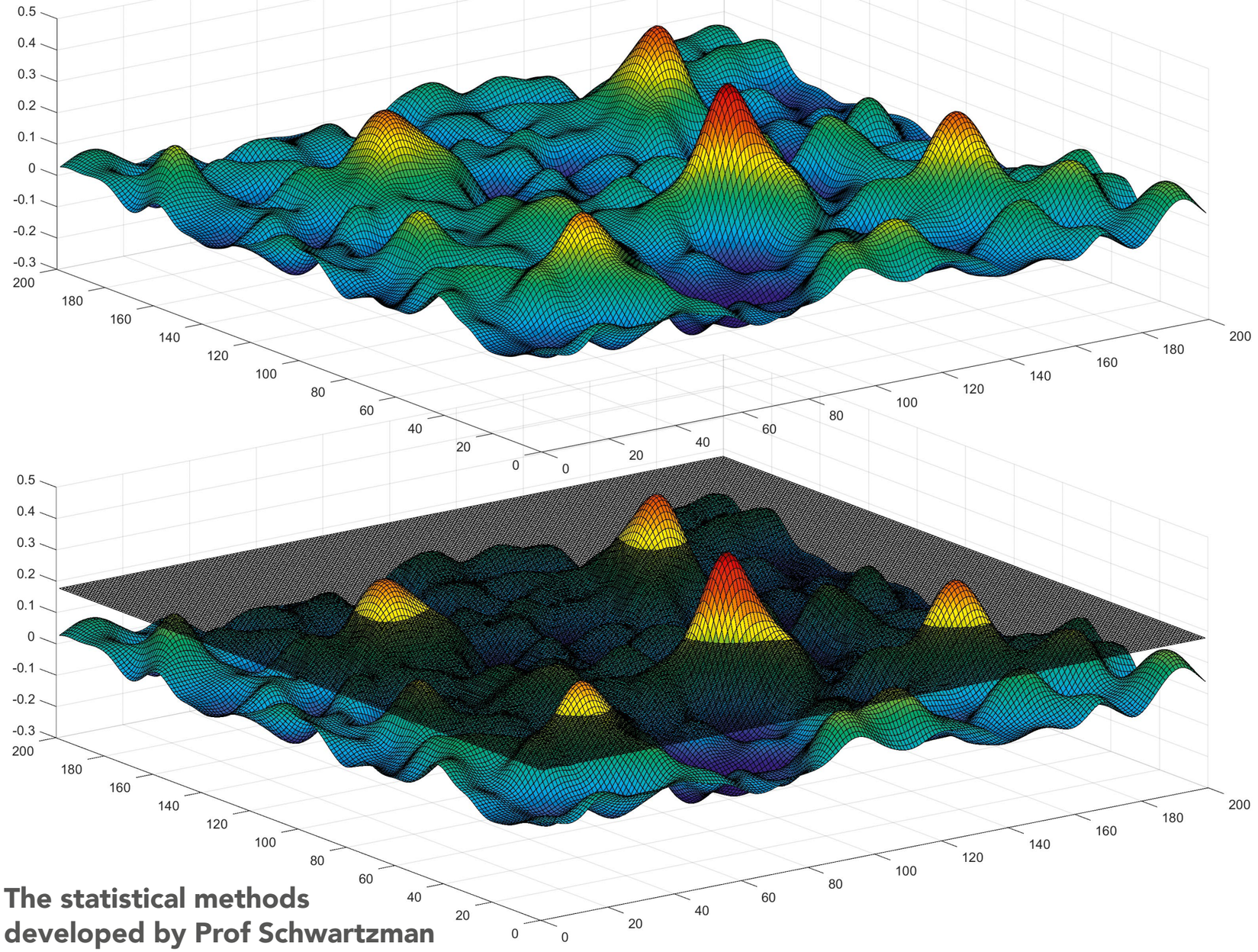
Statistics to simplify and understand complex imaging data



The statistical methods developed by Prof Schwartzman could greatly simplify analysis of imaging data within a variety of fields

When appropriately studied and analysed, large sets of data help researchers to observe changes over time or across conditions and reach important scientific breakthroughs. This is also true for imaging and signal-related data, collected by scientists within a variety of fields. **Professor Armin Schwartzman**, working at the University of California in San Diego, has dedicated his career to the development of simple but effective statistical methods for signal and imaging data analysis, which could have significant biomedical and environmental applications.

Much of the data collected by scientists comes in the form of images. This includes scientific images taken by medical imaging devices such as PET and MRI machines, as well as those collected by Earth-orbiting satellites, microscopes, or images generated by computer models. Making sense of this data can often be a tortuous and complicated task, requiring the use of advanced statistical methods.

Imaging data is extremely diverse; it may be two-dimensional (2D), three-dimensional (3D). It may be even more intricate, for instance depicting the celestial sphere on the convoluted surface of the brain. Even the human genome with its linear array structure can be thought of as a very long one dimensional (1D) image, where every base-pair plays the role of a genomic pixel. Images are often taken over time, adding yet another dimension to this rich and complex data source. Making sense of such complex data is a huge task.

This is where the research of Professor Schwartzman at the University of California, San Diego comes in. His team develops sophisticated mathematical models and algorithms that have the potential to improve data analysis for a vast range of applications, ranging from genomics, medical imaging and environmental monitoring – enabling scientific understanding and discovery.

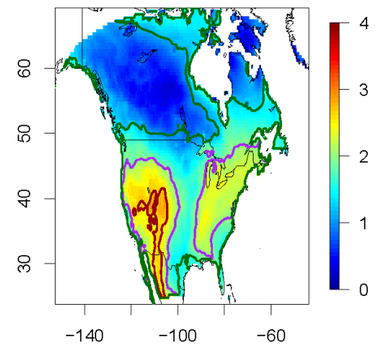
THE LOCALISATION CHALLENGE

One of the most challenging problems associated with complex, high-dimensional data image analysis is the accurate detection of sparse, localised true signals that are embedded within background signal, or 'noise.' This may be a signal involving cerebral regions where there is an activation in response to a stimulus, or a signal following the evaluation of an irradiated cancerous tissue after treatment.

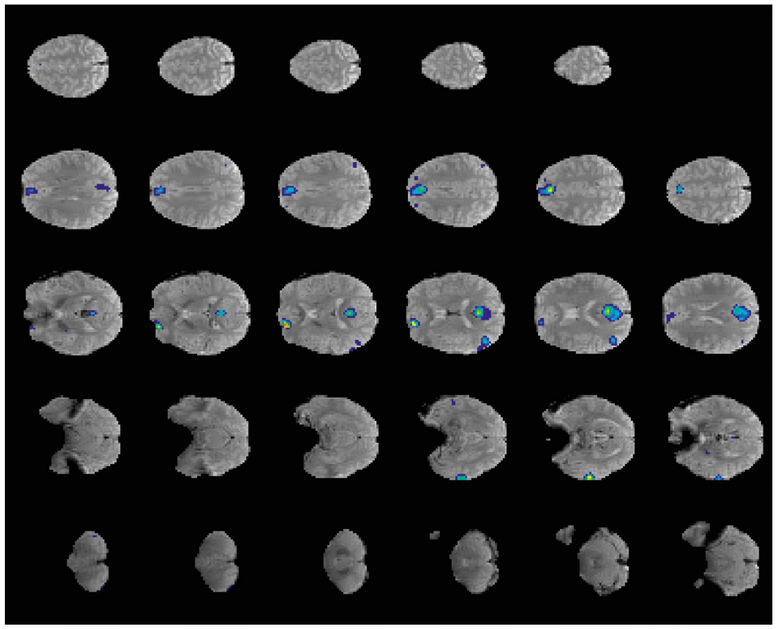
In the eyes of a statistician, the localisation challenge becomes a large-scale multiple testing exercise, where every location is tested for significance. The aim is to determine significant regions of the image that provide optimal quality information against a noisy background, to make correct, specific inferences about the quantity being measured. Given the sheer magnitude of the search across such large amounts of data, strict detection thresholds are needed to prevent too many false positives. There is the danger, however, that this compromises the ability to detect real signals.

THE STRUCTURE OF THE DATA IS KEY

To solve the localisation problem, Professor Schwartzman and his team take advantage of the data's structure. As Schwartzman says: "while statistical methods for multiple testing often assume independence between the tests, many real situations exhibit dependence and an underlying structure." Take the example of the human genome, with its 1D structure within each chromosome. Similarly, environmental



Quantifying spatial uncertainty of climate change projections. In this map of simulated temperature difference between late 20th century and mid-21st century, there is high confidence that the mean summer temperature *will* increase by more than 2 degrees Celsius inside the red contour and *will not* increase by more than 2 degrees Celsius outside the red contour.



Signal detection in functional Magnetic Resonance Imaging (fMRI): a slice-by-slice montage of the brain, showing statistically significant regions (in colour) involved in social cognitive processing.

data have a 2D spatial structure, and brain images a 3D spatial structure.

By taking advantage of this dependence structure, the teams' methods facilitate discovery and reduce the number of false findings. His research focuses on developing statistical methods for signal and image analysis that are proving vital for analysing data generated by biomedical and environmental fields. Specifically, his team are developing and efficiently applying multiple testing methods for random fields and high-dimensional data.

HIGH-DIMENSIONAL DEPENDENT DATA

For 2D and 3D images, the team model the image values as a smooth random field. Describing the theory of smooth Gaussian random fields, Professor Schwartzman explains "originally developed to model the surface of water, it is a beautiful mathematical theory that allows modelling of the spatial correlation structure, estimating error rates and determining detection thresholds." The novelty of his work is that such tests should not be performed at every observed location, but only at important interpretable topological features of the image, such as the local maxima (maximum point within a given range) of the observed data.

His research applies this theory, and adapts it for each data set. For 2D and 3D images the theory has been developed to model the height distribution, so that both localisation and inference can be measured from observed signal peaks. For other data sets, like the 1D genome which is not smooth, the observed distribution of test statistics is characterised under correlation, to inform the inference. For each individual case, the theory is tweaked accordingly to solve the practical problem at hand.

The key to solving them is to use mathematical and statistical tools that take advantage of the structure and geometry of the data

ADAPTING THE SAME TOOLS TO DIFFERENT FIELDS

Although these fields are different in nature, the problems he worked on solving have similar characteristics. Prof Schwartzman says: "The key to solving them is to use mathematical and statistical tools that take advantage of the structure and geometry of the data."

In order to do this, Schwartzman and his colleagues have developed a general theory as a guiding principle and then adapt their methods in accordance with the specific topic they are focusing on. Some data analysis requires more computational approaches, such as the image feature extraction, and machine learning tools used to characterise liver tumours. The team are also developing a software pipeline to estimate the retreat of mountain glaciers worldwide from Landsat images available through Google Earth Engine.

SIMPLE MODELS TO ANALYSE COMPLEX DATA

As complex algorithms require specialised training, the key to the teams' success is making these models simple enough so that they are accessible to all scientists. Indeed, Professor Schwartzman collaborates with researchers from a variety of different fields, enabling them to make inferences about specific types of complex data and identify significant patterns or effects.

Professor Schwartzman teams up with neuroscientists, using MRI scans to identify regions of the brain that respond to a stimulus; with radiologists, using CT to characterise liver tumours or PET to spot changes in response to cancer therapy. Working with climatologists, using climate simulations to identify regions where climate change might require a prompt intervention and with cosmologists, trying to achieve a better understanding of the early universe by analysing snapshots of the cosmos. His statistical models are also used by geneticists, for example to identify where mutations are associated with phenotypical traits.

Professor Schwartzman's research has far-reaching impact, helping the scientific world discover, extract and understand the information hidden within images.

Q&A

When you first started your academic research, could you have imagined that you would eventually develop a unified view of signal detection for random fields?

Not at all. When I arrived as a graduate student at Stanford, I did have the goal of understanding the signal processing concepts I had learned in Electrical Engineering from a more formal mathematical and statistical point of view. But I did not know which form that would take. I owe it to Jonathan Taylor, Robert Adler and Brad Efron. Jonathan got me curious about functional MRI and geometry, and through him I met Robert, from whom I learned random field theory. Brad Efron taught me to search for real applications and keep a wide angle of view. Then things came together as I wrote my first R01 grant application as a junior faculty member at Harvard.

Are there differences in your approach when using such statistical methods in different issues? For instance, when investigating satellite images compared to MRI signals for example?

The statistical models and methods need to reflect not only differences in the data, but more importantly, they need to address the scientific questions that are relevant to the data. For example, in terms of the data itself, satellite images are two-dimensional and exhibit clouds, while MRI images are three-dimensional and have other sources of noise such as motion artefacts. On the other hand, in satellite images we are interested in estimating time trends, while in MRI time trends may be transient and we may be more interested in the brain activity after the time trends have been removed.

How long do you believe it might take to introduce your statistical models within real-life settings, and in what field might their initial applications be?

Statistical methods are slow to be adopted. Statistical research has produced very advanced methodology, and yet a large part of science still uses very basic tools like linear regression and hypothesis tests for their inference. However, the need to properly quantify uncertainty is becoming more evident, particularly in areas where large amounts of complex data are being collected, and these are areas where modern statistical methods are more likely to take hold. In my case, brain image and genomic analysis are probably the areas with the most statistical analysis software tools already and eager to continue the trend. I am working with collaborators to incorporate our new tools into the standard software analysis platforms for widespread use by neuroscientists.

In the years to come, what role do you feel computational methods and machine learning algorithms will play for research purposes and within clinical settings?

While computers and algorithms have been at the forefront of research in highly technical areas such as engineering, they are now making their way into all areas of research, from biology to social sciences, helping us make sense of large and complex data. Medicine will be seeing this soon too as medical records become better organised and easier to analyse. However, while machine learning algorithms excel at performing complex tasks, we also desire scientific explanations of those complex phenomena and an understanding of the fundamental principles guiding them. For this reason, I believe mathematical and statistical models will continue to play a very important role in research alongside computational methods, both complementing each other and helping us better understand the workings of the world.

What are your plans for future investigation?

I want to expand the use of statistical image analysis tools in important areas that still need them. Environmental research, in particular, is becoming increasingly important and I believe it can greatly benefit from the advanced analysis tools we have produced by working in other areas of science.

Detail

RESEARCH OBJECTIVES

Prof Schwartzman's research focuses on developing statistical methods for signal and image analysis, with wide-ranging applications. His work provides a unified view of feature detection in high-dimensional data that applies to a large class of problms ranging from genomics to medical imaging to environmental monitoring.

BIO

Armin Schwartzman is Associate Professor of Biostatistics at the University of California, San Diego. His research interests are focused on signal and image analysis, with applications to biomedicine and the environment. He holds BS and MS degrees in Electrical Engineering from the Technion - Israel Institute of Technology and the California Institute of Technology, and a PhD degree in Statistics from Stanford University.

FUNDING

- National Cancer Institute (NCI)
- National Institutes of Health (NIH)

COLLABORATORS

Robert Adler, David Azriel (Technion - Israel Institute of Technology); Dan Cheng (Texas Tech University); Anders Dale, Fabian Telschow (University of California, San Diego); Joshua French (University of Colorado, Denver); Yulia Gavrilov (TechnoSTAT and Tel Aviv University); David Groisser (University of Florida); Andrew Jaffe (Johns Hopkins School of Public Health); Sungkyu Jung (University of Pittsburgh); Meng Li (Rice University); Xihong Lin, Cliff Meyer (Harvard School of Public Health); Domenico Marinucci (University of Rome, Tor Vergata); Nezamoddin Nezamoddini-Kachouie (Florida Institute of Technology); Thomas Nichols (University of Oxford);Lei Qi n (Dana-Farber Cancer Institute); Brian Reich (North Carolina State University); Philip Reiss (University of Haifa); Steve Sain (The Climate Corporation); Max Sommerfeld (University of Göttingen); Wenguang Sun (University of Southern California); Jeffrey Yap (University of Utah)