Supplementary information to "Clustering and Metaclustering with Nonnegative Matrix Decompositions"

Metaclustering gene expression data: the Meyerson lung cancer dataset

Liviu Badea
AI lab, National Institute for Research in Informatics

In the following we show that metaclustering is successful at biclustering a large lung cancer dataset from the Meyerson lab [1].

Using HG-U95Av2 Affymetrix oligonucleotide microarrays, Bhattacharjee et al. [1] have measured mRNA expression levels of 12600 transcript sequences (genes) in 186 lung tumor samples (139 adenocarcinomas, 21 squamous cell lung carcinomas, 6 small cell lung cancers, 20 pulmonary carcinoids) and 17 normal lung samples (203 samples in total).

Since the raw CEL files were presumably processed by the authors with Affymetrix MAS4 software, certain gene expression value estimations are negative[1]. Therefore, we apply separate additive corrections for genes having negative values so that all gene expression values become positive. (We avoid a global scaling of the samples since various forms of cancer may affect a large fraction of the genome.)

For testing our metaclustering algorithm, we first selected a subset of genes that are differentially expressed between the classes. ANOVA, pairwise t-tests or SAM [2] could have been used for this purpose, but we preferred the following SNR measure, since it discourages large intra-class STD in both classes:

$$SNR_{class} = \frac{\mu_{class} - \mu_{normal}}{\sigma_{class} + \sigma_{normal}} .$$

More precisely, we selected the genes with an average expression level over 100 [2] and having $|SNR_{class}| > 2$ for at least one of the classes (small cell, squamous or carcinoid). There were 251 such genes.

Since adenocarcinoma is a very heterogeneous disease, whose subclasses are poorly understood at the molecular level, we discarded the adeno samples from the dataset and used the histological classification of samples provided in the supplementary material to the original paper [1] as a gold standard for the evaluation of the biclustering results.

To eliminate the bias towards genes with high expression values, the resulting restricted gene expression matrix was then normalized by separate scalings of the genes such that their norms (uncentred STDs) become equal.

Although nonnegative factorizations have the advantage of obtaining sparse and easily interpretable[3] decompositions, they cannot directly account for gene down-regulation. To

---

[1] Since MAS 4 uses a $PM - MM$ model, where PM=perfect match and MM=mismatch.
[2] For Affymetrix chips, expression levels below 100 are considered to be too low to be reliable.
[3] Since no complex cancellations of positive and negative terms are allowed.

deal with gene down-regulation in the context of NMF, we extend the gene expression matrix with new "down-regulated genes" g' associated to the original genes g as follows:

$$g' = \mathrm{pos}(mean(g_{normal}) - g)$$

where $mean(g_{normal})$ is the average of the gene over the *normal* samples, while pos(·) is the Heaviside step function.[4]

We then used our metaclustering algorithm to factorize the extended gene expression matrix as follows (with n_c=4 and running PTF over 20 NMF runs):

$$X_{sg} \approx \sum_c A_{sc} \cdot S_{cg}$$

(The matrix X has 64 rows (samples) and 2×251=502 columns (extended genes).)

The following Figures 1 and 2 show the resulting sample (A) and gene cluster (S) matrices.

Note that the algorithm has recovered the sample clusters with high accuracy (as can be seen in Figure 1).

The relative error of the decomposition is $\varepsilon = \frac{\|X - AS\|_F}{\|X\|_F} = 0.2722$, while the relative errors of the 20 individual runs are slightly higher:

0.2723	0.2723	0.2723	0.2734	0.2723	0.2727	0.2723	0.2723	0.2723	0.2723
0.2723	0.2726	0.2726	0.2723	0.2728	0.2724	0.2723	0.2724	0.2724	0.2723

More details on the gene clusters can be found in the accompanying file http://www.ai.ici.ro/ecml05/gene_clusters.xls). Cluster membership degrees S_{cg} were considered significant if they were larger than the threshold

$$\theta_g = \frac{1}{\sqrt{n_g}}.$$

Note that the overlap between the small cell and carcinoid sample clusters[5] has a biological interpretation: both contain samples of tumors of neuroendocrine type. The low mixing coefficients indicate however that carcinoids are highly divergent from the malignant small cell tumors.

We also looked in detail at some known marker genes. For example, the known small cell marker *ASCL1* (achaete scute 1) is *specific* to the small cell cluster, while *KRT5* (keratin 5) is specific to the squamous cluster.

On the other hand, known proliferative markers like *PCNA* (proliferating cell nuclear antigen), *MCM2* and *MCM6* are *common* to small cell *and* squamous clusters, as expected.

Overall, our metaclustering algorithm proved quite robust at rediscovering the known histological classification of the various lung cancer types in the Meyerson dataset.

---

[4] pos(x) = x if $x > 0$ and 0 otherwise.

[5] Columns 3 and 1 of A in Figure 1.

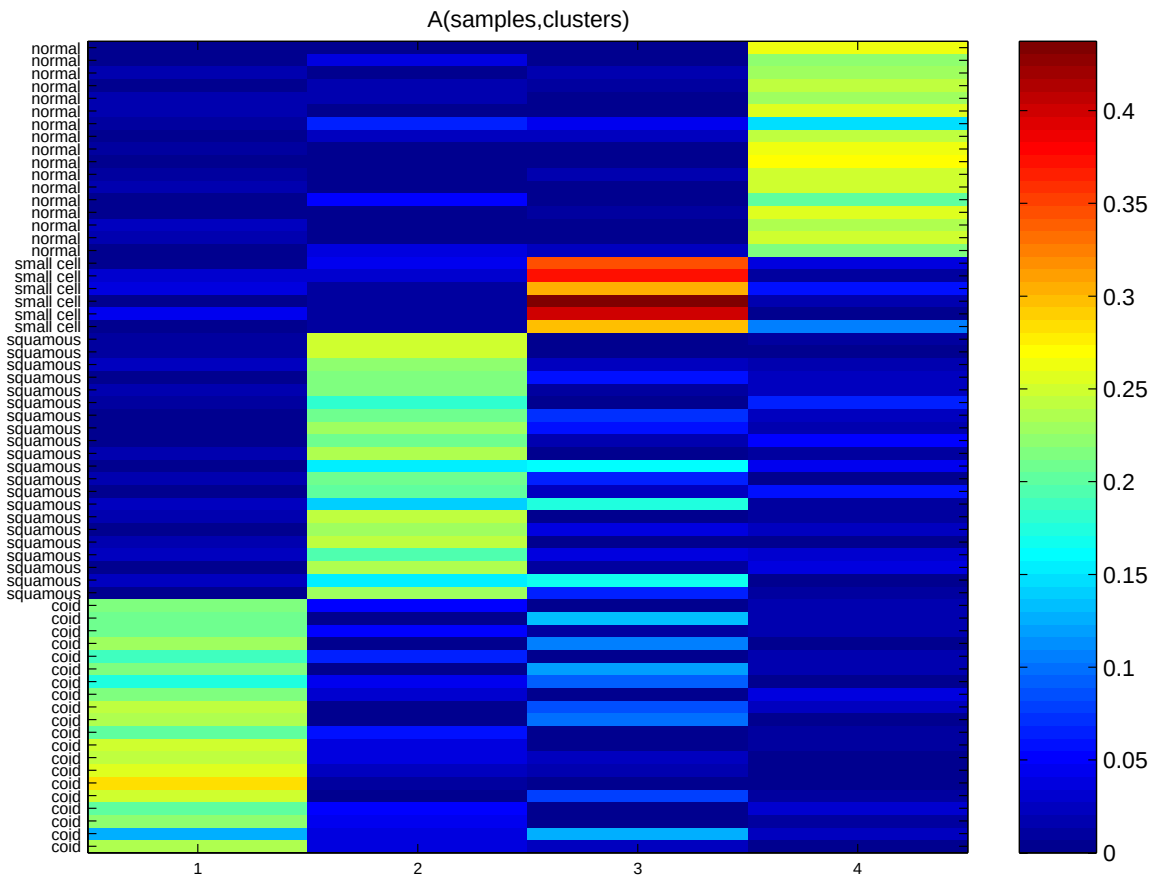

Figure 1. The sample clusters (significance threshold $\theta_s = \frac{1}{\sqrt{n_s}} = 0.125$)

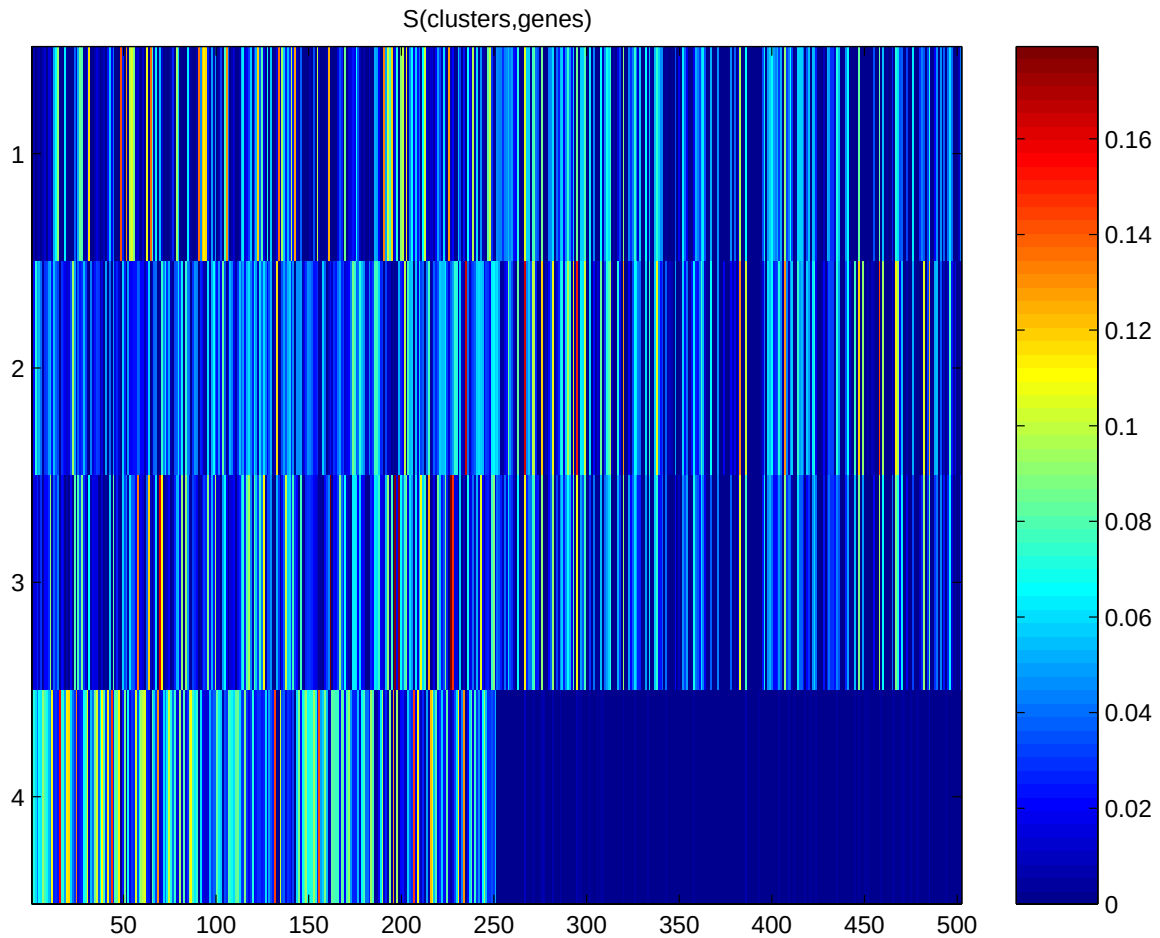

Figure 2. The gene clusters (the first 251 columns correspond to the original genes, while the last correspond to "down-regulated genes", as explained in the text; the significance threshold is $\theta_g = \frac{1}{\sqrt{n_g}} = 0.0446$)

References

1. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001 Nov 20;98(24):13790-5.
2. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001 Apr 24;98(9):5116-21. Epub 2001 Apr 17. Erratum in: Proc Natl Acad Sci U S A 2001 Aug 28;98(18):10515.