ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Statistics for big data: A perspective

Peter Bühlmann *, Sara van de Geer

Seminar for Statistics, ETH Zürich, Switzerland

ARTICLE INFO

Article history:
Available online xxxx

MSC:
primary 62-01
secondary 68-01

Keywords:
Heterogeneity
Large-scale data
Lasso
Learning theory
Mathematical theory
Negative results
Replicability
Reproducibility
Validation

ABSTRACT

We look at the role of statistics in data science. Two statisticians, two views. Besides the need of developing appropriate concepts, methodology and algorithms, the first one makes in Section 3 a case for validation and carefully designed simulation studies, while the second one writes in Section 4 that a mathematical underpinning of methods is fundamental. Both views converge to the same point: there should be more room for publishing negative findings.

1. A short introduction

"Big Data" is perhaps not a well-defined terminology. Wikipedia (https://en.wikipedia.org/wiki/Big_data) states the following: "Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time".

Computation, open software, open data and reproducibility. The computational issue mentioned above is certainly a relevant one. We publicize open source software development for large scale problems, open data access and corresponding reproducibility (see Section 2). Much has been written in the last few years about these topics and we want to re-emphasize their importance without going into more details.

From a more fundamental research perspective, we mention here the interesting framework of "computational statistical trade-offs" (Chandrasekaran and Jordan, 2013): it seems particularly interesting for addressing the problem of statistical optimality (optimal statistical information extraction) under computational constraints.

In the following, we will focus on issues and challenges beyond computation, namely: replicability; heterogeneity in large-scale datasets; causality and its connection to robust prediction; validation and a so-called NCP guidelines, where "NCP" stands for No Cherry Picking; and we close with a reflection in terms of an epilogue.

2. Inferential statistics and replicability

Perhaps *the* core statistical task in (traditional) statistics is inductive inference from data to models and scientific conclusions; and we believe that this core task is still very relevant in the advent of massive datasets. Thereby, the assessment

* Corresponding author.
E-mail addresses: buhlmann@stat.math.ethz.ch (P. Bühlmann), geer@stat.math.ethz.ch (S. van de Geer).

https://doi.org/10.1016/j.spl.2018.02.016
0167-7152/

of uncertainties is crucial and Bayesian or frequentist statistics offer a unique framework with a long history to address this issue. Ideally, the conclusions and findings are replicable.

Replicability should not be confused with reproducibility. Reproducibility when defined in a narrow technical sense means: the same data input leads to the same output and numerical results. For this, one needs open access to data and software with a clear protocol how one goes from the input to the output; see our comment in Section 1. On the other hand, replicability means that the findings in a study can be successfully replicated in another study under the same conditions, not exactly but up to statistical error. The statistical error might be quite large and failure of successful replication might be sometimes not so surprising: a-priori, it is difficult to judge unless we have an understanding about the statistical uncertainties!

A major challenge with Big Data is to come up with reasonable models, methods and algorithms which enable to quantify the statistical error — and hopefully, some findings from Big Data would indeed enable a fair amount of success for replication in other studies or datasets.

Stability and replicability. Related to replicability is the notion of stability (Yu, 2013). For example, one can consider the notion of replicability within sub-samples of a large dataset: if a finding from half of the data can be successfully replicated in the other half, then the finding seems to be more "promising" and "relevant". Stability Selection (Meinshausen and Bühlmann, 2010) provides a methodology for assessing such "relevance". It uses repeated splitting of a dataset into two halves and connects the notion of "stable findings" to a type I error rate in terms of the expected number of false positives, based on a very simple formula (Meinshausen and Bühlmann, 2010, formula (9)). We note that stability turns out to be beneficial for predictive performance as well (Breiman, 1996).

2.1. Heterogeneity

Stability and replicability usually require some form of homogeneity and independence. In the simplest terms, the first and the second dataset (or the first and second half of a single dataset) are independent and of the same nature. We believe that Big Data are often of heterogeneous nature, violating such a homogeneity assumption. In other words, large-scale data, a terminology which we use in the sequel for data which have both large sample size and dimensionality, is typically *not* generated as i.i.d. or stationary samples from a single population distribution and we have good reasons to expect that the data is heterogeneous, arising from different "scenarios", "regimes" or "sub-populations". Quantifying the statistical error and the amount of replicability which we can expect under such circumstances are important for drawing generalizing and scientific conclusions from large-scale data.

To make things more precise, consider a regression or classification setting with univariate response Y and p-dimensional covariate X. We assume that we have several "environments", "regimes" or "sub-populations" encoded by $e \in \mathcal{E}$, i.e., e encodes an environment and $\mathcal{E}$ is the space of environments which are present in the dataset (e and $\mathcal{E}$ are either observed or they are unknown): the dataset then consists of

$$(Y^e, X^e),\ e \in \mathcal{E} \tag{1}$$

with the understanding that the dataset is homogeneous within each environment e; here Y^e and X^e denote the data vector ($n_e \times 1$) or matrix ($n_e \times p$), measuring the same variables across all the environments.

An established way to assign uncertainties and the amount of replicability in heterogeneous data is given by *meta analysis* (Hedges and Olkin, 1985; Owen, 2009, cf.). A large-scale heterogeneous dataset as in (1) with known (observed) environments can be divided into smaller homogeneous sub-datasets, and in the language of meta-analysis these sub-datasets are the different studies from which one pursues an aggregation of significance statements. Meta analysis is a straightforward concept and computationally attractive since computing for every environment e can be trivially distributed.

Another kind of heterogeneity is mentioned in Secchi (2017) where different variables are measured in different datasets (which can be understood as different "sub-populations"). What we discuss below might be partially useful for such a setting as well.

2.2. Heterogeneity, causality and robustness against worst case scenarios

Heterogeneity in large-scale data seems to be an unpleasant obstacle. However, it can be advantageously exploited for causal inference and robustness against worst case scenarios.

Consider heterogeneous data as in (1). An interesting *invariance* assumption, related to stability, is as follows: there exists a subset $S \subseteq \{1, \ldots, p\}$ of covariates such that the conditional distribution

$$\mathcal{L}(Y^e | X^e_S) \text{ is invariant across all } e \in \mathcal{E}, \tag{2}$$

where X^e_S denotes the covariates of the subset S in environment e. There might be several such subsets S, and these subsets of covariates (which satisfy the invariance assumption in (2)) are interesting in their own right since they satisfy a stability or invariance assumption. That is, stability (and its related replicability) might occur in the form of a conditional distribution as in (2).

Please cite this article in press as: Bühlmann P., van de Geer S., Statistics for big data: A perspective. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.016.

Under some assumptions, the *causal* variables (for causality, cf. Pearl (2000)) satisfy the invariance assumption in (2); and vice-versa, the invariance or robustness within a heterogeneous large-scale dataset can be exploited in an algorithm for estimating causal variables for the response Y, including a type I error guarantee for guarding against false positive (causal) findings. For details we refer to Peters et al. (2016).

Causal models have a robustness property due to the so-called autonomy assumption (Aldrich, 1989). In terms of heterogeneous data as in (1) the following holds. Denote by $S_{\text{causal}} \subseteq \{1, \ldots, p\}$ the indices of the causal variables: the autonomy assumption says that invariance holds,

$$\mathcal{L}(Y^e | X^e_{\text{causal}}) \text{ is invariant across all } e \in \mathcal{F}, \tag{3}$$

where $\mathcal{F}$ denotes a large class of possible environments, typically much larger than the environments from $\mathcal{E}(\subset \mathcal{F})$ which are present in the data. This then implies for e.g. a linear model that

$$\operatorname{argmin}_{\beta} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e\beta|^2 \ni \beta_{\text{causal}}, \tag{4}$$

where β_{causal} are the causal parameters with support

$$\text{supp}(\beta_{\text{causal}}) = \{j;\ \beta_{\text{causal};j} \neq 0\} = S_{\text{causal}}.$$

Thus, the causal parameters – under some assumptions – insure against worst case scenarios (from $\mathcal{F}$). Other methodologies are useful to ensure against scenarios which are not exactly present in the data but where the data (the past) still carries useful information about new unseen data (the future): see for example the framework of maximin estimation (Meinshausen and Bühlmann, 2015) and maximin aggregation (Bühlmann and Meinshausen, 2016) which optimizes a worst case performance as in (4) where $\mathcal{F}$ is – roughly speaking – the convex hull of $\mathcal{E}$.

worst case is about the convex hull.

3. Validation and the NCP guidelines

An important task is to *validate* (machine learning or statistical) algorithms, for example their predictive performance or their ability to find important or even causal variables. Validation of prediction is certainly easier than assessing the performance to find important or causal variables. However, already the task of prediction is not trivial at all: standard cross-validation is inappropriate for heterogeneous data, and assessing the prediction performance for new unseen scenarios (from $\mathcal{F}$, see (3)) is highly non-trivial.

3.1. Simulations and no cherry picking!

In view of the difficulty for validation, a standard approach for checking the performance of algorithms and methods is based on synthetic data and running simulations. It is a reasonable first step. However, it is certainly a challenge to construct realistic simulation models for Big Data. A good example exists in genetics, where PLINK (Purcell et al., 2007) allows to simulate synthetic data based on a model driven by physical and biological laws.

The NCP guidelines. We propose here the No Cherry Picking (NCP) guidelines for validating algorithms and methods using synthetic data.

The NCP guidelines

1. The performance of algorithms and methods should be reported for a broad range of simulation models.
2. For algorithms or methods, especially when newly proposed, their strengths *and* weaknesses should be reported.
3. The simulation models in point 1 should include fairly realistic scenarios as well as extreme "cornerpoints". The latter are mainly serving to empirically illustrate point 2, describing the range of scenarios where the methods work or break down.
4. If possible, some mathematical intuition should be given to back up the empirical findings (see also below).

A precise mathematical description (see point 4. of the NCP guidelines) to understand the regime of problems where a new method works well is usually difficult since the corresponding conditions would then be sufficient and necessary. An example where this is possible is the irrepresentable condition for the Lasso for variable selection in high-dimensional linear models (Zou, 2006; Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Bühlmann and van de Geer, 2011). Whenever a precise mathematical characterization is not (or only partially) possible, one can resort to simulation studies. To understand the potential and limitations of a (new) method or algorithm, one should provide empirical results where the method works

well and where it fails (repeating point 2. in the NCP guidelines). Obviously, some empirical understanding of the range of problems for success and failure is highly informative (points 2 and 3 in the NCP guidelines).

Referees, associate editors and editors should encourage authors to provide such transparent and more informative empirical results: editors might even insist to report on problems where the methods fail. Such a transparent, honest and informative way of reporting should be (more) adopted within the culture of statistics: we, the community in statistics, should be among the first to improve on this simple but important point! In some cases, the importance or quality of a contribution, which is the main criterion for editorial decisions, can be improved by following the NCP guidelines.

4. Reflection: an epilogue

We have lots of data these days, many numbers (zeros and ones) stored in our computers. We have much more data than in the old days, when measurements of the orbit of a single planet took weeks. Then in the 18th century Gauss invented the least squares method. As soon as there are data, there comes along a statistical method. Because data are there to be analyzed. Think of your data sitting in your computer. If you do not get them out of there what is the sense of it all?

There are several things one might aim at. For example, if the aim is to know whether a person is creditworthy, one runs an algorithm on his or her data, and out comes an answer. The main point is that the answer should typically be correct. One might not be so much interested in how the algorithm works. One even might be not interested in the psychology of behavior and why a person with certain characteristics is not creditworthy.

A relatively recent term is learning theory. The difference with classical statistics is I believe that it is less model based, but rather algorithm based. Learning theory is perhaps also more directly focusing on some parameter of interest and does not model the nuisance parameters. We run the learning algorithm, and out comes an answer. We now see that the robot can walk and the doll can recognize our emotions. We are surrounded by very smart computers. The computer has learned, maybe in a supervised setting, maybe unsupervised or reinforced. But we, what have we learned?

Some algorithms are based on models and explicitly on statistical methods, others less so. I would in either case call algorithms the statistical skeleton. Where is the rest of the body? It used to be clear what a particular method for analyzing data did. The least squares method for example was and is well understood. Today's algorithms are complex and often hardly understood. Many are not understood by most of us, and some are not understood by any of us.

When trying to develop more insight the statistician is facing several questions. The first question concerns generalizability and replicability. If one imagines running the algorithm again, now on a new dataset, would the outcome be similar? The second question is how to find out what similarity actually means, how to access accuracy, to quantify uncertainty. The third task is verify whether this uncertainty is something we can measure.

These are generally hard questions. One can try to tackle them by testing the algorithms on pseudo random artificial data. But how to generate these data? From what models? Another way is to develop mathematical theory, based on abstract models which are more general and flexible than those used in simulation studies. Developing the math can take some time. The very popular Lasso for instance was introduced by Robert Tibshirani in 1994 (and published two years later Tibshirani, 1996) and still, more than 20 years later, mathematical statistical journals publish many papers on the Lasso. The mathematical understanding of algorithms is hard work and yet the practical consequences are less visible. Replicability is typically a must, and here statistics acts primarily as rough quality test. If the algorithm does not pass the test it will be abandoned. The statistician is only noticed (and then sometimes held responsible) as soon as the test is passed but replicability is actually not there in new real data studies. Also the uncertainty can be mathematically well understood with little practical consequence because this uncertainty is not something you can directly see from data. One needs another algorithm to access uncertainty, and this algorithm has its own uncertainty. The task of the statistician is to quantify how far we can go. The modern adaptive algorithms are often such that the better they are (generate outcomes close to the target), the harder it is to estimate their uncertainty. For the "best" algorithm it can be simply impossible to know its accuracy. To quote Marc Hoffmann and Oleg Lepski: "You know adaptive estimators converge very fast if the function is very smooth (or has a prescribed complexity) but you can tell nothing about the estimated function itself" (Hoffmann and Lepski, 2002). The statistical theory serves as guard against cheating with data: you cannot beat the uncertainty principle.

The scientific motivation for analyzing data is to increase knowledge. Think for example at finding out which gene is associated with a particular phenotype. A classical statistical uncertainty measure is the p-value. If it is small one has found something. Then the result is called significant and published in scientific journals. Lately, some panic is going on as it turned out that many findings are not replicable. The statistical uncertainty measure seems not to be reliable! The p-value is cheating! However, there is no reason for surprise as there is no outlet for publishing insignificant results (at least in domains where statistical significance is required). Maybe editors should invite authors to write about their journey along the path of insignificant results that led to the significant one. There is moreover also no reason for panic. False positives are part of the process of gaining knowledge and insight. The significant findings are to subjected to further studies and sense has to be made out of it. Thus statistics plays its role in understanding algorithms, the value of their outcomes and consequent decisions for new research directions.

Please cite this article in press as: Bühlmann P., van de Geer S., Statistics for big data: A perspective. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.016.

References

Aldrich, J., 1989. Autonomy. Oxford Econ. Pap. 41, 15–34.
Breiman, L., 1996. Heuristics of instability and stabilization in model selection. Ann. Statist. 24, 2350–2383.
Bühlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer.
Bühlmann, P., Meinshausen, N., 2016. Magging: maximin aggregation for inhomogeneous large-scale data. Proc. IEEE 104, 126–135.
Chandrasekaran, V., Jordan, M., 2013. Computational and statistical tradeoffs via convex relaxation. Proc. Natl. Acad. Sci. 110, E1181–E1190.
Hedges, L., Olkin, I., 1985. Statistical Methods for Meta-Analysis. Academic Press.
Hoffmann, M., Lepski, O., 2002. Random rates in anisotropic regression (with a discussion and a rejoinder by the authors). Ann. Statist. 30, 325–396.
Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the Lasso. Ann. Statist. 34, 1436–1462.
Meinshausen, N., Bühlmann, P., 2010. Stability Selection (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 72, 417–473.
Meinshausen, N., Bühlmann, P., 2015. Maximin effects in inhomogeneous large-scale data. Ann. Statist. 43, 1801–1830.
Owen, A., 2009. Karl Pearson's meta-analysis revisited. Ann. Statist. 37, 3867–3892.
Pearl, J., 2000. Causality: Models, Reasoning and Inference. Cambridge University Press.
Peters, J., Bühlmann, P., Meinshausen, N., 2016. Causal inference using invariant prediction: identification and confidence interval (with discussion). J. Roy. Statist. Soc. Ser B. 78, 947–1012.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M., Sham, P., 2007. PLINK: a toolset for whole-genome association and population-based, linkage analysis. Am. J. Hum. Genet. 81, 559–575.
Secchi, P., 2017. On the role of statistics in the era of big data: a call for a debate. Statist. Probab. Lett. Special Issue on The role of Statistics in the era of big data (in press).
Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 267–288.
Yu, B., 2013. Stability. Bernoulli 19, 1484–1500.
Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. J. Mach. Learn. Res. 7, 2541–2563.
Zou, H., 2006. The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.

Please cite this article in press as: Bühlmann P., van de Geer S., Statistics for big data: A perspective. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.016.