Biclustering of gene expression data by non-smooth non-negative matrix factorization

Chengyu Liu

Helsinki October 15, 2010

UNIVERSITY OF HELSINKI
Department of Computer Science
Biclustering of gene expression data by non-smooth non-negative matrix factorization

Microarray technology has been used widely, which has enabled the generation and accumulation of gene expression datasets containing expression levels of thousands of genes across different number of experiments conditions. The challenge we meet is how to discover local patterns that consist of genes and conditions and shows similarities. These patterns might provide valuable information to discover the nature of biological processes. In this paper, I introduce new methodology that can be used to cluster genes and conditions highly related in sub-portions of the data. The results shows that the method can identify localized features set of genes which show consistent pattern across subset of conditions.
1 Introduction

Microarray technologies have been used widely in the most medical and biological research areas, and have become an indispensable part of the modern biological and biomedical research areas. Microarray gene expression data is a matrix where each cell represents the gene expression level of a gene under a particular condition. In Section 2, I will introduce the basic idea of microarray. Due to the popularity of high-throughput technology in system biology, a large number of microarray data are generated, some of which share less common experimental conditions.

One of the most important tasks of microarray analysis is to identify simultaneously the groups of genes and groups of experimental conditions which shows the experimental pattern similarities. Clustering provides a solution to such an application to discover coherent subset of genes and experimental conditions and many modified versions of such clustering methods have been applied successfully, such as k-mean [TAV99] and hierarchical clustering [EIS98]. The identification of clusters plays a significantly important role in understanding the mechanism of biology. Veer, L. et al. used unsupervised hierarchical cluster algorithm to cluster 98 tumors on the basis of their similarities measured over approximately 5,000 significant genes (Figure 1). The two dimensional display shows that the samples can be divided into two groups based on the set of 4,968 genes. In the upper type only 34% of the patients developed distant metastasis within 5 years, however, 70% of the sporadic patients had progressive disease in the lower type. Consequently, unsupervised clustering methods does identify groups of similarly expressed genes or samples. But such classical clustering methods can’t find the patterns that are overlapped with others.

In this paper, I describe a new method [PED06] able to discover patterns even they are overlapped. This method simultaneously clusters genes and conditions highly related in sub-portions of the data. The main purpose of this paper is to show the potential of this method identifying gene expression patterns that share only local similarities in gene expression pattern.

This paper is organized as follows. In Section 2, I begin with introduction of microarray and definition of biclustering algorithm. I describe the non-negative matrix factorization which is the base of non-smooth non-negative matrix factorization in Section 3 and followed with the key algorithm of non-smooth non-negative matrix factorization in Section 4. I present the results of the experiment in Section 5 and make conclusion last in Section 6.

2 Background

2.1 Microarray

DNA microarray technology has been a revolutionary approach in biological and biomedical research areas, measuring the expression levels of at least ten thousands of genes in a single chip under various conditions. Microarray has been applied in the areas of biology, drug discovery and disease diagnosis.

In its most general form, a DNA array consists of a solid surface on which strands of nucleotides, called probe are deposited. What kinds of strands might be attached are dependent on the purpose of array. One might attach a number of regions dedicated to the individual genes,
if one wants to discover the function of gene set. However, arbitrary probes might be deposited for more general queries. Normally this query is done by washing the array using the solution containing strands called target. Target is complementary sequences to the probes of DNA array, which you want to query. The main idea is that the targets will hybridize with complementary sequence in solution. The key to interpretation is that the targets are labeled with fluorescent dye or other methods. Then hybridization spot can be detected and quantified easily (more details see [STO05]).

2.2 Biclustering methods

A various types of clustering methods, such as k-mean [TAV99] and hierarchical clustering [EIS98], have been extensively applied to analysis gene expression data which are obtained from microarray experiments. A major goal of clustering is to group genes or conditions into disjoint subset, which are similar to each other within subset. Formally given a set of conditions with a vector of genes for each condition, clustering methods partition the conditions into disjoint sets, known as clusters, in order that attribute vectors are similar, whereas vectors of disjoint clusters are dissimilar. For example, given a gene expression matrix with rows corresponding to genes and columns corresponding to conditions, we may apply clustering method to either genes or condition to identify groups of similarly expressed genes or conditions from gene expression data. Clustering methods make several assumptions which are not perfectly adequate in all applications, especially in the analysis of large and diverse conditions of gene expression data, although clustering methods are very useful and powerful, and have been successfully applied in the microarray data [VEE02]. First, clustering may be applied to either genes or conditions based on the global similarities in gene expression data. This implicates that analysis is directed to a particular part of the system under study. However a set of co-regulated genes are co-expressed in a subset of experimental conditions, and show non-relativity and independence in the rest. Second, clustering methods always group genes or conditions into disjoint clusters, assigning each gene or condition to a single cluster. Nevertheless one gene might be related to many different biological process, thus such a gene should be grouped into various clusters, not into a single cluster. Further, clustering one gene into only one cluster may not discover all the interrelationships among genes and fail to obtain a molecular fingerprint of tissues or cells in different biological conditions. During last decade, several new approaches have been proposed to conquer the limitations of clustering methods. Among those approaches, biclustering approaches have become an alternative algorithms [SMD03] [TSS02] to identify local structures from gene expression datasets. The term biclustering was first introduced by Cheng and Church [CCH00] in gene expression data analysis. It’s a different class of clustering algorithms, which perform clustering simultaneously to genes and conditions in order to discover a bicluster, while clustering algorithms are applied either to genes or conditions of gene expression matrix. Biclustering algorithms derive local structures while clustering algorithms discover global patterns. Further, biclustering algorithms have no assumptions on constructing biclusters, and genes and conditions can be grouped into more than one biclusters or into no bicluster(compare Figure 2).

A bicluster is defined as a submatrix of gene expression matrix, formed by a set of genes and a set of conditions. Each gene in a bicluster shows similarity of expressed pattern only across specific
Figure 1: Unsupervised two-dimensional cluster analysis of 98 breast tumors with 4,968 genes.

Figure 2: Clustering and biclustering of a gene expression matrix. Clusters correspond to disjoint strips in the matrix. A gene cluster must contain all columns, and a condition cluster must contain all rows. Biclusters correspond to arbitrary subsets of rows and columns, shown here as rectangles. Note that since gene (condition) clusters are disjoint, the rows (columns) of the matrix can be reordered so that each cluster is a contiguous strip. Similar reordering of rows and columns that shows all the biclusters as rectangles is usually impossible. [PED06]
subset of condition, and each condition in a bicluster show similar activity across a specific subset of genes. For a review of biclustering algorithms see [MOB04] [TSS05].

3 Non negative matrix factorization

Non-negative matrix factorization (NMF) was introduced by both Paatero and Tapper [PPT94] in 1994 and Lee and Seung [LS99] in 1999, which is used to reduce the dimensionality, to discover hidden pattern, and, further, to find the representation of the non-negative data. NMF has been applied to many different fields such as image analysis [LS99] and molecular pattern discovery [BPT04].

The principal difference between NMF and other classical factorization techniques, such as Principal Component Analysis and Independent Component Analysis is depending on the non-negativity constraints imposed on the matrix factors $W$ and $H$. Due to the properties of such constraints, only additive, not subtractive combination is permitted.

Formally given an $n \times m$ data matrix $V$ with $V_{ij} \geq 0$ and a pre-specified positive integer $r < \min(n,m)$, NMF finds two non-negative matrices $W \in \mathbb{R}^{n \times r}$ and $H \in \mathbb{R}^{r \times m}$ such that

$$V \approx WH$$  \hspace{1cm} (1)

where $W$ are the reduced $r$ bias factors and $H$ which is known as encoding vectors contains the coefficients of the linear combinations of the basis factors needed to reconstruct the origin data. One conventional approach to find $W$ and $H$ is that we minimize the difference between $V$ and $WH$:

$$D(V, WH) = \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij} \ln \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}$$  \hspace{1cm} (2)

The most easy way to solve Equation 2 is the multiplicative update algorithms proposed by Lee and Seung (2001) [LSH01], which is very easy to implement and produce good results. In each iteration, all the elements of $W$ and $H$ are multiplied by certain factors. As the zero elements are not updated, all the elements of the $W$ and $H$ are strictly non negative for iterations. When convergence is reached, sparseness matrices $W$ and $H$ are formed. Each basis produced by NMF contains relatively small set of rows with non-zero coefficient that determine a local structure. Those rows are strongly correlated with sub-portion of data. In the same way, those coefficient in basis are used to determine the set of columns of data associated with those module.

4 Non-smooth Non negative matrix factorization

Although NMF has been presented and applied successfully in many fields to find the underlying part-based structures of complex data, there is not any explicit proofs to support this trait except for the non-negativity constraints. In fact there is high degree of overlapping among those basis vectors that are negative evidence of the intuitive nature of the parts.

Non-smooth non-negative matrix factorization was proposed by Pascual-Montano [PCS06], 2006 which is a further evolution of non-negative matrix factorization model aiming to discover sparse
features that could better represent the non-negative multivariate data, whose definition given below:

\[ V \approx (WS)H = W(SH) \]  

(3)

where \( S \in \mathbb{R}^{r \times r} \) is a positive symmetric smoothness matrix defined as:

\[ S = (1 - \theta)I + \theta \frac{11^T}{r} \]  

(4)

where \( I \) is the identity matrix, \( 1 \) is a vector of ones, and the parameter \( \theta \) controls the sparseness of the model and satisfies \( 0 \leq \theta \leq 1 \).

The interpretation of \( S \) as a smoothing matrix is as follows: Let \( x \) be a positive vector. Consider the transformed vector \( y = Sx \). If \( \theta = 0 \), then \( y = x \) and no smoothing on \( x \) has occurred. However, as \( \theta \to 1 \), the vector \( y \) tends to the constant vector with all elements almost equal to the average of the elements of \( x \). This is the smoothest possible vector in the sense of 'nonsparseness' because all entries are equal to the same nonzero value, instead of having some values close to zero and others clearly nonzero[PCS06].

The objective function can be defined as following divergence function:

\[ D(V,WSH) = \sum_{i=1}^{m} \sum_{j=1}^{n} V_{ij} \ln \frac{V_{ij}}{(WSH)_{ij}} - V_{ij} + (WSH)_{ij} \]  

(5)

Further insight into the nature of the new nsNMF model, non-sparseness in the basis \( W \) will force sparseness in the encoding \( H \). At the same time, non-sparseness in the encoding \( H \) will force sparseness in the basis \( W \). Due to the simultaneity of both conditions, sparseness will be enforced on both basis and encoding parts. This trait of nsNMF, that is the capacity in extracting local patterns from data is the main motivation for using this algorithm for biclustering[PED06]. Additionally, \( \theta \) plays an important role, which control the sparseness of the model and it is impossible to estimate from functional. By performing numerous empirical test, Pedro et al.[PED06] found that the value of 0.5 leads to a reasonable results.

5 Experiment

5.1 nsNMF discover local structures

As described in previous section, nsNMF can be used to reproduce a gene matrix \( V \) with \( m \) genes and \( n \) samples as a product of two matrices of \( W \in \mathbb{R}^{m \times r} \) and \( H \in \mathbb{R}^{r \times n} \). The \( r \) columns of \( W \) are called factors and the columns of \( H \) are known as encoding vectors which are in one-to-one correspondence with a single experiment of gene expression matrix \( V \). Thus, each row of \( H \) denotes as 'basis gene'. Each factor contains a relatively small set of genes with non-zero coefficients that determine a local gene expression feature. Those genes are much correlated with sub-portion of data and constitute a gene module. In the same way, coefficients in basis genes determine the set of samples which are highly related to those modules. In other words, the set of genes and experimental conditions that show high values in the same basis experiment (\( l \)th column of \( W \)) and
Figure 3: General schema of the method nsNMF

Figure 4: (a) Original dataset with the two embedded patterns. (b) Dataset sorted by two-way hierarchical clustering. Dataset sorted by (c) the first basis gene and basis experiment and (d) the second basis gene and basis experiment yielded by nsNMF at \( r = 3 \). Conditions belonging to pattern P1a are marked in green and conditions belonging to pattern P2a are depicted in orange. The two plots over the heat maps represent the coefficients of conditions in each sorted basis gene.
its corresponding basis gene (1 th row of \( H \)) respectively are highly related in only a sub-portion of the data and constitute a gene expression bicluster [PED06].

Given a certain factor \( l \)th column of \( W \), according to the association of the local pattern extracted by this factor, all genes in the dataset can be sorted. Samples can be sorted simultaneously based on their coefficients in the corresponding basis gene \( l \)th row of \( H \). This operation is carried out in one-to-one correspondence through all columns in \( W \) and all rows in \( H \). Finally \( k \) local structures of the gene expression matrix are discovered, in which gens and samples are highly correlated in sub-portion, and are placed in the upper left corner. However we can not guarantee the structures we found are biologically meaningful, but are occurred more frequently. We still need a biologist to interpret and verify the findings, even they are statistically significant.

Figure 3 shows the general schema of nsNMF. nsNMF approximates the original matrix as a product of two submatrices, \( W \) and \( H \). Columns of \( W \) are basis experiments while rows of \( H \) constitute basis genes (columns of \( W \) and rows of \( H \) are separated for a better visibility). Coefficients in each pair of basis gene and experiment are used to sort conditions and genes in the original matrix. Conditions and genes with high values in the same basis gene and basis experiment are highly related in a sub-portion of the data and are co-clustered in the upper left corner of the sorted array.

5.2 Model Selection

One important consideration in the application of nsNMF is that how to select the number of factors \( r \) in order to better explain the data. Finding an appropriate value of \( r \) is depending on the specific applications and it is influenced by the nature of dataset itself. Generally speaking, more factors we use, we explain the data better. In our application of biclustering, it is enough to use only a reduced set of factors explaining the data without obscuring the biclusters information with too many details.

For any given rank \( r \), the nsNMF algorithm groups data into biclusters. For any \( r \) we can group \( r \) biclusters. The key issue is how to choose \( r \) so that the algorithm groups data into meaningful biclusters. nsNMF algorithm might or might not converge to the same solution on each run, depending on the random initial conditions. Brunet et al.[BPT04] proposed cophenetic correlation coefficient which is used to evaluate the stability of clustering associated with a given rank \( r \).

For each run, a connectivity matrix \( C \) of size \( M \times M \) is defined. Each entry \( c_{ij} \) is equal = 1 if samples \( i \) and \( j \) belong to the same cluster, otherwise \( c_{ij} \) is equal = 0. After that the consensus matrix \( \hat{C} \) is computed as the average connectivity matrix over many clustering runs. Then \( p_k(\hat{C}) \) is measured, which indicates the dispersion of the consensus matrix \( \hat{C} \) as Pearson correlation of two distance matrices: the first, I-\( \hat{C} \), is the distance between samples induced by the consensus matrix, and the second is the distance between samples induced by the linkage used in the reordering of \( \hat{C} \). In perfect consensus matrix (all entries = 0 or 1), the cophenetic correlation coefficient is 1. We observe how \( p_k \) changes as \( r \) increases. We select the value of \( r \) where the magnitude of the cophenetic correlation coefficient begins to fall.
Figure 5: (a) Original dataset with the three embedded patterns and (b) the same dataset sorted by two-way hierarchical clustering. Heatmaps of the original dataset sorted by the (c) first, (d) second, (e) third and (f) fourth basis genes and basis experiments yielded by nsNMF at \( r = 4 \) are shown in the bottom part of the figure. Non-overlapping conditions of Plb are marked in red, non-overlapping conditions of P2b are marked in green and non-overlapping conditions of P3b are marked in magenta. The overlapped area between Plb and P2b is marked in brown while the overlapped columns between P2b and P3b are marked in orange. Columns of P4b are marked in blue.
5.3 Performance

In order to show the potential of this method, two examples are included in this paper. The first example is that given a 100 × 20 noisy matrix which two embedded structures without any overlapping (p1a: 20 × 5 and p2a: 25 × 8) shown in Figure 4. As it was expected, these two structures are separated both by hierarchical clustering and nsNMF. At r = 3 calculated based on cophenetic correlation coefficient, the model has high stability.

Another example is used to test the potential of nsNFM model to identify overlapped structures. Figure 5 shows the results obtained in the analysis. The dataset contains four embedded structures (p1b, p2b, p3b and p4b of sizes 10 × 8, 15 × 9, 20 × 5, 10 × 3 respectively). p2b has three shared columns with p1b and two columns with p3b. Based cophenetic correlation coefficient it was found that the model has high robustness at r = 4. The final result shows that all four structures were found. However, two-way hierarchy clustering performs well at grouping samples and genes belonging to p4b, while it did not correctly identify overlapped structures. The reason for failure is the limitation of hierarchy clustering as presented in Section 2. Hierarchy clustering as well as other classical clusters groups objects into discrete clusters masking potential relationships among objects.

nsNMF model is tested using real microarray gene expression matrix with thousands of human genes across a panel of 79 human tissues. Three of the eight gene expression modules are noticeable consistency. For see [PED06]

6 Conclusion

In this paper, I described a new method of biclustering able to cluster DNA microarray gene expression data. Compared with other classical clustering algorithms, such as hierarchy clustering and PCA, it is able to discover the structures that has overlap and considers the relationship among genes and experimental conditions. This method is easy to implement and is new attempt of biclustering method.

Although the experiment results show the impressive capacity of discovering local structures both with overlap and without overlap, there is no comparison between this method with other biclustering methods.

References


CCH00 Cheng, Y. and Church, G., Biclustering of expression data. Proc. ISMB’00, pages 93 – 103.


MOB04 Madeira, S. and Oliveira, A., Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.


SMD03 Sheng, Q., Moreau, Y. and De Moor, B., Biclustering microarray data by gibbs sampling. *Bioinformatics*.


TSS02 Tanay, A., Sharan, R. and Shamir, R., Discovering statistically significant biclusters in gene expression data. *Bioinformatics*.