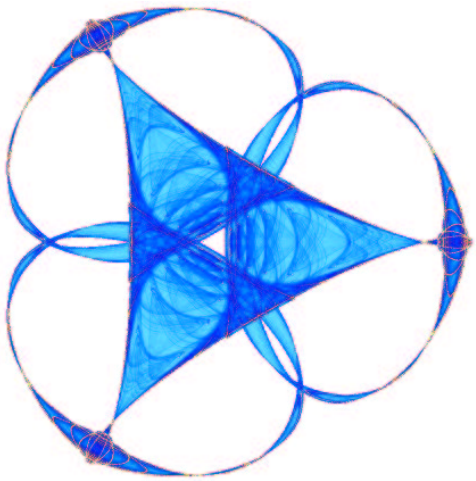
EXPLORATION AND REDUCTION OF HIGH DIMENSIONAL SPACES WITH INDEPENDENT COMPONENT ANALYSIS

By

Enrico Capobianco

IMA Preprint Series # 2013

(December 2004)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA
514 Vincent Hall
206 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612/624-6066 Fax: 612/626-7370
URL: http://www.ima.umn.edu

Exploration and Reduction of High Dimensional Spaces with Independent Component Analysis

Enrico Capobianco
Biomedical Engineering Department,
Boston University, 44 Cummington St., Boston, MA 02215 USA
ecapob@bu.edu

December 17, 2004

Abstract

The application of Independent Component Analysis (ICA) to genomic data is here considered. In recent years, microarrays have delivered to researchers huge series of measurements of gene expression levels under different experimental conditions. This work, for instance, emphasizes exploratory data analysis following experimental work on the popular *Escherichia coli*; the context is a typical one in which changes in gene expression values are observed after perturbing genes at an initial time and measuring the responses at regular time intervals until the steady state is achieved. The gene temporal patterns are as usual very short, and it is no exception for the application here described, as only six time points are available. This aspect combines with a very large feature space (i.e., the gene dimensionality). Thus, several kinds of fluctuations have to be monitored, and many are discarded because not significantly different from noise. ICA represents a very flexible signal processing tool which attempts to deal with noise as well, although the expected impact involves its most inherent property of delivering a decomposition of the gene profiles according to statistically independent sources of information. The latter are most likely linked to underlying biological processes regulating the genes, but it is not a goal of this paper to characterize these biological aspects, since this is a subject of ongoing research which requires further experimental investigation and tests before drawing any important conclusion. However, the initial computational results which have been obtained are very encouraging, and thus they are presented here together with some interesting problem-specific aspects.

i **Keywords**: *Independent Component Analysis; Microarrays; Dimensionality Reduction; Gene Selection; Redundancy; Entropy.*

1 Introduction

1.1 ICA and Genomics

Genomics has become a challenging research field for many computational scientists; during the recent years a vast variety of statistical techniques and machine learning algorithms have been inspiring cross-disciplinary joint work with computational and systems biologists.

From classical matrix analysis solutions like singular value decomposition [1] to clustering [12], gene selection and classification techniques [16, 4, 11], from dynamic modeling [18] to regularization approaches [17], these represent only a few examples of quantitative techniques and paradigms which have been extensively investigated in genomic applications by several authors in the recent years.

The rationale for applying ICA [6, 10] in genomics has been emphasized in some published work on microarray data [27, 24, 9, 26], gene feature detection [2] and classification [19, 20]. It is worth to mention that while for genetic regulatory networks it appears basically unknown the role that ICA can play, a recently published work [25] on network component analysis actually discusses the ICA impact compared to network topology-preserving techniques.

Among its possible definitions, ICA represents an approximate method for inducing variable decomposition and dimensionality reduction in latent variable models where the observed data are generated by a linear (or non-linear) mixing of unknown sources. Specifically for the purposes of the genomic application in this work, the decomposition of the gene profiles observed at each sample point is allowed to depend on two unknown variables: the set of independent components (or sources), and the matrix which is mixing them.

While both these quantities need to be estimated in order to yield an approximation of the gene expression dynamics through just a few dominant modes, gene selection via thresholding is also a key issue. In particular, through ICA one may find a viable route for effectively reducing the dimensionality of the original feature space by using the estimated independent components and their attached gene profiles before turning to standard significance tests for their corresponding expression levels.

1.2 Key Factors: Dimensionality Reduction and Sparsity

Usually, in genomic applications, the researcher deals with noisy highly dimensional feature spaces. Combined with lots of genes whose expression levels are experimentally measured, there are often only a few samples where these measurements occur, being these related to quite expensive experimental conditions and limited time points.

This unbalanced combination suggests that it might be hard for standard statistical inference techniques to come up with good general solutions, likewise for machine learning algorithms to avoid heavy computational loads. Thus, one naturally turns to two major aspects of the problem: sparsity and intrinsic dimensionality.

Many natural signals can be sparsely represented in a proper signal dictionary, where only a small number of decomposition coefficients remain significantly different from zero. Real world signals may inherit the sparsity property in time or frequency domain, but also time-scale decompositions have been shown to be very useful. ICA may potentially work well (see, among others, [31]) under any of these conditions, and this represents a great

advantage.

The truly intrinsic dimensionality of many kinds of input spaces (see for an excellent review [14]) is very often reducible to a small number; thus, together with a good approximation to the core dimension of the problem, one may also gain knowledge of what input variable are influencing the observed responses or targets, and what are instead highly collinear with them.

The simplest methodological assumption in ICA is that the observed mixture signal is nothing but a linear combination of latent informative sources whose dynamics are combined through the mixing matrix. When some of the sources result sparse, the simple structure of the model allows for clusters to be derived along the directions indicated by the columns of the mixing matrix; correspondingly, a consistent degree of collinearity is to be expected between the mixture signals and the directions related to the sources which are not sparse.

In other words, just a few modes which are regulated by the mixing matrix can be very helpful in approximating and revealing the complex dynamics behind complex biological network structures, in approaching its intrinsic dimensionality, or equivalently reducing the noisy high-dimensional input space to a low-dimensional feature embedding space. One might thus hope to identify through ICA just a few independent components able to detect gene groups connected at some biological level.

Exploiting the inherent level of sparsity is a key issue in genetic regulatory networks, where the connectivity matrix needs to account for the real links among genes and discard many redundancies. Most experimental evidence suggests that real gene-gene connections represent indeed a subset of what is usually mapped onto either a huge gene vector or a typically dense and highly structured network.

Inferring gene network connectivity from the expression levels represents a challenging inverse problem that is at present stimulating key research in biomedical engineering and system biology. The complication of having experimental measurements for the gene expression changes in only a limited number of time points poses important methodological questions, as a sample size much smaller than the gene dimensionality makes the inference problem under-determined.

However, several attempts have been made to describe gene networks with only limited interactions, thus exploiting the inherent sparsity of these systems [15]. This in turn suggests that a certain redundancy of links in gene networks, or equivalently the inherent sparsity structure of these systems, might let the essential connections be identified and the inverse problem be given both satisfactory definition and computationally efficient tractability.

1.3 Gene Selection

By using ICA one can simplify the problem of making inference about gene networks. This goal is obtained by approximating its intrinsic dimensionality with a small set of estimated components as much statistically independent as possible, and by then identifying homogeneous gene groups by the means of the information embedded in the same components combined with some form of thresholding able to discriminate information from noise suitably for gene selection.

A straightforward criterion is the following. Sets of genes endowed with outlying informative content (i.e., differentially expressed more or less than a certain descriptive statistic) are

measured with respect to the gene profile values attached to each component, and should validate the hypothesis that the former genes are subject to the influence of different biological processes underlying the same components.

The components represent approximately and independently these biological influences. It is by the same nature (i.e., constructive statistical properties) of the estimated components that the sets of genes result maximally independent from each other. Moreover, due to the fact that genes sharing the same biological pathway are expected to show relatively strong connectivity, it is also likely for these groups to be quite homogeneous, depending on the kind of influence that a biological process may exert through the corresponding component.

It is also possible, as it is evident in our computational experiments, that some genes belong to different groups, a fact that reflects the simultaneous presence of different underlying biological influences leading to co-regulated activity visible through the gene expression values.

2 Methodogy

ICA is a very popular procedure used in signal processing and machine learning applications. Among many possible model variations, we adopt the quite simple "non-parametric" view, which in statistical terms means that no assumptions are retained with regard to the distributional aspects of the gene data.

An ICA model is here summed up very concisely, as the subject would deserve coverage and treatment far beyond the scopes of the present work. Simply stated, consider at time t the system $X = AS$ with the signals X represented by $x_i, i = 1, \ldots, n$, and the independent and non-Gaussian sources S represented by $s_i, i = 1, \ldots, m$. The latter are mixed up linearly through the mixing matrix A; this system may be thought to approximate, under suitable conditions, a non-linear system such as $Y = f(X)$, with unknown f. The dimensions of the given model are thus $S \in R^m$, $X \in R^n$, and $A \in R^{n \times m}$.

Theoretical work suggests 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 ([6, 22]). ICA is very indicated when one deals with non-Gaussian distributions; compared to principal component analysis (PCA) that is targeted to Gaussian settings and thus only linear independence (through the second-order statistics), ICA exploits higher order statistics (from the moments of the involved distributions).

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 $X = AS$. This case requires constraining the sources or transforming the model in order to get an approximate solution.

In an attempt to estimate both the unknowns, i.e., the mixing matrix and the sources, several computational algorithms can be efficiently applied, such as the Joint Approximate Diagonalization of Eigenmatrices (for Real signals) or *JadeR* [8] and the *fastICA* [21, 22], probably the most popular ones.

These algorithms have been adopted here too; 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 via principal components, and for performing numerical optimization under different conditions. In particular, we have obtained the estimates for the separating or de-mixing matrix $B(= A^{-1})$, and thus through its pseudo-inverse $pinv(B)$ one would get back to an estimated version of the mixing matrix. About the sources, one finds $Y = BX$ that approximate S.

While under conditions of perfect separation $Y = BAS = S$ would make sense, the usual solutions hold only approximately, and in particular up to permutation P and scaling D matrices, i.e., $Y = DPS$. Despite these well-known ambiguities, the statistical independence is strong enough to make the sources identifiable without need to model their probability density functions.

Thus, de-correlation and rotation steps yield a set of approximated m independent components which often perform well in various and complicated application contexts, and not only in those ones where a physical mixing process is truly linear or a system is endowed with independent coordinates explaining the statistical properties of natural/artificial signals, and suggesting optimal information processing.

ICA has been successfully employed in many applications characterized by noise, high complexity of dynamics, and convoluted variables under non-standard statistical conditions, such as non-Gaussianity and non-stationarity.

3 Smoothness and Noise

In some cases the separation of a signal from the corrupting noise is just not possible, at least no adaptive techniques can always yield optimal separation of these two components. Regardless how well one deals with its presence, through an ad hoc filtering device or a statistical model, noise remains inevitably a strong conditioning factor in any signal processing application, including genomics.

The gene temporal patterns usually reveal various shapes and behaviors. The smoothly changing ones are of interest compared to those remaining constant over time or others showing a random signature dominant over more systematic effects. The presence of temporal dependence across time points can be more or less distinct from noise; sharp fluctuations lead to discarding the corresponding genes, and monitoring just those whose activity level changes smoothly across conditions or time.

For a gene expression matrix X where the rows, say, represent all the different genes involved in the regulatory network under exam, and the columns list the measurements at successive time points ($t_1, \ldots, t_6$ in our study), the problem here addressed is that of representing X through mixed sources of information, or factorizing the data into the unknown sources and their mixing mechanism. Thus, both A and S must be estimated.

In short, we might either consider for each available sample point (related to different conditions or observations at a given time) a certain mixture of genes, i.e., a gene profile of the gene vector dimensionality, or look at gene temporal patterns. This choice becomes one between working, say, with X^T instead of X when feeding the ICA algorithms. While in several applications the structure of data and the usually long temporal patterns simplify

this initial step, there are aspects in gene data that deserve some comments.

If one aims to produce independent components of the same dimensionality of the gene set, thus ending up with making gene selection through thresholding, one good strategy is that of looking at the differentially expressed values of the genes which are delivered by each estimated independent component. An interval of significance around the sample mean or median computed for each component can be used for testing which genes are manifesting a particularly active role.

In other terms, validating through thresholding the hypothesis that the observed expression values are or not significantly different from a test statistic implies the identification of those values located outside a mean or median interval which are significantly deviating, and thus supporting the rejection of the null hypothesis of no presence of outlying effects in the IC-based gene profile. The interesting or outlying genes are linked to specific independent components by the shared statistical properties and by the biological characterizing information conveyed by the same components.

4 Data Analysis: Transformations and Pre-Processing

The gene lists we have tested were of different size and correspondingly the genetic regulatory networks endowed with different complexity and order of magnitude in their gene-gene connections, but the number of identified and estimated independent components each time we have applied ICA has shown that scalability does not represent an issue, at least for the available data structures (n genes $\times$ 6 time points).

One might simply infer that the intrinsic compared to the initial dimensionality of the problem is small even with large available networks, which is indeed the hypothesis we are most interested in. Ideally, in order to thoroughly investigate this matter one would need more time points, a condition which is hard to satisfy. Thus, it is quite challenging to proceed under much more limited conditions, but the following are the steps we have pursued in order to optimize the performance of the employed algorithms.

The data (gene expression values) have been log-transformed. In all but one control case the data have been previously normalized to the initial time so as to allow for changes of comparable magnitude in gene expression values across time points, in relation to the perturbation experiments[1]. The data are then passed to logarithms[2] in order to account for the difference of magnitude in the expression values across genes. A certain stability is gained when implementing the ICA numerical algorithms.

The gene selection results are reported in relation to the estimated independent components, together with the corresponding z-scores. The z-scores are computed for each gene across the IC values. By implementing this column-wise standardization we expect a more balanced contribution of the component-related influences over the genes. Thus, the distributions become more uniform along the gene profiles.

The source gene profile values are now projected onto a sphere which reveals how the

[1]The cells were treated with 10μg/ml Norfloxacin at the initial time and then six measurements were taken every twelve minutes. The microarray data are subject to the Affy MAS 5.0 normalization algorithm.

[2]Natural, log_2 and log_{10} transforms have been tried, by just noting small changes in the degree of smoothing. Thus, natural logs have been taken.

mass of data tend to concentrate more or less uniformly, depending on the cases under exam. Monitoring the estimated sources with respect to the impact of relative changes in their contribution to every gene may help testing the robustness of the system against noise and validating the discriminatory power of thresholding.

The data may be used to select a suitable metric; gene temporal patterns may be divided in clusters according to some similarity measure and gene dimensionality reduction can be achieved by taking average profiles. While the method refers directly to the observed gene temporal patterns, it is weak in separating signal and noise contents in the measured gene values.

There are limitations from this approach which might as well affect ICA when used in the same way, i.e., across the genes. The main weakness of clustering, i.e. the arbitrary selection of parameters, would become an issue for ICA too, and establishing the number of independent components based on the exam of temporal patterns built over six time points would lead to a wide spectrum of solutions. Then, the problem of dealing with across-sample correlation and noise would likely lead the algorithm to a failure in distinguishing common from specific features of gene patterns.

When pre-processing the gene lists one usually perform a whitening or sphering step through PCA. This step means applying a transform matrix to the data vector so as to deliver unit variance uncorrelated variables. The method is typically delivering estimates of the eigenvalues of the data covariance matrix. Usually, consistency rules hold if the elements of an estimated matrix converge to the elements of an invertible corresponding matrix in probability[3].

We looked at the dimensions suggested by PCA while performing whitening over the small available sample vs the vast genomic gene profile. In the former case it produces minimal eigenvalues with order of magnitude equal to the maximal one obtained through the gene profile. While the ratios of the smallest and the largest eigenvalues are of comparable size (7.8/431 vs 0.02/5.6), choosing one or the other ways of performing PCA has different stability effects for the numerical performance of the ICA algorithm, with substantial deterioration in the first case where the order of magnitude of the eigenvalues is clearly not aligned with that of the data.

4.1 Signal + Noise

Singular value decomposition in linear algebra, and the related PCA and Karhunen-Loeve transform in statistics and signal processing, are popular data pre-processing techiques aimed 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, N, is thus decomposed in the M-component for the signal, and the $N - M$-component for the noise, depending on the relative magnitude of the singular values which are identified.

ICA can be implemented as a two-step algorithm, where a whitening step is followed by an optimization step. The first step is indeed carried out by principal components applied to the data covariance, or singular value decomposition applied to the data. Thus, orthogonal

[3]The ratio of the smallest to the biggest eigenvalues computed from the estimated matrix should converge to that obtained from the target matrix, which results a positive number. If this quantity is large, the *condition number* of the problem is large too, and numerical instability follows.

mixtures of sources are created.

The second step leads to the real essence of ICA, and involves a rotation matrix for which the components are statistically independent at higher order than two. It is at this level that one gains extra information from ICA compared to PCA, i.e., from exploiting the information delivered by the most non-Gaussian data directions. The so-called *Jade* algorithm is one solution among other possible choices that implement this concept.

Pre-processing is useful though because 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, and the elimination of the $N - M$ weaker sources related to the noise terms which combine with the mixture of main sources (for a detailed analysis see [28]).

Another great advantage is that numerical problems are avoided, or at least minimized, when the final optimization is performed on a clean signal space, as the presence of weak sources and the corresponding inclusion of too small eigenvalues lead to a singular mixing matrix. In this last case, the columns of the mixing matrix are linearly dependent and the source estimates are very poor. Monitoring the condition number of the mixing matrix can alert about when singularity is approached, thus helping the source selection process [13].

However, there is also another possible scenario which should be considered. Due to the limited information in the samples, one may actually not get the optimal signal space from the most informative sources selected by any pre-processing analysis, but just be given a dimension of this linear space smaller than the real number of sources characterizing the system dynamics.

In order to explore this possibility, we have embedded the system with some redundancy through the inclusion of an higher number of independent components compared to that suggested by the previously executed PCA step. Our results indicate that it is important to calibrate well the choice of the number of independent components allowed in the system.

Details are given later, but the robustness and redundancy in the system seems to depend clearly on the number of estimated independent components. If properly controlled, possible redundancy might turn out to be an advantage in terms of robustness to noise. However, we bring some evidence from the genome experiment that casts doubts and suggests the need to cautiously consider this procedure.

4.2 Thresholding and Bootstrap

The dimensionality problem is faced by ICA in two different steps. First a set of independent components is estimated. Then, correspondingly to these IC, groups of genes are identified by thresholding, which enables gene selection in a very straightforward way. It turns out that the global number of genes selected is orders of magnitude smaller than the initial sample, according to the type of intervals which are used, but anyhow substantially reducing the dimensions of the problem.

The thresholding step is at first pursued directly over the estimated components via fixed intervals, then through sequential or iterative outlier search procedures (following [27]), and average deviations from the medians. Bootstrapped statistics are also computed so as to calibrate the gene selection process. Gene selection is variable due to the randomness in the system and the ICA inherent ambiguities. The randomization of bootstrap represents

a way to overcome some of the computational difficulties in finding good estimators from individually estimated components.

Thus, we have sampled from each estimated independent component and re-computed the statistics used for thresholding. Both kinds of estimates, bootstrapped and not, are kept as one has also a price to pay for randomization, which is the loss of information. Regardless the initial sample size from which one builds new samples, the class of estimators which are computed as functions of the original sample comes out to be at least as good as the class of estimators obtained by bootstrap analysis. But of course one gets to know more about bias.

The exam of the bootstrap statistics for the estimated components shows that the bootstrap distributions are approximately Gaussian (1000 samples have been taken) and a small bias is present. In order to investigate this matter, we have computed bootstrap-based (percentile and percentile-t) confidence intervals for the sample statistics, and their tightness and mutual closeness suggests a very good asymptotic approximation to Normality and an accurate coverage (see Figure 7; supporting material is available from the author, while a short description is given in the appendix).

As a final remark, gene selection by thresholding based on bootstrap statistics is just indicative[4] but when comparing the bootstrap-derived groups with those formed by thresholding the estimated components, they are almost the same each time. This because similar cutoff limits could be retained with good approximation, due to the very good asymptotic and alignment properties of the bootstrap distributions.

5 Computational Experiments

5.1 Data Description

The gene set contains the *E.coli* genome, which appears reduced to 2330 genes from the initial 4345 ones. In each table a list of genes is reported after selection by thresholding, i.e., gene selection is computed for each independent component and based on the outlying genes. For instance, the selected genes have measured expression value exceeding a two-sided mean interval such as $\mu \pm (c \times \sigma)$, with c as confidence level cutoff limit, μ as sample mean, and σ as standard deviation of each estimated IC.

The symbol "*Int*" marks the limits of the intervals at two different levels indicated with "*m2*" or "*m3*", and corresponding to two or three times the standard deviation from the mean). The intervals involving bootstrap appear as "*boot-m2*" and "*boot-m3*", with the estimated statistics and thresholds which are based on mean and standard deviation values computed from bootstrapped components. These intervals, as said, are only aimed to calibrate our results.

It is worth stressing the fact that in the gene lists selected by the more relaxed intervals (i.e., the *m2*-based genes) the significance can be more or less borderline, thus not sharp for

[4]A preference has been given to the chance of finding as many significant values as possible, including occasional ones, instead of running into the risk of missing something. Bootstrap can be used also to determine frequencies with which each gene appears in the generated samples, so as to quantify the effects on the computed statistics and have a measure of reliability for the selected genes. It has been explored this option too, in separate work.

all genes. If one looks for more stringent intervals, the *m3* ones offer a valid alternative.

The next few tables summarize the results. Figures 1-7 are for descriptive results and diagnostic evidence from the estimated IC.

- The numbers in Table 1 refer to the size of the gene groups selected from each component, and not to the specific gene labels. The intervals for gene selection are obtained via thresholding by using the estimated components (*m2, m3*) and the bootstrapped ones (*boot-m2, boot-m3*) as before, but also a sequential or iterative outlier search which leads to the identification of both negatively (*dif-n*) and positively (*dif-p*) differentially expressed genes, and a robust method computed via the medians and average deviations from them (*adM2, adM3*);

- Table 2 deals with the results obtained under hypothesis of redundancy in the system, i.e., with more IC allowed to explain the data;

- Table 3 and Table 4 report different lists of genes, where one type indicates the size of the groups and the other the gene labels[5]. The "score by frequency" is used to to assign a relative score which counts how many genes appear in the different groups;

- Table 3 and 4 also report the results for the cases of redundancy in the system or not, depending on the number of independent components which have been identified and estimated.

5.2 Computational Results

First of all, no gene results from the z-scores, i.e., the transformed components, as they do not appear capable of detecting outlying gene values via thresholding. The effects caused by the standardization are such that the more uniform contribution of the estimated components to the genes now prevents from finding outliers. We investigate this aspect in the next section.

Int	**m2**	**m3**	**boot-m2**	**boot-m3**	**dif-n**	**dif-p**	**adM2**	**adM3**
IC-1	106	16	113	16	5	13	317	67
IC-2	79	26	81	26	39	21	236	81
IC-3	124	40	126	42	38	31	260	94
IC-4	104	34	104	34	63	17	252	99

Table 1: The *dif-n* and *dif-p* are computed $3 * \sigma$ from the mean, while *adM2* and *adM3* indicate intervals computed from (2 and 3 times) the deviations from the median.

The sizes of the *m2* and *boot-m2* intervals are compatible, likewise those of the *m3* and *boot-m3* intervals[6]. The *adM2* intervals appear too relaxed for suggesting any real discriminatory power, while *adM3* corrects this problem. The other intervals, *dif-p* and *dif-n* are of intermediate size between that of the *m3* (or *boot-m3*) and the *adM3* intervals.

[5]These labels appear as the identifying labels attached to the original list of genes.

[6]This follows from the fact that when we measure the mean of the estimated components and of the corresponding bootstrapped ones, the deviation is small.

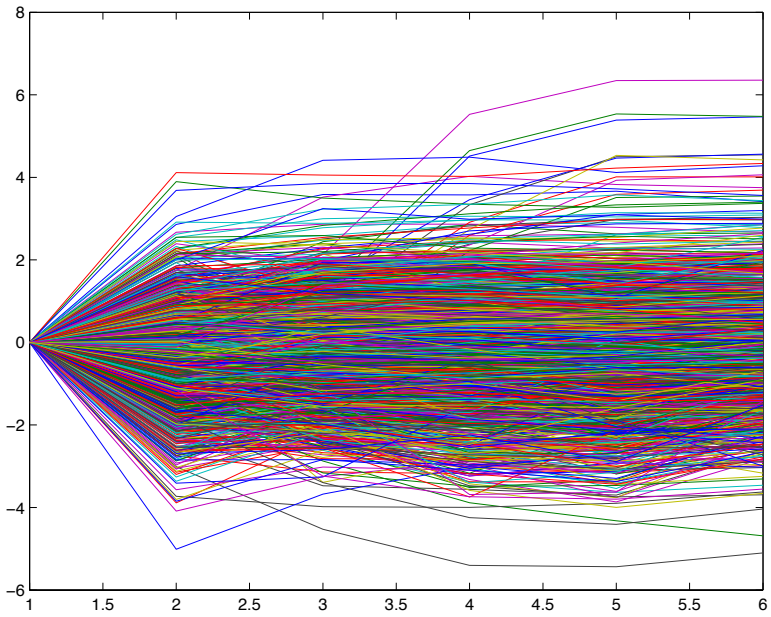


Figure 1: Genome: gene expression levels (y) and temporal patterns (x)

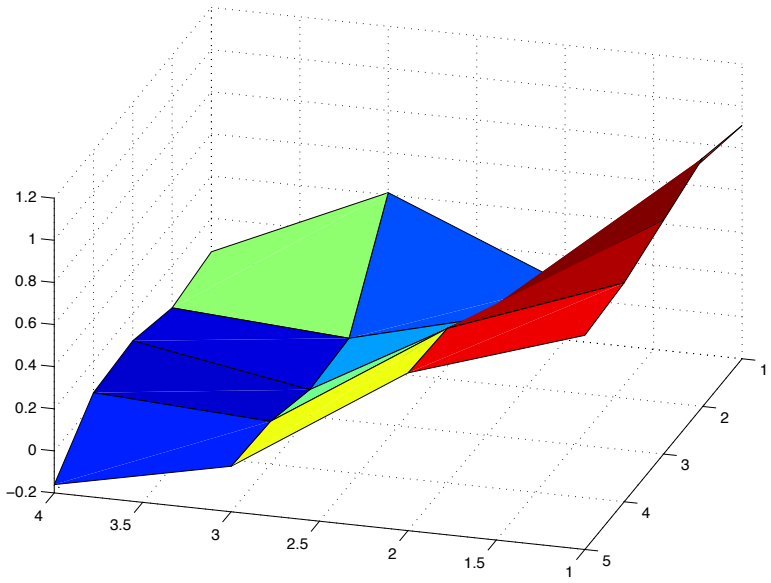


Figure 2: IC (x) Correlation Surface over five time periods (z)

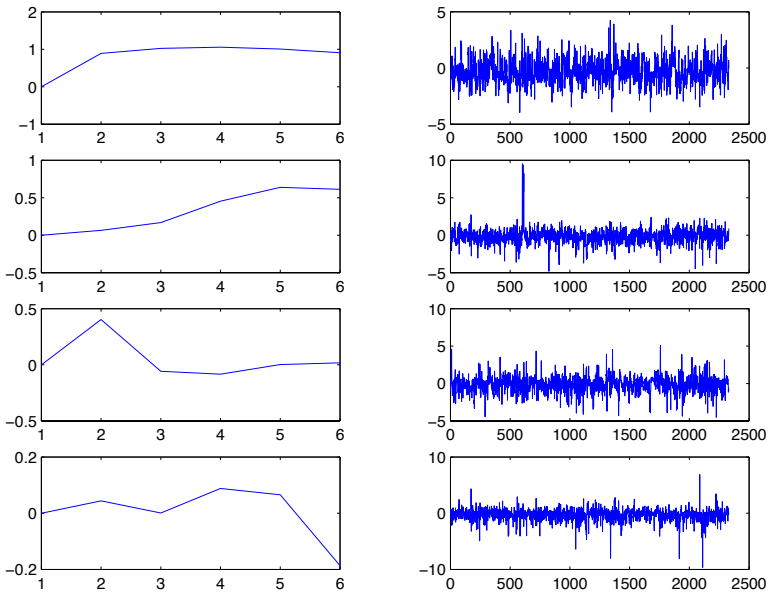


Figure 3: Columns(A) (left), IC signatures (right). From top to bottom in sequential order (first top).

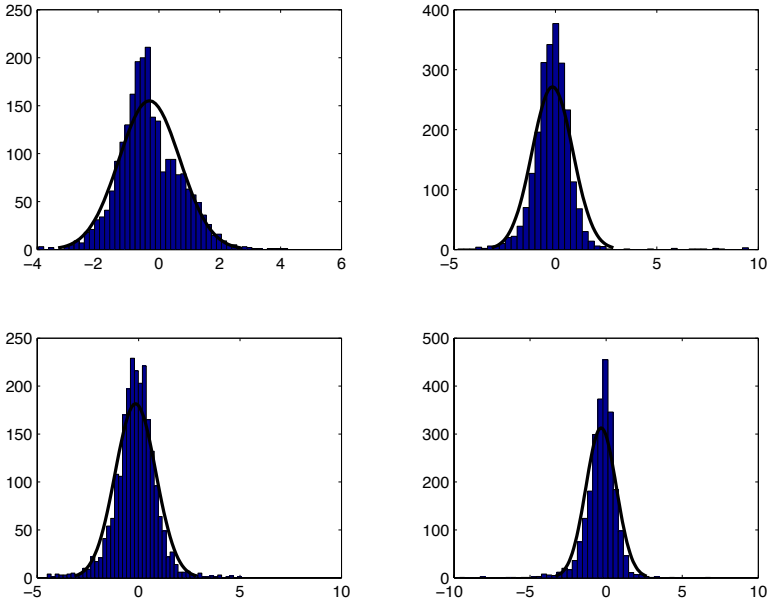


Figure 4: From IC-1 (top-left) to IC-4 (bottom-right): histograms and superimposed density fits

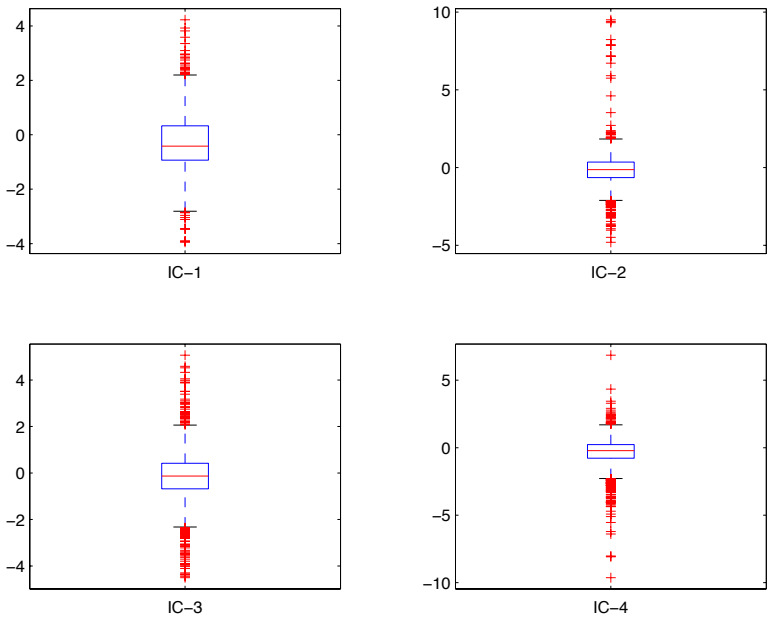


Figure 5: From IC-1 (top-left) to IC-4 (bottom-right): boxplots

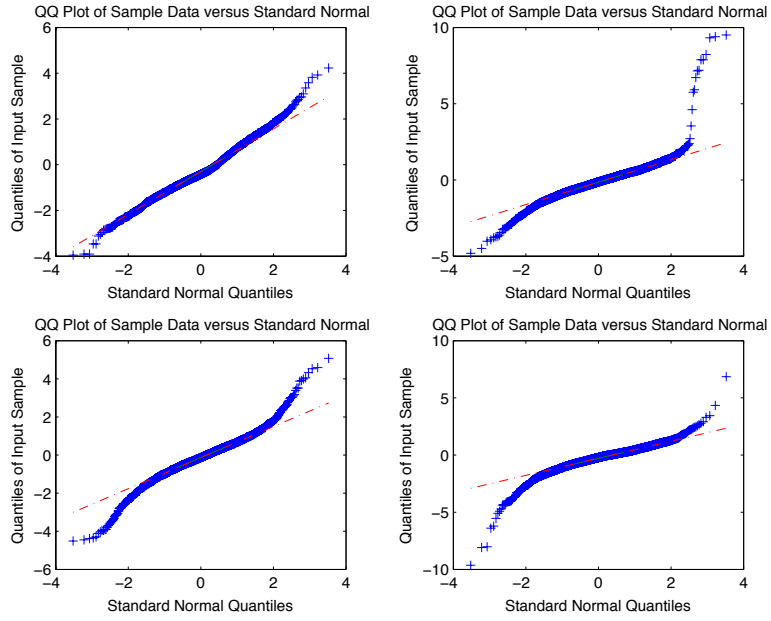


Figure 6: From IC-1 (top-left) to IC-4 (bottom-right): QQ-plots

Int	m2	m3	dif-n	dif-p	adM2	adM3
IC-1	112	17	5	15	306	58
IC-2	79	24	28	40	220	86
IC-3	120	41	31	31	263	104
IC-4	119	49	74	35	249	120
IC-5	100	35	25	45	266	93
ICz-1	46	−−				
ICz-2	66	−−				
ICz-3	19	−−				
ICz-4	29	−−				
ICz-5	55	−−				

Table 2: Z-scores are reported, and indicated with ICz-i. For *m3* no value has been reported.

What we observe in Table 2 is a re-distribution of genes among groups occurring when five instead of four IC are estimated; despite a comparable order of magnitude, more genes participate to the selection process[7]. The same size-based rank of before holds now for the most stringent *m3*) intervals.

Overall we can say that testing the significance of the genes for their selection in distinct groups leads in both circumstances, with four or five estimated components, to a quite relevant dimensionality reduction of the genome size, approximately of one or two order of magnitude, depending on the intervals.

In the 5-component case we have found a certain reduced discriminatory power in selecting genes also with z-scores, unlike with the 4-component case where there is no sign of such power. The *m2* groups are quite also different in size, and show just a moderate degree of alignment with the genes selected without z-scores[8], while no *m3*-selected values appear.

By looking at the groups formed with four or five components, we can account for the co-expressed genes which overlap the groups, and are thus selected by different estimated components because most likely regulated by various biological factors, up to some level of noise and approximation error.

The lists of these genes are reported in Table 3, with the numbers identifying the number of genes appearing two, three or four times across the estimated components. We find (see Table 4) that with four components only three genes are present in four groups, while a list of other 13 genes has score 3, and a list of 85 genes has score 2. They the reduce to zero, one and nine genes, respectively, when *m3* intervals are considered, the remaining only one gene with score 3.

With five components only three genes are present in four groups, i.e., have score 4, while 20 genes have score 3, and 142 genes have score 2. When *m3* intervals are considered, three genes have score 3, and a list of 21 genes has score 2. Among the genes with score 3, six genes are showing up in both the 4- and 5-component *m2* intervals, while one gene has score 3 with four components and score 4 with five components.

Through the sequential search, also reported in Table 3 and Table 4, different groups are

[7]No bootstrap has been reported here.

[8]Supporting evidence is available in some files.

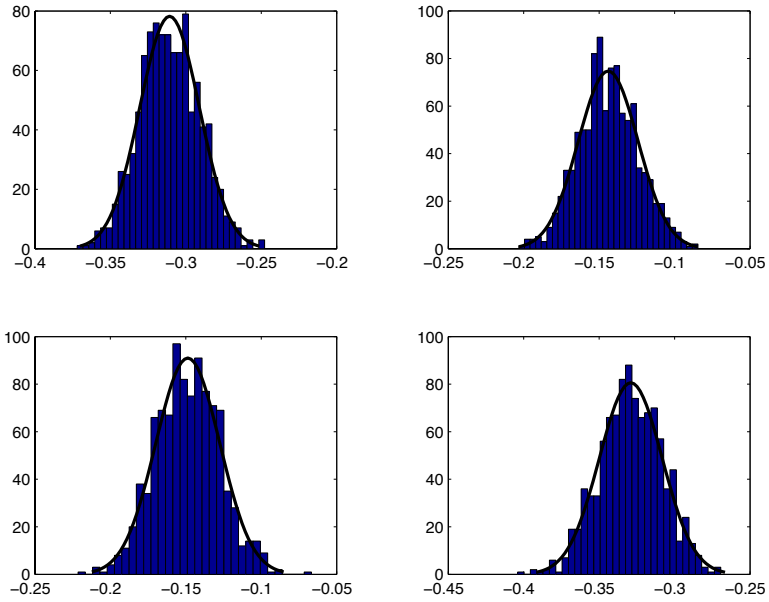


Figure 7: Bootstrap distributions of the component sample means (IC-1, top-left)

formed at more selective intervals ($3 * std$ from the mean), and thus three genes have score 3 with four IC, while in the case of five IC we find one gene with score 4, and a list of five genes with score 3.

When medians and average deviations from their values are computed, we have a new series of results. While two genes appear with score 4 in the case of five IC, one of the two has score 3 with four IC, and the other one has just score 2 (not reported). Two genes are then with score 3 in both cases, together with all different genes in the remaining lists of entries for the two cases.

Score by frequency	4	3	2
4 IC - *m2*	1	13	85
4 IC - *m3*	−−	1	9
4 IC - *seq/m3*	−−	3	39
4 IC - *adM/m3*	−−	7	60
5 IC - *m2*	3	20	142
5 IC - *m3*	−−	3	21
5 IC - *seq/m3*	1	5	54
5 IC - *adM/m3*	2	17	119

Table 3: Differentially expressed genes appear in 4,3,2 groups, according to the kind of interval considered. The numbers indicate the size of the groups.

We observe that apart from the *m2* intervals, the three different (and more stringent) intervals used for the significance tests have a different discriminatory power depending on a system with four or five components. By looking at the scores the groups' composition has changed; thus, the quality of the gene selection results is not only determined by the

Score by frequency	4	3
4 IC - *m2*	3766	296; 1068; 2096; 2393; 2980; 3672; 3914; 3927; 3989; 4068; 4189; 4239; 4354
4 IC - *m3*	−−	4239
4 IC - *seq/m3*	−−	1068; 3989; 4239
4 IC - *adM/m3*	−−	1068; 2980; 3914; 3927; 3989; 4068; 4239
5 IC - *m2*	1068; 2685; 3285	296; 306; 811; 952; 1109; 1154; 1157; 2279; 2280; 2282; 2364; 2980; 3763; 3914; 3927; 3989; 4239; 4240; 4251; 4253
5 IC - *m3*	−−	1154; 3989; 4239
5 IC - *seq/m3*	3285	1154; 1157; 2685; 3989; 4251
5 IC - *adM/m3*	1068; 3285	290; 306; 924; 952; 1154; 1157; 1819; 2279; 2280; 2282; 3763; 3914; 3989; 4239; 4240; 4251; 4253

Table 4: Differentially expressed genes appear in 4,3,2 groups, according to the kind of interval considered. The numbers identify each gene.

informative content of each estimated component but also by their number. Therefore, an issue of reliability becomes a key factor in order to accept a near-optimal number of components for approximating the system.

5.3 Redundancy

The presence of genes selected in intervals computed from the z-scored components has been matter of further investigation, likewise the issue of the possible benefits from the presence of redundancy of IC in the system.

Before investigating the effects of identifying and estimating extra components, some natural questions arise:

- Are the observed changes a result of the addition of an extra informative component or is it just a by-product of more randomness inserted in the system?
- Is the increased complexity of the system with more components yielding better group separability power or instead more instability?
- Is there more statistical accuracy gain, or instead just overfitting?

As we know from signal processing, redundancy may bring robustness to noise in the system, as often shown in applications. From the data pre-processing step enabled by PCA it appears that one may equivalently consider whether small eigenvalues could be kept together with the others of larger size, which brings of course the risk of embedding the system with noise, and thus overfitting power when ICA is estimated.

It turns out that despite its possible benefits and risks, redundancy needs to be carefully considered on the grounds of system stability and signal-to-noise ratio quality. Separability power may be lost with z-scores, as it occurs with four IC in the genome case, but not with

five IC. The fact of finding discriminatory power when analyzing systems with more than a minimal number of components may represent for these data a sign of the presence of undesirable redundancy in the system, i.e., the one carrying noise instead of informative signal.

6 Entropy

6.1 General Aspects

The well-known peculiarity of ICA is that its embedded independence property goes beyond the decorrelation power sufficient only for second-order dependent systems, as in order not to miss important data features under conditions departing from Gaussianity one needs more than simple decorrelation. Thus, by avoiding the limitation of an infinite number of linear transforms able to yield decorrelation, ICA leverages on mutual information.

From [7], if one uses linear ICA with normalized components, then minimizing the mutual information is equivalent to minimizing the sum of entropies of all the components. Furthermore, mutual information can be decomposed in a term which represent the decorrelation between the components plus another term dealing with the characterization of the system in terms of Gaussianity.

The mutual information between two random variables is given by:

$$I(x,y) = \int p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{1}$$

It can also be defined in relation to entropy, and for a random vector Z, as follows:

$$I(Z) = \sum_j H(z_j) - H(Z) \tag{2}$$

The $H(.)$ stands for the Shannon's differential entropy quantity, which for continuous random variables is given by:

$$H(Z) = -\int_Z P_Z(z) log P_Z(z) dz \tag{3}$$

It was noted [7] that since the entropy $H(Z)$ remains constant under a transformation given by the product of whitening and orthogonal matrices, thus $I(Z)$ varies with the remaining quantity in Eqn. (2), 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. We then show another example, also supported by graphical evidence, about the relation between linearly estimated components which are as much independent as possible, and thus result uncorrelated, but also as much non-Gaussian as possible, due to the ICA separation processing.

In other words, as dependence can be seen as a sum of correlation and marginal Gaussianity effects which can be strongly, weakly or even non present, we monitor the signature of the signals extracted from ICA compared to other patterns subject only to linear correlation,

so as to emphasize the capacity of ICA to filter out of the observed mixtures more distinct information.

6.2 Stability Check

In general, by mapping variables from one domain to another, one aims to reduce the variability linked to noise. An high degree of dispersion/variability in the projected domain is an unfortunate outcome, contradicting the strongly pursued paradigm of dimensionality reduction and missing the goal of techniques such as PCA or ICA targeted to improve the signal-to-noise ratio and help selecting input variables.

But the variance of a signal is not necessarily related to the relevance of the input variables, the latter depending on the statistical properties characterizing the input domain. Thus, without considering these properties the risk is that only variance-sensitive selected features have not much to do with the desired targets, which in our case are represented by differentially expressed genes.

Having made the distributions of the estimated and then transformed components more uniform affects the overall variability of the system, by changing the between- and within-source variability compared to before when normalized components were delivered by the ICA algorithm. The main consequence is that the new variability level gains power over other statistical properties such as non-Gaussianity and independence, and as a result decreases the discriminatory power of thresholding.

It is thus clear that with four components the fact of conditioning the discrimination power for detecting differentially expressed genes on their high- and low-variance contributions can limit the potential of finding more interesting relations. It is also fair to accept the fact that spurious dependencies may always arise in system because of noise, or by using linear approximations in highly complex networks of variables. Thus, one would ideally like to minimize the risk of including false features through the tests.

There is an interesting connection also with the entropy reduction method [29], where given that non-linearities in the data are better described when higher moments of the involved distributions are accounted for, the method suggests to concentrate the non-linear variation in a certain random variable over a restricted subset of the other variable 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.

Figure 8 illustrates the difference among a system with four (top panel) and five components, for both the estimated sources (left panel) and their z-scored versions.

The gene profiles attached to each estimated component have been z-scored; this has produced a convenient visualization outcome, where a sphere is now obtained from the z-scores and the transformed variables lie in it. We use this evidence for investigating the extent to which we might discriminate information from noise in the system, or in other words if the projected space is endowed with more or less instability.

This contributes to explaining why the z-scores produce values with five components and not with four. The link we make to support our findings is with entropy, as the more random the system is, the more likely one can expect to find an increased entropy. Thus, as entropy in discrete random systems is usually defined as $H(x) = -E_x[log\ p(x)]$ for a sequence x with probability $p(x)$, for a random variable it can be interpreted as the degree of information

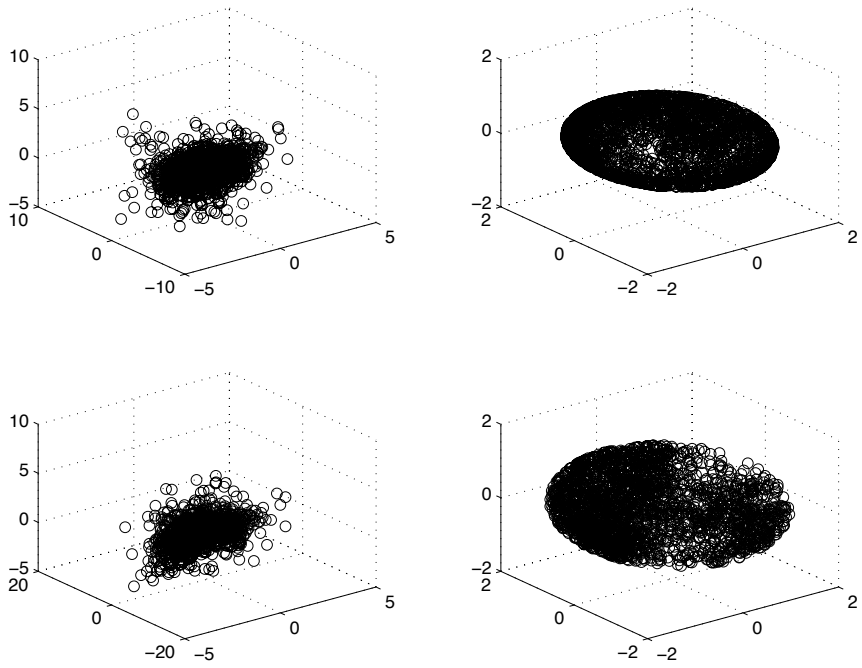


Figure 8: Scatter of IC_4 (top-left) and ICz_4 (top-right) vs corresponding plots of IC_5 and ICz_5.

delivered, where the more random or unpredictable the variable is the bigger is its entropy.

The *Sample Approximate Entropy* (SApEn, from [23]) 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.

In a sequence, if patterns of fluctuations are repeated the sequence becomes more predictable; *SApEn* reflects the chance that similar patterns are or not observed. When many of them are present, then *SApEn* estimate is relatively small, which makes more predictable the sequence. Otherwise a more complex and random sequence will deliver 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.

We found that the z-scores for the system with four or five IC indicate 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 verify this through some numerical computation of dispersion and empirical entropy measures. Table 5 is for the sample entropy. By looking at the estimated entropies, the system with five components is always more random for the representation via z-scores.

Table 6 shows the results for the sample variance estimate. Note that the estimated IC result always with unit variance, by construction, while their z-scored versions differ in way that shows how with an extra component the dispersion increases. As a further check we have also computed the empirical entropy variance (see [30]), an estimate for the quantity $var(H) = var[-log(p(x))] = E[log(p(x))^2]$. The results are reported in Table 6 together

with the other variances[9].

SApEn (at m)	1	2	3	4	5
$IC1_4$	2.09	1.82	1.77	1.72	1.6
$IC2_4$	1.91	1.73	1.7	1.67	1.63
$IC3_4$	2.03	1.81	1.77	1.69	1.59
$IC4_4$	1.93	1.82	1.77	1.7	1.7
$IC1_5$	2.09	1.82	1.77	1.72	1.62
$IC2_5$	1.89	1.73	1.7	1.69	1.7
$IC3_5$	2.02	1.78	1.75	1.7	1.63
$IC4_5$	1.88	1.80	1.75	1.74	1.71
$IC5_5$	1.96	1.84	1.79	1.74	1.79
$ICz1_4$	2.01	1.79	1.68	1.68	1.5
$ICz2_4$	2.0	1.74	1.7	1.66	1.61
$ICz3_4$	2.01	1, 77	1.7	1.59	1.49
$ICz4_4$	2.04	1.92	1.88	1.9	2.02
$ICz1_5$	2.07	1.83	1.75	1.7	1.71
$ICz2_5$	2.05	1.88	1.85	1.89	1.86
$ICz3_5$	2.08	1.93	1.9	1.87	1.87
$ICz4_5$	2.02	1.96	1.95	1.94	2.03
$ICz5_5$	2.0	1.86	1.85	1.87	1.91

Table 5: *Sample Approximate Entropy* for each IC and ICz at epoch $m = 1, \ldots, 5$. The subscript $_4$ or $_5$ indicates the number of components in the system.

6.3 Linear Correlation and Dependence

ICA finds components or modes as much statistically independent as possible. Thus, at each observation point one can measure what influence each component has on the observed gene profile by looking at the correlation coefficients computed between these two quantities.

In other words, tracking the pattern of the estimated coefficients at each time point reveals the contribution of each component across the different conditions. The correlation coefficient pattern is reported in Table 8 and a plot is presented together with additional considerations about this issue. The estimated Pearson coefficient is used to as to give an indication of the linear dependence of the estimated (or the bootstrapped) components and the time point-wise sample profiles.

However, further computations have been carried out; in Figure 9 and Figure 10 graphical evidence of these results is provided. First, from Figure 9 it is clear that IC-1 is strongly correlated with the gene profile across samples (see the first column). All the other components are more or less decorrelated, moderately for the case of IC-2 and more for the other two components.

Then, the genes which are differentially expressed across samples are put in relation with those delivered by the estimated components. When the correlation of sample and

[9]Only the case of $m = 1$ is proposed.

Sample Variance (at each IC)	1	2	3	4	5
IC_4	1	1	1	1	
ICz_4	0.83	0.69	0.68	0.68	
IC_5	1	1	1	1	1
ICz_5	0.92	0.73	0.75	0.65	0.68
Sample *SApEn* Variance (at each IC)	1	2	3	4	5
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 6: Sample Variances (top) and Entropy Variances (botom) for each IC. The subscript $(.)_4$ or $(.)_5$ indicates the number of components in the system. The letter z stands for z-scores.

IC-dependent gene profiles is high a stronger alignment of the patterns (i.e., the lists of extracted genes in the intervals) is observed. Table 7 shows the number of genes selected by thresholding sample-wise the gene profiles over the *m2* and *m3* intervals.

Sample profiles	x-2	x-3	x-4	x-5	x-6
m2	114	115	129	116	126
m3	17	26	17	21	27

Table 7: Differentially expressed genes according to the kind of interval considered and measured across time points (from time 2 to 6).

In Figure 10 some patterns are shown after considering the sequences of genes selected by thresholding, where the top four plots report sample-wise selections and the other plots refer to the lists of genes attached to the estimated components. These gene sequences are quite different in relation to the two domains (sample-wise and IC-based) where thresholding operates. The IC-1 and IC-2 patterns (at the left-hand third row panel) concentrate most of the variability, correspondingly to the sample dependent patterns.

However, pairwise comparisons for the other components and sample-wise patterns suggest that different mechanisms underlie the selection process, depending on the statistical properties behind the detection power of the outlying genes, and despite the likely presence of noise.

Given the strong form of dependence exploited by ICA, the estimated components select genes in groups by considering across-sample dynamics, and thus a more global dependence structure is accounted for, compared with linear correlation.

The sample-wise gene groups detected are instead specifically sample time-related. We thus expect these latter groups to be much more sensitive to the linear correlation pattern, and thus overlap to a much larger extent compared to the groups selected by the estimated components and endowed with statistical independence.

In other terms, it appears that following the correlation patterns one may get a better or worse alignment between gene groups, but the information brought by each component is not necessarily related to the one determined sample-wise. While one can say that the stronger

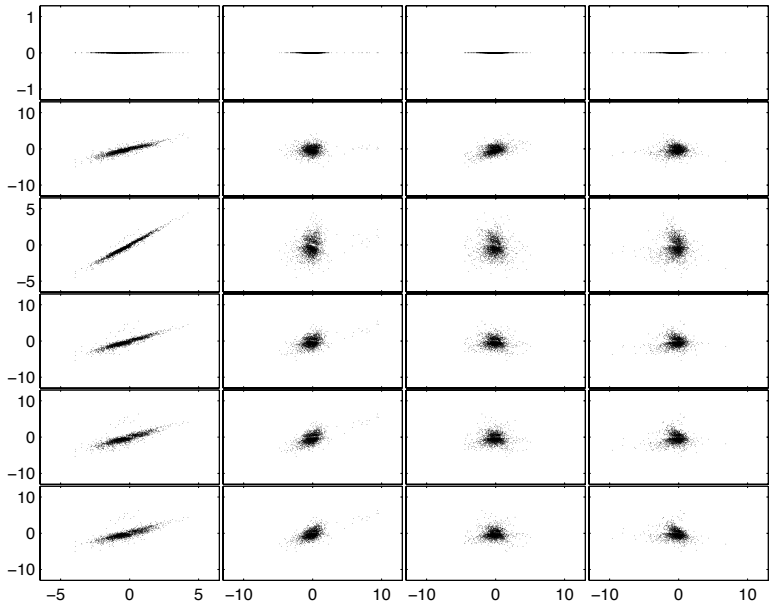


Figure 9: Linear Correlations: Independent Components (x) vs Sample Profiles(y)

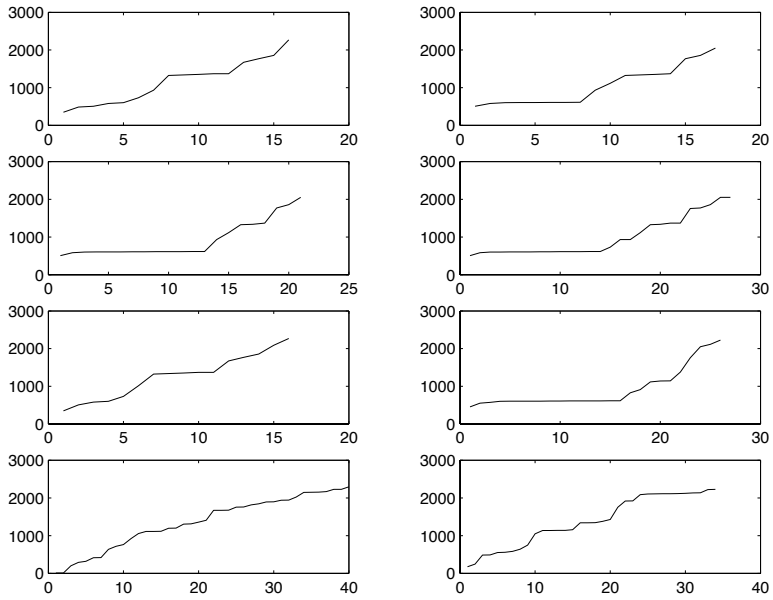


Figure 10: Sample profiles at time 3 (left top-most), 4 (right top-most) and 5, 6 (second row panels) vs IC profiles (from IC-1 third row at left, to IC-4 bottom at right). All selected genes at m3 intervals.

Samples	x-1	x-2	x-3	x-4	x-5	x-6
bootIC-1	−−	0.908	0.98	0.91	0.84	0.817
std	(−−)	(0.005)	(0.001)	(0.008)	(0.013)	(0.014)
IC-1	−−	0.908	0.98	0.91	0.84	0.817
bootIC-2	−−	0.067	0.16	0.389	0.0534	0.549
std	(−−)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
IC-2	−−	0.066	0.16	0.39	0.533	0.551
bootIC-3	−−	0.41	−0.055	−0.073	0.003	0.016
std	(−−)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
IC-3	−−	0.41	−0.055	−0.072	0.002	0.015
bootIC-4	−−	0.04	0.0006	0.07	0.05	−0.16
std	(−−)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
IC-4	−−	0.04	0.0006	0.07	0.05	−0.17

Table 8: Bootstrapped correlations (*bootIC-i*, with *std* taken from the man values.) vs data-based estimated correlation coefficients. Due to normalization and log-transformation, the first-sample coefficient is zero.

the correlation and the smaller is the fraction of data one needs to keep for interpreting the underlying dynamics, this fact my reduce too much the rank of the estimated mixing matrix with the risk of excluding the gene-related information carried by some components.

Related to this last aspects, and also to the noise issue, it is interesting to take a closer look at the estimated mixing matrix A for both the cases under exam, with the 4 or 5 components for the genome sets[10]. They reveal that the third and fourth columns (associated with four components) and the fourth and fifth ones (associated with five components) are quite sparse, which indicates a minor contribution of these sources in explaining the dynamics of the mixture profile observed each time.

Thus, a sort of control over the level of redundancy inserted through the extra component is active in the system, and allows for the noise effects to be smoothed out. This context has not even ideally a negative characterization, as many natural signals can be sparsely represented, and it has been demonstrated (see [31]) that not only the sparsity of source signals may help their separation in time domain, but when this is not the case for some kinds of signals, a sparse representation can be obtained through suitable signal dictionaries, usually not restricted to operate in the time domain (see, for instance, [3, 5]).

As a last comment, the sparse components have usefully contributed to the gene selection process as well, and unless the noise signature is clearly predominant over them, the corresponding gene lists are worth of close consideration.

7 Conclusions

Exploratory work about ICA is provided for genomic data obtained through experimental work. While the full biological relevance requires further tests and experiments, the initial

[10]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
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 9: Estimated Mixing Matrix with 4 components.

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 10: Estimated Mixing Matrix with 5 components.

results are promising and can suggest a strong role for ICA in this field. We envisage the following advantages from the use of ICA:

- Dimensionality reduction and gene selection can be efficiently obtained via thresholding over the estimated components and through various significance tests for the computed gene expression values;
- Co-expressed genes can be identified as those differentially expressed in a number of estimated components, and might be a useful informative source for determining the co-regulated activity of genes and the sparsity structure which characterizes the gene network interaction matrix;
- Redundancy in the system can in principle be supported by the introduction of more than minimal number of components, but it has to be monitored in its effects on gene selection and robustness to noise, as the insertion of extra components may just bring in weak sources which deteriorate the discriminatory power of thresholding and the quality of gene groups;
- Z-scores have misleadingly shown discriminatory power with five components in the system, as this becomes embedded with noise which is responsible for boosting the expression values in the significance region.

The conclusive remark concerns the role of ICA for gene network analysis. ICA is a powerful exploratory instrument which works as an approximating tool in systems with complex dynamics under very weak assumptions and constrains, and results computationally efficient even in its simple linear representation.

Both the dimensionality reduction and the gene selection tasks are in principle naturally combined by the ICA approach. This fact might lead to an excellent quality of exploratory analysis compared to other standard methodologies unable to exploit the same statistical properties of ICA.

By looking in perspective, ICA opens wide possibilities for future research in genetic regulatory network analysis and reverse engineering approaches, as it suggests an effective way of dealing with dimensionality reduction in noisy complex variable spaces, and suitably interpreting the mixing mechanisms which regulate the influences underlying the gene-gene interaction dynamics.

8 Acknowledgments

The contribution of Tim Gardner in introducing the author to the problem and explaining the experiments is gratefully acknowledged, as well as the data preparation done by Ilaria Mogno, whose accurate work is highly appreciated.

References

[1] Alter O., Brown P.O., and Botstein D., (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, 97, 10101-10106.

[2] Berger J.A., Hautaniemi S., Edgren H., Monni O., Mitra S.K., Yli-Harja O., and Astola J., (2003). Identifying underlying factors in breast cancer using independent component analysis. *Proc. of the 2003 IEEE Int. Work. NNSP*, Toulose, FR, 81-90.

[3] Bofill P., and Zibulevsky M., (2000). Blind separation of more sources than mixtures using sparsity of their short-time fouries transform. *Proc. ICA*, 87-92.

[4] Brown M.P.S. *et al* (2000). Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc. Natl. Acad. Sc. USA* 97(1), 262-267.

[5] Capobianco E. (2003). Independent Multiresolution Component Analysis and Matching Pursuit. *Comp. Statist. Data Anal.* 42(3) 385-402.

[6] Cardoso J. (1989). Source separation using higher order moments. *Proc. Int. Conf. ASSP*, 2109-2112.

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

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

[9] Chiappetta P., Roubaud M.C., and Torresani B., (2003). Blind Source Separation and the Analysis of Microarray Data. *J. Comp Biol.*, to appear.

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

[11] Dudoit S., Fridlyand J., aand Speed T.P., (2002). Comparison of discrimination methods for the classification of tumors suing gene expression data. *J. Am. Stat. Assoc.*, 97, 77-87.

[12] Eisen M.B., Spellman P.T., Brown P.O., and Botstein D., (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95, 14863-14868.

[13] Everson R. and Roberts S., (2001). Particle filters for non-stationary ICA. In: *Independnet Component Analysis: Principles and Practice*, Roberts S. and Everson R. (eds), Cambridge University Press, pp. 280-298.

[14] Friedman J., (2001). Greedy function approximation: a gradient boosting machine. Tech. Rep., Statistics Department, Stanford University, CA. *Ann. Statist.*, to appear.

[15] Gardner T., di Bernardo D., Lorenz D., and Collins J.J., (2003). Inferring genetic regulatory networks and identifying compound mode of action via expression profiling. *Science*, 301, 102-105.

[16] Golub T.R., *al* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537.

[17] Hastie T. and Tibshirani R., (2004). Efficient Quadratic Regularization for Expression Arrays. *Biostat.*. 5, 329-340.

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

[19] Hori G., Inoue M., Nishimura S., Nakahara H. (2001). Blind gene classification. An application of a signal separation method. *Genome Informatics*, 12, 255-256.

[20] Hori G., Inoue M., Nishimura S., Nakahara H. (2002). Blind gene classification. An ICA-based gene classification/clustering method. Riken BSI/BSIS Tech. Rep. N. 02-5.

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

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

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

[24] Lee S. and Batzoglou S., (2003). Application of independent component analysis to microarrays. *Genome Biology*, 4:R76.

[25] Liao J., Boscolo R., Yang Y., Tran L., Sabatti C., Roychowdhury V., (2003). Network component analysis: reconstruction of regulatory signals in biological systems. *PNAS*, 100(26) 15522-15527.

[26] Liao X., and Carin L., (2002). Constrained Independent Component Analysis of DNA microarray signals. *Proc. IEEE SAM Multichannel SP Workshop*, VA.

[27] Liebermeister W. (2002). Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18, 51-60.

[28] Nadal J.P., Korutcheva E., and Aires F., (2000). Blind source separation in the presence of weak sources. *Neural Networks*, 13, 589-596.

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

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

[31] Zibulewsky M. and Pearlmutter B.A. (2001). Blind Source Separation by Sparse Decomposition in a Signal Dictionary. *Neur. Comp.*, 13(4), 863-882.