Large Data Sets: Computation and Structure
Friday, November 13, 2009
Center for Theoretical & Mathematical Sciences, Duke University

9:00 – 9:45:

Approximate Sparse Recovery: Optimizing Time and Measurements

Anna Gilbert, Department of Mathematics, University of Michigan

A Euclidean approximate sparse recovery system consists of parameters epsilon, k, d, an n-by-d measurement matrix, Phi, and a decoding algorithm, D. Given a vector, x, the system approximates x by y=D(Phi x), which must satisfy ||y – x||_2 <= (1+epsilon)||yopt_k – x||_2, where yopt_k denotes the optimal k-term approximation to x. For each vector x, the system must succeed with probability at least ¾. Many previous works have addressed this problem. Among the goals are minimizing the number n of measurements and the runtime of the decoding algorithm, D.
In this result, we give a system with n=O(k log(d/k)/epsilon) measurements—matching a lower bound, up to constant factors—and decode time k*poly(log(d)/epsilon), matching a lower bound up to factors of log(d)/epsilon.

9:50 – 10:35

Object Oriented Data Analysis

Steve Marron, Department of Statistics, University of North Carolina at Chapel Hill

Object Oriented Data Analysis is the statistical analysis of populations of complex objects. In the special case of Functional Data Analysis, these data objects are curves, where standard Euclidean approaches, such as principal components analysis, have been very successful. Recent developments in medical image analysis motivate the statistical analysis of populations of more complex data objects which are elements of mildly non-Euclidean spaces, such as Lie Groups and Symmetric Spaces, or of strongly non-Euclidean spaces, such as spaces of tree-structured data objects. These new contexts for Object Oriented Data Analysis create several potentially large new interfaces between mathematics and statistics. Even in situations where Euclidean analysis makes sense, there are statistical challenges because of the High Dimension Low Sample Size problem, which motivates a new type of asymptotics leading to non-standard mathematical statistics.

10:55 – 11:30

Challenges of Visualizing Large Data

Rachael Brady, Department of Computer Science, Duke University

Visualization is used for both data exploration, a highly interactive task, and data presentation such as animations and still images. Historically, large data visualizations are non-interactive due to the slower rendering times and data fetching times. This limits one's ability to probe and query one's data. In an effort to make large data visualization interactive, techniques such as wire-frame rendering during camera movements and level of detail culling have improved rendering times, but one can always create a situation

where more data is within the field of view than can be rendered at interactive rates (faster than 10 Hz). Similarly, data pre-caching has proven successful when we know the structure of the data. Finding solutions for unstructured, time varying, streaming data is still an active area of research. This talk will review current techniques in large data visualization.

11:30 – 12:15

Eigenvalues of Large Dimensional Random Matrices

Jack W. Silverstein, Department of Mathematics, North Carolina State University

The talk will outline recent work on spectral properties of random matrices. One topic to be covered concerns the properties of individual eigenvalues of a class of matrices of sample covariance type. It is defined as $B_n = (1/N)T_n^{1/2} X_n X^*_n T_n^{1/2}$ where $X_n = (X_{ij})$ is n x N with i.i.d. complex standardized entries, and $T_n^{1/2}$ is a Hermitian square root of the nonnegative definite Hermitian matrix T_n. This matrix can be viewed as the sample covariance matrix of N i.i.d. samples of the n dimensional random vector $T_n^{1/2}(X_n)_{\cdot 1}$.
It is known that if $n/N \to c > 0$ and the empirical distribution function (e.d.f.) of the eigenvalues of T_n converge as $n \to \infty$, then the e.d.f. of the eigenvalues of B_n converges a.s. to a nonrandom limit. This result is relevant in situations in multivariate analysis where the vector dimension is large, but the number of samples to adequately approximate the population matrix (required in standard statistical procedures) cannot be attained.
Consider a finite number of eigenvalues of T_n which are outside the support of its limiting spectral e.d.f. This is referred to as a “spiked population model". Results are obtained for the limiting behavior of those eigenvalues of B_n which correspond to the “spiked" eigenvalues of T_n. An application is given to the detection problem in array signal processing: determining the number of sources (presumed large) impinging on a bank of sensors in a noise filled environment (joint work with Jinho Baik at University of Michigan, and with Raj Rao at MIT).
Another class of matrices of the form $C_n = (1/N)(R_n + \sigma X_n)(R_n + \sigma X_n)^*$ where X_n is as in B_n, $\sigma > 0$, and R_n is n x N random, independent of X_n with the spectral e.d.f. of $(1/N)R_n R_n$ converging to a nonrandom limit. These matrices model situations, such as in array signal processing, where information is contained in the sampling of the vectors $R_{\cdot 1} \cdots R_{\cdot N}$, but the received vector is contaminated by additive noise (the columns of σX_n).The e.d.f. of the eigenvalues of C_n also converges a.s. as $n \to \sigma$ (with $n/N \to c > 0$). Properites of the limiting distribution will be outlined. (joint work with Brent Dozier).
A third class to be discussed generalizes B_n. It is of the form $D_n = (1/N)T_n^{1/2} X_n S_n X^*_n T_n^{1/2}$ where S_n is N x N nonnegative definite Hermitian, and appears in the modeling of MIMO (multiple-input-multiple-out) systems in wireless communications (joint work with Debashis Paul at UC Davis).

1:15 – 2:00

Nonparametric Statistical Manifold Learning and Compressive Sensing

Lawrence Carin, Department of Electrical and Computer Engineering, Duke University

Nonparametric methods are considered for learning the statistics of high-dimensional data that reside on manifolds. The models infer the latent dimension of the manifold, and also constitute a statistical mapping from the high-dimensional data to an associated low-dimensional representation (an embedding). The model is generative, and therefore one may also synthesize high-dimensional data via a low-dimensional parametrization. The statistical framework is also employed to perform compressive-sensing inversion for signals that live on manifolds.

2:00 – 2:45

Large Matrices Beyond Singular Value Decomposition

Andrea Montanari, Departments of Electrical Engineering and Statistics, Stanford University

A number of data sets are naturally described in matrix form. Examples range from microarrays to collaborative filtering data. In many of these examples, singular value decomposition (SVD) provides an efficient way to construct a low-range approximation thus achieving a large dimensionality reduction. SVD is also an important tool in the design of approximate linear algebra algorithms for massive data sets. It is a recent discovery that – for 'generic' matrices – SVD is sub-optimal, and can be significantly improved upon. There has been considerable progress on this topic over the last year, partly spurred by interest in the Netflix challenge. I will overview this progress.

3:10 – 3:55

Algorithmic and Statistical Perspectives on Large-Scale Data Analysis

Michael Mahoney, Department of Mathematics, Stanford University

In recent years, motivated in large part by large-scale scientific and Internet data analysis problems, traditional ideas from statistics (such as leverage and influence and regularization) have begun to interact in increasingly sophisticated and fruitful ways with ideas in computer science having to do with the worst-case analysis of graph algorithms. As a by-product, we have seen the development of algorithms that simultaneously have strong worst-case algorithmic guarantees; have a solid statistical underpinning; and are useful for very applied data analysis problems. After providing an overview of these ideas, two specific examples will be described in detail. The first example involves choosing a set of "informative" or "influential" DNA SNPs from a dataset such as that provided by the HapMap project in order to perform reconstruction or classification or prediction. The second example involves taking advantage of regularization implicit in algorithms for intractable graph problems to attempt to find clusters or communities that are commonly-hypothesized to exist in large social and information networks. In both cases, understanding the statistical perspective on the worst-case algorithms allows one to make strong and sometimes surprising claims in the application area of interest.

3:55 – 4:40

Finding Significant Large-Average Submatrices in High Dimensional Data

Andrew Nobel, Department of Statistics, University of North Carolina at Chapel Hill

Exploratory analysis of high dimensional data often begins with independent clustering of samples and variables, yielding a partition of the available data matrix into disjoint row-column blocks (submatrices). Of particular interest in practice are submatrices whose entries are large on average. In conjunction with clinical and functional annotation, large average submatrices are frequently the starting point for subsequent analyses, such as the identification of genetic pathways and new disease subtypes in the study of gene expression data.

This talk describes a simple algorithm, belonging to the general category of biclustering methods, for identifying large average submatrices in high dimensional data. Like other biclustering methods, the algorithm improves on independent sample variable clustering in several respects: the submatrices it identifies can overlap and they need not cover the entire data matrix (features that better reflect the underlying structure of many problems), and the inclusion of samples and variables in a submatrix does not depend on their expression values outside the submatrix. The algorithm seeks to maximize a simple measure of statistical significance, which also provides an objective basis for comparing and selecting among submatrices of different sizes and average intensities. I will discuss the applications of the algorithm to a recent gene-expression based cancer study, and will provide a detailed comparison of its performance with several other biclustering method, including its application to semi-supervised classification.