High-dimensional data, random questions and random answers

Sara van de Geer

Statistics is crucial for dealing with the large amount of data available today. There are many machine learning algorithms around that help us to find specific topics with google search, new genes of that might be of interest, suspect aircrafts, or potential clients for our products. The task of the statistician is however also to provide estimates of the accuracy of the outcomes of an algorithm, or address the good-old question of significance of a finding. This question needs to be constantly reshaped. For example, due the pressure on scientists to publish, the number of journal papers increases but the number of insignificant - yet published as being significant - findings seems to increase even more. Or given a large set of genes, one is bound to find some of them important in one experiment, but no longer in the next one. We will consider some relatively new approaches to a statistical analysis of high-dimensional data. We then focus at a particular algorithm called the Lasso, which is much used for a selection of variables. Asking about significance of these is a random question. We will discuss the problems arising and then describe how the Lasso algorithm and its relatives can be invoked for obtaining random answers to nonrandom questions. We finalize with some random matrix theory.