Applying independent component analysis on preprocessing of yeast gene expression data

Ville Viitaniemi 44374L

31st January 2001

The method of independent component analysis has previously been applied to analysis of natural image patches. There basis functions have been found with resemblance to the working of the early visual system. This work takes first stepsi into direction applying similar methods to analysis of gene expression time series.

1 Introduction

Improvements in DNA microarray technology ([16][15]) have made it possible to monitor gene expression levels on genomic level in a straightforward manner. An organism can be subjected to a specific treatment during which its gene expression levels are monitored and collected as a time series.

This way one obtains vast amounts of experimental raw data about the behaviour of the organism under different environmental conditions. But possessing this raw data is still far away from the real goal, understanding the actions and interactions of various genes.

The large data volumes themselves make automated methods very attractive. Often clustering methods are applied to the data. As usual in knowledge extraction tasks, the role of data preprocessing is important. In describing the expression profile of a gene one often uses some heuristic features (such as peak location) or simplified profile models (triangle model, ref.)([2][3]). In general, these various measures perform pretty well. Anyway, it would be interesting to also have some mathematically more justified method.

This project takes a look at applying independent component analysis (ICA, [13][4][12]) to the gene expression data. The approach is to take short windows from the data and seek for such kind of linear transform (set of filters) that when representing the data as linear combination of some basis functions, the combination coefficients would be statitically independent.

The approach gets its inspiration from the statistical analysis of natural images, where the requirement of independence leads to filters that are very much like those observed in early mammal vision([14][7] [9][10]). Moreover, some properties of higher visual processing can also be thought to be derived from the condition of statistical independence [8].

As a extensive treatment of the subject is out of the reach of this project, we merely take look at one specific approach of applying the ICA. Moreover, as a goal-oriented project would preferably need consultations with a biologist to discuss the sensibility and significance of the results, we settle for just making couple of observations on the data and the method.

2 Methods

The data set considered was a publicly available data set describing the behaviour of expression levels of 2496 genes in yeast (Saccharomyces cerevisiae) during various treatments. The data set was obtained from HUT Bioinformatics course homepage ([5]).

Let x_i be a $1xL$ vector describing gene expression levels of a single gene during a time window of length L time points. Let $\{x_i\}_{i=1}^{N}$ be the N element set of all such vectors from all the different genes and with different starting times. Our data matrix $\mathbf{x}$ is then the concatenation of all the elements of the set:

$$\mathbf{x} = (x_1 x_2 \cdots x_N) \tag{1}$$

Thus, it is a LxM matrix. The independent component analysis finds such representation

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2}$$

($\mathbf{s}$ is a LxN and $\mathbf{A}$ a LxL matrix)
that the rows of $\mathbf{s}$ are statistically as independent of each other as possible.

The particular method used here is the FastICA algorithm[11] with a $g(u) = u^3$ nonlinearity to estimate the mixing matrix $\mathbf{A}$. In essence the method first normalises the data, starts from a random mixing matrix, and alternates the steps of approximate Newton iteration and of normalisation of the matrix until convergence. The actual FastICA algorithm is implemented by a publicly available Matlab library[6].

When using this patch approach the problem with the data set was that most of the treatments contained only small amount of data points. It was not wanted that a boundary between two treatments would occur inside one patch. Consequently only 3 longest time series with 14,15 and 18 data points (labeled Elu, cdc15 and alpha) and points were chosen to be examined.

The FastICA algorithm was run on each of the treatment data sets separetely and thereafter on each combination of these three. First a window length of 7 time points was used, then length of 14 points.

3 Results

Some of the results from the experiments are shown in figures 1 to 4. The plots show the mixing matrix columns, which act as basis time patches when reconstructing the time series from the ICA results. Since FastICA produces independent components of unit variance, the magnitude of filter coefficients directly translate to the importance of each filter.

Figures 1 and 2 show the results from the runs with window length 7. figure 1 displaying the results of the longest time series alone (treatment alpha), figure 2 has all the three treatments lumped together.

Figures 3 and 4 show the results for window length 14 in the same manner.

Couple of observations can be made from the pictures. Some of the filters are very localised in time. Others have smoother relationship between neighbouring time points. At least using one's imagination some filters can be recognized to have a selectivity for some temporal frequencies (e.g. figure 3, 3rd and 5th rows), possibly even spatially modulated in a Gabor-like fashion.

It seems to be the case that when shorter patches are used, it is more important to have a set of localised filters centered on most of the time points. In the ecperiments with longer time windows a larger variety of filter forms was observed. This is at least partly due to the fact that the longer time windows had essentially fixed locations in the time series since there was room for only small variation of starting time .

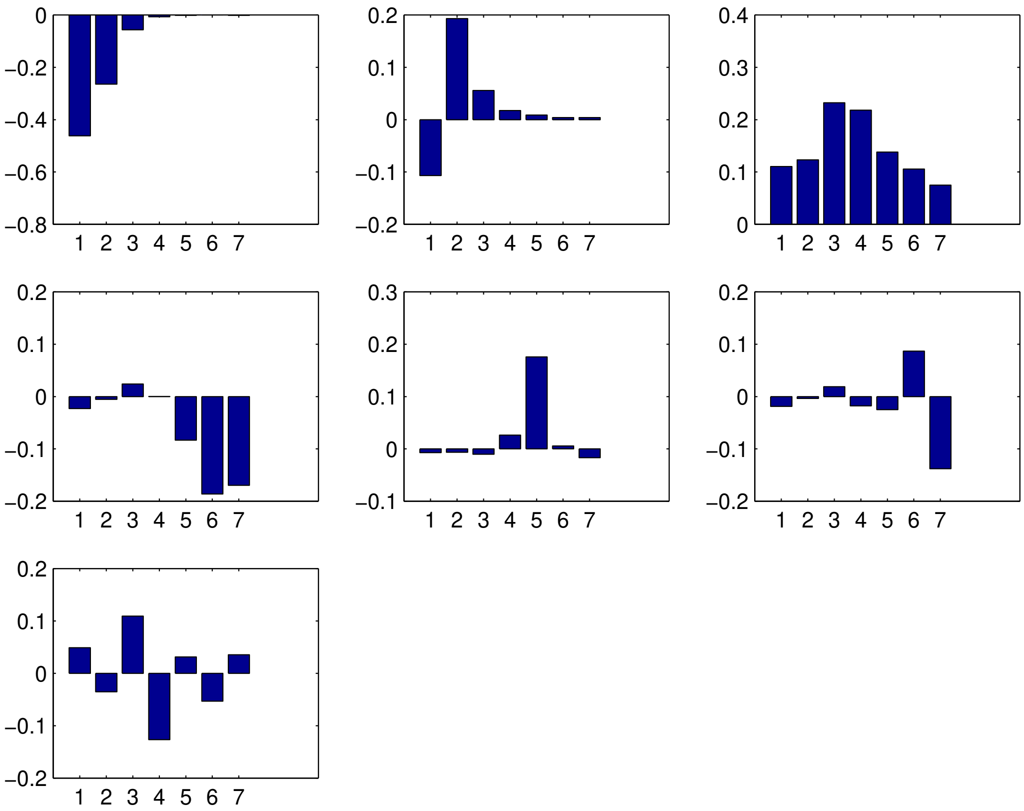

Figure 1: L=7, treatment alpha

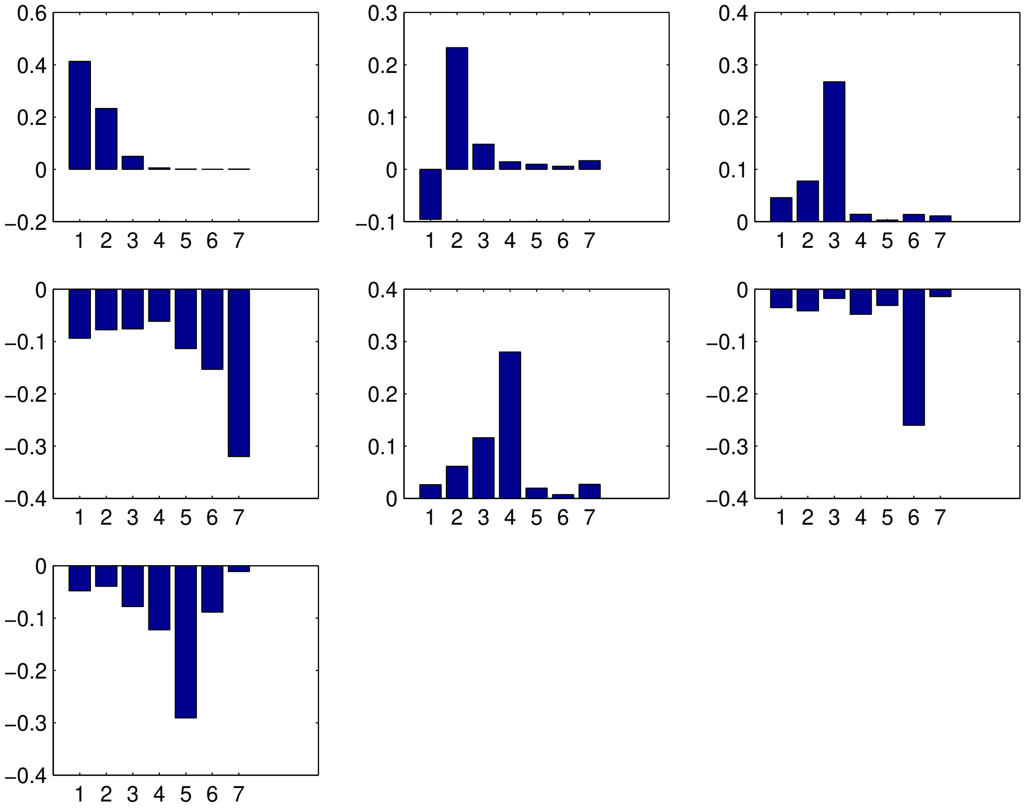

Figure 2: L=7, treatments alpha, Elu and cdc15

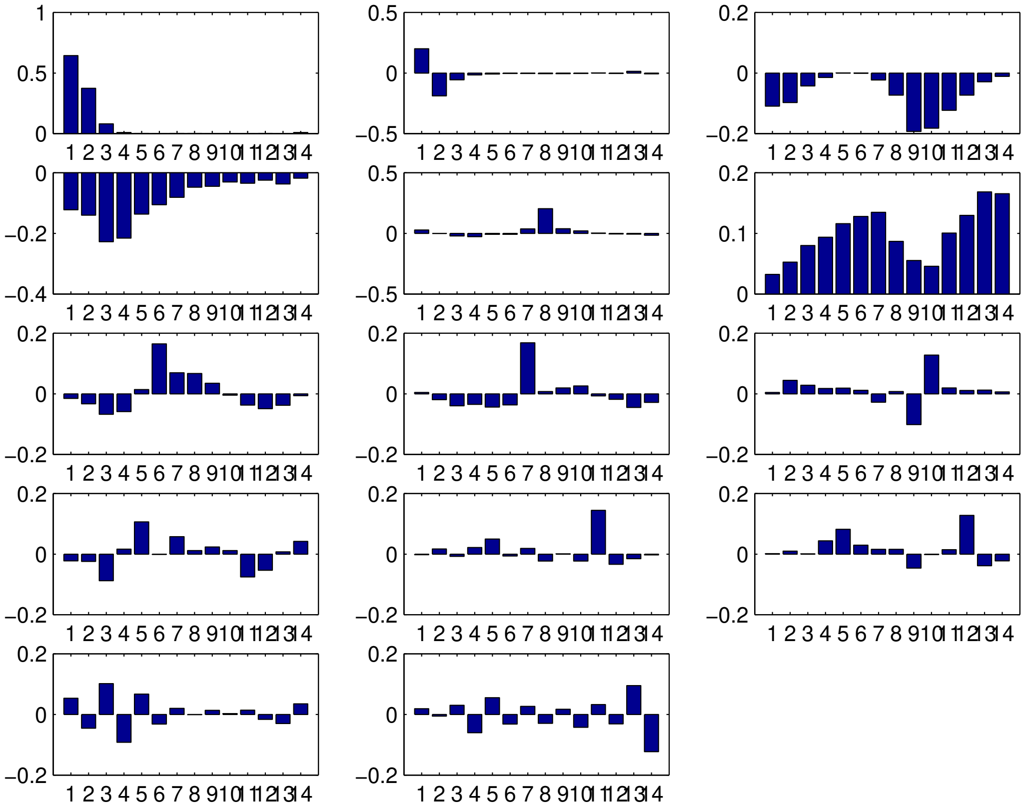

Figure 3: L=14, treatment alpha

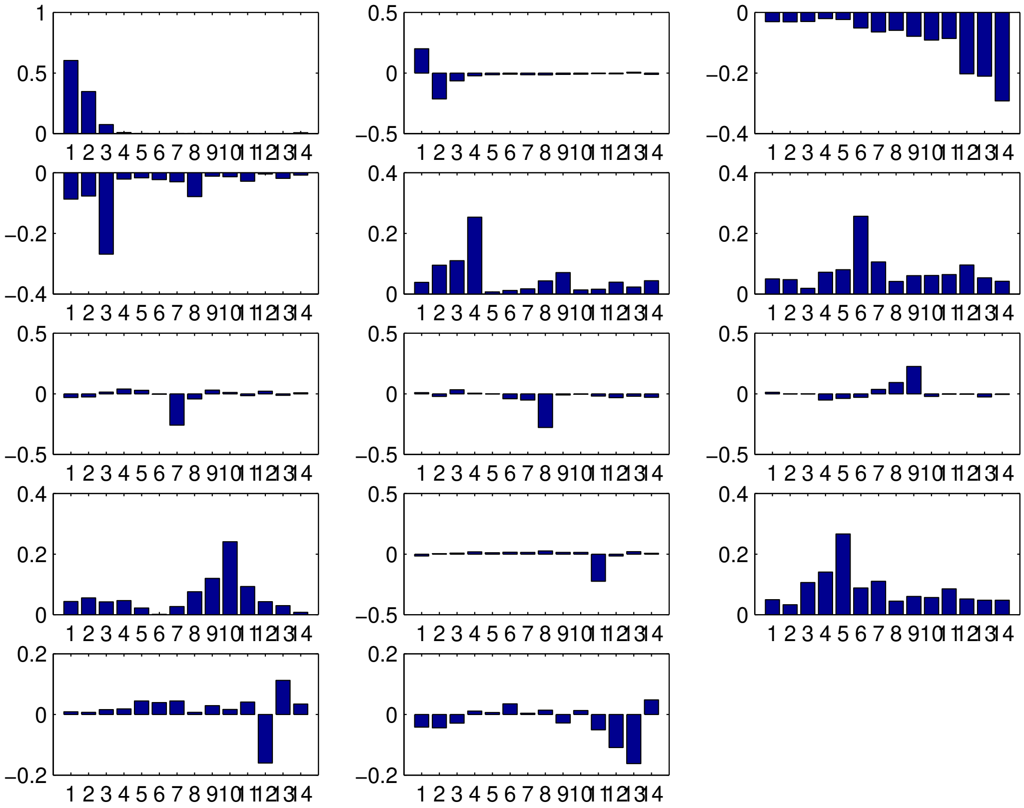

Figure 4: L=14, treatments alpha, Elu and cdc15

This should be tested with further experiments.

There could also to be tendency of richer, more specialised set of filters in the basis derived from a single treatment data compared to more general basis for many treatments (e.g. figure 1 versus figure 2). The number of experiments performed and data used was so small that this remains merely a possibility.

Up to now the actual contents of the independent components were completely ignored.

4 Conclusion and discussion

It seems initially to be that ICA of gene expression has some interesting characteristics. At least they warrant some more experiments, both confirming the the above results with more experiments and better data, and along partly different guidelines.

It's not known whether ICA turns out to be of any use when trying to understand gene expression. ICA may be able to provide complementary information to some other approaches. For example when compared with PCA based approaches (e.g.[1]) ICA can extract different kind of information since the basis functions are not constrained to be orthogonal.

There is at least two different approaches that could be employed when trying to characterise genes with ICA. One would be to obtain a general ICA basis from a large sample. Then the chacteristic features (of some set) would lie in the independent components (the expansion coefficients). In the other hand, one perform ICA on small subsets of data. Then the subsets could be compared by comparing the bases.

Assuming that the ICA algorithm works properly (as ot seems to) the ICA basis itself is a valid construct. It simply is a maximally independent basis for a linear transformation. But it would be interesting to know, how the well the ICA model describes data in reality . That is, how independent the linearly transformed components actually are.

It might be worthwhile to try some other approaches to independence if the basic linear model doesn't seem to be enough. Sensible extensions could be topographic ICA and nonlinear ICA models.

References

[1] Alter, O., Brown, P.O., Botstein, D., Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences 97(2000) 18, pp. 10101-10106

[2] Ben-Dor, A., Yakhini, Z., Clustering gene expression patterns. In: Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99), 1999, pp. 33-42

[3] Chen, T., Filkov, V., Skiena, S.S., Identifying Gene Regulatory Networks from Experimental Data In: Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99), 1999, pp. 94-103

[4] Comon, P., Independent component analysis - a new concept? Signal Processing 36 (1994) pp. 894-906

[5] Helsinki University of Technology, WWW page of the course Tik-61.181. [http://www.cis.hut.fi/Opinnot/Tik-61.181/projektityo.shtml], 31.1.2001

[6] The WWW page of FastICA Matlab package. [http://www.cis.hut.fi/projects/ica/fastica/], 31.1.2001

[7] Hoyer, P.O., Hyvärinen, A., Independent Component Analysis Applied to Feature Extraction from Colour and Stereo Images. Network: Computation in Neural Systems, 11(2000)3 pp.191-210

[8] Hyvärinen, A.Hoyer, P.O., Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. Neural Computation,(2000), in press

[9] Hubel, D., Wiesel, T., Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology, 160(1962), pp. 106-154

[10] Hubel, D., Wiesel, T., Receptive fields and functional architecture of monkey striate cortex. Journal of Physiology, 195(1968), pp. 215-243

[11] Hyvärinen, A., Oja, E., A fast fixed point algorithm for independent component analysis. Neural Computation, 7(1997) 9, pp. 1483-1492

[12] Hyvärinen, A., Survey on independent component analysis. Neural Computing surveys, 2(1999), pp. 94-128

[13] Jutten, C., Herault, J., Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Processing 24 (1991), pp.1-10

[14] Olshausen, B.A., Field, D.J., Emergence of single cell receptive field properties by learning a sparse code for natural images. Nature, (1996) 381, pp.607-609

[15] Ramsay, G., DNA chips: State-of-the-art. Nature Biotechnol. 16 (1998), p. 40-44

[16] Schena, M.,Shalon, D.,Davis, R.W.,Brown, P.O., Quantitative monitoring of gene expression patterns with a complementary DNA Microarray. Science 270 (1995), pp. 467-470

[17] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E.,Lander, E.S., Golub, T.R., Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA, 96 (1999)3, pp.2907-2912