AARMS Workshop
Statistical Learning and Health Data Analytics

October 15, 2017, Fredericton, New Brunswick, Canada
Room Tilley 404, University of New Brunswick

Session 1: Plenary Talk
Chair: Gary Sneddon, Mount Saint Vincent University

9:00 – 9:50am
Overview of Statistical Learning Methods
Hugh Chipman (President, Statistical Society of Canada; Acadia University)

9:50-10:00am **Coffee Break**

Session 2: Health Data Analytics
Chair: Tariqul Hasan, University of New Brunswick

10:00 -10:40am
Health administrative data in New Brunswick: facilitating evidence based policy development
Ted McDonald (Director, NB Institution for research, data and training (NB-IRDT); Economics, University of New Brunswick)

10:40-11:20am
Leveraging Medical Lexicons to Improve Health Data Analytics
Samuel Stewart (Director, Health Data Nova Scotia (HDNS); Department of Community Health and Epidemiology, Dalhousie University)

11:20-1:00pm
MHSCC proposal and AARMS CRG Updating (20 minutes**)**
Lunch Break (McConnell Hall)

Session 3: Statistical Learning Methods for Dependent Data with Applications in Medicine
Chair: Guohua Yan (University of New Brunswick)

1:00 -1:30pm
Graph-based change-point test for high-dimensional data
Amy Wu (York University)

1:30-2:00pm
Statistical learning methods in emergency diagnoses and human microbiome analysis
Hong Gu (Dalhousie, Dalhousie)

2:00-2:30pm
Analysis of dependent data of various correlation structures
Renjun Ma (University of New Brunswick)

2:30-3:00pm **AARMS CRG Discussion Continued**

Talks

Overview of Statistical Learning Methods
Hugh Chipman

Abstract: Like machine learning, the field of statistical learning seeks to "learn from data". I will review some of the central ideas that identify statistical learning, including regularization and the bias-variance trade-off, resampling methods such as cross-validation for selecting the amount of regularization, the role of a probability model for data, and quantification of uncertainty. These ideas will be discussed in the context of popular recent approaches to supervised learning.

Health administrative data in New Brunswick: facilitating evidence based policy development
Ted McDonald

Abstract: This presentation will outline recent developments in researcher access to individual-level population data through two secure facilities on the UNB Fredericton campus. Following a review of data holdings in each facility, access procedures will be briefly reviewed and then recent and ongoing research using those data will be showcased. The first facility is the NB Institute for Research, Data and Training, a data custodian and research data centre for provincial administrative data from the Departments of Health, Social Development and other agencies. The second is the NB Statistics Canada Research Data Centre, a repository containing a wide range of Census and survey files as well as an increasing array of complex linked administrative data files.

Leveraging Medical Lexicons to Improve Health Data Analytics
Samuel Stewart

Abstract: At Health Data Nova Scotia (HDNS) we work with a variety of administrative and clinical health data sets, with a focus on facilitating data driven research within Nova Scotia. Through its databases HDNS analysts work with many different data types, most notably working extensively with medical taxonomies such as those contained within the Unified Medical Language System (UMLS).

The purpose of this presentation will be to (a) present the concepts and added value of medical taxonomies to the health data analytics community, and (b) present novel analytic methods for leveraging the network structure that arises out of medical taxonomies to improve our health data analytics methods. Two novel analytic methods, the Balanced Information Content Genealogy Model (BICGM) and a Generalized Vector Space Model (GVSM) will be presented, and as a case study they will be applied to two medical discussion forums to identify similarity patterns and clusters of similar users.

Graph-based change-point test for high-dimensional data
Amy Wu

Abstract: A change-point test for high-dimensional data is presented by using a Bayesian-type statistic based on the shortest Hamiltonian path, and the change-point is estimated by using ratio cut. A permutation procedure is applied to approximate the significance of Bayesian-type statistics. The change-point test is proven to be consistent, and an error probability in change-point estimation is provided. The test is powerful against alternatives with a shift in variance and is accurate in change-point estimation, as shown in simulation studies. Its applicability in tracking cell division is illustrated.

Statistical learning methods in emergency diagnoses and human microbiome analysis
Hong Gu

Abstract: It is well known that the human microbiome plays an important role in maintaining a healthy state. Second generation sequencing has accumulated a large amount of data which holds the promise of solutions for many scientific questions. These include how the microbiome interacts with the host genome and environment; how the microbiome community is structured; how this structure changes through time; and how these changes are related to human health. These data have many features that make statistical analysis challenging. I will introduce some data analysis methods developed by my research group and collaborators.

A separate topic is on machine learning diagnoses of emergency department patients. Our aim is to use the data from an existing electronic database, which is currently underused, to build an automated diagnosis tool for physicians. This tool will provide a list of the most likely diagnoses for each patient. The physician could refer to this list to check for any possibilities that were overlooked. This diagnostic tool could also assess the most informative additional tests, which will reduce the time and resources used on unnecessary diagnostic tests. Our first attempt focusses on patients presenting with abdominal pain. One statistical issue is that different medical test results are available for different patients. We propose a model framework that can combine different models for groups of patients with different variables and compare the results with multiple imputation.

Analysis of dependent data of various correlation structures
Renjun Ma

Abstract: As the objective of this Collaborative Research Group is to bring researchers in the areas of statistical learning and dependent data to work together, in this talk, we present an overview of our team work on analysis of dependent data of various correlation structures. Specifically, we have developed best linear unbiased predictor approach to Tweedie mixed effects models for time series, clustered, longitudinal, spatial and spatiotemporal data. We have also been working on multivariate extension of these models. Our models can be applied to discrete, survival, continuous and semi-continuous data as well as clustered categorical and ordinal data of random cluster sizes. In addition, we have also studied zero-inflated time series, clustered, longitudinal and survival data.

The workshop is organized by an AARMS Collaborative Research Group (CRG), Statistical Learning for Dependent Data under the Administration of Ying Zhang. Contact Ying Zhang, Acadia University ying.zhang@acadiau.ca for any questions

Sponsors

We are grateful for the support of Science Atlantic, the Atlantic Association for Research in the Mathematical Sciences, Canadian Statistical Sciences Institute Health Science Collaborating Centres, and UNB Fredericton.

About our AARMS Collaborative Research Group (CRG)

Real-world problems are constantly challenging us to invent new statistical methods. Our group members are actively involved in methodological research and have numerous multidisciplinary collaborations tied to real-world problems. All the members have active interdisciplinary research programs funded externally by individual and/or partnership grants, are active in HQP training (undergraduate/graduate/PDF), and are publishing high quality papers. For the past decade, the group members have been solving problems involving complex dependent data, big data, and statistical learning, in important application areas such as environmental sciences, manufacturing quality control, public health, medical science research and bioinformatics. There have been active and successful collaborations among the members, modelling longitudinal, spatial, and clustered dependent data (Ma, Hasan, Yan, and Sneddon); testing trend in time series data (Cabilio and Zhang); data mining (Gu, Kenney, and Chipman); and bioinformatics/biostatistics (Gu, Kenney, Chipman, Zhang, and Peng).

The primary objective of the CRG is to develop a collaborative research program to share resources and coordinate activities in order to

a) address emerging statistical learning methods and computing issues motivated by multidisciplinary collaborations related to big data with complex dependency structure.
b) develop and distribute novel statistical software, popularizing our research tools in both statistics and application areas.
c) encourage collaboration, the exchange of ideas and joint supervision of HQP at all levels (undergraduate, graduate, PDF) among universities and in partnership with external organizations.