

Sydney Summer Statistics Workshop - 2004

(Jointly Sponsored by the University of NSW, Macquarie University and
the University of Sydney)

Theme : Multivariate and Graphical Methods

Friday, 27 February - 9.00am to 5.00pm

Stephen Roberts Lecture Theatre, The University of Sydney

Bayesian methods for Variable Selection and Covariance Selection in Multivariate Regression Models

Chris Carter, CSIRO

Abstract:

We consider the Bayesian estimation of regression parameters and an inverse covariance matrix from Gaussian data. Methods are given to construct priors for covariance selection models where the marginal distribution of the model size has a simple form. The priors have normalising constants for each possible model size, rather than for each possible model, which gives a tractable number of normalising constants that we estimate using Markov chain Monte Carlo methods and store offline. Our priors do not require the restriction that the corresponding graphical models are decomposable. The effectiveness of variable selection and covariance selection in estimating the multivariate regression model is assessed using several loss functions.

** This is joint work with Ed Cripps and Robert Kohn (The University of New South Wales) and Frederick Wong (The Australian Graduate School of Management)

EDA using Direct Manipulation Graphics

Di Cook, Department of Statistics, Iowa State University

Abstract:

Question: What is the primary goal of exploratory data analysis?

Answer: Insight. The clear (and often sudden) understanding of a complex situation, the power or act of seeing into a situation; Insight implies detecting and uncovering underlying structure in the data.

”The primary goal of EDA is to maximize the analyst’s insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:

1. a good-fitting, parsimonious model
2. a list of outliers
3. a sense of robustness of conclusions
4. estimates for parameters
5. uncertainties for those estimates
6. a ranked list of important factors
7. conclusions as to whether individual factors are statistically significant
8. optimal settings”

<http://www.itl.nist.gov/div898/handbook/>

With the evolution of fast, graphically enabled desktop computers, exploratory data analysis has evolved into a highly interactive, real-time, dynamic and visual process. To perceive multi-dimensional structure, the user displays multiple views of the data and links graphics objects in one plot with the other plots. Thus with basic building blocks such as histograms, scatter-plots, barcharts, mosaic and double decker plots the statistician can pull out a lot of information about high-dimensional relationships in data.

This talk will demonstrate the highly interactive data analysis process as applied to a microarray data set (groan?) collected at ISU. It’s interesting data! There are treatments, repeated measures, and model diagnostics. I began the analysis using graphics. A student began the analysis using statistical modeling. The lists of interesting genes that we produced had almost zero overlap! Why is this so? This talk will discuss how this is possible, what we learn from each approach and what we miss. And along the way we will see how to use multivariate graphics, some new complex linking between plots, and some surprising data findings.

This is joint work with Heike Hofmann, Eun-Kyung Lee, Hao Yang, Basil Nikolau and Eve Wurtele.

Dynamic graphics for microarray data

Harold Henderson, AgResearch Ruakura, NZ

Abstract

I will present some examples of the use of simple dynamic graphics for visualizing data from two colour cDNA microarrays. I find dynamic graphics useful in checking the data before normalisation. This checking easily reveals some unusual features of the data, for example, spots that are decidedly green on all (including dye reversal) slides and subsequently were found to be a contamination. Dynamic graphics are also useful in exploring the data after it has been normalised, in the subsequent analysis across slides, for example, in checking the consistency of ESTs with the same contig.

Model-Based Classification with High-Dimensional Data

Geoff McLachlan, Department of Mathematics and Institute for Molecular Bioscience, University of Queensland

Abstract

Finite mixture models are being increasingly used to model the distributions of a wide variety of random phenomena. In this talk, we consider the use of mixture distributions to provide a model-based approach to classification with high-dimensional data. The need to be able to classify such data arises in many applications ranging from biology to image processing. For the unsupervised classification problem (cluster analysis), we consider the use of mixtures of factor analyzers. This approach enables a normal mixture model to be fitted to data which have high dimension relative to the number of data points to be clustered. The number of free parameters is controlled through the dimension of the latent factor space. By working in this reduced space, it allows an interpolation in model complexities from isotropic to full covariance structures without any restrictions. We also consider the case of supervised classification (discriminant analysis). The methodology is illustrated by applications in the context of cancer diagnosis and treatment. One problem concerns classifying a relatively small number of tumour tissue samples containing the expression data on very many (possibly thousands) of genes from microarray experiments.

Graphics = The Human Eye + Data + Theory

John Maindonald, ANU

Abstract

The human eye/brain combination is the instrument that reads meaning into graphs. Data come both in the form of data for plotting, and experimental data on the processes of human perception. Theory comes both in the form of statistical theory that gives insights on what to plot, and a theory of human perception that indicates that some forms of presentation will be effective, and others ineffective. Graphs may fail both because they reveal too little, and because they lie and thereby reveal too much. In practical data analysis, graphical presentation goes hand in hand with model-based analysis. Both for graphical presentation and for analysis, parsimony and hierarchy help cope with complexity. Parsimony can be overdone, and checks are essential. This talk will explore these ideas in the context of multiple regression, discriminant analysis (here seen primarily as a dimension reduction technique) and principal components analysis.

Empirical Bayes model selection and analysis of microarray data

Matt Wand, UNSW

Abstract

An emerging area of research interest for statisticians has been in the analysis of data from DNA microarrays. DNA microarrays allow the simultaneous measurement of expression for thousands of genes in tissue samples. We consider analysis of data from DNA microarray experiments where scientific questions of interest can be framed in terms of comparison of a collection of linear models for each of the genes. We discuss ways in which we can use hierarchical and empirical Bayes methods for model selection to borrow strength across genes for making inferences and to deal with ques-

tions of multiple comparisons and model uncertainty. Differences between our methodology and existing approaches to the same problem will be discussed: in particular, our methodology doesn't require prespecification of a prior probability for differential expression and we take a model selection rather than hypothesis testing approach so that comparison of non-nested models for each gene can be undertaken. Application of the methods is considered for a series of experiments intended to study the contribution of genetic variation to control of gene transcription. This is joint work with Eva Chan, Chris Cotsapas, William Dunsmuir, Michael Kirk, Peter Little, Rohan Williams and Matt Wand.

** This work is joint with David Nott

Dynamic Balancing Randomisation: a tree-based method of randomisation and its properties

Stephane Heritier, Avinesh Pillai and Val Gebski, NHMRC Clinical Trial Centre, University of Sydney

Abstract

In the design of randomised clinical trials, balancing of treatment allocation across important prognostic factors (strata) improves the efficiency of the final comparisons. Whilst randomisation methods exist which attempt to balance treatments across the strata (permuted blocks, minimisation), these approaches assign equal importance for all the strata. Dynamic Balancing Randomisation (DBR) is a tree-based method proposed by Signorini et al. (1995) allowing different levels of imbalance in different strata ensuring a balance for each level of prognostic risk factors (conditional balance) whilst at the same time preserving randomness. We present a modification to the original approach to maintain a marginal balance over important strata and examine the properties of this modification. Two important measures of performance are used: a loss function, which can be interpreted as the squared norm of the imbalance vector, and a forcing index which conveys the degree of randomness. A comparison of the DBR with minimisation (a common randomisation procedure for clinical trials) is used to illustrate the advantages of the new method. A comparison to the biased coin approach for which the asymptotical behaviour has been studied but is not necessarily well known is also made.

END ABSTRACTS

Shelton Peiris, 93515764

Please visit :

<http://www.maths.usyd.edu.au:8000/u/shelton/2004usydstat.html>

for more details about Sydney University (including the site map).