Statistics and Probability Letters 136 (2018) 1–3

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

The role of Statistics in the era of Big Data

Big data have nowadays pervaded most domains of research, business and administration. What is the role of Statistics in this big data era?

This special issue gathers the opinions of leading scientists from statistics, as well as from machine learning, computer science, applied mathematics and engineering. It has been stimulated by a plenary lecture by David Dunson - Duke University, followed by a discussion by Piercesare Secchi - Politecnico di Milano, at the 48th Meeting of the Italian Statistical Society (Salerno, Italy, June 2016). It collects a total of 35 papers, carrying a large spectrum of opinions and varied perspectives on the role of Statistics in the era of big data. While some of the opinions and views are shared by many authors, other views are in quite strong opposition.

Many authors highlight the importance and centrality of the statistical model, possibly backed up by theoretical knowledge of the phenomenon under study; similarly, various contributors recommend the definition of clear investigation hypotheses, suggested by expert knowledge of the problem; see, e.g., the discussions in Secchi (2018), Scott (2018), Wit (2018), Smirnova et al. (2018), Bühlmann and van de Geer (2018), Bowman (2018), Olhede and Wolfe (2018), Quarteroni (2018), Lau et al. (2018) and Shi (2018). The idea that "data, either small or big, do not speak on their own", that they must be "appropriately interrogated" to extract meaningful information, that "correlation is not enough", as many authors word it, is of course deeply rooted in statistical thinking. But several contributors also remind us that this view is not shared by the whole scientific community and has for instance been strongly questioned by eminent computer scientists, who claim that the advent of big data makes models and theory useless; see, e.g., the many citations reported in Torrecilla and Romo (2018), Scott (2018), Quarteroni (2018) and Ceri (2018).

The increasing complexity of data, and not simply their sheer size, is highlighted in a number of contributions. As noted for example by Dunson (2018), Secchi (2018), Torrecilla and Romo (2018), Scott (2018), Smirnova et al. (2018), Wit (2018), Bühlmann and van de Geer (2018), Olhede and Wolfe (2018), Chung (2018), Yu et al. (2018), Fassò et al. (2018), Cox et al. (2018), Dryden and Hodge (2018), Giraldo et al. (2018), Shi (2018), Vieu (2018) and Vantini (2018), today's data are in fact more and more frequently highly dimensional, that is, they involve a very large number of variables (not simply of statistical units). Data may as well be intrinsically infinite-dimensional; this is the case for instance of functional data (possibly displaying complex dependencies in space and/or time), multidimensional images and videos, manifold data and other complex data objects that lie outside of the classical Euclidean paradigm. The analysis of so complex and highly dimensional data calls for the development of novel methods and models, fueling some of the most fascinating and fastest growing fields of modern statistics and creating new avenues of frontier research.

Besides the specific and varied problems posed by the sheer volume of the data, several other critical issues are discussed. Big data are often affected by selection bias and other sampling problems; see, e.g., Dunson (2018), Olhede and Wolfe (2018), Lau et al. (2018), Wong (2018), Dryden and Hodge (2018), Bivand and Krivoruchko (2018), Sharples (2018), Cox et al. (2018), Faraway and Augustin (2018) and Bartolucci et al. (2018). They are typically heterogeneous, coming from different sources and providers; see, e.g., Secchi (2018), Torrecilla and Romo (2018), Scott (2018), Wit (2018), Bühlmann and van de Geer (2018), Olhede and Wolfe (2018), Yu et al. (2018), Bivand and Krivoruchko (2018) and Sharples (2018). In general, the quality of massive datasets, whose recording rarely follows an experimental design, may in fact be lower than the quality of smaller datasets, especially when the latter are specifically collected with the purpose of investigating a given phenomenon, following a precise sampling scheme; see, e.g., Sharples (2018), Scott (2018), Cox et al. (2018), Faraway and Augustin (2018), Bühlmann and van de Geer (2018), Meng (2018), Bivand and Krivoruchko (2018) and Shi (2018). All these issues must be suitably taken into account and require the development of appropriate data analysis techniques.

To face the many challenges posed by the analysis of big and complex data, many of the authors, both coming from statistics, as well as from applied mathematics and computer science, feel the need for stronger collaborations among their

https://doi.org/10.1016/j.spl.2018.04.009

respective disciplines; see, e.g., Secchi (2018), Torrecilla and Romo (2018), Quarteroni (2018), Ceri (2018) and Wong (2018). Moreover, as emphasized by various contributors, it is crucial to build serious interdisciplinary collaborations with scientists who are expert in the phenomenon being investigated; see, e.g., Dunson (2018), Smirnova et al. (2018), Scott (2018), Meng (2018) and Shi (2018).

The applied fields considered by the authors of this special issue are varied, ranging from life sciences, to environmental sciences, physical sciences, engineering, business and administration. For instance, Smirnova et al. (2018) and Wit (2018) discuss the impact of big data on biostatistics, with examples concerning genomic data and accelerometry data. Genomic data are as well considered by Ceri (2018), Wong (2018) and Yu et al. (2018). Shi (2018) considers sensor data that record human movements. Cox et al. (2018) and Sharples (2018) bring the example of massive electronic health records. Chung (2018) and Dunson (2018) focus on the challenges posed by brain imaging data, and Yu et al. (2018) consider the integration of brain imaging data with genetic data, to study the influence of genetics on brain connectivity. Quarteroni (2018) describes an example in cardiovascular research. Scott (2018), Castruccio and Genton (2018), Bivand and Krivoruchko (2018), Fassò et al. (2018), Gupta et al. (2018) consider massive environmental and climate spatio-temporal data. Meng (2018) writes about the analysis of astronomical data. Lau et al. (2018) discuss an application to the analysis of sensor data used in structural engineering. Dryden and Hodge (2018) write about a large project involving the analysis of massive transport data. James (2018) and Giudici (2018) outline some aspects of big data in business and finance. Azzone (2018) presents the increasing impact of big data on the design of public policies.

Besides the techniques and models discussed by the authors in relation to specific applications, various other approaches are recommended to address the analysis of big an complex data. On the computational side, Bierkens et al. (2018) propose a novel MCMC algorithm that is particularly well suited to do bayesian inference in the contest of big datasets. Vantini (2018) and Vieu (2018) advocate the use of non-parametric and semi-parametric procedures for the analysis of functional and complex data. Quarteroni (2018) discusses how computational models, based on first principles from physics, can efficiently be coupled with real data, leading to major improvements in many applied fields. Olhede and Wolfe (2018) outline some of the challenges in network analysis. Bartolucci et al. (2018) describe the role of latent variable models for the analysis of big data. Secchi (2018), Torrecilla and Romo (2018), Bowman (2018), Scott (2018) and Reid (2018) note that the exploration and the analysis of big and complex data demands the development of innovative data visualization tools.

Finally, as highlighted by, e.g., Secchi (2018), Reid (2018), Olhede and Wolfe (2018), Meng (2018), Cao (2018) and James (2018), a crucial aspect to be considered, to defend the centrality of the role of Statistics in the big data era, and to avoid missing the data science boat, is the education of the next generations of statisticians. Since some years, many universities and institutions have started defining new undergraduate and graduate programs, that give students not only a strong statistical background, but also the competencies needed to efficiently interact with fellows mathematicians, computer scientists and domain experts, thus enabling those collaborations that are essentials for advancing our ability to analyze big data.

This issue offers a window on a debate that will be alive over the years, changing the ways and methods, and possibly the very core, of the disciplines that we now associate with the analysis of data. A debate that will culturally enrich us all, whether or not we will reach a consensus, and will greatly influence the scientific and technological developments of the years to come.

Let me conclude by heartily thanking the many authors who contributed to this special issue, as well as the reviewers, whose comments further helped to shape these articles. I am grateful to the Journal Manager and Managing Editor, for their precious assistance. Finally, I would like to thank the Journal Publishers, who warmly supported this initiative.

References

Azzone, G., 2018. Big data and public policies: opportunities and challenges. Statist. Probab. Lett. 136, 116–120. Special Issue on “The role of Statistics in the era of Big Data”.

Bartolucci, F., Bacci, S., Mira, A., 2018. On the role of latent variable models in the era of big data. Statist. Probab. Lett. 136, 165–169. Special Issue on “The role of Statistics in the era of Big Data”.

Bierkens, J., Bouchard-Côté, A., Doucet, A., Duncan, A.B., Fearnhead, P., Lienart, T., Roberts, R., Vollmer, S.J., 2018. Piecewise deterministic Markov processes for scalable Monte Carlo on restricted domains. Statist. Probab. Lett. 136, 148–154. Special Issue on “The role of Statistics in the era of Big Data”.

Bivand, R., Krivoruchko, K., 2018. Big data sampling and spatial analysis: “which of the two ladles, of fig-wood or gold, is appropriate to the soup and the pot?”. Statist. Probab. Lett. 136, 87–91. Special Issue on “The role of Statistics in the era of Big Data”.

Bowman, A., 2018. Big questions, informative data, excellent science. Statist. Probab. Lett. 136, 34–36. Special Issue on “The role of Statistics in the era of Big Data”.

Bühlmann, P., van de Geer, S., 2018. Statistics for big data: A perspective. Statist. Probab. Lett. 136, 37–41. Special Issue on “The role of Statistics in the era of Big Data”.

Cao, J., 2018. Statisticians can do better in the big data era. Statist. Probab. Lett. 136, 146–147. Special Issue on “The role of Statistics in the era of Big Data”.

Castruccio, S., Genton, M., 2018. Principles for statistical inference on big spatio-temporal data from climate models. Statist. Probab. Lett. 136, 92–96. Special Issue on “The role of Statistics in the era of Big Data”.

Ceri, S., 2018. On the role of statistics in the era of big data: a computer science perspective. Statist. Probab. Lett. 136, 68–72. Special Issue on “The role of Statistics in the era of Big Data”.

Chung, M., 2018. Statistical challenges of big brain network data. Statist. Probab. Lett. 136, 78–82. Special Issue on “The role of Statistics in the era of Big Data”.

Cox, D., Kartsonaki, C., Keogh, R.H., 2018. Big data: some statistical issues. Statist. Probab. Lett. 136, 111–115. Special Issue on “The role of Statistics in the era of Big Data”.

Dryden, I.L., Hodge, D.J., 2018. Journeys in big data statistics. Statist. Probab. Lett. 136, 121–125. Special Issue on "The role of Statistics in the era of Big Data".
Dunson, D., 2018. Statistics in the big data era: Failures of the machine. Statist. Probab. Lett. 136, 4–9. Special Issue on "The role of Statistics in the era of Big Data".
Faraway, J.J., Augustin, N., 2018. When small data beats big data. Statist. Probab. Lett. 136, 142–145. Special Issue on "The role of Statistics in the era of Big Data".
Fassò, A., Finazzi, F., Madonna, F., 2018. Statistical issues in radiosonde observation of atmospheric temperature and humidity profiles. Statist. Probab. Lett. 136, 97–100. Special Issue on "The role of Statistics in the era of Big Data".
Giraldo, R., Dabo-Niang, S., Martínez, S., 2018. Statistical modeling of spatial big data: an approach from a functional data analysis perspective. Statist. Probab. Lett. 136, 126–129. Special Issue on "The role of Statistics in the era of Big Data".
Giudici, P.S., 2018. Financial data science. Statist. Probab. Lett. 136, 160–164. Special Issue on "The role of Statistics in the era of Big Data".
Gupta, S., Degbelo, A., Mateu, J., Pebesma, E., 2018. Quality of life, big data and the power of statistics. Statist. Probab. Lett. 136, 101–104. Special Issue on "The role of Statistics in the era of Big Data".
James, G., 2018. Statistics within business in the era of big data. Statist. Probab. Lett. 136, 155–159. Special Issue on "The role of Statistics in the era of Big Data".
Lau, F.D.-H., Adams, N.M., Girolami, M.A., Butler, L.J., Elshafie, M.Z., 2018. The role of statistics in data-centric engineering. Statist. Probab. Lett. 136, 58–62. Special Issue on "The role of Statistics in the era of Big Data".
Meng, X., 2018. Conducting highly principled data science: A statistician's job and joy. Statist. Probab. Lett. 136, 51–57. Special Issue on "The role of Statistics in the era of Big Data".
Olhede, S.C., Wolfe, P.J., 2018. The future of statistics and data science. Statist. Probab. Lett. 136, 46–50. Special Issue on "The role of Statistics in the era of Big Data".
Quarteroni, A., 2018. The role of statistics in the era of big data: a computational scientist' perspective. Statist. Probab. Lett. 136, 63–67. Special Issue on "The role of Statistics in the era of Big Data".
Reid, N., 2018. Statistical science in the world of big data. Statist. Probab. Lett. 136, 42–45. Special Issue on "The role of Statistics in the era of Big Data".
Scott, M., 2018. The role of statistics in the era of big data: crucial, critical and under-valued. Statist. Probab. Lett. 136, 20–24. Special Issue on "The role of Statistics in the era of Big Data".
Secchi, P., 2018. On the role of statistics in the era of big data: a call for a debate. Statist. Probab. Lett. 136, 10–14. Special Issue on "The role of Statistics in the era of Big Data".
Sharples, L.D., 2018. The role of statistics in the era of big data: electronic health records for healthcare research. Statist. Probab. Lett. 136, 105–110. Special Issue on "The role of Statistics in the era of Big Data".
Shi, J.Q., 2018. How do statisticians analyse big data – our story. Statist. Probab. Lett. 136, 130–133. Special Issue on "The role of Statistics in the era of Big Data".
Smirnova, E., Ivanescu, A., Bai, J., Crainiceanu, C., 2018. A practical guide to big data. Statist. Probab. Lett. 136, 25–29. Special Issue on "The role of Statistics in the era of Big Data".
Torrecilla, J.L., Romo, J., 2018. Data learning from big data. Statist. Probab. Lett. 136, 15–19. Special Issue on "The role of Statistics in the era of Big Data".
Vantini, S., 2018. Wishing the non-parametric re-evolution. Statist. Probab. Lett. 136, 139–141. Special Issue on "The role of Statistics in the era of Big Data".
Vieu, P., 2018. On dimension reduction models for functional data. Statist. Probab. Lett. 136, 134–138. Special Issue on "The role of Statistics in the era of Big Data".
Wit, E.C., 2018. Big data and biostatistics: the death of the asymptotic Valhalla. Statist. Probab. Lett. 136, 30–33. Special Issue on "The role of Statistics in the era of Big Data".
Wong, L., 2018. Big data and a bewildered lay analyst. Statist. Probab. Lett. 136, 73–77. Special Issue on "The role of Statistics in the era of Big Data".
Yu, Z., Pluta, D., Shen, T., Chen, C., Xue, G., Ombao, H., 2018. Statistical methods and challenges in connectome genetics. Statist. Probab. Lett. 136, 83–86. Special Issue on "The role of Statistics in the era of Big Data".

Laura M. Sangalli
MOX - Dipartimento di Matematica, Politecnico di Milano, Italy
E-mail address: laura.sangalli@polimi.it.