Research Interaction Team Plan, Fall '02

Team Name: Statistics of Large Cross-Classified Datasets

Team Director: Eric Slud, × 5-5469, evs@math.umd.edu

Research Focus: Large datasets arise naturally in many areas of science, government, and business. Typically, as the size of a dataset gets large, the complexity of questions which one addresses with it also increases. Such problems range from Semiparametric Statistical Inference to Order-selection problems in regression and time series, to Classification and Clustering as in the Microarray Data problem-area in which I ran the AMSC seminar last fall and an RIT last spring. This is unlike the formal setting of most mathematical statistics, in which parameter-dimension is fixed and sample size increases to ∞. The contrast suggests the need for a new Asymptotics which explicitly recognizes the growth of the parameter- space of a probability model as a function of the size n of the dataset.

In Fall 2002, I will run a 'research interaction' seminar on mathematical/statistical topics in Large Cross-Classified Datasets, which broadly encompasses the overlap of my students' thesis projects and most of my own current research interests.

Desirable background: To benefit from this research activity, a graduate student should have completed Stat 700 and at least one of Stat 740, 741, 750, or 770, and have some familiarity with Statistical Computing at the level of Stat 430 (SAS programming) or Stat 798C (Splus and SAS).

Undergraduate Prerequisites: An interested undergraduate should have had at least one course in Mathematical Statistics (e.g. Stat 401 or 420) and considerable experience — either in courses or projects — with numerical computing or data analysis.

Graduate Program: Graduate students will be involved in collaboratively formulating the research questions, in searching out available literature references and publicly available software, and in implementing model-fits and goodness-of-fit evaluations.

Undergraduate Program: Undergraduate students will be involved primarily in experimenting with ideas for generating and comparing the validity and stability of clusters from fitted models on public data (particularly in DNA microarrays).

Meeting Schedule: We will run a bi-weekly seminar in the fall of 2002. Participants will be asked to pursue topics of individual interest, which may involve reading papers on specific topics in specific problem-areas (e.g., clustering of genes in cDNA Microarrays, classification of tongue images during speech based on Principal Components, model-fitting in a specific dataset, semiparametric models in large clinical trials, etc.) and to present brief reports or problem-statements. Our seminar will attempt to abstract asymptotic research questions from these topics, with the overall theme of complexity increasing with data-size.