STUDENT SPEAKERS ABSTRACTS:

Andriy Derkach, Ph.D. Candidate, Department of Statistics
Robust Association Tests for Rare Genetic Variants

Many association tests have been proposed for rare variants, however, there is much confusion about the practical choice of a good test due to limited insights on the underlying genetic models. We recently showed that previously proposed methods can be categorized as either linear statistics that have high power against very specific alternative hypotheses or quadratic statistics that are designed to have good power over wide ranges of alternatives. However, neither class of tests consistently outperforms the other or provides comparable power, a conclusion that has also been drawn by several other authors. To achieve robustness, we propose hybrid statistics that borrow strength from the two classes of tests using Fisher's method and the minimal p-value approach of combining p-values from the complementary linear and quadratic tests. Extensive simulation studies show that both methods are robust across genetic models with varying proportions of causal, deleterious and protective variants, and variant frequency and effect size. Moreover, in situations when both the linear and quadratic tests have some power, Fisher's method consistently outperforms the minimal p-value approach and has better power than both the linear and quadratic tests.

Paul Nguyen, Postdoctoral Fellow, Dalla Lana School of Public Health & Cancer Care Ontario
Mapping Cancer Risk in Southwestern Ontario with Changing Census Boundaries

Mapping cancer risk often involves working with data aggregated in space and time due to data being collected at an area level, such as postal codes or health regions. Further, studying rare cancers requires using data collected over a long time period, introducing the problem of boundaries of census regions changing over time. Adopting a local likelihood framework, we present a local-EM algorithm that permits multiple maps with different census boundaries to be combined. This algorithm estimates the risk surface on a tessellation where the cumulative intersections of the various maps are distinct regions. We also implement parametric bootstrapping procedures to make inference on the uncertainty of these risk surface estimates. These methods were demonstrated on mesothelioma lung cancer data collected in the counties of Lambton and Middlesex of Southwestern Ontario, Canada between 1985 and 2007. We identified several areas of elevated risk throughout Lambton County and areas of low risk throughout Middlesex County.

(Joint work with Patrick E. Brown, University of Toronto and Cancer Care Ontario and Jamie Stafford, University of Toronto)

Nitish Srivastava, M.Sc. Student, Department of Computer Science
Multimodal Learning with Deep Belief Nets

Information in the real world comes from multiple dependent data modalities. We propose a Deep Belief Network (DBN) architecture for learning joint features over multimodal data. Our model defines a joint probability distribution over the space of multimodal inputs and allows sampling from the conditional distributions over each data modality. Our experimental results on a collection of tagged images show that the Multimodal DBN can learn a good generative model of the joint space of image and text inputs that is useful for filling in missing data. We further demonstrate that the features discovered by the Multimodal DBN significantly outperform SVMs and LDA on various discriminative tasks.

(Joint work with Ruslan Salakhutdinov, University of Toronto)

Laurent Charlin, Ph.D. Candidate, Department of Computer Science
Matrix Completion for Recommendation Systems

The burgeoning interest in recommender systems has led to a plethora of techniques for predicting user preferences or ratings for unseen items. Rating predictions methods, which can be framed as a matrix completion problems, have attained impressive performance. In practice, however, recommendations must not only account for user preferences in isolation; one usually has to tradeoff preferences for recommended items with various constraints or objectives. In this talk I will focus on match-constrained recommendation, where the quality of a set of recommendations or matching is measured relative to constraints or objectives that account for the entire set of users to whom an item is recommended, the entire set of items recommended to a single user, or both. Furthermore, I will show how taking the matching constraints into account for query selection (active learning) is beneficial.

(Joint work with Richard Zemel and Craig Boutilier, University of Toronto).

KEYNOTE SPEAKERS ABSTRACTS:

Michael I. Jordan
Professor, Department of EECS and Statistics, University of California, Berkeley

At the Interface of Statistics and Computation: A Scalable Bootstrap, Matrix Completion and Stein's Method

There are many issues remaining to be addressed, or even formulated, at the interface of statistics and computation. The ideal would be classes of inferential procedures that provide theoretical guarantees that statistical risk decreases as the number of data points grows, without bound, even when we impose a fixed computational budget. We are far from such a framework in theory, and in practice massive data sets are often severely subsampled with uncontrolled effects on the quality of inference. In this presentation, I discuss some initial forays into this problem domain. The first is an exploration of the bootstrap in the regime of very large data sets, where it is computationally infeasible to obtain bootstrap resamples. I present a new procedure, the ``bag of little bootstraps," which inherits the favorable theoretical properties of the bootstrap but is also computationally scalable. The second is an exploration of divide-and-conquer strategies for matrix completion. Here the theoretical support is provided by concentration theorems for random matrices, and I present a new approach to this problem based on Stein's method.

(Joint work with Ariel Kleiner, Lester Mackey, Purna Sarkar, Ameet Talwalkar, Richard Chen, Brendan Farrell and Joel Tropp).

David Dunson
Professor, Department of Statistical Science, Duke University

Nonparametric Bayesian Learning from Big Data

In modern applications, data sets tend to be big and highly structured, with large p, small n problems commonly encountered. In such settings, sparse representations of the data are crucial and there is a rich frequentist literature focused on inducing sparsity through penalization (typically L1). Motivated by genetic epidemiology and imaging applications, we instead develop nonparametric Bayesian methods that avoid parametric assumptions while favoring low-dimensional representations of complex high-dimensional data. In this talk, the particular focus is on Bayesian probabilistic tensor factorizations, which generalize low rank matrix factorizations, such as SVD, to higher orders. The framework accommodates general joint modeling of object data of different types (images, text, categorical, real, etc) but for simplicity we focus on two applications: (1) high-dimensional multivariate categorical data analysis (contingency tables); (2) estimation of lower dimensional manifolds from point cloud data. In the contingency table case, we propose a collapsed Tucker factorization and develop associated methods for testing of associations and interactions in huge sparse tables. In the manifold learning case, we propose a tensor product of basis functions for estimating 3d closed surfaces. In both settings, theoretical results are provided on large support and asymptotic properties & efficient computational methods are developed, which scale to large data sets. Applications in genetics and imaging are presented.

Marina Meila
Associate Professor, Department of Statistics, University of Washington

(Bayesian) Statistics with Rankings

This talk is concerned with summarizing -- by means of statistical models -- of data that expresses preferences. This data is typically a set of rankings of n items by a panel of experts; the simplest summary is the "consensus ranking", or the "centroid" of the set of rankings. Such problems appear in many tasks, ranging from combining voter preferences to boosting of search engines.

We study the problem in its more general form of estimating a parametric model known as the Generalized Mallows (GM) model. I will present an exact estimation algorithm, non-polynomial in theory, but extremely effective in comparison with existing algorithms.

Then I will introduce the infinite GM model (IGM), corresponding to"rankings" over an infinite set of items, and show that this model is both elegant and of practical significance.

From a statistical perspective, we show that the (I)GM model is an exponential family model, present the Dirichlet Process Mixture of GM models and other statistical results.

Connections with other models over permutations will be discussed too.

(Joint work with Harr Chen, Alnur Ali, Bhushan Mandhani, Le Bao, Kapil Phadnis, Arthur Patterson, Brendan Murphy and Jeff Bilmes).

.