

STANFORD UNIVERSITY
DEPARTMENT OF STATISTICS
DEPARTMENTAL SEMINAR

4:15 p.m., Tuesday, February 4, 2003
Sequoia Hall Room 200
(Cookies at 3:45 in 1st Floor Lounge)

Eitan Greenshtein

Statistics Department, Haifa University

**Consistency in high dimensional linear predictor-selection
and the virtue of over parametrization**

Let $Z^i = (Y^i, X_1^i, \dots, X_m^i), i = 1, \dots, n$, be i.i.d. random vectors, Z^i are distributed F , where F is unknown and belongs to a family of distributions. It is desired to predict Y by $\sum \beta_j X_j$, where $(\beta_1, \dots, \beta_m) \in B^n$, under a prediction loss. Suppose that $m = n^\alpha, \alpha > 1$, i.e., there are much more explanatory variables than observations. We study the following asymptotics. How 'large' may the set B^n be, so that selecting nearly the best from it under F , based on Z^1, \dots, Z^n , is possible. 'Large' is in terms of the maximal number of non-zero coefficients among the members of B^n or the l_1 radius of B^n . Sharp bounds for orders of magnitudes are given under certain (parametric and non-parametric) assumptions about the family of distributions to which F belongs. Algorithmic complexity of such consistent procedures is also studied.

This is a joint work with Ya'acov Ritov.