Data-Adaptive Statistical Inference: foreword to the DGIJB special issue

A. Chambaz[1], A. Hubbard[2], M. J. van der Laan[2]

[1] Modal'X, Université Paris Ouest Nanterre, France
[2] Division of Biostatistics, University of California, Berkeley, CA, USA

The concomitant emergence of big data, explosion of ubiquitous computational resources and democratization of the access to more powerful computing make it necessary and possible to rethink pragmatically the practice of statistics. While numerous machine learning methods provide much ever easier access to data-mining tools and sophisticated prediction, there is a growing realization that *ad hoc* and non-prespecified approaches to high-dimensional problems lend themselves to a proliferation of "findings" of dubious reproducibility. This period of fast-paced evolution is thus a blessing for statistics. It is a golden opportunity to build upon more than a century of methodological research in statistics and five decades of methodological research in machine learning to bend the course of statistics in a new direction, away from the misuse of parametric models and reporting of non-robust inference, to tackle rigorously the challenges that we, as a community, are confronted with.

The foundation of statistics is incorporating knowledge about the data-generating experiment through the definition of a statistical model (a set of laws), formalizing the question of interest through the definition of an estimand seen as the value of a statistical parameter (a functional mapping the model to a parameter set) at the true law of the experiment and inferring the estimand based on data yielded by the experiment. Typically, one would construct an estimator of (a collection of key features) of the true law and evaluate the statistical parameter at its value. The present special issue broadly focuses on the inference of various statistical parameters in situations where either the data-generating law or the statistical parameter or both are data-adaptively defined and/or estimated. Statistical theory has advanced in sync with scientific computing and practical implementation is now possible for the resulting computationally challenging estimators.

We asked researchers currently engaged in cutting edge research on data-adaptive inferential methods to share their views with us. The result is a compelling collection of advances in statistical theory and practice.

The special issue consists of 19 articles. Its theoretical spectrum is wide. Semiparametric models and inference, empirical process theory and machine learning are the three major subfields explored in the articles. Across this special issue, the acceptation of the word inference covers the estimation of finite-dimensional parameters and the construction of confidence regions for them, the estimation of infinite-dimensional features (either as an endgame or as a means to an end); testing hypotheses (for the sake of making discoveries), identifying particular subgroups in a population, selecting (groups or clusters of) significant variables, comparing data-adaptive predictors. Cross-validating, decomposing a task in a series of sub-tasks (by partitioning or relying on a recurrence), fluctuating and weighting are the recurring technical concepts. Most articles are motivated by applications arising from medicine (analyzing neuroimages, comparing treatments, inferring optimal individualized treatment rules). The others address challenging theoretical questions. They shed light on delicate theoretical problems, offer guidance for better practice, open exciting new territories to explore.

We hope you will enjoy perusing the special issue and that it will serve as a useful pivot towards methods that can address new challenges in data science.

We wish to thank warmly the De Gruyter team for its unconditional scientific and technical support, in particular Theresa Haney, Spencer McGrath and John Wolfe.