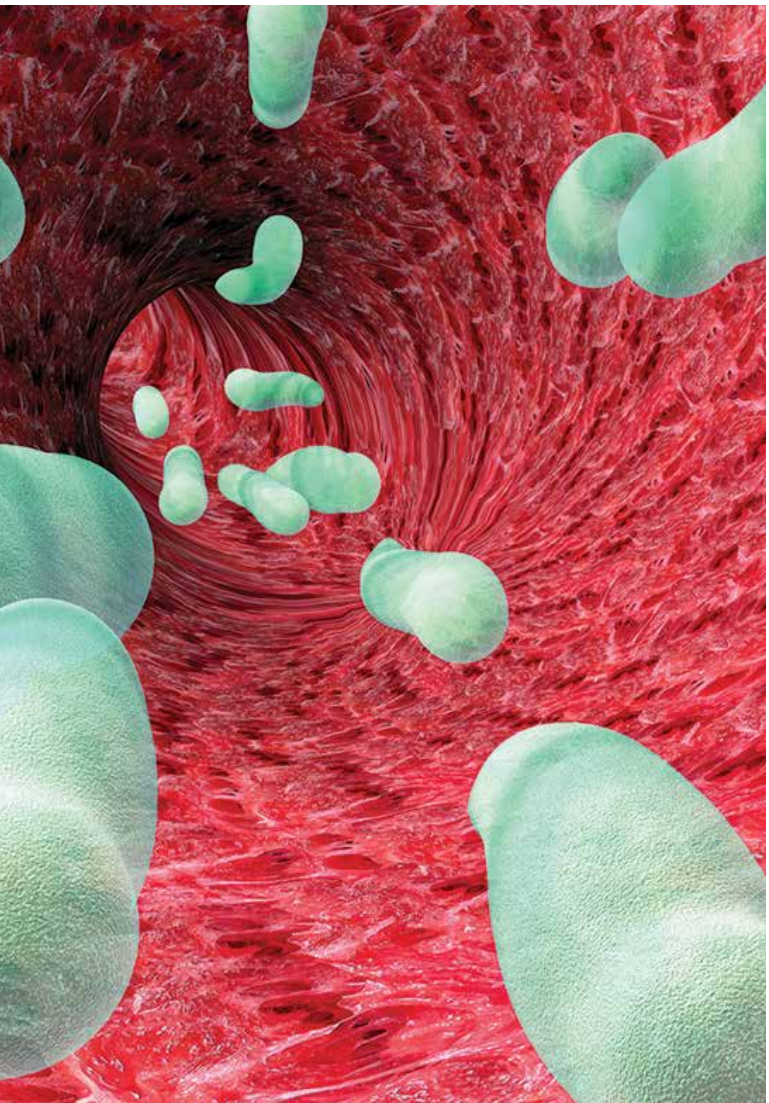
New Statistics for **Smelling Out Disease**

Professor Wenxuan Zhong
Professor Ping Ma



Scientia

NEW STATISTICS FOR SMELLING OUT DISEASE

Cutting-edge scientific techniques are generating a wealth of details about biological systems. This information could enable the rapid detection of disease and toxic substances, and provide insights into the mechanisms behind complex diseases. However, teasing out elusive signals in these biochemical datasets can be complicated, so statistician **Professor Wenxuan Zhong** and her team at the University of Georgia are developing new numerical methods to unlock their secrets.

The process of scientific discovery has always opened up exciting new insights into the natural world. As technology has developed, the amount of information scientists can collect has increased dramatically. This data offers unprecedented opportunities to understand the ways in which complicated biological and physical systems behave, to collect observations on how they change, and even to make predictions. These huge datasets record information on many scales: from telescopes monitoring the movements of stars, to the genetic sequence unique to every individual. The vast volumes of biochemical information now available to scientists have the potential to greatly improve healthcare. Identifying the specific data signatures associated with disease could allow the rapid detection of harmful substances, and unlocking the mechanisms of disease to opening up routes towards new treatments.

Smell-Seeing with the 'Optical Electronic Nose'

A striking invention whose powers of data detection have a variety of health-related applications is a little device sometimes known as the 'optical electronic nose'.

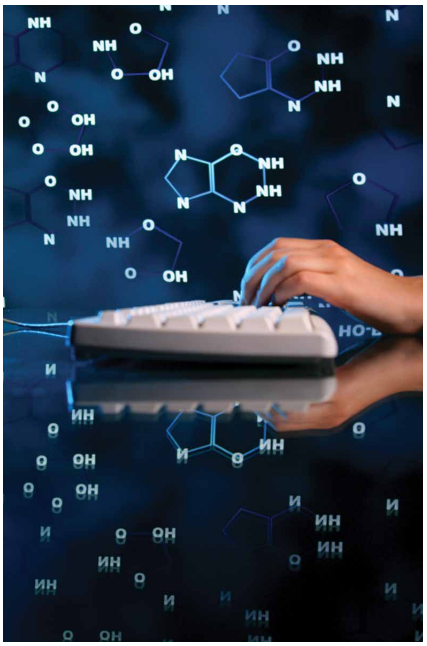
More technically called the colorimetric sensor array (CSA), it looks like a square with 36 coloured circles printed on it. But when the array is exposed to a mixture of chemicals, the dots change colour in a way that depends on the exact chemical profile of the mixture. Like the human nose and its ability to identify a smell based on the pattern of electrical signals it triggers in sensory receptors, the CSA produces a specific set of colour changes for the chemicals it detects, hence its nickname. CSAs can be small, cheap and accurate. Their many uses include detecting markers of disease in breath, including lung cancer and infections; identifying pathogenic bacteria based on the volatile substances they produce; and sensing volatile toxicants as part of security systems.

Noses to Numbers

The set of all the changes undergone by each dot produces a sort of fingerprint for the chemical signal, and can be represented numerically, as the change in each of the dot's red, green and blue colour components. This is written as a table with red, green and blue entries for each dot, also called a matrix. This representation allows scientists to compare several colour changes. In turn, knowing which colour changes correspond to which chemical mixture makes it possible to identify the chemicals in an unknown mixture.
However, classifying the chemicals accurately depends on statistical methods that can perform this classification well.

One common way of telling different groups apart from each other is a technique called linear discriminant analysis (LDA). LDA works by finding combinations of the properties, or predictors, of the group members, in such a way that when the groups are compared by these combinations of predictors, the differences between groups are larger than the differences between group members. However, this method is not ideal for CSA data, because the CSA data has a matrix structure.

To be suitable for LDA, the CSA data must be converted from its two-dimensional matrix



structure to a simple, one-dimensional form. This can lead to a substantial loss of the structural information we had when each dot's colour change in, say, red, was recorded separately. Another problem that comes up with this transformation is that each colour change for each dot is recorded as a separate property, or variable. This means that if we have 36 colour changes in each colour component, we end up with 108 separate variables. Thus, the parameters to be estimated are tripled. One of the main principles of statistical methods is that if we have a small set of experiments, but a large number of variables that we observe, it is very difficult to say anything meaningful about the differences between the variables we observe through our experiments. This is called the 'curse of dimensionality'. So, methods that are designed for the sort of data that CSA produces would be a great step forward in improving the accuracy of the data's interpretation, which is important in situations such as breath analysis to identify diseases.

New Methods for Making Sense of CSA Data

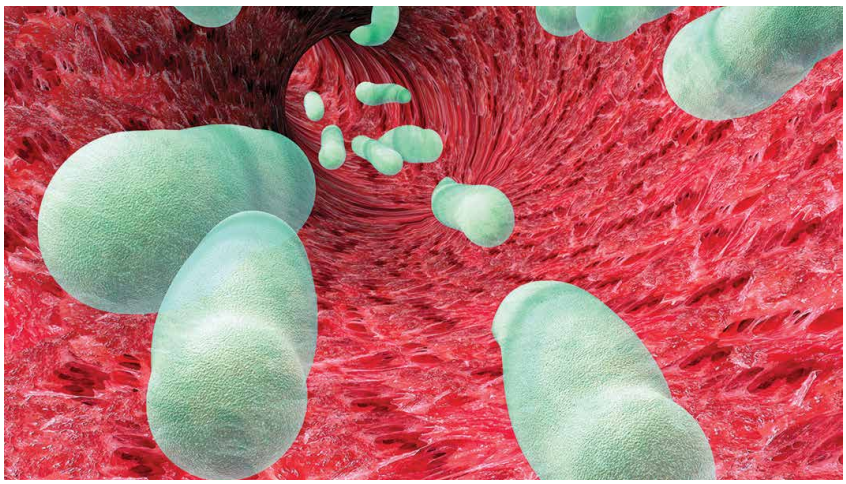
One scientist at the forefront of this type of statistical research is Professor Wenxuan Zhong. Professor Zhong runs the Big Data Analytics Lab at the University of Georgia, where she develops new statistical approaches to tackle the search for patterns in large, difficult datasets. She and Kenneth Suslick at the University of Illinois, who invented the original CSA technique, have created a new classification method tailored to CSA data. Presented in her 2014 publication in the journal *Technometrics*, Professor Zhong's 'matrix discriminant analysis' (MDA) approach takes the CSA matrix of colour changes and transforms it into two components, one for the rows, based on dyes, and one for the columns, based on colour changes. This preserves the information contained in the matrix structure, and allows us to escape the curse of dimensionality we'd get with 108 variables.

Professor Zhong has refined her MDA approach to work around another problem of CSA analysis, which is that not all dots produce a meaningful colour change for every chemical. So, incorporating the information from these misleading changes can lead to classification mistakes. The 'penalised MDA' approach reduces these errors, improving the accuracy of identifying chemicals. Professor Zhong tested the performance of her PMDA method in a real-life scenario – the detection of low levels of volatile toxic chemicals. Even in small quantities, long-term exposure to such chemicals can be dangerous. Professor Zhong's approach consistently outperformed the simpler LDA method, with lower classification errors.

The MDA approach to classifying CSA-like data can be taken a step further. Professor Zhong has developed an algorithm she calls 'Sequential Iterative Dimension Reduction' – SIDRA for short, which she describes in her 2015 publication in *Wiley Interdisciplinary Reviews: Computer Statistics.* Just like in the MDA method, SIDRA compresses the information from CSA-like data, more technically defined as a tensor, into lower dimension, which makes it easier to analyse. Importantly, the method ensures that no information about the data structure is lost in the process.

Understanding the Genetic Causes of Disease

Now Professor Zhong hopes to apply her statistical expertise to tackling some other tricky problems in biological data. She's still interested in improving health outcomes, but this time she's focussing on understanding



the causes of complex diseases, at the most basic level – DNA. As humans, our cells contain the genetic code for our species, but our bodies also host diverse microbial communities with their own DNA, such as the intestinal microflora. We now know that the exact microbial species we host, and their different abundances, can play an important part in regulating our health. This is thought to be a factor in conditions such as inflammatory bowel disease, type 2 diabetes, and obesity. Scientists can now take a sample of human gut flora and use next-generation sequencing techniques in a metagenomics approach to quickly build up a picture of the patient's complete microbial profile, without all the pitfalls of trying to cultivate microbes in the lab. However, there are still significant challenges to accurately identifying the species in samples like this.

Professor Zhong and her team have invented a powerful new method which sidesteps the need to use a reference genome, which can lead to biases in differentiating closely related species and in estimating their abundances. Metagenomic data consists of short DNA sequences from all the species in the sample. Unlike other methods, her MetaGen algorithm assumes that each species will have a unique relative abundance of a contig across the multiple samples – a sample profile. The sample profile of a contig should be the same as that of the species genome, and contigs with similar sample profiles are likely to be derived from the same genome. MetaGen uses this information to group the contigs into different species bins, and a method from information theory, the Bayesian information criterion, to determine the number of species. Professor Zhong hopes that MetaGen's fast and accurate approach to identifying microbial species and associated abundances will make it easier to associate microbial species profiles with specific diseases, and bring us a step closer to finding cures. Nick Nystrom, the Senior Director of Research and the Principal Investigator for the Bridges Supercomputer at Pittsburgh Supercomputing Center, commented that 'this was a really aggressive simulation or calculation where they were looking at 378 billion base pairs'.

Beyond Genes – Epigenetics

However, some diseases are thought to be shaped by DNA in ways that go beyond the genetic information it encodes. These 'epigenetic' mechanisms lead to potentially longer-term changes in gene transcription – the process in which RNA is produced from DNA, before being translated into proteins – without affecting the DNA sequence. Although some epigenetic mechanisms are a normal part of development, certain changes can lead to disease. Epigenetic processes can include the modification of histones, the proteins around which DNA is wrapped in the cell, and DNA methylation, in which a methyl molecular group is added to DNA.

DNA methylation is a relatively recent discovery, but has already been found to have an important role in essential processes such as the differential regulation of specific genes in tissues. While DNA methylation may occur in stable patterns that are inherited, it can also be altered in development and ageing, and by environmental factors. Faulty regulation of DNA methylation appears to be one of the causes of heart disease, diabetes and cancer. Next-generation genome sequencing techniques, first developed in the 1990s, have been further refined to allow the detailed mapping of DNA methylation across genomes.

Understanding how different DNA methylation patterns are associated with specific diseases could open up new routes to treatment, but presents a particular statistical challenge due to the high volume of data, and the interference of other signals, or 'noise'. Current methods of comparing methylation levels between cell types focus on aggregating the methylation levels at different genome sites into a single statistic, and comparing the statistics obtained between cell types.

However, this method overlooks the pattern of methylation at different sites, and this information is known to play an important part in regulating gene expression. The relationship between levels of DNA methylation and expression also appears to be complex, and classical correlation approaches do not accurately describe the way in which gene expression varies with methylation.

Professor Zhong thinks she may have invented a method that could accurately classify these challenging DNA methylation patterns, and allow us to identify how individual patterns could be characteristic of particular diseases. Their idea is based on building a mathematical model which predicts the methylation level of a gene, and uses information about its cell type and methylation levels at particular sites. They have already conducted a preliminary test to see whether the new method can distinguish between the different DNA methylation patterns in the genomes of patients with two variants of leukaemia. Comparing the graphs of DNA methylation levels that the model predicts from methylation sites and cell types, there are striking differences between the two sets of patients, proving that the model has the potential to perform well.

Professor Zhong now plans to upscale the model, identifying differentially methylated regions across the human genome and linking these to cell types. With some additional statistical ideas drawing on her earlier work on dimension reduction, this will allow her team to predict changes in gene expression linked to differential methylation patterns. Finally, she hopes to make the model widely available by designing a user-friendly piece of software that researchers can use, giving us another exciting new tool that could revolutionise prospects for treating a suite of diseases.

Meet the researchers

Professor Wenxuan Zhong

Department of Statistics
University of Georgia
Athens, USA

Professor Wenxuan Zhong completed a PhD in Statistics at Purdue University in 2005, and went on to carry out postdoctoral research in the Department of Statistics and FAS Center for Systems Biology at Harvard University until 2007. She then became Assistant Professor at the Department of Statistics of the University of Illinois at Urbana-Champaign. Professor Zhong is currently an Associate Professor at the Department of Statistics of the University of Georgia, where she leads research on developing statistical theory and methodology to address the challenges of analysing large volumes of genetic data. She is the founding director of the Big Data Analytics Lab, and its Principal Investigator together with Professor Ping Ma.

CONTACT

E: wenxuan@uga.edu
T: (+1) 706 542 0120
W: http://www.stat.uga.edu/people/wenxuan-zhong

Professor Ping Ma

Department of Statistics
University of Georgia
Athens, USA

Professor Ping Ma completed a PhD in Statistics at Purdue University in 2003, before carrying out a postdoc from 2003–2005 at Harvard University in Bioinformatics and Computational Biology. In 2005 he became Assistant Professor at the Department of Statistics of the University of Illinois, and was Associate Professor from 2011 to 2013. In 2014 he moved to the University of Georgia to take up the role of Associate Professor at the Department of Statistics, and since 2015 he has been Professor at the Department. Together with Professor Wenxuan Zhong, he is Principal Investigator of the Big Data Analytics Lab.

CONTACT

E: pingma@uga.edu
T: (+1) 706 542 0714
W: http://www.stat.uga.edu/people/ping-ma

FUNDING

National Institutes of Health
National Science Foundation

REFERENCES

W Zhong and K Suslick, Matrix Discriminant Analysis with Application to Colorimetric Sensor Array Data, Technometrics, 2014, 57, 524–534. DOI: 10.1080/00401706.2014.965347

W Zhong, X Xing, and K Suslick, Tensor sufficient dimension reduction, Wiley Interdisciplinary Reviews: Computational Statistics, 2015, 7, 178–184. DOI: 10.1002/wics.1350

Research featured in The International Conference for High Performance Computing, Networking, Storage and Analysis (SC16): http://on-demand.gputechconf.com/supercomputing/2016/presentation/sc6118-nystrom-nyawira-converged-hpc-big-data-system-hpc-research.pdf

Research featured in The Intersect360 Research Podcast 'This Week in HPC': http://www.intersect360.com/industry/podcast_transcripts/TWIHPC%20Transcript%20- %20Intel%20OmniPath%20Launch%20-%20May%202016.pdf

Research featured in Intel Company website: http://itpeernetwork.intel.com/hpc-innovation-delivers-best-in-class-industry-solutions/

Research featured in Pittsburgh Supercomputing Center Research Highlights: https://www.psc.edu/news-publications/30-years-of-psc/161-news/psc-highlights/2414-bridges-reveals-diabetes-gut-microbe-links-2

Research featured in The Extreme Science and Engineering Discovery Environment (Xsede) Science Successes https://www.xsede.org/opening-bridges

UNIVERSITY OF GEORGIA