Andrew Janke
Time-series data arise in many medical and biological imaging scenarios. In such images, a time-series is obtained at each of a large number of spatially dependent data units (e.g. electrodes, pixels, or voxels). It is often interesting to organize these data units into clusters that are modeled by an underlying probabilistic process. A two-stage procedure is presented for this task. In Stage 1, a mixture of autoregressive (MoAR) model is used to marginally cluster the time series arising at each data unit. The MoAR model is fitted using maximum marginal likelihood (MML) estimation via an MM (minorizationmaximization) algorithm. In Stage 2, a Markov random field (MRF) model is used to induce a spatial dependency structure onto the Stage 1 clustering. The MRF model is also fitted using maximum pseudolikelihood (MPL) estimation via an MM algorithm. A simulation study is presented to demonstrate the performance of the two-stage procedure. An application to the segmentation of a zebrafish brain calcium image is presented as a demonstration of the methodology.

EC1246: **Robust estimation for mixtures of skew data**
*Presenter:* **Francesca Greselin**, University of Milano Bicocca, Italy
*Co-authors:* Luis Angel Garcia-Escudero, Agustin Mayo-Iscar, Geoffrey McLachlan
Recently, observed departures from the classical Gaussian mixture model in real datasets have led to the introduction of more flexible tools for modeling heterogeneous skew data. Among the latest proposals in the literature, we consider mixtures of skew normal, to incorporate asymmetry in components, as well as mixtures of $t$, to down-weight the contribution of extremal observations. Clearly, mixtures of skew $t$ have widened the application of model based clustering and classification to great many real datasets, as they can adapt to both asymmetry and leptokurtosis in the grouped data. Unfortunately, when data contamination occurs far from the bulk of the data, or even between the groups, classical inference for these models is not reliable. Our proposal is to address robust estimation of mixtures of skew normal, to resist sparse outliers and even pointwise contamination that could arise in data collection. We introduce a constructive way to obtain a robust estimator for the mixture of skew normal model, by incorporating impartial trimming and constraints in the EM algorithm. At each E-step, a low percentage of less plausible observations, under the estimated model, is tentatively trimmed; at the M-step, constraints on the scatter matrices are employed to avoid singularities and reduce spurious maximizers. Some applications on artificial and real data show the effectiveness of our proposal, and the joint role of trimming and constraints to achieve robustness.

## EO304 Room MAL 532 HIGH-DIMENSIONAL DATA ANALYSIS BEYOND LINEAR MODELS Chair: Junhui Wang

EO1303: **Probability-enhanced sufficient dimension reduction for binary classification**
*Presenter:* **Hao Zhang**, University of Arizona, United States
Many sufficient dimension reduction (SDR) methods have been developed since the introduction of sliced inverse regression. For binary classification problems, SIR suffers the limitation of estimating at most one direction since only two slices are available. We propose a new and flexible probability-enhanced SDR method for binary classification problems using the weighted support vector machine (WSVM). The key idea is to slice the data based on conditional class probabilities of observations rather than their binary responses. We show that the central subspace based on the conditional class probability is the same as that based on the raw binary response, which justifies the proposed slicing scheme and assures no information loss. Furthermore, in order to implement the new slicing scheme, one does not need exact probability values since the only required information is the relative ordering of probability values. The new SDR bypasses the probability estimation and employs the WSVM to directly estimate the order of probability values, based on which the slicing is performed. The performance of the proposed probability-enhanced SDR scheme is evaluated by both simulated and real data examples.

EO0634: **Sparse quadratic discriminant analysis**
*Presenter:* **Chenlei Leng**, Warwick, United Kingdom
A novel QUadratic Discriminant Analysis procedure called QUDA is proposed for analysing high-dimensional data. The proposed method is able to identify quadratic interactions of the variables for classification. Under appropriate sparsity assumptions, we show that QUDA works even when the dimensionality is exponentially high with respect to the sample size. We develop an efficient algorithm based on the alternating direction method of multipliers method (ADMM) for finding interactions along the way, which is much faster than its competitor in the literature. The competitive performance of QUDA is illustrated via extensive simulation studies and the analysis of real datasets.

EO0217: **Sparse partially linear additive models**
*Presenter:* **Jacob Bien**, Cornell University, United States
*Co-authors:* Yin Lou, Rich Caruana, Johannes Gehrke
The generalized partially linear additive model (GPLAM) is a flexible and interpretable approach to building predictive models. It combines features in an additive manner, allowing each to have either a linear or nonlinear effect on the response. However, the choice of which features to treat as linear or nonlinear is typically assumed known. Thus, to make a GPLAM a viable approach in situations in which little is known a priori about the features, one must overcome two primary model selection challenges: deciding which features to include in the model and determining which of these features to treat nonlinearly. We introduce the sparse partially linear additive model (SPLAM), which combines model fitting and both of these model selection challenges into a single convex optimization problem. SPLAM provides a bridge between the lasso and sparse additive models. Through a statistical oracle inequality and thorough simulation, we demonstrate that SPLAM can outperform other methods across a broad spectrum of statistical regimes, including the high-dimensional setting. We develop efficient algorithms that are applied to real data sets with half a million samples and over 45,000 features with excellent predictive performance.

## EG067 Room MAL 402 CONTRIBUTIONS ON COMPLEX DATA Chair: Thaddeus Tarpey

EC1635: **Rate of uniform consistency for a class of mode regression on functional stationary ergodic data**
*Presenter:* **Mohamed Chaouch**, United Arab Emirates University, United Arab Emirates
*Co-authors:* Naamane Laib, Djamal Louani
The aim is to study the asymptotic properties of a class of kernel conditional mode estimates whenever functional stationary ergodic data are considered. To be more precise, in the ergodic data setting, we consider a random element $(X;Z)$ taking values in some semi-metric abstract space $E \times F$. For a real function $\varphi$ defined on the space F and $x \in E$, we consider the conditional mode of the real random variable $\varphi(Z)$ given the event $X = x$. While estimating the conditional mode function, say $\theta_\varphi(x)$, using the well-known kernel estimator, we establish the strong consistency with rate of this estimate uniformly over Vapnik-Chervonenkis classes of functions $\varphi$. Notice that the ergodic setting offers a more general framework than the usual mixing structure. Two applications to energy data are provided to illustrate some examples of the proposed approach in time series forecasting framework. The first one consists in forecasting the daily peak of electricity demand in France (measured in Giga-Watt). Whereas the second one deals with the short-term forecasting of the electrical energy (measured in Giga-Watt per Hour) that may be consumed over some time intervals that cover the peak demand.