An integrated toolkit for high-dimensional complex and time series data analysis

Big data can be too large and complex for traditional methods and conventional software packages to deal with. Dr Fang Han, Assistant Professor in Statistics at the University of Washington, Seattle, is meeting this challenge head on. He is creating an integrated statistical toolkit comprising robust statistical procedures, including distribution-free inference and rank-based methods, which can be applied to high-dimensional datasets. These methods are designed to offer robust as well as efficient solutions to various data analysis problems, revealing hidden patterns as Dr Han demonstrates with biological systems.

Advances in technology and the advent of big data mean that huge sets of high-dimensional, unstructured data are becoming common place; as is the need to collect, store and process them. For example, stock market analysis, genetic testing, and magnetic resonance imaging all produce massive amounts of high-dimensional data. These data sets are too large and too complex to be dealt with using traditional methods and conventional software packages often struggle to handle them.

COMPLEX DATA CHALLENGE

Statisticians are faced with new challenges from this complex data. These enormous quantities of very high-dimensional data can be skewed, exhibit nonlinear relationships and contain useless information or noise, preventing them from being analysed by traditional parametric or linear methods. Dr Fang Han, Assistant Professor of Statistics at the University of Washington, Seattle, is meeting this challenge head on. His research focuses on high-dimensional statistical theory and its application in order to resolve statistical problems in the fields of economics, finance, and science. He highlights a requirement for statistical methods that can be scaled up to handling large datasets and acknowledges that while these methods have to be able to capture the subtleties of the particular area of interest, they also have to be able to cope with different modelling assumptions and data contamination. He also draws attention to the development of statistical theory and methodology generally lagging behind the development of new technologies.

A NEW INTEGRATED TOOLKIT

Dr Han is creating a statistical toolkit by developing robust statistical procedures that can be applied to high-dimensional datasets. This collection of novel statistical methods combines advanced statistical modelling approaches and innovative probability procedures underpinned by high-dimensional statistical theory. Dr Han explains: "It lies on the fact that the theory can tell you when a method does or doesn't work, and in the latter case, how to lead you to a 'correct' solution."

HIGH-DIMENSIONAL DISTRIBUTION-FREE INFERENCE

Dr Han is particularly interested in high-dimensional distribution-free inference, where the aim is to carry out statistical inference, i.e. *statistical* analysis that infers properties of a population, on high-dimensional data, while imposing as few assumptions as possible. Consider the stock market data for example, if we want to know whether there is frequent interaction between particular stocks. Traditionally, statisticians would assume that the returns are distributed with a normal (bell-shaped curve) distribution so they would carry out associated hypothesis tests. This assumption, however, is unnecessary and likely to create problematic outcomes. Stock market returns have shown to be highly skewed, so the assumption of normality is unsound. Dr Han has published his work on distribution-free tests of independence in high-dimensional data, where he shows that a carefully designed, rank-based test can produce robust results to such problems with minimal assumptions.

High-dimensional time series data is routine in medical data collection.

NONPARAMETRIC/ SEMIPARAMETRIC REGRESSION

Another area of importance to Dr Han's work is nonparametric/semiparametric regression. Regression is a popular tool used to locate relationships between variables. Parametric regression involves estimation of a finite number of parameters (e.g., the linear relationship between two features). In contrast, nonparametric regression involves estimating parameters of infinite dimension, while semiparametric regression combines both parametric and nonparametric models. As containing a nonparametric component, the model involves parameters of infinite dimension; while the target, like the parametric regression, is to infer a particular parameter subset that contains only finitely many elements.

The objective of nonparametric/ semiparametric regression is to create a statistical procedure that takes observed data and enables the automatic selection of the most suitable function to represent the link between the variables. Dr Han is designing robust nonparametric/ semiparametric regression procedures for high-dimensional data where outliers, data contaminations, and model misspecification have minimal effects on the outcomes. For example, in genetic testing when we want to investigate the relation between a particular disease and a number of gene markers, the data could contain measurement errors (data contamination), the wrong statistical model could be applied to the data, and the results would be incorrect.

HIGH-DIMENSIONAL TIME SERIES ANALYSIS

Dr Han's research into the mathematical theories supporting time series analysis has revealed that high-dimensional time series data has received relatively little attention to date. Almost every scientific database, however, contains massive amounts of high-dimensional time series data. This highlights the urgent requirement for high-dimensional time series analysis tools. Dr Han has provided methods for monitoring parameters in both time and frequency that are upheld by both statistical estimation theory and innovative probability tools.

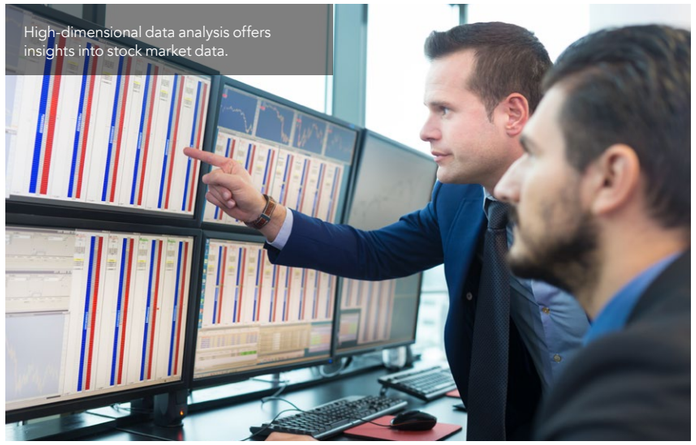
This collection of novel statistical methods combines advanced statistical modelling approaches and innovative probability procedures underpinned by high dimensional statistical theory.

Dr Han's peer reviewed research, however, demonstrates that his carefully designed nonparametric/semiparametric regression procedures can avoid these hazards and produce reliable outcomes.

Dr Han's statistical analysis has huge potential implications for scientific data analysis.

RANDOM MATRIX THEORY
The general mathematical theories that these statistical problems provoke have become a passion for Dr Han. This includes random matrix theory, exploring the properties of large matrices where the elements of the matrices are randomly generated. The properties of the eigenvalues and eigenvectors derived from these matrices, for instance, are of strong research interest. The theoretical results of random matrix theory can be applied to numerous problems in statistics, economics, and other fields with the potential to develop statistical techniques for high-dimensional problems. Dr Han has published and is working on more results to enhance our understanding of the mechanism of large random matrices.

INTEGRATED TOOLKIT APPLICATIONS
Within the integrated toolkit, Dr Han is developing and applying procedures such as high-dimensional generalised regression models, shape-constrained regression models, partially linear models, and copula time series models. These methods are designed to achieve optimality under a statistical criterion, offering robust statistics that can reveal hidden patterns in various scientific problems.


High-dimensional data analysis offers insights into stock market data.

To name one, rapid advances in technology, and in particular medical imaging, means that neuroscience has an enormous amount of data available in the form of functional magnetic resonance imaging (fMRI), electroencephalograph (EEG) and positron emission tomography (PET) data. The analysis of these data increases the potential for research into the brain connectivity network offering neuroscientists the opportunity to explore how the brain changes with conditions such as Alzheimer's disease. The analysis of this high-dimensional data can offer answers to rudimentary questions like brain functional connectivity.

These methods are designed to be optimal, offering robust statistics that can reveal hidden patterns in biological systems.

Colossal amounts of genome sequencing data are also being collected, and high-dimensional data analysis can also offer insights into issues such as how transcription factors control gene activities. A transcription factor is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA (ribonucleic acid). A transcription factor regulates genes, by turning them on and off, to ensure that the genes are expressed in the correct cell at the right time and in the appropriate amount throughout the life of the cell and the organism.

The integrated toolkit offers powerful analysis procedures, and when combined with neuroscientists' and biologists' expertise these questions can be investigated, their subtleties addressed and hidden patterns in the data can be further uncovered.

BROADER IMPACT
Dr Han is collaborating with genomic and neuroscience researchers at John Hopkins University, University of Washington, and Fred Hutchinson Cancer Research Center. It is anticipated that the outcomes will have a direct impact on the future scientific development of genomic and brain imaging data analysis. Dr Han is also designing software packages to allow easy access to his methods. Finally, he has developed and is developing courses on cutting-edge statistics, probability, and machine learning tools, which will provide a vehicle for disseminating the integrated toolkit for high-dimensional, complex and time series data analysis to scientists working in related areas.

Behind the Research
Dr Fang Han

E: fanghan@uw.edu **T:** 1-206-221-6560 **W:** https://www.stat.washington.edu/people/fanghan/

Research Objectives

Dr Han's research interests include high-dimensional statistics and robust statistics.

Detail

PDL Hall B-313
University of Washington
Seattle, WA 98195

Bio
Dr Han was awarded his B.S. degree in probability and statistics from Peking University, China, his M.S. and Ph.D. in biostatistics from University of Minnesota and Johns Hopkins University. He is currently an Assistant Professor in statistics at University of Washington, Seattle.

Funding
National Science Foundation

Collaborators
- Mathias Drton
- Yanqin Fan
- Daniela Witten
- Jianfeng Yao
- Cunhui Zhang
- Brian Caffo
- Chongzhi Di

References

Gao, C., Han, F. and Zhang, C. (in press). On Estimation of Isotonic Piecewise Constant Signals. *The Annals of Statistics.*

Han, F., Ji, H., Ji, Z. and Wang, H. (2017). A Provable Smoothing Approach for High Dimensional Generalized Regression with Applications in Genomics. *Electronic Journal of Statistics,* 11(2), 4347-4403. https://doi.org/10.1214/17-EJS1352.

Han, F., Chen, S. and Liu, H. (2017). Distribution-Free Tests of Independence in High Dimensions. *Biometrika,* 104(4), 813-828. https://doi.org/10.1093/biomet/asx050.

Personal Response

What initially prompted your interest in high-dimensional data?

"High-dimensional data are everywhere nowadays. From a practical perspective, these datasets call for new and careful statistical analyses. From a more theoretical perspective, these datasets often possess structures of rich hidden information. They motivate statisticians to develop innovative statistical thoughts as well as novel mathematical tools. I myself am constantly attracted by problems stemmed in high-dimensional data analysis from both perspectives."

What are your future plans for the integrated toolkit?

"There are two trajectories. First, I would love to dig deeper along the tracks of high-dimensional robust statistics. For example, I have seen the power of rank statistics in analysing big complex data, achieving the statistical goal with little information lost. It is now the right time for me to explore and exploit the potential here. Second, I would love to intensify my current collaborations with genomic and neuroscience experts as well as explore more opportunities for collaboration. The toolkit offers statistical tools tailored to specific problems. It is motivated by the collaboration with scientists, and it will also be used by them."