— ONE PAGER —

PROJECT TOPICS

BY: DR. P.J. (PAULO) DE ANDRADE SERRA

My field of research. Mathematical Statistics is a field of Mathematics that deals with the design and study of statistical procedures. I focus on non-parametric models – so high- or infinite dimensional models. The main goal is to carry out statistical analyses while avoiding to make potentially restrictive assumptions on the data.

The kind of work that I do. Once data have been collected and an appropriate (non-parametric) model selected, we typically have certain questions in mind. These typically fall into one of three different categories. 1) Point estimation – the goal is to produce estimates, e.g., "What is the mean recovery time for a patient undergoing treatment A?". 2) Uncertainty quantification – making *high-probability* statements about what to expect, e.g., "What is the maximal recovery time that I should expect for the majority (say 95%) of patients?" 3) Testing – picking between competing hypothesis, e.g., "Do patients undergoing treatment A recover faster than patients undergoing treatment B?" These questions are answered by learning the distribution (or certain features of the distribution) of the data, but this is challenging if a large model is being used.

The tools that I use. I work mainly with three different types of approaches, namely: penalised estimation (these allow controlling a tradeoff between complexity and predictive power), recursive estimation (a tool for online learning that can be used for statistical tracking), and Bayesian estimation (this is a powerful and general tool for statistical learning which allows expert knowledge to be incorporated into the inference in a natural way). I am mostly interested in fundamental work but always with an eye out for applications.

Some recent problems. Here are some problems that I've been working on recently:

- Modelling of PPG signals (Philips) – PPG signals are collected by inexpensive wrist/finger mounted devices. In this project we model these signal and detect certain cardiac events.
- Estimation of intrinsic dimension (Journal of Machine Learning Research) – Large datasets, where many variables are recorded for each individual are common, but often contain redundancies between measurements so it becomes relevant to reduce the dimensionality. This motivates the problem of estimating the so called intrinsic dimension.
- Uncertainty quantification using splines (Bayesian Analysis) – Splines are popular estimators and in this work we use a Bayesian approach to quantify their uncertainty as estimators.
- Traffic estimation and route selection (Under review) – We solve statistical problems on traffic networks, and optimise routes based on the outcomes.
- Sparse quantile regression (Ongoing) – Quantile regression is about estimating quantiles, rather than means (to get robust estimates), and sparse models depend on many parameters only a few of which are relevant.

Possible topics for projects.

- Quantile Regression in Practice.
- Joint, Nonparametric Mean-Covariance Estimation.
- Validity of Propensity Score Based Methods.
- Data-driven optimization of energy systems.

November 4, 2022

Quantile Regression in Practice

Supervisor: dr. P.J. (Paulo) de Andrade Serra.

Keywords: quantile regression.

Background

Regression models describe features of a dependent variable, conditional on a set of independent variables; classically, this feature is the conditional expectation. In quantile regression, on the other hand, the focus is on inferring conditional quantiles thus providing a more complete and robust description of the relation between the predictors and the response.

Description

An analysis based on quantile regression has quite a few advantages over traditional regressions. Sample quantile-based estimates are robust to outliers. When the distribution of the data is heavy-tailed, then analysing quantiles is more meaningful than analysing means. Also, estimating several quantiles simultaneously provides a more comprehensive description of the data.

In this project you will look at the potential of using quantile regression, as opposed to (mean) regression to analyse a dataset. You are free to pick the dataset (that is appropriate).

Goals

The goal is to carry out a full analysis of a dataset with a focus on comparing what a quantile- vs mean-regression would deliver.

Student profile

This topic is compatible with projects for BSc students. The student should have knowledge in Mathematical Statistics (e.g., XB0049), with a focus on Linear Regression.

Projects can focus on practical/computational issues, and can be pursued in a Bayesian context. There is also the possibility to look at theoretical aspects, if relevant.

References

Koenker, R. – *Quantile Regression*, Cambridge University Press, 2010.

Bijma, F., Jonker, M., van der Vaart, A. – *An Introduction to Mathematical Statistics*, Amsterdam University Press, 2016.

Joint, Nonparametric Mean-Covariance Estimation

Supervisor: dr. P.J. (Paulo) de Andrade Serra.

Keywords: Regression, covariance, asymptotics.

Background

Linear regression models prescribe a linear relation between a dependent (response) variable and a set of independent variables. For simplicity, the observations are often assumed to be independent but this is restrictive. It is then of interest to jointly estimate the mean and covariance of the data. These models are relevant in economics, and medicine.

Description

When we impose non-parametric models on the mean and covariance structure, estimating these becomes challenging. A possible way to analyse this problem is to assume that the regression function is periodic and that the noise process is stationary.

Goals

Develop and study non-parametric estimators for the mean and covariance structure, and study their asymptotics. Of particular interest is the so called long-range-dependent noise setting.

Student profile

This topic is compatible with projects for MSc students. The student should have knowledge in Fourier Analysis (e.g., XB0005), Mathematical Statistics (e.g., XB0049), and Asymptotic Statistics (e.g., Mastermath).

Projects will focus on theoretical aspects and can be pursued in a Bayesian context.

References

Stein, E.M., and Shakarchi, R. – *Fourier Analysis: An Introduction*, Princeton Lectures in Analysis, 2003.

Tsybakov, A.B. – *Introduction to Nonparametric Estimation*, Springer, 2009.

Validity of Propensity Score Based Methods

Supervisor: dr. P.J. (Paulo) de Andrade Serra.

Keywords: Causal inference, Propensity scores, Regression.

Background

Causal inference concerns uncovering causal relations between quantities of interest. Think: does exposure to a certain chemical cause cancer?

Randomised controlled trials (RCT) are the gold standard approach to collect data from which causal relations can be directly extracted. In these, a pool of units is randomly assigned to either a control group (used as a baseline), or an experimental group (which undergoes an intervention). For a number of reasons (e.g., ethical concerns, practicality, costs) RCT are often not an option and one has to fall back on observational data.

In observational data, units do not end up in the experimental group purely at random so it becomes challenging to assess if differences with the control group are due to the intervention or due to confounders – variables which affect both treatment assignment and responses.

Description

In this project you will become acquainted with the use of propensity scores, and how these can help mitigate the effect of confounders. The starting point are two papers from P.C. Austin (see below) which respectively review and apply propensity score based methods to study the effect of smoking cessation counseling on mortality.

There can be two possible goals to this project:

- To reproduce the same methodology on a new dataset. This new dataset concerns...
- To reproduce the same methodology on simulated data. Here we wish to assess how different dependency structures affect the validity pf propensity score methods.

Student profile

This topic is compatible with projects for MSc students. The student should have knowledge in Mathematical Statistics (e.g., XB0049), including multiple Linear Regression.

The project focuses on methodological aspects so it is crucial that the student is comfortable with, say, Python, working with datasets, performing simulations, but also with implementing statistical methodology.

References

Austin, P.C. – *An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies*, Multivariate Behavioral Research 46, 399-424, 2011.

Austin, P.C. – *A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality*, Multivariate Behavioral Research 46, 119-151, 2011.

Data-driven optimization of energy systems

Supervisor: dr. P.J. (Paulo) de Andrade Serra, dr. A. (Alessandro) Zocca.

Keywords: Linear programming, network optimization, sustainable energy systems, time series analysis, Kalman filter.

Background

Power grids are one of the most critical and complex networks in modern-day society and are required to function reliably at all times. These networks are rapidly evolving to meet ambitious sustainability goals. However, the increasing adoption of renewable energy sources, which depend on highly variable sun and wind, introduces a massive amount of uncertainty. Therefore, it is key to augment the current optimization procedures for energy dispatch operations to fully account for the uncertainty using comprehensive stochastic methods that can capture spatio-temporal correlations of renewable energy sources.

Description

In this project you will work at the interface of optimisation and statistics.

The optimisation component consists of a multi-stage linear network optimisation problem known as *Optimal Power Flow* augmented with a few so-called chance constrains to capture the uncertainty. To solve the problem efficiently, these probabilistic constraints need to be approximated in a data-driven fashion. Data consists of publicly available time series for measurements of solar and wind power production at different locations. There are non-trivial interactions between the time- and space component, and it is important to find a parsimonious model for such data.

Furthermore, there is likely a tradeoff between the sensitivity of the parameters of the optimal solution and uncertainty quantification of the statistical inferences which further emphasises the interdependence of the two problems.

Student profile

This topic is compatible with projects for MSc students. The student should have knowledge in Mathematical Statistics (e.g., XB0049), including multiple Linear Regression, and Mathematical Optimization (e.g., XM0051).

The project focuses on methodological aspects so it is crucial that the student is comfortable with, say, Python, working with datasets, performing simulations, but also with implementing statistical methodology and solving mathematical programs.

References

Bienstock, D., Chertkov, M. and Harnett, S. – *Chance-Constrained Optimal Power Flow: Risk-Aware Network Control under Uncertainty*, SIAM Review, 56(3), pp.461-495, 2014 [PDF available here]

Tajer, A., Perlaza, S.M., and Poor, H.V. – *Advanced data analytics for power systems*, Cambridge University Press, 2021 [PDF available here]

Brockwell, P.J., and Davis, R.A. – *Time Series: Theory and Methods*, Springer, 1991.