Inference in high dimensional regression

organized by
Peter Buehlmann, Andrea Montanari, and Jonathan Taylor

Workshop Summary

Organizers.

Peter Bühlmann (ETH Zürich)
Andrea Montanari (Stanford University)
Jonathan Taylor (Stanford University)

The workshop lasted four days. We organized the following activities: (*i*) Ten hour-length presentations on recent progress in the field; (*ii*) One open-problem session to identify a list of mathematical challenges that are considered important across the community; (*iii*) Five parallel working groups addressing a subset of these questions.

The talks spurred lively discussions about several topics including the following:

New challenges: to statistical methodology coming from modern scientific research, most notably in biological sciences.

Modern research in these area typically involve a large number of independent labs, strongly incentivized to produce new discoveries and publications. On the other hand reproducibility and replicability are not equally incentivized and/or widespread. This has lead to a rapid increase of false discoveries, as witnessed even by the popular press. This effect could have been anticipated on a purely statistical basis.

A stimulating exchange developed around the methodological and mathematical questions posed by these developments.

Selective inference: (or 'inference on the selected') is a rapidly evolving line of research that address some of the challenges posed by high-dimensional inference. The idea is to perform inference on a small subset of parameters selected *after* analyzing the data, and still account in a rigorous way for the selection effect.

The scope and limitations of existing methods for selective inference were lively discussed. Also, the undelyng mathematical assumptions, and connection to classical inference were the object of many exchanges and clarifications.

The open problems discussion was also very productive, and led to formulating a selection of special topics addressed in the working groups. These were

(1) *Relations between selective and classical inference.* Classical inference is focusing on the parameter in the full model. While classical inference is plagued by the high-dimensionality and high multiplicity of the inference problem, selective inference is much less exposed to these issues. Under some unrealistic mathematical assumptions (nice design and a beta-min condition), selective and classical inference become comparable. For settings with strong correlations among the variables, both approaches

face some (largely unsolved) challenges: the selection scheme (in selective inference) exhibits "instability" while classical inference is plagued by (near) non-identifiability.

(2) *Confidence intervals on predicted values.* The recent advances in inference for high-dimensional regression address the problem of computing confidence intervals or p-values for low dimensional parameters. In many applications –however– the main application of high-dimensional regression is in fact to prediction. In this case, it would be very valuable to complement such predictions with a measure of uncertainty, e.g. a confidence interval.

    Unfortunately, existing tools do not seem to extend in any obvious way to this very important problem.

(3) *Confidence intervals for matrix completion.* In matrix completion, the data analyst is given a large data matrix with a number of missing entries. In many interesting applications (e.g. to collaborative filtering) it is indeed the case that the vast majority of entries is missing. In order to fill the missing entries, the assumption is made that the underlying –unknown– matrix has a low-rank structure.

    Substantial work has been devoted to methods for computing point estimates of the missing entries. In applications, it would be very interesting to compute confidence intervals as well. This requires developing distributional characterizations of standard matrix completion methods.

(4) *Inference for statistical network models.* Many modern datasets take the form of graphs or networks. There has been significant interest into extracting some underlying features from such graphs (latent features, highly connected subgraphs, and so on). Performing rigorous statistical inference on such data is also a broadly open direction.