ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

Gene expression data analysis using spatiotemporal blind source separation

Matthieu Sainlez[1], P.-A. Absil[2], and Andrew E. Teschendorff[3] *

1- CRISIA, Haute Ecole Robert Schuman,
Chemin de Weyler 2, B-6700 Arlon, Belgium (matthieu.sainlez@hers.be)

2- Department of Mathematical Engineering, Université catholique de Louvain,
B-1348 Louvain-la-Neuve, Belgium (http://www.inma.ucl.ac.be/~absil/)

3- Medical Genomics Group, Paul O'Gorman Building, UCL Cancer Institute,
University College London, London WC1 6BT, UK

Abstract. We propose a "time-biased" and a "space-biased" method for spatiotemporal independent component analysis (ICA). The methods rely on computing an orthogonal approximate joint diagonalizer of a collection of covariance-like matrices. In the time-biased version, the time signatures of the ICA modes are imposed to be white, whereas the space-biased version imposes the same condition on the space signatures. We apply the two methods to the analysis of gene expression data, where the genes play the role of the space and the cell samples stand for the time. This study is a step towards addressing a question first raised by Liebermeister, on whether ICA methods for gene expression analysis should impose independence across genes or across cell samples. Our preliminary experiment indicates that both approaches have value, and that exploring the continuum between these two extremes can provide useful information about the interactions between genes and their impact on the phenotype.

1 Introduction

A gene can be thought of as a "recipe" that specifies how to assemble amino acids in order to build a certain protein. The process by which the gene is translated into the protein is known as *gene expression*. This process involves copying the genetic information into a molecule of messenger RNA (mRNA), therefore, measuring in a cell the quantity of a given mRNA transcript is a way to assess the level of expression of the associated gene. Nowadays, due to the rapid development microarray technology, gene expression levels of more than 20,000 human genes are available for thousands of different cells samples.

The amount of a given mRNA transcript in a cell is determined by a whole range of biological processes that act to reduce or increase this number. Therefore, it seems reasonable to model the level of mRNA transcripts as a weighted sum of activation patterns associated to various biological processes. In Matlab

*This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

notation, this reads

$$X(:,j) = \sum_{k=1}^{n} A(:,k)B(k,j), \quad j = 1, \dots, T,$$

where $X \in \mathbb{R}^{m \times T}$ is the *gene expression matrix* containing the expression levels of m genes over T cell samples, $A(:,k) \in \mathbb{R}^m$ is the kth *activation pattern*, and $B(k,j)$ is the *weight* of the kth activation pattern in the jth cell sample. This suggests that blind source separation (BSS) via independent component analysis (ICA)—which aims at recovering statistically independent sources from linear mixtures—has potential for extracting biologically meaningful activation patterns from gene expression data, in an unsupervised way. This was confirmed in several studies that go back to Liebermeister [1]; see [2] and references therein.

Most ICA algorithms rely on a *contrast function* that evaluates the "level of statistical independence" of a collection of signals, and on an optimization algorithm to maximize the contrast function and hence recover signals that are as independent as possible. Different choices for the contrast function and for the optimization algorithm lead to different ICA methods. The results in [2] show that the various ICA algorithms do not differ much from each other in their ability to produce activation patterns that are similar to known gene pathways or regulatory motifs.

There is another choice that is made when applying ICA to gene expression data, namely: should independence be imposed across genes or across cell samples. *Independence across genes* means that the activation patterns (i.e., the columns of A) should be as independent as possible. *Independence across samples* means that the weights attributed to the activation patterns (which are found in the corresponding rows of B) should be as independent as possible. This freedom was already pointed out by Liebermeister [1, p. 54].

In practice, the level of statistical independence must be evaluated along a dimension of X that has sufficient length for the evaluation to be significant. Indeed, BSS estimates are poor when the number of available observations is not much larger than the number of signals. In the early days of microarray technology, the best data sets had thousands of genes but less than a few dozens of cell samples. Because the dimension along cell samples was so small, independence across genes has been favored in the literature. However, with the wider availability of microarray technology, data sets with larger sample sizes are emerging, which makes statistics over cell samples reasonably reliable. For example, the Wang database that we analyze in this paper contains 285 samples.

In this paper, we investigate how the choice of imposing independence across genes or across samples impacts on the ability of ICA to recover known gene pathways. Moreover, we consider a middle way, in the spirit of the spatiotemporal ICA of Stone *et al.* [3], that makes it possible to mitigate between these two extremes. The contrast function of spatiotemporal ICA consists of a linear combination, tuned by a parameter α, of a statistical independence measure of the columns of A and a statistical independence measure of the rows of B. The term "spatiotemporal" comes from the application to fmri data where $X(i,j)$

gives the intensity of pixel number i at time j; by analogy, we will refer to independence across genes as "spatial" ICA ($\alpha = 0$) and independence across samples as "temporal" ICA ($\alpha = 1$).

Our version of spatiotemporal ICA uses a contrast function based on approximate joint diagonalization (JD) of covariance-like matrices, as in Theis *et al.* [4]. However, we impose the approximate joint diagonalizer to belong to the orthogonal group, which has the advantage of being a compact manifold (see, e.g., [5]). This leads us to propose two flavors of spatiotemporal ICA: a *time-biased* version, where the rows of B are uncorrelated ($BB^\top = I$), and a *space-biased* version, where the columns of A are uncorrelated ($A^\top A = I$).

2 Spatiotemporal ICA

Let $X \in \mathbb{R}^{m\times T}$ be a matrix containing the expression levels of m genes over T experiences (or cell samples). In this section, we present the concept of time-biased and space-biased spatiotemporal ICA for X. Because of the analogy with the better-known application to fmri [3], we refer to the first dimension of X (the dimension along genes) as the space dimension, and the second dimension (along cell samples) as the time dimension.

First, the spatial and temporal means are removed, which yields the new matrix

$$X := (I - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top)X(I - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^\top),$$

where $\mathbf{1}_m$ denotes the vector of all ones in $\mathbb{R}^m$ and the superscript $^\top$ denotes the transpose. Next, the dimension is reduced to n components using a truncated SVD, which yields

$$\hat{X} = U_n D_n V_n^\top$$

where $X = UDV^\top$ denotes an SVD of X with the elements of D in decreasing order, and $U_n = U(:,1:n)$, $D_n = D(1:n,1:n)$, $V_n = V(:,1:n)$. The ICA step *per se* is based on the observation that

$$\hat{X} = U_n D_n V_n^\top = \underbrace{U_n D_n W^{-1}}_{=:A}\underbrace{W V_n^\top}_{=:B},$$

for all $W \in \mathbb{R}_*^{n\times n}$, where $\mathbb{R}_*^{n\times n}$ denotes the set of all $n\times n$ invertible matrices. Given covariance-like functions C_i, $i = 1,\ldots,N$, and a *spatiotemporal parameter* $\alpha \in [0,1]$, spatiotemporal ICA seeks the matrix W that maximizes the contrast function

$$\begin{aligned}\tilde{f}_\alpha : \mathbb{R}_*^{n\times n} \to \mathbb{R} : W \mapsto &-\sum_{i=1}^{N}\left(\text{off}\,(\alpha C_i(B)) + \text{off}\left((1-\alpha)(C_i(A^\top))^{-1}\right)\right)\\ &= -\sum_{i=1}^{N}\left(\text{off}\left(\alpha W C_i(V_n^\top)W^\top\right) + \text{off}\left((1-\alpha)W(C_i(D_n U_n^\top))^{-1}W^\top\right)\right),\end{aligned}$$

where $\operatorname{off}(\cdot)$ returns the sum of the squares of the off-diagonal elements of its matrix argument, and the C_i's are covariance-like matrix-valued functions with the transformation property

$$C_i(W^\top V_n^\top) = W^\top C_i(V_n^\top) W,$$

for all W. Some examples of covariance-like functions are mentioned in [4, §4.1]. Once the optimal W is found, the matrices A of activation patterns and B of weights are given by $A = U_n D_n W^{-1}$ and $B = W V_n^\top$. The parameter α makes it possible to explore a continuum between the spatial model ($\alpha = 0$) and the temporal model ($\alpha = 1$). Note that when $\alpha = 1$, we recover the classical JD-based contrast function as, e.g., in JADE [6] (or see [7]).

Up to here, the development can be seen to be mathematically equivalent to the one in [4]. In this work, however, we restrict W to belong to the orthogonal group $\mathrm{O}(n) = \{W \in \mathbb{R}^{n\times n} : W^\top W = I\}$; that is, we maximize $f_\alpha := \tilde{f}_\alpha\big|_{\mathrm{O}(n)}$. This restriction is equivalent to imposing that B is "white", i.e., $BB^\top = I$. We call the method *time-biased* (TB-stICA), because the time signatures of the ICA modes are required to be white. Restricting W to $\mathrm{O}(n)$ has a computational advantage: $\mathrm{O}(n)$ is compact which, along with the continuity of the cost function, ensures that the maximum of f_α exists.

To obtain the space-biased flavor of JD-based spatiotemporal ICA (SB-stICA), we start from the decomposition

$$\hat{X} = U_n D_n V_n^\top = \underbrace{U_n W}_{=:A} \underbrace{W^{-1} D_n V_n^\top}_{=:B}$$

and follow a similar development. Observe that it is now A that is guaranteed to be white, in the sense that $A^\top A = I$.

3 Results and model validation

We applied our spatiotemporal ICA algorithms to the "Wang" [8] breast cancer data set. In order to assess the ability of the algorithm to recover existing pathways, we used a pathway enrichment index (PEI), as defined in [2]. Roughly speaking, the PEI counts the fraction of pathways that display a "sufficient matching" with at least one activation pattern. We used the same database of 536 pathways as in [2]. In our spatiotemporal ICA algorithms, the covariance-like matrices C_i are chosen as $\frac{n(n+1)}{2}$ JADE-like fourth-order cumulants, $n = 10$, and the approximate joint diagonalizer of f_α is sought using an algorithm based on Jacobi rotations, as in the original JADE algorithm [6]. Our Matlab implementation of the algorithm evolved from the stJADE algorithm of Theis [4].

Fig. 1 shows the obtained PEI as a function of the spatiotemporal parameter α. The leftmost point corresponds to spatial ICA (independence is assessed solely across genes, which is the usual practice as we mentioned in the introduction), and the rightmost point provides the PEI for temporal ICA (independence is evaluated purely across cell samples).

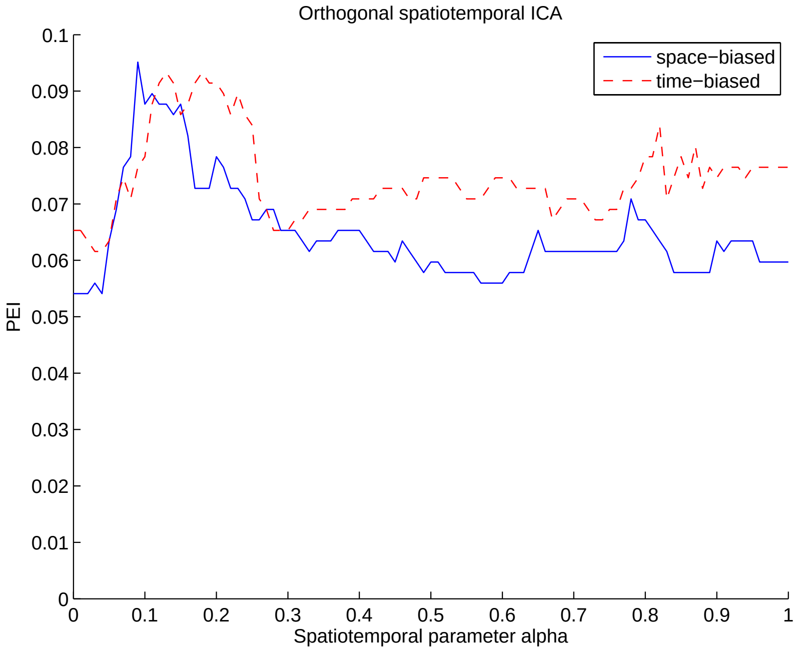

Fig. 1: PEI for database Wang: $X \in \mathbb{R}^{14913\times 285}$

One should not be led to believe that the features seen in Fig. 1 are generic. They have been obtained for a specific gene expression data set, a particular pathway database, and a spatiotemporal contrast function that involves several choices (a particular JD-based cost function, with $N = \frac{n(n+1)}{2}$ covariance-like matrices built as in JADE). Nevertheless, Fig. 1 shows some interesting features, which will need to be confirmed in forthcoming work. Notably, in this case, ICA across cell samples performed slightly better than ICA across genes. This suggests that the idea of imposing independence across cell samples is not to be dismissed. Another interesting finding is that introducing a bit of temporal ICA into spatial ICA may yield a PEI that is quite superior to the PEI obtained with the purely spatial ICA. Our preliminary experiments show that this peak around $\alpha = 0.1$ does not appear systematically with other gene expression data sets, however, in this case at least, spatiotemporal ICA has made it possible to reveal pathways (e.g., classicPathway, compPathway, AndrogenReceptor) that were not detected by purely spatial or purely temporal ICA. Since the measurement of gene expression remains a resource-consuming process, it is important to extract as much information as possible from the available data. Amongst the many data analysis methods available, spatiotemporal ICA clearly deserves attention.

4 Conclusion

We have proposed two kinds of spatiotemporal ICA algorithms that work by computing an orthogonal approximate joint diagonalizer of a set of covariance-like matrices. We have shown that the algorithms are valuable methods for discovering pathways in an unsupervised way from gene expression data. A more systematic analysis, with other gene expression data sets and other pathway databases, will be required to better evaluate how advantageous it may be to use a spatiotemporal ICA method rather than purely spatial or purely temporal ICA.

References

[1] Wolfram Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

[2] Andrew E. Teschendorff, Michel Journée, Pierre A. Absil, Rodolphe Sepulchre, and Carlos Caldas. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput. Biol.*, 3(8):1539–1554, 2007. doi:10.1371/journal.pcbi.0030161.

[3] J. V. Stone, J. Porrill, N. R. Porter, and I. D. Wilkinson. Spatiotemporal independent component analysis of event-related fmri data using skewed probability density functions. *NeuroImage*, 15:407–421, 2002.

[4] Fabian J. Theis, Peter Gruber, Ingo R. Keck, Anke Meyer-Bäse, and Elmar W. Lang. Spatiotemporal blind source separation using double-sided approximate joint diagonalization. In *Proc. EUSIPCO, Antalya, Turkey*, 2005. Available from http://fabian.theis.name/.

[5] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[6] J.F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings - F*, 140(6):362–370, 1993.

[7] Fabian J. Theis, Thomas P. Cason, and P.-A. Absil. Soft dimension reduction for ICA by joint diagonalization on the Stiefel manifold. Technical Report UCL-INMA-2008.155, Department of Mathematical Engineering, Université catholique de Louvain, 2008. Accepted for publication in the proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation (ICA2009).

[8] Yixin Wang, Jan GM Klij, Yi Zhang, and Anieta M Sieuwerts. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–679, 2005.