Research Proposals

PhD Course in Statistics

Department of Statistical Sciences

University of Padua, Italy
Title: Bayesian Expectiles Regression Trees.

Supervisor: Bernardi Mauro
Contact person: Mauro Bernardi <mauro.bernardi@unipd.it>

Short description: Random forests have been recently introduced in the machine learning literature by Breiman (2001) as a nonparametric technique for classification, clustering and regression. Their Bayesian counterparts, the Additive Regression Trees, Chipman et al. (2010), have been proven to accurately approximate the posterior distribution of the conditional mean regression of a response variable as a function of high-dimensional covariates. This research project aims to extend the Additive Regression Trees technique to learn about the whole conditional probability distribution of the response variable by means of the expectile regression. Similar to conditional quantiles, a series of conditional expectiles can summarize the relation between the explained variable Y and the predictors X. Newey and Powell (1987) showed that the conditional expectile can be estimated by Asymmetric Least Squares (ALS), for any confidence level. Bayesian Expectiles Regression Trees (BERT) is a nonparametric Bayesian regression technique that considers the joint estimation of multiple expectiles at different confidence levels. Since the likelihood function is not available when multiple quantiles are considered, the inferential procedure relies on Approximate Bayesian Inference (ABC), Beaumont et al. (2002). To this end, either ALS scores or the empirical likelihood methods can be exploited.

Duration: 6 months.

Research outcomes: 1 or 2 research papers.

Requirements: Bayesian methods, simulation methods and Markov chain Monte Carlo Methods. Knowledge of statistical software: R or Matlab.

Essential references


Title:
Small-Sample Confidence Intervals for the Reliability in a Bivariate Normal Distribution

Supervisors:
Alessandra R. Brazzale and Valentina Mameli
Contact person: Alessandra R. Brazzale <alessandra.brazzale@unipd.it>

Short description:
Inference concerning the strength-stress reliability $R=P(X>Y)$ of two random variables $X$ and $Y$ has been investigated by several authors in literature under various assumptions on the distribution of the variables $X$ and $Y$. Considerable attention has been devoted to the case of dependent variables. Different confidence sets have been proposed in the framework of the bivariate normal distribution. Barbiero (2010) reviews results on the estimation of $R$ in the bivariate normal case and compares three types of confidence sets. His study highlighted that the actual coverage probability of these intervals for small sample sizes is below the nominal level, although the coverage probability approaches the nominal level when the sample size increases. In terms of the average length, the three confidence sets show very similar results. The aim of this project is to use the recent developments in likelihood-based theory to obtain solutions which have third order of accuracy (Brazzale et al., 2007). In particular, we wish to provide confidence intervals for $R$ when $X$ and $Y$ follow the bivariate normal distribution. The proposed research path uses the theory developed in Fraser and Reid (1995) and Fraser et al. (1999), which involves the modified signed likelihood ratio. The project follows the pioneering work by Cortese and Ventura (2012) who construct confidence intervals for the reliability $R$ through third order theory in the case of independent normal variables.

Duration:
6 months if the candidate is familiar with the theory of higher order likelihood-based inference; 12 months otherwise

Research outcomes:
1. simulation study to compare the accuracy of the newly proposed solutions with the methods considered by Barbiero (2010)
2. re-analysis of the six river locations data set discussed in Barbiero (2010), and first presented in Devore (2003)
3. software implementation of the proposed methods and divulgence through an R package

Requirements:
✓ proficiency in large-sample likelihood asymptotics; possibile knowledge of small-sample asymptotics
✓ proven programming abilities especially in the numerical computing environment R (www.r-project.org)
✓ proficiency in spoken and written English

Essential references


Title:
Sex ratio at birth in Vietnam

Supervisors:
Maria Castiglioni and Fausta Ongaro
Contact person: Maria Castiglioni <maria.castiglioni@unipd.it>

Short description:
There has been increasing interest on the excess of young boys at birth in some Eastern and Southern Asian countries, often resulting from the abortion of female fetuses. The preference for sons has deep roots for cultural and economic reasons. For example, in India, once a daughter marries, she leaves the parental home to live with her in-laws. She is perceived to be of little economic benefit to her parents and will not support them in their old age.

The project aims to explore this issue in Vietnam in the light of the following research questions:
   a) Has Vietnam experienced in the last decades an excess of male births?
   b) Are there internal differences in sex ratio at birth by birth order, sex of previous children, geographical areas or by socio-economic family conditions?
   c) Are there institutional or cultural factors at the base of this behavior?
   d) How does it affect the population structure of Vietnam?

Duration:
Duration of the project (6-12 months).

Research outcomes:
The answers to the previous questions can be viewed as possible outcomes of the research.

Requirements:
Candidate are required to have interest in demographic topics, knowledge of mathematical or statistical methods able to analyse demographic data at the macro and at the micro level, familiarity with specific software like R, SAS, SPSS or STATA.

Essential references
• UNFPA, 2009, Recent change in sex ratio at birth in Viet Nam, A review of evidence.
• Arnold F, 1997, Gender preferences for children, DHS Comparative Studies, No.3
Title: Discrete choice analysis for the elicitation of graduates’ preferences for jobs

Supervisors: Luigi Fabbris and Manuela Scioni
Contact person: Luigi Fabbris <luigi.fabbris@unipd.it>

Short description: The research programme is aimed at gathering and analysing data about the criteria which early graduates adopt as guidance for choosing a job from those offered. The data collection will be carried out with a CAWI questionnaire designed to highlight what the graduates have in mind when they decide to either accept or refuse a job. An innovative questionnaire will be used to highlight the selection process of early graduates when they are offered a job within a given labour market. The research programme will be realised through the following steps:

a) Gathering data from graduates using a ‘conjoint measurement’ technique, which consists of asking the graduates what, within a depicted scenario, would be their choice if they were offered more than one job, each being characterised by a ‘conjoint’ set of attributes. This simulation will be repeated more than once in each questionnaire. The idea is that the graduates’ choices will reflect their existential values and their attitudes towards labour, conditional on their characteristics.

b) Estimation of the relative importance (also termed ‘utility’ or ‘part worth’) of job characteristics for the early graduates looking for a job; the data collected with this conjoint measurement approach will be useful in accomplishing this. The conjoint analysis will concern the part-worth estimation of each single characteristic and of some interactions between couples of characteristics, and should highlight the mental mechanisms which induce a university graduate either to accept or refuse an offered job. This analytical technique also allows the estimation of the so-called ‘willingness to accept’ a job defined by given characteristics.

Duration: Duration of the project: 12 months.

Research outcomes: The expected results concern both innovation in the methodology of choice data collection and that of data analysis with conjoint analysis techniques, on the one hand, and the interpretation of conjoint data aimed at eliciting the preferences for jobs of graduates, on the other hand.

Requirements: The competencies solicited from the candidate concern:
- Design and implementation of factorial experiments for the collection of data on large samples of graduates;
- The statistical analysis of multivariate data which may result from an experiment on early graduates using ad hoc software.
Essential references
Title:
Networks of Networks

Supervisor:
Bruno Scarpa
Contact person: Bruno Scarpa <Bruno.scarpa@unipd.it>

Short description:
Assume we have \( n \) subjects (nodes), for each subject \( k \) we have a brain network \( A_k \), and also we observe a network \( W \) among these subjects (like social relations among the \( n \) subjects). Hence our data are a \( n \times n \) social network \( W \), and a sequence of \( V \times V \) brain networks \( A_1, \ldots, A_n \); one for each actor in the network \( W \). Hence we end with a network \( (W) \) of networks \( A_1, \ldots, A_n \).

What can we do?: Suppose the \( n \) subjects are from several families, a cool question is for example: is it true that subjects having closer social relations in \( W \) (member in the same family) share some similar brain network structure \( A_k \). To do this my idea is:

1] Cluster brains \( A_1, \ldots, A_n \) using Population of Networks. So that we obtain group indicators \( G_1, \ldots, G_n \), with subjects in the same group, sharing the same brain network structure.
2] Model \( W \) (\( n \times n \)) via stochastic block-models, with node-specific regressor given by \( G_k \) and test whether \( G_k \) is informative on the stochastic blocks in \( W \).

Duration:
6 months.

Research outcomes:
A short list of results expected at the end of the projects.

Requirements:
Computer skills: good in programming, possibly knowledge and familiarity of R
Basics in functional analysis, matrix algebra, probability and theory of statistics

Prerequisites for the candidate (such as, for instance, necessary computer skills, or familiarity with specific software, background knowledge of theory and/or methods, or knowledge of specific models and/or methods).

Essential references
Title:
The Life Cycle of Ideas. Analyses of Keyword Counts in Large Corpora of Scientific Literature

Supervisor:
Arjuna Tuzzi (Department of Philosophy, Sociology, Education and Applied Psychology)
Contact person: Arjuna Tuzzi <arjuna.tuzzi@unipd.it>

Short description:
The main aim of this research project is exploring the opportunities of reading the history of a discipline by means of the temporal evolution of keywords’ frequencies within papers published by mainstream scientific journals. We assume that the “shape” of the keywords' trajectories reflects the relevance of the corresponding concepts and topics in the scientific discourse and, moreover, that there should be the opportunity to position the research activities of scholars as part of this frame. The project treads unchartered territory and provides great opportunities in terms of both theoretical and applied research.

The project aims at
- finding solutions for the normalization and standardization of diachronic textual data;
- achieving an overview of the relationship between time and keywords to check the existence of latent temporal patterns by means of explorative data analyses (EDA);
- identifying new statistical methods to assess the development of selected sets of keywords (e.g. splines and wavelets);
- identifying keywords showing prototypical temporal patterns and clustering keywords portraying similar temporal patterns by means of methods based on time series analysis (TSA), functional data analysis (FDA), cluster analysis and curve clustering (model-based and distance-based methods).

Duration:
12 months

Research outcomes:
- a critical analysis of different solutions for the normalization and standardization of diachronic textual data;
- new methods to assess the temporal evolution of selected sets of keywords;
- new methods to cluster keywords that show a similar temporal evolution;
- R packages to share methods among scholars and R users.

Requirements:
1. good familiarity with R-packages and R programming;
2. basic knowledge of EDA methods;
3. basic knowledge of FDA approaches.

Essential references


