Statistical Theory and Methods for Complex High-Dimensional Data
7 January to 27 June 2008

Report from the Organisers:
D. Banks (Duke), P. Bickel (Berkeley), I. Johnstone (Stanford) and M. Titterington (Glasgow)

Background

Most twentieth-century statistical theory was restricted to problems in which the number of 'unknowns', such as parameters, is much smaller than the number of experimental units. However, the practical environment has changed dramatically, with the need to analyse massive datasets, in areas such as image analysis, genomics, astronomy and climatology, in which the number of 'unknowns' can hugely exceed the number of experimental units. As a result, innovative core statistical theory is required, along with new methodology and novel approaches to graphical display that both cope with these important practical scenarios and exploit state-of-the-art computing capability. Key advances are being made by both mainstream statisticians and the growing machine-learning community.

The general aim of the programme was to promote research in areas such as strategies for dimension-reduction including latent-structure modelling and the exploitation of sparsity, classification methods for large-scale problems, asymptotics for increasing dimension and visualisation methods for complex datasets. Although the importance of theoretical breakthroughs was emphasised, their impact on applications was not to be neglected. The theory and methods covered by the programme are essential for the proper analysis of data in a wide range of contexts in modern-day life.

The organisers benefited from the advice, as Scientific Advisors, of C. Bishop (Microsoft), P. Hall (Melbourne), J. Shawe-Taylor (University College London) and S. van de Geer (Zurich). Welcome supplementary support for the usual sources of funding was provided by the Cambridge Philosophical Society, by the Leverhulme Trust, by the U.S. National Institutes of Health and National Science Foundation for workshop attendance of young researchers and, in particular, by Microsoft Research.

Structure

The programme attracted 88 visiting fellows and 16 programme participants, with about 75% of the total coming from outside the UK. Workshop numbers are listed separately. Most of the day-to-day activity was left to spontaneity but there were two or three more formal seminars per week, and dedicated short series on metabolomics, organised by D. Banks and Dianne Cook, and on 'looking at data', by Dianne Cook. C. Butucea, with B. Clarke, initiated a working group on performance bounds for inference.

Contemporary Frontiers in High-Dimensional Statistical Data Analysis
7-11 January 2008
Organisers : D. Banks, M. Titterington and S. van de Geer

The opening workshop, with 111 participants, laid out the main themes for the programme. Talks by D. Donoho, M. Wainwright, P. Niyogi and E. Candes addressed the information-theoretic limits that arise in problems where the number of observations is smaller than the number of potentially explanatory variables. S. van de Geer, A. Young, A. Lee, B. Yu, Dennis Cook, R. Samworth, B. Nadler, P. Bickel and M. Jordan described ways to approach those limits, using such methods as

the Lasso, high-dimensional bootstrapping, treelets, regularisation, dimensional reduction and kernel-based contrast functions. Visualisation was addressed by Dianne Cook, W. Stuetzle and E. Wegman, and applications in finance and bioinformatics were covered by M. West, E. Mammen and J. Fan. These presenters represented perspectives from statistics, mathematics and computer science, and that triune cross-disciplinary interaction has been the engine for recent progress in this field. Altogether there were 27 talks, all invited, and 20 contributed posters. The poster presenters were given the chance, in the 'Gong Show', of presenting a three-minute trailer for their poster, with the welcome reward of a glass of wine or juice.

High-dimensional Statistics in Biology
31 March - 4 April 2008
Organisers : P. Bickel, E. Birney, W. Huber and R. Durbin

This workshop took advantage of local strength in molecular biology in the University, the European Bioinformatics Institute and the Sanger Institute. There were 135 participants, 23 talks and 20 posters. Of the speakers 16 were self-described biologists and 7 mathematical scientists. This imbalance was deliberate, and led to achievement of the major aim of exposing mathematical scientists to the great variety and complexity of genomic data and to the underlying biological goals. Statistical or computational talks were given by Y. Benjamini on multiple comparisons, P. Bickel on the statistical complexity of the genome, H. Huang on biological database issues, and G. McLachlan and M. West on classification. N. Beerenwinkel, G. McVean, W. Huber and L. Pachter mixed computational and biological issues fairly equally. The other talks were primarily biological but almost all involved statistical concepts such as Hidden Markov Models, networks and graphs. The amount of data, already huge, is clearly on the threshold of another level of exponential growth, due largely to new, high-throughput, relatively cheap, sequencing technologies: R. Durbin described the 1000 Genomes project, which compares the genomes of 1000 humans at extremely fine resolution. Comparative genomics, inferring structure from high conservation between species, and phylogenetics, will have to incorporate new dimension-reduction ideas, as discussed by G. McVean, E. Margulies and E. Birney.

Bayesian Analysis of High-Dimensional Data
14-16 April 2008
Organisers : D. Banks, J. Griffin, F. Rigat and M. Steel

This Satellite Meeting at the University of Warwick provided the programme with a distinctive Bayesian element. There were about 63 participants, who enjoyed 20 invited talks, 3 contributed talks and 17 contributed posters. The workshop highlighted recent methodological and applied advances in the Bayesian analysis of complex data. Keynote and themed talks focussed on selected topics in biostatistics, computational systems biology and mathematical statistics. Methodological contributions were provided by N. Hjort, F. Liang, L. Li, A. Dobra, M. Guindani, R. Kass, Y. Teh and C. Rasmussen, and context-related modelling featured strongly in many talks. Those by V. Schmid, M. Ghosh, J. Morris, D. Wilkinson and V. Purutcuoglu all had a biological flavour and climatological applications were described by B. Sanso, D. Nychka and J. Haslett.

Inference and Estimation in Probabilistic Time-Series Models
18-20 June 2008
Organisers : D. Barber, S. Chiappa and T. Cemgil

The workshop, partially funded by the European Network of Excellence PASCAL2, encouraged cross-fertilisation of ideas among the machine-learning, engineering and statistics communities. For example, parameter estimation in state-space models often uses variational procedures in machine learning, Markov chain Monte Carlo methods in statistics and subspace methods based on Hankel matrices in engineering. M. Opper's application of deterministic approximate inference to continuous Markov processes contrasted very well with O. Papaspiliopoulos's talk on advanced MCMC-based techniques. The work of A. Sykulski and S. Olhede on approximating integrated volatility in stochastic differential equations was also directed related, as was M. Titsias's use of Gaussian Processes to infer latent external functions underlying continuous-time dynamical processes. Similar cross-fertilisation was apparent in Z. Ghahramani's 'infinite Hidden Markov Model', for which Papaspiliopoulos suggested potentially more powerful sampling schemes. S. Godsill raised doubts about the ultimate feasibility of inference for complex models of an environment, given that parameter estimation is compounded by strong temporal posterior correlations; this point was emphasised in R. Turner's talk on variational inference for time-series models. Altogether there were 19 talks, 4 posters and 80 participants. The contributed talks and posters are available from the website as an electronic proceedings.

Future Directions in High-Dimensional Data Analysis
23-27 June 2008
Organisers : D. Barber, I. Johnstone, R. Samworth and M. Titterington

The closing workshop looked both backwards and forwards, but with a strong emphasis on the future, with several talks based on advances made during the programme. There were 20 invited talks, 9 contributed talks, 13 posters and 110 participants altogether. The talks of V. Koltchinskii, N. Meinshausen, R. Tibshirani, A. Tsybakov and M. Yuan concerned algorithms and properties for large linear regression models when sparsity in the coefficients is expected and exploited. This approach continues to generate intense interest, but it is not without its critics, represented for example by the semiparametric-inference approach of J. Robins. High dimensionality also lurks within functional models for regression or principal component analysis (J.-L. Wang, J. Shi), and is even more pronounced when dependence between pairs of variables is assessed, through correlations (P. Hall) or matrices of covariances (E. Levina, G. Pan, Y. Wang). Machine-learning perspectives on sparse regression were provided by M. Seeger and J. Shawe-Taylor, and models for large graphs or networks were discussed by D. Barber, E. Wit, E. Airoldi, M. West and E. Xing. The workshop's subtitle was 'New Methodologies, New Data Types and New Applications'. The last two themes were reflected in the talks of A. Owen on 'transposable data' and J. Rice on challenges raised by large-scale astronomical surveys. Large-scale applications of regression, clustering and simultaneous testing in genetics and systems biology illustrated many talks, in particular that of K.C. Li on 'liquid association'.

The longer workshops were complemented by a variety of shorter events: a 'Financial Data Day', organised by C. Rogers, emphasised the new problems and opportunities raised by large-scale, high-frequency financial data, and attracted an audience of about 70 people, most coming from the finance industry itself; two afternoons were dedicated to talks presented by members of Cambridge groups closely related to the programme, one with the Machine Learning Group of the Department of Engineering and the other with the Statistical Laboratory; one afternoon brought together participants from the Institute's two parallel programmes, the other being that on Combinatorics and Statistical Mechanics, a synergistic exercise that generated one or two substantive collaborations; and finally the Institute held an 'Open for Business Day'. The objective of this event, co-organised with R.

Leese of the Smith Institute, was to alert senior industrial managers to the type of expertise and applicable research associated with the programme. The opportunity to meet programme participants individually was followed by expository talks on data-mining, by D. Hand, and machine learning, by C. Bishop, and then a Discussion Panel session at which issues of interest were developed further. The event attracted over 50 participants.

The most important individual presentation of the programme was the lecture entitled 'More unknowns than equations? Not a problem! Use sparsity!', by the Rothschild Visiting Professor, D. Donoho. In addition the programme provided, in the form of a well-received talk by D. Banks, the Institute's contribution to the Cambridge Science Festival.

Outcome and achievements

The main aims of the programme were to accelerate ideas in a crucial area of core statistical research and to bring together, through both the general programme and workshops, world leaders in both statistics and machine learning.

The following form a small sample of the topics on which progress was made: issues concerning the approximation and estimation of large matrices (T. Cai, A. Hero, G. Claeskens, P. Wolfe, H.-G. Mueller, M. Yuan, D. Donoho, I. Johnstone, H. Zhou); nonparametric Bayesian analysis (P. Bickel, B. Kleijn, A. Gamst); issues of sparsity (many participants, including V. Koltchinskii, E. Levina, B. Nadler, A. Rohde, A. Tsybakov); oracle inequalities and properties (F. Bunea, B. Nan, A. Tsybakov, M. Wegkamp); visualisation (Dianne Cook, A. Hero, S. Marron). The Rothschild Professor, D. Donoho, interacted with many people, in particular with J. Tanner on random polytopes, J. Jin on optimal feature selection and G. Kutyniok on the geometric separation problem. J. Wellner worked on Nemirovski's inequality with L. Duembgen and S. van de Geer, and on empirical process theory with A. van der Vaart. Microsoft Fellow Dennis Cook developed a general framework for sparse dimension reduction and, with J. Kent, looked into ways of unifying the understanding of regularisation methods. S. Olhede developed a new methods for estimating diffusion tensors in medical image analysis, with I. Dryden, and worked on bivariate curve analysis with A. Kovac. With D. Wagner of the Combinatorics and Statistical Mechanics programme, D. Banks worked on agent-based modelling. B. Rajaratnam and A. Young initiated a collaboration on bootstrap ideas.

The dialogue between statistics and machine learning was strongly apparent, in seminars, workshop talks and informal discussions. The latter involved, among others, J. Lafferty, N. Lawrence, M. Seeger, M. Pontil, D. Barber, Y. Teh, J. Shawe-Taylor, N. Cristianini, S. Roweis and I. Murray, on the machine-learning side, and B. Rajaratnam, J. Kent, G. McLachlan, C. Robert, R. Samworth, J. Shi, M. Yuan and M. Titterington from statistics. The hope is that seeds have been sown for productive interactions and further constructive blurring of this interface.

Feedback indicated universal praise for the scientific and social excellence of the Institute's environment and for the efficiency and friendliness of its Staff.

Publications

Work done during the programme will appear over the next few years in many relevant journals. In addition, two particular compilations are envisaged. The organisers of the Time-Series workshop have a provisional arrangement with Cambridge University Press for a book of contributed chapters based on the workshop. Secondly, the Royal Society have accepted a proposal from the programme organisers for a theme issue in the Society's Philosophical Transactions, Series A.