

**Stony Brook University
The Graduate School**

Doctoral Defense Announcement

Abstract

A Mutli-scale, Geometric Algorithm for Non-parametric Data Exploration with an
Application to Genomic Data

By

Joe McQuown

This thesis presents an efficient and adaptive multi-scale algorithm for analyzing measurement data, composed of two categories: a regular set of measurements that can be described by means of a dominant geometry, and a set of "outliers", i.e., measurements that deviate from this underlying geometry. The algorithm uses a stopping-time construction in order to identify local regions of different sizes where the data is concentrated around local lines (or d-planes) and controls the local percentages of putative outliers outside those regions. Thus, it is able to efficiently construct a description of the dominant "geometry" in terms of a curve (or d-dimensional graph). Using local geometric properties, it then detects the outliers. The approach need not make any assumption about the distributional properties of the noise and it exhibits robustness against noise and outliers. Furthermore, the speed of our algorithm is linear in the size of the data, a great advantage for large data sets. It can handle high-dimensional data without a blow-up of computational expense. Genomic expression data is an application that can be assayed quite well within this framework. This thesis explores the experimental results of such phenomena.

Date: October 5, 2007

Time: 11:00 AM

Place: Math Tower

Program: AMS

Dissertation Advisor: James Glimm