Modern Statistical Methods (M24)

Sergio Bacallado

The remarkable development of computing power and other technology now allows scientists and businesses to routinely collect datasets of immense size and complexity. Most classical statistical methods were designed for situations with many observations and a few, carefully chosen variables. However, we now often gather data with a huge numbers of variables, in an attempt to capture as much information as we can about anything which might conceivably have an influence on the phenomenon of interest. This dramatic increase in the number variables makes modern datasets strikingly different, as well-established traditional methods perform either very poorly, or often do not work at all.

Developing methods that are able to extract meaningful information from these large and challenging datasets has recently been an area of intense research in statistics, machine learning and computer science. In this course, we will study some of the methods that have been developed to study such datasets. We aim to cover the following topics.

- Kernel machines: Ridge regression, the kernel trick, kernel ridge regression, the support vector machine, the hashing trick.
- Penalised regression: Model selection, the Lasso, variants of the Lasso.
- High-dimensional covariance matrices: non-asymptotic error bounds in the operator norm and the effective rank, the spiked covariance model, PCA and sparse PCA, the graphical Lasso.
- Multiple testing and high-dimensional inference: the closed testing procedure and the Benjamini–Hochberg procedure, the debiased Lasso.

Pre-requisites

Basic knowledge of statistics, probability, linear algebra and real analysis. Some background in optimisation would be helpful but is not essential.

Literature

1. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning.* 2nd edition. Springer, 2001.
2. P. Bühlmann, S. van de Geer, *Statistics for High-Dimensional Data.* Springer, 2011.
3. C. Giraud, *Introduction to High-Dimensional Statistics.* CRC Press, 2014.
4. T. Hastie, R. Tibshirani and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations.* CRC Press, 2015.
5. M. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge University Press, 2019.

Additional support

Four examples sheets will be provided and four associated examples classes will be given. There will be a revision class in the Easter Term.