SAHD Workshop
Duke University
July 27 - 29, 2015

Featured Talk Abstracts

The Latent Variable - Autoregressive Latent Trajectory (LV-ALT) model: a general framework for longitudinal data analysis

Ken Bollen, University of North Carolina, Chapel Hill

In recent years, longitudinal data have become increasingly relevant in many applications, heightening interest in selecting the best longitudinal model to analyze them. Too often traditional practices rather than substantive theory guide the specific longitudinal model selected for the analysis. This opens the possibility that alternative models might better correspond to the data. In this regard, Bollen and Curran (2004) developed the Autoregressive Latent Trajectory (ALT) model. It captures the desirable features of both latent growth curve and autoregressive models, having the ability to discriminate between these two approaches to model panel data. The purpose of this paper is to develop the Latent Variable ALT (LV-ALT) model as a generalization of the autoregressive latent trajectory model. We show how the LV-ALT under constraints can specialize to a wide variety of other longitudinal models. Hence, if theory or prior work dictate the model, then latent variable ALT is likely capable of specializing to that structure. On the other hand, if there is little guidance on the best model, then the LV-ALT provides a way to empirically compare a wide variety of models and determine the most appropriate for the data. The latent variable ALT model also provides a framework which reveals the connections between many longitudinal models that were previously considered as distinct.

Graph matching: relax or not?

Alex Bronstein, Tel Aviv University, Duke University

Graphs are a ubiquitous mathematical abstraction employed in numerous problems in science and engineering. Of particular importance is the need to find best structure-preserving matching of graphs. Since graph matching is a computationally intractable problem, numerous heuristics exist to approximate its solution. An important class of graph matching heuristics is relaxationtechniques based on replacing the original problem by a continuous convex program. Conditions for applicability or inapplicability of such convex relaxations are poorly understood. In this talk, I will show easy to check spectral properties characterizing a wide family of graphs for which equivalence of convex relaxation to the exact graph matching is guaranteed to hold.

A universal primal-dual convex optimization framework

Volkan Cevher, EPFL

This talk proposes a new primal-dual algorithmic framework for a prototypical constrained convex optimization template. The algorithmic instances of our framework are universal since they can automatically adapt to the unknown Holder continuity properties within the template. They are also guaranteed to have optimal convergence rates in the objective residual and the feasibility gap for each smoothness level. In contrast to existing primal-dual algorithms, our framework avoids the proximity operator of the objective function altogether. We instead leverage computationally

Last Modified: July 27, 2015

cheaper, Fenchel-type operators, which are the main workhorses of the generalized conditional gradient (GCG)-type methods. In contrast to the GCG-type methods, our framework does not require the objective function to be differentiable, and can also process additional general linear inclusion constraints. Our analysis technique unifies Nesterov's universal gradient methods and GCG-type methods to address the more broadly applicable primal-dual setting. We provide numerical evidence to demonstrate the scalability of our framework in diverse applications, from Quantum Tomography to Phase Retrieval and from clustering to matrix completion.

Solving random quadratic systems of equations is nearly as easy as solving linear systems

Yuxin Chen, Stanford University

This talk is concerned with solving quadratic systems of equations in n variables. This generally NP-complete problem has many applications ranging from combinatorial optimization to the famous phase retrieval problem. We demonstrate that one can solve unstructured random quadratic systems in n variables from O(n) equations in linear time, that is, in time proportional to reading/evaluating the constraints. This is achieved by attempting to minimize a non-convex objective as in the Wirtinger flow approach. However, there are several key distinguishing features, most notably, a distinct objective functional and novel update rules, which operate in an adaptive fashion and drop terms bearing too much influence on the search direction. These careful selection rules provide a tighter initial guess, better descent directions, and thus enhanced practical performance.

Sub-Nyquist sampling without sparsity

Yonina Eldar, Technion

In recent years there has been an explosion of work on exploiting sparsity in order to reduce sampling rates in a wide-range of applications. In this talk, we consider several examples in which sub-Nyquist sampling is possible without assuming any structure on the signal being sampled. This is possible due to the fact that we are not interested in direct recovery of the signal itself, but rather of some function of the signal. First, we consider sampling a signal when we are interested in recovering its power spectrum. Next, we develop the minimal sampling rates required to achieve minimal distortion when representing an arbitrary signal by quantized samples. Finally, we consider sampling of ultrasound signals where the goal is to create a beamformed image from the given samples. In all cases we show that sampling at rates much lower than the Nyquist rate are possible, despite the fact that no structure is assumed on the input signal.

A theory of neural dimensionality, dynamics, and measurement

Surya Ganguli, Stanford University

In many experiments, neuroscientists tightly control behavior, record many trials, and obtain trial-averaged firing rates from hundreds of neurons in circuits containing millions of behaviorally relevant neurons. Dimensionality reduction has often shown that such datasets are strikingly simple; they can be described using a much smaller number of dimensions (principal components (PCs)) than the number of recorded neurons, and the resulting projections onto these components yield a remarkably insightful dynamical portrait of circuit computation. This ubiquitous simplicity raises several profound and timely conceptual questions. What is the origin of this simplicity and its implications for the complexity of brain dynamics? Would neuronal datasets become more complex if we recorded more neurons? How and when can we trust dynamical portraits obtained from only hundreds of neurons in circuits containing millions of neurons? We present a theory that answers these questions, and test it using data from reaching monkeys. Overall, this theory yields a picture of the neural measurement process as a random projection of neural dynamics, conceptual insights into how we can reliably recover dynamical portraits in such under-sampled measurement regimes, and quantitative guidelines for the design of future experiments.

LASSO with nonlinear measurements

Babak Hassibi, California Institute of Technology

We consider estimating an unknown, but structured, signal x from a vector of nonlinear observations y = g(Ax+v)+w, where A is a known measurement matrix, v and and w are unknown noise vectors, and g(.) is a (possibly unknown) nonlinear function. Such measurements could arise in applications where the measurement device has nonlinearities and uncertainties. They could also arise by design, e.g., when the measurements are quantized. Motivated by the classical work of Brillinger, and more recent work by Vershynin and Plan, we consider estimating x using the generalized LASSO applied directly to the vector y. While this approach naively ignores the nonlinear function g(.), earlier work shows that when A has iid standard normal entries, this is a good estimator of x (up to a constant of proportionality). We considerably strengthen these results by obtaining explicit and precise expressions for the mean-square error of x as the dimension of x and the number of measurements grow. A main result is that the performance of LASSO, when applied to nonlinear measurements, is the same as that of LASSO for linear measurements, provided the measurement noise variance is suitably modified. Our expressions, for example, yield the precise error performance of LASSO when applied to quantized measurements. One interesting consequence is that the optimal quantizer that minimizes the mean square error of LASSO is the celebrated Lloyd-Max quantizer.

Correlation mining from massive data: high dimensional sampling regimes

Al Hero, University of Michigan

Graphical modeling with the Bethe approximation

Tony Jebara, Columbia University

Inference (a canonical problem in graphical modeling) recovers a probability distribution over a subset of variables in a given model. It is known to be NP-hard for graphical models with cycles and large tree-width. Learning (another canonical problem) reduces to iterative marginal inference and is also NP-hard. How can we efficiently tackle these problems in practice? We will discuss the Bethe free energy as an approximation to the intractable partition function. Heuristics like loopy belief propagation (LBP) are often used to optimize the Bethe free energy. Unfortunately, in general, LBP may not converge at all, and if it does, it may not be to a global optimum. To do marginal inference, we instead explore a more principled treatment of the Bethe free energy using discrete optimization. We show that in attractive loopy models we can find the global optimum in polynomial time even though the resulting landscape is non-convex. To generalize to mixed loopy models, we use double-cover methods that bound the true Bethe global optimum in polynomial time. Finally, to do learning, we combine Bethe approximation with a Frank-Wolfe algorithm in the convex dual which circumvents the intractable partition function. The result is a new single-loop learning algorithm which is more efficient than previous double-loop methods that interleaved iterative inference with iterative parameter updates. We show applications of these methods in friendship link recommendation, in social influence estimation, in computer vision, and in power networks. We also combine the approaches with sparse structure learning to model several years of Bloomberg data. These graphical models capture financial and macro-economic variables and their response to news and social media topics.

Modeling ordered data by counting inversions

Marina Meila, University of Washington

How do we define and estimate "exponential family models" when data comes in the form of permutations, partial rankings, or other objects with rich combinatorial structure? I will start with the simple "consensus ranking" problem and describe how combinatorics, optimization and statistics can be used to build interpretable models and algorithms that are efficient in the presence of consensus.

At the center of the results is the *code* of a permutation. This natural mathematical way to represent a permutation is key both to fast computing and to better understanding the models we create. Yet, estimating the models from data leads to combinatorial optimization problems which are often NP-hard. The talk will introduce a class of algorithms to attack these problems, that are exact but intractable in the worst case, and will demonstrate that they are effective on real-world examples.

Joint work with Alnur Ali, Raman Arora, Le Bao, Jeff Bilmes, Harr Chen, Bhushan Mandhani, Chris Meek, Brendan Murphy, Kapil Phadnis, Arthur Patterson.

Semidefinite programming relaxations for graph estimation

Andrea Montanari, Stanford University

How sparsity and L1 optimization impacts "continuous" applied mathematics, physics and engineering

Stan Osher, University of California, Los Angeles

Scalable Bayesian nonparametric dictionary learning

John Paisley , Columbia University

We present a stochastic EM algorithm for scalable dictionary learning with the beta-Bernoulli process, a Bayesian nonparametric prior that learns the dictionary size in addition to the sparse coding of each signal. The core EM algorithm provides a new way for doing inference in nonparametric dictionary learning models and has a close similarity to other sparse coding methods such as K-SVD. Our stochastic extension for handling large data sets is closely related to stochastic variational inference, with the stochastic update for one parameter exactly that found using SVI. We show our algorithm compares well with K-SVD and total variation minimization on a denoising problem using several images.

Elementary estimators for "big-p" statistical models

Pradeep Ravikumar, University of Texas, Austin

We consider the problem of learning Big-p or high-dimensional statistical models, where the number of variables, typically denoted by p, could be potentially larger than the number of observations. This class of problems has attracted considerable attention over the last decade, with state of the art statistical estimators based on solving regularized convex programs. Scaling these typically non-smooth convex programs to the very large-scale problems of the Big Data era comprises an ongoing and rich area of research.

In contrast to this two-stage approach of first devising statistically efficient estimators, and then devising computationally efficient optimization methods to solve these estimators, we suggest the development of what we call comptastical estimators: that consider statistical as well as computational constraints at the outset. As an instance of this, we consider elementary closed-form estimators that can be computed in a very small number of simple steps. But can such elementary estimators come with statistical guarantees that are comparable to state of the art regularized likelihood estimators which require computationally intensive iterative optimization algorithms? Surprisingly, we show that under certain conditions, this question can be answered in the affirmative. We analyze our estimators in the high-dimensional setting, and moreover provide empirical corroboration of their statistical and computational performance guarantees.

Joint work with Eunho Yang and Aurelie Lozano.

Sparse if-then rule models

Cynthia Rudin, Massachusetts Institute of Technology

I aim to produce sparse and accurate logical machine learning models to compete with decision tree algorithms like CART and C5.0. My models are decision lists, which consist of a series of IF...THEN... statements (for example, ”if high blood pressure, then stroke”) that discretize a high-dimensional, multivariate feature space into a series of simple, readily interpretable decision statements. I will introduce a generative model called Bayesian Rule Lists that yields a posterior distribution over possible decision lists. It employs a novel prior structure to encourage sparsity. It does not use greedy splitting and pruning like decision trees. Our experiments show that Bayesian Rule Lists has (i) predictive accuracy on par with the current top algorithms for prediction in machine learning, (2) highly sparse solutions, (3) reasonable computational tractability, even for problems with thousands of features and observations.

Iteratively reweighted ℓ_1 approaches to sparse composite regularization

Phil Schniter, The Ohio State University

Motivated by the observation that a given signal $\boldsymbol{x}$ admits sparse representations in multiple dictionaries $\boldsymbol{\Psi}_d$ but with varying levels of sparsity across dictionaries, we propose two new algorithms for the reconstruction of (approximately) sparse signals from noisy linear measurements. Our first algorithm, Co-L1, extends the well-known lasso algorithm from the L1 regularizer $\|\boldsymbol{\Psi x}\|_1$ to composite regularizers of the form $\sum_d \lambda_d \|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1$ while self-adjusting the regularization weights λ_d. Our second algorithm, Co-IRW-L1, extends the well-known iteratively reweighted L1 algorithm to the same family of composite regularizers. We provide several interpretations of both algorithms: i) majorize-minimization (MM) applied to a non-convex log-sum-type penalty, ii) MM applied to an approximate ℓ_0-type penalty, iii) MM applied to fully-Bayesian inference under a particular hierarchical prior, and iv) variational expectation-maximization (VEM) under a particular prior with deterministic unknown parameters. A detailed numerical study suggests that our proposed algorithms yield significantly improved recovery SNR when compared to their non-composite L1 and IRW-L1 counterparts.

False discovery rate smoothing

James Scott, University of Texas, Austin

We present false discovery rate smoothing, an empirical-Bayes method for exploiting spatial structure in large multiple-testing problems. FDR smoothing automatically finds spatially localized regions of significant test statistics. It then relaxes the threshold of statistical significance within these regions, and tightens it elsewhere, in a manner that controls the overall false-discovery rate at a given level. This results in increased power and cleaner spatial separation of signals from noise. The approach requires solving a non-standard high-dimensional optimization problem, for which an efficient augmented-Lagrangian algorithm is presented. One of the key subroutines in our algorithm is the solution of a graph-fused lasso problem on an arbitrary graph, for which a new algorithm (of independent interest) is presented. We demonstrate that FDR smoothing exhibits state-of-the-art performance on simulated examples. We also apply the method to a data set from an fMRI experiment on spatial working memory, where it detects patterns that are much more biologically plausible than those detected by existing FDR-controlling methods. This is joint work with Wesley Tansey, Sanmi Koyejo, and Russ Poldrack.

High-dimensional biological sequences through simple models and posterior diagnostics

Marc Suchard, University of California, Los Angeles

High-dimensional statistical methods for comparing the relative rates of sequence mutation over time maintain a central role in annotating immunologically important genes through a evolutionary process called positive selection. To identify selection, researchers often estimate the ratio of these relative rates non synonymous to synonymous mutations at individual alignment sites. A reliable way to perform such estimation fits a codon-based evolutionary model that captures heterogeneity of dN / dS values across sites. Unfortunately, the large state space of possible codons makes codon-based models computationally prohibitive for massive data sets, containing hundreds or even thousands of sequences. Alternatives crudely estimate the numbers of synonymous and nonsynonymous substitutions at each site and use these counts to identify positively selected sites. Although these counting approaches scale well to massive data sets, they fail to account for ancestral state reconstruction uncertainty and to provide site-specific estimates. We propose a hybrid solution that borrows the computational strength of counting methods, but augments these methods with empirical Bayes modeling of synonymous and nonsynonymous substitution rates. The result is a fast and reliable method capable to identify sites under positive selection and to estimate site-specific dN / dS values in large data sets using posterior diagnostics. Importantly, our hybrid approach, set in a Bayesian framework, integrates over the posterior distribution of phylogenies and reconstructions to quantify uncertainty about site-specific dN / dS estimates. Comparisons with mixture codon-based models demonstrate that this hybrid method competes well with these more principal statistical procedures and in some cases even outperforms them. We illustrate the utility of our method using human immunodeficiency virus and influenza examples.

Parallel-ℓ_0, a fully parallel algorithm for combinatorial compressed sensing

Jared Tanner, University of Oxford

We consider the problem of solving for the sparsest solution of large underdetermined linear system of equations where the matrix is the adjacency matrix of an expander graph corresponding with at most d neighbours per node. We present a new combinatorial compressed sensing algorithm with provable recovery guarantees, fully parallel with computational runtime less than traditional compressed sensing algorithms, and able to recover sparse signals beyond l1-regularization. This work is joint with Rodrigo Mendoza-Smith.

Applied random matrix theory

Joel Tropp, California Institute of Technology

Non-Convex, Bayesian-inspired algorithms for sparse and low-rank estimation

David Wipf, Microsoft Research

Sparse estimation and related matrix rank minimization have emerged as important tools in diverse fields including computational neuroscience, signal/image processing, computer vision, and machine learning. Both for computational efficiency and theoretical accessibility, convex algorithms have come to dominate the practical optimization landscape. The typical recipe, which has been iterated in numerous application domains, involves first postulating some putatively ideal, combinatorial cost function favoring sparsity, low-rank, or both, forming the tightest convex relaxation, and then assessing equivalence conditions whereby the convex approximation will lead to solutions sufficiently close to the original non-convex problem. In particular, the recent popularity of compressive sensing and robust PCA have expedited the acceptance of such convex surrogates, largely because the requisite equivalence conditions can be naturally satisfied using, for example, simple randomized or incoherent designs. But in reality, many practical applications of sparsity and low rank matrices do not benefit from this luxury; rather, because of intrinsic correlations in the signal dictionary (or related structure in rank minimization problems), convex algorithms must be used in regimes where theoretical support no longer holds. Moreover, in some situations it has been shown that convex relaxations are in fact provably bad. Consequently, non-convex algorithms, while perhaps theoretically less accommodating, may nonetheless produce superior results. Here we will examine non-convex estimation algorithms, many of which originate from Bayesian machine learning ideas, that thrive in environments where more popular convex alternatives fail. In all cases, theoretical model justification will be provided independent of any assumed prior distributions. Illustrative examples related to robust PCA and rank minimization under affine constraints will be considered for evaluation purposes.

Network analysis and nonparametric statistics

Patrick Wolfe, University College London

Networks are ubiquitous in today's world. Any time we make observations about people, places, or things and the interactions between them, we have a network. Yet a quantitative understanding of real-world networks is in its infancy, and must be based on strong theoretical and methodological foundations. The goal of this talk is to provide some insight into these foundations from the perspective of nonparametric statistics, in particular how trade-offs between model complexity and parsimony can be balanced to yield practical algorithms with provable properties.