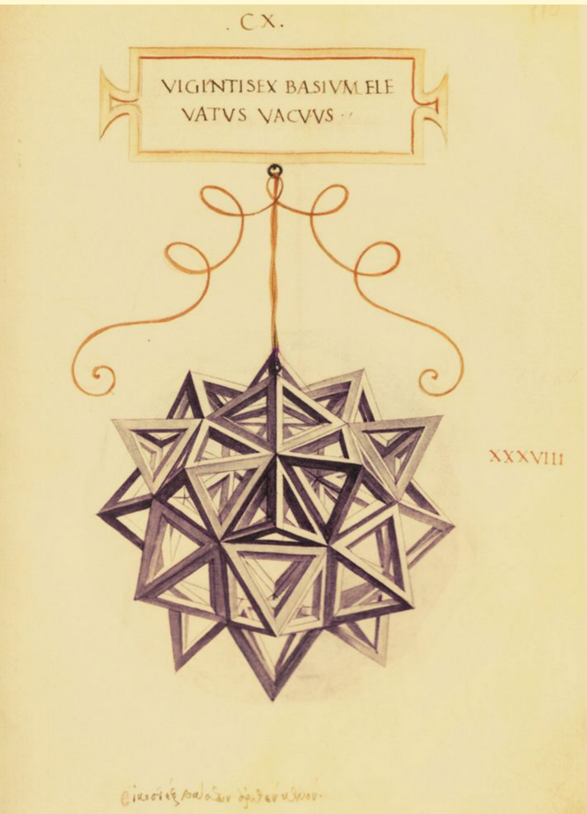

CLADAG 2021

ABSTRACTS

13-th Scientific Meeting Classification and Data Analysis Group

Firenze, September 9-11, 2021

VERIDICAL DATA SCIENCE: THE PRACTICE OF RESPONSIBLE DATA ANALYSIS AND DECISION-MAKING

Bin Yu[1]

[1] Departments of Statistics, and Electrical Engineering and Computer Sciences, UC Berkeley
(email: binyu@berkeley.edu)

> "A.I. is like nuclear energy -- both promising and dangerous"
>
> Bill Gates, 2019.

ABSTRACT: Data Science is a pillar of A.I. and has driven most of recent cutting-edge discoveries in biomedical research. In practice, Data Science has a life cycle (DSLC) that includes problem formulation, data collection, data cleaning, modeling, result interpretation and the drawing of conclusions. Human judgement calls: wq:ware ubiquitous at every step of this process, e.g., in choosing data cleaning methods, predictive algorithms and data perturbations. Such judgment calls are often responsible for the "dangers" of A.I. To maximally mitigate these dangers, we developed a framework based on three core principles: Predictability, Computability and Stability (PCS). Through a workflow and documentation (in R Markdown or Jupyter Notebook) that allows one to manage the whole DSLC, the PCS framework unifies, streamlines and expands on the best practices of machine learning and statistics – bringing us a step forward towards veridical Data Science.

In this lecture, we will illustrate the PCS framework through the development of iterative random forests for predictive and stable non-linear interaction discovery and that of epiTree, a pipeline to discover epistasis interactions from genomics data. We will also briefly discuss two on-going PCS-driven software developments: VeridicalFlow and simChef for ease of PCS-compliant data analysis and data-driven simulations, respectively.

USING SUBSET LOG-LIKELIHOODS TO TRIM OUTLIERS IN GAUSSIAN MIXTURE MODELS

Katharine M. Clark[1] and Paul D. McNicholas[1]

[1] McMaster University, Canada
(email: paulmc@mcmaster.ca)

ABSTRACT: Mixtures of Gaussian distributions are a popular choice in model-based clustering. Outliers can affect parameters estimation and, as such, must be accounted for. Predicting the proportion of outliers correctly is paramount as it minimizes misclassification error. It is proved that, for a finite Gaussian mixture model, the log-likelihoods of the subset models are distributed according to a mixture of beta distributions. An algorithm is then proposed that predicts the proportion of outliers by measuring the adherence of a set of subset log-likelihoods to a beta mixture reference distribution. This algorithm removes the least likely points, which are deemed outliers, until model assumptions are met.

KEYWORDS: Clustering, Mixture models, Outliers

ADDITIVE BAYESIAN VARIABLE SELECTION UNDER CENSORING AND MISSPECIFICATION

F. Javier Rubio[1]

[1] Department of Statistical Science, University College London
(email: f.j.rubio@ucl.ac.uk)

ABSTRACT: We discuss the role of misspecification and censoring on Bayesian model selection in the contexts of right-censored survival and concave log likelihood regression. Misspecification includes wrongly assuming the censoring mechanism to be non-informative. Emphasis is placed on additive accelerated failure time, Cox proportional hazards and probit models. We offer a theoretical treatment that includes local and non-local priors, and a general non-linear effect decomposition to improve power-sparsity trade-offs. We discuss a fundamental question: what solution can one hope to obtain when (inevitably) models are misspecified, and how to interpret it? Asymptotically, covariates that do not have predictive power for neither the outcome nor (for survival data) censoring times, in the sense of reducing a likelihood associated loss, are discarded. Misspecification and censoring have an asymptotically negligible effect on false positives, but their impact on power is exponential. We show that it can be advantageous to consider simple models that are computationally practical yet attain good power to detect potentially complex effects, including the use of a finite-dimensional basis to detect truly non-parametric effects. We also discuss algorithms to capitalize on sufficient statistics and fast likelihood approximations for Gaussian-based survival and binary models. This methodology is implemented in the R package mombf.

KEYWORDS: Additive regression, Generalized additive model, Misspecification, Model selection, Survival

MODELING CLUSTERS OF CORPORATE DEFAULTS: REGIME-SWITCHING MODELS SIGNIFICANTLY REDUCE THE CONTAGION SOURCE

Bård Støve[1], Geir D. Berentsen[2], Jan Bulla[1] and Antonello Maruotti[1,3]

[1] Department of Mathematics, University of Bergen
[2] Department of Business and Management Science, NNH Norges Handelshøyskole
[3] Dipartimento di Giurisprudenza, Economia, Politica e Lingue Moderne, LUMSA, Roma
(email: Bard.Stove@uib.no)

ABSTRACT: In this paper, we report robust evidence that the process of corporate defaults is time-dependent and can be modelled by extending an autoregressive count time series model class via the introduction of regime-switching. That is, some of the parameters of the model depend on the regime of an unobserved Markov chain, capturing the model changes during clusters observed for count time series in corporate defaults. Thus, the process of corporate defaults is more dynamic than previously believed. Moreover, the contagion effect - that current defaults affect the probability of other firms defaulting in the future - is reduced compared to models without regime-switching, and is only present in one regime. A two-regime model drives the counts of monthly corporate defaults in the United States of America (USA). To estimate the model, we introduce a novel quasi-maximum likelihood estimator by adapting the extended Hamilton-Grey algorithm for the Poisson autoregressive model.

KEYWORDS: Extended Hamilton-Grey algorithm, Integer valued time series, Poisson autoregression, RS-INGARCHX