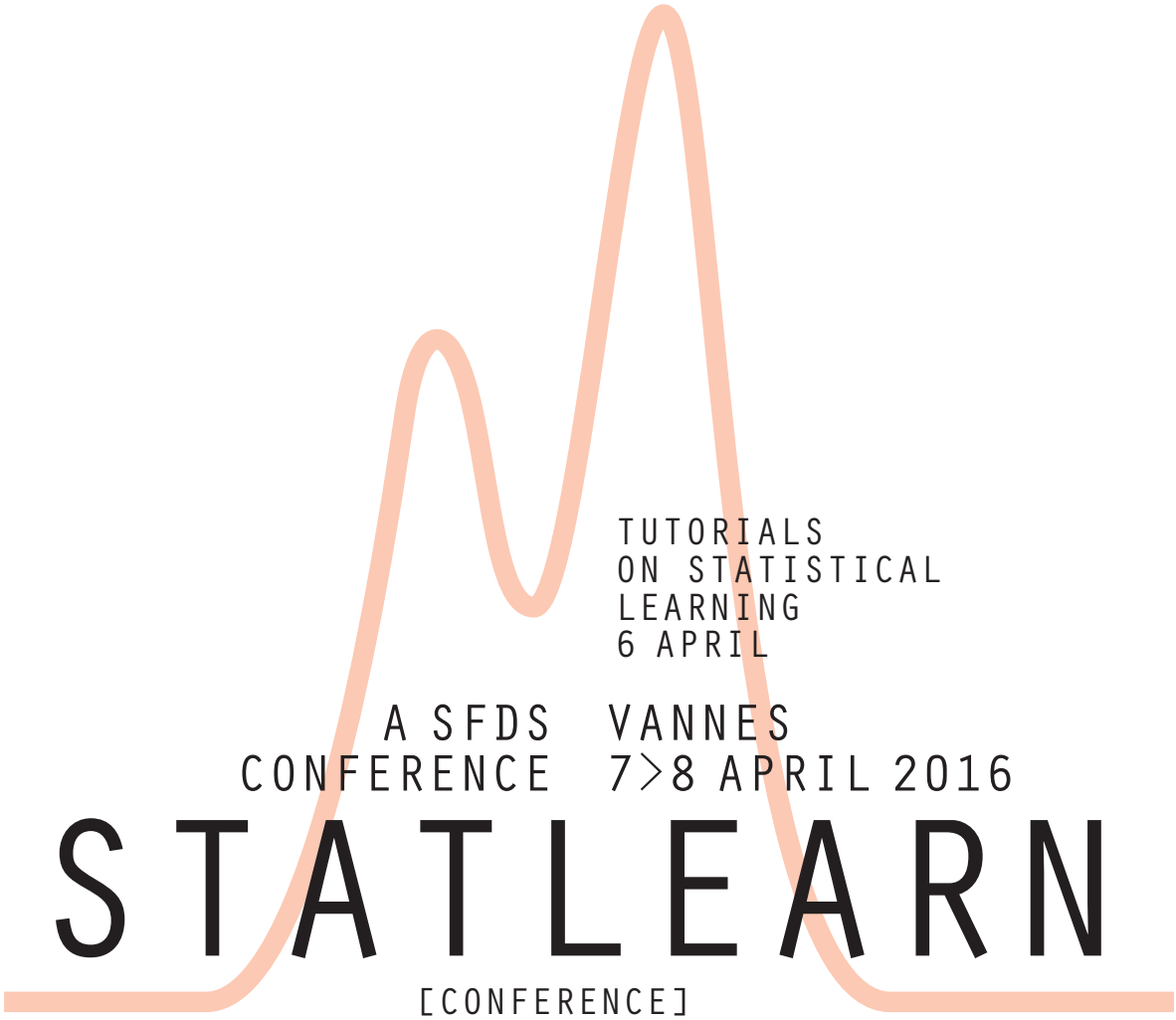

TUTORIALS
ON STATISTICAL
LEARNING
6 APRIL

A SFDS VANNES
CONFERENCE 7>8 APRIL 2016

STATLEARN

[CONFERENCE]

CHALLENGING PROBLEMS IN STATISTICAL LEARNING

Organized by
N. Béchet, C. Bouveyron, L. Chapel, N. Courty,
M. Emily, E. Fokoué, C. Friguet, J. Jacques & P. Latouche

Introduction

The 7th edition of the workshop on "Challenging problems in Statistical Learning" will be held in Vannes on April 7-8, 2016. This workshop aims to summarize the new and future problems in statistical learning and to give a good idea of what already exists for dealing with these problems. This two day workshop is split into 4 sessions :

— Session 1 : Topic modeling & text mining
— Session 2 : High dimension and applications
— Session 3 : Optimal transport and learning
— Session 4 : New and future problems in statistical learning

Organization

The workshop is an event of the Société Francaise de Statistique and is part of the thematic semester 2016 of the Centre Henri Lebesgue.

Dates and place

The workshop will take place on April 7-8, 2016 at Université Bretagne Sud, Campus de Tohannic, 56000 Vannes. The nearest bus station is "Université" on lines 6 and 2, and "Tohannic" on line 10.

The talks will be given in the Amphithéâtre 250 in Bâtiment DESG (main floor).

Acknowledgements

The organizers would like to thank the following organisms and persons who helped in preparing this workshop :

— Jean-Francois Dupuy (IRMAR/INSA Rennes, coordinateur du semestre Statistique du CHL),
— Véronique Vellet (LMBA),
— Xhensila Lachambre (coordinatrice CHL),
— Agnès Cottais (IRISA),
— l'IUT de Vannes pour son soutien logistique.

Wednesday April 6, 2016

10h00–18h00 : Tutorials on « Statistical learning and non-vectorial data »

— 9h15-10h00 : Registration & Coffee
— 10h00-12h00 : Pierre Latouche, *Statistical learning on graphs*
— 13h30-15h30 : Julien Jacques, *Statistical learning with functional data*
— 16h00-18h00 : Julien Velcin, *Topic models*

Thursday April 7, 2016

9h00–12h30 : Session « Topic modeling & text mining »

— 8h45-9h30 : Registration & Coffee
— 9h30-10h15 : Cédric Archambeau, *Latent IBP Compound Dirichlet Allocation : Sparse Topic Models Fit for Natural Languages*
— 10h15-11h00 : Julien Velcin, *Joint extraction of topics and sentiments*
— 11h00-11h45 : Coffee-break
— 11h45-12h30 : Quentin Pleplé, *Interactive Topic Modeling*

14h30–17h30 : Session « High dimension and applications »

— 14h30-15h15 : Ernest Fokoué, *Statistical Machine Learning with High Multidimensional Arrays of Data*
— 15h15-16h00 : Mathieu Fauvel, *Spectral-spatial classification of high-dimensional remote sensing images*
— 16h00-16h45 : Coffee-break
— 16h45-17h30 : Emeline Perthame, *Variable selection for correlated data in high dimension using decorrelation methods*

18h00–20h00 : Session « Poster »

Friday April 8, 2016

9h00–12h00 : Session « Optimal transport and learning »

— 9h00-9h45 : Marco Cuturi, *Regularized Optimal Transport and Applications*
— 9h45-10h30 : Jérémie Bigot, *A Forward-Backward algorithm for geodesic PCA of histograms in the Wasserstein space*
— 10h30-11h15 : Coffee-break
— 11h15-12h00 : Rémi Flamary, *Optimal transport for domain adaptation*

13h30–15h00 : Session « New and future problems in statistical learning »

— 13h30-14h15 : Patrick Perez, *On visual comparison*
— 14h15-15h00 : Vladimir Vapnik, *Learning with Intelligent Teacher : Similarity Control and Knowledge Transfer*

Tutoriel 1 – Statistical learning on graphs
Pierre Latouche

Graphs are highly used to characterize relationships between objects of interest. In this course, I will review the modularity based methods. I will also cover random graph models and focus on the latent position cluster model and the stochastic block model. Finally, inference techniques for these models, to do the clustering and the estimation of the parameters, will be investigated. In particular, I will talk about the variational expectation maximization (VEM) algorithm, the variational Bayes EM algorithm, Gibbs sampling, greedy search, and model selection. Examples (in R) of studies of real networks will be provided and R packages will be presented.

Tutoriel 2 – Statistical learning with functional data
Julien Jacques

Functional data generally refers to statistical analysis of data samples consisting of random functions. If functional data was for longtime inaccessible for statistics, today it becomes more and more easy to observe, to store and to process large amounts of such data. In this tutorial, the main tools for dealing with functional data will be presented and applied using the R software. The outline of the tutorial is :
- Motivations : why to consider such data as functions and not as temporal series,
- Technical tools : from discrete observations to functional data,
- Descriptive analysis : descriptive statistics and functional PCA,
- Clustering,
- Regression.

Tutoriel 3 – Topic models
Julien Velcin

Topic models constitue a major tool for dealing with textual data and beyond (e.g., images, meta-data, etc.). They are used by many different communities, from data mining to social sciences and humanities. The popular Latent Dirichlet Allocation (LDA) model is integrated into different platforms and programming languages (Java, R, Python ?), although its theoretical basis and limitations are usually not really appreciated by its users. In this tutorial, I will give an overview of the main applications of topic modeling and focus on various aspects of LDA (probabilistic model, parameter estimation, role of priors, choice of features). During the session, the audience will have the opportunity to test the model on their own textual dataset through the packages "rJava" and "mallet" in R.

Latent IBP Compound Dirichlet Allocation : Sparse Topic Models Fit for Natural Languages
Cédric Archambeau

Probabilistic topic models such as latent Dirichlet allocation are widespread tools to analyse and explore large text corpora. They postulate a generative model of text which ignores its sequential structure, but has been proven to be sufficient to capture the underlying semantics. However, the generative model is unable to handle out-of-vocabulary words and does not account for the power-law distribution of the vocabulary of natural languages. Among others, this leads in practice to the creation of an unnecessary large number of topics when modelling corpora of increasing size. We tackle this problem by introducing a probabilistic topic model based on the four-parameter IBP compound Dirichlet process, a stochastic process that generates sparse nonnegative vectors with potentially an unbounded number of entries. We call this new model the latent IBP compound Dirichlet allocation (LIDA), which enables us to model power-law distributions, both, in the number of topics summarising the documents and in the number of words defining each topic. It can be interpreted as a sparse variant of the hierarchical Pitman-Yor process when applied to topic modelling. We derive an efficient and simple collapsed Gibbs sampler closely related to the collapsed Gibbs sampler of latent Dirichlet allocation (LDA), making the model applicable in a wide range of domains. Our nonparametric Bayesian topic model compares favourably to the widely used hierarchical Dirichlet process and its heavy tailed version, the hierarchical Pitman-Yor process, on benchmark corpora. Experiments demonstrate that accounting for the power-distribution of real data is beneficial and that sparsity provides more interpretable results. This is joint work with Balaji Lakshminarayanan and Guillaume Bouchard.

Joint extraction of topics and sentiments
Julien Velcin

Faced with all the "big" data generated by users through the Web and the social media, we need efficient techniques to provide us with overviews such as categories and trends. Useful summaries can be calculated on textual data by using topic models, such as Latent Dirichlet Allocation (LDA) and variants able to deal with temporal dynamics (DTM, TOT). Besides, it turns out that topic models are also able to integrate the modeling of sentiments expressed towards these topics (ASUM, JST). Topic modeling and sentiment analysis (or opinion mining) are two popular tasks that have been treated separately in the past, even though they are complementary : sentiments usually target topics and topics can be the basis of subjective positions. In this talk, I will first give a brief overview on the various ways topics and opinions have been modeled together. I will then present an attempt to jointly extract topics and sentiments by extending the classical probabilistic LDA model. Based on two case studies experiments, I will show that our model named TTS is more fitted to capture the overall dynamics of the opinions expressed on short messages. In addition, I will show how such an hybrid topic-opinion model has been recently used to address the task of stock market prediction.

Interactive Topic Modeling
Quentin Pleplé

Topics discovered by the latent Dirichlet allocation (LDA) method are sometimes not meaningful for humans. The goal of our work is to improve the quality of topics presented to end-users. We present a novel method for interactive topic modeling. The method allows the user to give live feedback on the topics, and allows the inference algorithm to use that feedback to guide the LDA parameter search. The user can indicate that words should be removed from a topic, that topics should be merged, and/or that a topic should be split, or deleted. After each item of user feedback, we change the internal state of the variational EM algorithm in a way that preserves correctness, then re-run the algorithm until convergence. Experiments show that both contributions are successful in practice.

Statistical Machine Learning with High Multidimensional Arrays of Data
Ernest Fokoué

This lecture intends to give the audience a cursory tour of some of the most common tools used for learning the patterns underlying high multidimensional arrays of data. I will first focus on the ubiquitous two dimensional array (matrix) setting with n denoting the sample size or number of p dimensional vectors under consideration, and I will touch on some of the most recent statistical and computational methods for dealing with both the n»> p and the n «< p scenarios. Among other things, I will talk about the techniques of regularization/penalization, selection and projection that help circumvent the inferential and prediction challenges inherent in high dimensional regression, classification and clustering. I will also cover ensemble techniques like random forest and random subspace learning in general that have proved formidable in mitigating many learning challenges arising in high dimensional predictive modelling. Throughout this presentation, the concept of ?high ? will remain relative, typically approached from pure commonsense but also with respect to the computational architecture ultimately used for implementing the devised methods. Indeed, some of the solutions to the high dimensional data modelling conundrum will come from making the most of high performance parallel computation now available through multicore CPU architectures now standard on all our desktop computers, or Graphics Processing Units (GPU) that can be easily added or even clusters of computers more and more commonly used by statisticians. For most of the methods, techniques and algorithms mentioned earlier, I will point to the existing parallel implementation wherever possible. If time allows it, I will give an overview of some of the most promising results and applications of multidimensional arrays (tensors) in machine learning, namely the use of m-tensors in image processing, topic modelling, recommender systems and community detection just to name a few.

Spectral-spatial classification of high-dimensional remote sensing images

Mathieu Fauvel

In this talk, the classification of high dimensional remote sensing images will be discussed. By high dimensional, we mean that the number of features is high, typically several hundreds, while the number of training samples remains low. Feature will be of three kinds : ? Spectral : they are related to the reflectance in different wavelength domain acquire by the sensor. ? Spatial : they are related to some geometric feature extracted from the data. ? Temporal : they are related to different acquisition over the time. Recent advances in spectral-spatial classification of high dimensional remote sensing images will be presented in this talk. Several techniques are investigated for combining both spatial and spectral information. Spatial information is extracted at the object (set of pixels) level rather than at the conventional pixel level. Mathematical morphology, Markov random field, and object classification will be discussed in a kernel methods framework.

Variable selection for correlated data in high dimension using decorrelation methods

Emeline Perthame

The analysis of high throughput data has renewed the statistical methodology for feature selection. Such data are both characterized by their high dimension and their heterogeneity, as the true signal and several confusing factors are often observed at the same time. In such a framework, the usual statistical approaches are questioned and can lead to misleading decisions as they are initially designed under independence assumption among variables. In this talk, I will present some improvements of variable selection methods in regression and supervised classification issues, by accounting for the dependence between selection statistics. The methods proposed in this talk are based on a factor model of covariates, which assumes that variables are conditionally independent given a vector of latent variables. During this talk, I will illustrate the impact of dependence on the stability on some usual selection procedures. Next, I will particularly focus on the analysis of event-related potentials data (ERP) which are widely collected in psychological research to determine the time courses of mental events. Such data are characterized by a temporal dependence pattern both strong and complex which can be modeled by the mentioned above factor model.

Regularized Optimal Transport and Applications
Marco Cuturi

Optimal transport (OT) theory provides geometric tools to compare probability measures. After reviewing the basics of OT distances (a.k.a Wasserstein or Earth Mover ?s), I will show how an adequate regularization of the OT problem can result in substantially faster (GPU parallel) and much better behaved (strongly convex) numerical computations. I will then show how this regularization can enable several applications of OT to learn from probability measures. I will focus on in particular on the computation of Wasserstein barycenters and inverse problem (regression) in the simplex with the OT geometry (the latter being joint work with G. Peyré and N. Bonneel).

A Forward-Backward algorithm for geodesic PCA of histograms in the Wasserstein space
Jérémie Bigot

Principal Component Analysis (PCA) in a linear space is certainly the most widely used approach in multivariate statistics to summarize efficiently the information in a data set. In this talk, we are concerned by the statistical analysis of data sets whose elements are histograms with support on the real line. For the purpose of dimension reduction and data visualization of variables in the space of histograms, it is of interest to compute their principal modes of variation around a mean element. However, since the number, size or locations of significant bins may vary from one histogram to another, using PCA in an Euclidean space is not an appropriate tool. In this work, an histogram is modeled as a probability density function (pdf) with support included in an interval of the real line, and the Wasserstein metric is used to measure the distance between two histograms. In this setting, the variability in a set of histograms can be analyzed via the notion of Geodesic PCA (GPCA) of probability measures in the Wasserstein space. However, the implementation of GPCA for data analysis remains a challenging task even in the simplest case of pdf supported on the real line. The main purpose of this talk is thus to present a fast algorithm which performs an exact GPCA of pdf with support on the real line, and to show its usefulness for the statistical analysis of histograms of surnames over years in France.

Optimal transport for domain adaptation
Rémi Flamary

Domain adaptation is one of the most challenging tasks of modern data analytics. If the adaptation is done correctly, models built on a specific data representations become more robust when confronted to data depicting the same semantic concepts (the classes), but observed by another observation system with its own specificities. Among the many strategies proposed to adapt a domain to another, finding domain-invariant representations has shown excellent properties, as a single classifier can use labelled samples from the source domain under this representation to predict the unlabelled samples of the target domain. In this paper, we propose a regularized unsupervised optimal transportation model to perform the alignment of the representations in the source and target domains. We learn a transportation plan matching both PDFs, which constrains labelled samples in the source domain to remain close during transport. This way, we exploit at the same time the few labeled information in the source and distributions of the input/observation variables observed in both domains. Experiments in toy and challenging real visual adaptation examples show the interest of the method, that consistently outperforms state of the art approaches.

On visual comparison
Patrick Perez

Deciding if two pieces of visual content are somehow related, or to which degree they are related, is an ubiquitous problem when searching, processing and analyzing images. Such visual comparisons can be conducted among small fragments (e.g., point matching for tracking or 3D reconstruction, patch-based image processing, mid-level feature mining), object-level fragments (e.g., face verification or face clustering) or whole images (copy detection, image retrieval, picture linking). To this end, visual content is usually turned into a fixed size, high-dimensional vector representation and a suitable similarity measure is defined between such vectors. Focusing on large scale example-based search and on face verification, we shall discuss how parts of this ?description-comparison ? pipeline can be learned, with or without supervision, in order to speed up comparisons or to make them more meaningful.

Learning with Intelligent Teacher : Similarity Control and Knowledge Transfer

Vladimir Vapnik

In the talk, I will introduce a model of learning with Intelligent Teacher. In this model, Intelligent Teacher supplies (some) training examples $(x_i, y_i),\ i = 1, ..., \ell,\ x_i \in \mathcal{X},\ y_i \in \{-1, 1\}$ with additional (privileged) information) $x_i^* \in \mathcal{X}^*$ forming training triplets $(x_i, x_i^*, y_i),\ i = 1, ..., \ell$. Privileged information is available only for training examples and not available for text examples. Using privileged information it is required to find a better training processes (that use less examples or more accurate with the same number of examples) than the classical ones. In this lecture, I will present two additional mechanisms that exist in learning with Intelligent Teacher :
- The mechanism to control Student's concept of examples similarity,
- The mechanism to transfer knowledge that can be obtained in space of privileged information to the desired space of decision rules.
Privileged information exists for many inference problem and Student-Teacher interaction can be considered as the basic element of intelligent behavior.

List of accepted posters

Poster	Authors	Title
1	Hussein AL-NATSHEH	A generic sentence semantic semilarity model
2	Ignacio Arroyo-Fernandez, Gerardo Sierra and Juan-Manuel Torres-Moreno	Learning Kernels for Semantic Clustering
3	Pierre-Alexandre Mattei, Charles Bouveyron, Pierre Latouche	Globally Sparse Probabilistic PCA
4	Yanwei CUI	A Subpath Kernel for Learning Hierarchical Image Representations
5	Adeline Bailly, Simon Malinowski, Romain Tavenard, Laetitia Chapel, Thomas Guyet	Dense Bag-of-Temporal-SIFT-Words for Time Series Classification
6	Arnaud de Myttenaere, Bénédicte Le Grand, Fabrice Rossi	Supervised classification under explanatory shift
7	Marco Corneli, Pierre Latouche and Fabrice Rossi	Clustering in dynamic networks via non-homogeneous Poisson processes and exact ICL
8	Romain Huet, Nicolas Courty and Sebastien Lefevre	A new penalisation term for image retrieval in clique neural networks
9	C. Bouveyron, L. Bergé, P. Latouche and R. Zreik	Topic-conditional Stochastic Block Model for the Simultaneous Clustering of Vertices and Textual Edges in Communication Networks
10	Villain Jonathan, Durrieu Gilles, Bureau Ronan	Robustness in machine learning: application to chemoinformatic.
11	Maxime Sangnier, Olivier Fercoq and Florence D'Alché-Buc	Joint quantile regression in vector-valued RKHSs
12	Yosra BEN SLIMEN, Julien JACQUES, Sylvain ALLIO	Model-Based Co-clustering for Functional Data
13	Maël Chiapino and Anne Sabourin	Feature clustering for extreme events analysis, with application to extreme streamflow data
14	Julie Soulas	Habit monitoring over sensor streams
15	J. Salotti, R. Billot, NE. El Faouzi, S. Fenet, C. Solnon	Toward the use of causal graph for better short-term road traffic forecast.
16	Nicolas Audebert, Bertrand Le Saux, Séabstien Lefèvre	Deep learning for aerial cartography
17	Valérie Robert, Gilles Celeux, Christine Keribin	Latent Block Model and model selection with application in pharmacovigilance
18	Mélina Gallopin, Emilie Devijver	Block-diagonal covariance selection for high-dimensional Gaussian graphical models.

Statlearn'16 is organized with the support of