Speaker: Alexandre d'ASPREMONT
CMAP Ecole Polytechnique

1. Introduction

A simple version of the variable selection problem will allow us here to discuss our research plan in more details. When data samples are independent, the most studied variable selection problem is least-squares regression with a constraint on the cardinality of the solution. In this case, we are given n data samples denoted by (X_1,Y_1), …, (X_n,Y_n) where X_i contains ``predictor vectors'' and Y_i contains the ``output'' or measured variables. We would like to predict Y using a linear predictor, and we want this predictor to be sparse, i.e. have only a few non-zero coefficients, to improve interpretability. This variable selection problem is numerically challenging (NP-hard in fact). A classic way of addressing this issue involves penalized regression techniques such as the LASSO which is formulated as minimizing a least-squares ``loss function'' with a penalty on the l_1 norm (sum of absolute values) of the prediction coefficients.

In the above context, the problem is to ascertain conditions on the data X, which guarantee that, if the true linear model generating the samples has some sparsity pattern, then the above Lasso approximation does recover this pattern. Recent sparse recovery results on penalized regression have received a considerable amount of attention. Many of these results hinge on ``sparse extremal eigenvalues'' of the covariance matrix of the data points, i.e. eigenvalues this symmetric matrix computed with a restriction on the number of nonzero coefficients in the corresponding eigenvector. For example, in the field of compressed sensing, sparse extremal eigenvalues were used by Candès and Tao in 2005 (under the alternate name of ``restricted isometry constants'') to prove perfect recovery in linear coding problems. The concept also serves to prove model selection consistency (which is about recovering the true model for the large sample case), and to bound optimal error rates in Lasso type regressions, when the true linear model is sparse. Thus, sparse extremal eigenvalues are the key quantities controlling theoretical properties of Lasso-type variable selection, i.e. the discovery of a ``true'' model, under the assumptions that this model is both linear and sparse. Bounding these eigenvalues in a computationally efficient manner mostly remains, however, an open problem.

2. Tractable Recovery Conditions

Several other conditions have been derived as substitutes to sparse eigenvalues to guarantee exact recovery of sparse solutions to regression or decoding problems. However, all of them so far are combinatorial in nature and cannot be tested in polynomial time (to be fair, hardness results are only known for a few of them). Like the restricted isometry property, these conditions can thus only be enforced with high probability on random data matrices, which severely limits their practical impact.

In this talk, we will start by a brief primer on compressed sensing. We will then focus on a particular type of recovery condition with an explicit geometrical interpretation and discuss both positive and negative results on the complexity of checking these properties. Finally, we will conclude with a list of open problems related to these questions.