Special feature: data science—present and future

Manabu Iwasaki

Published online: 15 October 2022
© The Author(s) under exclusive licence to Japanese Federation of Statistical Science Associations 2022

This is the second special feature on data science in the *Japanese Journal of Statistics and Data Science (JJSD)*. The first special feature "Perspectives on data science for advanced statistics" was appeared in the very first issue of *JJSD*. In the preface of the special feature, quoting the Wikipedia definition of data science as 'to extract knowledge or insights from data', the organizer Aoshima wrote that such a definition of data science is the definition of statistics that he has been using for 30 years in his elementary statistics courses (Aoshima, 2018).

If data science is not statistics itself, what is the difference between them? As a statistician, perhaps I should be called a classical statistician. In recent years, I have defined data science in my writing and frequent talks, as

\[(\text{statistics} + \text{computer science}) \times \text{social application}\]

Specifically, I have argued that the theoretical foundations of data science lie in statistics and information science. However, I stressed that this alone is not sufficient, and it is quite important to apply this formula to various problems in the real world to obtain rational and useful solutions. In this definition, it is worth noting that (statistics + information science) and social application are multiplied. And so, if one of the two terms is zero, the whole will be zero! This means that theoretical studies in statistics and information science alone are insufficient, and that it is useless to simply seek ad hoc solutions that do not have a theoretical basis for individual real-world problems.

In recent years, it is often said the "data everywhere", and hence the scope of data science has become more and more diverse. In this special feature, nine papers are collected papers that deal with various subjects in a variety of fields, including computer science, economics, medicine, sports, and so forth. In addition to such research, data science education is another extremely important topic. Recent trends show that data science education should not be limited to classroom lectures, but should also include a variety of activities with appropriate exercises and practical training using real data. This kind of educational methodology is of course one of the research fields of data science.

---

1 Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

\[\text{Springer}\]
In the past few years, especially after the first special feature mentioned above (Aoshima, 2018), new departments and/or courses with the name of “data science” have been established in Japan, which is not only a Japanese trend but also a global one. In addition, many data science education and training courses for working people are being developed. It is expected that a special feature on data science will be organized in the next few years, in which a wider variety of papers will be collected.

The following is a brief introduction of the papers collected in this special feature in the order of acceptance.

The first paper is “Converting ADMM to a proximal gradient for efficient sparse estimation” by Shimmura and Suzuki, which discusses sparse estimation methodology, a topic that is often discussed especially when dealing with recent large-scale data. This paper proposes a general method for converting the alternating direction method of multipliers (ADMM) solution to the proximal gradient method. Then, they apply it to sparse estimation problems, such as sparse convex clustering and trend filtering. It is shown by a numerical study that their method shows a significant improvement in terms of efficiency.

The second paper is “Shiga University’s endeavor to promote human resources development for data science in Japan” by Tanaka, Himeno and Fueda, all from Shiga University. In 2017, Shiga University established the Department of Data Science as the first department in Japan specializing in data science and statistics. This paper reports the historical context, curricula, and collaboration with industry and other universities. The career paths of the graduates and the massive open online courses and textbooks provided by the department are also summarized.

The third paper “Experience of distance education for project-based learning in data science” by Sakamaki, Taguri, Nishiuchi, Akimoto and Koizumi also deal with education in data science. They stress the importance of a project-based learning (PBL) approach in education because students can learn data science practices based on real-world problems and real-world data. In this paper, the authors explain how they developed and conducted the PBL approach and also provide survey results from students. Under the unfortunate COVID-19 pandemic, it might be good news that the skill of distance learning improved effectively among faculty members.

The fourth paper “Real world data and data science in medical research: present and future” by Togo and Yonemoto, who are both from a pharmaceutical company. They discuss real world data (RWD). This article summarizes the overview and challenges of RWD analysis in medical fields from methodological perspectives. The authors first argue that data sources of RWD should be comprehended. Then they point out that the selection of appropriate statistical and epidemiological methods is highly critical for RWD analysis, compared to those for randomized clinical trials. Lastly, they discuss the future of RWD in terms of overcoming limited data by proxy confounders, using unstructured text data, and the linking of multiple databases.

Next, Takahashi discusses “A new robust ratio estimator by modified Cook’s distance for missing data imputation”, which deals with missing data inevitable in recent data science with large datasets. The author proposes a new robust ratio estimator, named the TC-ratio estimator (ratio estimator with trimming based on Cook’s distance), which is robust against outliers on the vertical axis (variable $y$), on the horizontal
axis (variable $x$), and on both axes ($x$ and $y$), for missing data imputation. He also suggests a novel way to automatically determine the number of outliers. To assess the performance of the new method, Monte Carlo simulations are conducted under 160 different data generation processes, each repeated in 10,000 simulation runs. The relative superiority of the new method is shown against the traditional robust ratio imputation methods.

Kunitomo and Sato, both econometricians, discuss “Local SIML estimation of some Brownian and jump functionals under market micro-structure noise”. To estimate Brownian and jump functionals from high-frequency financial data under market microstructure noise, they introduce a new local estimation method of the integrated volatility and higher-order variation of Ito’s semi-martingale processes. In this study, they develop the local SIML (LSIML) method, which is an extension of the existing separating information maximum likelihood (SIML) method. The new LSIML method is simple, and it is shown that the LSIML estimator has some desirable asymptotic properties and reasonable finite sample properties.

The paper “R package DCchoice for dichotomous choice contingent valuation: a contribution to open scientific software and its impact” by Aizaki, Nakatani, Sato and Fogarty is rather computational. They developed an R package “DCchoice” that is designed to mitigate programing-related barriers to the application of dichotomous choice contingent valuation (DCCV) methods in empirical studies. This paper introduces the current version of DCchoice which supports single-, one-and-one-half-, and double-bounded DCCVs, with and without a spike. Additionally, the willingness-to-pay and its confidence intervals can be calculated for a representative respondent as well as for a user-defined specific respondent using the current version.

“Analysis and visualization of team performances of football games” by Obata and Izumi deals with sports data. It is well known that the number of goals in a football game approximately follows a Poisson distribution. Since teams belonging to a professional football league have many games through a season, the performances of such teams should be considered not to be constant throughout a season. In this paper, a Poisson regression model with varying coefficients is proposed to analyze and visualize the time-varying performance indices of football teams. The estimated performance up to the middle of a season can be used to estimate the final performance of the teams. The proposed method is demonstrated with real data of the Japanese professional football league.

The final paper “measuring global flow of funds: who-to-whom matrix and financial network” by Nan Zhang deals with global flow of funds (GFF) statistics to measure global financial stability at the national and cross-border sectoral levels. After investigating the data sources and rebuilding the statistical framework to establish the GFF statistical matrix of G20 countries, the author evaluates their financial risks and influences. The analytical results systematically show the financial relationships among G20 countries in the GFF, the characteristics of overseas investment among China, Japan, and the United States, and external shocks and internal influences.
Reference

Aoshima, M. (2018). Special feature: Perspectives on data science for advanced statistics. *Japanese Journal of Statistics and Data Science, 1*, 39–40.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.