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...
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 mathematical 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 characterize 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 specialized 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 with a variety of different fields, enabling them to make inferences about specific types of images or to 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. 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 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. Machine learning 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.