Sparse Matrix Factorization for Analyzing Gene Expression Patterns

Nathan Srebro Tommi Jaakkola
nati,tommi@ai.mit.edu

Abstract

Motivated by the analysis of gene expression data, we develop a new unsupervised modeling technique. Specifically, we study how such data can be modeled via *sparse matrix factorization* (SMF).

Unsupervised modeling using constrained matrix factorization has been studied by Lee and Seung [1, 2, 3]. Under this approach, one unveils structure in a data matrix $A \in \mathcal{R}^{n \times d}$, by approximating it as a product of two matrices $A \approx C \cdot F$, $C \in \mathcal{R}^{n \times k}, F \in R^{k \times d}$, subject to various (e.g., non-negative) constraints on C and F. We suggest explicit sparsity constraint on C. Specifically, each row of C is to have at most m non-zero entries. Setting $m = 1$, we obtain a clustering of the data rows, where the rows of F indicate the cluster *centers*. At the other extreme, setting $m = k$, leaves C unconstrained and yields a low-rank approximation, specified by the leading components of the singular value decomposition.

Focusing on small values of m, and viewing the rows of F as *factors*, each row of the data matrix A is approximated as a linear combination of only m of the k factors. In the context of gene expression analysis, where, e.g., the rows of the data matrix correspond to genes, and the columns to different experiments, we get a model in which the expression pattern of a gene is explained as a (linear) combination of a few (at most m) underlying factors. This model model allows us to capture combinatorial effects and genes which take part in more than one underlying process. Constraining to sparse C permits us to recover a higher number of interpretable factors than what is possible with singular value decomposition [4, 5, 6].

When $m < k$, finding the best SMF (i.e. finding appropriate C, F that best approximate A) is a difficult optimization task. We formulate and investigate several iterative (alternating) maximization techniques in this context. Alternatively, the hard sparsity constraint can be relaxed to regularization penalties on the rows of matrix C, yielding a continuous, and thus easier to handle, optimization problem.

We study the statistical problem of reconstructing a sparse matrix factorization in the presence of noise. We determine the conditions under which the factors in F can be reconstructed, and study the problem of recovering the pattern of zeroes in C as an error correcting code, whose error correction properties can be determined as a function of the noise level. We also address the model

selection problem of identifying meaningful settings of the number of factors k and the polymorphicity m.

The primary goal of this work is to provide a large scale functional genomics analysis tool using gene expression and other data sources. Beyond using SMF to recover underlying factors, and structure among genes, we use SMF to extend partial factor realizations (some factors fixed according to known profiles of transcriptional activators). Moreover, we recover expression profiles for factors identified by common sequence motifs.

We also explore the connection of SMF to other learning and inference tasks. SMF can also be viewed as a class of probability relational models (PRMs), similar to the ones suggested by Segal *et al* [7] for analyzing gene expression data. Moreover, SMF can be seen as a technique for independent component analysis (ICA), where the sparsity requirement serves as an additional (symmetry breaking) regularization constraint. Lee and Seung suggested viewing constrained matrix factorization as a coding of the rows in A, the point of view explicitly taken in our analysis.

References

[1] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] Eran Segal, Ben Taskar, Audrey Gasch, Nir Friedman, and Daphne Koller. Rich probabilistic models for gene expression. In *Proceedings of the 9th International Conference on Intelligent Systems For Molecular Biology*, July 2001.