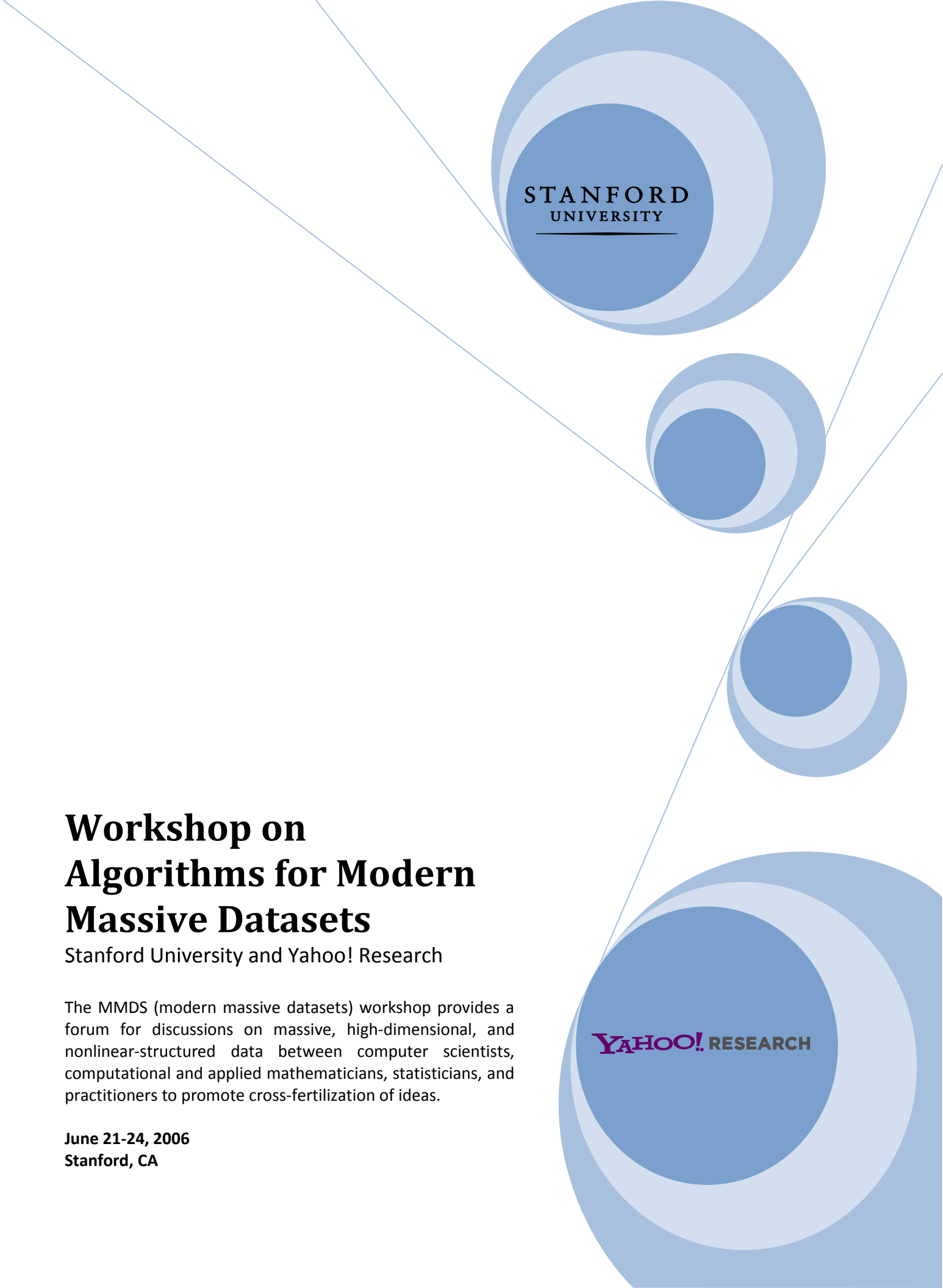


Workshop on Algorithms for Modern Massive Datasets

Stanford University and Yahoo! Research

The MMDS (modern massive datasets) workshop provides a forum for discussions on massive, high-dimensional, and nonlinear-structured data between computer scientists, computational and applied mathematicians, statisticians, and practitioners to promote cross-fertilization of ideas.

June 21-24, 2006
Stanford, CA

Wednesday: June 21, 2006

Linear Algebra Basics

Time	Event
9:00-10:00	**Breakfast/Registration**
10:00-11:00	Tutorial: Ravi Kannan *Sampling in large matrices*
11:00-11:30	Santosh Vempala *Sampling methods for low rank approximation*
11:30-12:00	Petros Drineas *Subspace sampling and relative error matrix approximation*
12:00-1:30	**Lunch (on your own)**
1:30-2:30	Tutorial: Dianne O'Leary *Matrix factorizations for information retrieval*
2:30-3:00	Pete Stewart *Sparse reduced rank approximations to sparse matrices*
3:00-3:30	Haesun Park *Adaptive discriminant analysis by regularized minimum squared errors*
3:30-4:00	**Break**
4:00-4:30	Michael Mahoney *CUR matrix decompositions for improved data analysis*
4:30-5:00	Daniel Spielman *Fast algorithms for graph partitioning, sparsification, and solving SDD systems*
5:00-5:30	Anna Gilbert/Martin Strauss *List decoding of noisy Reed-Muller-like codes*
5:30-6:00	Bob Plemmons *Low-rank nonnegative factorizations for spectral imaging applications*
6:00-6:30	Art Owen *A hybrid of multivariate regression and factor analysis*
6:30-8:00	**Reception (at New Guinea Garden)**

Thursday: June 22, 2006

Industrial Applications and Sampling Methods

Time	Event
9:00-10:00	Tutorial: Prabhakar Raghavan *The changing face of web search*
10:00-10:30	Tong Zhang *Statistical ranking problem*
10:30-11:00	**Break**
11:00-11:30	Michael Berry *Text mining approaches for email surveillance*
11:30-12:00	Hongyaun Zha *Incorporating query difference for learning retrieval functions*
12:00-12:30	Trevor Hastie/Ping Li *Efficient L^2 and L^1 dimension reduction in massive databases*
12:30-2:00	**Lunch (on your own)**
2:00-3:00	Tutorial: Muthu Muthukrishnan *An algorithmicist's look at compressed sensing problems*
3:00-3:30	Inderjit Dhillon *Kernel learning with Bregman matrix divergences*
3:30-4:00	Bruce Hendrickson *Latent semantic analysis and Fiedler retrieval*
4:00-4:30	**Break**
4:30-5:00	Piotr Indyk *Near-optimal hashing for approximate nearest neighbor problems*
5:00-5:30	Moses Charikar *Compact representations for data*
5:30-6:00	Sudipto Guha *At the confluence of streams, order, information, and signals*
6:00-6:30	Frank McSherry *Preserving privacy in large-scale data analysis*

Friday: June 23, 2006

Kernel Learning and Applications

Time	Event
9:00-10:00	Tutorial: Dimitris Achlioptas *Applications of random matrices in spectral computations and machine learning*
10:00-10:30	Tomaso Poggio *Learning from data: Theory, engineering applications and (a bit of) neuroscience*
10:30-11:00	**Break**
11:00-11:30	Stephen Smale *Geometry and topology of data*
11:30-12:00	Gunnar Carlsson *Algebraic topology and analysis of high dimensional data*
12:00-12:30	Vin de Silva *Point-cloud topology via harmonic forms*
12:30-2:00	**Lunch (on your own)**
2:00-2:30	Dan Boley *Fast clustering leads to fast support vector machine training and more*
2:30-3:00	Chris Ding *On the equivalence of (semi-)nonnegative matrix factorization and k-means*
3:00-3:30	Al Inselberg *Parallel coordinates: visualization & data mining for high dimensional datasets*
3:30-4:00	Joel Tropp *One sketch for all: a sublinear approximation scheme for heavy hitters*
4:00-4:30	**Break**
4:30-5:00	David Donoho *Needles in haystacks: Finding relevant variables among many useless ones*
5:00-5:30	Rob Tibshirani *Prediction by supervised principal components*
5:30-6:00	Tao Yang/Apostolos Gerasoulis *Page ranking for large-scale internet search: Ask.com's experiences*
6:00-8:00	**Poster Session/Banquet (at Wallenberg Hall)**

Saturday: June 24, 2006

Tensor-Based Data Applications

Time	Event
9:00-10:00	**Breakfast**
10:00-11:00	Tutorial: Lek-Heng Lim *Tensors, symmetric tensors and nonnegative tensors in data analysis*
11:00-11:30	Eugene Tyrtyshnikov *Tensor compression of petabyte-size data*
11:30-12:00	Lieven De Lathauwer *The decomposition of a tensor in a sum of rank (R_1, R_2, R_3) terms*
12:00-1:30	**Lunch**
1:30-2:00	Orly Alter *Matrix and tensor computations for reconstructing cellular pathways*
2:00-2:30	Shmuel Friedland *Tensors: Ranks and approximations*
2:30-3:00	Tammy Kolda *Multilinear algebra for analyzing data with multiple linkages*
3:00-3:30	Lars Eldén *Computing the best rank-(R_1, R_2, R_3) approximation of a tensor*
3:30-4:00	**Break**
4:00-4:30	Liqun Qi *Eigenvalues of tensors and their applications*
4:30-5:00	Brett Bader *Analysis of latent relationships in semantic graphs using DEDICOM*
5:00-5:30	Alex Vasilescu *Multilinear (tensor) algebraic framework for computer vision and graphics*
5:30-6:00	Rasmus Bro *Multi-way analysis of bioinformatic data*
~~**6:00-6:30**~~	~~Pierre Comon~~ ~~*Independent component analysis viewed as a tensor decomposition*~~
6:30-8:00	**Closing Reception**

Wednesday: June 21, 2006 (Linear Algebra Basics)

10:00-11:00

SAMPLING IN LARGE MATRICES

In many application areas, the input data matrix is too large to be handled by traditional Linear Algebra algorithms. A number of recent results address this. One proves that a small random sample of columns and a small random sample of rows are sufficient to approximate any matrix provided the sampling probabilities are judiciously chosen. Also, from a good low-rank approximation (LRA) to the sampled sub-matrices, one can derive a good LRA to the whole matrix. These approximations are suitable for a host of numerical applications which go under the name of Principal Component Analysis.

We also discuss applications of these methods to a broad class of combinatorial problems of which a typical one is to maximize a low-degree n-variable polynomial over the corners of the unit n-cube. We describe an efficient algorithm for finding a rough analog of LRA to a tensor which then helps us estimate the maximum.

Ravi Kannan
Department of Computer Science
Yale University

11:00-11:30

SAMPLING METHODS FOR LOW RANK APPROXIMATION

It is well-known that the sum of the first k terms of the Singular Value Decomposition of a matrix gives its optimal rank-k approximation (under the 2-norm or the Frobenius norm). Is computing the SVD essential for low-rank approximation?

It was shown in 1998 (FKV) that a small sample of rows chosen according to their squared lengths can be used to compute a low-rank approximation whose error is at most the best possible plus a term that depends on the sample size and the norm of the original matrix. This leads to a randomized algorithm to compute such an approximation in time linear in the number of non-zero entries of the matrix.

In this talk, we discuss this approach and present two ways of generalizing it, called adaptive sampling and volume sampling. Together, they show that a sample of $O(k/c + k \log k)$ rows of any matrix can generate a rank-k matrix whose error is at most $(1+c)$ times that of the optimum rank-k matrix. The algorithm based on this computes such a multiplicative low-rank approximation, also in linear time.

Santosh Vempala
Department of Mathematics
Massachusetts Institute of Technology

11:30-12:00

SUBSPACE SAMPLING AND RELATIVE ERROR MATRIX APPROXIMATION

In this talk we will focus on low-rank matrix decompositions that are explicitly expressed in terms of a small number of actual columns and/or rows of a matrix, as opposed to, e.g., linear combinations of up to all the columns and/or rows of the matrix, such as provided by truncating the singular value decomposition (SVD). Motivations for studying such matrix decompositions include (i) the efficient decomposition of large low-rank matrices that posses additional structure such as sparsity or non-negativity, (ii) expressing Gram matrices in statistical learning theory in terms of a small number of actual data points, (iii) the improved data interpretability that such decompositions provide for datasets in the Internet domain, the social sciences, biology, chemistry, medicine, etc., and (iv) the efficient computation of low-rank matrix approximations in space-constrained settings.

We shall discuss two such decompositions. Given an m-by-n matrix A, the first one approximates A by the product CX, where C consists of a few columns of A, and X is a coefficient matrix. The second one is of the form CUR, where C consists of a few columns of A, R consists of a few rows of A, and U is a carefully

constructed, constant-sized matrix. Previous such matrix decompositions only guaranteed additive error approximations. Our algorithms for constructing *C*, *X*, *U*, and *R* take low polynomial time and return approximations that are almost the best possible in a relative error sense. They employ the subspace sampling technique; in particular, rows and columns of *A* are sampled with probabilities that depend on the lengths of the rows of the top few singular vectors of *A*.

Petros Drineas
Department of Computer Science
Rensselaer Polytechnic Institute

1:30-2:30

MATRIX FACTORIZATIONS FOR INFORMATION RETRIEVAL

This tutorial will survey a wide variety of matrix decompositions that have been proposed for use in information retrieval. Comparisons among the methods and their applicability to massive data sets will be emphasized.

Dianne O'Leary
Department of Computer Science and Institute for Advanced Computer Studies
University of Maryland, College Park

2:30-3:00

SPARSE REDUCED-RANK APPROXIMATIONS TO SPARSE MATRICES

The purpose of this talk is to describe an algorithm for producing a reduced rank approximation to a sparse *m x n* matrix *A* in which the factors are either small or sparse. When the sparseness of the factorization is not an issue, there are two decompositions that give reduced-rank decompositions: the singular value decomposition and the pivoted QR decomposition. This talk focuses on the latter.

We will show how a variant of the pivoted Gram--Schmidt algorithm can produce approximations to A of the form *XS*, where *X* is an *m x p* matrix consisting of columns of *X* and *S* is a *p x n* dense matrix. Thus if *p* is sufficiently small, this factorization is suitable for the case where *m* is much larger than *n*. The same algorithm applied to the transpose of *A* an approximation of the form *TY*, where *Y* is an *p x n* matrix consisting of rows of *A* and *T* is an dense *m x p* matrix. If both approximations have been computed, they may be combined into an approximation of the form *XUY*, where *U* is a *p x p* matrix--an approximation suitable for the case where both *m* and *n* are large.

The approximations can be computed a column (or row) at a time. At each stage the error in the approximation can be computed essentially for free, thus giving a rigorous criterion for stopping the process.

G.W. Stewart
Department of Computer Science
University of Maryland, College Park

3:00-4:00

ADAPTIVE DISCRIMINANT ANALYSIS BY REGULARIZED MINIMUM SQUARED ERRORS

Fisher's discriminant analysis for binary class dimension reduction and the binary classifier design based on the minimum squared error formulation (MSE) have been widely utilized for handling high-dimensional clustered data sets. As the data sets get modified from incorporating new data points and deleting obsolete data points, there is a need to develop efficient updating and downdating algorithms for these methods to avoid expensive recomputation of the solution from scratch.

In this talk, an efficient algorithm for adaptive linear and nonlinear kernel discriminant analysis based on regularized MSE, called KDA/RMSE, is introduced. In adaptive KDA/RMSE, updating and downdating of the computationally expensive eigenvalue decomposition (EVD) or singular value decomposition (SVD) is approximated by updating and downdating of the QR decomposition achieving an order of magnitude speed up. This fast algorithm for adaptive kernelized discriminant analysis is designed by utilizing regularization and some relationship between linear and nonlinear discriminant analysis for dimension reduction and the MSE for classifier design. An efficient algorithm for computing leave-one-out cross validation is also introduced by utilizing downdating of KDA/RMSE.

Haesun Park
College of Computing
Georgia Institute of Technology

4:00-4:30

CUR MATRIX DECOMPOSITIONS FOR IMPROVED DATA ANALYSIS

Much recent work in theoretical computer science, numerical linear algebra, machine learning, and data analysis has considered low-rank matrix decompositions of the following form: given an *m*-by-*n* matrix *A*, decompose it as a product of three matrices, *C*, *U*, and *R*, where *C* consists of a few columns of *A*, *R* consists of a few rows of *A*, and *U* is a small carefully constructed matrix that guarantees that the product CUR is "close," either provably or empirically, to *A*. Applications of such decompositions include the computation of compressed “sketches” for large matrices in a pass-efficient manner, matrix reconstruction, speeding up kernel-based statistical learning computations, sparsity-preservation in low-rank approximations, and improved interpretability of data analysis methods. We shall review the motivation for these decompositions, discuss various choices for the matrices *C*, *U*, and *R* that are appropriate in different application domains, and then discuss the application of *CUR* decompositions to three diverse application domains: human genetics, medical imaging, and internet recommendation systems. In each of these applications, the columns and rows have a natural interpretation in terms of the processes generating the data, and CUR decompositions can be used to solve reconstruction, clustering, classification, or prediction problems in each of these areas.

Michael Mahoney
Yahoo! Research

4:30-5:00

FAST ALGORITHMS FOR GRAPH PARTITIONING AND SPARSIFICATION, AND THE SOLUTION OF SYMMETRIC DIAGONALLY-DOMINANT LINEAR SYSTEMS

Motivated by the problem of solving systems of linear equations, we develop nearly-linear time algorithms for partitioning graphs and for building strong sparsifiers. The graph partitioning algorithm is the first provably fast graph partitioning algorithm that finds sparse cuts of nearly optimal balance. It is based on an algorithm for finding small clusters in massive graphs that runs in time proportional to the size of the cluster it outputs. Using the graph partitioning algorithm and random sampling, we

show how to spectrally approximate any graph by a sparse subgraph.

We then apply these algorithms, along with constructions of low-stretch spanning trees, to optimize a preconditioner construction introduced by Vaidya. We thereby obtain a nearly-linear time algorithm for approximately solving systems of linear equations of the form $Ax=b$, where A is symmetric and diagonally dominant.

Joint work with Shang-Hua Teng.

Daniel Spielman
Department of Computer Science
Yale University

5:00-5:30

LIST DECODING OF NOISY REED-MULLER-LIKE CODES

Coding theory has played a central role in the development of computer science. One critical point of interaction is decoding error-correcting codes. First- and second-order Reed-Muller (RM(1) and RM(2), respectively) codes are two fundamental error-correcting codes which arise in communication as well as in probabilistically checkable proofs and learning. (The theory of first-order Reed-Muller codes is also known as Fourier analysis on the Boolean cube.)

In this paper, we take the first steps toward extending the decoding tools of RM(1) into the realm of quadratic binary codes. We show how to recover a substantial subcode of RM(2) called a Kerdock code, in the presence of significant noise. The Kerdock codes are a well-studied family of codes for coding theory, radar signaling, and spread spectrum communication. Our result is a list-decoding result for Kerdock codes which is roughly analogous to that of RM(1). In addition, we present a new algorithmic characterization of Kerdock codes that we hope will be more useful to the algorithmic (and coding theory) community than the classic descriptions.

Joint work with Robert Calderbank.

Anna Gilbert/Martin Strauss
Department of Mathematics
University of Michigan, Ann Arbor

5:30-6:00

LOW-RANK NONNEGATIVE FACTORIZATIONS FOR SPECTRAL IMAGING APPLICATIONS

Nonnegative matrix factorization (NMF) is a technique with numerous applications that has been used extensively in recent years for approximating high dimensional data where the data are composed of nonnegative entries. The process involves low-rank nonnegative approximate matrix factorizations to allow the user to work with reduced dimensional models and to facilitate more efficient statistical classification of data. In spectral imaging, we will describe our use of NMF to obtain spectral signatures for space object identification using a point source, imaged at different spectral frequencies. In other applications, we are also concerned with multispectral data to be processed for extended object identification and analysis, and for processing stacks of correlated 2D images. We wish to find a nonnegative factorization of a multiway array, i.e., a tensor. The non-negative tensor factorization (NTF) problem has a unique set of difficulties for our applications that we will discuss.

Joint work with Christos Boutsidis, Misha Kilmer, and Paul Pauca.

Bob Plemmons
Department of Mathematics and
Computer Science
Wake Forest University

6:00-6:30

A HYBRID OF MULTIVARIATE REGRESSION AND FACTOR ANALYSIS

Multivariate regression is a useful tool for identifying age related genes from microarray data. Introducing latent variables into the regression provides a diagnostic tool for spotting observations that are outliers and for identifying missing variables. An example of the former is a patient whose kidney appears, from gene expression, to be that of a much younger patient. The latent variables alone yield a factor analysis model while the measured variables alone are a multivariate regression. The criterion for the combined model is a Frobenious norm on the residual. The set of minimizers can be characterized in terms of a singular value decomposition. A power iteration appears to be faster the SVD in some real data.

Joint work with Stuart Kim and Jacob Zahn.

Art B. Owen
Department of Statistics
Stanford University

Thursday: June 22, 2006 (Industrial Applications and Sampling Methods)

9:00-10:00

THE CHANGING FACE OF WEB SEARCH

Web search has come to dominate our consciousness as a convenience we take for granted, as a medium for connecting advertisers and buyers, and as a fast-growing revenue source for the companies that provide this service. Following a brief overview of the state of the art and how we got there, this talk covers a spectrum of technical challenges arising in web search - ranging from social search to auction design and incentive mechanisms.

Prabhakar Raghavan
Yahoo! Research

10:00-10:30

STATISTICAL RANKING PROBLEM

Ranking has important applications in web-search, advertising, user recommender systems, etc, and is essential for internet companies such as Yahoo.

In this talk, I will first go over some formulations and methods of ranking in the statistical literature. Then I will focus on a formulation suitable for web search, and talk about training relevance models based on DCG (discounted cumulated gain) optimization. Under this metric, the system output quality is naturally determined by the performance near the top of its rank-list. I will mainly discuss various theoretical issues in this learning problem. They reflect real issues encountered at Yahoo when building a machine learning based web-search ranking system.

Joint work with David Cossock at Yahoo.

Tong Zhang
Yahoo! Research

11:00-11:30

TEXT MINING APPROACHES FOR EMAIL SURVEILLANCE

Automated approaches for the identification and clustering of semantic features or topics are highly desired for text mining applications. Using a low rank non-negative matrix factorization algorithm to retain natural data non-negativity, we eliminate the need to use subtractive basis vector and encoding calculations present in techniques such as principal component analysis for semantic feature abstraction. Existing techniques for non-negative matrix factorization are briefly reviewed and a new approach for non-negative matrix factorization is presented. A demonstration of the use of this approach for topic (or discussion) detection and tracking is presented using the Enron Email Collection.

Joint work with Murray Browne.

Michael Berry
Department of Computer Science
University of Tennessee, Knoxville

11:30-12:00

INCORPORATING QUERY DIFFERENCE FOR LEARNING RETRIEVAL FUNCTIONS

In this talk we discuss information retrieval methods that aim at serving a diverse stream of user queries. We propose methods that emphasize the importance of taking into consideration of query difference in learning effective retrieval functions. We formulate the problem as a multi-task learning problem using a risk minimization framework. In particular, we show how to calibrate the empirical risk to incorporate query difference in terms of introducing nuisance parameters in the statistical models, and we also propose an alternating optimization method to simultaneously learn the retrieval function and the nuisance parameters. We work out the details for both L^1 and L^2 regularization

cases. We illustrate the effectiveness of the proposed methods using modeling data extracted from a commercial search engine.

Joint work with Zhaohui Zheng, Haoying Fu and Gordon Sun.

Hongyuan Zha
Department of Computer Science and Engineering
Pennsylvania State University

12:00-12:30

EFFICIENT L^2 AND L^1 DIMENSION REDUCTION IN MASSIVE DATABASES

Sampling and sketching (including random projections) are two basic strategies for dimension reduction. For dimension reduction in L^2 using random projections, we show that the accuracy can be improved by taking advantage of the marginal information; and the efficiency can be improved dramatically using a very sparse random projection matrix. For dimension reduction in L^1, we prove an analog of the JL lemma using nonlinear estimators. Previous known results have proved that no analog of the JL lemma exists for L^1 if restricted to linear estimators. As a competitor to random projections, a sketch-based sampling algorithm is very efficient and highly flexible in sparse data. This sampling algorithm generates random samples online from sketches. These various methods are useful for mining huge databases (e.g., association rules), distance-based clustering, nearest neighbor searching, as well as efficiently computing the SVM kernels.

Joint work with Kenneth Church.

Trevor Hastie/Ping Li
Department of Statistics
Stanford University

2:00-3:00

AN ALGORITHMICIST'S LOOK AT COMPRESSED SENSING PROBLEMS

Recently Donoho introduced the problem of Compressed Sensing and proved an interesting result that there exists a set of roughly k log n vectors of dimension n each such that any n dimensional vector can be recovered to nearly best accuracy possible using k terms, from only the inner products of the vector with the k log n vectors. This result has partial precursors in algorithms research and many postcursors. I will provide an overview of Compressed Sensing, its pre- and postcursors, from an algorithmicist's point of view.

Muthu Muthukrishnan
Google Inc.

3:00-3:30

KERNEL LEARNING WITH BREGMAN MATRIX DIVERGENCES

Bregman divergences offer a rich variety of distortion measures that are applicable to varied applications in data mining and machine learning. Popular Bregman vector divergences are the squared Euclidean distance, relative entropy and the Itakura-Saito distance. These divergence measures can be extended to Hermitian matrices, yielding as special cases, the squared Frobenius distance, von Neumann divergence and the Burg divergence. In this talk, I will talk about the kernel learning problem that uses these Bregman matrix divergences in a natural way. A key advantage is in obtaining efficient update algorithms for learning low-rank kernel matrices.

Joint work with Brian Kulis, Matyas Sustik and Joel Tropp.

Inderjit Dhillon
Department of Computer Sciences
University of Texas, Austin

3:30-4:00

LATENT SEMANTIC ANALYSIS AND FIEDLER RETRIEVAL

Latent semantic analysis (LSA) is a method for information retrieval and processing which is based upon the singular value decomposition. It has a geometric interpretation in which objects (e.g. documents and keywords) are placed in a low-dimensional geometric space. In this talk, we derive a new algebraic/geometric method for placing objects in space to facilitate information analysis. Our approach uses an alternative motivation, and reduces to the computation of eigenvectors of Laplacian matrices. We show that our method, which we call Fiedler retrieval, is closely related to LSA, and essentially equivalent for particular choices of scaling parameters. We then show that Fiedler retrieval supports a number of generalizations and extensions that existing LSA approaches cannot handle, including unified text and link analysis.

Bruce Hendrickson
Sandia National Laboratories

4:30-5:00

NEAR-OPTIMAL HASHING ALGORITHM FOR THE APPROXIMATE NEAREST NEIGHBOR PROBLEM

The nearest neighbor problem is defined as follows: given a set of n data points in a d-dimensional space, construct a data structure which, given a query point q, returns the data point closest to the query.

The problem is of importance in multiple areas of computer science. Nearest neighbor algorithms can be used for: classification (in machine learning), similarity search (in biological, text or visual databases), data quantization (in signal processing and compression), etc. Unfortunately, classic algorithms for geometric search problems, such as kd-trees, do not scale well as the dimension increases.

In recent years, there has been extensive research done on *approximate* variants of the nearest neighbor problem. One of the algorithms, based on the idea of Locality-Sensitive Hashing (LSH), has been successfully used in several applied scenarios.

In this talk I will review that work, as well as describe recent developments in this area. In particular, I will show a new LSH-based algorithm, whose running time significantly improves over the earlier bounds. The running time of the new algorithm is provably close to the best possible in the class of hashing-based algorithms.

Piotr Indyk
Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology

5:00-5:30

COMPACT REPRESENTATIONS FOR DATA

Several algorithmic techniques have been devised recently to deal with large volumes of data. At the heart of many of these techniques are ingenious schemes to represent data compactly. This talk will present several examples of such compact representation schemes. Some of these are inspired by techniques devised in the field of approximation algorithms for “rounding” solutions of LP and SDP relaxations. We will also see how such compact representation schemes lead to efficient, one-pass algorithms for processing large volumes of data (streaming algorithms).

Moses Charikar
Department of Computer Science
Princeton University

5:30-6:00

AT THE CONFLUENCE OF STREAMS; ORDER, INFORMATION AND SIGNALS

In this talk we will explore several new directions in data streams, particularly at the intersection of streams, information and signal processing. One of the most important features of a data stream is the meaning or the semantics of the elements streaming by. They typically either define a distribution over a domain or represent a signal. We will investigate a few problems corresponding to these two scenarios. The first set of problems will consider modeling a distribution as a stream of updates, and investigate space efficient algorithms for computing various information theoretic properties, divergences etc. The second set of problems would consider wavelet processing on data streams.

Sudipto Guha
Department of Computer and Information Science
University of Pennsylvania

6:00-6:30

PRESERVING PRIVACY IN LARGE-SCALE DATA ANALYSIS

We discuss privacy issues associated with the analysis of sensitive data sets, motivated by the general inaccessibility of sensitive data due to privacy concerns. We present a definition of "differential privacy," in which each user is assured that no conclusion reached by the data analyst is much more likely than if than user had not participated in the study. The unequivocal nature of the statement, quantified over all possible conclusions and all prior knowledge of the analyst, gives very strong privacy guarantees. Additionally, we show detail how many common analyses can be faithfully implemented in a framework that yields differential privacy, and discuss the opportunities for incorporating new analyses.

Joint work with Cynthia Dwork.

Frank McSherry
Microsoft Research

Friday: June 23, 2006 (Kernel Learning and Applications)

9:00-10:00

APPLICATIONS OF RANDOM MATRICES IN SPECTRAL COMPUTATIONS AND MACHINE LEARNING

We will see how carefully crafted random matrices can be used to enhance a wide variety of computational tasks, including: dimensionality reduction, spectral computations, and kernel methods in machine learning. Several examples will be considered, including the following two.

Imagine that we want to compute the top few eigenvectors of a matrix *A*, hoping to extract the "important" features in *A*. We will prove that either this is not worth doing, or that we can begin by randomly throwing away a large fraction of *A*'s entries.

A famous result of Johnson and Lindenstrauss asserts that any set of n points in *d*-dimensional Euclidean space can be embedded in *k*- dimensional Euclidean space, where $k = O(\log n)$, such that all pairwise distances are preserved with arbitrarily good accuracy. We prove that to construct such an embedding it suffices to multiply the $n \times d$ matrix of points with a random $d \times k$ matrix, whose entries are set to ±1 independently, with equal probability.

Dimitris Achlioptas
Department of Computer Science
University of California, Santa Cruz

10:00-10:30

LEARNING FROM DATA: THEORY, ENGINEERING APPLICATIONS AND (A BIT OF) NEUROSCIENCE

The problem of learning is one of the main gateways to making intelligent machines and to understanding how the brain works. In this talk I will give a brief overview of recent work on learning theory and on general conditions for predictivity of an algorithm and ultimately for science itself. I will then show a few examples of our efforts in developing machines that learn for applications such as visual recognition and graphics and speech synthesis. Finally, I will sketch a new theory of the ventral stream of the visual cortex, describing how the brain may learn to recognize objects, and show that the resulting model is capable of performing recognition on datasets of complex images at the level of human performance in rapid categorization tasks. The model performs as well or better than most state-of-art computer vision systems in recognition of complex images.

Tomaso Poggio
Center for Biological and Computational Learning,
Computer Science and Artificial Intelligence Laboratory, and McGovern Institute for Brain Research
Massachusetts Institute of Technology

11:00-11:30

GEOMETRY AND TOPOLOGY OF DATA

The following questions are discussed. Suppose points are drawn at random from a submanifold *M* of Euclidean space. How may one reconstruct the topology and the geometry of *M*? and with what confidence?

Stephen Smale
Department of Mathematics
University of California, Berkeley and
Toyota Technological Institute,
University of Chicago

11:30-12:00

ALGEBRAIC TOPOLOGY AND ANALYSIS OF HIGH DIMENSIONAL DATA

In recent years techniques have been developed for evaluating Betti numbers of spaces given only a finite but large set of points (a "point cloud") sampled from the space. Such techniques can be applied in statistical settings where the data being considered is high dimensional (and hence cannot be visualized) and non-linear in nature. This talk will describe the techniques required to obtain the Betti numbers, as well as discuss some actual examples where they have been applied.

Gunnar Carlsson
Department of Mathematics and
Institute for Computational and Mathematical Engineering
Stanford University

12:00-12:30

POINT-CLOUD TOPOLOGY VIA HARMONIC FORMS

Point-cloud data sets are discrete objects which vary continuously, whereas topological spaces are continuous objects which vary discretely. It follows that the discrete invariants of classical algebraic topology (such as homology) are not immediately useful for point-cloud topology. It is always necessary to make those invariants continuous in some way. One very successful paradigm for this is the theory of persistent homology. I will discuss an alternative paradigm which uses harmonic forms defined by discrete Laplacian operators.

Vin de Silva
Department of Mathematics and
Computer Science
Pomona College

2:00-2:30

FAST CLUSTERING LEADS TO FAST SUPPORT VECTOR MACHINE TRAINING AND MORE

We discuss how a fast deterministic clustering algorithm enables a variety of applications, including the training of support vector machines and the computation of a low memory factored representation which admits the application of many standard algorithms on massive datasets. Some theoretical and experimental properties of the clustering methods and of the resulting support vector machines will be provided. We will show that the clustering method compares favorably with *k*-means and that the resulting SVM compares favorably with that computed in the usual way directly from the data.

Daniel Boley
Department of Computer Science
University of Minnesota, Twin Cities

2:30-3:00

ON THE EQUIVALENCE OF (SEMI-) NONNEGATIVE MATRIX FACTORIZATION (NMF) AND K-MEANS/SPECTRAL CLUSTERING

We show that the objective of NMF $X=FG'$ is equivalent to K-means clustering: G is cluster indicator and F contains cluster centroids. This can be generalized to semi-NMF where X, F contain mixed-sign data, while G being nonnegative. We further propose convex-NMF by restricting F be convex combinations of data points, ensuring F to be meaningful cluster centroids. We also show that the symmetric NMF $W=HH'$ is equivalent to Kernel K-means clustering and the Laplacian-based spectral clustering. All these follow by rewriting K-means objective as a trace of quadratic function whose continuous relaxation solution are given by PCA components. We emphasize orthogonality and nonnegativity in matrix based clustering. We derive the updating algorithms for semi-NMF, convex-NMF, symmetric-NMF and prove their convergence. We

present experiments on face images, newsgroups, web log and text data to show the effectiveness of these NMF based clustering.

Based on joint work with Xiaofeng He, Horst Simon, Tao Li and Michael Jordan.

Chris Ding
Lawrence Berkeley National Laboratory

3:00-3:30

PARALLEL COORDINATES: VISUALIZATION & DATA MINING FOR HIGH DIMENSIONAL DATASETS

A dataset with *M* items has *2M* subsets anyone of which may be the one we really want. With a good data display our fantastic pattern-recognition ability can not only cut great swaths searching through this combinatorial explosion but also extract insights from the visual patterns. These are the core reasons for data visualization. With Parallel Coordinates (abbr. k-coords) the search for multivariate relations in high dimensional datasets is transformed into a 2-D pattern recognition problem. After a short overview of k-coords, guidelines and strategies for knowledge discovery are illustrated on different real datasets, one with 400 variables from a manufacturing process. A geometric classification algorithm based on k-coords is presented and applied to complex datasets. It has low computational complexity providing the classification rule explicitly and visually. The minimal set of variables required to state the rule is found and ordered by their predictive value. A visual economic model of a real country is constructed and analyzed to illustrate how multivariate relations can be modeled by means of hypersurfaces, understanding trade-offs, sensitivities and interelations. Parallel coordinates is also used in collision avoidance algorithms for air traffic control.

Al Inselberg
School of Mathematical Sciences
Tel Aviv University

3:30-4:00

ONE SKETCH FOR ALL: A SUBLINEAR APPROXIMATION SCHEME FOR HEAVY HITTERS

The heavy hitters problem elicits a list of the m largest-magnitude components in a signal of length d. Although this problem is easy when the signal is presented explicitly, it becomes much more challenging in the setting of streaming data, where the signal is presented implicitly as a sequence of additive updates. One approach maintains a small sketch of the data that can be used to approximate the heavy hitters quickly. In previous work, this sketch is essentially a random linear projection of the data that fails with small probability for each signal. It is often desirable that the sketch succeed simultaneously for ALL signals from a given class, a requirement that may be called uniform heavy hitters. It arises, for example, when the signal is queried a large number of times or when the signal updates are stochastically dependent.

This talk describes a random linear sketch for uniform heavy hitters that succeeds with high probability. The recovery algorithm produces a list of heavy hitters that approximates the input signal with an l_2 error that is optimal, except for an additive term that depends on the optimal l_1 error and a controllable parameter ε. The recovery algorithm requires space m*poly(log(d)/ε) and time m^2*poly(log(d)/ε) to produce the list of heavy hitters. In the turnstile model, the time per update is m*poly(log(d)/ε). Up to logarithmic factors, the performance of this algorithm is the best possible with respect to several resources.

Joint work with Anna Gilbert, Martin Strauss, and Roman Vershynin.

Joel Tropp
Department of Mathematics
University of Michigan

4:30-5:00

NEEDLES IN HAYSTACKS: FINDING THE RELEVANT VARIABLES AMONG MANY USELESS ONES

Many modern datasets have more variables than observations. Among those variables it is thought that only a small number are really useful for linear prediction, but we don't know which. We could proceed by combinatorial search to try all possible subsets and build prediction models with those, but that's computationally very heavy.

In my talk I will describe some fun results showing that often one can get just as good results without all-subsets search. For example, often one can penalize models by the l_1-norm on the coefficients and find the same model that combinatorial optimization would have found. There are maybe 100 papers related to this over the last 18 months, and I'll try to give a flavor of what's being done and why.

David Donoho
Department of Statistics and
Institute for Computational and Mathematical Engineering
Stanford University

5:00-5:30

PREDICTION BY SUPERVISED PRINCIPAL COMPONENTS

In regression problems where the number of predictors greatly exceeds the number of observations, conventional regression techniques may produce unsatisfactory results. We describe a technique called supervised principal components that can be applied to this type of problem. Supervised principal components is similar to conventional principal components analysis except that it uses a subset of the predictors that are selected based on their association with the outcome. Supervised principal components can be applied to regression and generalized regression problems such as survival analysis. It compares favorably to other techniques for this type of problem, and can also account for the effects of other covariates and help identify which predictor variables are most important. We also provide asymptotic consistency results to help support our empirical findings. These methods could become important tools for DNA microarray data, where they may be used to more accurately diagnose and treat cancer.

Joint work with Eric Bair, Trevor Hastie, and Debashis Paul.

Robert Tibshirani
Department of Statistics
Stanford University

5:30-6:00

PAGE RANKING FOR LARGE-SCALE INTERNET SEARCH: ASK.COM'S EXPERIENCES

Ask.com has been developing a comprehensive suite of search and question-answering technology and differentiated products to help users to find what they are looking for. This talk gives an overview of Ask.com's search engine and describes many of challenges faced in seeking relevant answers from billions of documents for tens of millions of users everyday. Particularly, this talk will discuss our ExpertRank algorithm which provides relevant search results by identifying topics and authoritative sites on the Web through query-specific clustering and expert analysis.

Joint work with Apostolos Gerasoulis, Wei Wang, and other Ask.com engineers.

Apostolos Gerasoulis/Tao Yang
Ask.com and
Department of Computer Science,
Rutgers Unversity, and
University of California, Santa Barbara

Saturday: June 24, 2006 (Tensor-Based Data Applications)

10:00-11:00

MULTILINEAR ALGEBRA IN DATA ANALYSIS: TENSORS, SYMMETRIC TENSORS, AND NONNEGATIVE TENSORS

A simple recipe for creating multilinear models is to take a bilinear model, say, a term-by-document matrix, and include one or more additional "modes" to get, say, a term-by-document-by-URL tensor (which may come from, for example, a collection of term-by-document matrices across different URLs). Now, in analogy with matrix methods, one finds a low-rank approximation to the tensor in question and hope that it reveals interesting features about the data. Here, rank may mean the outer-product rank or the multilinear rank (the respective models are sometimes known as CANDECOMP/PARAFAC and the Tucker model) or indeed any other notions of tensor rank.

An immediate question is: why should this work? Many problems arise once we go from order-2 (matrices) to higher orders (tensors). Not least among which is the fact that the problem of finding a best rank-r approximation for tensors often do not even have a solution. Furthermore, there seems to be no proper statistical foundation or such models --- what is one really doing when decomposing a tensor of data values into a sum of outer product of vectors? why should these vectors tell us anything about the data?

In this talk, we will discuss these and other related questions. We will also address similar issues for two classes of tensors that are important in data-analytic applications: symmetric tensors (Independent Component Analysis) and nonnegative tensors (Nonnegative Tensor Factorization). We will show how these topics have recently been enriched by ideas from the ancient (eg. Cayley's formula for the hyperdeterminant of a 2x2x2 tensor) to the present (eg. Donoho's work in l^0-approximations). Furthermore, we shall see how some existing computational, mathematical, and statistical tools could be employed to study multilinear models. Examples include: using Matroid Theory to prove uniqueness of tensor decompositions; using properties of higher secant varieties of Segre/Veronese variety to shed light on the aforementioned non-existence of best rank-r approximations; using SDP to solve relaxed convex versions of the best rank-r tensor approximation problem; using the Log Linear Model, or more generally, Graphical Models and Bayesian Networks to provide a statistical foundation.

Time permitting, we will also discuss Multilinear Spectral Theory and show how the eigenvalues of symmetric tensors may be used to obtain basic results in Spectral Hypergraph Theory and how one could prove a Perron-Frobenius Theorem for irreducible non-negative tensors.

Lek-Heng Lim
Institute for Computational and Mathematical Engineering
Stanford University

11:00-11:30

TENSOR COMPRESSION OF PETABYTE-SIZE DATA

In some applications we are interested to work with huge amounts of data for which standard methods of numerical analysis cannot be applied. For instance, consider a 3-dimensional array of size of several thousand in every dimension. Is it possible to work with this size array any efficiently? In the talk, we are going to show that the answer is "yes" - however, under certain assumptions on the data.

The main assumption is that the data admits a sufficiently accurate low-tensor-rank approximation. Then, we present a special cross-approximation technique using a relatively small amount of the entries of the given array and providing the so-called

Tucker decomposition with the complexity linear in one-dimensional size of the array. The technique generalizes the low-rank approximation technique for matrices to 3-dimensional arrays. The Tucker core can be further approximated by a trilinear (canonical) decomposition. Development of good algorithms solving the latter problem is still topical. To this end, we present a new algorithm that seems to be competitive with the best known algorithms.

Joint work with I. Oseledets and D. Savostianov.

Eugene Tyrtyshnikov
Institute of Numerical Mathematics
Russian Academy of Sciences

11:30-12:00

THE DECOMPOSITION OF A TENSOR IN A SUM OF RANK (R_1, R_2, R_3) TERMS

The Parallel Factor decomposition (PARAFAC) in multilinear algebra decomposes a higher-order tensor in a sum of rank-1 terms. The decomposition can be essentially unique without imposing orthogonality constraints on the rank-1 terms. Tucker's decomposition is a second way to generalize the Singular Value Decomposition (SVD) of matrices. Here, the transformation matrices are orthogonal. Their columns correspond to the directions of extremal oriented energy in column space, row space, etc. However, the decomposition does in general not reduce the tensor to diagonal form. In this talk we introduce a new tensor decomposition, of which Tucker's decomposition and PARAFAC are special cases. The decomposition thus generalizes the SVD of matrices in two ways. The result sheds new light on very fundamental aspects of tensor algebra, including tensor rank.

Lieven De Lathauwer
Lab ETIS
CNRS,
Ecole Nationale Superieure d'Electronique et
de ses Applications, and
University of Cergy-Pontoise

1:30-2:00

MATRIX AND TENSOR COMPUTATIONS FOR RECONSTRUCTING THE PATHWAYS OF A CELLULAR SYSTEM FROM GENOME-SCALE SIGNALS

DNA microarrays make it possible to record, for the first time, the complete genomic signals, such as mRNA expression and DNA-bound proteins' occupancy levels, that are generated and sensed by cellular systems. The underlying genome-scale networks of relations among all genes of the cellular systems can be computed from these signals. These relations among the activities of genes, not only the activities of the genes alone, are known to be pathway-dependent, i.e., conditioned by the biological and experimental settings in which they are observed.

I will describe the use of the matrix eigenvalue decomposition (EVD) and pseudoinverse projection and a tensor higher-order EVD (HOEVD) in reconstructing the pathways, or genome-scale pathway-dependent relations among the genes of a cellular system, from nondirectional networks of correlations, which are computed from measured genomic signals and tabulated in symmetric matrices [Alter & Golub, PNAS 2005].

EVD formulates a genes x genes network, which is computed from a “data” signal, as a linear superposition of genes x genes decorrelated and decoupled rank-1 subnetworks. Significant EVD subnetworks might represent functionally independent pathways that are manifest in the data signal.

The integrative pseudoinverse projection of a network, computed from a data signal, onto a designated “basis” signal approximates the network as a linear superposition of only the subnetworks that are common to both signals, i.e., pseudoinverse projection filters off the network the subnetworks that are exclusive to the data signal. The pseudoinverse-projected network simulates observation of only the pathways that are manifest

under both sets of the biological and experimental conditions where the data and basis signals are measured.

I will define a comparative HOEVD, that formulates a series of networks computed from a series of signals as linear superpositions of decorrelated rank-1 subnetworks and the rank-2 couplings among these subnetworks. Significant HOEVD subnetworks and couplings might represent independent pathways or transitions among them common to all or exclusive to a subset of the signals.

I will illustrate the EVD, pseudoinverse projection and HOEVD of genome- scale networks with analyses of mRNA expression data from the yeast *Saccharomyces cerevisiae* during its cell cycle and DNA-binding data of yeast transcription factors that are involved in cell cycle, development and biosynthesis programs. Boolean functions of the discretized subnetworks and couplings highlight known and novel differential, i.e., pathway-dependent relations between genes.

Orly Alter
Department of Biomedical Engineering and
Institute for Cellular and Molecular Biology
University of Texas, Austin

2:00-2:30

TENSORS: RANKS AND APPROXIMATIONS

The theory of k (≥ 3) tensors became of great interest in recent applications. In particular, the rank of a tensor and an approximation of a given tensor by low rank tensors, emerged as the most outstanding problems in data processing with many parameters.

In this talk I will point out some directions of study of these problems using analytical methods: as algebraic geometry combined with matrix theory, and numerical methods: as SVD for matrices combined with Newton algorithms. Some very preliminary results can be found in Chapter 5 of my notes: http://www2.math.uic.edu/~friedlan/matrlecsum.pdf

Shmuel Friedland
Department of Mathematics,
Statistics and Computer Science
University of Illinois, Chicago

2:30-3:00

MULTILINEAR ALGEBRA FOR ANALYZING DATA WITH MULTIPLE LINKAGES

Link analysis typically focuses on a single type of connection, e.g., two journal papers are linked because they are written by the same author. However, often we want to analyze data that has multiple linkages between objects, e.g., two papers may have the same keywords and one may cite the other. The goal of this paper is to show that multilinear algebra provides a tool for multi-link analysis. We analyze five years of publication data from journals published by the Society for Industrial and Applied Mathematics. We explore how papers can be grouped in the context of multiple link types using a tensor to represent all the links between them. A PARAFAC decomposition on the resulting tensor yields information similar to the SVD decomposition of a standard adjacency matrix. We show how the PARAFAC decomposition can be used to understand the structure of the document space and define paper-paper similarities based on multiple linkages. Examples are presented where the decomposed tensor data is used to find papers similar to a body of work (e.g., related by topic or similar to a particular author's papers), find related authors using linkages other than explicit co-authorship or citations, distinguish between papers written by different authors with the same name, and predict the journal in which a paper was published.

Joint work with Daniel M. Dunlavy and W. Philip Kegelmeyer.

Tammy Kolda
Sandia National Laboratories

3:00-3:30

COMPUTING THE BEST RANK-(R_1, R_2, R_3) APPROXIMATION OF A TENSOR

We investigate various properties of the best rank-(R_1, R_2, R_3) approximation of a tensor, and their implications in the development of algorithms for computing the approximation. In particular we discuss the Grassmann-Newton method for solving the problem, and its relation to the Grassmann-Rayleigh quotient iteration of de Lathauwer et al.

Joint work with Berkant Savas.

Lars Eldén
Department of Mathematics
Linköping University

4:00-4:30

EIGENVALUES OF TENSORS AND THEIR APPLICATIONS

Motivated by the positive definiteness issue in automatic control, I defined eigenvalues, *E*-eigenvalues, *H*-eigenvalues and *Z*-eigenvalues for a real completely symmetric tensor. An *m*th order *n*-dimensional real completely symmetric tensor has $n(m-1)^{n-1}$ eigenvalues and at most and at most $\sum_{k=0}^{n-1}(m-1)^k$ E-eigenvalues. They are roots of two one-dimensional polynomials, which are called the characteristic polynomial and the *E*-characteristic polynomial respectively. The sum of the eigenvalues is equal to the trace of the tensor, multiplied with $(m-1)^{n-1}$. The eigenvalues obey some Gerschgorin-type theorems, while the *E*-eigenvalues are invariant under orthogonal transformation. The latter property indicates that *E*-eigenvalues are invariants of practical tensors in physics and mechanics. Real eigenvalues (*E*-eigenvalues) with real eigenvectors (*E*-eigenvectors) are called *H*-eigenvalues (Z-eigenvalues). *Z*-eigenvalues always exist. When *m* is even, *H*-eigenvalues always exist. A real completely symmetric tensor is positive definite if and only if all of its *H*-eigenvalues (*Z*-eigenvalues) are positive. Based on the resultant theory, *H*-eigenvalue and *Z*-eigenvalue methods are proposed for judging positive definiteness of an even degree multivariate form when m and n are small. Independently, Lek-Heng Lim has also explored the definitions and applications of *H*-eigenvalues and *Z*-eigenvalues. In particular, he proposed a multilinear generalization of the Perron-Frobenius theorem based upon *H*-eigenvalues.

Liqun Qi
Department of Mathematics
City University of Hong Kong

4:30-5:00

ANALYSIS OF LATENT RELATIONSHIPS IN SEMANTIC GRAPHS USING DEDICOM

This presentation introduces the DEDICOM (DEcomposition into Directional COMponents) family of models from the psychometrics literature for analyzing asymmetric relationships in data and applies it to new applications in data mining, in particular social network analysis. In this work we present an improved algorithm for computing the 3-way DEDICOM model with modifications that make it possible to handle large, sparse data sets. We demonstrate the capabilities of DEDICOM as a new tool for reporting latent relationships in large semantic graphs, which we represent by an adjacency tensor. For an application we consider the email communications of former Enron employees that were made public, and posted to the web, by the U.S. Federal Energy Regulatory Commission during its investigation of Enron. We represent the Enron email network as a directed graph with edges labeled by time. Using the three-way DEDICOM model on this data, we show unique latent relationships that exist between types of employees and study their communication patterns over time.

Joint work with Richard Harshman and Tamara Kolda.

Brett Bader
Sandia National Laboratories

5:00-5:30

MULTILINEAR (TENSOR) ALGEBRAIC FRAMEWORK FOR COMPUTER VISION AND GRAPHICS

Principal components analysis (PCA) is one of the most valuable results from applied linear algebra. It is used ubiquitously in all forms of data analysis -- in data mining, biometrics, psychometrics, chemometrics, bioinformatics, computer vision, computer graphics, etc. This is because it is a simple, non-parametric method for extracting relevant information through the dimensionality reduction of high-dimensional datasets in order to reveal hidden underlying variables. PCA is a linear method, however, and as such it has severe limitations when applied to real world data. We are addressing this shortcoming via multilinear algebra, the algebra of higher order tensors.

In the context of computer vision and graphics, we deal with natural images which are the consequence of multiple factors related to scene structure, illumination, and imaging. Multilinear algebra offers a potent mathematical framework for explicitly dealing with multifactor image datasets. I will present two multilinear models that learn (nonlinear) manifold representations of image ensembles in which the multiple constituent factors (or modes) are disentangled and analyzed explicitly. Our nonlinear models are computed via a tensor decomposition, known as the M-mode SVD, which is an extension to tensors of the conventional matrix singular value decomposition (SVD), or through a generalization of conventional (linear) ICA called Multilinear Independent Components Analysis (MICA).

I will demonstrate the potency of our novel statistical learning approach in the context of facial image biometrics, where the relevant factors include different facial geometries, expressions, lighting conditions, and viewpoints. When applied to the difficult problem of automated face recognition, our multilinear representations, called TensorFaces (M-mode PCA) and Independent TensorFaces (MICA), yields significantly improved recognition rates relative to the standard PCA and ICA approaches. Recognition is achieved with a novel Multilinear Projection Operator.

In computer graphics, our image-based rendering technique, called TensorTextures, is a multilinear generative model that, from a sparse set of example images of a surface, learns the interaction between viewpoint, illumination and geometry, which determines surface appearance, including complex details such as self-occlusion and self shadowing. Our tensor algebraic framework is also applicable to human motion data in order to extract "human motion signatures" that are useful in graphical animation synthesis and motion recognition.

Alex Vasilescu
Media Laboratory
Massachusetts Institute of Technology

5:30-6:00

MULTI-WAY ANALYSIS OF BIOINFORMATIC DATA

In many metabonomic (bioinformatic) investigations, the data suffer from specific problems such as baseline issues, shifts in time etc. Additionally, such data are often characterized by being overwhelmingly information rich which is a problem in mathematical and statistical terms but even more so from a scientific point of view because the desire is often to obtain valid causal explanations for complex biological problems. We will show how tailored multi-way analysis tools can help in analysis of such complex data and show several examples of different origin.

Rasmus Bro
Chemometrics Group,
Department of Dairy and Food Science
The Royal Veterinary and Agricultural University

~~6:00-6:30~~

~~INDEPENDENT COMPONENT ANALYSIS VIEWED AS A TENSOR DECOMPOSITION~~

The problem of identifying linear mixtures of independent random variables only from outputs can be traced back to 1953 with the works of Darmois or Skitovich. They pointed out that when data are non Gaussian, a lot more can be said about the mixture. In practice, Blind Identification of linear mixtures is useful especially in Factor Analysis, Signal & Image Processing, Digital Communications, Biomedical, and Complexity Theory.

Harshman and Carroll provided independently in the seventies numerical algorithms to decompose a data record stored in a 3-way array into elementary arrays, each representing the contribution of a single underlying factor. The main difference with the well known Principal Component Analysis is that the mixture is not imposed to be a unitary matrix. The Parafac ALS algorithm, widely used since that time, theoretically does not converge for topological reasons, but yields very usable results after a finite number of iterations, under rank conditions however.

Independently, the problem of Blind Source Separation (BSS) arose around 1985 and was solved with the help of High-Order Statistics (HOS), which are actually tensors. It gave rapidly birth to the more general problem of Independent Component Analysis (ICA) in 1991. ICA is a tool that can be used to extract Factors when the physical diversity does not allow to store directly and efficiently the data in tensor format, in other words when the data cannot be uniquely decomposed directly. The problem then consists of decomposing a data cumulant tensor, of arbitrarily chosen order, into a linear combination of rank-one symmetric tensors. For this purpose, several algorithms can be thought of.

Pierre Comon
Laboratoire I3S
CNRS and University of Nice, Sophia-Antipolis

Posters

COLLECTIVE SAMPLING AND ANALYSIS OF HIGH ORDER TENSORS FOR CHATROOM COMMUNICATIONS

This work investigates the accuracy and efficiency tradeoffs between centralized and collective algorithms for (i) sampling and (ii) n-way data analysis techniques in multidimensional stream data, such as Internet chatroom communications. Its contributions are threefold. First, we use the Kolmogorov-Smirnov goodness-of-fit test to demonstrate that statistical differences between real data obtained by collective sampling in time dimension from multiple servers and that of obtained from a single server are insignificant. Second, we show using the real data that collective data analysis of 3-way data arrays (users x keywords x time) is more efficient than centralized algorithms with respect to both space and computational cost. Third, we examine the sensitivity of collective constructions and analysis of high order data arrays to the choice of server selection and sampling window size. We construct 4-way datasets (users x keywords x time x servers) and analyze them to show the impact of server and window size selections on the results.

Evrim Acar
Department of Computer Science
Rensselaer Polytechnic Institute

CONVEX OPTIMIZATION TECHNIQUES FOR LARGE-SCALE COVARIANCE SELECTION

We consider the problem of fitting a large-scale covariance matrix to multivariate Gaussian data in such a way that the inverse is sparse, thus providing model selection. Beginning with a dense empirical covariance matrix, we solve a maximum likelihood problem with an l_1-norm penalty term added to encourage sparsity in the inverse. For models with tens of nodes, the resulting problem can be solved using standard interior-point algorithms for convex optimization, but these methods scale poorly with problem size. We present two new algorithms aimed at solving problems with a thousand nodes. The first, based on Nesterov's first-order algorithm, yields a rigorous complexity estimate for the problem, with a much better dependence on problem size than interior-point methods. Our second algorithm uses block coordinate descent, updating row/columns of the covariance matrix sequentially. Experiments with genomic data show that our method is able to uncover biologically interpretable connections among genes.

Onureena Banerjee
Department of Electrical Engineering and Computer Sciences
University of California at Berkeley

PALSIR: A NEW APPROACH TO NONNEGATIVE TENSOR FACTORIZATION

PALSIR (Projected Alternating Least Squares with Initialization and Regularization) is a new approach to Nonnegative Tensor Factorization (NTF). PALSIR is designed to decompose a nonnegative (D_1 x D_2 x D_3) tensor T into a sum of 'k' nonnegative (D_1 x D_2 x D_3) rank-1 tensors T_i each of which can be written as the outer product of three nonnegative vectors: x_i, y_i and z_i of dimensions (D_1 x 1), (D_2 x 1) and (D_3 x 1) respectively, $i=1:k$. PALSIR consists of the following phases: i) initialization of 2 out of the 3 vector groups x_i's, y_i's, z_i's; ii) an iterative tri-alternating procedure where, at each stage, two of the three groups of vectors remain fixed and a nonnegative solution is computed with respect to the third group. Each of these stages requires solving {D_1, D_2 or D_3} constrained least squares (LS) problems. These are solved by first solving a linear ill-posed inverse problem in the least squares sense, using Tikhonov regularization and then projecting the solutions back to the feasible solution - instead of nonnegative LS.

We applied PALSIR on a 3d cube of images of the space shuttle Columbia taken by an Air Force telescope system (in its last orbit before disintegration upon re-entry in February 2003) in order to identify a parts-based representation of the image collection. PALSIR appears to be competitive compared to other available NTF algorithms both in terms of computational cost and approximation accuracy.

Joint work with E. Gallopoulos, P. Zhang and R.J. Plemmons.

Christos Boutsidis
Computer Engineering and
Informatics Department
University of Patras, Greece

SUPPORT VECTOR MACHINE TRAINING WITH A FEW REPRESENTATIVES

We study algorithms that speed up the training process of support vector machines by using only a relatively small number of representatives. We show how kernel K-means usually can be expected to yield a good set of representatives. The effectiveness is demonstrated with experiments on some real datasets and a theoretical PAC-style generalization bound.

Dongwei Cao
Department of Computer Science
University of Minnesota at Twin Cities

FAST RELATIVE LOW-RANK MATRIX APPROXIMATION

Low-rank approximation using Singular Value Decomposition (SVD) is computationally expensive for certain applications involving large matrices.

Frieze-Kannan-Vempala (FKV) showed that from a small sample of rows of the given matrix we can compute a low-rank approximation, which is (in expectation) only an additive error worse than the "best" low-rank approximation. This can be converted into a randomized algorithm to compute this additive low-rank approximation in "linear" (in the number of non-zero entries) time. But in general, their additive error can be unbounded compared to the error of the "best" low-rank approximation.

Using some generalizations of the FKV sampling scheme, we strengthen their results for low-rank approximation within multiplicative error. Based on this we get a randomized algorithm to find such an approximation in "linear" time.

Joint work with Luis Rademacher, Santosh Vempala and Grant Wang.

Amit Deshpande
Department of Mathematics
Massachusetts Institute of Technology

IMPOSING INDEPENDENCE CONSTRAINTS TO THE CP-MODEL

Data-driven decomposition techniques (like Independent Component Analysis (ICA), Canonical Correlation Analysis (CCA), Canonical Decomposition/PARAFAC (CP)) received in the last decades an increasing amount of attention as an exploratory tool in biomedical signal processing as opposed to model-driven techniques. Recently, "tensor-ICA", a combination of ICA and the CP model was introduced as a new concept for the decomposition of functional Magnetic Resonance data (fMRI) (Beckmann et al, 2005). The trilinear structure was in that study imposed after the computation of the independent components. We propose another algorithm to compute a trilinear decomposition with supposed independence in one mode. In this algorithm, independent component and trilinear structure constraints are imposed at the same time. We also show that this new algorithm outperforms the previously proposed tensor-ICA.

Joint research with Lieven De Lathauwer and Sabine Van Huffel.

Maarten De Vos
Department of Electrical Engineering
Katholieke Universiteit Leuven

RANK-(R_1, R_2, R_3) REDUCTION OF TENSORS BASED ON THE RIEMANNIAN TRUST-REGION SCHEME

We consider unstructured third-order tensors and look for the best (R_1, R_2, R_3) low-rank approximation. In the matrix case, low-rank approximation can be obtained from the truncated Singular value decomposition (SVD). However, in the tensor case, the truncated Higher-order SVD (HOSVD) gives a suboptimal low-rank approximation of a tensor, which can only be used as a starting value for iterative algorithms. The algorithm we present is based on the Riemannian trust-region method. We express the tensor approximation problem as minimizing a cost function on a proper manifold. Making use of second order information about the cost function, superlinear convergence is achieved.

Joint work with Lieven De Lathauwer, Pierre-Antoine Absil, Rodolphe Sepulchre, Sabine Van Huffel.

Mariya Ishteva
Department of Electrical Engineering
Katholieke Universiteit Leuven

SOLVING SECULAR EQUATIONS FOR LARGE TOTAL LEAST SQUARES/DATA LEAST SQUARES PROBLEMS BY MEANS OF GAUSS QUADRATURE RULES

We approximate secular equations for total least squares (TLS) and data least squares (DLS) problems by means of Lanczos tri-diagonalization processes. Based on Gaussian Quadrature (GQ) rules, the best number of Lanczos steps is determined with a given tolerance by investigating the bounds of the secular equations. The numerical example shows the efficacy of the GQ approach to solving the large TLS/DLS problems. We also discuss some implementation issues such as bisection, stabilization, Q-less QR preprocessing, and indefinite systems for TLS/DLS problems.

Joint work with Gene Golub and Zheng Su.

SungEun Jo
Department of Computer Science
Stanford University

MULTISCALE SPECTRAL GRAPH PARTITIONING AND IMAGE SEGMENTATION

Spectral methods for graph partitioning, based on numerical solution of eigenvalue problems with the graph Laplacian, are well known to produce high quality partitioning, but are also considered to be expensive. We discuss modern preconditioned eigensolvers for computing the Fiedler vector of large scale eigenvalue problems. The ultimate goal is to find a method with a linear complexity, i.e. a method with computational costs that scale linearly with the problem size. We advocate the locally optimal block preconditioned conjugate gradient method (LOBPCG), suggested by the presenter, as a promising candidate, if matched with a high quality preconditioner. We provide preliminary numerical results, e.g., we show that a Fiedler vector for a 24 megapixel image can be computed in seconds on IBM's BlueGene/L using BLOPEX in our BLOPEX software with Hypre algebraic multigrid preconditioning.

Andrew Knyazev
Department of Mathematical Sciences
University of Colorado at Denver

RETRIEVAL OF BIOLOGICAL IMAGES BASED ON REGION SIMILARITY

The sub-regions of an image may be more 'interesting' than an entire image. For e.g., an image of a retina has only a few regions that depict detachment of proteins in a certain fashion. To detect the presence of such protein detachments that are similar to a given image pattern, from a large database of retinal images, is a non-trivial task. The optimal algorithms for such pattern-matching are practically infeasible and unscalable. The goal of this project is to develop efficient heuristics for real time detection of such matching regions.

Rajesh Kumar and Swaroop Jagadish
Department of Computer Science
University of California, Santa Barbara

STOCHASTIC PROCESS METHODS FOR INFORMATION EXTRACTION

We consider several stochastic process methods for performing canonical correlation analysis (CCA). The first uses a Gaussian Process formulation of regression in which we use the current projection of one data set as the target for the other and then repeat in the opposite direction. The second uses a method which relies on probabilistically sphering the data, concatenating the two streams and then performing a probabilistic PCA. The third gets the canonical correlation projections directly without having to calculate the filters first. We also investigate nonlinearity and sparsification of these methods. Finally, we use a Dirichlet process of Gaussian models in which the Gaussian models are determined by Probabilistic CCA in order to model nonlinear relationships with a mixture of linear correlations where the number of mixtures is not pre-determined.

Pei Ling Lai
Department of Electronics Engineering
Southern Taiwan University of Technology

SHIFT SENSITIVITY OF EIGENFACE, EIGENPHASE, AND EIGENMAGNITUDE

Eigenface method applies Principal Component Analysis on a set of learning images from which eigenface are extracted. It's widely used and studied in statistical image recognition. We demonstrate that this method is sensitive to shift in the learning images. The correlation of images with the eigenvector, also known as eigenfeatures, are shifting as the learning images shift. We also compare it with eigen-phase and eigen-magnitude methods.

Wanjun Mi
Institute for Computational and
Mathematical Engineering
Stanford University

PROBABILISTIC FINGERPRINTS FOR SHAPES

We propose a new probabilistic framework for the efficient estimation of similarity between 3D shapes. Our framework is based on local shape signatures and is designed to allow for quick pruning of dissimilar shapes, while guaranteeing not to miss any shape with significant similarities to the query model in shape database retrieval applications. Since directly evaluating 3D similarity for massive collections of signatures on shapes is expensive and impractical, we propose a suitable but compact approximation based on probabilistic fingerprints which are computed from the shape signatures using Rabin's hashing scheme and a small set of random permutations. We provide a probabilistic analysis that shows that while the preprocessing time depends on the complexity of the model, the fingerprint size and hence the query time depends only on the desired confidence in our estimated similarity. Our method is robust to noise, invariant to rigid transforms, handles articulated deformations, and effectively detects partial matches. In addition, it provides important hints about correspondences across shapes which can then significantly benefit other algorithms that explicitly align the models. We demonstrate extension of our algorithm to

streaming data application. We demonstrate the utility of our method on a wide variety of geometry processing applications.

A preliminary version of the work has been submitted to Symposium of Geometry Processing (2006).

Niloy J. Mitra
Department of Computer Science
Stanford University

EXTENSIONS OF NON-NEGATIVE MATRIX FACTORIZATION (NMF) TO HIGHER ORDER DATA

Higher order matrix (tensor) decompositions are mainly used in psychometrics, chemometrics, image analysis, graph analysis and signal processing. For higher order data the two most commonly used decompositions are the PARAFAC and the TUCKER model. If the data analyzed is non-negative it may be relevant to consider additive non-negative components. We here extend non-negative matrix factorization (NMF) to form algorithms for non-negative TUCKER and PARAFAC decompositions. Furthermore, we extend the PARAFAC model to account for shift and echo effects in the data. To improve uniqueness of the decompositions we use updates that can impose sparseness in any combination of modalities. The algorithms developed are demonstrated on a range of datasets spanning from electroencephalography to sound and chemometry signals.

Morten Mørup
Department of Signal Processing
Technical University of Denmark

OUTLINES OF NON-CLASSICAL TIKHONOV METHOD

Today we know that ill-posed problems, for which the solution does not continuously depend on the variations in the data, arise naturally from real physical problems and are not, in most cases, as a result of incorrect modeling, but due to the inherent characteristics of the original physical problems.

In solving the corresponding Linear Least squares Problems, resulting from discretization of the original models, the usual solution methods such as QR/SVD or Normal equation Algorithm, would result in solutions with very little relevance to the exact solutions. The remedy, for the solution of these ill-conditioned least squares problem, is application of some regularization methods. The most used regularization methods are TSVD and Tikhonov regularization methods. In using Tikhonov regularization method, the most important steps are the choice of priori and the regularization parameter, which controls the level of regularization for the problem. Classical Tikhonov method is based on global priori and regularization parameter. This approach underestimates the local features of the solution and may result to oversmoothing of the original solution. To address this difficulty, a more global approach is needed. In this work, we consider this approach in the solution of the Tikhonov regularization method.

Jointly with Gene H. Golub.

Kourosh Modarresi
Institute for Computational and
Mathematical Engineering
Stanford University

EXPLORING PHYLOGENETIC RELATIONSHIPS IN SEQUENCE ALIGNMENTS THROUGH SINGULAR VALUE DECOMPOSITION

The 16S ribosomal RNA is a highly conserved molecule used to derive phylogenetic relationships among organisms, by traditional methods for phylogeny reconstruction. We describe the SVD analysis of an alignment of 16S rRNA sequences from organisms belonging to different phylogenetic domains. The dataset is transformed from a matrix of positions x organisms to a tensor of positions x code x organisms through a binary encoding that takes into account the nucleotide at each position. The tensor is flattened into a matrix of (positions x code) x organisms and singular value decomposition is applied, to obtain a representation in the "eigenpositions" x "eigenorganisms" space. These eigenpositions and eigenorganisms are unique orthonormal superpositions of the positions and organisms respectively. We show that the significant eigenpositions correlate with the underlying phylogenetic relationships among the organisms examined. The specific positions that contribute to each of these relationships, identified from the eigenorganisms, correlate with known sequence and structure motifs in the data, which are associated with functions like RNA-- or protein--binding. Among others, we identify unpaired adenosines as significant contributors to phylogenetic distinctions. These adenosine nucleotides, unpaired in the secondary (2D) structure, have been shown to be involved in a variety of tertiary (3D) structural motifs, some of which are believed to play a role in RNA folding [Gutell et al., *RNA* 2000].

Joint work with Robin R. Gutell, Gene H. Golub, Orly Alter.

Chaitanya Muralidhara
Department of Cellular and Molecular Biology
University of Texas at Austin

A TENSOR HIGHER-ORDER SINGULAR VALUE DECOMPOSITION FOR INTEGRATIVE ANALYSIS OF DNA MICROARRAY DATA

The structure of DNA microarray data is often of an order higher than that of a matrix, especially when integrating data from different studies. Flattened into a matrix format, much of the information in the data is lost. We describe the use of a higher-order singular value decomposition (HOSVD) in transforming a tensor of genes x arrays x studies, which tabulates a series of DNA microarray datasets from different studies, to a "core tensor" of "eigengenes" x "eigenarrays" x "eigenstudies," where the eigengenes, eigenarrays and eigenstudies are unique orthonormal superpositions of the genes, arrays and studies, respectively. This HOSVD, also known as N-mode SVD, formulates the tensor as a linear superposition of all possible outer products of an eigengene with an eigenarray with an eigenstudy, i.e., rank-1 "subtensors," the superposition coefficients of which are tabulated in the core tensor. Each coefficient indicates the significance of the corresponding subtensor in terms of the overall information that this subtensor captures in the data. We show that significant rank-1 subtensors can be associated with independent biological processes, which are manifested in the data tensor. Filtering out the insignificant subtensors off the data tensor simulates experimental observation of only those processes associated with the significant subtensors. Sorting the data according to the eigengenes, eigenarrays and eigenstudies appears to classify the genes, arrays and studies, respectively, into groups of similar underlying biology. We illustrate this HOSVD with an integration of genome-scale mRNA expression data from yeast cell cycle time courses, each of which is under a different oxidative stress

condition. Novel correlation between the DNA-binding of a transcription factor and the difference inthe effects of these oxidative stresses on the progress of the cell cycle is predicted.

Joint work with Gene H. Golub, Orly Alter.

Larsson Omberg
Department of Physics
University of Texas at Austin

A NOVEL HIGHER-ORDER GENERALIZED SINGULAR VALUE DECOMPOSITION FOR COMPARATIVE ANALYSIS OF MULTIPLE GENOME-SCALE DATASETS

We define a higher-order generalized singular value decomposition (GSVD) of two or more matrices D_i of the same number of columns and, in general, different numbers of rows. Each matrix is factored into a product $U_i\Sigma_i X^{-1}$ of a matrix U_i composed of the normalized column basis vectors, a diagonal matrix X^{-1}, and a nonsingular matrix X^{-1} composed of the normalized row basis vectors. The matrix X^{-1} is identical in all the matrix factorizations. The row basis vectors are the eigenvectors of C, the arithmetic mean of all quotients of the correlation matrices $D_i^T D_i$. The nth diagonal element of Σ_i, $\Sigma_{i,n}$, indicates the significance of the nth row basis vector in the ith matrix in terms of the overall information that the nth row basis vector captures in the ith matrix. The ratio $\Sigma_{i,n}/\Sigma_{j,n}$ indicates the relative significance of the nth row basis vector in the ith matrix relative to the jth matrix. We show that the eigenvalues of C that correspond to row basis vectors of equal significance in all matrices D_i, such that $\Sigma_{i,n}/\Sigma_{j,n}$, are equal to 1; the eigenvalues that correspond to row basis vectors which are approximately insignificant in one or more matrices D_i relative to all the other matrices D_j, such that $\Sigma_{i,n}/\Sigma_{j,n}\approx 0$, are $>>1$. We show that the column basis vector $U_{i,n}$ is orthogonal to all other column basis vectors if the corresponding nth row basis vector is of equal significance in all matrices, such that the corresponding eigenvalue of C is 1. These properties of the GSVD of two matrices [Golub & Van Loan, Johns Hopkins Univ. Press 1996] are preserved in this higher-order GSVD of two or more matrices.

Recently we showed that the mathematical row basis vectors uncovered in the GSVD of two genome-scale mRNA expression datasets from two different organisms, human and the yeast *Saccharomyces cerevisiae*, during their cell cycle, correspond to the similar and dissimilar among the biological programs that compose each of the two datasets [Alter, Brown & Botstein, *PNAS* 2003]. We now show that the mathematical row basis vectors uncovered in this higher-order GSVD of five genome-scale mRNA expression datasets from five different organisms, human, the yeast *Saccharomyces cerevisiae*, the yeast *Schizosacchomyces pombe*, bacteria and plant during their cell cycle, correspond to the similar and dissimilar among the biological programs that compose each of the five datasets. Row basis vectors of equal significance in all datasets correspond to the cell cycle program which is common to all organisms; row basis vectors which are approximately insignificant in one or more of the datasets correspond to biological programs, such as synchronization responses, that are exclusively manifested in all the other datasets and might be exclusive to the corresponding organisms. Such comparative analysis of genome-scale mRNA data among two or more model organisms, that is not limited to orthologous or homologous genes across the different organisms, promises to enhance fundamental understanding of the universality as well as the specialization of molecular biological mechanisms.

Joint work with Gene H. Golub, Orly Alter.

Sri Priya Ponnapalli
Department of Electrical and
Computer Engineering
University of Texas at Austin

MATRIX APPROXIMATION AND PROJECTIVE CLUSTERING VIA ADAPTIVE SAMPLING

Frieze, Kannan and Vempala proved that a small sample of rows of a given matrix contains a low rank approximation that minimizes the distance in terms of the Frobenius norm to within a small additive error, and the sampling can be done efficiently using just two passes over the matrix. We generalize this work by showing that the additive error drops exponentially by iterating the sampling in an adaptive manner. This result is one of the ingredients of the linear time algorithm for multiplicative low-rank approximation by Deshpande and Vempala.

The existence of a small certificate for multiplicative low-rank approximation leads to a PTAS for the following projective clustering problem: Given a set of points in Euclidean space and integers *k* and j, find *j* subspaces of dimension *k* that minimize the sum over the points of squared distances of each point to the nearest subspace.

Joint work with Amit Deshpande, Santosh Vempala and Grant Wang.

Luis Rademacher
Department of Mathematics
Massachusetts Institute of Technology

MANIFOLD-VALUED DATA MINING

New types of sensors and devices are being built everyday. Not only are we measuring huge amount of data but also new types of data, highly geometric in nature and inherently different than traditional Euclidian-valued data. Typical examples include human motion data in animation and diffusion tensor data in medical imaging. These types of data takes values on special Riemannian manifolds, called Symmetric Spaces. We call it 'Manifold-Valued' data. As the amount M-valued data touches terabyte mark, tools are needed that can efficiently mine this type of data. For example, radiologist in medical imaging might be interested in searching many terabytes of diffusion tensor data to find those matching, in a certain sense, given DT image. In animation one might be interested in extracting those motion clips, from huge motion capture database, that matches given query clip or description given by the animator. Hence all the traditional supervised and unsupervised learning issues, like clustering, classification, indexing, retrieval, searching etc, become important for M-valued data. Learning algorithm for Euclidian-valued data may not be appropriate and directly applicable because of highly geometric and non-linear nature of M-valued data. In this work, we discuss new wavelet like transform, that we have developed for M-valued data and its applicability in facilitating above cited learning and data mining task.

Inam Ur Rahman
Institute for Computational and
Mathematical Engineering
Stanford University

MATRIX DECOMPOSITIONS AND SECURITY PROBLEMS

Problems in counterterrorism, fraud, law enforcement, and organizational attempts to watch for corporate malfeasance all require looking for traces in large datasets. Sophisticated `bad guys' face two countervailing pressures: the needs of the task or mission, and the need to remain concealed. Because they are sophisticated, they do not show up as outliers; because they are unusual, they don't show up as mainstream either. Instead, they are likely to appear at the “edges” of structures in the data.

Several properties of matrix decompositions make them superb tools to look in the “edges” or “corners” of datasets. For example, SVD transforms data into a space in which distance from the origin corresponds, in a useful sense, to interestingness; data from innocent people provides a picture of normal correlation, against which unusual correlation stands out; and the symmetry between records and attributes makes it possible to investigate how “edge” records differ from normal ones. The machinery of spectral graph partitioning can also be used to look for unusual records, or

values using link prediction. Unresolved issues of normalization and removing stationarity are critical to making these approaches work.

Other matrix decompositions also have a role to play. ICA, for example, is powerfully able to discover small tightly-knit subgroups within a dataset. It was able to discover cells within a dataset of al Qaeda links. The importance of textual data suggests a role (as yet unfulfilled) for NNMF.

David Skillicorn
School of Computing
Queen's University

BEYOND STREAMS AND GRAPHS: DYNAMIC TENSOR ANALYSIS

Time-evolving data models have been widely studied in data mining field such as time-series, data streams and graphs over time. We argue that all these canonical examples can be covered and enriched using a flexible model *tensor stream*, that is a sequence of tensors growing over time.

Under this model, we propose two streaming algorithms for tensor PCA, a generalization of PCA for a sequence of tensors instead of vectors. We applied them in two real settings, namely, anomaly detection and multi-way latent semantic indexing. We used two real, large datasets, one on network flow data (100GB over 1 month) and one from DBLP (200MB over 25 years). Our experiments show that our methods are fast, accurate and that they find interesting patterns and outliers on the real datasets.

Jimeng Sun
Department of Computer Science
Carnegie Mellon University

EIGENCLUSTER: FINDING INNATE CLUSTERINGS ON THE FLY

We present a spectral algorithm for clustering massive data sets based on pairwise similarities. The algorithm has guarantees on the quality of the clustering found. The algorithm is especially well-suited for the common case where data objects are encoded as sparse feature vectors and the pairwise similarity between objects is the inner product between their feature vectors; here, the algorithm runs in space linear in the number of nonzeros in the object-feature matrix. The spectral algorithm outputs a hierarchical clustering tree. We show how to use dynamic programming to find the optimal tree-respecting clustering for many natural clustering objective functions, such as k-means, k-median, min-diameter, and correlation clustering. We evaluate the algorithm on a handful of real-world datasets; the results show our method compares favorably with known results. We also give an implementation of a meta-search engine that clusters results from web searches.

This is joint work with David Cheng, Ravi Kannan, and Santosh Vempala.

Grant Wang
Computer Science and
Artificial Intelligence Laboratory
Massachusetts Institute of Technology

STRUCTURED MATRIX METHODS FOR THE COMPUTATION OF A RANK REDUCED SYLVESTER MATRIX

The Sylvester resultant matrix *S(p,q)* is a structured matrix that can be used to determine if two polynomials *p=p(y)* and *q=q(y)* are coprime, and if they are not coprime, it allows their greatest common divisor (GCD) to be computed. In particular, the rank loss of *S(p,q)* is equal to the degree of the GCD of *p(y)* and *q(y)*, and the GCD can be obtained by reducing *S(p,q)* to row echelon form.

The computation of the GCD of two polynomials arises in many applications, including computer graphics, control theory and geometric modeling. Experimental errors imply that the data consists of noisy realizations of the exact polynomials *p(y)* and *q(y)*, and thus even if *p(y)* and *q(y)* have a non-constant GCD, their noisy realizations, *f(y)* and *g(y)* respectively, are coprime. It is therefore only possible to compute an approximate GCD, that is, a GCD of the polynomials *f*(y)* and *g*(y)* that are obtained by small perturbations of *f(y)* and *g(y)*. Different perturbations of *f(y)* and *g(y)* yield different approximate GCDs, all of which are legitimate if the magnitude of these perturbations is smaller than the noise in the coefficients. It follows that *f*(y)* and *g*(y)* have a non-constant GCD, and thus the Sylvester resultant matrix *S(f*, g*)* is a low rank approximation of the Sylvester matrix *S(f,g)*.

The method of structured total least norm (STLN) is used to compute the rank reduced Sylvester resultant matrix *S(f*, g*)*, given inexact polynomials *f(y)* and *g(y)*. Although this problem has been considered previously, there exist several issues that have not been addressed, and that these issues have a considerable effect on the computed approximate GCD.

The GCD of *f(y)* and *g(y)* is equal (up to a scalar multiplier) to the GCD of *f(y)* and *ag(y)*, where a is an arbitrary constant, and it is shown that a has a significant effect on the computed results. In particular, although the GCD of *f(y)* and *ag(y)* is independent (up to an arbitrary constant) of a, an incorrect value of a leads to unsatisfactory numerical answers. This dependence on the value of *a* has not been considered previously, and methods for the determination of its optimal value are considered. It is shown that a termination criterion of the optimization algorithm that is based on a small normalized residual may lead to incorrect results, and that it is also necessary to monitor the singular values of *S(f*,g*)* in order to achieve good results. Several non-trivial examples are used to illustrate the importance of *a*, and the effectiveness of a termination criterion that is based on the normalized residual and the singular values of *S(f*,g*)*.

The dependence of the computed solution on the value of *a* has implications for the method that is used for the solution of the least squares equality (LSE) problem that arises from the method of STLN. In particular, this problem is usually solved by the penalty method (method of weights), which requires that the value of the weight be set, but its value is defined heuristically, that is, it is independent of the data (the coefficients of the polynomials). As noted above, the value of the parameter *a* is crucial to the success or failure of the computed solution, and thus the presence of a parameter that is defined heuristically is not satisfactory. The QR decomposition, which does not suffer from this disadvantage, is therefore used to solve the LSE problem.

Joint work with John D. Allan.

Joab Winkler
Department of Computer Science
University of Sheffield

Index of Contributors

Acknowledgements

This workshop received financial support from the following sources:

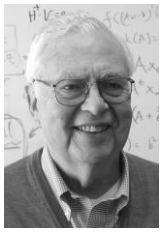
The National Science Foundation
Ask.com
Yahoo! Research

The organizers would like to thank the following people for their kind assistance in making this workshop possible:

Jillian Anderson (Stanford Computer Forum), Indira Choudhury (Stanford ICME), Adelaide Dawes (Wallenberg Hall), David Gleich (Stanford ICME), Sharad Khanal (Stanford CS), Lin Koh (Yahoo!), Mirella Machuca (Stanford CS), Wanjun Mi (Stanford ICME), Chana Motobu (Stanford ICME), Patty Namba (Yahoo!), Kelly Hamilton (Yahoo!).

Organizers

The conference was organized by Gene Golub, Michael Mahoney, Lek-Heng Lim, and Petros Drineas.



Gene Golub

Michael Mahoney

Lek-Heng Lim

Petros Drineas