TITLE: Statistical Models and Neural Nets: Supervised classification and prediction via soft trees
Speaker: Antonio Ciampi, McGill University

Abstract: It is well known that any statistical model for supervised or unsupervised classification can be realized as a neural network. This talk is devoted to supervised classification and therefore the essential framework is the family of feed-forward nets. An approach to unsupervised classification based on the link between Kohonen maps and statistical models for density estimations is also being developed (see Lechevallier's talk).

Ciampi & Lechevallier have studied 2- and 3-hidden layer feed-forward neural nets that are equivalent to trees, characterized by neurons with 'hard' thresholds. Softening the thresholds has lead to more general models. Also, neural nets which realize additive models have been studied, as well as networks of networks which represent a 'mixed' classifier (predictor) consisting of a tree component and an additive component. Various 'dependent' variables have been studied, including the case of censored survival times.

A new development has recently been proposed: the soft tree. A soft tree can be represented as a particular type of hierarchy of experts. This representation can be shown to be equivalent to that of Ciampi & Lechevallier. However, it leads to an appealing interpretation, to other possible generalizations and to a new approach to training. Soft trees for classification and prediction of a continuous variable will be presented. Comparisons between conventional trees (trees with hard-thresholds) and soft trees will be discussed and it will be shown that the soft trees achieve better predictions than the hard tree.

Title: Gibbs Sampling for Pharmacophore Identification
Speaker: S. Stanley Young, National Institute of Statistical Sciences

Abstract: The binding of a small molecule to a protein is inherently a three dimensional matching problem. As crystal structures are not available for most drug targets, there is a need to be able to infer the key binding features of small molecules and their disposition in space, the pharmacophore from bioassay data. We use fingerprints of 3D features and a modification of Gibbs sampling to determine the common pharmacophore parts for a set of compounds. We use a clique detection method to map the features back onto the binding conformations. The method works for known pharmacophores. We show the basic algorithm is fast, 15 minutes for 15 molecules, and it can easily deal with a hundred compounds and tens of thousands of conformations. We demonstrate the successful use of PharmID on a multiple binding mode problem. Being able to sort out multiple pharmacophores from the same data set is potentially useful in cell-based assays where different molecules could be hitting different biological targets. Knowing the 3D pharmacophore for a biological target is a key for more efficient compound design and 3D database searching.

Title: Modeling HIV complex clinical and proteomic data structures
Speaker: George Hatzakis, McGill University

Abstract: Within the context of Clinical/Bio-Informatics, we use numerical techniques to model and optimize

clinical management of patients treated for Human Immunodeficiency Virus-1 (HIV-1). HIV infection
is for the most part chronic and asymptomatic. Optimal therapy should suppress the HIV-virus, prev
ent the emergence of antiviral drug-resistance and control long-term side effects. In this present
ation we will address the former 2 aspects. To achieve virus-suppression we have to longitudinally
follow and understand how an HIV-patient progresses. However, clinical and laboratory follow-up i
nformation is non-stationary and characterized by transients and trends. We use Artificial Intelli
gence based models to follow the progression of a subset of patients from the Electronic Anti-Retr
oviral Therapy (EARTH) International-cohort and address several what-if scenarios related to morbi
dity and mortality. Also, to identify those patients that could mostly benefit from the new class
of drugs based on the CCR5 and CXCR4 chemokine inhibitors, we analyze the proteomic sequences of t
he V3 loop on over 1000 patients coming from the HOMER BC-cohort. Clustering techniques are herein
presented.

Title: Data Mining at CSE
Speaker: Shirley Mills, Carleton University

Abstract: Data mining in the world of communications security offers a wealth of challenges to the research community in Canada. This talk will focus, at an unclassified level, on some of the issues faced and some of the collaborations possible.

Title: Clustering Categorical Data

Speaker: Steven X. Wang, York University

Abstract: In the talk, we will review current algorithms for clustering categorical data including the advantages of different algorithms and their computational complexities.
We also introduce a new clustering algorithm for categorical data which does not require any convergence criterion. The proposed algorithm produces a unique partition since it is insensitive to the input order. It also does not require the knowledge of number of parameters. We will also discuss the challenges in clustering categorical data and future works.

Title: The use of high throughput genomics to predict disease risk and treatment response
Speaker: Brent Zanke, University of Toronto, OCRN

Abstract:
The coincidence of functionally relevant polymorphisms in genes that are part of a single
pathophysiologic pathway may cause significant risk for an individual and collectively account

for a large proportion of population attributable risk. For instance, the activities of phase I
enzymes such as the cytochrome P450s, phase II enzymes, such as glutathione-s-transferases, DNA
repair enzymes, cell cycle control enzymes and apoptosis effectors. Polymorphisms in each of
these enzymes that individually would confer only minor increased disease susceptibility could
collectively cause significant individual risk. Many case-control studies evaluating isolated
polymorphisms have failed to identify significant disease association, potential victims of
underpowered study designs.

The Human HapMap. In anticipation of genome-wide disease association studies an international
human variation-mapping project was launched in October 2002 to catalogue blocks of LD and
haplotype diversity (http://genome.gov/10005336). As much as 85% of the human genome may be
organised into haplotype blocks that are 10,000 bases or larger. The exact pattern of SNP
variants within a given haplotype block differs among individuals, though for most less than 5
distinct haplotype clades exist. This limited haplotype diversity makes complete genotyping of
individuals of Northern European or Asian descent possible with measurement of as few as
50,000-100,000 haplotyping SNPs (htSNPs) and measurement of approximately 250,000-500,000
htSNPs in individuals of African descent.

We are studying haplotype diversity in patients with colon cancer and controls to detect
associations with the presence of the disease and with treatment response to those with cancer
receiving chemotherapy. Tests such as these will reduce health care costs and reduce the social
cost of cancer. With an investment from Genome Canada our group will measure over 1 billion
SNPs in 2400 individuals over the next 6 months. The statistical analysis of this data set will
present new issues in multiple testing correction and multivariate analysis.

Title: Initial analysis of ARCTIC data: dimensionality reduction and knowledge pooling
Speaker: Celia Greenwood and Rafal Kustra, University of Toronto

Abstract: We will present an initial analysis of the first batch of ARCTIC data in an international effort to derive a prognostic test of colon cancer using dense maps (hundreds of thousands) of genetic markers and detailed clinical and lifestyle data. In the second half of the talk we will discuss our attempts in building a predictive, multivariate model using

boosting and propose a dimension reduction techniques motivated by statistical and evolutionary genetics. Our untested proposal is intended to spark discussion on dealing with huge dimensionality of genomic data in the presence of highly refined existing knowledge about genetics, knowledge which could potentialy be used to construct more succesful predictive models.

Title: Issues in Multiple Testing and Effect Estimation for Candidate Gene and Genome-wide Studies of Complex Traits

Speaker: Shelley Bull, University of Toronto

Abstract:
While it is well-recognized that the examination of multiple hypotheses corresponding to multiple SNPs within a candidate gene and/or to multiple genes or genetic markers across the genome can lead to inflated false positive rates and failure to replicate findings in an independent sample, the impact of multiple testing and strict type I error control on effect estimation has received less attention. To put these issues in context, I^Òll begin by introducing some background concerning gene discovery and gene characterization, and the related data structures. Approaches to multiple testing adjustments in genetic linkage and association analyses, whether family-based or case-control designs, can usefully depend on the correlation structure among neighbouring genetic loci. However, multiple testing and stringent type I error control typically induce bias in the associated effect parameter estimates. We propose a bootstrap algorithm and resampling-based estimators that yield bias-reduced estimates from the original sample in general settings.

Title: Statistical problems in gene clustering from high-throughput data
Speaker: Jenny Bryan, UBC

Abstract:
We use the term "high-throughput data" to encompass a large variety of current assays in which a response is measured across a range of condition or subjects for a large number genes (often for practically an entire genome). This certainly includes transcriptional profiling via microarrays, as well as highly parallel phenotypic studies in, for example, the yeast deletion set. A common use of such data is to cluster genes, with the hope that apparent gene clusters will have substantial overlap with biological gene groups, such as pathways, protein complexes, or regulons. I will cast this problem in the form of a traditional statistical inference problem and draw some practical conclusions about preferred algorithms and such matters. If time allows, we can use this framework to generate group discussion on the general "disparate data" unification problem in gene clustering (should we create meta-datasets and then cluster? should we cluster datasets

separately and then merge? should we use biclustering-type
techniques?).

Title: Data Analysis, Software, and Pedagogy in Big Science Biology
Speaker: Bob Nadon, McGill University

Abstract:
Big science biology is generating massive data sets that provide motivation
for algorithm development and potential for long-term funding. This
continuing collaboration between biology and the computational sciences will
be most productive if knowledge and tools are made available in formats
readily accessible to applied scientists. I describe such a project in
microarray analysis that integrates software, pedagogy, and data analysis
research.

Title: Longitudinal and Clustered Data Analysis (LCDA)
Speaker: Peter Song, University of Waterloo

Abstract:
I would like to offer a personal overview of
the methodological development in the LCDA.
Arguably, the methodology of the LCDA has
provided powerful tools to practitioners
for their subject-matter innovative research in
past two decades or so. In this talk, I plan to
cover both Liang and Zeger's marginal models and
the generalized linear mixed models. I will use
a few real world data sets as running examples
to enchance my discussions. This presentation is
intended to give a summary of methodology developments
in the LCDA, including successful methods as well
as challenges to be confronted.

Title: Forests, Fires and Stochastic Modelling
Speaker: John Braun, UNiversity of Western Ontario

Abstract:
Statisticians have an important role to play in the study of various
aspects of forestry. This talk will begin with a description of how an
NPCDS workshop can facilitate interactions between statisticians and
researchers into wildfire behaviour as well as forest ecology and
hydrology.

The latter half of the talk will describe a work in progress connected
with a problem of forest fire prediction given observed lightning strokes.
The prediction problem is not solved; however, the talk will describe how
interactions with forest fire researchers have spurred development
of statistical methodology.

Title: Marine Ecology: An Overview
Speaker: Chris Field, Joanna Flemming, Dalhousie University

Abstract:
Will give a general overview of the Marine Ecology Workshop to be held

at Dalhousie in August, 2005 as part of the NPCDS programme. Will give a fairly brief description of a problem involving the dynamics of plankton levels in the tropical Pacific. We conclude by providing a more detailed analysis of a problem involving tracking data of leatherback turtles, a long distance migrant.

Title: Issues in Model Calibration
Speaker: Jason L. Loeppky, UBC

Abstract: Computer models are widely used in engineering and science to simulate physical phenomena. Before using a computer model, for example to optimize systems, a natural first step is often to assess whether it reliably represents the real world. Data from the computer model are compared with data from field measurements. Similarly, field data may be used to calibrate or tune unknown constants in the computer model.

Calibration is particularly problematic in the presence of systematic discrepancies between the computer model and field observations. In this talk we will introduce a likelihood based approach to the estimation of the calibration parameters and will further show how one could use this to test the reliability of the computer model. The approach and the test will be illustrated through a series of examples, and compared to the results of a Bayesian implementation.

Title: BUILDING SURROGATE MODELS BASED ON DETAILED AND APPROXIMATE SIMULATIONS
Speaker: Zhiguang Qian, Georgia Institute of Technology

Abstract:

Title: Designing efficient simulations for exploring features of response surfaces
Speaker: Pritam Ranjan, SFU

Abstract: In many engineering applications, one is interested in identifying the Values of the inputs in computer experiments that lead to a response above a pre-specified threshold. In this talk we introduce statistical methodology that identifies the desired contour in the input space. The proposed approach has three main components. Firstly, a stochastic model is used to approximate the global response surface. The model is used as a surrogate for the underlying computer model and provides an estimate of the contour together with a measure of uncertainty, given the current set of computer trials. Then, a strategy for choosing subsequent computer experiments that improve the estimation of the contour is outlined. Finally, we discuss how the contour is extracted and represented. The methodology is illustrated with an example from a multi-class queueing system.

Title:
Speaker: Derek Bingham, SFU

Abstract:

Latin hypercube sampling is a popular method for evaluating the expectation of functions in computer experiments. However, when the expectation of interest is taken with respect to a non-uniform distribution, the usual transformation to the probability space can cause relatively smooth functions to become extremely variable in areas of low probability. Consequently, the equal probability cells inherent in hypercube methods often tend to sample an insufficient proportion of the total points in these areas. In this paper we introduce Latin hyper-rectangle sampling to address this problem. Latin hyper-rectangle sampling is a generalization of Latin hypercube sampling that allows for non-equal cell probabilities. A number of examples are given illustrating the improvement of the proposed methodology over Latin hypercube sampling with respect to the variance of the resulting estimators. Extensions to orthogonal-array based Latin hypercube sampling, stratified Latin hypercube sampling and scrambled nets are also described.

Title: Inert Column Variable Selection
Speaker: Crystal Linkletter, Department of Statistics and Actuarial Science Simon Fraser University

Abstract:
In many situations, simulation of complex phenomena requires a large number of inputs and is computationally expensive. Identifying which inputs most impact the system can be a critical step in the scientific endeavour so that these factors can be further investigated. In computer experiments, it is common to use a Gaussian spatial process to model the output of the simulator. In this article, we introduce a new, simple method for identifying active factors in computer screening experiments. The approach is Bayesian and only requires the generation of a new inert variable in the analysis. The posterior distribution of the inert factor is used as a reference distribution with which we assess the importance of the experiment factors. The methodology is demonstrated on simulated examples as well as an application in material science.

Title: Generalized Quasilikelihood Approach for Survey Based Incomplete Longitudinal Binary Data
Speaker: Brajendra Sutradhar, Memorial university of Newfoundland

Abstract:
When the response variable in a longitudinal model is subject to missing completely at random (MCAR) or missing at random (MAR), the existing `independence' or `working correlations' based generalized estimating equations (GEE) approaches yield consistent estimators for the effects of the covariates. These GEE based estimators may, however, be inefficient. There also exists a true correlation structure based GEE approach to deal with exponential family based longitudinal responses subject to MCAR or MAR. The existing correlation models used in such incomplete data analysis are, however, quite restricted. In this paper, we exploit a robust correlation model based generalized quasilikelihood (GQL) approach, where the correlation model can accommodate AR(1), MA(1) and exchangeable correlation structures for longitudinal binary responses. Furthermore, for the cases when individuals are

selected based on a complex survey sampling scheme rather
than simple random sampling, it becomes necessary to incorporate the
survey weights in the estimation approach. For this purpose, we
develop a survey design based GQL (DBGQL) estimating equation approach
as a generalisation of the GQL approach. The DBGQL estimation approach
is illustrated by analysing a real life binary longitudinal data set
subject to MCAR or MAR.

Title: Latent Regression with Social Science Data: A Comparison of Various Methods Using Simulation and C
omplex Survey Data Examples

Speaker: Roland Thomas, Carleton and Irene Lu, York University

Abstract:
The paper will focus on methods for estimating regression coefficients for the linear latent regre
ssion models frequently encountered in social science research. In the social sciences, latent var
iables are typically measured using batteries of questionnaire items from which latent variable sc
ores can be predicted in numerous ways. These scores comprise fallible estimates of the underlyin
g latent variables, and it is well known that naïve methods of analysis based on these scores are
likely to result in biased estimates. These biases are quantified not only for simple scoring meth
ods, but also for methods based on Item Response Theory (IRT). The conclusion is that the use of s
cores, no matter how sophisticated, yields unacceptably large bias and should be avoided. An alter
native approach via discrete structural equation modeling (SEM) is also evaluated. This approach,
which implicitly includes the IRT model structure, is shown to provide lower levels of regression
parameter bias, though its bias cannot be ignored for the smaller sample sizes. Finally, the paper
describes a recent adaptation (Bollen, 1996) of the instrumental variables approach to social sci
ence data, and shows that this simple approach provides low levels of parameter bias comparable to
the more computationally involved discrete SEM method. The performance of the various approaches
is compared using simulation, and is also illustrated on complex survey data from Statistics Canad
as Youth in Transition Survey.

Title: Survey Bootstrap Methods and Analysis of Survey Data
Speaker: Milorad Kovacevic, Statistics Canada

Abstract:
A variety of approaches for estimating design-based variances of
estimated model parameters are suggested in literature. The particular

approach of bootstrapping through the rescaling of the survey weights - which we call the survey bootstrap, is gaining popularity due to its portability. Namely, once bootstrap samples have been taken and bootstrap weights calculated, the user estimates the quantities of interest in exactly the same way with the full sample and with each of the bootstrap samples, and then combines these estimates to obtain variance estimates. There are situations, however, in which this direct variance estimator may be unstable. Recently, methods have been proposed for making inferences using an estimating function bootstrap in a model-based setting, which seem to provide more stable results. These methods have been adapted to produce different design-based estimating function survey bootstraps. In this presentation we will cover some of these new developments. Results of a simulation study motivated by a real-life analysis will be presented.

Title: A Nonparametric Test for Association of Interval Censored Event Times in the National Population Health Survey.
Speaker: Norberto Pantoja Galicia

Abstract:
Outcomes from the questionnaire of the National Population Health Survey (NPHS), a longitudinal survey conducted by Statistics Canada, offer the necessary information to explore the relationship between smoking cessation and pregnancy. A formal nonparametric test for association is presented. This test requires estimation of the joint density for interval censored event times, which takes into account complexities of the sample design.