IE 529
Stats of Big Data and Clustering
Group Project

1. NOV 9: "Scalable K-Means++," by Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii.

   Over half a century old and showing no signs of aging, k-means remains one of the most popular data processing algorithms. As is well-known, a proper initialization of k-means is crucial for obtaining a good final solution. The recently proposed k-means++ initialization algorithm achieves this, obtaining an initial set of centers that is provably close to the optimum solution. A major downside of the k-means++ is its inherent sequential nature, which limits its applicability to massive data: one must make k passes over the data to find a good initial set of centers. In this work we show how to drastically reduce the number of passes needed to obtain, in parallel, a good initialization. This is unlike prevailing efforts on parallelizing k-means that have mostly focused on the post-initialization phases of k-means. We prove that our proposed initialization algorithm k-means$\|$ obtains a nearly optimal solution after a logarithmic number of passes, and then show that in practice a constant number of passes suffices. Experimental evaluation on real-world large-scale data demonstrates that k-means$\|$ outperforms k-means++ in both sequential and parallel settings.

2. NOV 11: "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," by Bradley Efron.

   Current scientific techniques in genomics and image processing routinely produce hypothesis testing problems with hundreds or thousands of cases to consider simultaneously. This poses new difficulties for the statistician, but also opens new opportunities. In particular, it allows empirical estimation of an appropriate null hypothesis. The empirical null may be considerably more dispersed than the usual theoretical null distribution that would be used for any one case considered separately. An empirical Bayes analysis plan for this situation is developed, using a local version of the false discovery rate to examine the inference issues. Two genomics problems are used as examples to show the importance of correctly choosing the null hypothesis.

3. NOV 14: "Regularization and Variable Selection via the Elastic Net," by Hui Zou and Trevor Hastie.

   We propose the elastic net, a new regularization and variable selection method. Real world data and a simulation study show that the elastic net often outperforms the lasso, while enjoying a similar sparsity of representation. In addition, the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. The elastic net is particularly useful when the number of predictors (p) is much bigger than the number of observations (n). By contrast, the lasso is not a very satisfactory variable selection method in the $p \gg n$ case. An algorithm called LARS-EN is proposed for computing elastic net regularization paths efficiently, much like algorithm LARS does for the lasso.

4. NOV 16: "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," by Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller.

A new method for performing a nonlinear form of principal component analysis is proposed. By the use of integral operator kernel functions, one can efficiently compute principal components in high-dimensional feature spaces, related to input space by some nonlinear mapfor instance, the space of all possible five-pixel products in 16×16 images. We give the derivation of the method and present experimental results on polynomial feature extraction for pattern recognition.

5. NOV 18: "Exact Post Model Selection Inference for Marginal Screening," by Jason D. Lee and Jonathan E. Taylor.

   We develop a framework for post model selection inference, via marginal screening, in linear regression. At the core of this framework is a result that characterizes the exact distribution of linear functions of the response y, conditional on the model being selected (condition on selection" framework). This allows us to construct valid confidence intervals and hypothesis tests for regression coefficients that account for the selection procedure. In contrast to recent work in high-dimensional statistics, our results are exact (non-asymptotic) and require no eigenvalue-like assumptions on the design matrix X. Furthermore, the computational cost of marginal regression, constructing confidence intervals and hypothesis testing is negligible compared to the cost of linear regression, thus making our methods particularly suitable for extremely large datasets. Although we focus on marginal screening to illustrate the applicability of the condition on selection framework, this framework is much more broadly applicable. We show how to apply the proposed framework to several other selection procedures including orthogonal matching pursuit and marginal screening+Lasso.

6. NOV 28: "Gossip PCA," by Satish Babu Korada, Andrea Montanari, and Sewoong Oh.

   Eigenvectors of data matrices play an important role in many computational problems, ranging from signal processing to machine learning and control. For instance, algorithms that compute positions of the nodes of a wireless network on the basis of pairwise distance measurements require a few leading eigenvectors of the distances matrix. While eigenvector calculation is a standard topic in numerical linear algebra, it becomes challenging under severe communication or computation constraints, or in absence of central scheduling. In this paper we investigate the possibility of computing the leading eigenvectors of a large data matrix through gossip algorithms.

   The proposed algorithm amounts to iteratively multiplying a vector by independent random sparsification of the original matrix and averaging the resulting normalized vectors. This can be viewed as a generalization of gossip algorithms for consensus, but the resulting dynamics is significantly more intricate. Our analysis is based on controlling the convergence to stationarity of the associated Kesten-Furstenberg Markov chain.

7. NOV 30: "Randomized Algorithms for Low-Rank Matrix Factorizations: Sharp Performance Bounds," by Rafi Witten and Emmanuel Candès.

   The development of randomized algorithms for numerical linear algebra, e.g. for computing approximate QR and SVD factorizations, has recently become an intense area of research. This paper studies one of the most frequently discussed algorithms in the literature for dimensionality reduction—specifically for approximating an input matrix with a low-rank element. We introduce a novel and rather intuitive analysis of the algorithm in [6], which allows us to derive sharp estimates and give new insights about its performance. This analysis yields theoretical

guarantees about the approximation error and at the same time, ultimate limits of performance (lower bounds) showing that our upper bounds are tight. Numerical experiments complement our study and show the tightness of our predictions compared with empirical observations.

8. DEC 2: "Matrix Completion from Noisy Entries," by Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh.

   Given a matrix M of low-rank, we consider the problem of reconstructing it from noisy observations of a small, random subset of its entries. The problem arises in a variety of applications, from collaborative filtering (the Netflix problem) to structure-from-motion and positioning. We study a low complexity algorithm introduced by Keshavan, Montanari, and Oh (2010), based on a combination of spectral techniques and manifold optimization, that we call here OPTSPACE. We prove performance guarantees that are order-optimal in a number of circumstances.

9. DEC 5: "Distributed k-Means and k-Median Clustering on General Topologies," by Maria Florina Balcan, Steven Ehrlichy, and Yingyu Liangz.

   This paper provides new algorithms for distributed clustering for two popular center-based objectives, k-median and k-means. These algorithms have provable guarantees and improve communication complexity over existing approaches. Following a classic approach in clustering by [16], we reduce the problem of finding a clustering with low cost to the problem of finding a coreset of small size. We provide a distributed method for constructing a global coreset which improves over the previous methods by reducing the communication complexity, and which works over general communication topologies. Experimental results on large scale data sets show that this approach outperforms other coreset-based distributed clustering algorithms.