An Introduction to the Teaching Reform of Data Science Courses in the Context of Big Data

Du Shang[1,*], Shuai Su[1]

[1] School of Automation and Intelligence, Beijing Jiaotong University, Beijing, China

Abstract. In the big data age, it's imperative for data science courses to undergo transformation and embrace innovation. Building upon the foundation of existing data science education, we should introduce a big data-informed perspective for curriculum reform. This reform should encompass specific strategies and measures across various dimensions, including curriculum content, pedagogical approaches, theoretical instruction, and practical application scenarios. The goal of this curriculum overhaul is to foster an integration of modern statistical techniques with big data analytics, balancing theoretical knowledge with hands-on experience. It aims to develop students' statistical acumen and their ability to manage complex datasets within the big data landscape. By doing so, it seeks to elevate students' overall proficiency and empower them with the skills to apply contemporary mathematical and statistical methods to tackle real-world challenges effectively.

1 Introduction

Big data encompasses datasets so vast, intense, and intricate that they surpass the processing and analytical capabilities of conventional software tools [1]. The study of big data dates back to the mid-1990s. It arises from countless online interactions among individuals, transactions between people and systems, and machines equipped with sensors. Internet search engines produce billions of data points daily [2]. The scale of data generated today is measured not in gigabytes (GB) or terabytes (TB), but in zettabytes (ZB), expanding at a rate of 40% daily. In big data analytics, the 3V model is widely recognized: Volume, Velocity, and Variety [3]. Variety introduces non-traditional and unstructured data forms, such as social media sentiments and internet map usage, demanding innovative approaches to understanding data structures and formulating insightful research queries. The sheer volume and speed of data can pose challenges such as scalability and storage constraints, accumulation of noise, false correlations, stochastic endogeneity, and measurement errors, which in turn raise concerns over the computational viability and stability of algorithms.

The challenges of big data transcend traditional realms of mathematics and statistics, necessitating interdisciplinary solutions that combine domain expertise, computational proficiency, and statistical acumen. While mathematics and statistics have significantly contributed to the theoretical frameworks and practical applications of big data, science and engineering students must expand their skills to address the less structured or ambiguous real-world problems posed by big data [4]. This includes providing structure to ill-defined problems or devising new models for emerging data types like images or networks. It has been suggested that to excel in the field of big data or one's specific data science domain, one must engage with practical problems, with relevant methodologies and theories emerging as a natural consequence.

Although the perception of big data in media or business may differ from that of academic scientists, as a field of study, it requires the seamless integration of classical statistical methods, traditional mathematical theories, and contemporary big data technologies [5]. This integration is essential to meet the challenges of the big data era, enhance the statistical quality and depth of complex data, reduce statistical costs, and infuse mathematical and statistical theories with new opportunities and vigour. Big data broadens the knowledge base of mathematical and statistical disciplines and elevates the significance and stature of data science within the social and natural sciences. It represents a significant new opportunity and source of vitality for traditional mathematics and classical statistics.

2 Challenges faced by data science courses in the context of big data

The advent of big data has bestowed upon contemporary society a trove of unique opportunities, even as it confronts data scientists with formidable challenges. Grasping the unique attributes of big data is essential, given its propensity to instigate revolutionary changes in statistical and computational techniques, as well as in the underlying computing infrastructure. Data science curricula must introduce innovative approaches to big data analysis and computation. While big data holds the promise of revealing insights impossible with smaller datasets, its vast samples and high dimensionality introduce computational and statistical hurdles such as scalability, storage issues, noise, spurious correlations,

* Corresponding author: dushang@bjtu.edu.cn

and measurement errors [6]. These challenges demand the creation of new computational and statistical frameworks, particularly focusing on the feasibility of sparse solutions within high-confidence parameters. The intrinsic endogeneity present in big data techniques may result in dubious statistical deductions and, consequently, flawed conclusions.

The analysis of big data, marked by extensive dimensionality and substantial sample volumes, encounters three primary difficulties: (a) the elevated dimensionality results in the buildup of noise, false correlations, and uniformity; (b) when high dimensionality is paired with large datasets, it leads to increased computational expenses and the instability of algorithms; (c) the aggregation of massive samples from various sources and times introduces heterogeneity, experimental variation, and statistical bias, necessitating more adaptive and robust analytical programs [7].

The vast and multidimensional nature of big data brings considerable computational challenges and the need for transformation in large-scale optimization techniques. The straightforward use of penalized likelihood estimation on high-dimensional datasets entails tackling large-scale optimization issues, which are typically expensive, slow, and prone to numerical convergence instability. There is a critical need for scalable solutions to large-scale non-smooth optimization processes [8]. Additionally, industries such as genomics, neuroinformatics, marketing, and online social media, which handle big data samples ranging into the millions or billions, face significant computational demands for data management and querying. This situation necessitates the adoption of parallel computing, stochastic algorithms, approximation methods, and streamlined implementations. When developing statistical programs, it is essential to meticulously consider the scalability of statistical methods to accommodate high-dimensional and large-sample datasets.

To overcome the challenges of big data, data science curricula must integrate advanced statistical perspectives and computational techniques. Conventional approaches that are effective for moderate-sized datasets frequently fail to scale effectively with the vastness of big data. Likewise, statistical techniques that work well with low-dimensional data face considerable obstacles when applied to high-dimensional data. In crafting robust statistical programs aimed at exploring and predicting big data, it is imperative to address challenges such as heterogeneity, noise, spurious correlations, and incidental endogeneity, all while maintaining a balance between statistical precision and computational efficiency.

When crafting data science courses, it's essential to integrate content that fosters the development of computational methods and statistical techniques capable of scaling up to big data challenges. Achieving a balance between statistical accuracy and computational efficiency necessitates a holistic strategy that addresses big data management from both a statistical and computational standpoint.

3 Reflections on the reform of data science curriculum in the context of Big Data

3.1. A deep understanding of modern numerical computing techniques is necessary to meet the challenges of big data.

In practical applications, the foundation of Big Data analytics is built upon parallel and distributed computing, which drives the development of modern Big Data algorithms, software, and system architectures. Across various systems, an array of platforms have become popular, such as clusters, cloud computing, multi-core processors, and notably, graphics processing units (GPUs). The use of parallel and distributed databases, NoSQL databases for managing non-relational data like graphs and documents, and data flow management systems is widespread and covers numerous applications.

Regarding computational efficiency, Big Data has catalysed the evolution of novel computing infrastructures and data storage methodologies. The optimization process is pivotal in Big Data analytics, serving as a tool rather than an end goal, thus facilitating the creation of swift algorithms capable of managing high-dimensional, vast datasets.

Ensuring data quality is of utmost importance, emphasizing the need for stringent quality control measures, standardization processes, tracing the origins of data, and annotating metadata. From a computational standpoint, there is an ongoing effort to develop and put into practice new performance benchmarks. In terms of algorithms, there is a notable upsurge in the popularity of machine learning approaches such as deep learning and reinforcement learning. Additionally, areas like computational learning theory and differential privacy are poised to gain significant advantages from the statistical foundations laid by Big Data.

In the realm of applied statistics, analysts face the challenge of computational bottlenecks when dealing with Big Data. They must strike a balance between the demands for accuracy and the constraints of computation time, ensuring that the statistical methods employed are both precise and computationally feasible.

3.2 In-depth knowledge of the latest developments in statistical methods and software is necessary to meet the challenges of big data.

Big data is universally recognized for its vast volume, intensity, and complexity, which outstrip the capacity of traditional analytical tools. This phenomenon presents statisticians with both opportunities and challenges. It's essential to grasp the significance of big data analytics in scientific discovery and to stay updated on the latest advancements in statistical methods and software to tackle the complexities of big data. Key methodologies include subsampling-based approaches, divide-and-conquer strategies, and online updating methods for streaming data. Online updating methods, a novel

technique in big data, have been expanded for variable selection and standardization, with their efficacy evaluated through simulation studies and models for streaming data.

Software packages should incorporate state-of-the-art tools that overcome limitations in computer memory and computational power. Demonstrating these tools through case studies, such as logistic regression analysis of flight delay data, can be instructive. Clustering methods are crucial for tackling the challenge of dimensionality reduction in big data. Principal Component Analysis (PCA) is a widely used technique that employs a subset of key components to represent the majority of the variation within complex, high-dimensional data sets. Nonetheless, this traditional method can still face computational hurdles when dealing with the vast scope of big data.

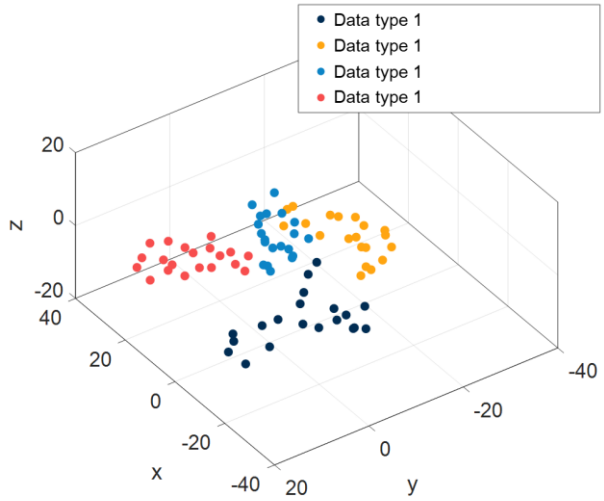

Figure.1 A demonstration of the demensional reduction of the Multidimensional Scaling method.

Linear and nonlinear Multidimensional Scaling (MDS) methods, which depend on the full distance matrix among all data points in high-dimensional space to generate a lower-dimensional model that preserves the proximity of data points, similarly confront significant computational obstacles in the realm of big data. A novel spectral MDS technique enhances the aggregation of global neighbourhood structures by focusing on more intriguing local neighbourhoods. Random projection, based on the Johnson-Linden Strauss lemma, is another method for preserving positions. It guarantees that data points, even in very high dimensions, can be efficiently mapped into a lower-dimensional space while approximately maintaining the initial relationships among the points, which is illustrated in Fig.1.

3.3 The problem of downscaling and clustering of high-dimensional data in big data must be addressed.

Clustering is indeed the most prevalent method for depicting large datasets. It's an unsupervised process that groups similar data points to minimize the sum of distances within clusters and maximize the distances between them. Cluster representations, such as k-means from classical clustering, offer a simplified and lucid view of dataset structures. Even with a multitude of points and noise, strategies have been crafted to refine traditional clustering methods for large datasets. These strategies tackle challenges such as the necessity for iterative calculations that involve a massive $O(n^2)$ distance matrix, or the prerequisite of having the entire dataset readily available for such computations.

In the realm of high-dimensional data analysis, the reduction of data scale and selective variable choice are essential in combating the buildup of noise. Within the context of high-dimensional classification, conventional methods that utilize all available features might not outperform random chance due to the interference caused by noise, where a demonstration of the noise affecting data need analysis is shown in Fig.2. This has spurred the development of new regularization methods and deterministic independent screening. High dimensionality can also lead to spurious correlations, causing false statistical inferences and misleading scientific conclusions. Additionally, it can result in collateral endogeneity, where many uncorrelated covariates might incidentally correlate with residual noise.

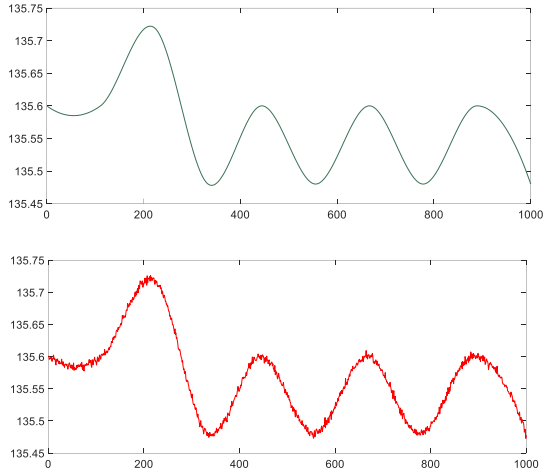

Figure.2 A demonstration of the data under analysis is affected by noise.

In conclusion, in the era of rapid technological advancements and the prevalence of big data, data science course reforms are of utmost significance and should be meticulously designed within the comprehensive and dynamic big data context. These reforms ought to be firmly grounded in and build upon the existing teachings, taking into account the knowledge and skills that have already been imparted to students. The reform initiative should not only identify but also propose highly specific and practical paths and measures in multiple crucial aspects. In terms of content, it should be updated and refined to incorporate the latest trends and essential elements of big data analytics. Regarding teaching techniques, innovative and interactive methods need to be introduced to engage students more effectively and enhance their learning experience. For theoretical methods, a more in-depth and integrated approach should be adopted to equip students with a solid foundation. When it comes to application cases, a

wide range of real-world and industry-relevant examples should be presented to help students understand the practical implications and applications of data science. In the area of faculty development, continuous training and professional growth opportunities should be provided to ensure that instructors are well-versed in the latest advancements and can guide students proficiently. The ultimate and overarching goal of this reform is to systematically foster students' statistical thinking and their remarkable ability to process complex data. By doing so, it aims to enhance the overall research literacy of graduates, enabling them to make significant contributions in the fields of data analysis, research, and various industries that rely heavily on data-driven decision-making. This will not only benefit the individual students in their future careers but also have a profound and positive impact on the entire academic and professional community related to data science.

4 Concluding remarks

There is no doubt that big data analytics, a central and frequently debated subject in the realm of data science, has seen a rapid surge in data creation across various sectors and disciplines. The volume and velocity of data being generated have been truly staggering, with industries such as finance, healthcare, and e-commerce amassing vast amounts of information at an unprecedented rate. The coming times are expected to bring a collective effort to tackle the complex challenges posed by Big Data, with researchers potentially working together to understand the risks, advantages, and compromises involved. This collaborative endeavour might result in the rise of a new breed of data scientists who are proficient in both theoretical understanding and practical abilities, providing optimism for tackling substantial obstacles. These data scientists would need to possess a deep knowledge of statistical analysis, machine learning algorithms, and data visualization techniques, among others. The onset of the big data era necessitates a more vigorous development of top-tier talent. As the demand for data-driven insights and solutions continues to soar, the need for highly skilled professionals becomes even more pronounced. At the same time, data science programs in higher education must keep pace with the evolving landscape. It is crucial to refine curriculum design to match the distinctive features of big data. This could involve incorporating more hands-on projects, real-world case studies, and interdisciplinary courses. In the future, there is a need to strengthen the study of all state-of-the-art big data techniques available to us and to persistently innovate and advance sophisticated data analysis methods and theories. The ultimate goal is to stimulate national economic growth and meet the diverse requirements across a broad spectrum of scientific and technological fields. By leveraging the power of big data, businesses can optimize their operations, governments can make more informed decisions, and scientific research can reach new heights.

References

1. Y. Wu, et al, 33rd IEEE International Conference on Computer Communications and Networks. 10637520 (2024).
2. Y. Himeur et al, Artif. Intell. Rev. 56, 6 (2023).
3. Z. Geng, Stat. Res. 31, 1 (2014).
4. D. Qiu, Stat. Res. 31, 1 (2014).
5. K. Crawford, HBR Blog Network 1, (2013).
6. S. Ma, Stat. Res. 34, 1, (2017).
7. S. Tian, Z. Zhang, X. Xie, C. Yu, Adv. Prod. Eng. Manag. 17, 3 (2022).
8. I. Arshad, SH. Alsamhi, W. Afzal, CMC-Comp. Mater. & Con. 74, 2 (2023).