Statistics and Probability Letters 136 (2018) 139–141

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Wishing the Non-parametric Re-evolution

Simone Vantini

MOX - Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milan, Italy

ARTICLE INFO

Article history:
Available online 21 February 2018

MSC:
62A01
62G99
01A65

Keywords:
Non-parametric
Permutation test
Complex data
Big data

ABSTRACT

This short paper presents some personal considerations on the challenge that the outbreak of big data (which I prefer to call complex data) has posed to statistics as a discipline. I also unpretentiously indicate a possible (hopefully winning) strategy to tackle this challenge.

Few years ago I listened to a talk by Steve Marron (recognized as one of the fathers of Object Oriented Data Analysis). I remember one statement: "Data of the future will not be numbers, nor vectors, nor functions. They will be complex metric objects". Last year I listened to another talk, this time by Jim Ramsay (one of the fathers of Functional Data Analysis). At the beginning of his talk, while speaking about the future of functional data analysis (i.e., the statistical analysis of data sets in which each sample unit is not associated to a number but to a function), he said: "In the future, statisticians should get rid of any reference system and focus on spaces themselves". In the summary of Victor Panaretos' ERC project "Statistics for Complex Data", complex data are defined as "data that cannot be described in the standard Euclidean context of statistics, but that rather needs to be thought of as an element of an abstract mathematical space with special properties". Many other statisticians (myself included) share this point of view about the future of statistics: data that statisticians are currently and will be asked to analyse in the coming years will be so complex and so far from the reassuring Euclidean framework – which statisticians have been always used to – that most of the classical statistical modelling tools developed in the last century will become nearly worthless in many scientific fields. Some say that this is the end of statistics. On contrary, I believe this is one of the most exciting challenges that statistics has ever met: that is, developing efficient tools to analyse extremely complex data such as curves, networks, trees, images, texts, and many more.

The typical answer that statisticians have given to the challenge of "complex data" was "complex modelling". As a result, in the literature this has lead to an unseen-before proliferation of highly complex but also highly application-specific models and procedures. The likely risk of this prevalent approach is to trigger a never-ending and unsuccessful chase of models after data. Technology is indeed nowadays undoubtedly running much faster than science. Thus, if we statisticians want statistics to gain back the role it had in the history of modern science – that is driving science and not being exclusively driven by science – a strong change of direction is needed. Similarly to Christopher Columbus, statistics needs to stop going East searching for new routes towards the Indies, and start moving West: the answer to the challenge of data becoming more and more complex is not "more complex modelling" but rather "less modelling". With that, I do not mean to get rid of any modelling (that would be indeed the end of statistics), but – in line with Occam's Razor principle – trying to get rid of any unnecessary modelling superstructure pertaining to data which has been inherited from the past. There are many important

E-mail address: simone.vantini@polimi.it.

https://doi.org/10.1016/j.spl.2018.02.045
0167-7152/

examples of these superstructures that were firstly introduced to foster the progress of statistical thinking and that have now become instead barriers to further development. I here just mention two of them, probably the most abused and the most recurring assumptions in the statistical literature which has become dramatically critical in the context of complex data: "data distribution is Gaussian" (which is currently and unfortunately still at the foundations of most of the parametric literature for complex data) and "sample size is large" (which is at the foundations of most of the asymptotic literature for complex data). The definition of Gaussianity for "complex data" (being typically based on the concept of one-dimensional projections) requires indeed the embedding of data in a Hilbert space sufficiently rich to capture the information carried by data. The identification of a sufficiently rich embedding could be unfeasible (e.g., tree-valued or network-valued data) or – when possible – definitively not unique (e.g., functional data). In many applications, infinite or high dimensionality is another facet of "complex data" (e.g., functional data). In this latter setting no realistic finite sample size can be assumed to be "sufficiently close" to infinite and thus sufficient to justify the use of asymptotic theory.

In this new context, there is evidence of a need for rethinking statistical thinking. Statistics needs indeed to stop, reset, look at its history, go over parametric and asymptotic inference, take into consideration the huge (in absolute terms) but small (in relative terms) corpus of literature in the field non-parametric statistics, and – starting from there – rediscover and embrace one of its original missions, that is aiming indeed at developing both general and mathematically sound inferential tools for the data-driven inspection of phenomena for which a both realistic and mathematically tractable model is hard or even impossible to identify. I would synthetically name this regeneration process that could take to a sort of Renaissance of statistics: the *Non-parametric Re-evolution*.

To make my perspective clearer to those statisticians who had not the chance (or willingness) to experience the non-parametric fashion of doing statistics, I would like to undertake a short journey back to 1936, when Fisher published his paper "The Coefficient of Racial Likeness and the Future of Craniometry". In that paper – which is likely to contain one of the first proposals in the literature in the field of non-parametric statistics – Fisher presents a "very simple" but "very tedious" way to perform hypothesis testing simply based on data permutations. He considered the method "very simple" in the sense that it would have not required the introduction of any strong probabilistic modelling assumption and "very tedious" in the sense that it would have required an enormous (even though finite) number of computations. Using his own words:

> *"Let us suppose, for example, that we have measurements of the stature of a hundred Englishmen and a hundred Frenchmen. It may be that the first group are, on the average, an inch taller than the second, although the two sets of heights will overlap widely. [...] The simplest way of understanding quite rigorously, yet without mathematics, what the calculations of the test of significance amount to, is to consider what would happen if our two hundred actual measurements were written on cards, shuffled without regard to nationality, and divided at random into two new groups of a hundred each. This division could be done in an enormous number of ways, but though the number is enormous it is a finite and a calculable number. We may suppose that for each of these ways the difference between the two average statures is calculated. Sometimes it will be less than an inch, sometimes greater. If it is very seldom greater than an inch, in only one hundredth, for example, of the ways in which the sub-division can possibly be made, the statistician will have been right in saying that the samples differed significantly. For if, in fact, the two populations were homogeneous, there would be nothing to distinguish the particular subdivision in which the Frenchmen are separated from the Englishmen from among the aggregate of the other possible separations which might have been made. Actually, the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method"* .
>
> (Fisher, 1936)

The unexpected end of the story is: (i) on the one hand, the possibility of proving by means of Neyman–Pearson Fundamental Lemma and Fisher–Neyman Factorization Theorem – without the necessity of introducing any model-specific assumption on the probabilistic data-generative model – that this computationally intensive and only apparently heuristic approach – which technically exclusively requires a notion of distance between to sample units – is instead statistically very sound and able to provide exact inference; (ii) and, on the other hand, the fact that the first binary computer was created a few years after 1936. This idea was thus not pursued by the statistical community and the opposite approach (i.e., parametric inference) – based on strong modelling assumptions and low computational power – has become prevalent and since remained the standard, with the non-parametric approach being forced to play the secondary role of the second-best option.

One might ask why in the second half of the twentieth century with the mass outbreak computers non-parametric methods (even though rediscovered and further developed) never became mainstream among the *makers of Statistics*. My take on this is that it was, most likely, simply not the right time. Indeed, until a couple of decades ago data were typically low-dimensional and likely to fit well the Euclidean framework, and thus standard parametric methods were still working well enough that this change of perspective would not have been worth the effort. However, I believe that the current coexistence of complex data and enormous computational power has created the perfect setting for an extensive and rewarding exploitation of non-parametric inference. One of the reviewers further provocatively commented on this, pointing out that non-parametric methods are indeed surprisingly much more popular among the *users of Statistics* (e.g., experts of other fields and statistical consultants) than among the *makers of Statistics*. They think – and I feel to agree with them – that this trend could also possibly be perpetuated by the fact that it is easier to publish highly complex methods with (often) unrealistic assumptions than simple methods with realistic assumptions. For a reviewer in charge of evaluating and checking

a highly complex method, it is indeed both more difficult to spot mistakes and to fairly assess the true rate of novelty of the method.

If the *Non-parametric Re-evolution* will be embraced by the entire statistical community, the impacts in all fields of science will be ground-breaking. Scientists will be provided indeed with a sound mathematical framework for analysing even extremely complex data for which a realistic and mathematically tractable probabilistic generative model would be hard or even impossible to identify. Problems like deviations from Gaussianity or asymptotic approximations will become memories from the past. It is just a matter of leaving or not leaving:

> "*You can never cross the ocean until you have the courage to lose sight of the shore*".
>
> Christopher Columbus

Acknowledgements

My perspective on the role of statistics in the era of big data is grounded in my working experience at the laboratory MOX for Modelling and Scientific Computing of the Department of Mathematics of Politecnico di Milano. In this laboratory I had indeed the opportunity of working for more than a decade on a special type of complex data (i.e., functional data) in different research projects often in between purely methodological and industry-driven research in collaboration with colleagues from other fields from academia and the business sectors. My point of view has been of course also strongly forged by the many formal and informal discussions with all my statistician colleagues from MOX (in particular with Alessia Pini, Laura Maria Sangalli, Piercesare Secchi, and Aymeric Stamm) with whom I have had the pleasure and honour of debating and sharing ideas about the opportunities and the risks related to being a statistician at the beginning of the twenty-first century.

References

Fisher, R.A., 1936. The coefficient of racial likeness and the future of craniometry. J. Anthropol. Inst. Great Britain Ireland 66, 57–63.