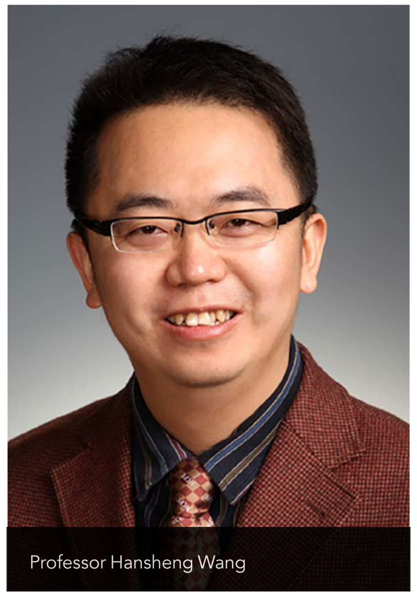
PopTika/Shutterstock.com

Professor Hansheng Wang on the importance of statistical analysis

Professor Hansheng Wang is Director of the Business Analytics Programme at the prestigious Guanghua School of Management, Peking University, China. Recently named one of China's most highly cited researchers, he has written over 100 English research papers, and over 20 Chinese papers. He is now a member of the International Statistical Institute, the American Statistical Association, the Institute of Mathematical Statistics, the Royal Statistical Society, and the International Chinese Statistical Association. His varied statistical and econometrical research covers everything from sample size collection to business analysis. Research Features were privileged to speak with him about his research.

It has become increasingly apparent that rigorous statistical insight is essential in understanding everything from deep learning algorithms to everyday consumer data. The influence of these forms of analysis can seem so ingrained in our theoretical understanding that we overlook their importance. Professor Hansheng Wang is well-equipped to unpack the complex terminology and methodology behind statistical analysis. We spoke to him about the relationship between statistics and deep learning, the future of the field, and about his own serendipitous entry into the profession.



Professor Hansheng Wang

Could you introduce us to some of the research you are currently undertaking?

My current research mainly focuses on developing a statistical understanding of deep learning methods, both theoretically and computationally. Deep learning methods (ie, various deep neural networks) are arguably the most important and successful machine learning methods used for various AI-related applications. Despite their successful application, these methods suffer from incredibly complex model structures, and an extraordinarily large number of parameters need to be estimated when using them. Furthermore, there is a lack of theoretical explanation for the widespread success of these methods.

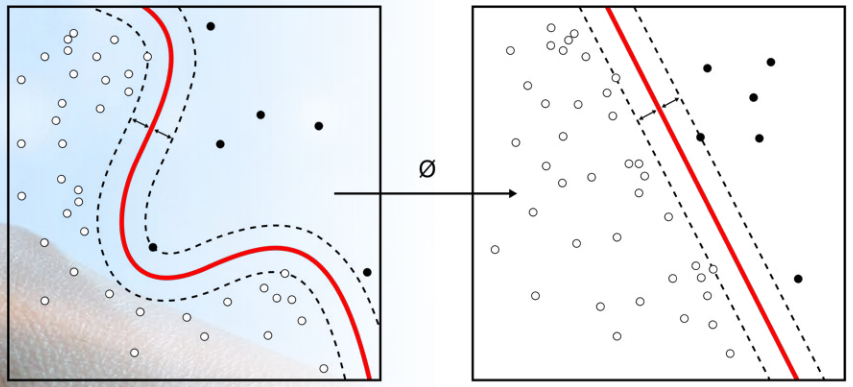


Photo Credit: Original: Alisneaky Vector: Zirguezi, CC BY-SA 4.0, via Wikimedia Commons

Gradient descent is the algorithm with which most deep learning methods are trained. This scatterplot represents a machine's so-called 'decision boundary'.

I believe a good theoretical understanding about deep learning from a statistical point of view is of great importance. Firstly, this form of understanding about deep learning might lead to simplified model structures and a significantly reduced number of model parameters to be estimated. In the meantime, empirical performances should not be sacrificed significantly. Secondly, a good statistical understanding might lead to better training methods for model parameter estimation. For example, most deep learning methods have been trained by a method known as 'gradient descent'. To implement this algorithm, an important tuning parameter (ie, the so-called 'learning rate') needs to be subjectively specified. A good statistical theory might provide an optimal and objective selection of this important tuning parameter. To summarise, I believe a good statistical understanding about deep learning models should be important for both statistical theory and deep learning research.

What originally appealed to you about the study of statistics?

I came to study statistics purely by accident. When I applied for my undergraduate studies at Peking University, I had no idea about what to

> **Professor Hansheng Wang is well-equipped to unpack the complex terminology and methodology behind statistical analysis.**

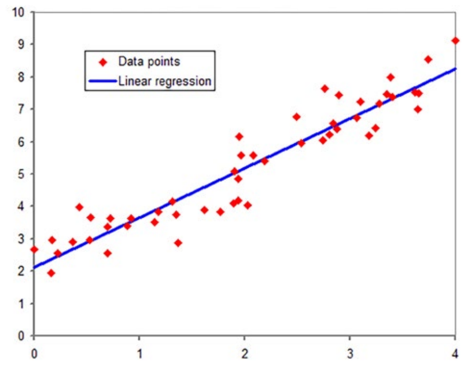
batjaket/Shutterstock.com

Ultra-high dimensional data analysis offers more nuanced assessment than traditional analysis.

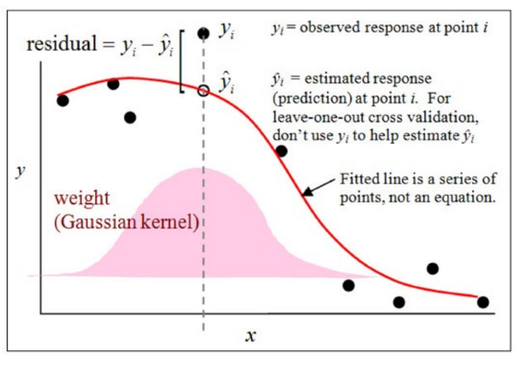
learn in the future and I had no real idea what the study of statistics is all about. So, I followed other people's advice and applied for International Business and Computer Science. At that time, those were highly popular undergraduate majors and a good many people were applying. The competition was very keen, and I obviously failed to get in. Then, the admission committee of the university made the decision for me and assigned me to the Department of Probability & Statistics, partially because my mathematics performance was good. So, I came to statistics by a beautiful accident, and it turns out I love it so much that I made it my lifetime career.

What does 'ultra-high dimensional data analysis' involve?

Consider, for example, that we are analysing a consumer-related dataset, which was collected by large online retailers (for example, Amazon, or the Chinese e-commerce company JingDong). The marketing manager might want to promote a new product and he/she wants to understand what kind of consumers might be more likely to purchase the product than others. Consequently, the manager needs to build up a 'regression model' to establish the dependent relationship between consumer purchase behaviours (that is, whether they buy or not) with a number of consumer- and product-related features.



A regression line.



residual = $y_i - \hat{y}_i$
y_i
y_i = observed response at point i
$\hat{y}_i$
$\hat{y}_i$ = estimated response (prediction) at point i. For leave-one-out cross validation, don't use y_i to help estimate $\hat{y}_i$
Fitted line is a series of points, not an equation.
y
weight (Gaussian kernel)
x

Photo Credit: Brmccune, CC BY-SA 3.0, via Wikimedia Commons

An example of Gaussian kernel smoothing.

Traditional data analysis might not be able to collect many features about (for example) the consumer. These traditional features are most likely demographic variables, and these are not overly varied or complex. However, with the rapid development of information technology, the retailer should be able to derive a huge number of features from the consumer's online behaviour (including their browsing path, their product comments, social networks, etc). Those features constitute a feature vector of ultra-high dimension analysis. In many cases, this feature dimension might be even larger than the sample size. That creates a serious challenge regarding the related statistical models and calls for novel methodology. That, in a nutshell, is what is involved in 'ultra-high dimensional data analysis'.

Could you introduce our readers to some of the non-parametric models you use in your statistical research?

To address this question, we might need to define two terminologies carefully. First, what is a regression model? Second, what is a nonparametric regression model? By regression model, we mean any model which attempts to establish a dependent relationship between a response variable, 'Y', and a set of features, 'X'. Technically, how to establish this dependent relationship mathematically becomes a problem of great importance. In fact, there exist many different choices, ranging from very simple to highly complicated.

> A good statistical understanding might lead to better training methods for model parameter estimation.

A relatively simple model would attempt to assume a very specific model structure for X and Y, which is often one of a variety of linear model structures. Those models are typically referred to as 'parametric models'. The key feature of a parametric model is that the total number of unknown parameters is relatively small compared to its nonparametric counterpart. With nonparametric models, we place no stringent model structure between X and Y. Since no stringent model structure is assumed, practically we have to rely on a working model with a larger number of unknown parameters to approximate it. Then, 'what kind of working model should we use?' becomes the key question.

By making different choices about the working model, different nonparametric models can be developed. For example: there are a range of models known as 'kernel smoothing', 'wavelet', 'spline', 'tree', 'support vector machine', and neural networks. Different researchers often have their preferences. For me, I like nonparametric kernel smoothing and neural networks very much. I like the former because it has a very deep and beautiful statistical theory to support it. I like the latter because it is the fundamental building block of all deep learning methods.

Can you predict any trends in statistical research, or areas of study which you think will become increasingly important in the coming months and years?

Given the fact that various so-called 'unstructured' data (including images and natural languages) are becoming increasingly available, I believe that statistical research related to these applications is becoming more and more important. In this regard, various deep learning models have proved their efficacy, and explaining their tremendous success through statistical analysis will become a very important direction for intensive research. That is why my research now focuses primarily on the statistical understanding of deep learning methods.

What are your future research plans?

Going forward, my research will continue to focus on a statistical approach to deep learning methods. The research in this regard is very preliminary, but the progress we have made so far is very encouraging. So, I would expect to continue in this research direction for the next five or even ten years.

E: hansheng@pku.edu.cn