Developing powerful statistical methods for the analysis of large-scale genomic studies

Analysing large amounts of DNA sequences poses a huge challenge for scientists. The research of **Professor Xinge (Jessie) Jeng** at North Carolina State University focuses on cutting-edge techniques in modern statistics which are applied to genomic research. Her work provides efficient statistical tools and powerful computational methods to identify causative mutations at the single-nucleotide resolution.

With the emergence of high-throughput technologies, it is now possible to detect information in large-scale and high-dimensional datasets. The demand for methods of detection and estimation of sparse signals has never been greater. Such tools have potential to impact upon a wide spectrum of applications including genomic data analysis, identification of astrophysical sources, detection of covert communication, and monitoring of outbreaks in disease surveillance. This type of data analysis however is riddled with conceptual and technical challenges. The need for computationally efficient and statistically optimal methods to discover information-bearing signals in big datasets is what motivates Dr Xinge Jessie Jeng at North Carolina State University and her work in this field has been well-recognised.

A fundamental challenge of interpreting high-dimensional data is that the relevant, true signals are obscured by irrelevant data, or noise. For low- to moderate-dimensional data, several classical statistical methods can be used, which focus on the identification of strong, true signals. Such methods, however, cannot be extended to cases where data dimension is much larger than the sample size and weak true signals are surrounded by significant amounts of noise (as is the case for modern genomic analysis). In addition, the range of noise tends to increase with the dimensionality of data, often rendering existing methods to be impractical.

TRICHOTOMOUS FRAMEWORK

Dr Jeng's long-term aim is to address these difficulties. A leader in this field she is developing new approaches to analyse high-dimensional data. In particular, she has introduced a novel method seeking to facilitate the detection of weak signals. One of the significant upshots of this approach is that a high proportion of weak signals can be retained for follow-up study. The proposed procedure is well-adapted to deal with unknown features of the datasets including signal intensity and sparsity. Dr Jeng tested her proposal by applying it to real-world data stemming from genomic research. A series of simulation studies evaluated the efficiency and computational speed of the newly developed method. She observed that high-dimensional data can be considered as being composed of three disjointed subsets. The first subset includes strong signals, the second one random noise, while the third relatively weak signals intertwined indistinguishably with noise. Dr Jeng has designated this type of approach as the 'trichotomous framework'.

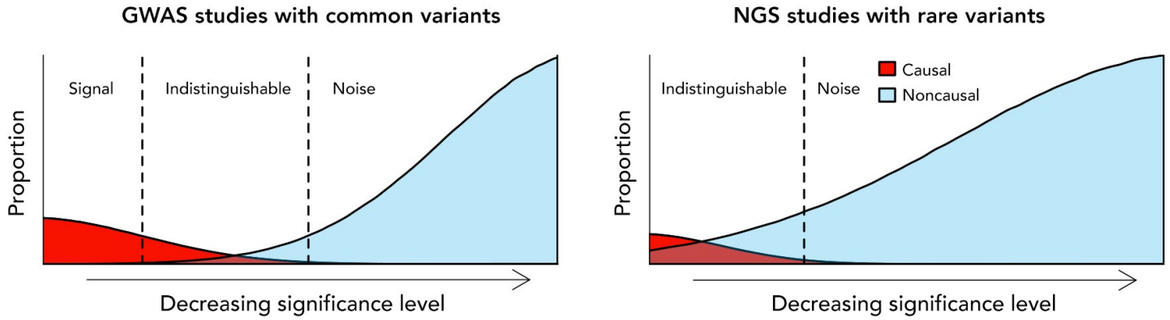
DETECTING THE TRUE SIGNALS

Dr Jeng then focused on the challenge of detecting the true signals which belong to the latter subset, i.e. the signals that are indistinguishably mixed with noise. She proposed a data-driven screening procedure to identify the mixed and noise subsets, and a procedure to retain the relatively weak signals in the mixed subset. Dr Jeng provided a theoretical analysis which shows that the methods she uses retain true signals with high probability. Moreover, the trichotomous framework has been also successfully tested in practice by being applied to certain problems concerning analysis of large DNA datasets.

UNDERSTANDING GENETICS

Much of Dr Jeng's work is focused on the problem of detecting and identifying sparse short segments in long one-dimensional

The challenge of interpreting high-dimensional data is that the relevant, true signals are obscured by irrelevant data, or noise

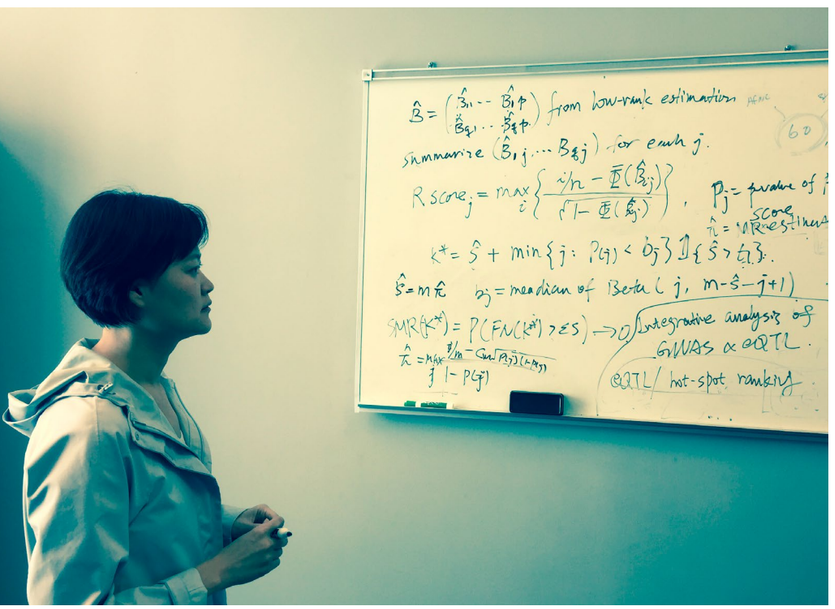


GWAS: genome-wide association study. NGS: next-generation sequencing

sequence of data. This problem is very closely related to human genetics where the central goal is to understand the inherited basis of human variation in phenotypes. Of crucial importance in this context is the so-called DNA copy number variation (CNV) analysis. CNV is defined as duplication or deletion of a segment of DNA sequence compared to a reference genome data.

CNV occurs when sections of the genome are repeated or lost. Evidence now exists showing that CNV might influence human phenotypes in a variety of manners. It has long been known that sporadic diseases, termed `genomic disorders', are caused by structural alterations of the genome. Apart from causing sporadic diseases, CNVs play a role in the formation of mendelian diseases.

Dr Jeng and her team focus their efforts on providing methods for high-throughput data analysis, bordering on the brink of what is statistically possible



Dr Jeng's approach yields an effective method of detection for CNV segments. Motivated by the problem of CNV detection and other applications, she considered the general problem of detecting sparse and short segments from a long sequence of noisy data. Under realistic assumptions, she obtained a number of results that pertain to statistical research in several areas. Genomic analysis is notoriously challenging, since the true signals are very sparse (that is, both the number and the lengths of signal segments are very small). To get a sense of the numbers involved, it is estimated that there are about 500,000 to 1,000,000 numerical observations along the genome of an individual human. In contrast, there are usually less than 100 CNV segments. Detection is not an easy task.

Dr Jeng and her colleagues have developed a statistical characterisation of identifiable region of a signal segment. They proposed a selection procedure, the likelihood ratio selection (LRS) procedure, to identify the signal segments. Their simulation results demonstrate that this tool is more powerful and greatly outperforms current signal identification methods. A detection method is optimal if it works in all the situations where successful detection is possible. Dr Jeng's newly developed method does just that.

As high-throughput technologies progress in the scientific world, the tools to analyse results from these techniques need to progress at a similar rate. Until now research has only focused on optimal signal detection for relatively ideal situations. Dr Jeng's research is leading the way in developing new statistical methods for complex, real-world problems. Her research borders on the brink of what is statistically possible,

Q&A

Your research into large-scale next-generation sequencing will have a huge impact on genomic data analysis. How did you become interested in this field?
When I was doing my postdoc at the University of Pennsylvania, I had the chance to analyse next-generation sequencing datasets collected by the Children's Hospital of Philadelphia for a neuroblastoma study. Neuroblastoma is a rare type of childhood cancer. The advent of new genotyping technology makes it possible to explore the genetic association of rare variants to this complex disease. I felt the big potential of next-generation sequencing and, at the same time, became concerned about the limitations of the available analytic tools.

What has been the most enjoyable and interesting aspect of your work?
The most interesting aspect of my work has been the feeling of dedication to a scientific discipline from my unique perspective. The most enjoyable moments are the occasions when beautiful connections are established between mathematical derivations and real-world problems.

Regarding detection of CNV segments, how hard do you think it would be to extend your methods to more general scenarios (e.g. identification using data from multiple sequences)?
My team has extended the methodology framework to multiple sequences for the discovery of recurrent CNVs. Notable groups at the University of Pennsylvania and Harvard have made important strides by building upon the framework to detect short signal segments with arbitrary signal shapes and signal regions in genome-wide association studies.

What is the most challenging problem that you are currently trying to solve?
It is hard to identify the most challenging problem among the research problems that I'm working with. Each research project deals with a non-trivial challenge which could evolve even further as we learn more about the underlying features of the problem. In a more general sense and as expounded by the physicist and philosopher Thomas Kuhn, it is often challenging to introduce new analytic frameworks when scientific practice is dominated by a prevailing one. However, I am glad that the wider scientific community is gradually recognising that classical tools often do not solve their problems and the need for new thinking in the data sciences.

What is next for your research?
The driving force for my research is the feeling of uneasiness when current techniques hit a boundary. A current source of the uneasiness comes from how limited one method based on one dataset can contribute to signal discovery under high-dimensionality. I plan to explore multi-stage methods, integrative data analysis, and cross-sample replicability in future research.

The driving force for my research is the feeling of uneasiness when current techniques hit a boundary

spanning important areas of statistics including high-dimensional regression, sparse and weak signal discovery, large-scale hypothesis testing, dimension reduction, and robust inference. Her work has been well-recognised. In 2015 she received the Young Investigator award from the National Security Agency and the 2010 David P. Byar Young Investigator Award from the American Statistical Association. Her ultimate aim is to develop a novel analytic framework to evaluate the estimation uncertainty of high-dimensional data, that will have a far-reaching impact on genetic data analysis.

Detail

RESEARCH OBJECTIVES
Dr Jeng's research interests include High-Dimensional Inference, Multiple Testing, Model Selection and Bioinformatics. Her current research builds on her background in high-dimensional inference and sparse signal detection for the analysis of large-scale genomic data. Her aim is to provide powerful and efficient new statistical tools and computational methods to allow investigators to identify causative mutations at the single-nucleotide resolution.

FUNDING
- National Institutes of Health (NIH)

COLLABORATORS
- Prof T. Tony Cai
- Dr Z. John Daye
- Prof Jiashun Jin
- Prof Hongzhe Li
- Prof Wenbin Lu
- Prof Jung-Ying Tzeng

BIO
Dr Jeng completed her doctorate in Statistics at Purdue University in 2009 before joining the Wharton School and the Perelman School of Medicine at the University of Pennsylvania as a postdoctoral researcher. Since 2012, she has been an Assistant Professor in the Department of Statistics at North Carolina State University.

CONTACT
Dr X. Jessie Jeng
Assistant Professor
Department of Statistics
North Carolina State University
NCSU Statistics Department
2311 Stinson Drive
Campus Box 8203
Raleigh, NC 27695-8203
USA

E: xjjeng@ncsu.edu
T: +1 919 515 0612
W: https://sites.google.com/site/xingejeng/