Stochastic Systems, Probability, and Other Mathematical Aspects of Data Science
Les systèmes stochastiques, les probabilités et d'autres aspects mathématiques de la science des données
(Org: **Martin Lysy** (Waterloo))

MEIXI CHEN, University of Waterloo
Decoding Neural Population Dynamics Through Latent Factor Models

The human brain contains some hundred billion nerve cells (a.k.a. neurons) which communicate through electrochemical waves called spikes. A sequence of consecutive spikes from a neuron is called a spike train, which encode information about firing rates. Over the past few decades, mathematical and statistical models for neuronal activities have played an important role in helping neuroscientists shed light on neuroscientific phenomena such as the interactions among multiple neurons over time. However, scalability and interpretability of these models are still a challenge in computational/theoretical neuroscience. We present a novel latent factor model for studying the spike train interactions of multiple neurons recorded simultaneously. In the proposed model, the activities of the neuron population are described by correlated Wiener processes, which themselves depend on a small number of latent factors determining the neuronal clustering. We demonstrate how to tackle the computational challenges of high dimensional integration of latent variables and large matrix inversions. We show that our model is highly scalable and can accurately recover neuronal clusters when applied on simulated data. Finally, we apply our model to a set of experimental data obtained from rats' medial prefrontal cortex.

SANJEENA DANG, Carleton University
Clustering matrix-variate count data

Three-way data structures or matrix-variate data are frequent in biological studies. In RNA sequencing, three-way data structures are obtained when high-throughput transcriptome sequencing data are collected for n genes across p conditions at r occasions. Matrix variate distributions offer a natural way to model three-way data and mixtures of matrix variate distributions can be used to cluster three-way data. Clustering of gene expression data is carried out as means of discovering gene co-expression networks. In this work, a mixture of matrix variate Poisson-log normal distributions is proposed for clustering read counts from RNA sequencing. By considering the matrix variate structure, the number of covariance parameters to be estimated is reduced and the components of resulting covariance matrices provide a meaningful interpretation. We propose three different frameworks for parameter estimation - a Markov chain Monte Carlo based approach, a variational Gaussian approximation-based approach, and a hybrid approach. The models are applied to both real and simulated data, and we demonstrate that the proposed approaches can recover the underlying cluster structure. In simulation studies where the true model parameters are known, our proposed approach shows good parameter recovery.

OSVALDO ESPIN GARCIA, Western University
Using genetic algorithms in the design of two-phase studies

The two-phase study is a cost-effective way to leverage available information in phase 1 of the study to strategically select a most informative subset in phase 2. Expensive information is then collected in the phase 2 subset only, reducing the overall cost. Last, information from both study phases is jointly analyzed by performing statistical inference. Two-phase studies provide a desirable trade-off by economically and strategically using limited resources such as budget without compromising statistical performance by leveraging missing-by-design data methods. A main challenge lies in identifying such an informative subset, which can rely on both outcome and (inexpensive) phase 1 covariates. Genetic algorithms (GAs) are stochastic optimization techniques that mimic nature's evolutionary process. Often used in discrete optimization, GAs offer wide flexibility and ease of implementation. However, these advantages also come with some obstacles, for instance lack of a unique solution and unclear converge criteria are two of the main critiques of these approaches. In this talk, I will present my work on using a GA to identify an informative sample for two-phase fine-mapping studies. I will discuss some of the mathematical and computational challenges found as well as potential future work.

JESSE GRONSBELL, University of Toronto
Leveraging electronic health records for data science

The adoption of electronic health records (EHRs) has generated massive amounts of routinely collected medical data with potential to improve our understanding of healthcare delivery and disease processes. However, the analysis of EHR data remains both practically and methodologically challenging as it is recorded as a byproduct of clinical care and billing, and not for research purposes. For example, outcome information, such as presence of a disease or treatment response, is often missing or poorly annotated in patient records, which brings challenges to statistical learning and inference. In this talk, I will discuss predictive modeling in settings with an extremely limited amount of outcome information and demonstrate the advantages of semi-supervised learning methods that incorporate large volumes of unlabeled data into model estimation and evaluation.

CAMERON JAKUB, University of Guelph
The Angle Process in Deep Neural Networks and the Bessel Numbers of the Second Kind

A mysterious property of deep neural networks is that, on initialization, the inputs tend to get more and more correlated as the network gets deeper and deeper. In this talk, we investigate fully connected networks with the ReLU non-linearity, and we discover how the angle between any two inputs evolves as a function of network depth. The formula involves the joint moments of the ReLU function applied to Gaussian random variables. We take a combinatorial approach to explicitly solve for these joint moments and doing so reveals a surprising connection to the Bessel numbers of the second kind. We are able to accurately predict the joint distribution of each layer on initialization given only the inputs into the network. The formula becomes more exact as the width of each layer tends to infinity. Both the mathematical theory behind the formula as well as simulations to validate our results are presented.

HANNA JANKOWSKI, York University
The isotonic single index model under fixed and random designs

To quote L. Wasserman, probability theory asks "Given a data generation process, what are the properties of the outcomes?", while statistics ask "Given the outcomes, what can we say about the process that generated the data?", where the latter question can be viewed as solving an inverse problem.

In the first part of the talk I will motivate shape-constrained methods of estimation in the context of solving the inverse problem while striking a balance between robustness (bias) and efficiency (variance), the two key sources of error in statistical estimation. I will then discuss some recent results on the monotone single index model, a dimension reduction model. This is joint work with Fadoua Balabdaoui (ETHZ) and Cecile Durot (Paris X).

In the monotone single index model a real response variable Y is linked to a multivariate covariate X through the relationship $E[Y|X = x] = f_0(\alpha_0^T x)$ almost surely. Both the ridge function, f_0, and the index parameter, α_0, are unknown and the ridge function is assumed to be monotone. Under random design, we show that the rate of convergence of the estimator of the bundled function $f_0(\alpha_0^T x)$ is $n^{1/3}$. For the fixed design setting, we show that the rate of convergence is parametric, as expected. Throughout the talk I will illustrate the methodology on several real data sets.

LOBNA KHADRAOUI, Ottawa University
Large graph limit for an epidemic evolution process in random network with heterogeneous age, variant and connectivity

We consider a stochastic epidemic model on a random network in which each node corresponds to an individual, and each individual is classed as either Susceptible, Infectious, or Recovered ("SIR"). While the nodes are fixed, the edges evolve randomly. Volz [1] used popular heuristics to derive corresponding deterministic ordinary differential equations as the population size goes to infinity, following the work of Newmann [2]. Later, Decreusefond [3] proved weak convergence to this ODE system in the large-population limit. In this talk, we will present a similar convergence result for a more general model including a "Death" state (compartment) together with some additional variables: degree, age and disease variant. Of particular interest, the continuous nature of the age and variant variables leads to a limiting system of partial differential equations (PDEs) in place of

the ODEs considered by earlier authors. In order to prove weak convergence, a rescaled process is used, and the infinitesimal generator and the martingal properties are provided. Finally, we propose several numerical simulations in order to illustrate the convergence of compartment sizes for a large population, the distribution of age variable and the evolution of waves of the disease using the variant variable.

DAN ROY, University of Toronto

Admissibility is Bayes optimality with infinitesimals

In this talk, I'll summarize recent work exploiting tools in mathematical logic to resolve longstanding open problems in statistical decision theory. I'll focus on an exact characterization of admissibility in terms of Bayes optimality in the nonstandard extension of the original decision problem, as introduced by Duanmu and Roy (Ann. Statist. 49(4): DOI:10.1214/20-AOS2026). Unlike the consideration of improper priors or other generalized notions of Bayes optimality, the nonstandard extension is distinguished, in part, by having priors that can assign "infinitesimal" mass in a sense that is made rigorous using results from nonstandard analysis. With these additional priors, we find that, informally speaking, a decision procedure δ_0 is admissible in the original statistical decision problem if and only if, in the nonstandard extension, the nonstandard extension of δ_0 is Bayes optimal among the extensions of standard decision procedures with respect to a nonstandard prior assigning at least infinitesimal mass to every standard parameter value. We use this theorem to give further characterizations of admissibility, one related to Blyth's method and another related to a condition due to Stein that characterizes admissibility under regularity. Our results imply that Blyth's method is a sound and complete method for establishing admissibility. Buoyed by this result, we revisit the univariate two-sample common-mean problem, and show that the Graybill–Deal estimator is admissible among a certain class of unbiased decision procedures.

Joint work with Haosui Duanmu and David Schrittesser.

MOHAN WU, University of Waterloo

Parameter Inference for Differential Equations Using the Kalman Filter

Parameter inference for ordinary differential equations (ODEs) involves the evaluation of the likelihood function for each ODE solution. While this solution is typically approximated by deterministic algorithms, new research indicates that probabilistic solvers produce more reliable estimates by better considerations of numerical errors. A particularly effective probabilistic method, Fenrir, uses Kalman filtering in an efficient manner to obtain the ODE solution. However, it is constrained by the assumption of normally distributed observed data. We extend this method by allowing for observations not necessarily normally distributed. Several examples are used to demonstrate the effectiveness of this approach.