IMPACT

BIG DATA DELIVERS REAL SOLUTIONS

EPSRC
Engineering and Physical Sciences Research Council

UNIVERSITY OF CAMBRIDGE

http://www.statslab.cam.ac.uk/~rjs57

Impact Objectives

- Provide training in innovative, robust and scalable statistical methods to address contemporary 'big data' challenges within academia and many different industry sectors

Big data delivers real solutions

***Professor Richard Samworth**, statistician and Director of the **University of Cambridge's Statistical Laboratory**, discusses the far reaching impacts of statistics on society and his effort to share this knowledge*

What type of research is the Statistical Laboratory at the Centre for Mathematical Sciences involved with?

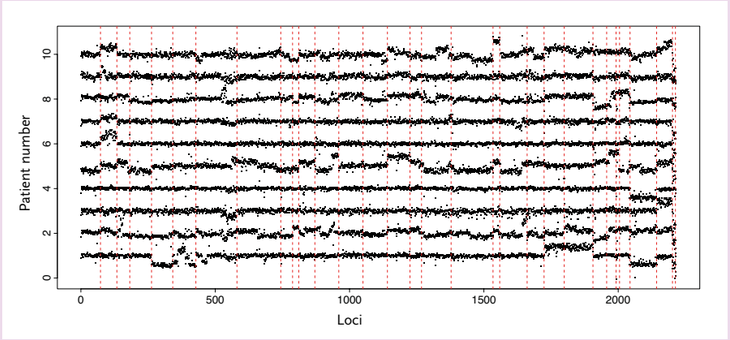
We cover four main areas: Statistics, Probability, Operations Research and Mathematical Finance. This is quite a wide intellectual diversity spread over 17 academic staff, ranging from researchers who would consider themselves to be pure mathematicians, through to people who regularly collaborate with practitioners from science and industry. Nevertheless, we have much in common, and are all motivated in some way to understand random phenomena.

How does understanding more about statistics help society generally? What are the impacts and benefits of your work?

Modern technology now allows the routine collection of data on previously unimaginable scales. This has the potential to yield dramatic benefits across so many different disciplines, but only if we develop appropriate robust and scalable methods to extract useful information from this data deluge. Issues such as how Oyster card data can be used to improve London's transport network and how reliable algorithms can be developed to underpin driverless car technology need to be addressed through statistical analysis. In science, fields like medical imaging, particle physics, climate science, astrostatistics, social science and many others stand on the verge of being able to answer questions that would have seemed well out of reach only a decade ago. To give an illustration, I am working with collaborators at Cancer Research UK to combine data from many different sources and provide accurate, near real-time diagnosis and treatment of cancer.

What tools do you use in your research?

The justification for the methods I develop often involves tools from different areas. Optimisation, approximation theory, differential geometry, analysis and probability have all played significant roles in some of my recent papers.



Log-intensities of fluorescence measurements of DNA fragments at different loci on the genome of ten patients with bladder tumours. The dotted red lines indicate estimated change points

What future directions do you see your work heading?

This is very difficult to predict, and certainly if you'd asked me this question five years ago, my answer then would have turned out to be very inaccurate! One direction that I think is very important and I'd like to explore further is that of so-called online algorithms. These occur typically when data are collected over time, often in vast quantities, and decisions are needed as new data are observed, rather than, say, waiting for the whole study to be completed. As a specific example, my post-doctoral research associate, Tengyao Wang, and I have recently been interested in methods for identifying change points in situations where many data streams are being monitored simultaneously, such as levels of traffic at multiple internet routers. This existing work considers only the offline setting, where the entire data set is available. I would like to see if it is possible to develop analogous methods in settings where one wants to identify a change as soon as possible.

What kind of efforts are you taking to make the knowledge gathered in your research more open?

I do this in several ways. I post papers in the arXiv (a preprint repository) as soon as I submit them, to encourage rapid dissemination of the work. Whenever I develop new statistical methods, I follow the now common practice of making the algorithms publicly available by means of packages written in the free, open-source statistical programming language called R. Where relevant, I upload data to the journal website to accompany publications. I also give many talks on my work all over the world, including recent ones in China, the US, Finland and France.

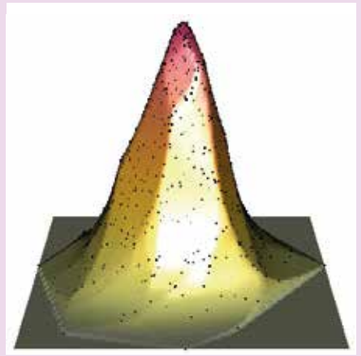
Delivering world class training

*The **University of Cambridge's Statistical Laboratory** is taking on the new challenges of interpreting big data. Whether it's for medical data, traffic solutions or even the arts, researchers are providing real world solutions and valuable training to young statisticians*

In a world awash with data and information, making sense of statistics and numbers is a daunting task. Professor Richard Samworth, Director of the Statistical Laboratory at the University of Cambridge, believes 'that statistics is seen as a difficult and challenging topic partly due to current fears of a reproducibility crisis in scientific publication and a post-truth society drowning in alternative facts'.

Although this seems a grim assessment it is exactly why Samworth is excited about the field of statistics and views this as an opportune time to enter the field: 'It is an incredibly exciting time to be a statistician. Even over the course of my relatively short career to date, it has become clear that the field of statistics is now recognised as far more important and influential than was historically the case'. It is with this optimism and enthusiasm that he heads his research group and the University's Statistical Laboratory.

Statistics are used to interpret data and find the significant patterns or trends. It allows us to sift through the noise and isolate the information that is most important for the question being asked. This makes statistics an essential component of nearly all other scientific disciplines. As fields like genetics and physics improve their data collection techniques they are increasingly faced with larger and larger datasets and new statistical methods for dealing with this so called 'big data' will be required to make sense of it all.



A density estimator called the log-concave maximum likelihood estimator, computed based on a sample of 1000 points

The concepts of big data stretch beyond traditional scientific endeavours as well. City planners and governments need to track traffic data and use this to improve transport networks. Increasingly, businesses and health agencies are collecting personal data which need to be handled in a safe and ethical manner. The quest for driverless cars is dependent on statistics, as robust algorithms are needed to rapidly sort the immense amounts of incoming data that will keep these cars on the road and away from collisions.

DIVERSE RESEARCH, APPLICABLE SOLUTIONS

The Statistical Laboratory, which is a sub-department of the Department of Pure Mathematics and Mathematical Statistics, comprises of 17 University professors, readers and lecturers who supervise 21 research students and 11 postdoctoral students. This Laboratory houses a diverse group of experts, which is reflected in the seminars and courses that they offer as well as the research topics being tackled: Statistics, Probability, Operations Research and Mathematical Finance.

The research of Samworth alone spans theoretical, methodological, computational and applied aspects of statistics and he is regularly published in the top international journals. His work is leading to practical applications for big data problems as well. A method he developed with Rajen Shah called 'Complementary Pairs Stability Selection', which was published in the Journal of the Royal Statistical Society Series B, can be applied to large gene datasets in order to quickly and accurately determine how many out of potentially thousands of genes are associated with a particular disease. He has also published methods that are applicable to email filters; helping them to better decide if a message is genuine or fake.

Through collaboration the members of the Statistical Laboratory are looking to make an impact on an endless number of disciplines. 'Almost all of my research projects are collaborative, and every collaboration is different! Most of my work consists of developing new methods, with appropriate theoretical justification, to tackle questions that may have applications in several different fields,' explains Samworth.

SHARING KNOWLEDGE

Along with publications, the Laboratory has been successfully running the Statistics Clinic for members of the University of Cambridge. Since 2009, a clinic has been held once a fortnight offering free statistical advice to anyone at the university. Knowing how to properly interpret data and apply the correct statistics is a challenging task for most researchers and students, evidenced by the sheer number of clients that have been helped since the clinic began. 'We have provided assistance to well over 1000 researchers since the clinic started,' notes Samworth. As may be expected these researchers often come from the life sciences but not always.

‘The field of statistics is now recognised as far more important and influential’

‘We also find musicologists, archaeologists, linguists, essentially every academic subject studied in the university.’ One of the benefits of programmes like this one is that they occasionally lead to collaborations and although Samworth thinks this is fantastic he insists it is ‘certainly not essential for a successful visit’.

One added benefit he does find particularly important is the training opportunity this provides students because training new students is a key goal of the Laboratory. On the teaching aspect, Samworth says ‘the clinic provides wonderful training for my research group and the other helpers. It takes real skill to understand quickly the key essence of a statistical question in a field with which you may be unfamiliar and to describe suggestions at a level appropriate for the client’.

Through activities such as this and the opportunity to work with supervisors on the varied research projects Samworth is training the next generation of statisticians to tackle the emerging data driven challenges. ‘Much of my research time is spent discussing problems with my PhD students and post-docs. I find these discussions one of the most rewarding aspects of the job. It is a real pleasure to see them experience the same excitement that I feel about doing research, and to see them develop and forge their own independent research careers.’ Both within the fields of statistics and branching out to others the Statistical Laboratory is impacting senior researchers and students alike.

A BRIGHT FUTURE

The field of statistics is a rapidly evolving one and this will require the statisticians of the future to be armed with a range of skills. Increasingly, the term ‘Data Science’ is being used to describe the ever expanding areas covered by statistics. From biology to computer science, signal processing and even ‘astrostatistics’, sound methodological approaches are needed to interpret and validate the latest discoveries.

For Samworth the path to this field was through his final year as an undergraduate, and it was when he was studying Part III mathematics in Cambridge that he thought about statistics as a career option. Hopefully, he is successful as it appears as though the thirst for data will not be quenched anytime soon. Whether it’s in the sciences, engineering, government or business resources like the Statistical Laboratory and the generations of statisticians they are training will be placed in a position to make a significant difference on the future world.

Project Insights

CURRENT PROJECTS

Engineering and Physical Sciences Research Council (EPSRC) Early Career Fellowship ‘Statistical methodology and theory for the Big Data era’ - Dec 2017-Nov 2020

CONTACT

Richard Samworth
Director

T: +44 1223337950
E: r.samworth@statslab.cam.ac.uk
W: http://www.statslab.cam.ac.uk/~rjs57

PROJECT COORDINATOR BIO

Professor Richard Samworth is Professor of Statistical Science and Director of the Statistical Laboratory at the University of Cambridge. He obtained his PhD, also from the University of Cambridge, in 2004. His research interests are in developing methodology and theory for modern statistical challenges. Samworth’s honours and awards include an Institute of Mathematical Statistics Medallion Lecture (scheduled for 2018), the Adams prize (2017, shared with Graham Cormode), a Philip Leverhulme prize (2014) and the Royal Statistical Society Guy Medal in Bronze (2012).

EPSRC
Engineering and Physical Sciences Research Council

UNIVERSITY OF CAMBRIDGE