Some Considerations on Mathematical Statistics for Dealing with Big Data from the Viewpoint of Educational Innovation

Wei Wang

School of Information, Renmin University of China, Beijing 100872, China

Abstract. Mathematical statistics is an important branch of mathematics. It is also an important tool to other sciences, such as social science, management science, etc. With the development of information science and technology, on one hand, the appearance of softwares, such as SAS, SPSS, etc., the tasks of computation which related to statistical analysis can be carried out by softwares, on the other hand, the appearance of big data brings about new challenges for mathematical statistics. What should we learn or teach in the future? In this paper, from the viewpoint of educational innovation, and based on the understanding of big data, we consider the possible reformation on the related methods of mathematical statistics in the future. We intend to provide some suggestions on the methods of statistical analysis which will be furtherly improved or developed, especially the analysis methods for the data with a whole sample, and statistical inferences for figures, sounds and words.

Keywords: Mathematical statistics; Educational innovation; Big data; Statistical analysis/inference.

1. Introduction

It is well known that mathematical statistics (MS) concerns about the application of mathematics to statistics, which was originally conceived as the science of the state, i.e., the collection and analysis of facts about a country: its economy, land, population, property, military, and so forth. Mathematical statistics has been inspired by and has extended many options in applied statistics [1].

As a science, MS is concerned with the planning of studies, especially with the design of randomized experiments and with the planning of surveys using random sampling. The initial analysis of the data from properly randomized studies often follows the proper study protocol. The data from a randomized study can be analyzed to consider secondary hypotheses or to suggest new ideas. A secondary analysis of the data from a planned study uses tools from data analysis. It is the practical and programmable methods that make MS a useful branch of mathematics and with applications in many fields [2].

However, with the development of information science and technology, on one hand, the appearance of softwares, such as SAS, SPSS, etc., the tasks of computation related with statistics can be carried out by using softwares, on the other hand, the appearance of big data arouses new challenges for mathematical statistics [3, 4]. What should we learn or teach in the future?

In this paper, from the viewpoint of educational innovation, and as a further consideration of the ones we discussed in [5], we consider the possible aspects that MS will be reformed or developed in the future, especially the methods on statistical analysis or statistical inference for big data. In section II, the necessary on the reformation is considered. In section III, the methods, which with neural engineering and dynamical system theory, which may be used for the analysis of big data with a whole sample, are considered. In section IV, the methods of statistical inferences related to figures, sounds and words are discussed. Followed are some conclusions.

2. The challenges of mathematical statistics with the appearance of big data

With the evolution of sciences and various modern technologies, we are in the midst of what is popularly called information revolution. Intentionally and/or accidentally, the generation of the data is inevitable. Huge amount of data is being constantly generated and collected around us, and we are living in a so-called world of knowledge. As a result, large data, broadly characterized by three Vs, i.e. large volume, velocity and variety, popularly known as ‘big data’, is becoming a fancy word [3]. At the same time, the analysis, access and store of these data are now central to various scientific

innovation, public health and welfare, public security, and so on. Moreover, big data are highly complex in nature, and most of the information is heterogeneous, time varying, redundant, uncertain and imprecise. The mining the data or information is not straight forward. To reason, understand and mine useful knowledge from these data is becoming a great challenge, especially for mathematical statistics [4]. Owing to the possibilities that big data may bring with us, scientists from the academia, industry, and open source community have been trying for improved reasoning and understanding of big data and to provide better scope to solve several social and natural problems. The scholars in computer science are working on the machine learning. It may be mentioned that while analytics over big data is playing a leading role, there has been a shortage of deep analytical talent globally ease of use. There also propose some urgent problems for mathematical statistics to deal with.

The problem here is that, in such circumstances, how can MS be improved or developed to meet the needs for the analyzing or mining of big data? In the following sections, we will concern about the possible aspects of mathematical statistics which can be reformed, or the possible novel methods of statistical inference which can be developed for the emergence of big data.

3. Statistical methods for big data which are heterogeneous and the data with a whole sample

In this section, we consider the basic methods for dealing with big data which are heterogeneous or with a whole sample.

3.1 Statistical analysis for data which are the mixtured components

Based on the ideology that human brain is itself the mechanism for dealing with the information which is the mixture of different components, such as figures, sounds and words, etc., we believe firmly that the methods based on neural science and cognitive science will be the proper methods for dealing with big data. So in this subsection, the methods for mathematical statistics are discussed for big data based on neural science [6, 7] and cognitive science [8].

From the basic facts of neural science, information processing depends not only on the anatomical substrates of synaptic circuits but also on the electrophysiological properties of neurons or on their dynamical properties. Even if two neurons in different regions of the nervous system possess features which are the identical morphological ones, they may respond to the same synaptic input in very different manners because of each cell’s intrinsic properties or bifurcation dynamics.

An outstanding open question of how neural coding supports Bayesian inference includes how sensory cues are optimally integrated over time. Bayesian statistical inference describes how sensory and prior information can be combined optimally to guide behavior. The conclusions in [7] address what neural response properties allow a neural system to perform Bayesian prediction, i.e., predicting where a source will be in the near future given sensory information and prior assumptions. The work shows that the population vector decoder will perform Bayesian prediction when the receptive fields of the neurons encode the target dynamics with shifting receptive fields.

Meanwhile, from the viewpoints of cognitive science [8], capturing nature's statistical structure in behavioral responses is at the core of the ability to function adaptively in the environment. One highly active area of cognitive modeling is concerned with the question of how we learn to categorize perceptual objects.

It is based on the results stated above that the more general statistical method for big data based on neural engineering or cognitive science will be developed in the future. And the discovery for the corresponding statistical method will be a completely original one, and it will be a prospective field not only for mathematical statistics but also for artificial intelligence.

3.2 Statistical inference for big data with a whole sample

In this subsection, we discuss the possible method for dealing with the data with a whole sample.

Generally speaking, in mathematics, a dynamical system is a system in which the evolution with time for a point in a geometrical space can be described by a function, for example, the mathematical models that describe the flow of water in a pipe, the swinging of a clock pendulum, and the number of fish each spring time in a lake [9].

The concept of a dynamical system has its origins in Newtonian mechanics. There, as in other natural sciences and engineering disciplines, the evolution rule of dynamical systems is an implicit relation that gives the state of the system for only a short time into the future. The relation to describe the evolution rule is either a differential equation, difference equation or other time series.

It is the fundamental characteristics of dynamical systems that we can use them as the tools for describing the revolution rule of certain kinds of big data with a whole sample, especially with the combination of the internal evolutive features of the data. The method of dynamical system will be irreplaceable for data with whole sample.

At the same time, the methods of dynamical systems will also be the method for the compression of big data. That is to say, we may use a simple dynamical model to describe some complex evolution process, and that will also release a huge amount of memory space.

3.3 The modelling of time series based on internal model principle

The drawback of conventional statistical methods for time series is that its model is mainly based the factors that we observed. And from the control theory of dynamical systems, such a model is just the reflection of the factors that are not only observability but also controllability. On the other hand, with the increase of data, it will make the discovery possible for finding more relations among the internal factors. That will make the improvement of modelling possible with the help of internal model principle. It is based on the considerations stated above that, in [10], a novel modelling method for time series was proposed. Because the influences of the internal or external factors are often inevitable which result in the irregular change of the series, especially in many of the financial time series, the conventional modelling methods, which depend on the autocorrelation of the series, cannot meet the requirements or have a passive response to the factors. To find an effective method which can reflect the intrinsic essence or external influences of the series, the internal model principle, which is used in control system synthesis for revealing the features of the internal factors, was considered.

It may open up broad prospects for improving the fitness or prediction ability of the models under the circumstances mentioned above. Especially, with the increasing or accumulating of data in certain fields, and with the more internal relations are explored, the method will find more and more applications in the future.

4. Statistical inferences for figures, sounds, words and something like that

In this section, we will propose some of the possible improvements of statistical inferences, such as inferences based on figures, sounds, words and something like that.

4.1 The necessary for the renovation of statistical inference

With the development of information technologies, the necessaries for statistical inferences for dealing with big data are becoming increasing constantly. With the prevailing of big data, the contents of information are the mixture or the combination of figures or graphs, sounds, words, etc. The data presentation methods used most in meteorology by using visualization is an example.

Another field for statistical inference based on figures or graphs is the airborne drones, which use geometry to make decisions on how to get to a final destination [11]. At the same time, robotics and autonomous systems represent key areas of scientific and engineering endeavour now and in the future, see [11] and the references therein.

Based on the requirements stated above, we can conclude that the revolution for mathematical statistics on statistical inferences for big data will have a fundamental influence to the development of mathematical statistics. We will discuss the related fields in the following subsections.

4.2 Statistical inference for figures

The statistic inference for figures and/or graphs is a very important area. As a general data structure, graphs have become increasingly important in modeling sophisticated structures and their interactions, with broad applications including chemical informatics, bioinformatics, computer vision, video indexing, text retrieval, and Web analysis [12]. Mining frequent subgraph patterns for further

characterization, discrimination, classification, and cluster analysis becomes an important task. Moreover, graphs that link many nodes together may form different kinds of networks, such as tele-communication networks, computer networks, biological networks, and Web and social community networks [12]. Furthermore, in a relational database, objects are semantically linked across multiple relations. Mining in a relational database often requires mining across multiple interconnected relations, which is similar to mining in connected graphs or networks. Such kind of mining across data relations is considered multi relational data mining [13].

So, in the near future, the proper methods related with the problems of statistical inferences on graphs or figures must be extended, especially the method based on extraction the features for figures or graphs will be developed. It also means that the methods of mathematical statistics based on the visualization will be furtherly developed.

4.3 Statistical inference for sounds

The second important area needed to be explored is the statistical inference for sounds. There are still a lot of examples for the exploration. As a matter of fact, robust sound source localization is performed by the human auditory system even in challenging acoustic conditions, complex scenarios. In [14], a computational binaural localization model is proposed that possesses mechanisms for handling of corrupted or unreliable localization cues and generalization across different acoustic situations. Bayesian computation of the direction-of-arrival probability map naturally leads to coherence-weighted integration of location cues across frequency and time.

From the example, we know that the methods on statistical inference for sounds will also need to be developed in the future. It is not only for the development of mathematical statistics itself but also for the development of artificial intelligence or robotics.

4.4 Statistical inference for words

The third important area in statistical inference for big data is the inference based on words, or on semantics and syntax, respectively. Language has syntax, which controls how it is broken up, and can be expressed in corpora (samples) of 'real world' text. There have been certain related scientific results for these, such as, the logical computation or the computation for events in probability theory.

A former important consideration in this field is the one that concerns about computing with words (CWW), termed sometimes also in a longer form as computing with words and perceptions (CWP), originated by Zadeh in the early or mid-1990s, which has rapidly attracted much attention from scholars and researchers from around the world. As a result, a two-volume book on CWW edited by Zadeh and his coauthor was edited in 1990 and appeared in the present book series [15]. CWW is a real breakthrough and will have a lasting impact on many areas in which systems and processes should be modeled. People are to operate in a human centric context in the sense that human specific features are crucial, among them the human willingness and propensity to use natural language.

In recent years there are still improvements for such problems, for example, a novel approach to linguistic mutual inference was described in the paper [16]. As being considered in [16], the inference can be established based on two modules, the motion language model and the natural language model.

From the viewpoint on the improvement of statistical inference, maybe in a near future, an improvement algorithm as σ-algebra on events or the method based on the extraction of features for words will be proposed.

In author's opinion, with the development of computer science and robotics' technology, such areas should be reformed extensively.

5. Conclusion

In this paper, we consider the possible areas that mathematical statistics should be reformed or developed in the future, especially the analysis methods for the whole sample, and statistical inferences for figures, graphs, sounds and words, and so on. The intention of the paper is to serve as a modest spur to induce someone to come forward with his valuable contributions. The future education on mathematical statistics should also pay certain attention on such areas.

References

[1] Mark J. Schervish, Theory of statistics (Corr. 2nd print. ed.). New York: Springer. 1995.

[2] F. Liese, and Klaus-J. Miescke, Statistical Decision Theory: Estimation, Testing, and Selection. Springer. 2008.

[3] Sankar K. Pal, Saroj K. Meher and A. Skowron, Data science, big data and granular mining, Pattern Recognition Letters, vol. 67, pp. 109-112, 2015.

[4] Cheikh K. Emani, N. Cullot, and C. Nicolle, Understandable big data: a survey, Computer Science Review, vol. 17, pp. 70-81, 2015.

[5] Wei Wang, The contents of some mathematical courses need to be improved: a viewpoint for education reformation, Advances in social science, education and humanities, vol. 28, pp. 382-386, 2015.

[6] C. Eliasmith and Charles H. Anderson, Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems, Cambridge, MA: MIT Press, 2003.

[7] D. Rich, F. Cazettes, Y.Y. Wang, J.L. Pena, and B.J. Fischer, Neural representation of probabilities for Bayesian inference, Journal of Computional Neuroscience, vol. 38, pp. 315-323, Apr. 2015.

[8] Robert E. Wray and Ronald S. Chong, Comparing cognitive models and human behavior representations: Computational tools for expressing human behavior, in Collection of Technical Papers - InfoTech at Aerospace: Advancing Contemporary Aerospace Technologies and Their Integration, vol 1, pp. 608-620, 2005.

[9] S. Lynch, Dynamical Systems with Applications using Maple, 2nd Ed. Springer. 2010.

[10] Wei Wang, The novel modelling method of time series based on the internal model principle, Proc. of International Conference on Artificial Intelligence and Industrial Engineering (AIIE 2015), Advances in Intelligent Systems Research, vol. 123, pp. 573-576, 2015.

[11] The Royal Society, Robotics and autonomous systems – vision challenges and actions, Part of the conference series breakthrough science and technologies transforming our future, On 13 November 2015.

[12] S. Ceri, Emanuele D. Valle, D. Pedreschi, and R. Trasarti, Mega- modeling for big data analytics, conceptual modeling, in Proc. of 31st International Conference, ER 2012, Florence, Italy, October 15-18, 2012, Springer-Verlag Berlin Heidelberg, 2012.

[13] S. Modafferi, A. Chakravarthy, and Z. Sabeur, Multi-level data fusion of environmental data in future internet applications, in Proc. of 21st Italian Symposium on Advanced Database Systems, SEBD 2013, Roccella Jonica, Reggio Calabria, Italy, pp.297-304.

[14] H. Kayser, V. Hohmann, Stephan D. Ewert, B. Kollmeier, and J. Anemuller, Robust auditory localization using probabilistic inference and coherence-based weighting of interaural cues, Journal of the Acoustical Society of America, vol. 138, pp. 2635-2648, 2015.

[15] Lotfi A. Zadeh, Computing with Words: Principal Concepts and Ideas, New York: Springer, 2012.

[16] W. Takano, and Y. Nakamura, Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions, International Journal of Robotics Research, vol. 34, pp. 1314-1328, 2015.