Scientific Program and Abstracts

International Workshop on Perspectives on
High Dimensional Data VII

HDDA VII

Guanajuato, Mexico
June 15-18, 2017

Program

Thursday June 15th

08:15 - 08:40 **Registration and Coffee**

08:40 - 09:00 **Opening Remarks HDDA-VII**
Ejaz Ahmed, *Brock University*
Victor Rivero, *CIMAT (Director General)*

09:00 - 09:50 **Plenary Talk**
Michael Daniels, *University of Texas at Austin*
Title: Bayesian Nonparametric Generative Models for Causual Inference with Missing at Random Covariates
Chair: Ejaz Ahmed

09:50 - 10:30 **Tanujit Dey**, *Cleveland Clinic*
Title: Variable screening, selection and prediction in high dimensions
Chair: Ejaz Ahmed

10:30 - 10:50 **Coffee Break**

10:50 - 11:30 **Carolina Euan**, *King Abdullah University of Science and Technology*
Title: Spectral Analysis approach to visualizing Clustering Patters of Winds and Waves in the Red Sea
Chair: Michael Daniels

11:30 - 12:10	**Vyacheslav Lyubchich**, *University of Maryland Cener for Environmental Science* **Title**: Dynamic Spatio-Temporal Clustering of Water Quality Trends Chair: Michael Daniels
12:10 - 12:50	**Xiaoli Gao**, *University of North Carolina at Greensboro* **Title**: Penalized Weighted Least Absolute Deviation Regression Chair: Michael Daniels
12:50 - 14:20	**Lunch Break**
14:20 - 14:50	**Miguel Villalobos**, *Universidad Anahuac* **Title**: A Statistical and Machine Learning Model to Detect Money Laundering Chair: Carolina Euan
14:50 - 15:10	**Coffee Break**
15:10 - 15:40	**Faisal Maqbool Zahid**, *Ludwig Maximilians University, Munich, Germany* **Title**: Variable Selection Techniques after Multiple Imputation in High-dimensional data Chair: Israel Martinez
15:40 - 16:20	**Ekaterina Smirnova**, *University of Montana* **Title**: Methods for Sparsity Adjustments in Microbiome Data Chair: Carolina Euan

Friday June 16th

08:30 - 09:00	**Coffee**
09:00 - 09:50	**Plenary Talk** **Nozer D. Singpurwalla**, *The City University of Hong* **Title**: Subjective Probability: Its Axioms and Acrobatics Chair: Martin Lysy
09:50 - 10:30	**Ivor Cribbean**, *University of Alberta School of Business* **Title**: Methods for Network Change Point Detection Chair: Martin Lysy
10:30 - 10:50	**Coffee Break**
10:50 - 11:30	**Yuan Huang**,*Yale University* **Title**: Promote Sign Consistency in the joint Estimation of Precision Matrices Chair: Martin Lysy

11:30 - 12:10 **Vahid Partovi Nia**, *Ecole Polytechnique de Montreal*
Title: Bayesian Clustering of Cell Shapes
Chair: Vahid Partovi Nia

12:10 - 12:50 **Ribana Roscher**, *University of Bonn*
Title: Sparse Representation-based Analysis of Hyperspectral Remote Sensing Data
Chair: Vahid Partovi Nia

12:50 - 13:00 **Group Photograph**

13:00 - 14:20 **Lunch Break**

14:20 - 15:00 **Alexander Shestopaloff**, *University of Toronto*
Title: Sampling Latent States for High-Dimensional Non-Linear State Space Models with the embedded HMM Method
Chair: Vahid Partovi Nia

15:00 - 15:20 **Coffee Break**

15:20 - 16:00 **Martin Lysy**, *University of Waterloo*
Title: Efficient Computational Inference for Microparticle Tracking in Biological Fluids
Chair: Vahid Partovi Nia

16:00 - 17:00 **Poster Session**

Saturday June 17th

08:30 - 09:00 **Coffee**

09:00 - 09:50 **Plenary Talk**
George Michailidis, *University of Florida*
Title: Regularized Estimation and Testing for High-Dimensional
Chair: Shuangge Ma

09:50 - 10:30 **Luis Javier Álvarez**, *Institute of Mathematics-UNAM*
Title: Applying deterministic chaos theory for analyzing high-dimensional and complex data
Chair: Luis Javier Álvarez

10:30 - 10:50 **Coffee Break**

10:50 - 11:30 **Igor Barahona**, *Institute of Mathematics-UNAM*
Title: From digital to nonlinear time series. Quantifying the degree of chaos in a volcanic eruption episode
Chair: Luis Javier Álvarez

11:30 - 12:10 **Elizabeth Santiago**, *Institute of Mathematics-UNAM*
Title: Community structure for analyzing the flow capability in fracture networks in rock
Chair: Luis Javier Álvarez

12:10 - 12:50 **Antonio Sarmiento Galan**, *Institute Of Mathematics - UNAM*
Title: Breathers and Thermal Relaxation as a Temporal Process
Chair: Luis Javier Álvarez

12:50 - 14:20 **Lunch Break**

14:20 - **Free Afternoon**

Sunday June 18th

08:30 - 09:00 **Coffee**

09:00 - 09:50 **Plenary Talk**
Peter Song, *University of Michigan*
Title: Statistical Inference with Estimating Functions via the MapReduce Scheme
Chair: Leticia Ramirez

09:50 - 10:30 **Fabian Martinez Martinez**, *CIMAT*
Title: Topology and Statistics, can they get along?
Chair: Peter Song

10:30 - 10:50 **Coffee Break**

10:50 - 11:40 **Plenary Talk**
Shuangge Ma, *Yale University*
Title: Analysis of Cancer Gene Expression Data with an Assisted Robust Marker Identification Approach
Chair: Igor Barahona

11:40 - 12:20 **Israel Martinez Hernandez**, *CIMAT*
Title: Dynamic Factor Model for Functional Time series
Chair: Igor Barahona

12:20 - 12:50 **Asad Lodhia**, *University of Michigan Ann Arbor*
Title: Marchenko-Pastur Law for Kendall's Tau
Chair: Igor Barahona

12:50 - 14:20 **Lunch Break**

14:20 - 14:50 **Shahla Faisal**, *Ludwig Maximilians University, Munich, Germany*
Title: Imputation for Missing Values in High-Dimensional Data Structures
Chair: Vyacheslav Lyubchich

14:50 - 15:10 **Coffee Break**

15:10 - 15:50 **Jingyi Jessica Li**, *University of Carlifornia, Los Angeles*
Title: A bootstrap Lasso+Partial Ridge Method to Contruct Confidence Intervals for High-Dimensional Linear Model Coefficients
Chair: Vyacheslav Lyubchich

15:50 - 16:30 **Idir Ouassou**, *Cadi Ayyad University, Morocco*
Title:
Chair: Vyacheslav Lyubchich

Abstracts

Speaker : Ivor Cribben, University of Alberta School of Business

Title: Methods for Network Change Point Detection

Coauthor(s): Yi Yu, University of Bristol

Abstract: Recently, there has been an increased interest in the estimation of change points in a high-dimensional time series context. In this work, we consider methods which detect change points in the network structure of a multivariate time series, with each component of the time series represented by a node in the network. We examine data-driven methods that use binary segmentation and shutter-windows as well as with and without spectral clustering (Luxburg, 2007). We apply the new methods to various simulated data sets and to a resting-state functional Magnetic Resonance Imaging (fMRI) data. In this framework, we picture the brain as an integrated system and wish to study its large-scale characterizations and dynamics.

Plenary Speaker : Michael Daniels, University of Texas at Austin

Title: Bayesian Nonparametric Generative Models for Causal Inference with Missing at Random Covariates

Coauthor(s): Jason Roy, Kirsten Lum, Bret Zeldow, Jordan Dworkin, University of Pennsylvania

Abstract: We propose a general Bayesian nonparametric (BNP) approach to causal inference in the point treatment setting. The joint distribution of the observed data (outcome, treatment, and confounders) is modeled using an enriched Dirichlet process. The combination of the observed data model and causal assumptions allows us to identify any type of causal effect - differences, ratios, or quantile effects, either marginally or for subpopulations of interest. The proposed BNP model is well-suited for causal inference problems, as it does not require parametric assumptions about the distribution of confounders and naturally leads to a computationally efficient Gibbs sampling algorithm for both large n and p. By flexibly modeling the joint distribution, we are also able to impute (via data augmentation) values for missing covariates within the algorithm under an assumption of ignorable missingness, obviating the need to create separate imputed data sets. This approach for imputing the missing covariates has the additional advantage of guaranteeing congeniality between the imputation model and the analysis model, and because we use a BNP approach, parametric models are avoided for imputation.

The performance of the method is assessed using simulation studies. The method is applied to data from a cohort study of human immunodeficiency virus/hepatitis C virus co-infected patients.

Speaker : Tanujit Dey, Cleveland Clinic

Title: Variable screening, selection and prediction in high dimensions

Coauthor(s): Ian Dryden, University of Nottingham and Daniel Vasiliu, College of William and Mary.

Abstract: We consider an alternative method for variable screening, selection and prediction in linear regression problems where the number of predictors are much higher than the number of observations. The method involves minimizing a penalized Euclidean distance with a newly defined penalty term. This particular formulation exhibits a grouping effect, which is useful for screening out predictors in higher or ultra-high dimensional problems. Practical performances of variable selection and prediction are evaluated through simulation studies and the analysis of a dataset of mass spectrometry scans from melanoma patients, where excellent predictive performance is obtained.

Speaker : Carolina Euan, King Abdullah University of Science and Technology (KAUST)

Title: Spectral analysis approach to visualizing clustering patters of winds and waves in the Red Sea

Coauthor(s): Ying Sun

Abstract: Wind and wave modeling in the Red Sea offers many challenges due to the physical structure. Along the Red Sea, a non-stationary spatial behavior of both phenomena can be observed. We consider clustering techniques to show the non-stationary. We consider a model that describes the similarities between wave and wind directional spectra. A well characterization of the patterns of wind and wave energy will be very important for harvesting energy. Even more, a well characterization based on a complete energy profile will result on a better identification of potential sources of energy. We developed an application to visualize the clustering patterns of wind and wave, which contributes in a better understanding of the relations between them.

C Speaker : Shahla Faisal, Ludwig Maximilians Universität, München, Deutchland

Title: Imputation for Missing Values in HighDimensional Data Structures Confirmed

Coauthor(s): Gerhard Tutz

In modern research, the data often contains a large number of variables but the information on some variables is missing. The techniques to analyse such highthroughput data often require a complete data without any missing values. Imputation is a common solution where the downstream analyses require a complete data matrix. A number of imputation methods are available that work under some distributional assumptions. We propose an improvement over the popular nonparametric nearest neighbours imputation method that requires no particular assumptions. The proposed method makes practical and effective use of the information on association among the variables. In particular we propose a weighted version of Minkowski's distance that uses the information from a subset of important variables only. The performance of the proposed method is investigated under a variety of data settings containing 10% to 40% missing values. The results show that the proposed method outperforms the compared approaches with a smaller imputation error. Furthermore, it works efficiently even when the number of samples is less than the number of variables

Speaker : Xiaoli Gao, University of North Carolina at Greensboro

Title: Penalize weighted least absolute deviation regression

Coauthor(s): Yang Feng, Columbia University

Abstract: In a linear model where the data is contaminated or the random error is heavy-tailed, least absolute deviation (LAD) regression has been widely used as an alternative approach to least squares (LS) regression. However, it is well known that LAD regression is not robust to outliers in the explanatory variables. When the data includes some leverage points, LAD regression may perform even worse than LS regression. In this manuscript, we propose to improve LAD regression in a penalized weighted least absolute deviation (PWLAD) framework. The main idea is to associate each observation with a weight reflecting the degree of outlying and leverage effect and obtain both the weight and coefficient vector estimation simultaneously and adaptively. The proposed PWLAD is able to provide regression coefficients estimate with strong robustness, and perform outlier detection at the same time, even when the random error does not have finite variances. We provide sufficient conditions under which PWLAD is able to identify true outliers consistently. The performance of the proposed estimator is demonstrated via extensive simulation studies and real examples.

Speaker : Yuan Huang, Yale University

Title: Promote sign consistency in the joint estimation of precision matrices Confirmed

Coauthor(s): Qingzhao Zhang and Shuangge Ma

Abstract: The Gaussian graphical model is a popular tool for inferring the relationships among random variables, where the precision matrix has a natural interpretation of conditional independence. With high-dimensional data, sparsity of the precision matrix is often assumed, and various regularization methods have been applied for estimation. Under quite a few important scenarios, it is desirable to conduct the joint estimation of multiple precision matrices. In joint estimation, entries corresponding to the same element of multiple precision matrices form a group, and group regularization methods have been applied for estimation and identification of the sparsity structures. For many practical examples, it can be difficult to interpret the results when parameters within the same group have conflicting signs. Unfortunately, the existing methods lack an explicit mechanism concerning with the sign consistency of group parameters. To tackle this problem, we develop a novel regularization method for the joint estimation of multiple precision matrices. It effectively promotes the sign consistency of group parameters and hence can lead to more interpretable results, while still allowing for conflicting signs to achieve full flexibility. Its consistency properties are rigorously established. Simulation shows that the proposed method outperforms the competing alternatives under a variety of settings.

Speaker : Jingyi Jessica Li, University of California, Los Angeles

Title: A bootstrap Lasso+Partial ridge method to construct confidence intervals for highdimensional linear model coefficients

Coauthor(s): Hanzhong Liu, Tsinghua University; Xin Xu, Yale University; Bin Yu, University of California, Berkeley

Abstract: For high dimensional sparse linear models, how to construct confidence intervals for feature coefficients remains a difficult question. The main reason is the complicated limiting distributions of common estimators, such as the Lasso estimator. A series of methods have been developed to tackle this problem. Among them, Bootstrap Lasso+OLS is notable for its simple technicality, good interpretability, and reasonable performance as compared with more complicated methods. Yet Bootstrap Lasso+OLS depends on the betamin assumption, a theoretic criterion that is often violated in practice. In this paper, we introduce a new method called Bootstrap Lasso+Partial Ridge (LPR) to relax the betamin assumption. LPR is a twostage estimator: first using Lasso to select features and second using Partial Ridge to reestimate the feature coefficients. Simulation results show that our proposed Bootstrap LPR method outperforms Bootstrap Lasso+OLS when there exist small but nonzero coefficients, a common situation that violates the betamin assumption. For such coefficients, compared to Bootstrap Lasso+OLS, confidence intervals constructed by Bootstrap LPR have on average 50% larger coverage probabilities. Also compared to the desparsified Lasso methods, Bootstrap LPR has on average 35% 50% shorter confidence interval lengths with comparable coverage probabilities, regardless of whether linear models are misspecified. Additionally, we provide theoretical guarantees of Bootstrap LPR under appropriate conditions. Finally, we implement and distribute Bootstrap LPR in the R package "LPR".

C Speaker : Asad Lodhia, University of Michigan Ann Arbor

Title: Marchenko-Pastur Law for Kendall's Tau

Coauthor(s): Afonso Bandeira and Philippe Rigollet

We prove that Kendall's Rank correlation matrix converges to the Marchenko-Pastur law, under the assumption that the observations are i.i.d random vectors $X_1, \ldots, X_n$ with components that are independent and absolutely continuous with respect to the Lebesgue measure. This is the first result on the empirical spectral distribution of a multivariate U-statistic.

Speaker : Martin Lysy, University of Waterloo

Title: Efficient Computational Inference for Microparticle Tracking in Biological Fluids

Abstract: State-of-the-art techniques in passive particle-tracking microscopy provide high-resolution path trajectories of diverse foreign particles in biological fluids. In order to analyze experiments often tracking thousands of particles at once, scientists typically must account for many sources of unwanted variability, such as heterogeneity of the fluid environment and measurement error. To this end, sophisticated models must be fit to large amounts of time-dependent data, imposing a monumental computational burden. This talk presents several strategies to mitigate this problem. In particular, we consider a versatile family of stochastic models for which many parameters can be analytically profiled out; present a novel method for "superfast" likelihood evaluations with stationary Gaussian data; and discuss a distributed-computing approach to inference for hierarchical models. These techniques are illustrated in the context of quantifying subdiffusive mobility of tracer particles in human lung mucus.

Speaker : Vyacheslav Lyubchich, University of Maryland Center for Environmental Science

Title: Dynamic spatio-temporal clustering of water quality trends

Coauthor(s): Yulia R. Gel, Jeremy M. Testa, Xin Huang, Iliyan Iliev, and Qian Zhang

Abstract: Modern climate data sets, including paleoreconstructions, long-term weather monitoring records, and remote sensing data, contain a wealth of space-time information that leads to a variety of challenges related to data storage, management, and analysis. This has sparked an interest in dynamic space-time clustering algorithms that are particularly suitable for the analysis of large data streams. The trend-based clustering algorithm TRUST allows segmentation of space-time processes in real time, but requires the user to set multiple tuning parameters, and this step is usually performed in a subjective manner. Here we propose data-driven automatic approaches to simultaneously select the tuning parameters. We focus on the two most important parameters of the TRUST algorithm, which define short-term closeness of observations across locations and long-term persistence of such closeness within an analyzed time window. We demonstrate the performance of the enhanced clustering procedures using simulated time series, and illustrate their applicability using long-term water quality records in Chesapeake Bay.

Plenary Speaker : Shuangge Ma, Yale University

Title: Analysis of Cancer Gene Expression Data with an Assisted Robust Marker Identification Approach

Abstract: Gene expression studies have been playing a critical role in cancer research. Despite tremendous effort, the analysis results are still often unsatisfactory, because of the weak signals and high data dimensionality. Analysis is often further challenged by the long-tailed distributions of the outcome variables. In recent multidimensional studies, data have been collected on gene expressions as well as their regulators (for example, copy number alterations, methylation, and microRNAs), which can provide additional information on the associations between gene expressions and cancer outcomes. In this study, we develop an ARMI (Assisted Robust Marker Identification) approach for analyzing cancer studies with measurements on gene expressions as well as regulators. The proposed approach borrows information from regulators and can be more effective than analyzing gene expression data alone. A robust objective function is adopted to accommodate long-tailed distributions. Marker identification is effectively realized using penalization. The proposed approach has an intuitive formulation and is computationally much affordable. Simulation shows its satisfactory performance under a variety of settings. TCGA (The Cancer Genome Atlas) data on melanoma and lung cancer are analyzed, which leads to biologically plausible marker identification and superior prediction.

C Speaker : Faisal Maqbool Zahid, Ludwig Maximilians University, Munich, Germany

Title: Variable Selection Techniques after Multiple Imputation in Highdimensional Data

Coauthor(s): Shahla Faisal, Christian Heumann

Missing data is a common issue in almost all research fields. Missing data, when handled inappropriately, can cause difficulties in the data analysis and may lead to the misleading results. In recent years multiple imputation (MI) has emerged as most attractive approach for handling missing data. MI replaces the missing values with more than one simulated imputed values and thus providing different plausible versions of the complete data. The variable selection is an important aspect in the analysis of highdimensional data structure. The use of a variable selection technique independently on each of the multiply imputed data set may provide different model for each imputed data set. The question about how to use the variable selection techniques on different multiply imputed data sets is still not clearly answered in the literature. This paper proposes and compares different alternatives of performing variable selection on multiplyimputed data sets to get a consistent set of selected variables across the multiply imputed data sets. We are considering two different approaches: first approach puts a constraint on the frequency of each selected variable and the second approach focuses on selecting the consistent set of variables by defining a threshold for the magnitude of the parameter estimates. The results of different variations of these two approaches are compared with the MILasso approach proposed by Chen and Wang (2013) and the variable selection applied on the true complete data in a simulation study. The approaches are compared using Hit Rate (proportion of selected important variables among the true important variables) and False Alarm Rate (proportion of selected unimportant variables among the true unimportant variables). The considered techniques are also compared with respect to the prediction error for the model

Key Words: Highdimensional data, Multiple imputation, Regularization, Rubin's rules, Variable selection.

Speaker : Israel Martínez Hernández, CIMAT

Title: Dynamic Factor Model for Functional Time Series Confirmed

Coauthor(s): Jesus Gonzalo and Graciela Gonzalez Farias

Abstract: In many phenomena, identify the factors that drives the behavior of observed temporal dependent data is important, even more when the data is functional. In this work we propose a new approach the Dynamic Factor Model for Functional Time Series. Our model will characterize the dynamic of the data through the factors and the functional behavior of the data via the factor loadings. Via simulation studies, we show that the factors capture the dynamic of the process and share the same properties of the functional data, such as stationarity. We propose a methodology to estimate the factors and we investigate finite sample performance using Monte Carlo studies.

Speaker : Fabian Martínez Martínez, CIMAT

Title: Topology and Statistics, can they get along?

Abstract: Topological data analysis (TDA) is a modern field of research aiming to find structure in data. In particular, it is of interest to model topological structures like connected components, holes and their analogous in higher dimensions. In this talk, we provide an introduction to TDA via persistent homology, focusing on the class of mathematical objects obtained from the data cloud. These topological statistics resemble first attempts to describe populations through exploratory data analyses, and as a natural extension to it, we are now concentrated on the challenging task of providing probabilistic and statistical models adequate to make inferences. Thus, we present a current work dealing with this lack of probability distributions, based on life times, which additionally seems to provide useful information to make inferences not only topologically speaking but also in terms of the geometry of the data.

Plenary Speaker : George Michailidis, University of Florida

Title: Regularized Estimation and Testing for High-Dimensional

Abstract: Dynamical systems comprising of multiple components originate in many scientific areas. A pertinent example is the interactions between financial assets and macroeconomic indicators, which has been studied at an aggregate level in the macroeconomics literature. A key shortcoming of this approach is that it ignores potential influences from other related components (e.g. Gross Domestic Product) that may exert influence on the system's dynamics and structure and thus produces incorrect results. To mitigate this issue, we consider a multi-block linear dynamic system with Granger-causal ordering between blocks, wherein the blocks temporal dynamics are described by vector autoregressive processes and are influenced by blocks higher in the system hierarchy. We obtain the maximum likelihood estimator for the posited model for Gaussian data in the high-dimensional setting with appropriate regularization schemes. To optimize the non-convex likelihood function, we develop an iterative algorithm with convergence guarantees. We establish theoretical properties of the maximum likelihood estimates, leveraging the decomposability of the regularizers and a careful analysis of the iterates of the proposed algorithm. Finally, we develop testing procedures for the null hypothesis of whether a block ?Granger-causes? another block of variables. The performance of the model and the testing procedures are evaluated on synthetic data, and illustrated on a data set involving log-returns of the US S&P 100 component stocks and key macroeconomic variables for the 2001–16 period.

Speaker : Idir Ouassou, Cadi Ayyad University, Morocco

Title:

Coauthor(s):

Abstract:

Speaker : Vahid Partovi Nia, Ecole Polytechnique de Montreal

Title: Bayesian clustering of cell shapes

Abstract: Statistical clustering, or unsupervised machine learning divides a heterogeneous data into homogenous subsets. Living cells are heterogeneous in shapes, specially while cancer in under development. Most of the available clustering algorithms of shapes use distance methods. Looking at cell shapes as a closed curve, and employing model-based clustering allows us to treat clustering shapes through mathematical functions. This new view extends the Bayesian information criterion towards cluster selection.

Speaker : Ribana Roscher, University of Bonn

Title: Sparse representation-based analysis of hyperspectral remote sensing data

Abstract: Hyperspectral imaging, also known as imaging spectroscopy, is a major research area in remote sensing, since it captures valuable spatial and spectral information of a scene. A hyperspectral image can contain up to hundreds of spectral bands across the electromagnetic spectrum, providing a unique spectral signature for each image pixel. However, the analysis of these images is challenging due to their high feature dimensionality or the oftentimes low spatial resolution resulting in mixed pixels capturing multiple materials. In this context, sparse representation turned out to be a versatile tool for a variety of application like signal unmixing, anomaly detection, change detection and classification. This work will present the basic sparse representation formulation with different extensions which are adapted to specific application examples such as disease detection in close range images of plants and interpretation of satellite images.

Speaker : Alexander Shestopaloff, University of Toronto

Title: Sampling latent states for high-dimensional non-linear state space models with the embedded HMM method

Abstract: We propose a new scheme for selecting pool states for the embedded Hidden Markov Model (HMM) Markov Chain Monte Carlo (MCMC) method. This new scheme allows the embedded HMM method to be used for efficient sampling of state sequences in state space models where the state can be high-dimensional. Previously, embedded HMM methods were only applied to models with a one-dimensional state space. We demonstrate that using our proposed pool state selection scheme, an embedded HMM sampler can have similar performance to a well-tuned sampler that uses a combination of Particle Gibbs with Backward Sampling (PGBS) and Metropolis updates. The scaling to higher dimensions is made possible by selecting pool states locally near the current value of the state sequence. The proposed pool state selection scheme also allows each iteration of the embedded HMM sampler to take time linear in the number of the pool states, as opposed to quadratic as in the original embedded HMM sampler. We also consider a model with a multimodal posterior, and show how a technique we term “mirroring” can be used to efficiently move between the modes. (Joint work with Radford M. Neal).

Plenary Speaker : Nozer D. Singpurwalla, The City University of Hong

Title: Subjective Probability: Its Axioms and Acrobatics

Abstract: Probability is a foundation for statistical inference and its offshoot, data analysis, machine learning, signal processing. But what does probability mean? This is an intriguing question that has baffled mathematicians, philosophers, and physicists. The Bayesian claim to fame rests on the paradigm that Bayesian inference is coherent because it is solely based on the calculus of probability and nothing more. But where does probability come from and how must one interpret it?. The answer matters because it provides a framework for assigning meaning to the results of data analysis be it high dimensional or otherwise. This is an expository talk at a conversational level which leads to the position that the only logically defensible interpretation of probability is that it is personal.

Speaker : Ekaterina Smirnova, University of Montana

Title: Methods for sparsity adjustments in microbiome data

Abstract: In microbiome studies, a sample provides counts of DNA fragments, which are grouped into species-level operational taxonomic units (OTUs), and are summarized as a table of OTU counts. Sparsity of observations in the OTU vector because many OTUs are not observed in a large proportion of samples is a characteristic feature of microbiome data sets. This has serious implications on data analysis, where first line approaches, such as PCA, SVD, or ICA, are not designed to deal with such extreme levels of sparsity. Left unaddressed, this could dramatically bias the estimates of species diversity; pose a problem in methods for identifying differentially expressed OTUs across the groups of control and disease samples, and bias low dimensional representation of the data. Treating sparsity as missing data, which could have been observed given a better sampling effort, we propose methods to: 1) estimate the proportion of mass assigned to empty cells; and 2) distribute the estimated mass according to the rules of probability for each OTU. Finally, we evaluate the performance of these adjustments for improving methods originally developed for RNA-seq data, and dimension reduction-based techniques.

Plenary Speaker : Peter Song, University of Michigan

Title: Statistical Inference with Estimating Functions via the MapReduce Scheme

Coauthor(s): Ling Zhou

Abstract: The theory of statistical inference along with the strategy of divide-and-combine for large-scale data analysis has recently attracted considerable interest due to great popularity of the MapReduce scheme in the Hadoop platform. The key to the development of statistical inference lies in the method of combining results yielded from separately mapped data batches. One seminal solution based on the confidence distribution has been proposed in the setting of maximum likelihood estimation in the literature. We consider a more general inferential methodology based on estimating functions, of which the maximum likelihood is a special case. This generalization allows us to perform regression analyses of massive complex data via the MapReduce scheme, such as longitudinal data, survival data and quantile regression, which cannot be done using the maximum likelihood method. The proposed statistical inference inherits many key large-sample properties of estimating functions. In addition, because the proposed method is closely connected to the generalized method of moments (GMM) and Crowder?s optimality, its optimality over the existing methods is conveniently verified. Our method provides a unified framework for many kinds of statistical models and data types, which is illustrated via numerical examples in both simulation studies and real-world data analyses.

C Speaker : Miguel Villalobos, Universidad Anáhuac

Title: A Statistical and Machine Learning Model to Detect Money Laundering

Coauthor(s): José E. Silva

Money laundering poses significant challenges to financial institutions, not only due to its economic impact, but also, due to the volume and complexity of financial data, and the speed at which financial criminals evolve their strategies and tactics, to conceal the nature of their actions and avoid detection. While financial institutions in Mexico have stablished rulebased mechanisms to detect money laundering, these fail to detect efficiently the fastevolving tactics employed by criminals. Given the size of money laundering profits in Mexico, which, in 2016 were estimated to exceed $25 Billion USD, and motivated by the implementation of the System of Interbank Payments in Dollars, SPID for its name in Spanish, the Central Bank (Banxico) is now requiring Financial Institutions to comply with additional regulations, which demand the implementation of what Banxico calls an “Additional Risks Model”. This article presents a Statistical and Machine Learning Model to detect money laundering that addresses the Central Bank’s SPID regulation. The solution combines several models and finds, under prespecified criteria, the “best” model ensemble to minimize false positives, while enabling early detection of concept drift, and the incorporation of a case feedback loop of continuous learning. The article also presents a proposed functional architecture as well as a recommended general process flow to complement current bank anti money laundering (AML) systems and to retrofeed and adjust scoring at the customer, contract and transaction levels.

Mini simposioum

Organizer : Dr Luis Javier Álvarez. Institute of Mathematics, UNAM.

Title: Applying deterministic chaos theory for analyzing high-dimensional and complex data

We discuss some techniques of nonlinear time series analysis on the basis of deterministic chaos theory. Algorithms for data representation, dimension reduction and Lyapunov exponents estimation, are provided with emphasis on optimal parameters selection. Four illustrative cases study complement our approach.

Speaker : Dr Luis Javier Álvarez. Institute of Mathematics, UNAM.

Title: Dollar–peso exchange rate and its relation with macroeconomics indicators. A nonlinear analysis based on Lyapunov exponents.

Speaker : Igor Barahona, Institute of Mathematics, UNAM

Title: From digital images to nonlinear time series. Quantifying the degree of chaos in a volcanic eruption episode

Abstract: We discuss some techniques of nonlinear time series analysis on the basis of deterministic chaos theory. Algorithms for data representation, dimension reduction and Lyapunov exponents estimation, are provided with emphasis on optimal parameters selection. Four illustrative cases study complement our approach.

Speaker : Elizabeth Santiago Del Angel, Institute of Mathematics, UNAM

Title: Community structure for analyzing the flow capability in fracture networks in rock

Abstract: In this talk, a community approach used in complex network is applied for characterizing regions that have the capability of transporting efficiently any fluid in fracture networks in rocks, this by means of the detection of highly connected communities. Firstly, the fracture networks in rocks are transformed into complex networks and are formally described as graphs. For this problem, the interpretation of nodes and edges associated to the resulting graph is defined as follows. The nodes represent the intersections of the segments of the fractures, and the links are the segments that connect to the cross points. The methodology is based on the construction of communities (clustering) of nodes through a technique of modularity optimization and centrality measures. Basically, the modularity is used for the formation of communities, that is, subgraphs with high interconnectivity among their nodes, and low connectivity among communities, where the modularity is computed by the strength that each node has with respect to the nodes belonging to the different modules. Later, the modules are characterized by the application of centrality measures where the communities with the highest values are selected. Finally, the results of some samples of fracture networks in rocks are presented, showing the communities that represent the regions with the most relevant nodes into the whole network.

Speaker : Antonio Sarmiento Gálan, Institute of Mathematics, UNAM.

Title: Breathers and thermal relaxation as a temporal process

Coauthor(s): Alfonso Castrejón Pita

Abstract: Breather stability and longevity in thermally relaxing nonlinear arrays is investigated under the scrutiny of the analysis and tools employed for time series and state reconstruction of a dynamical system. We briefly review the methods used in the analysis and characterize a breather in terms of the results obtained with such methods. Our present work focuses on spontaneously appearing breathers in thermal Fermi-Pasta-Ulam arrays but we believe that the conclusions are general enough to describe many other related situations; the particular case described in detail is presented as another example of systems where three incommensurable frequencies dominate their chaotic dynamics sreminiscent of the Ruelle-Takens scenario for the appearance of chaotic behavior in nonlinear systems. This characterization may also be of great help for the discovery of breathers in experimental situations where the temporal evolution of a local variable like the site energy is the only available/measured data.

Posters

Presenter : Jesus Daniel Arroyo Relion, University of Michigan

Title: Overlapping Community Detection via Sparse PCA

Coauthor(s): Elizaveta Levina

Community detection in networks, the problem of finding groups of nodes that have more connections to each other than to the rest of the network, has received a lot of attention in the literature, but many methods only allow for a node to belong to exactly one community. In practice, nodes in a network may belong to multiple communities. Here we propose a new efficient algorithm for overlapping community detection based on sparse principal component analysis. The algorithm has a computational cost similar to that of estimating the largest eigenvectors of the adjacency matrix, and does not require an additional clustering step like spectral clustering techniques. We show that our method is consistent in selecting the community memberships under an overlapping version of the stochastic blockmodel and evaluate the method empirically on simulated and real world networks, showing good statistical performance and computational efficiency.

Presenter : Addy Bolivar-Cimé, Universidad Juárez Autónoma de Tabasco

Title: Geometric Representation of High Dimensional Data and Binary Discrimination Methods

Coauthor(s): Luis Miguel Cordova-Rodriguez

Four binary discrimination methods are studied in the context of HDLSS data with an asymptotic geometric representation, when the dimension increases while the sample sizes of the classes are fixed. We show that the methods Support Vector Machine, Mean Difference, Distance Weighted Discrimination and Maximal Data Piling have the same asymptotic behavior as the dimension increases. We study the consistent, inconsistent and strongly inconsistent cases in terms of angles of the normal vetors of the separating hyperplanes of the methods and the optimal direction for classification. A simulation study is done to assess the theoretical results.

Presenter : Rosa-Isela Hernández-Zamora, Universidad Autónoma de Nuevo Len (UANL)

Title: A clustering algorithm based on neighborhood search

Coauthor(s): Alvaro Eduardo Cordero Franco, UANL; José Fernando Camacho Vallejo, UANL; Diana Barraza Barraza, Universidad Juárez del Estado de Durango

Clustering is the process of partitioning observations into natural groups named clusters, each cluster containing very similar observation within them and, at the same time, different to the observations assigned to the rest of the cluster. Clustering algorithms, such as, k-means, expectation-maximization and density-based algorithms are suitable for certain structure of the data, and their performance can be negatively affected when the data consists of clusters that are of diverse shapes, densities, and sizes. In this work, we develop a new heuristic algorithm for clustering purposes. The algorithm starts with a solution constructed by a randomized greedy algorithm using a restricted list of candidates. After that, new solutions are found by neighborhood searching, where a neighborhood is defined as all the solutions that satisfy a criterion according to a selected metric. The proposed heuristic includes a parameter that regulates randomness and greediness of the selected observations to avoid bias in the results of clusterization. Different literature and new simulated dataset are present to evaluate the performance of the proposed clustering algorithm.

Presenter : Elizabeth Hou, University of Michigan

Title: Efficient Distributed Estimation of Inverse Covariance Matrices

Coauthor(s): Jesús Arroyo

In distributed systems, communication is a major concern due to issues such as its vulnerability or efficiency. In this paper, we are interested in estimating sparse inverse covariance matrices when samples are distributed into different machines. We address communication efficiency by proposing a method where, in a single round of communication, each machine transfers a small subset of the entries of the inverse covariance matrix. We show that, with this efficient distributed method, the error rates can be comparable with estimation in a non-distributed setting, and correct model selection is still possible. Practical performance is shown through simulations.