Statistics and Probability Letters 136 (2018) 30–33

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Big data and biostatistics: The death of the asymptotic Valhalla

Ernst C. Wit

Johann Bernoulli Institute, Nijenborgh 9, 9747 AG Groningen, Netherlands

ARTICLE INFO

Article history:
Available online 21 February 2018

Keywords:
Biostatistics
Big data
High-dimensional inference
Model complexity

ABSTRACT

Despite the ubiquity of Big Data in the modern scientific discourse, most references describe storage and query considerations and rarely full-flexed analyses. In this article, we propose another definition with particular relevance to biometrics. We argue that the complexity of the generating measure of biological process means that the model complexity of any statistical model will have to be smaller. Only, when the model is used for prediction can we have any hope that the number of available features reasonably outnumbers the desired complexity of the model.

1. Introduction

Biostatistics stood at the cradle of the modern era of statistics about a century ago. It is often said that the work or Pearson and Fisher revolutionized statistical thinking and put it on a firm mathematical foundation, however it is also true that their work had a general biometric flavour. A century later, we ask ourselves how biostatistics is shaping up in response to the next challenge: Big Data.

Although from recent discussions it is clear that there is no universally accepted definition of Big Data, we will still attempt to define what we will understand by Big Data in a biometric setting, especially because this is different from more computer science definitions. In particular, we focus on two separate interpretations. The first harks back to the high-dimensional interpretation of big data with the aim of doing prediction. Twenty years ago microarrays revolutionized genomic screening. Large numbers of gene expressions were measured simultaneously with the possibility to relate it to some particular biological processes or clinical outcomes. The aim was to find genetic finger prints for the outcome of interest, preferably consisting of only a few genes. High-dimensional inference has become a fertile field of biometric research and have already been used in screening and prediction. At the same time, biologists have long been involved in an attempt to describe the biological reality by means of a overarching model. Dating back to the time soon after Newton, some biologists felt that they had to go beyond cataloguing the natural world and present the underlying laws governing it, just like Newton did for physics. The hope is that Big Data will shed more light on these processes and allow biostatisticians to model the process in an increasingly finer detail. This will require new, more complex mathematical models, such as various network models.

2. What is big biometric data?

Jacobs (2009) defines big data as "data too large to be … analyzed with the help of a desktop statistics … ackage". Although it stops short from saying that big data is data that is not amenable to statistical analyses, for modern statisticians the desktop statistics package is our weapon of choice and it would be disappointing if we did not have anything to say anymore about

E-mail address: e.c.wit@rug.nl.

https://doi.org/10.1016/j.spl.2018.02.039
0167-7152/

Big Data. So, whereas most common definitions focus on sample size n, potentially augmented by time and space, we will focus on model complexity p in relation to true process complexity p_0. We will distinguish two types of big data, which computer scientists may snarl at with contempt:

1. $p \gg p_0$: high-dimensional data for feature detection and
2. $p_0 \gg p$: dense system-wide data of a complex process.

This definition of big data makes no reference to their size. In fact, the data could possibly only as big as a few GB that may even fit in the RAM of a standard desktop computer. But it is not sheer size that I would like to discuss here, but their computational complexity and epistemic intricacies.

In traditional statistics, sample size is a mere storage issue, perhaps with the only additional consideration of being able to calculate a model's sufficient statistics. As long the latter can be calculated efficiently, statistical inference can proceed without further reference to the original sample. We do not want to downplay the skill and ingenuity that is required to deal with these issues: computer scientists have been making enormous progress with both issues and although there will always be, at any given time, a physical limit to what storage capabilities and computational operations are possible, this limit is continuously growing with the help of new information technologies and ingenious solutions.

However, as Bühlmann and Meinshausen (2016) point out, in most biometric and other applications with increasing n, the data will exhibit "inhomogeneities", due to changes in sampling design, the introduction of novel measuring technologies or even changes in the underlying population. Perhaps one of the most disappointing revelations of Big Data has been the fact that increasing n does not bring us closer to an asymptotic Valhalla, but instead to a non-ergodic Hell full of sampling errors, changing definitions and drifting realities. Before becoming too downtrodden, we will refocus our attention on the purpose of a statistical model. In as much it is aimed at learning about the underlying truth, a model itself is a tool, like a hook, to get a grip on that same reality. For biologists, this may be creating a coherent paradigm of the natural world, whereas for e.g. medical practitioners this may be to find predictive features of the phenomenon of interest.

3. Inference of complex biological models

Modern biology has been revolutionized by the genomic revolution. Not only the impressive development of genomic technologies is a sign of this innovation, but even the biological jargon itself is steeped in this revolution. Observable biological processes, which were the subject of biological treatises for centuries, are nowadays often referred to as phenotypes, as if they are mere shadows or manifestations of a more fundamental, underlying genomic reality. However, further study has revealed that even that was a simplification and that important biological phenomena occur at different scales, which cannot be reduced to one another. Although we cannot exclude the philosophical possibility that one day in a distant future we might come up with a Theory of Everything, which includes biological, medical and environmental processes, *now* we have to come to terms with the epistemic reality that our current scientific gaze will only be able to capture snapshots of the underlying biological complexity.

When we translate this to (bio)statistics, there are some important practical consequences.

First, we should keep in mind that our model will always be wrong. We will express this by the shorthand that the model complexity p will always be smaller that the complexity of underlying process p_0. One special case to illustrate this condensation $p \ll p_0$ is the scenario in which we consider a simple regression problem with p explanatory variables, but where there are $p_0 - p$ unobserved confounders. However, the statement goes actually further: the generating measure of the process we observe in our data will never be a regression model. The biological reality is so complex that the generating measure will always escape our epistemic conceptualizations in some (stochastic) model. In a way, this is another way of describing the all important dose of scientific scepticism that every scientist should possess.

Secondly, we should not forget that some models are useful (Box, 1976). Models are abstractions of reality and as such highlight a simplified version of that reality that allows us to make sense and use of it. Practically, this means that we have to choose between a variety of models. Within a biometrical setting, this choice is typically, on the one hand, theory-driven and on the other data-driven. To make this idea clear, we will focus on the example of a genomic network.

Networks have become an important paradigm to describe genomic systems: from describing the physical, molecular interactions between proteins to the abstract interactions functional genetic units, the vocabulary of networks has been adopted eagerly by biologists tasked with studying complex biological system. For example, Costanzo et al. (2016) argue that a global genetic interaction network highlights the functional organization of a cell and provides a resource for predicting gene and pathway function. The language of networks encapsulate the notion of the nodes, i.e. "genes", connected by edges, i.e. "genetic interactions".

The sampling scheme and design of a genomic experiment should match the type of model that is used for analysing it. We outline three modelling strategies, that are useful in describing various aspects of a "genomic network". A system of stochastic differential equations can describe single cell interactions, which takes into account the underlying stochasticity of particle interactions (Wilkinson, 2006; Purutçuoğlu and Wit, 2008). Often, however, genomic data is collected at either a more agglomerated level or across a number of cells that are destructively sampled. In those cases, temporal models are more appropriately described by means of ordinary differential equations (Papoutsakis, 1984). In large genomic systems, both SDE and ODE descriptions can be unstable or computationally prohibitive. In such cases, vector autoregressive models are useful (Abegaz and Wit, 2013).

All these models are inherently dynamic, describe the same underlying process, but highlight different aspects thereof. After making a choice with respect to the type of model to consider, more data-driven model selection criteria can be used to narrow down the models for further consideration. Models with good predictive performance are found through Kullback–Leibler type measures, such as AIC and cross-validation, whereas individual relevant model-components are selected through posterior probability criteria, such as BIC (Wit et al., 2012).

4. High-dimensional inference of biostatistical phenomena

There are nevertheless scenarios where, irrespective of the underlying biological complexity, the focus is about finding a predictive model. This scenario is very common in biomedical settings, where the aim is to find predictive variables for some clinical outcome. The true predictive model – not to be confused with the true generative model, as conveniently available consequences can be part of the predictive model, but not of the generative model – in such cases is assumed to be of a complexity p_0 that is much smaller than the number of available features p in the data. We consider the example of cancer survival.

Advances in genomic technologies have meant that many new clinical studies in cancer survival include a variety of genomic measurements, ranging from gene expression to SNP data. Studying the relationship between survival and genomic markers can be useful for a variety of reasons. If a genomic signature can be found, then patients can be given more accurate survival information. Furthermore, treatment and care may be adjusted to the prospects of an individual patient.

Sparse inference in the past two decades has been dominated by methods that penalize typically convex likelihoods by functions of the parameters that happen to induce solutions with many zeros. The lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), l_0 (Rippe et al., 2012) and the SCAD (Fan and Li, 2001) penalties are examples of such penalties that, depending on some tuning parameter, conveniently shrink estimates to exact zeros. Also in survival analysis these methods have been introduced. Tibshirani (1997) applied the lasso penalty to the Cox proportional hazards model.

Whereas penalized inference is convenient, justification of the penalty is somewhat problematic. Interpreting the solution as a Bayesian MAP estimator with a particular prior on the parameters seems to merely reformulate the problem, rather than solving it. Furthermore, the methods suffer from being not invariant under scale transformations of the explanatory variables. This means that measuring, e.g., height in centimetres or inches can and probably will result in dramatically different answers. Therefore, most penalized regression methods start their exposition by assuming that the variables are appropriately renormalized. This is clearly a merely algorithmic device and simply begs the question of invariance. Clearly the strongest argument in favour of some of these methods are their asymptotic properties. Nevertheless, what this means in the small sample settings encountered in practice is also problematic. Recent developments approach sparsity directly from a likelihood point of view. The angle between the covariates and the tangent residual vector within the likelihood manifold provides a direct and scale-invariant way to assess the importance of the individual covariates (Augugliaro et al., 2013, 2016).

5. Discussion

Naturally all the usual storage and retrieval issues of Big Data, as emphasized in computer science definitions, are also an issue for biostatistics. However, the biggest impact of Big Data on Biostatistics is the realization that increasing n means increasing complexity: changing sampling protocols, new measurement technologies and non-ergodic realities means that old fashioned asymptotics fail. This means that for explanatory models, we have to find a fine balance between model complexity and expressiveness, whereas for predictive models, novel work on regularized inference is particularly promising.

Acknowledgements

This work is part of the research programmes Mathematics for Planet Earth (NWO 657.014.005) and the European Cooperation for Statistics of Network Data Science (CA15109) which are financed by the Netherlands Organisation for Scientific Research (NWO) and the European Cooperation in Science and Technology, respectively.

References

Abegaz, F., Wit, E., 2013. Sparse time series chain graphical models for reconstructing genetic networks. Biostatistics 14 (3), 586–599.
Augugliaro, L., Mineo, A., Wit, E., 2013. Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (3), 471–498.
Augugliaro, L., Mineo, A.M., Wit, E.C., 2016. A differential geometric approach to generalized linear models with grouped predictors. Biometrika 103 (3), 563–577.
Box, G.E., 1976. Science and statistics. J. Amer. Statist. Assoc. 71 (356), 791–799.
Bühlmann, P., Meinshausen, N., 2016. Magging: maximin aggregation for inhomogeneous large-scale data. Proc. IEEE 104 (1), 126–135.
Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al., 2016. A global genetic interaction network maps a wiring diagram of cellular function. Science 353 (6306), aaf1420.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.
Jacobs, A., 2009. The pathologies of big data. Commun. ACM 52 (8), 36–44.

Papoutsakis, E.T., 1984. Equations and calculations for fermentations of butyric acid bacteria. Biotechnol. Bioeng. 26 (2), 174–187.
Purutçuoğlu, V., Wit, E., 2008. Bayesian inference for the mapk/erk pathway by considering the dependency of the kinetic parameters. Bayesian Anal. 3 (4), 851–886.
Rippe, R.C., Meulman, J.J., Eilers, P.H., 2012. Visualization of genomic changes by segmented smoothing using an l 0 penalty. PLoS One 7 (6), e38230.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
Tibshirani, R., 1997. The lasso method for variable selection in the cox model. Stat. Med. 16, 385–395.
Wilkinson, D.J., 2006. Stochastic Modelling for Systems Biology. Chapman and Hall/CRC.
Wit, E., Heuvel, E.V.D., Romeijn, J.-W., 2012. All models are wrong...: an introduction to model uncertainty. Stat. Neerl. 66 (3), 217–236.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2), 301–320.