BOSTON UNIVERSITY

Boston University Statistics and Probability Seminar Series

High Dimensional Sparse Regression and Structure Estimation

Shuheng Zhou
Computer Science Department
Carnegie Mellon University

Thursday, March 20, 2008, 4-5pm
Mathematics and Computer Science (MCS) Building, Room 149
111 Cummington Street, Boston

Tea and Cookies at 3:30pm in MCS 153

Abstract: Recent research has demonstrated that sparsity is a powerful technique in signal reconstruction and in statistical inference. Recent work shows that ℓ_1-regularized least squares regression can accurately estimate a sparse model from n noisy samples in p dimensions, even if p is much larger than n. My work in this area focuses on studying the role of sparsity in high dimensional regression when the original noisy samples are compressed, and on structure estimation in Gaussian graphical models when the graphs evolve over time.

In high-dimensional regression, the sparse object is a vector $\beta \in Y = X\beta + \epsilon$, where X is n by p matrix such that $n << p$, $\beta \in R^p$ and $\epsilon \in R^n$ consists of i.i.d. random noise. In the classic setting, this problem is ill-imposed for $p > n$ even for the case when $\epsilon = 0$. However, when the vector β is sparse, one can recover an empirical $\hat{\beta}$ that is consistent in terms of its support with true β. In joint work with John Lafferty and Larry Wasserman, we studied the regression problem under the setting that the original n input variables are compressed by a random Gaussian ensemble to m examples in p dimensions, where $m << n$ or p. A primary motivation for this compression procedure is to anonymize the data and preserve privacy by revealing little information about the original data. We established sufficient mutual incoherence conditions on X, under which a sparse linear model can be successfully recovered from the compressed data. We characterized the number of random projections that are required for ℓ_1-regularized compressed regression to identify the nonzero coefficients in the true model with probability approaching one. In addition, we showed that ℓ_1-regularized compressed regression asymptotically predicts as well as an oracle linear model, a property called "persistence". Finally, we established upper bounds on the mutual information between the compressed and uncompressed data that decay to zero.

For directions and maps, please see *http://math.bu.edu/research/statistics/statseminar.html.*