TEST
https://doi.org/10.1007/s11749-019-00646-6

DISCUSSION

Comments on: Data science, big data and statistics

Peter Bühlmann[1]

Abstract
We congratulate Pedro Galeano and Daniel Peña for a nice paper on the emerging theme of data science and the role of statistics.

Keywords Big data · Causal inference · Data science · Heterogeneity · High-dimensional statistics · Robustness

Mathematics Subject Classification 62-01

1 A summary of the paper

Galeano and Peña, referred in the sequel as "GP," present (aspects of) their interesting view of analyzing "big data": they approach the vast and broad theme by focusing on seven important points, they provide a large amount of references covering different cultures (Breiman 2001b), and they illustrate and exemplify their view by two large-scale data analyses.

1.1 Other data sources

Among the seven points from GP, I would like to re-emphasize the importance of new sources of information. Indeed, images, videos and audios are typically cheap devices to record data. GP do not mention recent progress with autoencoders (Hinton and Salakhutdinov 2006; Vincent et al. 2010): when using such techniques, one would again end up with numeric features which can then be used for further downstream analysis using techniques from high-dimensional statistics or statistical machine learning (Bühlmann and van de Geer 2011; Hastie et al. 2015, cf.).

This comment refers to the invited paper available at: https://doi.org/10.1007/s11749-019-00651-9.

✉ Peter Bühlmann
buhlmann@stat.math.ethz.ch

[1] Seminar for Statistics, ETH Zürich, 8092 Zurich, Switzerland

Published online: 08 April 2019

Single cell biology is an interesting example, not mentioned by GP, where imaging techniques are used as automated and low-cost devices to collect a vast amount of information (Carpenter et al. 2006; Kamentsky et al. 2011, cf.). Images can describe directly a phenotype or response like "cell is infected" or "cancerous tissue," at the price of less direct understanding of how genes or proteins function (Rämö et al. 2014). It is an interesting challenge to extract more interpretable systems—insights from image data.

1.2 Beyond multiple testing and false discovery rate: stability

Multiple testing has indeed become standard practice in many applications. If one really wants to find out about real potential importance or relevance of a variable or a group of variables, some additional sensitivity and stability analysis is advised to do, whether one uses a Bayesian or frequentist framework. An easy way to do this is via subsampling or bootstrapping, thus checking the stability with respect to perturbing the random sampling from a data generating distribution. This is in the spirit of Breiman for prediction (Breiman 1996a, b, 2001a) but is formalized and further developed for feature or variable selection with stability selection (Meinshausen and Bühlmann 2010); Yu (2013) developed related ideas for such settings. Another kind of stability is discussed next.

2 Heterogeneity, another stability and causality

GP point out the important issue that data are often heterogeneous. Indeed, the data might come from different subpopulations or clusters, as mentioned also by GP. This seems at first sight an obstacle and a nuisance, but we believe that it can be also a real blessing!

Consider the setting with a response Y and covariates X, as in regression or classification. We assume that we have various subgroups or environments, denoted by $e \in \mathcal{E}$ where $\mathcal{E}$ is the space of observed environments (e.g. $\mathcal{E} = \{1, 2, \ldots, 10\}$ encodes for 10 different countries in the observed dataset): the variables in each environment are then denoted by Y^e and X^e. We can look at stability or invariance across the different environments with respect to *conditional* distributions. We require that there is a subset of covariates S^* such that

$$\mathcal{L}(Y^e | X^e_{S^*}) \text{ is the same for all } e \in \mathcal{E}. \tag{1}$$

For a linear model, this translates as follows: there exists a single parameter vector β^* with $\text{supp}(\beta^*) = S^*$ and a single distribution F_ε such that for all $e \in \mathcal{E}$,

$$Y^e = X^e \beta^* + \varepsilon^e, \ \varepsilon^e \sim F_\varepsilon, \ \varepsilon^e \text{ independent of} X^e_{S^*}. \tag{2}$$

The set of covariates S^* is interesting as it stabilizes the regression coefficients and residual distributions. Furthermore, if the set of environments $\mathcal{E}$ is sufficiently rich and fulfills certain conditions, S^* equals the set of causal variables for Y. The invariance

principle in (1) or (2) can be used to infer stable solutions and causality from heterogeneous data (Peters et al. 2016; Heinze-Deml et al. 2018): even if inferring causality is impossible, one can gain predictive and distributional robustness by constructing methods and algorithms which are empirical estimates of a worst case risk optimizer

$$\beta^*(\mathcal{F}) = \text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}[(Y^e - (X^e)^T \beta)^2]. \tag{3}$$

The class $\mathcal{F}$ denotes a space of environments which is typically (much) larger than $\mathcal{E}$ which is observed in the data. Thus, (3) leads to robustness for future scenarios or environments outside the data range (Rothenhäusler et al. 2018; Bühlmann 2018). The robustness in (3) is a certain kind of distributional robustness (Meinshausen 2018), a theme which is important in adversarial training of deep networks (Sinha et al. 2017, cf.).

The key tool for achieving such robustness, stability or even causality is the invariance assumption in (1) or versions of it. Only because the data are heterogeneous, there is a possibility to estimate invariance and related properties as one needs to inspect over different observed environments $e \in \mathcal{E}$. Even if $\mathcal{E}$ is not known, one can estimate the heterogeneities from data: it is a kind of change point or cluster estimation problem (Pfister et al. 2018).

Another approach for achieving a vaguely related robustness as in (3) based on heterogeneous data, but with neither a causal-type model nor a corresponding interpretation, has been proposed using some aggregation techniques (Meinshausen and Bühlmann 2015; Bühlmann and Meinshausen 2016).

3 The role of statistics in data science

Statistics has the longest tradition in data analysis and extraction of meaningful information from data. The automatic and large-scale data collection in some (but not all) applications in science and engineering calls for new methodology and algorithms: statistics at the interface and together with machine learning, artificial intelligence, and in the broader sense together with mathematical, information and computer sciences, has a central role to play. The amount of problems to be addressed is huge: too large for one subcommunity to deal with, especially for the utterly important core task of teaching and education in data science!

When shaping a modern curriculum in statistics and data science, we should rely on the expertise and work force from the other related communities, and vice versa! And this principle also applies to research and development. GP stepped out of the well-known home and moved forward into the land of big data analysis: Congratulations!

References

Breiman L (1996a) Bagging predictors. Mach Learn 24:123–140
Breiman L (1996b) Heuristics of instability and stabilization in model selection. Ann Stat 24:2350–2383
Breiman L (2001a) Random forests. Mach Learn 45:5–32

Breiman L (2001b) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16:199–231
Bühlmann P (2018) Invariance, causality and robustness. Preprint arXiv:1812.08233
Bühlmann P, Meinshausen N (2016) Magging: maximin aggregation for inhomogeneous large-scale data. Proc IEEE 104:126–135
Bühlmann P, van de Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer, Berlin
Carpenter A, Jones T, Lamprecht M, Clarke C, Kang I, Friman O, Guertin D, Chang J, Lindquist R, Moffat J, Golland P, Sabatini D (2006) Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol 7:R100
Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the lasso and generalizations. CRC Press, Boca Raton
Heinze-Deml C, Peters J, Meinshausen N (2018) Invariant causal prediction for nonlinear models. J Causal Inference 6:20170016. https://doi.org/10.1515/jci-2017-0016
Hinton G, Salakhutdinov R (2006) Reducing the dimensionality of data with neural networks. Science 313:504–507
Kamentsky L, Jones T, Fraser A, Bray M, Logan D, Madden K, Ljosa V, Rueden C, Eliceiri K, Carpenter A (2011) Improved structure, function and compatibility for cellprofiler: modular high-throughput image analysis software. Bioinformatics 27:1179–1180
Meinshausen N (2018) Causality from a distributional robustness point of view. In: 2018 IEEE data science workshop (DSW). IEEE, pp 6–10
Meinshausen N, Bühlmann P (2010) Stability selection (with discussion). J R Stat Soc Ser B 72:417–473
Meinshausen N, Bühlmann P (2015) Maximin effects in inhomogeneous large-scale data. Ann Stat 43:1801–1830
Peters J, Bühlmann P, Meinshausen N (2016) Causal inference using invariant prediction: identification and confidence interval (with discussion). J R Stat Soc Ser B 78:947–1012
Pfister N, Bühlmann P, Peters J (2018) Invariant causal prediction for sequential data. J Am Stat Assoc 2018. https://doi.org/10.1080/01621459.2018.1491403
Rämö P, Drewek A, Arrieumerlou C, Beerenwinkel N, Ben-Tekaya H, Cardel B, Casanova A, Conde-Alvarez R, Cossart P, Csúcs G, Eicher S, Emmenlauer M, Greber U, Hardt W-D, Helenius A, Kasper C, Kaufmann A, Kreibich S, Kühbacher A, Kunszt P, Low S, Mercer J, Mudrak S, Muntwiler S, Pelkmans L, Pizarro-Cerda J, Podvinec M, Pujadas E, Rinn B, Rouilly V, Schmich F, Siebourg-Polster J, Snijder B, Stebler M, Studer G, Szczurek E, Truttmann M, von Mering C, Vonderheit A, Yakimovich A, Bühlmann P, Dehio C (2014) Simultaneous analysis of large-scale RNAi screens for pathogen entry. BMC Genomics 15(1):1162
Rothenhäusler D, Meinshausen N, Bühlmann P, Peters J (2018) Anchor regression: heterogeneous data meets causality. Preprint arXiv:1801.06229
Sinha A, Namkoong H, Duchi J (2017) Certifiable distributional robustness with principled adversarial training. Preprint arXiv:1710.10571. Presented at sixth international conference on learning representations (ICLR 2018)
Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11:3371–3408
Yu B (2013) Stability. Bernoulli 19:1484–1500