Statistical Embedding in Complex Biosystems

Enrico Capobianco*

Polaris
Science & Technology Park of Sardinia
Pula (CA), Italy

Summary

Complex high-dimensional systems represent an important area of interdisciplinary research in systems biology. Gene expression values obtained by microarray data represent a good example, owing to their various features that depend on biological network dynamics. This work emphasizes the role of blind source separation for dealing with dimensionality reduction and feature selection, and their useful combination with fuzzy rules, embedding principles and entropic measures. In particular, entropy and embedding are useful tools for controlling the robustness and stability of the decomposition of a system with larger than intrinsic dimensionality. As a result, the convergence to a small intrinsic dimensionality occurs by the means of least dependent components, seen as a minimal number of salient features.

1 Introduction

Information retrieval and signal extraction from high-dimensional data require a comprehensive effort aimed to integrate different biological information sources and obtain computationally efficient fusion of data streams in an automated way.

A general strategy to better understand the complex interactions occurring at the cellular level relies on building quantitative models from genomics (the study of genes and their functions), transcriptomics (the quantitative measurement of gene expression in a cell or tissue of a given organism), and proteomics (the study of the set of proteins encoded by a genome) data obtained by the means of high throughput technologies.

High dimensionality and redundancy are typical features of many real world complex systems, especially biological ones. Usually, through some form of dimensionality reduction, computational and methodological improvements can be achieved, and statistical inference can contribute to shed light on the interactions between variables.

We describe the role that a technique well-known in signal processing, machine learning and statistics, i.e. independent component analysis (ICA) [5, 7, 9], can play in genomics. In particular, it works as a dimensionality reduction methodology that allows for a convergence of the observed to the intrinsic genome data dimensionality, thus dealing with the typical *large variable and small sample* trade-off; it combines model flexibility with computational efficiency in decomposing the original signal space into statistically independent components, thus leading to a natural attempt to identify genes with different biological process, pathway, function

*This work has been conducted while the author was at Boston University, working at the Biomedical Engineering Department, where the experimental data have been provided.

characterization; last, it offers a straightforward feature selection tool, which allows for gene grouping according to co-expressed activity and for inference about the underlying regulatory dynamics.

We deal with microarray genomic data obtained from perturbation experiments where, in general, specific drugs are injected in the cells so as to observe the changes in the gene expression levels. It is normally accepted to treat the measurements via mining, normalization, and preprocessing steps, and we do that; but we also study possible ways to control the stability of the genomic system through fuzzy rules, embedding theory and entropy principles. We thus want to characterize a stable and robust least dependent decomposition of the genomic system under study that reveals the complex dynamics determined by the perturbation experiments.

The paper is organized as follows: Section 2 introduces the main methodological issues; Section 3 refers to data analysis aspects; Section 4 describes entropy fluctuation analysis, and Section 5 concludes the paper.

2 Methodological Aspects

2.1 Model Building

ICA is mainly designed for extracting latent structure from observed data mixtures; it is linked to projection pursuit regression (see for review of concepts and applications linked to ICA [10, 11, 12, 15, 18, 16, 17]), semiparametric modelling, and estimating equations theory (see, among others, [2]). Recently, ICA has started to play a role in genomics too, and interesting results have been obtained by [3, 4, 8, 20, 21, 14].

We believe that this method can simplify inference by approximating the gene system intrinsic dimensionality with a small set of estimated components. These components, when found to exist and are estimated, turn out to be statistically independent, or approximately so. This property would allow an efficient decomposition of the gene network dynamics and simplify their reconstruction, a task typically addressed in inverse (or ill-posed) systems biology problems by reverse engineering techniques.

A particular property of ICA is that the unobserved components that are identified (all but one) should be non-Gaussian for uniqueness and stability of solutions to be guaranteed. Intuitively, non-Gaussian directions in the projected space are considered the most interesting ones, as in this case the structure is maximally deviating from randomness and thus delivers relevant information.

However, the key aspect concerns the role that ICA naturally plays in high-dimensional problems, as it searches for sparsity and approaches the system's intrinsic dimensionality, or equivalently reduces its redundancy. In computational genomics, it is key to know what input variables (i.e., genes, regulators) are influencing the observed responses (target genes, pathways etc.), or which output variables are instead redundant because highly collinear with the former ones, or what kind of cascade effects can be observed at the genome level after perturbation or knock-out experiments.

The general ICA model has a simple structure. A mixture signal X includes $k-$dimensional vectors $x_i, i = 1, \ldots, n$ represented as independent non-Gaussian sources of information (or

components) S, consisting of $k-$dimensional vectors $s_j, j = 1, \ldots, m$, mixed linearly or non-linearly according to a mixing matrix A. The underlying random system may be non-linear, such as $Y = f(X) + error$; thus, a linear ICA is just an approximation for the system's dynamics, where both S and A must be estimated, given $S \in R^m$, $X \in R^n$, and $A \in R^{n \times m}$.

More specifically, consider at time t:

$$X = AS + \epsilon \tag{1}$$

ICA deals with this model through an approximate method that induces suitable variable decomposition by a linear (or non-linear) model representation. Consider a whitening transform, say, $A = U\Sigma V^t$, with $U \in R^{n \times m}$ and $V \in R^{m \times m}$ matrices with orthonormal columns, and Σ an $m-$ diagonal matrix. Apply then a rotational step, i.e. $G = QA = Q\Sigma^{-1}U^t$, for Q orthogonal under ICA and $Q = I$ under PCA. U^t projects the data into an $m-$dimensional source space, and Σ^{-1} scales the projections to leave unit variance. Furthermore, the final step is a numerical optimization search of the solution space, based on some criteria (from statistics, information theory, etc.) and depending on an algorithm (numerical, iterative, etc.).

Theoretical studies have stated that if the number of observed mixtures (n) is greater or equal than the number of sources (m), then it is possible to separate statistically independent sources provided that at most one of them is Gaussian ([5, 16]). The assumption underlying all ICA models is that the sources are statistically independent; since the sources S together generate a multi-dimensional probability density function (*pdf*) $P(S)$, in order to satisfy the basic assumption of statistical independence one needs the following joint source density factorization (at time *t*):

$$P[S(t)] = \Pi_{i=1}^{m} p[s_i(t)] \tag{2}$$

The independence condition is hard to verify in practice, as we can only approximate the unknown variables and their distributions, up to a certain degree. Also, noise and source interference can justify variable deviation degrees from a perfect source separation.

Various statistical assumptions may play an important role for identification and estimation purposes, even if one may adopt the quite simple non-parametric view of statistical inference, and thus work under relaxed assumptions. In particular, this last aspect has an important consequence: neither a parametric model nor a Gaussian context is required in order to carry out ICA.

Furthermore, ICA overcomes some limitations of principal component analysis (PCA); the latter is valuable under Gaussianity and deals with covariance-based or second-order statistics, thus only with linear independence (i.e. it can only decorrelate the data), while ICA exploits the higher order moments of the data distribution, or the the statistical information beyond the covariance function, and thus can detect stronger forms of dependence in the data.

Once the mixing matrix is estimated, the sources can be readily obtained by the inverse matrix (if $n = m$) or by the pseudo-inverse (when $n > m$). In the under-determined case when $m > n$ there is no unique inverse, which means that it exists an infinite number of independent components which are solutions of the linear problem. In an attempt to estimate both the

unknowns, the mixing matrix and the sources, several numerical algorithms can be efficiently applied.

The most popular choices are the *JadeR* or *Joint Approximate Diagonalization of Eigenmatrices* [7] and the *fastICA* [16, 17]; while the former offers a better control of the sequence of operations done by ICA, the latter is particularly useful for doing data pre-processing and for performing numerical optimization under different conditions.

Once obtained the estimates for the separating or de-mixing matrix $W(= A^{-1})$ through its pseudo-inverse $pinv(W)$, one would get back to an estimated version of the mixing matrix, and through $Y = WX$ retrieve S. While under conditions of perfect separation (and no noise) it is expected that $Y = WAS = S$, usually the solution can only be approximately true, up to permutation P and scaling D matrices, i.e. $Y = DPS$.

In other words, the model structure implies well-known ICA ambiguities; however, the scale of the sources (which are usually normalized) can always be adjusted by *ad hoc* factors inserted in the mixing matrix, while the order of the sources can be fixed according to some criterion (like the energy content, as in the *JadeR* algorithm).

2.2 Dimensionality

The intrinsic dimensionality, roughly speaking, represents the least number of functionally independent parameters needed to identify any observation or signal from a given class. Correspondingly, one goal in applications is to employ approximating systems with relatively small degrees of freedom such that only a few principal modes can be used to infer or reconstruct the internal dynamics, select the features of interest, and obtain reliable predictions for the target variables.

With ICA, the original dimensionality can be substantially reduced so as to reflect either the number of column vectors of the estimated mixing matrix (i.e. the directions of the data projections along which one hopes to detect structure), or equivalently the number of the component signals that are identified. The general interpretation is that the high-dimensionality of the observed data manifold can be gradually absorbed by a linear approximating subspace where the number of dominant singular values establishes the intrinsic dimensionality.

In real-world applications, decomposing systems with ICA leads to configurations with reduced dimensionality, in either the dominant modes which are ideally statistically independent or in possibly least dependent modes. Note that linearity in ICA is not a limiting structural property, as the usually required numerical optimization step implies a non-linear search for a solution (locally optimal). As the relations among variables in complex systems are likely to be non-linear, one goals is thus to find the information on the curvature of the manifold embedded by the data, instead of just examining locally linear regions.

2.3 Shrinkage

One problem in mining high-dimensional data spaces involves the separation of signal from noise, which often requires some kind of shrinkage. This idea refers to an established statistical technique for smoothing parameter estimates in stochastic systems endowed with sparse data

structure and thus obtain more robust and/or parsimonious signal reconstructions. A thresholding step is usually found to be a built-in component of estimation procedures where it is key to discriminate between informative and non-informative signal content according to some statistical criteria.

In genomics, experimentalists perturb genetic networks because they need to identify genes sensitive to some drugs, check cascade effects in the regulatory or transcriptional networks, distinguish between primary and secondary effects induced by the drugs on the genome. One thus needs to discriminate between the specific inherent biological characterization of the potential gene groups, and also with regard to noise-dependent dynamics.

Thresholding is the instrument that may validate hypotheses about the observed expression values, i.e. whether they are or not significantly different from an average test statistic, as done in significance tests, for instance. Shrinkage is a suitable tool for gene selection, by exploiting the discriminatory power of each estimated component; the genes significantly deviating from, say, the mean value computed over the gene profile in each estimated component, support with a certain confidence level the rejection of the null hypothesis of no presence of outlying effects.

The genes considered as outliers are thus endowed with a differentially expressed value compared to the average value computed in a specific component. The gene selection can be done via fixed intervals, i.e. by taking a scalar q times the standard deviation from the mean (approximately the 95% or 99% confidence levels), and thus identifying as outliers the genes whose values are outside the computed interval. A simple algorithm, described in Table 1, is proposed; it can be repeated for each estimated component through the following steps:

Gene Selection Algorithm
Step 1: *a* **pivot component** j *is considered;*
Step 2: *a* **confidence interval** *centered on a statistic and a related measure of deviation is computed from the available sample, while a critical level* q *is fixed depending on the desired degree of stringency in* **isolating the outliers**;
Step 3: *a* **thresholding step** *is then operated gene-wise along the* $k-$*dimensional genomic profile which is attached to each component.*

Table 1: Example of shrinkage-based gene selection.

A possible formulation for the confidence intervals is also introduced. When the mean and standard deviation are computed for the gene expression $k-$dimensional values s_j (given the source j), the interval is straightforward:

$$mq_k^j = mean[s_k^j] \pm q * std[s_k^j] \tag{3}$$

As mq_k^j delivers lower and upper limits l and u, respectively, the outlying genes are $out^j \notin (l, u)$.

Depending on the stringency of the intervals, one can expect that the groups of genes are maximally independent or least dependent on each other. This happens because of the statistical properties of ICA; the same property cannot hold in general with PCA, unless the data are Gaussian.

However, the selected genes may be subject to the influence of different biological processes underlying the same components that approximately and independently represent them. Equivalently, some genes might be detected by more than one component. Moreover, due to the fact that genes sharing the same biological pathway are expected to show relatively strong connectivity, it is expected that the component-selected groups include co-expressed genes, and also co-regulated ones (even if not necessarily).

Conversely, the fact that some genes may belong to different groups reflects the simultaneous presence of different underlying biological influences in each component. As a result, a certain level of source interference should be observed, and whose biological characterization (natural interference) should be ideally distinguished by the noise-dependent one (artificial interference).

Consequently, the presence of a certain degree of natural interference suggests that the hypothesis of least dependence fits well in a biological context. Thus, biological reasons support the fact that a small degree of dependence remains from ICA in the overlapping groups, without preventing ICA from the ability to perform an accurate gene feature selection.

3 Data Analysis

3.1 Pre-processing

The *Escherischia coli* genome has been here analyzed to monitor the *SOS* pathway response to DNA damages induced by perturbation experiments. These are carried out by calibrating drug (Norfloxacin, $10\mu g/ml$) injection in the cells at a certain initial time, and taking records of the gene expression changes at regular (twelve minutes) time intervals, until the steady state is reached. Thus, during one hour there are six measurements that have been considered, while two extra conditions refer to no-drug states.

The microarray data have been treated by the *Affy MAS 5.0* normalization algorithm. Then, after running the usual *t* and *rank* tests to eliminate the genes showing random or no variation at all, a reduced genome of 2330 genes (approximately half of the original size) remains for further investigation aimed to find significant gene relationships.

Next, pre-processing has involved log-transformation of the gene measurements so as to smooth out wild fluctuations among gene expression values, followed by normalization to the initial time so as to obtain comparable orders of magnitude for the gene expression changes across time points, relatively to the effects of the perturbation experiments.

3.2 Fuzzy Rules

The application of the ICA algorithm (JadeR) to the data yields four independent components, based on the spectrum of eigenvalues given by a PCA pre-processing step. The four components represent a near-minimal (up to inherent noise) system able to capture the gene network dynamics along the main paths of variation. The inclusion of an extra eigenvalue would allow for five components to be considered, thus testing the system's robustness to noise. We have checked these sensitivity aspects too.

Another useful instrument here employed is the *z-score* transform, usually aimed to standardize the data. In our study, we adapt it to perform a change of coordinates that is useful for controlling how different variability laws may affect the performance of the thresholding computed over the components, and consequently the gene group selection power.

The z-scores are computed for each gene in the $k-$dimensional genomic profile of each of the $j = 1, \ldots, m$ sources. From a profile indicated by s_{j_k}, its transformed value is obtained as follows:

$$\tilde{s}_{j_k} = \frac{(s_{j_k} - mean[s_{j_k}])}{std[s_{j_k}]} \tag{4}$$

As these calculations assign a membership probability for each gene with respect to every estimated component, they basically apply a fuzzy rule. In other words, the gene expression values are now assigned to the estimated components simultaneously and according to certain normalized weights, due to the z-scores.

The adapted z-scores measure the log-probability that a gene expression value computed by each component would be obtained by chance, rather than according to the outlying information content assigned originally to that gene by the components.

3.3 Mixing Parameter Estimation

The estimated mixing matrix values computed at each time point for each component is reported in Table 2[1]. By taking a closer look at A, the third and fourth columns (associated with a four-component system) are quite sparse, which indicates a minor contribution of these sources in explaining the dynamics of the mixture profile observed at each time. Thus, it is likely that incremental redundancy would be found by adding extra components; Table 3 reports the estimated matrix with one additional component identified. This extension, as we might see below, can have consequences.

Estimates	IC-1	IC-2	IC-3	IC-4
t=1	0	0	0	0
t=2	0.8914	0.065	0.4032	0.0437
t=3	1.0206	0.1678	−0.0577	0.0006
t=4	1.0554	0.4532	−0.0834	0.0882
t=5	1.006	0.6381	0.0029	−0.0654
t=6	0.908	0.6125	0.0171	−0.1867

Table 2: Estimated Mixing Matrix with 4 IC.

We believe that some level of control of the redundancy inserted through the extra component would be useful for understanding the system's dynamics and smoothing out the noise effects. Therefore, numerical experiments have been conducted to test what happens when four or five components are considered, in the attempt to discriminate between their structural versus noisy content.

[1]The first row of the estimated mixing matrix lists zeros due to data normalization and log-transformation.

Estimates	IC-1	IC-2	IC-3	IC-4	IC-5
t=1	0	0	0	0	0
t=2	0.9208	−0.00603	0.3283	0.0569	−0.031
t=3	1.0208	−0.1423	0.1491	0.0509	−0.0236
t=4	1.0479	−0.4643	−0.1619	−0.0005	−0.0644
t=5	1.0105	−0.5861	−0.1132	0.2282	0.0112
t=6	0.9105	−0.5822	−0.0686	0.0817	0.2365

Table 3: Estimated Mixing Matrix with 5 IC.

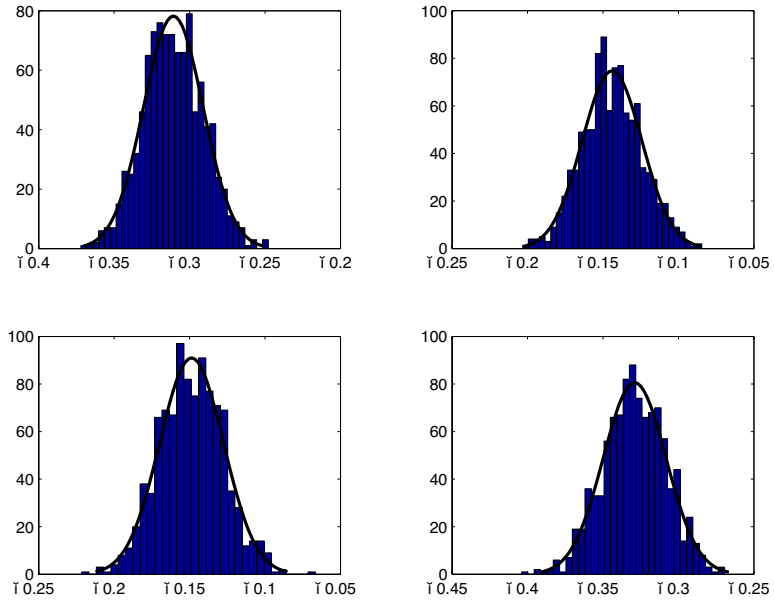


Figure 1: Bootstrap distributions of the component sample means. IC-1, top-left; IC-2 top-right; IC-3 bottom-left; IC-4 bottom-right.

3.4 Bootstrap

Bootstrap analysis has also been carried to verify that the computed statistics can calibrate the gene selection process by accounting for random variation. The gene selection is variable due to the randomness in the system, and also due to the ICA inherent ambiguities. Thus, the randomization step by bootstrap represents a way to overcome some of the computational difficulties in finding good estimators from individually estimated components.

The gene expression values have been sampled from each estimated independent component, and the statistics used for thresholding have been re-computed from each bootstrap sample. The examination of the bootstrap means for the estimated components shows that their distributions are approximately Gaussian when 1000 samples are taken, and only a small bias is present in our outcomes, due to the asymptotic and alignment properties of the bootstrap distributions (Figure 1). When comparing the bootstrap-derived gene groups with those formed by thresholding from the estimated components, the differences are almost negligible (see Table 4).

3.5 Gene Feature Selection

The following tables and figures report descriptive and diagnostic evidence from the estimated components. Table 4 and 5 present gene groups[2], as selected by thresholding; these values are computed from the estimated independent components, and appear together with those computed from the z-scores.

Thresh	m2	m3	boot-m2	boot-m3
IC-1	106	16	113	16
IC-2	79	26	81	26
IC-3	124	40	126	42
IC-4	104	34	104	34
ICz-1	−−	−−		
ICz-2	−−	−−		
ICz-3	−−	−−		
ICz-4	−−	−−		

Table 4: Outliers from the estimated components (m2 and m3), and from bootstrapped components (boot-m2 and boot-m3). Bootstrap values are reported.

Table 5 reports the results obtained under the hypothesis of redundancy in the system (only one extra component is here allowed, and no bootstrap is computed). Interestingly, from the z-scores there seems to be evidence of grouping among genes, as a result of the induced standardization, but unlike what seen with less redundancy (four components).

Thresh	m2	m3
IC-1	112	17
IC-2	79	24
IC-3	120	41
IC-4	119	49
IC-5	100	35
ICz-1	46	−−
ICz-2	66	−−
ICz-3	19	−−
ICz-4	29	−−
ICz-5	55	−−

Table 5: Outliers from m2 and m3. No bootstrap here performed. The z-scores are reported too.

By looking at the groups selected under the two cases, with four or five components, we want to account for the differentially expressed genes which overlap the groups and are thus selected by different estimated components. We believe that these genes have more chances to be possibly co-regulated by different biological factors. The lists of these genes are reported in Table 6, where the numbers identify the frequency with which genes overlap between 4, 3, or 2 groups.

With regard to the qualitative information content (i.e. biological validation) obtained from our gene grouping, note that in the first component of Table 4, and for *m2* intervals, quite a good

[2]The numbers refer to the size of the gene groups selected from each component (estimated and/or bootstrapped).

Score by frequency	4	3	2
4 IC (m2)	1	13	85
4 IC (m3)	−−	1	9
5 IC (m2)	3	20	142
5 IC (m3)	−−	3	21

Table 6: Score by frequency: differentially expressed genes in 4,3,2 components, from m2 and m3 intervals. The numbers indicate the size of the gene groups.

number of genes belonging to the SOS response system is detected, such as *dinA*, *ybfE*, *uvrB*, *sulA*, *dinI*, *yebG*, *recN*, *recX*, *recA*, *dinD*, and *uvrA*. At the *m3* intervals five of these genes are again detected, as they are highly expressed (i.e. *ybfE, sulA, recN, recX, recA*). Concerning the overlapping genes, there is no presence of SOS-related genes, but instead of a few other strong regulators. The full annotation can be found in [4].

3.6 How many good sources?

A good practice in many applications with complex data structures is turning the data to an uncorrelated form, via whitening or sphering transforms. A classical decorrelation method is based on sorting the eigenvalues from a Singular Value Decomposition (SVD) ([1, 13]) of the data, or equivalently after an application of PCA to their covariance matrix.

Such transforms are popular in statistics and signal processing, where often the goal is to obtain the smallest possible signal subspace (S) from the originally noisy space (Y) where the data lie, i.e., $Y = S + \epsilon$. The whole space rank, say n, is thus decomposed by the m-component for the signal part, and the $n - m$-component for the noise part, depending on the relative magnitude of the singular values which are identified.

In order to exploit this subspace split, ICA can be implemented as a two-step algorithm. A whitening (projection) step is followed by a rotation step directing the final step, i.e. optimization (based on diagonalization of empirical statistical functionals, mutual information minimization, etc.). As a result, from orthogonal mixtures of sources one finds components that are statistically independent (thus, beyond the simple decorrelation achieved under linear dependence).

In particular, projecting the data onto the space generated by the principal components enables the detection of those m strong sources able to capture the most relevant information in the data. It thus follows the possible elimination of the $n - m$ weaker sources related to the noise terms (see [22] for a detailed analysis).

Furthermore, numerical problems can be avoided, or at least minimized, when the final optimization is performed on a clean signal space; conversely, the presence of weak sources and the corresponding inclusion of too small eigenvalues lead to a singular mixing matrix.

However, due to the limited information from the samples, one may be able to achieve only a sub-optimal signal subspace from the most informative sources, with a dimension smaller than the optimal one (delivered by the true sources characterizing the system dynamics). In order to explore this possibility, we found useful to explore further the redundancy previously introduced.

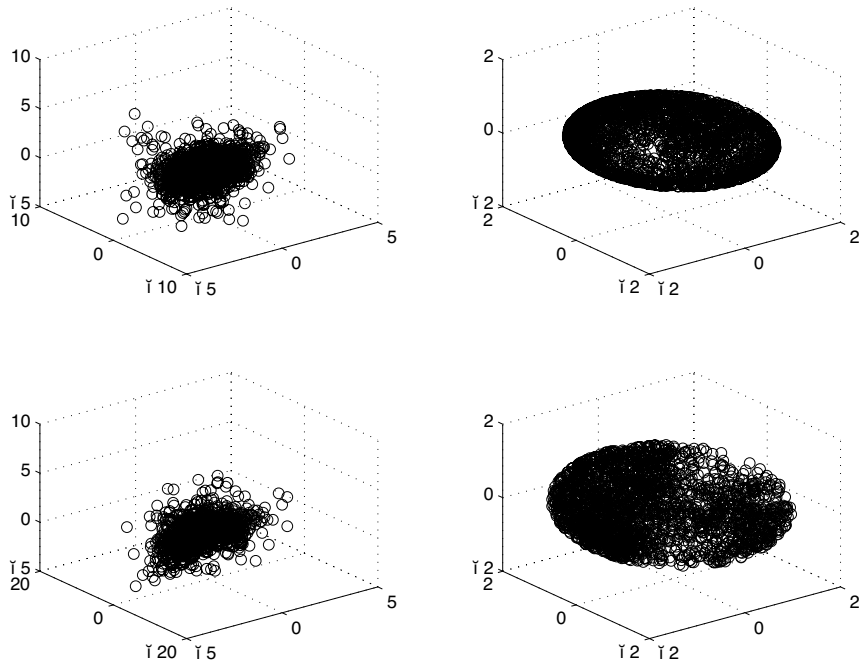


Figure 2: Scatter plot of the three first components. IC_4 **(top-left) and** ICz_4 **(top-right) vs corresponding plots of** IC_5 **and** ICz_5 **(bottom).**

Figure 2 illustrates the difference between a system with four (top) and five components, for both the estimated sources (left) and the z-scores. Through this visualization we investigate ways to discriminate information from noise in the system, or equivalently explore whether the projected space is endowed with more or less instability.

A well-known fact from signal processing is that the presence of redundancy in a system may increase the robustness to noise level, as often shown in several applications of regularization problems. Noise reduction and better component separation power may be obtained by inserting variables independent from those already included in the system; otherwise, when dependent components are observed, redundancy becomes only detrimental to the system, and no additional information is gained.

3.7 Numerical Perturbation Analysis

As a consequence of having inserted an extra component, we would also like to know whether the increased complexity of the system yields a better group separation power by ICA, or just more instability.

The exploratory work has been integrated by numerical perturbation analysis, and is reported with Figures 3 and 4. The method requires injection of zero mean Gaussian noise with variable standard deviation (see Figure 3). The noise has thus a different magnitude (std 1, 0.1 and 0.01) and is added directly to the estimated components before taking the z-scores.

When computed at each target gene, the adapted z-scores represent new contributions of the components to the gene expression values. To visualize this aspect, we consider projections onto a sphere that make the values more compact in terms of variability range (or energy dis-

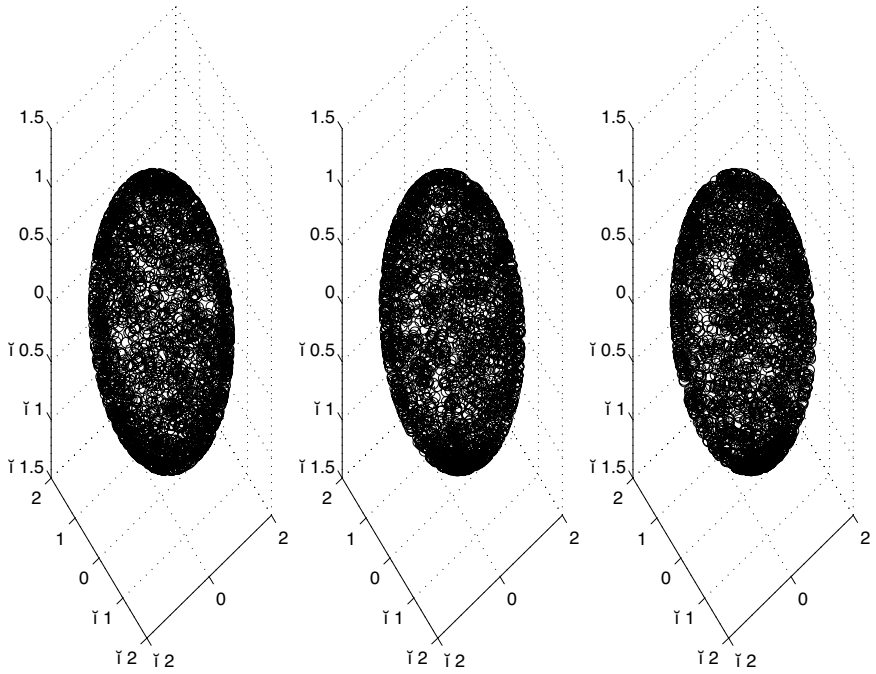


Figure 3: Scatter of perturbed IC_4 with zero mean Gaussian with std 0.01 (left), 0.1 and 1 (right)

persion). Note that four more than five components appear to yield z-scores more concentrated around the border, and thus define a smoother sphere's surface (see Figure 4).

In our experiments, we also applied the perturbations directly to the gene data; then, the sources are estimated and the z-scores computed. The difference in the energy dispersion of the values projected onto the sphere indicates that a different degree of inherent randomness is characterizing the two systems. In particular, less noise seems to characterize the system with four components.

Overall, systems with more than a minimal number of components indicate the presence of undesirable redundancy due to a noise cover masking the informative signal.

4 Entropy Fluctuation Analysis

In order to avoid the limitation of an infinite number of linear transforms able to yield decorrelation, ICA can leverage on dependence measure such as mutual information (MI), as explained in this section, but other contrast functions based on high-order distributional moments are also capable of capturing extra dependence in the data.

It is straightforward to show that the independence between the recovered sources can be measured by their MI, which is a generalized correlation compared to the classical Pearson linear correlation coefficient. The MI is defined in terms of entropies measuring the average amount of information that the estimated S yield. Given the differential entropy of an m-dimensional random variable S with *pdf* $P(S)$,we have:

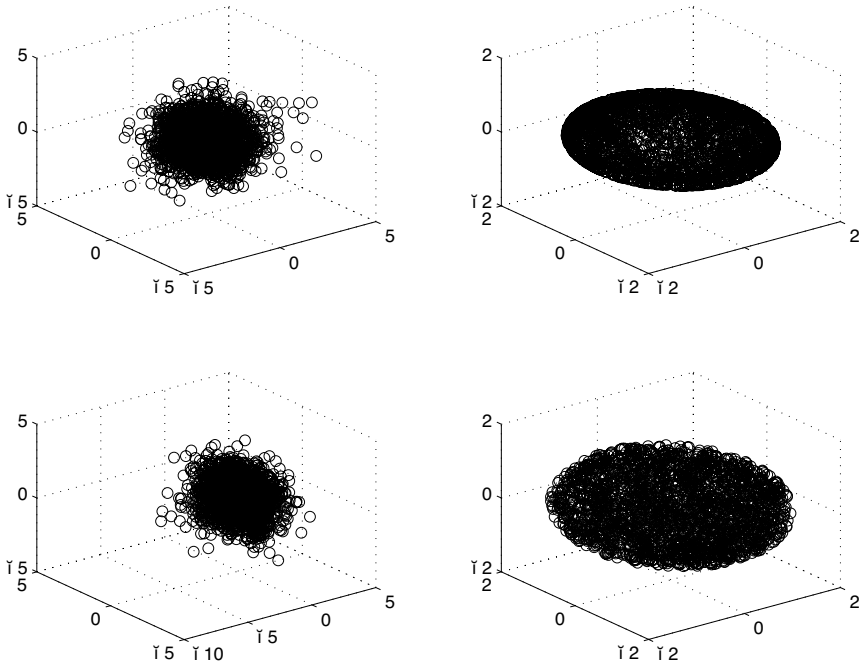


Figure 4: Scatter of three IC under system perturbation: IC_4 **(top-left),** ICz_4 **(top-right), and corresponding plots for** IC_5 **and** ICz_5 **(bottom)**

$$H[S] = -\int P(S) log P(S) dS \tag{5}$$

In order to deal with our application, consider a partition of the sources, $S = [\bar{S}, \tilde{S}]$, where the split into only two subsets is used to simplify the representation. The conditional entropy of $\bar{S}$ given $\tilde{S}$ is then defined as:

$$H[\bar{S} \mid \tilde{S}] = -\int P(\bar{S}, \tilde{S}) log P(\bar{S} \mid \tilde{S}) d\bar{S} d\tilde{S} \tag{6}$$

and the MI can now be obtained as:

$$I[\bar{S}, \tilde{S}] = H[\bar{S}] + H[\tilde{S}] - H[\bar{S}, \tilde{S}] = H[\bar{S}] - H[\bar{S} \mid \tilde{S}] \tag{7}$$

In general, the relationships with entropy can be given; for a random vector Z it works as follows:

$$I(Z) = \sum_j H(z_j) - H(Z) \tag{8}$$

Since the MI represents the difference in information obtained by observing separately (component-wise) or jointly the random vector, it is equal to zero if and only if $\bar{S}$ and $\tilde{S}$ are independent, i.e., $P(\bar{S}, \tilde{S}) = P(\bar{S})P(\tilde{S})$. Moreover, this quantity is non-negative, as more information comes from separated rather than jointly observed variables.

The same principles can be applied to the estimated components, and the MI can be associated to redundancy. When the recovered sources are independent, this quantity is zero and the source separation results successful. ICA thus finds a separating transform, or matrix under linear mixing, that minimizes the MI between the estimated sources.

Since the entropy $H(Z)$ remains constant under a transformation given by the product of whitening and orthogonal matrices, $I(Z)$ varies with the remaining quantity in Eqn. (8), i.e., the sum of the marginal entropies. We show with a simple example how this aspect helps in discriminating between redundant and non-redundant systems.

There is an interesting connection between ICA and the entropy reduction method of [23]. Given that non-linearities in the data are better described when higher moments of the involved distributions are accounted for, the method of these authors suggests to concentrate the non-linear variation in a certain random variable over a restricted subset of the other variables in the systems. While this is basically the same principle behind ICA, it works by minimizing the differential entropy of the original variable over the concentrated ones.

The Sample Approximate Entropy (*SApEn*), from [19], is an *ad-hoc* statistic quantifying the degree of unpredictability or randomness in a sequence of observations, and can be used as a plug-in empirical estimate for the entropy functional.

A sequence of values becomes more predictable if patterns of fluctuations are repeated; thus, the *SApEn* values reflect the chance that similar patterns are observed or not, and when many repeated patterns are present, the *SApEn* estimate is relatively small, which makes more predictable the sequence. Otherwise a more complex and random sequence delivers a bigger *SApEn* value. The parameter "m" indicates epochs or sub-sequences which are used to test the similarities of patterns, and more specifically it addresses their length.

This procedure mimics a time delay embedding technique; through a time lag τ a sampling rate is determined, while the embedding dimension m governs the length of the filter. Consequently, a sequence of data can be represented in phase-space by a set of delayed vectors $x_{\tau,m}(k) = [x_{k-\tau}, \ldots, x_{k-m\tau}]$.

If the term $m\tau$ is too small, then the signal variation within the vector is dominated by noise; thus, an adjustment is required. For instance, $\tau = 1$ and m is set to some value, so as to find a good compromise between the scale of the hidden dynamics and the inherent complexity of the sequences of interest, i.e. the IC-based gene profiles.

The parameter values that have been selected, $\tau = 1$ combined with $m = 1$ up to $m = 5$, yield a phase space representation that best reflects the dynamics of the system and identifies a near optimal embedding, i.e. one endowed with minimal differential entropy (or minimal disorder). Tables 7 and 8 report these results.

As we are in a linear setting by model construction, the embedding attempts to localize the linear structure and let the changes occur smoothly over the phase space. When switching between the two systems, we might expect that the modes or eigen-vectors vary not so smoothly across the points, particularly because of the degeneracy of some eigen-values and their noise sensitivity.

One may hope that the extra structures inserted in the approximating model brings in also a correspondingly higher degree of smoothness sufficient to balance some potential instability. When we measure the information with a redundant system, we might lose stability as we

SApEn at *m*=	1	2	3	4	5
$IC-1$	2.09	1.82	1.77	1.72	1.6
$IC-2$	1.91	1.73	1.7	1.67	1.63
$IC-3$	2.03	1.81	1.77	1.69	1.59
$IC-4$	1.93	1.82	1.77	1.7	1.7
$IC-1$	2.09	1.82	1.77	1.72	1.62
$IC-2$	1.89	1.73	1.7	1.69	1.7
$IC-3$	2.02	1.78	1.75	1.7	1.63
$IC-4$	1.88	1.80	1.75	1.74	1.71
$IC-5$	1.96	1.84	1.79	1.74	1.79
$ICz-1$	2.01	1.79	1.68	1.68	1.5
$ICz-2$	2.0	1.74	1.7	1.66	1.61
$ICz-3$	2.01	1.77	1.7	1.59	1.49
$ICz-4$	2.04	1.92	1.88	1.9	2.02
$ICz-1$	2.07	1.83	1.75	1.7	1.71
$ICz-2$	2.05	1.88	1.85	1.89	1.86
$ICz-3$	2.08	1.93	1.9	1.87	1.87
$ICz-4$	2.02	1.96	1.95	1.94	2.03
$ICz-5$	2.0	1.86	1.85	1.87	1.91

Table 7: *Sample Approximate Entropy* for each IC and ICz at epoch $m = 1, \ldots, 5$.

move away from the desired minimum.

It was mentioned before that the z-scores for the system with four or five components address a different degree of stability; the rougher shape of the sphere's surface for the case of five components reflects the presence of instability due to noise. We validate that graphical evidence with numerical computation of dispersion and empirical entropy measures.

In Table 7, by looking at the estimated values, the five-component system appears always more random in the representation obtained by the z-scores compared to the four-component system. We note a peak of magnitude at the fourth component when z-scores are considered, followed by a decrease at the fifth component, probably due to a loss of stability.

Furthermore, the z-scores with four or five components show a decreasing pattern along *m* for every component, with a stable pattern (first decreasing first and then increasing) at each value, except for ICz4 where a flat range of values appears across m. Once more, the fourth component is the crucial one for stability.

In Table 8, where the estimated entropy refers to perturbed sources, we underline two aspects: A) in the system with four components there is a jump of the entropy for components other than IC1 at every *m*; B) the error propagation in general masks the previously observed decreasing patterns across *m*, even if with z-scores this feature appears again, except for ICz4. The same can be observed for the system with five components, where the error propagation effects are a consequence of the noise injection that makes the pattern across m even more random than before.

Table 9 shows instead the results referred to both the sample variance computed for each estimated component and the *SApEn* variance estimates. The results are reported only for $m = 1$. While the estimated components result always with unit variance, by construction, their z-

SApEn at *m*=	1	2	3	4	5
$ICp-1$	2.12	2.06	2.04	2.05	2.09
$ICp-2$	2.17	2.17	2.16	2.2	2.24
$ICp-3$	2.17	2.17	2.18	2.18	2.2
$ICp-4$	2.17	2.16	2.18	2.18	2.14
$ICp-1$	2.12	2.07	2.05	2.07	2.05
$ICp-2$	2.18	2.18	2.17	2.16	2.23
$ICp-3$	2.17	2.17	2.16	2.16	2.27
$ICp-4$	2.17	2.17	2.19	2.17	2.36
$ICp-5$	2.18	2.19	2.19	2.12	2.26
$ICzp-1$	2.02	2.0	1.98	1.93	1.88
$ICzp-2$	2.03	2.02	2.02	2.03	2.01
$ICzp-3$	1.98	1.94	1.93	1.88	1.85
$ICzp-4$	2.04	2.04	2.03	2.02	2.13
$ICzp-1$	2.08	2.06	2.05	2.05	2.07
$ICzp-2$	2.1	2.1	2.12	2.19	2.21
$ICzp-3$	2.08	2.06	2.07	2.05	2.04
$ICzp-4$	2.1	2.11	2.13	2.17	2.24
$ICzp-5$	2.11	2.1	2.11	2.07	2.07

Table 8: ***Sample Approximate Entropy*** **for each IC and ICz under perturbation and with epoch** $m = 1, \ldots, 5$**. The symbol** p **indicates a perturbed system.**

scored versions differ in way that shows how the dispersion increases due to the presence of an extra component. As a further form of control, we have also computed the empirical entropy variance [24] as an estimate for the quantity $var(H) = var[-log(p(x))] = E[log(p(x))^2]$.

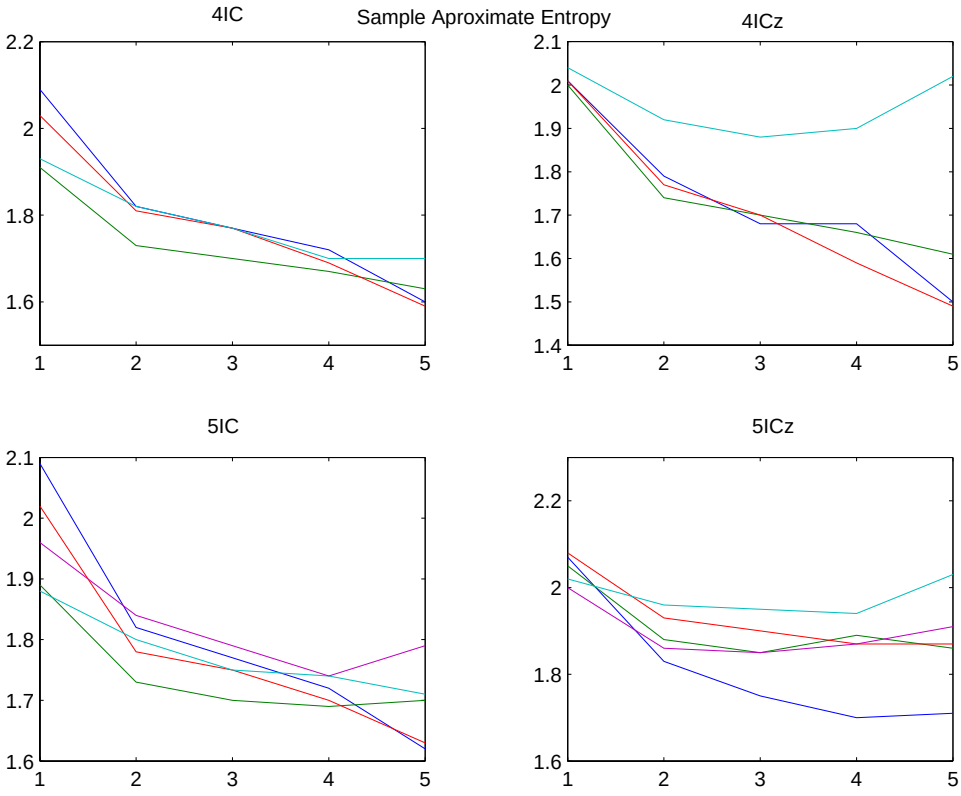
Figure 5 shows the entropy fluctuations of Tables 7 and 8. It seems even harder to detect a system switch from the sample variances of Table 9. We may note jumps in IC4 with the *SApEn* variance estimates, in both the four- and five-component system, with and without z-scores. Then, it appears an increasing trend in the system with five components due to error propagation effects combined with the inherent component variability.

5 Conclusions

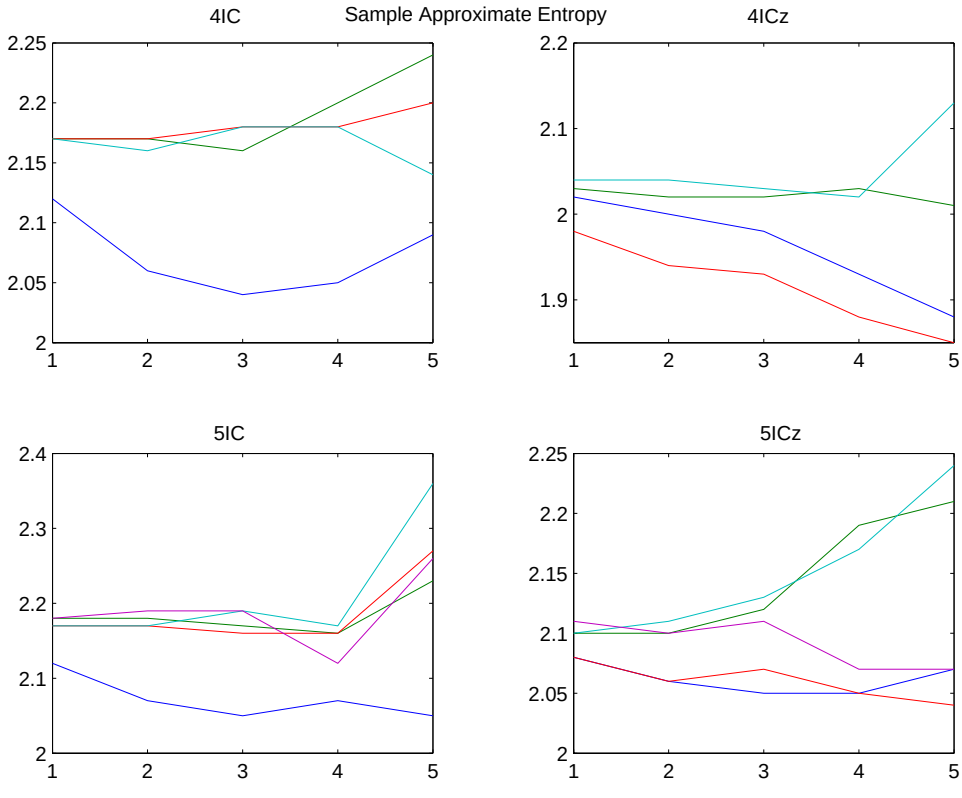
Overall, ICA achieves substantial dimensionality reduction and high-quality performance in gene selection via thresholding computed over the gene profile of each estimated components (or their bootstrapped versions), and through significance tests built around simple IC-based statistics so as to isolate outlying genes.

Even in its simplest linear form, ICA appears endowed with a great power of approximating the complex gene network dynamics, due to the very limited number of parameters which are estimated in the mixing matrix, or correspondingly to the capacity of decomposing the high-dimensional, and likely non-linear, system in just a few modes influencing each other almost independently.

It has been shown how the number of components can be estimated with quite good reliability, and tools like numerical perturbation and sample entropy measures have been computed to



A.



.B

Figure 5: Table 7 (A) and Table 8 (B) representations of entropy fluctuations and embedding. The system in B is under perturbation.

Sample Variance (at each IC)	IC1	IC2	IC3	IC4	IC5
IC_4	1	1	1	1	
ICz_4	0.83	0.69	0.68	0.68	
$ICzp_4$	0.72	0.72	0.72	0.74	
IC_5	1	1	1	1	1
ICz_5	0.92	0.73	0.75	0.65	0.68
$ICzp_5$	0.76	0.79	0.76	0.79	0.79
Sample *SApEn* Variance (at each IC)	IC1	IC2	IC3	IC4	IC5
IC_4	3.48	4.38	4.27	5.15	
ICz_4	3.78	3.61	3.76	4.21	
IC_5	3.34	4.40	4.42	4.99	4.86
ICz_5	3.39	4.26	4.21	4.38	3.76

Table 9: Sample Variances (top) and Entropy Variances (bottom) for each IC. The subscript $(.)_4$ or $(.)_5$ indicates the number of components in the system. The symbol *zp* stands for the perturbed z-scores. Only the case $m = 1$ is considered.

support these findings.

Redundancy in the system can in principle hold when more than a minimal number of components is present, but this has effects on the quality of the gene selection and on the system robustness to noise. The insertion of extra components may just bring into the system weak sources that deteriorate the discriminatory power of thresholding when forming gene groups.

The z-scores play an interesting role when they are computed for each gene across the estimated components in a fuzzy fashion; they have revealed that a misleading discriminatory power can appear in a system embedded with noise, which is then responsible for boosting the expression values in the significance region.

References

[1] O. Alter, P. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, *em PNAS*, 97(18), 10101-10106, 2000.

[2] S. Amari, J. Cardoso, Blind Source Separation - Semiparametric Statistical Approach, *IEEE Trans. Sign. Proc.*, 45, 2692-2700, 1997.

[3] J.A. Berger, S. Hautaniemi, H. Edgren, O. Monni, S.K. Mitra, O. Yli-Harja, J. Astola, Identifying underlying factors in breast cancer using independent component analysis, *Proc. IEEE Neur. Net. Sign. Proc.*, 81-90, 2003.

[4] E. Capobianco, Mining Time-Dependent Gene Features, *J. Bioinf. Computat. Biol.*, 3(5), 1191-1205, 2005.

[5] J. Cardoso, Source separation using higher order moments, *Proc. ASSP.* 2109-2112, 1989.

[6] J. Cardoso, Dependence, Correlation and Gaussianity in Independent Component Analysis, *J. Mach. Learn. Res.* 4, 1177-1203, 2003.

[7] J. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, *IEE Proc. F.*, 140(6), 771-774, 1993.

[8] P. Chiappetta, M.C. Roubaud, S. Torresani, Blind Source Separation and the Analysis of Microarray Data, *J. Computat. Biol.*, 11(6), 1090-1109, 2004.

[9] P. Comon, Independent Component Analysis - a new concept? *Sign. Proc.*, 36(3), 287-314, 1994.

[10] P. Diaconis, J.H. Friedman, Asymptotics of Graphical Projection Pursuit, *Ann. Statist.*, 12(3), 793-815, 1984.

[11] J.H. Friedman, Exploratory Projection Pursuit, *J. Amer. Statist. Assoc.*, 82(397), 249-266, 1987.

[12] J.H. Friedman, J.W. Tukey, A projection pursuit algorithm for exploratory data analysis, *IEEE Trans. Comp.*, C23(9), 881-890, 1974.

[13] N. Holter, A. Maritan, M. Cieplak, N. Fedoroff, J. Banavar, Dynamic modeling of gene expression data, *PNAS*, 98 (4), 1693-1698, 2001.

[14] G. Hori, M. Inoue, S. Nishimura, H. Nakahara, Blind gene classification. An application of a signal separation method., *Gen. Inform.*, 12, 255-256, 2001.

[15] P.J. Huber, Projection Pursuit (with discussion), *Ann. Statist.*, 13, 435-525, 1985.

[16] A. Hyvarinen, E. Oja, A fast fixed-point algorithm for Independent Component Analysis, *Neur. Computat.*, 9(7), 1483-1492, 1997.

[17] A. Hyvarinen, Fast and robust fixed-point algorithms for Independent Component Analysis, *IEEE Trans. Neur. Net.*, 10(3), 626-634, 1999.

[18] M.C. Jones, R. Sibson, What is projection pursuit? (with discussion), *J. R. Statist. Soc. A*, 150, 1-36, 1987.

[19] D.E. Lake, J.S. Richman, M.P. Griffin, J.R. Moorman, Sample entropy analysis of neonatal heart rate variability, *Amer. J. Physiol.*, 278(6), H2039-2049, 2002.

[20] S. Lee, S. Batzoglou, Application of independent component analysis to microarrays, *Gen. Biol.*, 4:R76, 2003.

[21] W. Liebermeister, W., Linear modes of gene expression determined by independent component analysis, *Bioinform.*, 18, 51-60, 2002.

[22] J.P. Nadal, E. Korutcheva, F. Aires, Blind source separation in the presence of weak sources, *Neur. Net.*, 13, 589-596, 2000.

[23] M. Samoilov, A. Arkin, J. Ross, On the deduction of chemical reaction pathways from measurements of time series of concentrations, *Chaos*, 1(11), 108-114, 2001.

[24] A.J. Wyner, D. Foster, On the lower limits of entropy estimation, Tech. Rep. Dept. Statistics, Wharton School, Un. Pennsylvania, 2003.