Sparse Matrix Factorization of Gene Expression Data

Nathan Srebro & Tommi Jaakkola

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Motivated by the analysis of gene expression data, we develop a new unsupervised modeling technique. Specifically, we study how data can be modeled as a *sparse matrix factorization*.

Motivation: Gene expression data consists of expression level reads for thousands of genes across dozens of experimental conditions, time points, cell types or repeated experiments. The goal of unsupervised modeling of such data is to find some underlying organization, structure or redundancy in the data, such as similarity or dependency between genes or between experiments. Such structure can then be used to assist in biological study of the gene expression patterns, or as a pre-processing step for classification and prediction tasks.

Previous Work: Clustering of genes, including hierarchical clustering, and clustering of experiments are unsupervised modeling approaches currently in common use for interpreting gene expression data [2]. Eigen-decomposition, or equivalently low-rank modeling using the singular value decomposition, has also been proposed [1, 3, 7].

Unsupervised modeling using constrained matrix factorization has been studied by Lee and Seung [4, 5, 6]. They suggested various forms of positivity constraints (convex, conic and non-negative matrix factorizations), demonstrated the utility of such methods in analyzing images and documents, and discussed their connection to human perception.

Approach: We introduce the following notion of a sparse matrix factorization. A matrix factorization of a given data matrix $A \in Re^{n \times d}$ is a factorization of A as a product of two matrices $A = C \cdot F$, where $C \in \mathcal{R}^{n \times k}$ and $F \in \mathcal{R}^{k \times d}$. A sparse matrix factorization is a matrix factorization with the added constraint that each row of C has at most m non-zero entries. The complexity of the factorization is controlled by the number of factors, k, and factor polymorphicity m.

We propose analyzing data by studying the sparse matrix factorization of a given complexity (k, m) that best approximates the data.

A (k, m) sparse matrix factorization can be thought of as an explanation of the data rows using k factors. Each row is a linear combination of k factors. However, each row can only be affected by a small number of factors, and is thus limited to being a linear combination of at most m of the k factors. This view might be appropriate for gene expression analysis, where expression of genes may be affected by many factors (representing either direct molecular factors or more abstract conditions), but each gene is only affected by a few of these factors.

The choice of the parameters k and m plays a crucial role in the structure of the decomposition. In particular, for the choice $m = k$, the factorization is simply a low rank approximation. For the choice $m = 1$, each row of the data matrix is associated with only a single factor, and the sparse matrix factorization is a clustering of the data rows.

Algorithms: Given a data matrix, we are faced with the problem of finding the sparse matrix factorization that best approximates it. For the case of $m = k$, this is the low rank approximation of the matrix, which can be found directly using the singular value decomposition.

However, for the cases we are interested in, where $m < k$, there is no known method for finding the sparse matrix factorization that best approximates the data. We propose several iterative maximization approaches, which iteratively maximize different subsets of the parameters, leaving the rest constant. These approaches generalize the standard maximization-maximization method for k-means clustering. As for the k-means methods, they may converge to non-optimal decompositions, and the rate of convergence is not guaranteed. The various approaches vary in the partitioning of the parameters into subsets which will be optimized together.

We are currently studying these optimization approaches and are developing heuristics that can reasonably

reconstruct the best-fit decomposition. We are analyzing the heuristics by observing their performance on synthetic "planted" instances.

Impact: Signal decomposition is an important part of data analysis. The sparse matrix factorization approach described here provides a combinatorial account of the observed data in terms of a few underlying factors (themselves unknown and inferred). We envision a useful application of this method to explaining differential gene expression in cancer tissues by uncovering more subtle and previously unknown variations in cancer types. Moreover, the method provides a means for more subtle attribution of gene function in the context of, e.g., pathogen response measurements.

Similarly to non-negative matrix factorization, the sparse factorization approach is not limited to any specific application area but instead offers a more widely applicable methodology for signal decomposition.

Future Work: We are currently extending the estimation algorithms and acquiring a better understanding of the factorization method. We are analyzing under what conditions the underlying factors and the sparse dependency patterns can be accurately recovered from limited observations. Moreover, we are exploring how the estimated/recovered factors are affected by different settings of the main parameters, the number of factors (k) and polymorphicity (m).

In terms of applications, we use the sparse factorization method to disambiguate underlying tissue type variations in human cancer cells, identify gene functions on a global scale in yeast, as well as compare our sparse factorization approach to related methods such as non-negative matrix factorization.

Research Support: The authors acknowledge support from Nippon Telegraph and Telephone Corporation and from ARO MURI grant DAAD19-00-1-0466. N.S. was supported in part by an NIH Genome Training Grant.

References:

[1] Orly Alter, Patrick O. Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*, 97(18):10101–10106, 2000.

[2] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863–14868, 1998.

[3] Neal S. Holter, Madhusmita Mitra, Amos Maritan, Marek Cieplak, Jayanth R. Banavar, and Nina V. Fedoroff. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *PNAS*, 97(15):8409–8414, 2000.

[4] D. Lee and H. Seung. Unsupervised learning by convex and conic coding. In *Advances in Neural Information Processing Systems*, volume 9, pages 515–521, 1997.

[5] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[6] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13, pages 556–562, 2001.

[7] S. Raychaudhuri, J.M. Stuart, and R.B. Altman. Principal components analysis to summarize microarray experiments: Application to sporulation time series. In *Pacific Symposium on Biocomputing*, volume 5, pages 452–463, 2000.