AN INITIATIVE OF THE PROJECT PRIN 2008AHWTJ4
SUPPORTED BY MIUR – ITALY

WORKSHOP

NEW ROBUST METHODS FOR THE ANALYSIS OF COMPLEX DATA

SEGIS Department
University of Sannio
Benevento, 10-13 September 2012

The workshop is organized by the PRIN 2008 research project on *NEW ROBUST METHODS FOR THE ANALYSIS OF COMPLEX DATA*. It is also financially supported by the University of Sannio

Project Partners

University of Parma
University of Sannio
University of Verona

Monday, 10th September 2012

Chairperson	Domenico Piccolo	
	9:30 – 10:15	Elvezio Ronchetti
	10:15 – 11:00	Stefan Van Aelst
	11:00 - 11:30	***Coffee break***
Chairperson	Marco Riani	
	11:30 – 12:00	Luca Greco
	12:00 – 12:30	Bruno Scarpa

Elvezio Ronchetti

Robust Inference for Marginal Longitudinal Generalized Linear Models

Longitudinal models are commonly used for studying data collected on individuals repeatedly through time and classical statistical methods are readily available to carry out estimation and inference.
However, in the presence of small deviations from the assumed model, these techniques can lead to biased estimates, p-values, and confidence intervals. Robust statistics deals with this problem and develops techniques that are not unduly influenced by such deviations.
In this talk we first review several robust estimators for marginal longitudinal GLM which have been proposed in the literature together with the corresponding robust inferential procedures. Then we discuss robust variable selection procedures (including a generalized version of Mallows's Cp) and we examine their performance in the longitudinal setup.
Finally, longitudinal data typically derive from medical or other large-scale studies where often large numbers of potential explanatory variables and hence even larger numbers of candidate models must be considered. In this case we discuss a cross-validation Markov Chain Monte Carlo procedure as a general variable selection tool which avoids the need to visit all candidate models. Inclusion of a ``one-standard error'' rule provides users with a collection of good models.

Stefan Van Aelst

Robust ANOVA tests for linear models

ANOVA tests are the standard tests to compare nested linear models fitted by least squares. These tests are equivalent to likelihood ratio tests and thus they are very powerful. However, least squares estimators are very vulnerable to outliers, and thus also the related ANOVA type tests are also extremely sensitive to outliers. Therefore, we consider regression tau-estimators to estimate the parameters in the linear models. Regression tau-estimators combine high robustness with high efficiency which makes them a suitable choice for robust inference beyond parameter estimation. In this talk we introduce robust likelihood ratio type test statistics based on the tau-estimates of the error scale in the linear models. To determine the null distribution of the test statistics we either use an asymptotic approximation or the fast and robust bootstrap. We study

the robustness and power of the resulting robust likelihood ratio type tests for nested linear models.

Luca Greco

S-estimation of Hidden Markov Models

A method for robust estimation of dynamic mixtures of multivariate distributions is proposed, with attention devoted to hidden Markov models.
The estimation problem is often addressed by maximum likelihood and the EM algorithm. This methodology is particularly sensitive to the occurrence of outliers, anomalous values that deviate from the general pattern of the bulk of the data.
Outliers may break down component specific parameter estimates with a consequent bias in the estimate of location and an inflation in the estimate of scatter. The EM algorithm is modified by replacing the classical M-step with high breakdown S-estimation of location and scatter. In particular, the bisquare multivariate S-estimator is considered. The estimates are computed by solving a system of estimating equations that are characterized by component specific sets of weights, based on the Mahalanobis distances.

Bruno Scarpa

Enriched Stick Breaking Processes for Functional Data

In many applications involving functional data, prior information is available about the proportion of curves having different attributes. It is not straightforward to include such information in existing procedures for functional data analysis. Generalizing the functional Dirichlet process (FDP), we propose a class of stick-breaking priors for distributions of functions.
These priors incorporate functional atoms drawn from a Gaussian process.
The stick-breaking weights are specified to allow user-specified prior probabilities for curve attributes, with hyperpriors accommodating uncertainty.
Compared with the FDP, the random distribution is enriched for curves having attributes known to be common. Theoretical properties are considered, methods are developed for posterior computation, and the approach is illustrated using data on temperature curves in menstrual cycles.

Tuesday, 11th September 2012

Chairperson	Elvezio Ronchetti	
	9:30 – 10:15	Anthony Atkinson
	10:15 – 11:00	Marco Riani
	11:00 - 11:30	***Coffee break***
Chairperson	Stefan Van Aelst	
	11:30 – 12:00	Aldo Corbellini
	12:00 – 12:30	Tiziano Bellini
	12:30 – 13:00	Domenico Perrotta

ANTHONY ATKINSON

A Framework for Comparing Methods of Very Robust Regression

Methods for regression are often compared using the breakdown point, which is the minimumamount of contamination which can cause the parameter estimates to become unbounded. For least squares this is asymptotically (as *n* gets large) zero as the outliers themselves go to infinity. The very robust estimators which are the subject of the talk all, on the other hand, have an asymptotic breakdown point of 50%. However, the definition of breakdown point describes behaviour for very remote contamination. In fact, the properties of very robust regression estimators depends on the distance between the regression data and the outliers. It is hard to be precise about the nature of the dependence. In order to give structure to the problem we introduce a parameter λ that defines a parametric path in the space of models and enables us to study, in a systematic way, the properties of estimators as the groups of data move from being far apart to close together. We examine, as a function of λ, the variance and squared bias of five estimators and we also consider their power when used in the detection of outliers. This systematic approach provides tools for gaining knowledge and better understanding of the properties of robust estimators.

Our motivating example comes from trade data when there may be a mixture of regression models. So an alternative to very robust fitting of one model is the clustering of regression models. However, as a first stage, we do require that the single fitted model reveal outliers and structural inadequacies.

Marco Riani

Robust Regression analysis with partially labelled regressors: carbon dating of the Shroud of Turin

The twelve results from the 1988 radio carbon dating of the Shroud of Turin show surprising heterogeneity.
We try to explain this lack of homogeneity by regression on spatial coordinates. However, although the locations of the samples sent to the three laboratories involved are known, the locations of the 12 subsamples within these samples are not. We consider all 387,072 plausible spatial allocations and analyse the resulting distributions of statistics. Plots of robust regression residuals from the forward search indicate that some sets of allocations are implausible. We establish the existence of a trend in the results and suggest how better experimental design would have enabled stronger conclusions to have been drawn from this multi-centre experiment.

Aldo Corbellini

Robust Bayesian Methods for Multiple Outliers Detection

In this paper we show how the application of the bayesian framework to the Forward search procedure enables us to detect groups of multiple outliers.
We first outline the Bayesian normal regression model and describe how, by means of Gibbs Sampling, a Monte Carlo class algorithm, we are able to simulate from the joint distribution of regression parameters and error variance for the predictive distribution of future observations.
Zellner, first, proposed a simple way of inputting prior information in a regression model, through Zellner's class of g priors which takes into account the variance of the independent variables, Agliari and Parisetti, later, proposed a normal coniugate informative called A-prior which takes into account different degrees of certainty about the different independent variables.
The crucial idea of the forward search is to monitor how the fitted model changes whenever a new statistical unit is added to the subset and we do this by measuring the minimum deletion residual (MDR) among the units not belonging to the subset. With our flexible trimming we can appraise the effect that each statistical unit (once it is introduced into the subset), outlier or not, exerts on the fitted model.

Tiziano Bellini

Robust Pairs Trading

Pairs trading has become a popular trading strategy in the financial industry. The main objective of pairs trading is to exploit the correlation of security movements in order to profit on their mispricing. A long-run equilibrium price relationship is then estimated for the identified trading pairs, and the resulting mean-reverting residual spread is modeled. When one security is traded up and the other is traded down, the trader sells the outperforming security and buys the underperforming ones. We exploit the cointegration approach in an attempt to model pairs trading. We exploit the Johansen test for

cointegration to select trading pairs to be used within a pairs trading framework. A long-run equilibrium price relationship is then estimated for the identified trading pairs, and the resulting mean-reverting residual spread is modeled as a Vector-Error-Correction model (VECM). It is evident that the presence of atypical observations affects the whole analysis. Thus we introduce robust techniques in order to evaluate the impact of outliers on parameter estimates and check the effectiveness of pairs trading strategies.

Domenico Perrotta

FSDA: A MATLAB toolbox for robust analysis and interactive data exploration

We present the FSDA (Forward Search for Data Analysis) toolbox, a new software library that extends MATLAB and its Statistics Toolbox to support a robust and efficient analysis of complex datasets, affected by different sources of heterogeneity. As the name of the library indicates, the project was born around the Forward Search approach, but it has evolved to include the main traditional robust multivariate and regression techniques, including LMS, LTS, MCD, MVE, MM and S estimation. To address problems where data deviate from typical model assumptions, tools are available for robust data transformation and robust model selection. When different views of the data are available, e.g. a scatterplot of units and a plot of distances of such units from a fitted model, FSDA links such views and offers the possibility to interact with them. For example, selections of objects in a plot are highlighted in the other plots. This considerably simplifies the exploration of the data in view of extracting information and detecting patterns.

Wednesday, 12th September 2012

Chairperson	Masanobu Taniguchi	
	9:30 – 10:15	Thomas J. DiCiccio
	10:15 – 11:00	Mike Waterson
	11:00 - 11:30	***Coffee break***
Chairperson	Mike Waterson	
	11:30 – 12:00	Angelica Gianfreda
	12:00 – 12:30	Gianluca Morelli
	12:30 – 13:00	Francesca Torti

Thomas J. DiCiccio

Adjusted nonparametric profile likelihood

As is the case in the parametric context, a nonparametric profile likelihood is not a genuine likelihood. In particular, the expectation of the profile score is not identically zero; it does not even vanish asymptotically.
We propose adjusted versions of nonparametric profile likelihood, based on an asymptotic expansion of the expectation of the profile score. Use of the adjusted profile likelihood is investigated in examples from two directions: the effect on estimators, and the accuracy of asymptotic approximations to the distributions of test statistics. Cases where the appropriate estimator is understood suggest that the adjustments do indeed produce the desired result. However, by contrast to what is seen in the parametric context, the asymptotic approximation to the distribution of the likelihood ratio test statistic based on adjusted nonparametric profile likelihood can be inadequate. Therefore, bootstrap calibration may be necessary to replace the asymptotic distribution. We attempt to clarify the gains to be made from bootstrapping the adjusted nonparametric profile likelihood ratio statistic rather than bootstrapping the unadjusted one, which is already known to provide very accurate inference.

Mike Waterson

Making use of events to analyse energy markets

Exogenous events can provide a very useful source of insight into otherwise complex problems of assessing economic impact. In this presentation, I discuss three examples. The first is the impact of what are called the New Electricity Trading Arrangements in Great Britain, where the electricity pool was superseded by bilateral bargaining. Here, the exogenous factor is that Scotland had the change later than England and Wales, so it provides a good counterfactual. The second is the value of storage, in this case gas storage. Here, a significant unexpected fire which dramatically reduced storage to around 20% of its previous level is used to analyse the value of storage as revealed in the impact on the market over the relevant period.. The third is the Fukushima earthquake. The impact in question is on European generation companies, through the policy fallout of increased constraints on European nuclear plant and the impact on profits and other key variables such as carbon prices.

ANGELICA GIANFREDA

Forecasting Italian Electricity Zonal Prices with Exogenous Variables

In the last few years we have observed deregulation in electricity markets and an increasing interest of price dynamics has been developed especially to consider all stylized facts shown by spot prices. Only few papers have considered the Italian Electricity Spot market since it has been deregulated recently.
Therefore, this contribution is an investigation with emphasis on price dynamics accounting for technologies, market concentration, congestions and volumes.
We aim to understand how these factors affect the zonal prices since these ones combine to bring about the single national price (prezzo unico d'acquisto, PUN). Hence, understanding its features is important for drawing policy indications referred to production planning and selection of generation sources, pricing and risk–hedging problems, monitoring of market power positions and finally to motivate investment strategies in new power plants and grid interconnections. Implementing Reg–ARFIMA–GARCH models, we assess the forecasting performance of selected models showing that they perform better when these factors are considered.

Gianluca Morelli

Comparison of robust clustering methods for optimal units allocation

The units allocation for n-groups clustering relies on arbitrary choices of both parameters and algorithms. So the units allocation depends by several subjective choices. These choices could give different results even when analyzing the same dataset. To avoid those troubles it is appropriate to obtain a classification method based on data structure such that the consequence of the arbitrary choices is negligible. In this work we test several classification methods on two different dataset. The first one composed by categorical data with finite support, and the second one composed by continuous data with strong outliers. The aim of this work is to present a comparison between different methods to underline, if it exist surprisingly, the stronger classification method, based on data structure, to get an optimal allocation for each dataset. To achieve this target we compare existing methods with new "forward-search" based classifiers which have shown high efficiency in many simulations performed so far.

FRANCESCA TORTI

FSDA in action

The FSDA (Forward Search for Data Analysis) is a new software library that extends MATLAB and its Statistics Toolbox to the robust analysis of complex datasets. This contribution complements the overview given by the same authors in the presentation 'FSDA: A MATLAB toolbox for robust analysis and interactive data exploration'. Here,

the focus is on the FSDA application on simulated as well as real datasets, in multivariate and regression contexts. FSDA will be concretely demonstrated, possibly using selected datasets from the audience.

Thursady, 13th September 2012

Chairperson	Anthony Atkinson	
	9:30 – 10:15	Masanobu Taniguchi
	10:15 – 11:00	Siem Jan Koopman
	11:00 - 11:30	***Coffee break***
Chairperson	Siem Jan Koopman	
	11:30 – 12:00	Lisa Crosato
	12:00 – 12:30	Fabrizio Laurini

MASANOBU TANIGUCHI

Robust portfolio estimation under skew-normal return processes

In the theory of portfolio analysis, optimal portfolios are determined by the mean and variance of the portfolio return. Several authors proposed estimators of the optimal portfolios as functions of the sample mean and the sample variance for independent returns of assets. However, empirical studies show that financial return processes are often dependent. Shiraishi and Taniguchi (2007) discussed the asymptotic efficiency for optimal portfolios when returns are non-Gaussian stationary processes.

For cases where the assumption of normality is not tenable, more flexible models can be adopted to accommodate skewness and heavy tails which are introduced by DiCiccio and Monti (2004). Flexible models that include the normal distribution as a special case are especially important, because they allow continuous variation from normality to nonnormality. Thus such models can adapt to distributions that are in a neighborhood of the normal model. Using a flexible model to handle nonnormality has certain benefits. When the deviation from normality involves skewness, a model is needed to establish meaningful location and scale parameters.

The skew-normal distribution is a family of distributions including the normal one, but with an extra parameter to regulate skewness. It allows a continuous variation from normality to non-normality, which is useful in many situations. The class of the multivariate skew-normal distributions represents a mathematically tractable extension of the multivariate normal distribution with addition of a vector parameter to regulate skewness.

In this talk, we discuss influence of skewness parameter of skew-normal distribution for return processes and the central sequence of the local asymptotic normality (LAN) of the family of probability distribution of observations. Based on the asymptotic distribution of portfolio estimators for Non-Gaussian dependent return process, we evaluate an influence of on the asymptotic variance of . From this, we can see robustness and sensitivity of portfolio estimator, and provide numerical examples, which show some interesting features of the influence. Next, we discuss an influence IMC of skewness on the central sequence of LAN assuming that mean vector and coefficient matrices which are specied by unknown parameter. Here we use some fundamental LAN results by Taniguchi and Kakizawa (2000) and Shiraishi and Taniguchi (2007). We also provide numerical examples for IMC.

The talk is organized as follows. In Section 2, we derive a class of linear processes, optimal portfolios function and the multivariate skew-normal distribution. Also we evaluate the asmyptotic distribution of the estimated portfolio when the return process is a linear process generated by skew-normal innovations. In Section 3, we evaluate robustness and sensitivity of asymptotic variance of portfolio estimators with respect to skew-normal perturbation. From the results, we examine the sensitiveness numerically. In Section 4, we discuss an influence of skewness on the central sequence of the local asymptotic normal (LAN) property.

Siem Jan Koopman

Long Memory Dynamics for Multivariate Dependence under Heavy Tails

We develop a new simultaneous time series model for volatility and dependence with long memory (fractionally integrated) dynamics and heavy-tailed densities. Our new multivariate model accounts for typical empirical features in financial time series while being robust to outliers or jumps in the data. In the empirical study for four Dow Jones equities, we find that the degree of memory in the volatilities of the equity return series is similar, while the degree of memory in correlations between the series varies significantly. The forecasts from our model are compared with high-frequency realised volatility and dependence measures. The forecast accuracy is overall higher compared to those from some well-known competing benchmark models.

Lisa Crosato

Correcting outliers in Garch models: a weighted forward approach

The purpose of this paper is to apply the forward search (Atkinson and Riani, 2000) to GARCH models to order observations according to their degree of agreement with the model. An exact application of the forward search to time series would lead to artificial gaps in the natural time ordering of the data. Rather, the autocorrelation structure characterizing time series data must be taken into account. In order to solve this problem, Grossi (2004) proposed to substitute observations missing at a given step of the search, and covering the time span between observations included in the search, with data simulated according to the GARCH model estimated at the previous step. In this work we suggest instead to fill the gaps in the series created by the forward ordering with the original observations weighed according to their degree of outlyingness established by the GARCH model estimated at the previous step. The weighs can then be used to correct detected outliers and proceed to robust estimates.

Fabrizio Laurini

Robust estimation of CVaR for optimal asset allocation of shares

In the classical mean-variance approach to portfolio selection, estimates of the expected returns and the covariance matrix provide the optimal asset allocation under Gaussian assumptions. However, it is well known that asset returns are not normal. Therefore, the mean and the variance alone do not fully describe the characteristics of the joint asset distribution. As a consequence, especially in cases of strong non normality, the classical

mean-variance approach will not be a satisfactory portfolio allocation model. Among the reasons of this drawback a relevant role is played by the influence of extreme returns which have strong influence on the estimation of the mean vector and covariance matrix. Another criticism which is commonly made to the mean-variance is the use of the historical standard deviation as a measure of risk. Several alternative measures of risk have been proposed. Among these alternatives, the value of risk (VaR) and the expected shortfall, which is also known as
conditional value at risk (CVaR), have become increasingly popular. Some of these measures are more appropriate than sample variance. The computation of optimal portfolios based on these measures does not even require the return covariance matrix, although being quantile-based estimation, the influence of extremes is still very relevant.

In this paper we discuss the problem of statistical robustness of optimization methods based on Spectral Risk Measures and show that the latter are not robust, meaning that a few extreme assets prices or returns can lead to ``sub-optimal'' portfolios. We then introduce a robust estimator based on the forward search in the maximization procedure and show that it is far more stable than the classical version based on maximum likelihood estimator and more efficient than existing robust estimators.

List of participants

Anthony ATKINSON
London School of Economics, UK.
a.c.atkinson@lse.ac.uk

Tiziano BELLINI
University of Parma, Italy
tiziano.bellini@prometeia.com

Aldo CORBELLINI
University of Parma, Italy
aldo@netline.it

Lisa CROSATO
University Milano-Bicocca, Italy
lisa.crosato@unimib.it

Thomas J. DICICCIO
Cornell University, New York, USA
tjd9@cornell.edu

Angelica GIANFREDA
European University Institute
angelica.gianfreda@univr.it

Luca GRECO
University of Sannio
luca.greco@unisannio.it

Luigi GROSSI
University of Verona, Italy
luigi.grossi@univr.it

Siem Jan KOOPMAN
University of Amsterdam, Netherlands
s.j.koopman@feweb.vu.nl

Fabrizio LAURINI
University of Parma, Italy
fabrizio.laurini@unipr.it

Anna Clara MONTI
University of Sannio, Italy
acmonti@unisannio.it

Gianluca MORELLI
University of Parma, Italy
gianluca.morelli@unipr.it

Domenico PERROTTA
European Commission, Joint Research Centre, Ispra, Italy
Domenico.Perrotta@ec.europa.eu

Domenico PICCOLO
University of Naples Federico II, Italy
domenico.piccolo@unina.it

Elvezio RONCHETTI
University of Geneva, Switzerland
Elvezio.Ronchetti@unige.ch

Bruno SCARPA
University of Padua, Italy
Bruno.scarpa@unipd.it

Masanobu TANIGUCHI
Waseda University, Japan
taniguchi@waseda.jp

Francesca TORTI
University of Parma, Italy
francesca.torti@nemo.unipr.it

Stefan VAN AELST
Ghent University, Belgium
Stefan.VanAelst@UGent.be

Mike WATERSON
University of Warwick, UK
michael.waterson@warwick.ac.uk