NMF finds Connections in Complex Data

Clare Lee\textsuperscript{1}, Des Higham\textsuperscript{1}, Keith Vass\textsuperscript{2}, Dan Crowther\textsuperscript{2,3}

4th September 2010

\textsuperscript{1} Department of Mathematics & Statistics, University of Strathclyde.
\textsuperscript{2} Translational Medicine Research Collaboration, University of Dundee.
\textsuperscript{3} Pfizer Inc.
Contents

Background and Algorithms

Results

An Extension to Multiple Data Sets

Biological Literature

Discussion
Aim

The aim is to take non-negative data sets, for example microarray data and to reorder or cluster the data to find hidden features using non-negative matrix factorisation (NMF).
NMF Algorithms

There are many different algorithms to compute a NMF. Typically they compute two factors so that

\[ A = WH + \text{error} \]  \text{minimises}  \| A - WH \| \]

with all entries of \( W \) and \( H \) being non-negative.

If \( A \) is of size \( m \times n \), then \( W \) is \( m \times k \) and \( H \) is \( k \times n \), where \( k \ll m \) or \( n \).

In the iterative approach

\[ W_{i+1} = W_i \sum_{\text{samples}} \left( \begin{array}{c} \text{under/over-estimate factor for this sample} \\ \text{importance of sample in cluster} \end{array} \right) \]

\( H \) is found in an analogous way.
Feature finding, Ordering and Clustering

\[ A \approx WH = \sum_{j=1}^{k} w_j h_j^T, \]

for \( W = [w_1, \ldots, w_k] \), and \( H = [h_1, \ldots, h_k]^T \). Each rank-one non-negative matrix \( w_j h_j^T \) expresses a “feature” of the data. As shown by Lee and Seung [Nature(1999)]
Feature finding, Ordering and Clustering

Therefore, for example in the microarray applications

- the columns of the first factor $W$ are referred to as “eigen-genes”
- the rows of the second factor $H$ are equivalently “eigen-samples”

Since each column/row expresses one feature we can locate this in the data by re-ordering the individual vectors to put the largest value in the bottom right corner.

We can also combine these features into one ordering ...
Feature finding, Ordering and Clustering

Each row in $W$ is assigned to a cluster corresponding to the largest element in that row.
Feature finding, Ordering and Clustering

Each row in $W$ is assigned to a cluster corresponding to the largest element in that row.

Row 1 is assigned to cluster 1
Row 2 is assigned to cluster 2
Row 3 is assigned to cluster 1
...
Feature finding, Ordering and Clustering

A row ordering then comes from stacking the clusters and sorting each cluster by size of that column.
Feature finding, Ordering and Clustering

The same is done with H and the columns.
Feature finding, Ordering and Clustering

These orderings can then be applied to the original matrix.
Colon Cancer Gene Correlation: $k=75$
Colon Cancer Gene Correlation: $k=75$
Comparing with known information
Genes suppressed by oncogene HRas

This cluster’s “Ras signature” contains many proteins found in the “extracellular region”. The cluster includes

ADAMTS5, C1S, CADM1, CH25H, COL11A1, COL1A2, COL3A1, COL4A1, COL4A2, COL5A1, COL5A2, CRISPLD2, DCN, EFEMP1, ELN, IGFBP3, LUM, MXRA5, POSTN, SPOCK1, SULF1
Comparing with known information

vec5 - probesets associated with cell division and DNA copying

This cluster’s “DNA replication and cell-division” set is enriched in proteins for the “nucleus” and the “mitochondrion”.
Comparing with known information

vec5 - probesets associated with cell division and DNA copying

This cluster’s “DNA replication and cell-division” set is enriched in proteins for the “nucleus” and the “mitochondrion”.

Comparing with known information

C2 set of genes associate with Notch pathway being active
Comparing with known information

HRas suppressed, vec5 and C2 genes
Using Multiple Data Sources

In some situations it may be advantageous to use more than one data source to improve our results or to look for differences and similarities between data sources.

For this there is Simultaneous NMF to factorise data matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times n}$ so that

$$A \approx WH \quad \text{and} \quad B \approx SH$$

with $W \in \mathbb{R}^{m \times k}$, $S \in \mathbb{R}^{p \times k}$ and $H \in \mathbb{R}^{k \times n}$. Producing a matching ordering/clustering of the columns of the two matrices.


We have extended this further to take any number of matrices...
Colon Cancer Correlation Matrices: Four data sets $k = 12$
Chromosomal Location
Colon Cancer Correlation Matrices: Four data sets $k = 12$
Colon Cancer: Adding extra information, $k = 32$
Colon Cancer: Adding extra information, $k = 32$
Colon Cancer: Adding extra information, $k = 32$
Looking to the literature

Chromosome 1 p34.1-p34.2 – cluster 11

Deletion of the Alu-VpA/MycL1 (1p34.2) Locus is a Negative Prognostic Sign in Human Colorectal Cancer


Blokhin Cancer Research Center, Russian Academy of Medical Sciences, Moscow; 115478 Russia

E-mail: kashkin@barch.ru

Received February 3, 2000

Deletion of 1p34.3 is negative in cancer prognosis

Fig. 4. Relapse-free life span of patients with Alu-VpA/MycL1 disturbances in the Kaplan–Meier assessment. (A-DEL) Loss of the Alu-VpA/MycL1 allele; (A-MSI) Alu-VpA/MycL1 instability; (A-HET) heterozygous tumors indistinguishable in Alu-VpA/MycL1 from the normal mucosa; (c) censored without relapse; (P) Cox test for A-HET and A-DEL (A-MSI was not assessed because of the short terms of observation).
Looking to the literature

Chromosome 17 q21.2-q23.2 – cluster 9

In conclusion, fluorescent PCR technology coupled with an automated DNA sequencer appeared to be a very accurate and reliable method for detection of microsatellite alterations in genomic DNA extracted from paraffin-embedded material. Genomic alterations in the 17q21-23 region may affect prognosis of CRC as well as regulation of the mmm23 protein expression via an unknown underlying mechanism. Finally, the area flanking the D17S579 and MPO loci is likely to contain potential tumour suppressor gene(s) in which mutational inactivation play(s) a significant role for development and/or progression of at least some sporadic colorectal tumours.
Looking to the literature

Chromosome 17 q25 – cluster 12
Both MEN1 and HRASLS3, known tumour suppressors, are included in the identified cluster.
Discussion

Linking chromosomal information with correlation analysis we find that;

- the chromosomal information changes the gene-expression only clustering.
- the gene-expression data only links *some* of the gene neighbourhoods.
- many of the clusters have been previously described in either colorectal or another form of cancer.
The ongoing investigations include

- Looking at the CRC gene-expression directly rather than the correlation matrices.
- Considering how normalising the data could affect the ordering.
- Considering ways of picking an “optimal” number of clusters if one exists.