Workshop on Complex Data Structures

Introductory Remarks

Projects, and pilot projects, within the National Program on Complex Data Structures (NPCDS) met April 9 – 14 at the Banff International Research Station. Leaders in Computer Experiments, Data Mining, Genomics and Survey Methods each organized a day of activity in their respective fields. An additional day was devoted to three pilot projects that have inaugural workshops later this year in the areas of Biomedicine, Forestry and Marine Ecology. Research presentations were incredibly varied and included topics that concerned pharmacophore identification, complex HIV proteomic data structures, communications security, studies of complex traits, social behaviour, forest fires, high throughput genomics, tracking of leatherback turtles, turbulence, and so on. Underlying such a diverse set of topics was a genuine common interest in complex data, regardless of its origin. This, in effect, bonded participants in their vision of what NPCDS can bring to the statistical sciences community in Canada. As such the event was instrumental in generating considerable enthusiasm for the Program's model. Concretely, the establishment of interdisciplinary projects with quantitative leadership was viewed as a vehicle that gives our community a greater voice in the research agenda's of other disciplines. These projects have the potential to create a culture in our discipline where training takes place in intensely interdisciplinary environments ensuring young researchers become effective collaborators in the long run. This was evident by the number of excellent presentations given by graduate students including Norberto Pantoja Galicia, Jason Loeppky, Pritam Ranjan and so on.

The Science

Data Mining

Certainly this workshop started in extremely strong fashion setting the tone for the remainder of the week. The first day's topic was *data mining,* a field with many connotations though organizers were able to encapsulate much of the research in this area through a focus on the rare target problem. Data mining is a new and fast-changing discipline, which aims at the discovery of unusual and unexpected patterns in large volumes of data. It came to life in response to the challenges and opportunities provided by the increasing number of large, complex, high-dimensional databases covering important areas of human activity, coming from the industrial, economical, social and biomedical sectors.

Stan Young of the National Institute of Statistical Sciences addressed the use of Gibbs sampling for pharmacophore identification, a problem where large libraries of molecules are searched for comparatively similar reactive properties. Here 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. They 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. They demonstrated 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 was a key for more efficient compound design and 3D database searching.

Stan's talk was followed up by a mesmerizing demonstration of the predatory behaviour of the Human Immunodeficiency Virus (HIV). This was in the context of George Hatzakis's (McGill University) lecture concerning modeling HIV complex clinical and proteomic data structures. Within the context of Clinical/Bio-Informatics, it is common to 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, prevent the emergence of antiviral drug-resistance and control long-term side effects. In George's presentation he addressed the former 2 aspects. To achieve virus-suppression one has to longitudinally follow and understand how an HIV-patient progresses. However, clinical and laboratory follow-up information is non-stationary and characterized by transients and trends. George used Artificial Intelligence based models to follow the progression of a subset of patients from the Electronic Anti-Retroviral Therapy (EARTH) International-cohort and addressed several what-if scenarios related to morbidity 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, he analyzed the proteomic sequences of the V3 loop on over 1000 patients coming from the HOMER BC-cohort. Clustering techniques were presented.

Further presentations concerned the development of particular data mining tools as demonstrated most effectively by Antonio Ciampi and Steven Wang who spoke on soft classification trees and clustering categorical data respectively. Perhaps one of the most compelling presentations was given by Shirley Mills and Ted Normington, both of Carleton University, who are involved in various research projects in consultation with the Communications Security Establishment: Data mining in action leading to secure national borders (we hope).

Genomics and Statistical Genetics

Not to be outdone by the data miners the second day of the workshop was devoted to the genetic revolution that is taking both the medical world and our imaginations by storm. Advances in many areas of Genomics have become the most exciting story in the biological, life, and health sciences in recent years, and have captured the imagination of

the public at large. One of the most interesting technological breakthroughs in genomics has been the miniaturization of classical experimentation techniques in molecular biology. This has led to the ability to conduct massively parallel experiments on the scale of the whole genome. The most widely known examples of such technology are various kinds of microarrays or DNA chips, which can now measure the expression activity of most of the predicted genes in humans. There exist similar high-throughput technologies to detect Single Nucleotide Polymorphisms (SNP chips), protein abundance (proteome chips), RNA activity, protein-protein interaction systems, and others.

For the first time in history, biologists are facing huge volumes of noisy data. The challenge of analyzing this data has been described as the biggest bottleneck in modern biology. Huge dimensionality and small sample size creates a challenge throughout an experiment, from the design, visualization and exploratory phases, to the analysis itself.

The genetics/genomics theme at the meeting was led by Dr. Brent Zanke, VP of the Ontario Cancer Research Network (OCRN) who spoke on the use of high throughput genomics to predict disease risk and treatment response. Here 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 at 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.

In anticipation of genome-wide disease association studies an international human variation-mapping (HapMap) 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.

Brent and colleagues 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.

Rafal Kustra and Celia Greenwood are leading efforts to confront the new statistical issues in this context. They present an initial analysis of the first batch of 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. They discussed attempts in building a predictive, multivariate model using boosting and proposed a dimension reduction techniques motivated by statistical and evolutionary genetics. Their 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 potentially be used to construct more successful predictive models.

Rafal and Celia were followed up by Shelley Bull who addressed issues in multiple testing and effect estimation for candidate gene and genome-wide studies of complex traits. 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 Shelley first considered 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. They proposed a bootstrap algorithm and resampling-based estimators that yield bias-reduced estimates from the original sample in general settings.

Jenny Bryan then spoke on statistical problems in gene clustering from high-throughput data. The term "high-throughput data" encompasses 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. Jenny cast this problem in the form of a traditional statistical inference problem and drew some practical conclusions about preferred algorithms and such matters. She used this framework effectively to generate group discussions 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?).

The final genomics speaker, Bob Nadon, addressed data analysis, software, and pedagogy in big science biology. 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. Bob described such a project in microarray analysis that integrates software, pedagogy, and data analysis research.

Pilot projects

After two exciting days devoted to Data Mining and Genomics attendees were treated to a short day composed a simply a morning session in which nascent NPCDS projects were featured. This was extremely rewarding for both the speakers and audience, the latter be given a sense of future directions of NPCDS while the former received an abundance of helpful advice and suggestions to assist their endeavours.

The first of these speakers, John Braun, discussed forest ecology under the title Forests, Fires and Stochastic Modelling. He asserted that statisticians have an important role to play in the study of various aspects of forestry. The talk began with a description of how an upcoming NPCDS workshop would facilitate interactions between statisticians and researchers into wildfire behaviour as well as forest ecology and hydrology. This was followed up by a description of 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.

John's lecture was followed up by a joint effort from Chris Field and Joanna Flemming who both gave an overview of statistical methods in marine ecology. This included a general overview of the Marine Ecology Workshop to be held at Dalhousie in August, 2005 as part of the NPCDS programme. They also gave brief descriptions of example problems involving the dynamics of plankton levels in the tropical Pacific and a more detailed analysis of a problem involving tracking data of leatherback turtles, a long distance migrant.

The third presentation was given by Peter Song who discussed an array of methods he has developed for use in biomedical research. This included a personal overview of the methodological development in the of longitudinal and clustered data analysis (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 his talk, he covered both Liang and Zeger's marginal models and the generalized linear mixed models. Peter used a few real world data sets as running examples to enhance discussions.

Computer Experiments for Complex Systems

The design and analysis of experiments continue to make important and far-reaching contributions to scientific investigation. Historically, experimenters have relied on physical experiments to help understand processes. The rapid growth in computing power has made the computational simulation of complex systems feasible and has helped avoid physical experimentation that might otherwise be too time consuming, costly, or even

impossible to observe. The advent of such widespread computer experiments raises a host of challenging statistical issues, which this project will explore.

The fourth day of the workshop was devoted to the topic of computer experiments and was marked by a large number of student presentations which were all quite excellent. Jason L. Loeppky, a postdoctoral student at UBC, addressed issues in model calibration. 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 Jason's talk he introduced a likelihood based approach to the estimation of the calibration parameters and further showed how one could use this to test the reliability of the computer model. The approach and the test were illustrated through a series of examples, and compared to the results of a Bayesian implementation.

Zhiguang Qian, a graduate student at Georgia Institute of Technology, discussed building surrogate models based on detailed and approximate simulations, while Pritam Ranjan, a graduate student from SFU, discussed designing efficient simulations for exploring features of response surfaces. Pritnam's talk was particularly interesting as 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 his talk he introduce statistical methodology that identifies the desired contour in the input space. The proposed approach had 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, he discussed how the contour is extracted and represented. The methodology is illustrated with an example from a multi-class queueing system.

Yet another graduate student, Crystal Linkletter of SFU, presented where she discussed inert column variable selection. 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. Crystal introduced 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 was demonstrated on simulated examples as well as an application in material science.

The final speaker of the day was Derek Bingham, who leads the NPCDS project in this area and supervised a number of the students who presented. In Derek's talk latin hypercube sampling was presented as 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. Derek introduced 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 were 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 were also described.

Complex survey data

Survey data now being collected by many government, health and social science organizations have increasingly complex structures precipitating an urgent demand for new statistical methodology to further research in substantive areas. In cross-sectional studies, which are taken at one point in time, it is typical to use very complex sampling designs, involving stratification and clustering as the components of random sampling. There can also be complexities in the resulting data file due to the patterns of nonresponse. In longitudinal studies, which follow individuals or groups of individuals over time, there is additional complexity stemming from possible complex correlation structures induced by repeated measurements on the same sampling unit, by irregularly spaced data and differing numbers of repeated observations on individuals. This datatype, with all it various complexities, is increasingly common in substantive areas due to its power to infer causality, to separate individual and population trends and to detect changes in time.

The final day of the workshop was devoted to the efforts of the survey methods project within NPCDS, although due to many of the team members being drawn to the meeting of the International Statistical Institute in Australia, the session was limited to four speakers. Nevertheless this is a very active project, involving many graduate students one whom presented in this session.

The first speaker was Milorad Kovacevic of Statistics Canada who discussed survey bootstrap methods and analysis of survey data. Here a variety of approaches for estimating design-based variances of estimated model parameters were reviewed. The particular approach of bootstrapping through the rescaling of the survey weights - which he call the survey bootstrap, was presented as 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 Milorad's presentation he covered some of these new developments. Results of a simulation study motivated by a real-life analysis were presented.

The next speaker, Brajendra Sutradhar considered generalized quasilikelihood approachs for survey based incomplete longitudinal binary data. 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 Brajendra's presentation he exploited 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, Brajendra developed a survey design based GQL (DBGQL) estimating equation approach as a generalisation of the GQL approach. The DBGQL estimation approach was illustrated by analysing a real life binary longitudinal data set subject to MCAR or MAR.

The next talk, a joint effort by Roland Thomas and Irene Lu, was of particular interest as the research resulted from a collaboration borne out of NPCDS/SAMSI joint efforts in the context of the SAMSI thematic program: Latent Variable Models in the Social Sciences. The title of their talk was "Latent Regression with Social Science Data: A Comparison of Various Methods Using Simulation and Complex Survey Data Examples" The presentation focused on methods for estimating regression coefficients for the linear latent regression models frequently encountered in social science research. In the social sciences, latent variables are typically measured using batteries of questionnaire items from which latent variable scores can be predicted in numerous ways. These scores comprise fallible estimates of the underlying latent variables, and it is well known that naive methods of analysis based on these scores are likely to result in biased estimates. These biases are quantified not only for simple scoring methods, but also for methods based on Item Response Theory (IRT). The conclusion is that the use of scores, no matter how sophisticated, yields unacceptably large bias and should be avoided. An alternative 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 speakers described a recent adaptation (Bollen, 1996) of the instrumental variables approach to social science 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 was compared using simulation, and is also illustrated on complex survey data from Statistics Canada's Youth in Transition Survey.

The day, and indeed the workshop, ended in fine form with a presentation from yet another graduate student, this time Norberto Pantoja Galicia from the University of Waterloo who was one of the participants of the internship program that is jointly funded by NPCDS and Statistics Canada. Norberto discussed a nonparametric test for association of interval censored event times in the National Population Health Survey (NPHS). Here outcomes from the questionnaire of the 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 was presented. This test requires estimation of the joint density for interval censored event times, which takes into account complexities of the sample design.

Concluding Remarks

For NPCDS this event at BIRS was timely as the Program is currently entering the second half of its four-year funding cycle and it offered an opportunity for participants to assess what has been accomplished thus far. The general view was "a lot!": with potentially seven national projects established in a two year span the Program has engaged the broader community in a robust way. Credit ***must*** be attributed to the many individual researchers who are investing time and energy into this endeavour. During the week at Banff, general meetings were held where progress, and the future of the program, was discussed openly. For example, issues concerning capacity led to consideration of an RFP for training initiatives, which is now being actively pursued. In addition, plans for the renewal of the program have been set in motion.