Statistical model selection

Organizer: Florencia Leonardi, Universidade de São Paulo, Brasil

Model selection methods are important tools for statistical modelling and have gained a lot of attention in recent years for its usefulness in analysing high dimensional datasets in many applied areas. While classical statistical analysis focuses on models with a given set of parameters, model selection methods are used to select the most suitable “dimensionality" for the model, from a set of candidate spaces, given the data. Different model selection techniques have been proposed in the literature, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), that are methods based on the penalisation of the likelihood function. On the other hand, other methods based on the optimisation of the prediction error has been proposed, as for example Cross-validation (CV). The goal of this session is to present, from a theoretical point of view, recent results in the field of statistical model selection.

Program

Learning communities in weighted networks (Joint work with Elizaveta Levina)
Andressa Cerqueira (UNICAMP; BR)

Network models have received an increasing attention from the statistical community, in particular in the context of analyzing and describing the interactions of complex random systems. In this context, community structures can be observed in many networks where the nodes are clustered in groups with the same connection patterns. In this talk, we will address the community detection problem for weighted networks in the case where, conditionally on the node labels, the edge weights are drawn independently from a Gaussian random variable with mean and variance depending on the community labels of the edge endpoints. We will present a fast and tractable EM algorithm to recover the community labels that achieves the optimal error rate.

Hidden Markov random field models applied to color homogeneity evaluation in dyed textiles images (joint work with Victor Freguglia and Juliano L. Bicas)
Nancy Garcia (UNICAMP, BR)

Color is one of the most important features in any textile material. Due to its competitive price, most of the colorants currently used for textile dyeing are synthetic, originated from non-renewable sources and highly pollutant. There is an increasing interest for natural processes to dye fabrics. When new textile dyeing technologies are developed, evaluating the quality of these techniques involves measuring the resulting color homogeneity using digital images. The presence of a texture effect, caused by the interlacing of wart and weft yarns as well as small displacement of the fabric, creates a sophisticated dependence structure in pixels coloring. A hidden Markov random field model is applied to dyed textile image data where a statistical model is necessary in order to separate the signal from the dyeing effect (fixed effect described by smooth functions) and warp and weft texture effect (Gaussian mixture driven by a Markov random field), allowing an evaluation of color homogeneity in dyed textiles without the signal from the texture.

Mean-field models for deep neural networks
Roberto Imbuzeiro Oliveira (IMPA, BR)

Deep neural networks (DNN) are behind many of the recent advances in artificial intelligence and pattern recognition, from natural language processing to image labelling. However, we lack a good theoretical understanding of what makes DNNs work. It is not even understood why "overparameterization" -- ie. adding more and more neurons to a network -- allows for better performance, even though it presumably increases the risk of overfitting. This talk will present a potential approach to understanding DNNs with a very large number of parameters. In a nutshell, we show that certain overparameterized DNNs may be approximated by a new class of mean-field models that are amendable to further analysis. This partially generalizes previous results on shallow neural nets by several different authors. In particular, our result suggests that overparameterization has a kind of regularizing effect on the network. Our mean-field model is related to mean-field games and many other classes of models from several areas that carry the "mean-field" or "McKean-Vlasov" prefix. Still, we will not assume any previous familiarity with the topic of mean-field systems. Joint work with Dyego Araújo and Daniel Yukimura from IMPA.

Strong structure recovery for partially observed discrete Markov random fields on graphs (joint work with Lara Frondana and Rodrigo R.S. Carvalho)
Florencia Leonardi (USP, BR)

Discrete Markov random fields defined on graphs, usually called graphical models in the statistical literature, have received much attention from researchers in recent years, especially due to its flexibility to capture conditional dependence relationships between variables. Graphical models are in some sense "finite" versions of general random fields or Gibbs distributions, classical models in stochastic processes and statistical mechanics theory. In this talk we will focus on discrete Markov random field models (with a countable infinite set of variables), where the set of random variables takes values on a finite alphabet. One of the main statistical questions for this type of models is how to recover the underlying graph; that is, the graph determined by the conditional dependence relationships between the variables. We propose a penalized maximum likelihood criterion to estimate the graph of conditional dependencies, that can be partially observed. We prove the almost sure convergence of the estimator in the case of a finite or countable infinite set of variables. In the finite case the underlying graph can be recovered with probability one, while in the countable infinite case we can recover any finite subgraph with probability one, by allowing the candidate neighbourhoods to grow with the sample size. Our method requires minimal assumptions on the probability distribution and contrary to other approaches in the literature, the usual positivity condition is not needed.