Peter Hall Conference 2019
UC Davis Statistics Department
May 10 & 11, 2019

Speaker Titles and Abstracts

Samory Kpotufe

Measuring transferability: some recent insights

Training data is often not fully representative of the target population due to bias in the sampling mechanism; in such situations, we aim to ’transfer’ relevant information from the training data (a.k.a. source data) to the target application. How much information is in the source data? How much target data should we collect if any? These are all practical questions that depend crucially on 'how far' the source domain is from the target. However, it remains generally unclear how to properly measure 'distance' between source and target.

In this talk we will argue that much of the traditional notions of 'distance' (e.g. KL-divergence, extensions of TV such as D_A discrepancy, and even density-ratios) can yield an over-pessimistic picture of transferability. In fact, much of these measures are ill-defined or too large in common situations where, intuitively, transfer should be possible (e.g. situations with structured data of differing dimensions, or situations where the target distribution puts significant mass in regions of low source mass). Instead, we show that a notion of 'relative dimension' between source and target (which we simply term the 'transfer-exponent') captures a continuum from easy to hard transfer. The transfer-exponent uncovers a rich set of situations where transfer is possible even at fast rates, helps answer questions such as the benefit of unlabeled data, and has interesting implications for related problems such as multi-task learning.

Finally, the transfer-exponent yields sharp guidance as to when and how to sample target data and guarantee fast improvement over source data alone. We illustrate these new insights through various simulations on controlled data, and on the popular CIFAR-10 image dataset.

The talk is based on work with Guillaume Martinet, and ongoing work with Steve Hanneke.

Shiva Kasiviswanathan

Contextual online false discovery rate control

Multiple hypothesis testing, a situation when we wish to consider many hypotheses, is a core problem in statistical inference that arises in almost every scientific field. In this setting, controlling the false discovery rate (FDR), which is the expected proportion of type I error, is an important challenge for making meaningful inferences. In this talk, we consider a setting where an ordered (possibly infinite) sequence of hypotheses arrives in a stream, and for each hypothesis we observe a p-value along with a set of features specific to that hypothesis. The decision whether or not to reject the current hypothesis must be made immediately at each timestep, before the next hypothesis is observed. We propose a new class of powerful online testing procedures, where the rejection thresholds are learned sequentially by incorporating contextual information and previous results. Any rule in this class controls online FDR under some standard assumptions. We will also discuss how our proposed procedures would lead to an increase of statistical power over a popular online testing procedure.

Wuchen Li

Learning via Wasserstein information geometry

In this talk, we review several differential structures from optimal transport (Wasserstein metric). Based on it, We will introduce the Wasserstein natural gradient in parametric models. The L2-Wasserstein metric tensor in probability density space is pulled back to the one on parameter space, under which the parameter space forms a Riemannian manifold. The Wasserstein gradient flows and proximal operator in parameter space are derived. We demonstrate that the Wasserstein natural gradient works efficiently in several machine learning models, including Boltzmann machine, generative adversary models (GANs) and variational Bayesian statistics.

Justin Solomon

Correspondence and optimal transport for geometric data processing

Correspondence problems involving matching of two or more geometric domains find application across disciplines, from machine learning to computer vision. A theoretical framework involving correspondence along geometric domains is optimal transport (OT). Dating back to early economic applications, the OT problem has received renewed interest thanks to its applicability to correspondence problems machine learning, computer graphics, geometry, and other disciplines. The main barrier to wide adoption of OT as a modeling tool is the expense of optimization in OT problems. In this talk, I will introduce efforts to make large-scale transport tractable over a variety of domains and application scenarios, helping transition OT from theory to practice. In addition, I will show how OT can be used as a unit in systems for correspondence problems involving the processing of geometrically-structured data.

Cun-hui Zhang

Factor models for high-dimensional tensor time series

Large tensor data are now routinely collected in a wide range of applications due to rapid development of information technologies and their broad implementation in our era. Often such observations are taken over time, forming tensor time series. In this paper we present a factor model approach for analyzing high-dimensional dynamic tensor time series and multi-category dynamic transport networks. Two estimation procedures are presented along with their theoretical properties and simulation results. Real applications are used to illustrate the model and its interpretations. This is joint work with Rong Chen and Dan Yang.

Patrice Koehl

Optimal transport at finite temperature

(joint work with Henri Orland, IPhT, CEA, Saclay, France, and Marc Delarue, Institut Pasteur, Paris, France)

We develop a new method for solving the discrete optimal transport (OT) problem using techniques adapted from statistical physics. We derive a strongly concave free energy function that captures the constraints of the optimal transport problem at a finite temperature. Its maximum defines an optimal transport plan, or registration between the two discrete probability measures that are compared, as well as a distance between those measures that satisfies the triangular inequality. This temperature dependent OT distance decreases monotonically to the standard OT distance, providing a robust framework for temperature scaling. I will illustrate applications of this framework to the problem of comparing images, as well as to the problem of protein fold recognition based on sequence information only.

Tamara Kolda

Stochastic optimization for large-scale tensor decomposition

Tensor decomposition is a fundamental unsupervised machine learning method in data science, with applications including network analysis and sensor data processing. Recently, the generalized canonical polyadic (GCP) low-rank tensor decomposition has extended tensor decomposition to loss functions besides squared error. For instance, GCP can use logistic loss or Kullback-Leibler divergence, enabling tensor decomposition for binary or count data. However, GCP tensor decomposition can be prohibitively expensive for large-scale and/or sparse tensors. In this talk, we conceptualize a stochastic optimization framework for efficient computation of the GCP decomposition. Our focus is on computing a stochastic gradient, including adaptations for stratified sampling. In particular, we motivate and show how to independently sample zeros and nonzeros for sparse tensors, in contrast to recommender system scenarios that sample only nonzeros. Our framework is amenable to missing data and distributed parallelism. We consider pragmatic questions such as how many samples to use for the stochastic gradient, and we apply it to real-world examples. This is joint work with David Hong (Michigan) and Jed Duersch (Sandia).

Ali Shojaie

Generalized matrix decomposition for two-way structured data: from exploratory analysis to inference

Two-Way structured data arise naturally in many scientific applications, from social and biological networks, to brain imaging and microbiome studies. Two common characteristics of these applications are that (i) distances among observations are informed by non-Euclidean measures and/or (ii) external information about relationships among covariates needs to be taken into account.

In a recent paper, Allen and Taylor (2014, JASA) extended the approach of Escoufier (1977) and proposed the Generalized Matrix Decomposition (GMD) as a natural alternative to classical dimension reduction methods, such as principal component analysis (PCA), which do not account for the row-and-column structure in the data. We discuss two important extensions of their work: (i) we propose the GMD biplot as an effective tool for exploratory data analysis in two-way structured data; and (ii) we develop GMD regression (GMDR) as an efficient estimation and inference framework for high-dimensional regressions with two-way structured data.

Mladen Kolar

Two-sample inference for high-dimensional Markov networks

Markov networks are frequently used in sciences to represent conditional independence relationships underlying observed variables arising from a complex system. It is often of interest to understand how an underlying network differs between two conditions. We develop methodology for performing valid statistical inference for difference between parameters of Markov network in a high-dimensional setting where the number of observed variables is allowed to be larger than the sample size. Our proposal is based on the regularized Kullback-Leibler Importance Estimation Procedure that allows us to directly learn the parameters of the differential network, without requiring for separate or joint estimation of the individual Markov network parameters. This allows for applications in cases where individual networks are not sparse, such as networks that contain hub nodes, but the differential network is sparse. We prove that our estimator is regular and its distribution can be well approximated by a Normal under wide range of data generating processes and, in particular, is not sensitive to model selection mistakes. Furthermore, we develop a new testing procedure for equality of Markov networks, which is based on a max-type statistics. A valid bootstrap procedure is developed that approximates quantiles of the test statistics. The performance of the methodology is illustrated through extensive simulations and real data examples.

Peter Bartlett

Overparameterization and benign overfitting in linear regression

Classical results that address the statistical performance of methods like deep neural networks involve a tradeoff between the fit to the training data and the complexity of the prediction rule. Deep learning seems to operate outside the regime where these results are informative, since deep networks can perform well even with a perfect fit to the training data. Motivated by the important challenge of understanding benign overfitting, we consider when a perfect fit to training data in high dimensional linear regression is compatible with good predictive accuracy. We give a characterization of gaussian linear regression problems for which the minimum norm interpolating prediction rule has near-optimal prediction accuracy. The characterization is in terms of two notions of effective rank of the data covariance. It shows that overparameterization is essential for benign overfitting in this setting: the number of directions in parameter space that are unimportant for prediction must significantly exceed the sample size.

Joint work with Philip Long, Gábor Lugosi, and Alexander Tsigler.

Misha Belkin

Fit without fear: from classical statistics to modern machine learning

“A model with zero training error is overfit to the training data and will typically generalize poorly" goes statistical textbook wisdom. Yet, in modern practice, over-parametrized deep networks with near perfect fit on training data still show excellent test performance.
As I will discuss in my talk, this apparent contradiction is key to understanding modern machine learning. While classical methods rely on the bias-variance trade-off where the complexity of a predictor is balanced with the training error, "modern" models are best described by interpolation, where a predictor is chosen among functions that fit the training data exactly, according to a certain inductive bias. Furthermore, classical and modern models can be unified within a single "double descent" risk curve, which extends the usual U-shaped bias-variance trade-off curve beyond the point of interpolation. This understanding of model performance delineates the limits of classical analyses and opens new lines of enquiry into computational, statistical, and mathematical properties of models. A number of implications for model selection with respect to generalization and optimization will be discussed.

Kenji Sagae

Multilingual analysis with language embeddings

Understanding the relationships, differences and similarities among the languages of the world is one of the central challenges in linguistics. Recent developments in computational linguistics, coupled with the availability of an abundance of text in various languages, have created the opportunity to examine these relationships from a data-driven perspective. Using large text collections in 20 languages, we characterize the grammar of each language as a vector that captures elements of the language's morphology and syntax. We show that these vectors, which are embeddings extracted from a multilingual neural language model, encode various known language-specific properties and language relationships, and can be used to improve multilingual language technology applications.

Purnamrita Sarkar

Overlapping clustering models, and one (class) SVM to bind them all

People belong to multiple communities, words belong to multiple topics, and books cover multiple genres; overlapping clusters are commonplace. Many existing overlapping clustering methods model each person (or word, or book) as a non-negative weighted combination of "exemplars" who belong solely to one community, with some small noise. Geometrically, each person is a point on a cone whose corners are these exemplars. This basic form encompasses the widely used Mixed Membership Stochastic Blockmodel of networks (Airoldi et al., 2008) and its degree-corrected variants (Jin et al., 2017), as well as topic models such as LDA (Blei et al., 2003). We show that a simple one-class SVM yields provably consistent parameter inference for all such models, and scales to large datasets. Experimental results on several simulated and real datasets show our algorithm (called SVM-cone) is both accurate and scalable.

James Sharpnack

Fused density estimation

In this talk, we introduce a method for nonparametric density estimation on geometric networks. We define fused density estimators as solutions to a total variation regularized maximum-likelihood density estimation problem. We provide theoretical support for fused density estimation by proving that the squared Hellinger rate of convergence for the estimator achieves the minimax bound over univariate densities of log-bounded variation. We reduce the original variational formulation in order to transform it into a tractable, finite-dimensional quadratic program. Because random variables on geometric networks are simple generalizations of the univariate case, this method also provides a useful tool for univariate density estimation. Lastly, we apply this method and assess its performance on examples in the univariate and geometric network setting. We compare the performance of different optimization techniques to solve the problem, and use these results to inform recommendations for the computation of fused density estimators. Joint work with Robert Bassett (Naval Postgraduate School).