

Abstracts

Session 1. 9 - 10.30am

Graham Wood: A procedure for the normalization of ratio data

Quantitative comparison of protein samples by mass spectrometry is rapidly increasing our understanding of biological systems, the mechanisms behind human disease and the detection of biomarkers for disease presence and prognosis. Commonly, these comparative experiments use relative quantitation to observe changes in the abundance of individual proteins. This data is usually expressed as a ratio of the abundance of a protein/peptide in a given state over the abundance of the same protein/peptide in another state (for example, abundance in mutated cells compared to abundance in normal cells). In such experiments a ratio greater than one indicates protein induction, or up-regulation and a ratio less than one indicates protein repression, or down-regulation. Statistical tests are then applied to assess whether a large or small protein ratio is other than a chance observation, so providing a robust platform for conclusions and the planning of future investigations. The fundamental basis of relative quantitation relies on equal loading of each sample under comparison. In practice, however, this is difficult to control accurately, requiring the data to be "normalized" for loading prior to comparative analysis. One approach is to use internal standards, proteins/peptides already present in the sample that are considered not to show radical differences in abundance levels between samples. For example, beta-actin and GAPDH are commonly accepted internal standards for Western blotting and RNA profiling experiments respectively. Alternatively, synthetic external standards can be accurately added to both samples. The uses of internal standards for normalization is favourable as this approach avoids complications that could arise from adjusting the composition of the sample through addition of exogenous analytes. Nonetheless, in either case these proteins should be present in approximately equal quantities in both samples, and thus should approximately represent a true comparative ratio of one. Any technical perturbation affecting the abundance of the standards requires us to devise a normalization rule, which can then be applied to all the data. The question is "How do we optimally perform this correction?" Approaches to data normalization exist in the literature. The purpose of this talk is to introduce a novel normalization method with certain optimal properties.

David Bulger: Stopping and Restarting Criteria for Adaptive Search Methods

Two common questions when one uses a stochastic global optimisation algorithm, e.g., simulated annealing, are when to stop a run of the algorithm, and how often to restart when the algorithm gets trapped at some local optimum. I'll describe joint work with recent Departmental visitor Zelda Zabinsky on this problem. We introduced a parameterised stochastic process to model the sequence of objective function values sampled by a stochastic optimisation method. During application of a real optimisation method, model parameters are estimated by MLE, in the hope of recognising, firstly, when the method is trapped and should be restarted, and secondly, when it has sufficiently probably sampled a small target region around the global optimum. Numerical results on test problems support the approach.

Sheenal Srivastava: Rules for cotranslational protein folding

It is generally acknowledged by experimentalists that proteins fold as they are produced in the ribosome. The potential power of such cotranslational folding, however, has not been used in currently available protein folding algorithms. The aim of this research project will be to establish key rules for cotranslational folding. For example,

1. Investigating whether the conformation of a peptide fragment depends upon its closeness to the nitrogen terminus, and
2. Investigating the impact of the speed (which is related to codon usage) with which a fragment is extruded from the ribosome on the final conformation of the protein.

Once established, such rules will be built into future structure prediction algorithms.

Peter Humburg: HMM Analysis of Histone Modifications in Arabidopsis

Tiling arrays are an important tool for genome-wide experiments. Here we consider their application to the study of histone modifications in Arabidopsis. Histones are a group of DNA binding proteins that has long been known to be responsible for the packaging of DNA in living cells. More recently the role of histones in the regulation of gene expression has begun to emerge and several mechanisms involving various histone modifications are now known. In this study a hidden Markov model is used to locate regions of histone H3 lysine 27 tri-methylation (H3K27me3) in Arabidopsis. In particular we focus on the length distribution of these regions. A previous analysis of these data, using a different HMM, indicates that H3K27me3 regions may be quite short. This result is challenged by the new analysis. We will briefly compare the two models and discuss the different predictions made by them.

Tim Peters: Calibrating Adaptive-Margin Support Vector Machines For Large-Dataset Feature Selection

Support Vector Machines (SVMs) have shown promise in recent times as robust, flexible tools for bipartite class separation. Adaptive-margin SVMs (AM-SVMs) (Weston and Herbrich, 2000), allow for the construction of a classifier which weights the summed magnitude of error of trained misclassified observations against the size of the separating margin via a smoothing parameter, γ . Results of a grid-search on γ using a steepest-descent optimisation on a large microarray dataset are presented with a view to deriving a suitable objective function to use in a Monte-Carlo optimisation algorithm.

Reference: J. Weston and R. Herbrich. Adaptive margin support vector machines. In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 281–295, Cambridge, MA, 2000.

Braddon Lance: Modelling the scope for improvement in template-based structure prediction

A long-standing challenge in template-based modelling is generating protein structure predictions better than the best template. I present in detail a generalised linear model of the best possible predictions achievable in template-based modelling, using a suite of protein structure properties as predictor variables.

Session 2. 11am - 1pm

Nan Carter: Other people's data sets

This is a glimpse into the diverse problems that I have worked on since coming to Macquarie University. The studies here are not experimental, but observational; the data are usually already collected, sometimes as responses to a commercially available questionnaire. A common theme is that the studies are about people.

Suzanne Curtis: Public school sampling strategy for Discipline surveys - simple random sampling versus stratified random sampling

School populations are stratified by class (as well as things like years attended etc.). Discipline problems are very likely to be class dependent (for example, some could be teacher related). Any survey should capture at least one respondent from each class. Simple random sampling (SRS) does not guarantee to do this. Stratified random sampling would guarantee a more representative sample across all classes (if proper follow-up procedures for non-respondents are implemented). Simple descriptive summaries can be generated in the same way for a stratified random sample as for a SRS. A multivariate hypergeometric distribution can be used to show that there is a substantial probability that in a SRS at least one class is not included in the sample. This probability is shown to increase with the number of classes. An easily implemented sampling procedure has been devised making use of a very simple XL spreadsheet.

Hilary Green: The Spot Price of Electricity

The spot price of electricity is determined by National Electricity Market Management Company Limited (NEMMCO) at half hourly intervals. Electricity is a non-storable traded commodity and demand is driven by seasonal and environmental factors. The spot price is aligned with peak and off-peak demand patterns. At periods of peak demand, the spot price can be extremely high. High price volatility of electricity in the energy market makes the modelling of spot price and its driving factors imperative. The proposed time series model explores the spot price dependency on the load in the electricity network and temperature. Various illustrative graphs are used to explore these processes.

Nino Kordzakhia: On parameter estimation of generalised long memory processes

Limit theorems for the conditional sum of square estimates of parameters of generalised ARFIMA processes are studied. Numerical analysis of asymptotic efficiency of sample mean versus BLUE is provided.

Andrzej Kozek: On Variance Reduction for Augmented Data in Heteroscedastic Ordinary Least Squares

We show that for a simple linear regression model the estimators obtained by the Ordinary Least Squares method have the following finite sample property: the variance of estimators is being reduced by each augmentation of data, assuming that the variances of the subsequently augmented data are non-increasing and the regressor values are increasing.

Karol Binkowski: Pricing of European options using empirical characteristic functions

Pricing problems of financial derivatives are among the most important questions of quantitative finance. Since 1973 when Nobel prize winning model was introduced by Black, Merton and Scholes the Brownian Motion (BM) process gained huge attention of professionals. It is known, however, that market stock log-returns are not fit well by the very popular BM process. Derivatives pricing models which are based on more general Levy processes tend to perform better. Carr & Madan (1999) and Lewis (2001) (CML) developed a method for vanilla options valuation based on a characteristic function of asset log-returns. Recognizing that the problem is in modeling the distribution of the underlying price process, we use instead a nonparametric approach in the CML formula, and base options valuation on Empirical Characteristic Functions (ECF). We consider four modifications of this model based on the ECF. The first modification

requires only historical log-returns of the underlying price process. The other three modifications of the model need, in addition, real option prices. We compare their performance based on data of DAX index and ODAX options written on the index between 1 June 2006 and 17 May 2007. Resulted pricing errors shows that our model performs better in some cases, than that of CML.

Barry Quinn: An odd simple linear regression problem

Consider the simple linear regression model

$$Y_j = a + bX_j + e_j .$$

We all know how to estimate a and b , and we know the properties of the estimators when the e_j are i.i.d. If the X_j are unknown, we've got problems. But what if we know $\{X_j\}$ is a non-decreasing set of integers? This is a real signal processing problem from Communications.

Adrian Barker: Time Series Analysis of properties of intra-day Financial Market Return Distributions

In this presentation, distributions of the log of returns of the ASX200 index are estimated at intervals from 2 minutes to 12 minutes for each days trading. Whilst the mean of these distributions, equivalent to a constant multiple of the daily return, has a random noise autocorrelation structure, quantiles and quantile ranges of these distributions have significant autocorrelation coefficients at several lags.

An ARMA(1,1) model provides a reasonable fit to the autocorrelation structure of most of these time series, however some residuals exhibit nonlinear properties. GARCH and bilinear time series models are considered as alternatives to account for these nonlinearities.

Session 3. 2 - 4.30pm

Peter Petocz and Tania Prvan: Relationship between Obstructive Sleep Apnea, facial and body structure, and ethnicity

Obstructive Sleep Apnea (OSA) is a disorder of sleep that is characterised by pauses in breathing that interfere with normal sleeping and cause daytime fatigue and more serious symptoms: the extent of the problem is measured by the AHI (apnea-hypopnea index). We investigate a set of data collected to investigate the relationship between OSA and facial/body measurements in two ethnic groups (Caucasian and Asian). We show that good models can be built from the data using techniques of logistic regression and classification trees. However, an important part of the modelling process is model validation, and this results in less optimistic modelling!

Petra Graham: Methodological issues with a meta-analytic data set

In many meta-analyses found in the literature, the methods are applied without a great deal of regard to the potential issues therein. This talk explores meta-analysis of a problematic data set with plenty of issues and shows how some of the problems may be dealt with.

Jun Ma: Simultaneous estimation of baseline hazard and regression coefficients in a proportional hazard model by directly maximizing the penalized likelihood

Since its introduction by Cox, the proportional hazard model becomes one of the most widely used model for analysing time to event data. Regression coefficients of proportional hazard models, however, are usually estimated by maximizing the partial likelihood function. The

popularity of the partial likelihood is mainly due to its ability to avoid estimating the baseline hazard function (treated as nuisance) when the main focus is on estimating the regression coefficients.

However, in many circumstances, the partial likelihood needs further modifications or even is difficult to implement directly. In this paper we propose an approach to estimating simultaneously the regression coefficients and the baseline hazard function by maximizing directly the penalized log-likelihood function.

Gillian Heller: Ordinal regression models for visual analogue scale measurements

Intangible qualities such as pain and quality of life are often measured on visual analogue scales (VAS). The patient makes a mark on a linear scale, and the distance of the mark from the origin is measured and analysed as a continuous response. There are obvious problems of subjectivity, nonlinearity and non-normality with these responses. Discretisation to an ordinal scale is a conservative way of dealing with the issue. Ordinal regression models are discussed, in the context of a clinical trial for the use of laser treatment for patients with neck pain.

Maurizio Manuguerra: Ordinal response models for continuous scales

Ordinal regression analysis is a convenient tool to analyze ordinal response variables in the presence of covariates. We extend this methodology to the case of continuous scales, deriving the likelihood as a function of the marginal distribution. As this distribution is usually unknown, we propose an approximate version of the original likelihood, taking into account the cognitive aspect of how pain perception changes with pain level. This method protects the estimation of intercepts from model misspecification and has an advantage over the standard ordinal regression model in studies with many categories or sparse data. Using VAS data from a study on the efficacy of low-level laser therapy in the treatment of chronic pain, we show that our formulation performs better than the standard one, gives similar results and is able to describe the nonlinear nature of the VAS scale.

Sangdao Wongsai: Cluster analysis of phytoplankton with similar temporal and spatial patterns

Cluster analysis is a multivariate analysis method that is widely used in ecological studies for data reduction. In this study, we aim to assign an assemblage of phytoplankton into clusters with dissimilar temporal and spatial patterns, based on eight-year data collected at the major reservoirs (Lake Yarrunga, Fitzroy Falls reservoir, and Wingecarribee reservoir) in New South Wales, Australia. Principal component analysis was used to estimate orthogonal components for the temporal and spatial variability of phytoplankton community structure. It turned out that the first four components accounted for about 72% of the total variation. The interaction effects between phytoplankton and these four principal components were then estimated using a multiplicative regression model. Subsequently, a (dis)similarity matrix is formed, containing the squared Euclidean distances between the estimated model parameters for every pair of phytoplankton studied. The significance of dissimilarity among clusters is further evaluated using a chi-squared test. The results indicated that the phytoplankton community composition and abundance were influenced by the temporal and spatial variability. Cyanobacteria were the most dominant group of phytoplankton across the reservoirs, and cell concentrations were relatively abundant in the Fitzroy Falls reservoir and Wingecarribee reservoir. Cryptomonads and golden-brown algae were patchy in the Lake Yarrunga, while green algae and diatoms were observed predominantly in the Fitzroy Falls reservoir and Wingecarribee reservoir.

Annette Kifley: Multidimensional and mixed outcomes in quality of life studies

Many clinical studies now have the objective of comparing treatments based on combinations of quality of life and/or time-to-event outcomes. This often involves assessing the overall balance of positive and negative impacts of conditions and treatments on different quality of life dimensions, along with other outcomes and patients own preferences. While single summary measures are often unsatisfying in this context they are frequently required and used. We aim to develop a cohesive and flexible approach to this problem using latent variable modeling.

Kehui Luo: Modelling Dengue Haemorrhagic Fever Counts

Several models, including generalised linear models (GLMs) based on the Poisson and negative binomial distributions and zero-inflated Poisson and negative binomial GLMs, were fitted to the Dengue Haemorrhagic Fever (DHF) counts with a large number of zeros. Up to three lags of the DHF incidence rates were considered in each of those models. All the models also contain additive effects associated with the season of the year, district and age group. To allow for possible spatial correlations between observations on different districts at the same time, and also for similar correlations between different age groups, additional terms allowing for these effects were also considered. Among all models fitted, both the negative binomial generalized linear (time series) model and its zero inflated version provide a good fit to the data.

Ayse Bilgin: A Historical Survey of PhD Students in Statistics

The governments push for shorter PhD completion times is shaping higher education institutions in Australia. The first step to ensure the completion times are shorter than before is to have a clear picture of past performances to compare. When the completion times are known for the past students, then the quality improvement to shorten completion times can start (if required). This paper aims to document PhD completion times for one statistics department in Australia by using a historical survey of PhD completions along with some of the characteristics of students (such as gender, nationality) that might be helpful to identify why the completion times are as they are (short or long compared to average or expectations). The descriptive statistics show that there is no difference between males and females completion times where on average international students (regardless of gender) complete their studies 8.4 months earlier than their local counterparts.