42nd Gregynog Statistical Conference Programme

(2006)

The talks will take place in Seminar Room 1 (2nd Floor, far end).

Friday	16.00	*Tea*	
21 April	17.30	Prof. Wally Gilks	Leeds University
		Introduction to Bioinformatics	
	19.00	*Dinner*	
Saturday	08.00	*Breakfast*	
22 April	09.30	Dr. Christine Currie	Southampton University
		Balancing bias and variance in the optimisation of simulation models	
	11.00	*Coffee*	
	11.30	Prof. Wally Gilks	Leeds University
		Fusing microarray data	
	13.00	*Lunch*	
		Afternoon free	
	16.00	*Tea*	
	17.00	Prof. Mike Kenward	London School of Hygiene and Tropical Medicine
		Some practical applications, and issues, with multiple imputation	
	18.30	*Dinner*	
	19.45	Prof. Chris Glasbey	Biomathematics & Statistics Scotland, Edinburgh
		Image restoration, segmentation and warping using generalizations of dynamic programming	
Sunday	08.00	*Breakfast*	
23 April	09.30	Prof. Charles Taylor	Leeds University
		Kernel methods in statistical learning	
	11.00	*Coffee*	
	11.30	Prof. Wally Gilks	Leeds University
		A statistical approach to distance-matrix phylogenetics	
	13.00	*Lunch and finish*	

Speakers

Dr. Christine Currie	Southampton
Prof. Wally Gilks	Leeds
Prof. Chris Glasbey	Biomathematics & Statistics Scotland, Edinburgh
Prof. Mike Kenward	London School of Hygiene and Tropical Medicine
Prof. Charles Taylor	Leeds

	Staff	Students	
Aberystwyth			
Alan Jones	John Lane		
Bangor			
Chris Whitaker	Rhiannon Whitaker		
Prof Michael Cain	Takeshi Kikuchi		
Cardiff			
Terry Iles		Emily Hamilton	Jonathan Gillard
Daniel Farewell		Jessica Read	Mark Kelly
Sofia Pedro		Venkat Timmaraju	
Keele			
Prof Peter Jones		Milica Blagojevic	
John Preater		Charis Emmett	
Nottingham Trent			
Prof Neville Davies			
Southampton			
Prof Russell Cheng			
Swansea			
Alan Meyer		Owen Bodger	See Ju Chua
Alan Watkins		Hannah Finselbach	
Warwick			
Prof John Copas		Ian Gallagher	Miguel Belmonte
Mohand Feddag		Theodore Papamarkou	Demetris Lamnisos
John Fenlon		Beatriz Penaloza	Christopher Howitt
Masayuki Henmi		Katherine Boyd	Claudia Lozada-Can
Prof Jane Hutton		Maria Costa	Michalis Kolossiatis
Jen Marsh		Peter Kimani	Patrick Ho
Prof Mark Steel			Miland Joshi
Wenjuan Zhang			

Abstracts

Three sessions on Bioinformatics

Professor Wally Gilks
Leeds University

1. Introduction to Bioinformatics

The field of bioinformatics has grown up around the huge and rapidly expanding genomic and related databases, including the human genome sequence. These open access databases and the many servers providing tools to interact with them present fresh challenges to statisticians. I will present a very brief overview of the field, and indicate where statisticians might try to get involved.

2. Fusing microarray data

Microarrays can be used to measure the activity of tens or hundreds of thousands of genes simultaneously.
The data generated are typically very noisy and microarray experiments may be poorly reproducible.
Ideally, therefore, experiments should be replicated in several ways and in several laboratories before scientific conclusions are drawn. To make the most of such data, we propose a method, based on multivariate regression, for microarray data fusion. We apply the method to data arising from cell cycle experiments in yeast.

3. A statistical approach to distance-matrix phylogenetics

Phylogenetics is the study of the evolutionary relationships between different species, typically represented in the form of a tree. Some quite sophisticated statistical methods have been employed to estimate phylogenetic trees on the basis of DNA sequences from each species. However, for very large phylogenies, such methods are too slow. More computationally tractable methods are based on a distance matrix constructed from the sequence data.
However, such methods have weak statistical underpinnings, in particular with regard to handling uncertainty.
We propose an agglomerative regression-based method for distance-matrix phylogenetics, having a sound statistical foundation, but still retaining the computational efficiency of simpler methods.

Balancing bias and variance in the optimisation of simulation models

Dr. Christine Currie
Southampton University

We consider the problem of identifying the optimal point of an objective in simulation experiments where the objective is measured with error. We describe some simple simulation experimental designs that emphasize the statistical aspects of the process. When the objective can be represented by a Taylor series near the optimum, we show that the best rate of convergence of the mean square error is when the variance and bias components balance each other. More specifically, when the objective can be approximated by a quadratic with a cubic bias, then the fastest decline in the mean square error achievable is n-2/3. Some elementary theory as well as numerical examples will be presented. Comparisons between the method described here and the established algorithm of Stochastic Approximation will also be made.

Image restoration, segmentation and warping using generalisations of dynamic programming

Chris Glasbey
Biomathematics and Statistics Scotland
(Pdf of talk available from http://www.bioss.ac.uk/~chris)

Dynamic programming (DP) is a fast, elegant method for finding the global solution to a class of optimisation problems. For example, it can be used to find maximum a posteriori (MAP) estimators of boundaries, to automatically segment 2-D medical images into anatomical regions (Glasbey and Young, 2002). A variant, dynamic time warping, can also be used to align pairs of tracks in 1-D electrophoresis gels. However, for many image problems, including 3-D segmentation, 2-D warping and image restoration, DP is not possible.

We consider three generalisations of DP for image restoration, segmentation and warping. The first approach is a greedy algorithm first proposed by Leung et al (2004), termed iterated dynamic programming (IDP), where DP is used to recursively solve each of a sequence of lower-dimensional problems in turn, to find a local optimum. A second algorithm replaces DP by a more computationally intensive Forward-Backwards Gibbs Sampler (Scott, 2002), and uses a simulated annealing cooling schedule to guarantee the optimal solution. The final approach is an empirical, stochastic optimiser, which is implemented by adding noise to IDP. Results are compared with existing pixel-by-pixel methods of iterated conditional modes (ICM) and simulated annealing, and illustrated using data from synthetic aperture radar (SAR), 3-D X-ray computed tomography and 1-D and 2-D electrophoresis gels.

Glasbey, C.A. and Young, M.J. (2002). Maximum a posteriori estimation of image boundaries by dynamic programming. Applied Statistics, 51, 209-221.
Leung, C., Appleton, B. and Sun, C. (2004). Fast stereo matching by Iterated Dynamic Programming and quadtree subregioning. British Machine Vision Conference (Eds. A Hoppe, S Barman and T Ellis) 1, 97-106.
Scott, S.L. (2002). Bayesian methods for Hidden Markov Models: recursive computing in the 21st century. Journal of the American Statistical Association, 97, 337-351.