

STANFORD UNIVERSITY
DEPARTMENT OF STATISTICS
DEPARTMENTAL SEMINAR

4:15 p.m., Tuesday, April 8, 2008
Sequoia Hall Room 200
(Cookies at 3:45 in 1st Floor Lounge)

Art B. Owen
Department of Statistics
Stanford University

Transposably invariant sample reuse methods

Sample reuse methods like the bootstrap and cross-validation are widely used in statistics and machine learning. They have some face value validity and they don't depend on strong model assumptions. These methods depend on repeating or omitting cases, while keeping all the variables in those cases. But for many data sets, it is not obvious whether the rows are cases and columns are variables, or vice versa. For example, with movie ratings organized by movie and customer, both movie and customer IDs can be thought of as variables.

This talk looks at bootstrap and cross-validation methods that treat rows and columns of the matrix symmetrically. We get the same answer on X as on X' . McCullagh has proved that no exact bootstrap exists in a certain framework of this type (crossed random effects). We show that a method based on resampling both rows and columns of the data matrix tracks the true error, for some simple statistics applied to large data matrices.

Similarly we look at a method of cross-validation that leaves out blocks of the data matrix, generalizing a proposal due to Gabriel that is used in the crop science literature. We find empirically that this approach provides a good way to choose the number of terms in a truncated SVD model or a non-negative matrix factorization. We also apply some recent results in random matrix theory to the truncated SVD case.

This work is joint with Patrick Perry.