Nonnegative Decompositions with Resampling for Improving Gene Expression Data Biclustering Stability

Liviu Badea[1] and **Doina Ţilivea**[1]

Abstract. The small sample sizes and high dimensionality of gene expression datasets pose significant problems for unsupervised subgroup discovery. While the stability of *unidimensional* clustering algorithms has been previously addressed, generalizing existing approaches to *biclustering* has proved extremely difficult. Despite these difficulties, developing a stable biclustering algorithm is essential for analyzing gene expression data, where genes tend to be co-expressed only for *subsets* of samples, in certain specific biological contexts, so that both gene and sample dimensions have to be taken into account simultaneously.

In this paper, we describe an elegant approach for ensuring bicluster stability that combines three ideas. A slight modification of nonnegative matrix factorization that allows intercepts for genes has proved to be superior to other biclustering methods and is used for base-level clustering. A continuous-weight resampling method for samples is employed to generate slight perturbations of the dataset without sacrificing data and a positive tensor factorization is used to extract the biclusters that are common to the various runs. Finally, we present an application to a large colon cancer dataset for which we find 5 stable subclasses.

1 INTRODUCTION

Many real-life application domains, such as bioinformatics, text mining and image processing involve data with very high dimensionality. For example, gene expression datasets contain measurements of the expression levels for virtually all genes of a given organism (tens of thousands in eukaryotes), while the number of samples is still limited to at most a few hundreds.

Clustering is one of the most frequently used unsupervised data analysis methods in the field of gene expression data analysis. However, clustering such high-dimension small-sample data is meaningful only if a certain *stability* of the resulting clusters can be achieved. Unfortunately however, virtually all clustering methods that are currently used in this field tend to produce highly unstable clusters, especially when clustering genes. (The instability manifests itself either w.r.t. the initialization of the algorithm, as in the case of k-means, or w.r.t. small perturbations of the dataset, in the case of deterministic algorithms, such as hierarchical clustering.)

The stability of clustering has been addressed in previous work mainly for *unidimensional* clustering (dealing with either genes[2] or samples) [e.g.12]. The main idea of these approaches is to construct a *consensus* among a number of different clusterings obtained either by slight perturbations of the input dataset or due to different initializations in the case of nondeterministic algorithms. To construct the consensus, one needs the correspondence between the clusters of different clusterings. Most of the above mentioned approaches avoid determining the cluster correspondence by working with so-called connectivity matrices. Such a *connectivity matrix* T_{g1g2} has non-zero entries for the items g_1, g_2 that belong to a common cluster. The *consensus matrix* M_{g1g2} is then the average of the connectivity matrices for the different clusterings obtained in different runs.

Unidimensional clustering is not fully satisfactory for gene expression data analysis, where genes tend to be co-expressed only for certain *subsets* of samples, corresponding to specific biological contexts. Therefore, both the gene and the sample dimension have to be taken into account simultaneously.

Unfortunately, the above-mentioned approach based on consensus matrices cannot be applied to *bidimensional* clustering. This is due to the fact that in the case of biclustering one cannot simply deal with separate gene and sample connectivity matrices. To appreciate this in more detail we need a few notations. Let X_{sg} represent the gene expression matrix value for gene g in sample s, $S^{(i)}_{cg}$ the *membership degree* of gene g in cluster c of clustering i and $A^{(i)}_{sc}$ the mean *expression level of cluster* (biological process) c in sample s. Then, the connectivity matrices for genes and samples are $CS=S^{(i)T}\cdot S^{(i)}$ and respectively $CA=A^{(i)}\cdot A^{(i)T}$.

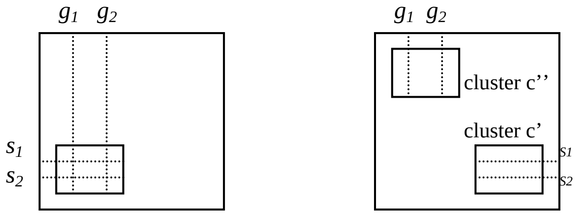


Figure 1.

Consider the two situations presented in Figure 1. In both situations genes (g_1, g_2) belong to the same gene cluster, so CS_{g1g2} is non-zero. Similarly, samples (s_1,s_2) belong to the same sample cluster, so CA_{s1s2} is non-zero in both situations as well. However, since (g_1,s_1) and (g_2,s_2) belong to the same *bicluster* only in the first situation (Figure 1, left), dealing with separate gene and sample connectivity matrices (CS and CA) would miss this essential distinction. The correct generalization of connectivity matrices to bidimensional clustering is what we call a *connectivity 4-tensor*:

$$C_{(s_1g_1)(s_2g_2)} = \sum_c (A_{s_1c} \cdot S_{cg_1}) \cdot (A_{s_2c} \cdot S_{cg_2}) \qquad (1)$$

which is in general not reducible to a (tensor) product of connectivity matrices: $C_{(s_1g_1)(s_2g_2)} \neq (\sum_{c'} A_{s_1c'} S_{c'g_1})(\sum_{c''} A_{s_2c''} S_{c''g_2})$.

[1] National Institute for Research in Informatics, email: badea@ici.ro

[2] In the following, we will refer to the items to be clustered as 'genes' and occasionally use other domain-specific terminology. However, the approach can easily be applied to other domains.

The *consensus 4-tensor* associated to different biclustering runs *i* would then be the average of the associated connectivity tensors:

$$M_{(s_1g_1)(s_2g_2)} = \sum\nolimits_i C^{(i)}_{(s_1g_1)(s_2g_2)} \Big/ r \qquad (2)$$

Unfortunately, explicitly computing and storing these connectivity and consensus tensors is practically infeasible for large gene expression datasets. Unlike the unidimensional case where the connectivity and consensus matrices are of sizes quadratic in the numbers of items to be clustered (e.g. genes), the 4-tensors above are of sizes $(n_s \cdot n_g)^2$, where n_s, n_g are the numbers of samples and genes respectively. In the colon cancer dataset analyzed below $n_s \approx 200$ and $n_g \approx 3000$, so we would have to deal with tensors of size $3.6 \cdot 10^{11}$.

Fortunately, there is a better way of constructing *stable biclusters*. Let us note that connectivity tensors are highly redundant (i.e. are of lower rank), the only reason for constructing them being due to the difficulty of determining the correspondence between similar biclusters in different clustering runs, especially when dealing with soft clustering algorithms.

To deal with this problem, we use the meta-clustering approach from [8,14], which is based on a positive tensor factorization (PTF) of the biclusters obtained in clustering runs *i*. This meta-clustering approach based on PTF solves in an elegant manner the cluster correspondence problem and tends to produce stable biclusters, but is still sub-optimal in certain respects. First, it uses nonnegative matrix factorization (NMF) [1,2] as base-level clustering algorithm. NMF performs very well for biclustering gene expression data, even for data with many irrelevant genes[1], but it tends to reconstruct the average expression levels of such irrelevant genes as superpositions of induced clusters. While this reduces the reconstruction error, it also produces artificial cluster membership coefficients for such irrelevant genes. Here, we solve this problem by slightly generalizing NMF to allow for "gene intercepts".

Secondly, PTF simultaneously determines the bicluster correspondence and constructs a consensus of the biclusters obtained in several runs of NMF *starting with different initializations*. Here, we consider an additional type of perturbation to the data based on resampling to ensure an increased stability of the resulting clusters. Various methods based on resampling have been applied in the context of unidimensional clustering (e.g. [12, etc]). Unfortunately, virtually all proposed approaches have significant drawbacks. For example, in *bootstrapping*, approximately one third of the original samples are discarded, potentially affecting the final results, especially in the small-sample case. The same holds for other *subsampling* approaches. On the other hand, methods based on resampling with replacement may be affected by spurious clusters constructed from sample replicates. Recently, Dresen et al [13] introduced a resampling method based on so-called *continuous weights* that avoids these problems by a *simulated* resampling, in which the (integer) numbers of resamplings of each sample are replaced with continuous weights. The difficult part consists in adapting the specific clustering algorithm[2] to work with such weighted samples instead of the resampled ones.

In this paper, we show how NMF can be generalized to deal with continous-weight resampling. We apply our approach to a large colon adenocarcinoma dataset [10,11] for which we discover 5 stable clusters, one of these containing normal colon samples.

2 BICLUSTERING USING NONNEGATIVE MATRIX FACTORIZATIONS WITH INTERCEPT

An elegant method of biclustering consists in factorizing the gene expression matrix *X* as a product of an $n_s \times n_c$ (samples × clusters) matrix *A* and an $n_c \times n_g$ (clusters × genes) matrix *S* [3]

$$X_{sg} \approx \sum\nolimits_c A_{sc} \cdot S_{cg} + So_g \qquad (3)$$

subject to additional nonnegativity constraints:

$$A_{sc} \geq 0,\, S_{cg} \geq 0,\, So_g \geq 0 \qquad (4)$$

which express the obvious fact that expression levels and cluster membership degrees cannot be negative.

Factorization (3) differs from the standard NMF factorization [1,2] by the additional "*gene intercept*" *So*, whose main role consists in absorbing the constant expression levels of genes, thereby making the cluster samples S_{cg} "cleaner".

The factorization (3-4) can be regarded more formally as a constrained optimization problem:

$$\min f(A,S,So) = \frac{1}{2} \| X - A \cdot S - e \cdot So \|_F^2 = \frac{1}{2} \sum_{s,g} \left(X - A \cdot S - e \cdot So \right)_{sg}^2 \qquad (5)$$

subject to the nonnegativity constraints (4). This problem can be solved using an iterative algorithm with the following multiplicative update rules (which can be easily derived using the method of Lee and Seung [2]):

$$\begin{aligned} A_{sc} &\leftarrow A_{sc} \frac{\left(X \cdot S^T \right)_{sc}}{\left(\left(A \cdot S + e \cdot So \right) \cdot S^T \right)_{sc} + \varepsilon} \\ S_{cg} &\leftarrow S_{cg} \frac{\left(A^T \cdot X \right)_{cg}}{\left(A^T \cdot \left(A \cdot S + e \cdot So \right) \right)_{cg} + \varepsilon} \qquad (6) \\ So_g &\leftarrow So_g \frac{\left(e^T \cdot X \right)_g}{\left(e^T \cdot \left(A \cdot S + e \cdot So \right) \right)_g + \varepsilon} \end{aligned}$$

where *e* is a column vector of 1 of size equal to the number of samples and ε is a regularization parameter (a very small positive number).

The algorithm initializes *A*, *S* and *So* with random entries, so that (slightly) different solutions may be obtained in different runs. (This is due to the non-convex nature of the optimization problem (5), which in general has many different local minima.)

We can view the different solutions obtained by the generalized NMF_i algorithm as *overfitted* solutions, whose *consensus* we'll need to construct. For combatting overfitting, we consider additional perturbations using continuous weight resampling as explained below.

We have observed experimentally that adding intercepts to standard NMF leads to significant improvements in the quality of the recovered clusters. More precisely, the genes with little variation are reconstructed by the standard NMF algorithm from combinations of clusters, while NMF_i uses the additional degrees of freedom *So* to produce null cluster membership degrees S_{cg} for these genes. Moreover, NMF_i recovers with much more accuracy than standard *NMF* the original sample clusters, the standard NMF

[1] i.e. genes that show little co-variation with other genes.

[2] [13] shows how to deal with correlation-based hierarchical clustering in this context.

[3] Recall from the introduction that X_{sg} represents the gene expression level of gene *g* in sample *s*, S_{cg} the *membership degree* of gene *g* in cluster *c* and A_{sc} the mean *expression level of cluster* (biological process) *c* in sample *s*.

algorithm being confused by the cluster overlaps. (See the Figure in Supplementary material at www.ai.ici.ro/ecai08/). This improvement in recovery of the original clusters is very important in our application, where we aim at a correct sub-classification of samples.

3 NMF WITH CONTINUOUS WEIGHT RESAMPLING

A frequently used method to obtain more stable clusters consists in building a consensus of several individual clusterings constructed from perturbations of the original dataset. As already mentioned in the Introduction, various types of perturbations based on *resampling* have been applied in the context of one-way clustering (e.g. [12]). However, all of these have drawbacks related either to loss of precious original data (a problem which is exacerbated in the case of small sample sizes), or to potential spurious clusters built from replicates of samples resampled several times. Recently, Dresen et al. [13] have addressed this problem by generalizing the (integer) numbers of resamplings of each sample to *continuous weights*. This retains the full dimensionality of the original data and has proved superior to bootstrapping especially for small numbers of samples. However, the approach requires modifying the original clustering algorithm to simulate working with "continuous numbers of samples". While [13] show how this can be done with correlation-based hierarchical clustering (by modifying Pearson correlation to take into account weighted samples), generalizing this approach to NMF factorization is non-trivial.

In the following we show how NMF$_i$ can be adapted to deal with continuous weight resampling.

The distribution of a drawing with replacement is the binomial distribution, which is approximated by the Poisson distribution for large numbers of observations. Since in a bootstrap sample the expected value and the variance is 1, [13] used a continuous approximation of the Poisson distribution, namely a log-normal distribution with mean and variance 1.

In the following, we assume that the continuous sample weights w_s are drawn from a log-normal distribution with equal mean and variance *M*. (The results improve as *M* is increased.)

Generalizing NMF$_i$ to deal with continuous weight resampling amounts to replacing the optimization problem (5) by the following:

$$\min f(A,S,So) = \frac{1}{2}\sum_{s,g} w_s \left(X - A\cdot S - e\cdot So\right)^2_{sg} \qquad (7)$$

The associated multiplicative update rules can be easily shown to take the following form:

$$A_{sc} \leftarrow A_{sc} \frac{\left(X\cdot S^T\right)_{sc}}{\left(\left(A\cdot S + e\cdot So\right)\cdot S^T\right)_{sc} + \varepsilon} \qquad (8.1)$$

$$S_{cg} \leftarrow S_{cg} \frac{\left(A^T\cdot W\cdot X\right)_{cg}}{\left(A^T\cdot W\cdot\left(A\cdot S + e\cdot So\right)\right)_{cg} + \varepsilon} \qquad (8.2)$$

$$So_g \leftarrow So_g \frac{\left(e^T\cdot W\cdot X\right)_g}{\left(e^T\cdot W\cdot\left(A\cdot S + e\cdot So\right)\right)_g + \varepsilon} \qquad (8.3)$$

where $W = \mathrm{diag}(w_s)$ is the diagonal matrix with w_s on the diagonal. We will call the factorization obtained by solving the optimization problem (7) a *w*-factorization and the corresponding algorithm NMF$_{ir}$.

It is interesting to note that *w*-factorizations can be reduced to standard NMF factorizations, but only in the absence of intercepts. More precisely, we have the following result.

Proposition. In the case of no intercepts, (*A*,*S*) is a *w*-factorization of *X* if and only if ($V\cdot A$,*S*) is a standard factorization of $V\cdot X$, where $V = diag\left(\sqrt{w_s}\right)$.

The fact that intercepts interact with resampling weights shows that the generalization is non-trivial.

4 CONSENSUS CLUSTERING WITH PTF

Starting with a number of NMF$_{ir}$ runs

$$X \approx A^{(i)}\cdot S^{(i)} + So^{(i)} \qquad i = 1,\ldots, r \qquad (9)$$

we construct a *consensus biclustering* using a *Positive Tensor Factorization* (PTF) [3] of the biclusters[4], which simultaneously determines the bicluster correpondence α and the consensus biclustering (β,γ) [8,14]:

$$A_{s(ic)}\cdot S_{(ic)g} \approx \sum_{k=1}^{n_c} \alpha_{(ic)k}\cdot\beta_{sk}\cdot\gamma_{kg} \qquad (10)$$

where *s* are samples, *g*, genes, *c* clusters and *k* metaclusters (or "consensus clusters").[5] β and γ represent the *consensus* of $A^{(i)}$ and $S^{(i)}$ respectively. More precisely, the columns $\beta_{\cdot k}$ of β and the corresponding rows $\gamma_{k\cdot}$ of γ make up a *base set of bicluster prototypes* $\beta_{\cdot k}\cdot\gamma_{k\cdot}$ out of which all biclusters of all individual runs can be recomposed, while α encodes the *(bi)cluster-metacluster correspondence*. The factorization (10) can be computed using the following multiplicative update rules [8,14]:

$$\alpha \leftarrow \alpha * \frac{(A^T\cdot\beta)*(S\cdot\gamma^T)}{\alpha\cdot[(\beta^T\cdot\beta)*(\gamma\cdot\gamma^T)]}$$

$$\beta \leftarrow \beta * \frac{A\cdot[\alpha*(S\cdot\gamma^T)]}{\beta\cdot[(\alpha^T\cdot\alpha)*(\gamma\cdot\gamma^T)]} \qquad (11)$$

$$\gamma \leftarrow \gamma * \frac{[\alpha*(A^T\cdot\beta)]^T\cdot S}{[(\alpha^T\cdot\alpha)*(\beta^T\cdot\beta)]^T\cdot\gamma}$$

where '*' and '—' represent element-wise multiplication and division of matrices, while '·' is ordinary matrix multiplication. After convergence of the PTF update rules, the rows of γ are normalized to unit norm to make the gene clusters directly comparable to each other, whereas the columns of α are normalized such that $\sum_{i,c}\alpha_{(ic)k} = r$ (*r* is the number of runs). Then, NMF$_{ir}$ initialized with (β,γ,γ_0) is run[6] to produce the final factorization $X \approx A\cdot S + e\cdot So$.

The nonnegativity constraints of PTF meta-clustering are essential both for allowing the interpretation of $\beta_{\cdot k}\cdot\gamma_{k\cdot}$ as consensus biclusters, as well as for obtaining sparse factorizations. In practice, the rows of the correspondence matrix α tend to contain typically one or only very few significant entries. Therefore, α can be used to assess the *stability* of the individual clusterings ($A^{(i)}$,$S^{(i)}$,$So^{(i)}$).

To do this, we diagonalize all $\alpha_{(i)}$ by row permutations

$$\alpha'_{(ic')k} = \sum_c P^{(i)}_{c'c}\cdot\alpha_{(ic)k} \qquad (12)$$

[4] A tensor factorization is needed instead of a matrix factorization since biclusters are matrices.

[5] To simplify the notation, the indices *i* and *c* were merged into a single index (*ic*).

[6] γ_0 is obtained from the 1-dimensional NMF decomposition $So^{(i)}_g = \alpha^{(i)}_{00}\gamma_{0g}$ with the normalization $\sum_i \alpha^{(i)}_{00} = r$.

such that the largest elements[7] of $\alpha_{(i)}$ end up on the diagonal $\alpha_{(ik)k}$.

We then apply these row permutations to the gene cluster matrices $S^{(i)}$ which are thereby synchronized with the consensus matrix γ

$$S'^{(i)}_{c'k} = \sum_c P^{(i)}_{c'c} \cdot S^{(i)}_{ck} \,. \tag{13}$$

At this point, we can estimate the stabilities of individual entries of the gene cluster matrices using the following instability measure:

$$instab(S'^{(\cdot)}_{kg}) = \sqrt{\sum_i (S'^{(i)}_{kg} - \gamma_{kg})^2 \Big/ r} \Big/ \gamma_{kg} \,, \tag{14}$$

which gauges the deviation of the individual runs from the consensus. It can be easily shown that (14) is equivalent to

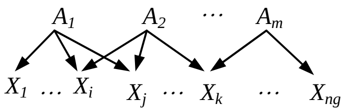
$$instab(S'^{(\cdot)}_{kg}) = \sqrt{std(S'^{(\cdot)}_{kg})^2 + \left(mean(S'^{(\cdot)}_{kg}) - \gamma_{kg}\right)^2} \Big/ \gamma_{kg} \tag{14'}$$

A similar measure can be defined for the sample matrices $A^{(i)}$.

While using α we can discard entire unstable clusters, our instability measure may be used to gauge our confidence in the individual gene or sample cluster values obtained.

5 EXPERIMENTAL EVALUATION

We first evaluated our approach on *simulated data* generated according to the following hidden-variable graphical model



in which the hidden variables correspond to potentially overlapping biclusters: $X = A \cdot S + \varepsilon$. The test contains 50 samples, 100 genes and the structure is random with 10 samples and 20 genes per bicluster. The logarithms of the hidden variables A were normally distributed with μ_{signal} ranging between 4 and 8, σ_{signal}=1 in the clusters and μ_{bkg} =3, σ_{bkgl}=1 outside.

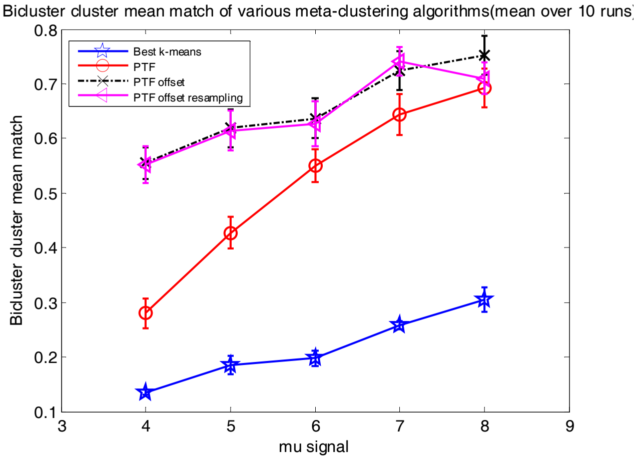


Figure 2. Variation of bicluster match with signal/noise ratio

Although all algorithms produce quite low relative errors $\varepsilon_{rel} = \| X - A \cdot S - e \cdot So \| / \| X \|$ (k-means – slightly higher ones), they behave differently when it comes to recovering the original clusters. Since the *match* of the recovered clusters with the original ones is more important than the relative error (see [8] for our definition of the *match* between two sets of possibly *overlapping* clusters), Figure 2 shows the variation of the match with the signal to background ratio. PTF with intercept and PTF with intercept and resampling behave very similarly, but outperform simple PTF [14] as well as k-means. Although we could not show that resampling is essentially better than PTF with intercepts in simulated data, we believe that it is useful for estimating cluster confidence factors in our real-life application.

Colon cancer dataset. The most frequent colon cancer type, *sporadic colon adenocarcinoma*, is very heterogeneous and its best current classification based on the presence or absence of microsatellite instabilities (MSI-L, MSI-H and MSS) [9] is far from ideal from the point of view of gene expression. To obtain a more accurate subclassification based on gene expression profiles, we have applied our approach to a large colon cancer dataset (204 samples) containing 182 colon adenocarcinoma samples from the *expO* database [10] and 22 control ("normal") samples from [11]. (All of these had been measured on Affymetrix U133 Plus 2.0 chips.) The combined raw scanning data was preprocessed with the RMA normalization and summarization algorithm. (The logarithmic form of the gene expression matrix was subsequently used, since gene expression values are approximately log-normally distributed.) After eliminating the probe-sets (genes) with relatively low expression as well as those with a nearly constant expression value[8], we were left with 3708 probe-sets. Finally, the Euclidean norms of the expression levels for the individual genes were normalized to 1 to disallow genes with higher absolute expression values to overshadow the other genes in the factorization.

An important parameter of the factorization is its internal dimensionality (the number of clusters n_c). To avoid overfitting, we estimated the number of clusters n_c as the largest number of dimensions around which the change in relative error $\frac{d\varepsilon}{dn_c}$ of the factorization of the real data is still significantly larger than the change in relative error obtained for a randomized dataset[9] (similar to [5]) – see also Figure 3 below. Using this analysis we estimated the internal dimensionality of the dataset to be around 5.

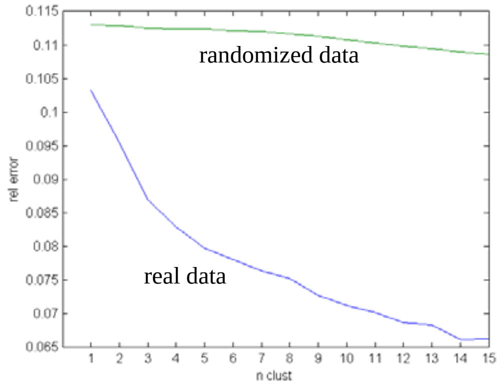


Figure 3. Determining the internal dimensionality of the dataset

We then ran PTF with 50 NMF$_{ir}$ iterations and n_c=5. Figure 4 depicts the sample cluster matrix A. Note that cluster 5 corresponds

[7] Since entire rows are permuted in the process of bringing the largest values of α on the diagonal, the largest value on a given column may not end up on the diagonal if it occurs on a row that had been permuted previously.

[8] Only genes with an average expression value over 100 and with a standard deviation above 150 were retained.

[9] The randomized dataset was obtained by randomly permuting for each gene its expression levels in the various samples. The original distribution of the gene expression levels is thereby preserved.

to the normal control samples from [11]. To make sure that this "normal cluster" is not a "batch effect" (due to the fact that we have combined two different datasets), we first looked at the expression of known housekeeping genes across the two datasets – overall, these turned out to have no particular dataset bias. Furthermore, the dataset from [11] contains besides normal samples from healthy individuals, also "normal" samples from individuals afflicted by early-onset colon cancer. We interpret the fact that a few of these cancer susceptibility samples but none of the samples from healthy individuals cluster in the colon cancer classes 1-4 as evidence against a systematic batch bias.

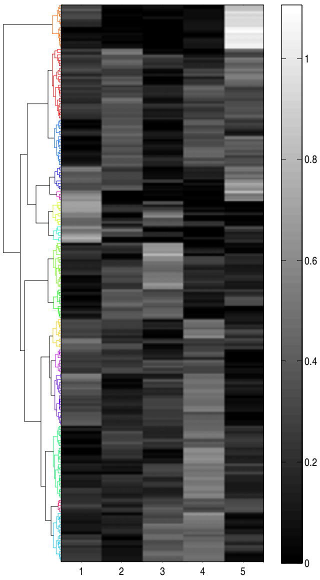


Figure 4. The sample cluster matrix *A*

The gene clusters contain genes with a well known involvement in colon cancer. For example, cluster 2 contains the regenerating islet-derived family member 4 *REG4*, which is known to be involved in inflammatory and metaplastic responses of the gastrointestinal epithelium[10] [PMID:12819006], its overexpression being an early event in colorectal carcinogenesis [PMID: 14550954]. Cluster 2 contains three additional genes from the same family, with documented oncogenic properties: *REG1B*, *REG1A*, *REG3A*. Cluster 3 contains several genes involved in the TGF-beta pathway: osteopontin (*SPP1*), activin A (*INHBA*), thrombospondin 1 (*THSB1*), the plasminogen activator inhibitor type 1 (*SERPINE1*), etc. Cluster 4 contains (with a high membership coefficient) the teratocarcinoma-derived growth factor 1 *TDGF1*, which has been proposed as a biomarker for colon and breast carcinoma [PMID:16951234]. TDGF1 expression has been recently shown to be controlled by the canonical Wnt/beta-catenin/TCF signaling pathway (the "classical" textbook pathway in colon cancer) [PMID:17291450], as well as by TGF-beta-like pathways [PMID: 17941089]. The cluster 1 gene *MYH11* has been very recently linked to microsatellite-stable HNPCC and sporadic colon cancer [PMID:17950328], while a polymorphism in the chemokine ligand 12 *CXCL12* has been found in colon cancer patients [PMID: 17143542]. Finally, the "normal" class 5 is characterized by genes down-regulated in colon cancer, such as the carcinoembryonic antigen-related cell adhesion molecule 7 *CEACAM7*, whose downregulation is known to be an early event in colorectal tumorigenesis [PMID:9135022]. (More details on the biclusters and the associated genes can be found in the supplementary material at www.ai.ici.ro/ecai08/.)

6 CONCLUSIONS

Soft biclustering is particularly difficult in the case of overlapping clusters, which are ubiquitous for gene expression data. Nonnegative factorizations like NMF are good for this purpose, but we show that they can be improved by adding intercepts. On the other hand, NMF factorizations depend on their initialization. Instead of regarding this as a drawback, we used PTF to construct a consensus factorization that hopefully reduces overfitting. Generating perturbations of the data by simulated resampling allow estimations of bicluster stability, which is especially important when looking at gene expression biclusters that typically contain hundreds of genes. Finally, we have applied the approach to a large colon cancer dataset, for which our approach finds 5 stable biclusters (one of which contains the genes active in the normal samples and down-regulated in colon cancer). Among the genes with the most significant coefficients, we find many with a known involvement in colon cancer. Our subclassification could thus be used to systematize the roles of these genes in the various subtypes.

7 REFERENCES

1. Lee D.D., H.S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, vol. 401, no. 6755, pp. 788-791, 1999.
2. Lee D.D., H.S. Seung. Algorithms for non-negative matrix factorization. Proc. NIPS 2000, MIT Press, 2001.
3. Welling M., Weber M. Positive tensor factorization. Pattern Recognition Letters 22(12): 1255-1261 (2001).
4. Cheng Y. Church G. Biclustering of expression data. Proc. ISMB-2000, 93-103.
5. Kim P.M., Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res. 2003 Jul;13(7):1706-18.
6. Brunet J.P., Tamayo P., Golub T.R., Mesirov J.P. Metagenes and molecular pattern discovery using matrix factorization. PNAS 101(12):4164-9, 2004, Mar 23.
7. Cheng Y, Church GM. Biclustering of expression data. Proc. ISMB 2000; 8:93-103.
8. Badea L. Clustering and Metaclustering with Nonnegative Matrix Decompositions. Proc. ECML-05, Vol. 3720, pp. 10-20.
9. Jass JR, et al. Characterisation of a subtype of colorectal cancer combining features of the suppressor and mild mutator pathways. J.Clin.Pathol. 52: 455-460, 1999.
10. expO. Expression Project for Oncology http://expo.intgen.org/expo/geo/goHome.do
11. Hong Y, Ho KS, Eu KW, Cheah PY. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. Clin. Cancer Res. 2007 Feb 15;13(4):1107-14.
12. Monti et al. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning, Vol. 52, No. 1-2. (2003), pp. 91-118.
13. Dresen Gana IM, et al. New resampling method for evaluating stability of clusters. BMC Bioinformatics. 2008, Jan. 24;9(1):42.
14. Badea L, Tilivea D. Stable Biclustering of Gene Expression Data with Nonnegative Matrix Factorizations. Proc. IJCAI-07, pp. 2651-2656.

[10] Due to lack of space, we refer to medical publications by their PubmedID.