DAVID BRILLINGER
University of California, Berkeley

Assessing connections in networks with point process input and output with a biological example

Networks with vector-valued point process input and output are considered, although the techniques apply to ordinary time series as well. Both time-side and frequency-side methods are presented, contrasted and combined. Data from a biological experiment on the muscle spindle are investigated. The experiment involves two input and two output neurons. The analysis combines the results of a time domain approach with those of a frequency domain approach to obtain "much new informatiom about the behaviour of the muscle." The latter work is joint with K. A. Lindsay and J. R. Rosenberg of the University of Glasgow, one of whom I am quoting.

MICHEL CHAVANCE (SPEAKER) AND SYLVIE ESCOLANO
INSERM, France

Misspecifying random effects in generalized linear mixed effects models

Correlated outcomes are common in biostatistics, originating from longitudinal, multi center or geographical studies. Marginal or mixed generalized linear models are generally used to take these correlations into account. In the former, inference commonly relies on the "sandwich" estimate of the estimator covariance matrix which is robust to misspecifications of the outcomes covariance structure. By contrast, inference in mixed models most commonly relies on the "naive" estimate of the estimator covariance matrix which is sensitive to such misspecifications. As a matter of fact usual softwares for mixed models do not propose any sandwich estimate of the covariance matrix.

We illustrate the problems that can be encountered with real data about the performance of transplantation centers. A log linear random intercept model was fitted to the annual number of procured organs in each center Y_{ij} (for center i, i = 1, 158, and year j, j = 1, 4). Nine characteristics of the center or of the hospital were included in the vector of explanatory variables with fixed effects and the random intercept bi was interpreted as a measure of the performance status of center i.

$$\log[E(Y_{ij})] = X_{ij}\beta + b_i.$$

However, for two explanatory variables, the lognumber of admissions in intensive care unit ($\beta = 0.22$) and the existence of a neurochirurgy ward in the hospital ($\beta = 0.62$), the sandwich estimate of the standard error was almost twice the naive estimate (respectively .10 versus .05 and .25 versus .13).

We also use simulations to quantify the risk of inference errors and to investigate how the discrepancy between the naive and the sandwich estimates of variance varies with an index of adequation between the observed and predicted covariance matrix of outcomes defined as

$$1 - \frac{\sum_i tr[(\Sigma_i^{-1} S_i - I)^2]}{\sum_i tr[(\Sigma_i^{-1} S_i)^2]}$$

SALEHIN CHOWDHURY
Carleton University

Association analysis of disease status with a candidate gene using generalized linear mixed models

We are interested in an association analysis of disease status with a candidate gene using Generalized Linear Mixed Model (GLMM). We have considered each of the families in the population as an independent cluster with intra cluster correlation. A binary logistic random intercept model has been assumed to simulate the primary data. Here the random intercept is the random effect of the individual families. The maximum likelihood method has been used to estimate the fixed and random effects of the model. We investigate the impact of model violations on the estimate of the coefficients of this model as well. For comparison, we used two strategies from a hierarchical nonparametric bootstrap approach. One strategy (Strategy 1) samples family units, preserving the structure and correlation within each family. The second strategy (Strategy 2) also samples family units but then randomly samples offspring with replacement in each family. Specifically, we evaluate the coverage probability of 95 percent confidence intervals, mean length of the 95 percent confidence interval and the asymptotic relative bias that results from correct and incorrect assumptions regarding the random effects.

RICHARD COOK
University of Waterloo

A copula random effect model for multi-type recurrent events under event-dependent censoring

In many chronic disease processes subjects are at risk of two or more types of events, which may differ in their location and severity. We describe a bivariate mixed Poisson model in which a copula function is used to induce an association between two random variables which have gamma margins. Unconditionally, we obtain a bivariate negative binomial process in which each event marginally follows a negative binomial process. Estimation for parametric and semiparametric models are described. This model is also used to investigate the consequences of an event-dependent censoring scheme arising in

many clinical trials in which subjects are withdrawn from a study when they experience a specified number of the more severe events. The asymptotic biases from naive marginal estimates are discussed, as well as the power implications of such censoring strategies.

KARELYN DAVIS[1] (SPEAKER), CHUL G. PARK[2], AND SANJOY K. SINHA[2]
[1]Carleton University and Statistics Canada, [2]Carleton University

Inequality constraints in generalized linear and mixed models with missing data

Generalized linear, and mixed models are commonly used for analyzing non-normally distributed response variables with or without the presence of clustering among the observations. While many authors proposed different algorithms for fitting such models with no restrictions on the model parameters, only a few authors have considered estimation and hypothesis testing in the presence of parameter constraints. This presentation will discuss maximum likelihood inference and algorithms for such models under linear inequality constraints and will demonstrate their performances through a simulation study. Further extensions to estimation with missing data using the EM algorithm will also be discussed.

CHARMAINE B. DEAN
Simon Fraser University

Spatio-temporal and mixture models for multi-state processes

The development of methods for spatio-temporal analyses has seen tremendous growth over the last two decades. There has also been considerable impact on disease monitoring and surveillance and on exploratory analyses to investigate etiology. This talk discusses spatial-temporal models and methods for analysis with specific emphasis on spatio-temporal analysis of rates and multi-state processes and estimation of ranks. The spatial analysis of rates as considered here is an exploratory analysis which focuses on visually describing the spatial distribution of rates over a region. Multi-state models can be useful in longitudinal studies where at any point in time, an individual may be said to occupy one of a discrete set of states and interest centers on determining what influences transitions between states. For example, states may refer to the number of recurrences of an event, or the stages of a disease. Statistical methodology for the analysis of this type of longitudinal data is presented with the important features of examining how the rates of transitions over states differ spatially over a region. Additionally, in the special case of survival analysis, mixture methods for multi-phase analysis will be discussed. These assume that there exist latent sub-populations which experience different risks of failure. A variety of epidemiological studies will illustrate the methods.

VINZENZ ERHARDT (SPEAKER) AND CLAUDIA CZADO
Munich University of Technology

Generalized estimating equations for generalized Poisson count data with regression effects on the mean and dispersion level applied to patent outsourcing rates

This work focuses on models for longitudinal count data. We discuss Generalized Estimating Equations (GEE) for Generalized Poisson (GP) data. The GP distribution introduced by (Consul and Jain 1970) has a more flexible variance function than Poisson and hence allows to model dispersion by an additional parameter. In addition to GP models considered by several authors (such as Czado et al. (2007)), we now allow for regression on both the mean and overdispersion parameters. Generalized Estimating Equations introduced by Liang and Zeger (1986) estimate model parameters by finding a solution of estimating equations. Since these equations are based on sums of weighted residuals, only parameters influencing the mean can be estimated. For parameters influencing the dispersion, however, additional estimating equations are necessary. A concept by Prentice and Zhao (1991) will be utilized for the GP model: it works with sums of weighted residuals between empirical and predicted covariances for counts at different times. These estimating equations will be solved using a Fisher-Scoring approach.

Our models will be applied to data dealing with outsourcing of patent filing processes. Asymptotic normality of the GEE estimates in this non-exponential setting is proven. Standard errors are estimated using the asymptotic normality of the estimates. For the given data, our GP GEE regression model will prove to be superior over Poisson GEE and even over GP GEE with overall dispersion demonstrating the usefulness of our proposed extensions.

Keywords: Panel count data; Generalized Poisson regression; Overdispersion; Generalized estimating equation; Patent outsourcing

References: Consul, P. C. and G. C. Jain (1970). On the generalization of Poisson distribution. Ann. Math. Statist. 41 (4), 1387.

Czado, C., V. Erhardt, A. Min, and S. Wagner (2007). Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. To appear in Statistical Modelling,

http://wwwm4.ma.tum.de/Papers/index.de.html.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. Biometrika 73, 1322.

Prentice, R. L. and L. P. Zhao (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics 47 (3), 825839.

HYANG MI KIM[1] (SPEAKER), DAVID RICHARDSON[2], DANA LOOMIS[3], MARTIE VAN TONGEREN[4], AND IGOR BURSTYN[5]
[1]University of Calgary, [2]University of North Carolina, [3]University of Nevada, [4]Institute of Occupational Medicine,UK, [5]University of Alberta

Bias in the estimation of exposure effects with individual-or group-based exposure assessment

In epidemiological studies, exposure scores are often assigned via group-based methods in which each subject is assigned a score that is posited to represent the mean exposure of all subjects in a group. In contrast, an individual-based exposure assessment implies that we have an exposure estimate for each subject. We develop models of bias in estimates of exposure-disease associations for analyses that use group- and individual-based exposure assessments in conjunction with logistic regression models when exposures are measured with error. The bias (attenuation) in risk estimates with individual-based assessment is smaller when the between- group and the between-subject variances are large. With group-based assessment, the estimates are less attenuated when the sample size is moderately large, the between- group variability is large and the between- worker variability is small. Grouping results in loss of precision, which is less when between group variance is large. We illustrate this in an application to analyses of the association between exposure to carbon black and respiratory symptoms.

SUBHASH R. LELE
University of Alberta

Data cloning: a simple approach for computing maximum likelihood estimates for mixed models

Maximum likelihood estimation for Linear Mixed Models (LMM) and Generalized Linear Mixed Models (GLMM), an important class of statistical models with substantial applications in epidemiology, medical statistics and many other fields, poses significant computational difficulties. In this paper, we suggest a simple method that exploits advances in Bayesian computation, in particular the Markov Chain Monte Carlo method, to obtain maximum likelihood estimators of the parameters in these models. This method also leads to a simple estimator of the asymptotic variance of the maximum likelihood estimators. Surprisingly, computation of the maximum likelihood estimator and the standard errors requires neither the maximization of a function nor the calculation of its derivatives. Furthermore, the proof indicates that the method necessarily finds a global maximum and not a local maximum. This approach is based on a simple modification of the well-known result that as the sample size increases, the posterior distribution converges to a Normal distribution with mean equal to the maximum likelihood estimator and variance equal to the inverse of the Fisher information matrix. We call our method

data cloning because replicate copies of the same data are used in the implementation of the method. We illustrate the use of the method by analyzing the Logistic-Normal model for over-dispersed binary data, the Poisson-Normal model for repeated and spatial counts data.

J.C. LOREDO-OSTI
Memorial University

Pedigree complexity and genetic inference via likelihood

There is a large class of important problems in genetics for which all the data consist of one or very few large and complex pedigrees. Each pedigree can be seen as one single realization of the genetic process. Although the pedigree may have many members, in general, from a genetic perspective, they are of little value as singletons. Thus, we are faced with the dilemma of carrying out the inference in situations where we cannot collect a large number of families but at most a handful of large pedigrees with marker/phenotype data recorded only for the individuals in most recent generations. Under these conditions the statistical modeling of genetic factors in the pedigree becomes crucial and the likelihood is the natural way to go. However, the computation of the exact likelihood of genetic linkage on large complex pedigrees is a known problem that has daunted the full use of pedigree analysis. The alternatives are to use Monte Carlo methods to estimate the likelihood or to use of rough models associated to methods (like variance components) amenable for computation. Some researchers resort to taking only one of the possible trees of descent into account to link known carriers of an allele to a common ancestor and do 'exact' computations on the selected subset. So, the question would be how the pedigree should be trimmed to retain the salient features of the genetic inheritance process. The solution to this sort of problems can be addressed through computational methods developed in the context of graph theory. There is a connection between finding the most likely paths of descent of a genetic variant and the reduction in pedigree complexity. Under a recessive-lethal mode of inheritance, maximum reduction in complexity is achieved by using a tree of minimal size, that is, a Steiner tree - a well known problem in graph theory - and this solution can be seen as a constrained approximation to the solution to the paths of descent problem. These and related problems in a genetic linkage framework will be discussed in this presentation.

This research is done in collaboration with Prof. K Morgan (McGill University) and CIHR and MITACS support.

FIELDS

ABSTRACTS 1.2

CHARLES E. MCCULLOCH
University of California, San Francisco

Prediction of random effects and effects of misspecifying their distribution

Statistical models that include random effects are commonly used to analyze longitudinal and dependent data, often with the assumption that random effects follow a Gaussian distribution. What are the consequences of an incorrect specification of this distribution? Using a mix of theoretical calculations and simulation I investigate the impact with a focus on prediction of the realized values of the random effects. The results are illustrated using data from the Heart and Estrogen/Progestin Replacement Study using models to predict whether patients are likely to develop high blood pressure.

J. N. K. RAO
Carleton University

Bootstrap methods for analyzing complex sample survey data with dependent structures

Data obtained from complex large-scale sample surveys, such as the National Population Health Survey (NPHS) of Canada, typically involve stratified multi-stage cluster sampling, leading to dependencies among sample elements, and unequal probabilities of selection, leading to unequal design weights. Moreover, design weights are often calibrated to known population totals of auxiliary variables as well as subjected to non-response adjustments, leading to final weights different from design weights. As a result, application of traditional statistical methods to survey data without accounting for dependent design structures and weight adjustments could lead to incorrect inferences even for large samples. We propose to use weighted estimating equations (WEE) to estimate model parameters of interest. For variance estimation and confidence intervals, we employ re-sampling methods, in particular bootstrap methods developed for survey data. Bootstrap offers an attractive option to the analyst for taking account of the survey design because it is easy to implement and more flexible than other re-sampling methods currently used. Data file for bootstrap implementation consists of the full-sample final weights and associated bootstrap final weights for a large number of bootstrap replicated as well as the observed data on the sample elements. We show how such data files can be routinely used to analyze complex survey data, cross-sectional as well as longitudinal. We also provide a one-step estimating function (EF) bootstrap method that avoids a difficulty with the bootstrap. We illustrate the bootstrap methods on the NPHS data. Statistics Canada currently uses bootstrap methods for analyzing data from several large-scale surveys, including longitudinal surveys.

GEOFF ROWE
Statistics Canada

Child birth, labour market transitions, and residential mobility: use of correlated event data to introduce heterogeneity in the LifePaths microsimulation model

Statistics Canada's LifePaths microsimulation model is a model of complete life courses for Canadian birth cohorts. The model is implemented in continuous time with dynamics governed by conditional waiting time distributions or hazard functions estimated from exceptionally rich data resources.

The LifePaths microsimulation model usually represents life course events as conditionally independent. Often, the conditional independence assumption is made given the absence of established alternative methods or analytical techniques. This presentation is based on recent developmental work documented in the working paper "Background to the LifePaths Fertility Module". That paper describes novel use of census data to estimate new fertility equations for the LifePaths model. The new equations make use of empirical relationships among child birth, labour market transitions, and residential mobility to capture more heterogeneity in the timing of birth events than would be possible if the correlations among these events were ignored.

DAVID SANKOFF
University of Ottawa

The generalized adjacency criterion in comparative genomics

We study a parametrized definition of a cluster of terms in a permutation, representing the clustering of genes on a chromosome in both of two genomes. The parameter allows us to control the trade-off between increasing gene content versus conserving gene order within a cluster. This is based on the notion of generalized adjacency (GA), which is the property shared by any two genes no farther apart, in the linear order of a chromosome, than a fixed threshold parameter θ. A cluster in two or more genomes is then just a maximal set of genes, where in each genome these markers form a connected GA chain. We study the statistical propertis expected number of clusters of a given size, limiting distributions, distribution of the largest cluster size of GA clusters under the null hypothesis that the n genes are ordered randomly on the genomes. We discover that the trend from small to large clusters as a function of the parameter θ exhibits a cut-off phenomenon at or near $\sqrt{n}$ as n increases. Using data on five yeast genomes, we also study the extent to which GA clusters are preserved from ancestor to descendant in a phylogenetic tree. We do this by dynamic programming optimization of the presence of individual GA at the ancestral nodes of the phylogeny. Finally, we search for the right value of θ to use with genomes of

size n, exploring the hypothesis that this should maximize, under appropriate constraints, the GA for random genomes. Again we find a square root law for θ.

BRAJENDRA C. SUTRADHAR
Memorial University

GQL inferences in stationary versus non-stationary GLLMs

When covariates of an individual in a longitudinal set up are not time dependent, Sutradhar (2003, Statistical Science) has introduced a common auto correlation structure based GQL (generalized quasi-likelihood) inference technique to obtain consistent and efficient estimates for the regression parameters involved in a class of stationary correlation models. In a non-stationary longitudinal set up, it is however not possible to construct a common auto-correlation structure that may be shared by the correlation models under that class. It will be demonstrated in this talk that one may still use the GQL inference under a given model but the estimates would be confirmed after selecting the model by applying a simple forecasting principle based diagnostic operation. Some simulation results and a data analysis will be provided to illustrate the proposed inference technique.

VICKNESWARY TAGORE (SPEAKER) AND BRAJENDRA SUTRADHAR
Memorial University

Conditional inference in linear versus non-linear models for binary time series

The modelling of discrete such as binary time series, unlike the continuous time series, is not easy. This is due to the fact that there is no unique way to model the correlation structure of the repeated binary data. Also some models may provide complicated correlation structure with narrow ranges for the correlations. In this paper, we consider a nonlinear dynamic binary time series model that provides a correlation structure which is easy to interpret and the correlations under this model satisfy the full -1 to 1 range. For the estimation of the parameters of this nonlinear model we use a conditional generalized quasilikelihood (CGQL) approach which provides the same estimates as that of the well-known maximum likelihood (ML) approach. Furthermore, we consider a competitive linear dynamic binary time series model and examine the performance of the CGQL approach through a simulation study in estimating the parameters of this linear model. The model mis-specification effects on estimation as well as forecasting are also examined through simulations.

Keywords: Consistency; Dynamic models; Forecasting; Model mis-specification effects; Generalized quasilikelihood; Goodness of fit.

ARAFAT TAYEB[1],[2] (SPEAKER), A. BUREAU[1],[3], J. CROTEAU1, C. MERETTE[1],[4], A. LABBE[1], 2
[1]Centre de Recherche UL-Robert Giffard, Dept. of [2]Math and Stat, [3]Social and Preventive Medicine, and [4]Psychiatry, Laval University

Latent class model under familial dependence with missing data

Complex diseases collect many health issues simultaneously. This makes specifying susceptibility genes a difficult task and obliges specialists to accumulate many clinical measurements. In order to solve the issue of heterogeneity, one idea is to put subjects with similar patterns in the same disease sub-group, where only a few number of genes are influent and which are thus easier to detect. Latent Class Analysis is an adapted tool for sub-grouping the disease in more homogeneous sub-types.

We have previously developed a LC Model with dependence between latent disease class status of relatives within families and proposed strategies to incorporate the posterior probability of membership to a latent disease class in linkage analysis. We, previously, implemented this model for nuclear families and simple extended pedigrees. Simulations showed that our approach is more powerful to detect disease genes tan the standard heterogeneity approach of Smith and IBD sharing methods.

In this work, we extend our methods to extended pedigree with missing measurements for some subjects. We present an algorithm to perform computations of the LC Model. This algorithm works like the Elston-Stewart Algorithm and processes nuclear families in a upward and a downward steps. Model fitting uses an EM algorithm with missing data. A set of simulations under different pedigree structures is performed. A study of a set of Bp-Sz data of CRULRG will also be presented.

LEILEI ZENG
Simon Fraser University

Methods for clustered/correlated failure time data

We consider a study of time to the onset of clinical damage in joints of patients registered in the Psoriatic Arthritis (PsA) Clinic at the University of Toronto. Correlated/clustered failure time data arise due to the fact that all the joints from the same patients are related, and the complex correlation structure between event times must be taken into account. One complication of this study is that the failure times were interval censored since the joints damage was assessed at periodically scheduled clinic visits. Another issue is that no damage was ever observed for some joints during the study, so the joints may contain a subgroup of insusceptible ones. We propose to use mixed-effect proportional hazard models with multivariate random-effects for the analysis of correlated interval censored data. We also discuss the use of spatial associations for the event times between joints and

describe the ways to incorporate the susceptivity of the joints into the models. Bayesian analysis using Monte Carlo Markov Chain algorithm was conducted for inference.