

The University of Chicago

Department of Statistics

Seminar

David W. Scott

Department of Statistics, Rice University

“Remarks on Handling Outliers in Data and Regression”

Monday, October 14, 2002 at 4:00 PM
133 Eckhart Hall, 5734 S. University Avenue

ABSTRACT

Maximum likelihood estimation is universal but sensitive to model misspecification and outliers. In this talk, I describe an alternative approach to formulating some robust estimates. Robust estimation provides a powerful solution to practical problems in applied statistics. Simple tasks such as data cleaning may be prohibitively expensive with large datasets. Our techniques can also handle the difficult situation where a dataset contains large clusters of outliers. For example, a multi-component normal mixture model may be estimated with the expectation that several components will identify groups of outliers. We examine this latter idea when we deliberately fit a mixture model with fewer components than required to pick up all outliers.

In our formulation, maximum likelihood is replaced by a data-based minimum-distance criterion. The usual M-estimator specification of the shape and scale of the influence function is replaced by a single choice of a distribution function for the data. This idea is illustrated for several common choices of data, including Gaussian.

Similar ideas have application in regression. I am interested not only in the case of outlier-contaminated regression but also in the case of mixtures of regressions, with outliers. Examples of our approach will be given.