# Finding structured gene signatures

Sofia Mosci
DIFI -DISI
Università degli Studi di Genova
via Dodecaneso 35, I-16146 Genova, Italy
mosci@disi.unige.it

Annalisa Barla
DISI
Università degli Studi di Genova
via Dodecaneso 35, I-16146 Genova, Italy
barla@disi.unige.it

Alessandro Verri
DISI
Università degli Studi di Genova
via Dodecaneso 35, I-16146 Genova, Italy
verri@disi.unige.it

Lorenzo Rosasco
CBCL - Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139
lrosasco@mit.edu

## Abstract

*In the context of gene signature identification from microarray data, a main problem is devising statistical and visual tools to interpret and understand the biological meaning of the selected genes. Most available statistical tools for gene signature extraction typically provide unstructured list of genes and lack the capability of handling correlation among genes. Recently an algorithm for feature selection, namely elastic net, was proposed allowing to deal with correlated genes in a transparent way. In this work we exploit the form of the output given by elastic net, as used in [3], to obtain a structured gene signature where genes are disposed in block of intra-correlated genes and the blocks are ranked according to a measure of the block discrminative power. After recalling how elastic net can be used to define nested lists of increasingly intra-correlated genes, we propose an ad hoc agglomerative clustering technique able to refine such a nested output by explicitly identifying modules of correlated genes. We take advantage of such a structure to visualize the correlation patterns underlying the data. The proposed procedure is validated on both synthetic data and applied to real gene expression datasets.*

## 1 Introduction

The analysis of microarray gene expression data has gained a central role in the process of understanding biological processes. Many algorithms have been recently proposed in order to deal with high-throughput data and to select the genes most relevant to characterize a given task. Indeed, genomics analysis has shifted from focusing on one gene at a time to a more complex situation where the action and interaction of many thousands of genes can be measured simultaneously. In other words, techniques such as microarrays are able to quantify the activity of thousands of genes at once, generating a global picture of the cellular function. In this context, one main goal of data analysis is to provide reliable statistical tools, that work in the typical *-omics* scenario of a small number of samples represented in a high dimensional space and that are able to capture the complex interactions among genes. To this aim many techniques have been proposed, including dimensionality reduction methods, projecting the data onto lower-dimensional spaces. For instance, in the case of Principal Component Analysis new linearly independent and non-redundant features are defined in terms of the original ones in a unsupervised way. This approach has been proved useful in many studies [11, 9], but the biological interpretation of the obtained meta-features is a non-trivial task. On the other side, supervised feature selection algorithms select a subset of the original variables identifying the most relevant ones and assigning to each of them a relative weigh. For example a common task is that of determining a classification rule able to assign patients to one among different disease sub-types. The variable selection problem in this case corresponds to finding a subset of genes *relevant* for sub-type classification. A more precise definition of what *relevant* means leads to defining different variable selection algorithms. A popular class of techniques is that based on defining relevant genes as those showing differential expression across classes [5, 10]. Corrected hypothesis testing are the statistical core of these methods. A second class of techniques is that

based on defining subsets of relevant genes as those having discriminative power in terms of classification properties. This latter class of methods also includes wrapper methods and embedded methods, see [7] for references. Unlike many other heuristically motivated schemes, recently a class of theoretically founded embedded methods has been proposed based on the concept of sparse regularization. Most notable examples are probably the well-known *Lasso* [13], *(naïve) elastic net* [14] algorithms, and *forward stage-wise regression* [8].

Despite the availability of feature selection techniques, the biological interpretability of the selected gene lists is often a major problem. This is mainly due to the lack of structure in the selected gene signatures and of complementary visualization tools. Most feature selection algorithms produce (sometimes ordered) lists of genes but are usually unable to handle correlation among genes. Typically in a group of intra-correlated genes only one representative is selected unless ad hoc heuristics are considered. Recently the so called elastic net regularization was proposed to deal with correlation among genes and was shown to be able to select *all intra-correlated genes* rather then only one. A limitation of elastic net is that though we know that the obtained list comprises groups of correlated genes such groups are not explicitly given.

The main idea of our work is to take a step further with respect to elastic net and propose a robust statistical analysis framework able to extract gene signatures that are *structured in modules of correlated genes ranked according to their discriminative power*. The obtained signature can be effectively visualized making the search of *interesting* biological patterns easier.

The preliminary step to our procedure is based on obtaining several raw gene signatures by means of the elastic net regularization, originally proposed by [14] and further studied and used in [2, 4, 3]. Elastic net regularization has several interesting properties. It was proved to be a consistent variable selection scheme [2] (as the number of available training samples increases the best possible estimator is eventually reached. Moreover it takes into account the multivariate effect of many genes together, and avoids discarding correlated variables. We use the two-stage approach to elastic net regularization presented in [3] as a supervised preprocessing of the data. The appealing property of such an approach is that the algorithm output is a one parameter family of nested lists with equivalent prediction ability and increasing correlation among genes.

Such a two-stage procedure, though, does not suggest which are the blocks of intra correlated genes, and a second step is thus needed. To this aim we apply a variation of an agglomerative clustering technique that exploits the nested output of the supervised preprocessing. We start from the first list which is *minimal*, the least number of discriminative genes is selected regardless of intra genes correlation. Such genes are used as *centroids*. Considering the subsequent lists with increasing size and comprising correlated genes we grow clusters of correlated genes around the obtained centroids. The main characteristic is that, differently from usual clustering approaches, the centroids (prototypes) are given and have a meaning in terms of classification ability. In this way we can extract and visualize a more structured genes signature, which captures and make evident the correlation patterns in the data. This provides a richer model that can be used to gain a better understanding of the genes function, possibly leading to new biological hypotheses.

The paper is organized as follows: in section 2 we briefly review the problem of variable selection and illustrate elastic net optimization and its two-stage approach; we then introduce our proposal for determining the genes modules, for ranking them in terms of their prediction power and for visualizing them. In section 3, the proposed methodology is applied on synthetic data sets and real gene expression data.

## 2 Methods

In this section we start with a preliminary description of how to use the elastic net algorithm in order to obtain the nested lists of genes that are at the basis of the extraction of structured signatures. We then show how to apply an appropriate ad hoc clustering technique to find a structure within these gene lists and discuss how the obtained clusters can be ranked according to their discriminative power. Finally we present two customized visualization tools that allow to better interpret and appreciate the results.

### 2.1 Gene Signature Extraction

Given a prediction (classification) task the problem of variable selection amounts to detecting the factors determining good prediction. Indeed this problem, which is often encountered in the context of gene-array analysis, is a classical problem in statistics but classical methods are usually not tailored to the analysis of very high dimensional data, when the number of variables (genes) $d$ is much larger than the available set of examples (patients) $n$.
To tackle the gene signature extraction, we adopt the elastic net approach [14]. More specifically, given a response vector $\mathbf{Y} = (y_1, y_2, ..., y_n)$ of labels and a $n \times d$ gene expression matrix $\mathbf{X}$, where each row $\mathbf{x} \in \mathbb{R}^d$ is an exam-

ple (patient, cell line, treatment) and each column a gene, we consider the linear classification rule $sign(\boldsymbol{\beta} \cdot \mathbf{x})$ where the coefficient vector $\boldsymbol{\beta}$ is unknown and assigns a weight to each gene. The elastic net solution is obtained by minimizing w.r.t. $\boldsymbol{\beta}$ the following functional:

$$\Phi_{\lambda,\epsilon}(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda(\sum_{j=1}^{d} |\beta_j| + \epsilon \sum_{j=1}^{d} |\beta_j|^2). \quad (1)$$

The above functional was introduced to overcome some drawbacks of the Lasso approach [13] which solution is obtained by minimizing $\Phi_{\lambda,0}$. Indeed the lasso penalty $\sum_{j=1}^{d} |\beta_j|$, namely the $\ell_1$-norm of $\boldsymbol{\beta}$, favors gene selection by setting most coefficients on $\boldsymbol{\beta}$ to zero but has shortcomings in the presence of correlated genes. *Classification performance does not change if we select one or more relevant but correlated genes*. On the other hand, adding the $\ell_2$ penalty $\sum_{j=1}^{d} |\beta_j|^2$ can be shown to enforce correlated genes to have similar weights, so that a whole genes group is either discarded or selected. The parameter $\epsilon > 0$ is a threshold determining the correlation level above which genes are to be considered as belonging to the same group.

While in [14] the two parameters $\lambda$ and $\epsilon$ are chosen via cross-validation, hence selecting only one list of significant variables, we follow the two stage procedure described in [3], where the tuning of the $\epsilon$ parameter allows to get an output with more structure than a simple list of genes. Here we briefly recall the main concepts of this procedure. In stage I, the optimal $\lambda$ in (1) is chosen via cross validation, [1]. In stage II, the algorithm is ran with the same value of $\lambda$ for increasing values of $\epsilon = \epsilon_1, \ldots, \epsilon_{\max}$, and for each value of $\epsilon$ a different set of weights $\boldsymbol{\beta}$ is obtained. The first set of selected genes ($\epsilon = 0$) is minimal, most coefficients are set to zero and no correlation is considered. Increasing $\epsilon$ causes correlated genes to be selected (they are assigned non zero weights). As the output of the procedure we have several genes lists of increasing size with genes ranked according to their weights. Though we know that genes are correlated, the algorithm does not give an explicit grouping. In the next section we discuss a simple way to extract such groups from the output of elastic net.

## 2.2 Structured clustering of two-stage elastic net output

The two stage approach described above has proved to be an efficient technique for gene selection [3, 1]. Moreover, the nested structure of the selected gene lists allows to choose the desired level of complexity to be used in the biological investigation of the underlying phenomenon. In other words, we can choose how many genes we want to consider for further studies, maintaining all the information extracted from the data. For example, when interested in finding a set of biomarkers to be used on large scale diagnostic tests, one might prefer a small panel of significant genes due to time, cost and resource limitations. On the other hand, when the main goal is the comprehension of the entire cell response to an external factor, such as a particular drug treatment, the maximal list ($\epsilon$ max) is preferable.

Despite this appealing property, the raw structure of the gene lists does not provide any understanding on how to group correlated genes. All we have is a set of lists ordered according to their size. Starting from the first minimal list of genes the following lists include more correlated genes. The main contribution of the present work is to proceed one step further towards the biological interpretability of the selected gene signatures. The idea is to build blocks of correlated genes using a variation of well known agglomerative clustering techniques. The technique is based on the Pearson distance:

$$d(X_{j_1}, X_{j_2}) = \frac{corr(X_{j_1}, X_{j_2})}{\sqrt{var(X_{j_1})var(X_{j_2})}}$$

evaluating the (normalized) correlation between the $j_1$-th and the $j_2$-th columns (gene expressions) of the data matrix $\mathbf{X}$.

The $B$ genes in the minimal list can be assumed to be independent. Therefore the structured gene signature will be constitued by $B$ blocks. Each gene $p_b$ in the minimal list will be used as a *prototype* for the $b$-th block. For each gene in the maximal list, correlation is calculated with each prototype and hence the gene is assigned to the block (prototype) associated to higher correlation. In this way we populate the blocks corresponding to the prototype genes in the minimal list with all the genes coming from the maximal list and encompassing all its sublists. Within each block we sort its genes according to the order of appearance while performing the two stage elastic net selection, for increasing $\epsilon = \epsilon_1, \ldots, \epsilon_{\max}$. In block $b$, the prototype $p_b$, selected for $\epsilon_1$, will be ranked first, its correlated genes appearing at $\epsilon_2$ will be ranked second (with ties) and so on.

## 2.3 Our Ranking Criterion

We now propose a criterion for ranking the gene blocks according to their prediction power with respect to the supervised problem under study. Since the estimator $y \sim f(\mathbf{x})$ is a weighted sum of the genes expressions, $f(\mathbf{x}) = \sum_{j=1}^{d} X_j \beta_j$, we can define the score, $s_b$ for block $b$ as the contribution given to $f(x)$ by the prototype of block $b$:

$$s_b = \left\| \mathbf{X}_b^{(1)} \boldsymbol{\beta}_b^{(1)} \right\|$$

where $X^{(1)}$ is the gene expression matrix resctricted to the genes selected with the lowest value of $\epsilon$, and $\boldsymbol{\beta}^{(1)}$ is the corresponding optimal weight vector.

### 2.4 Output Visualization

Biological data analysis and visualisation have traditionally been approached as independent problems. Relatively little attention has been given to the integration and visualisation of information and models. However, the integration of these areas facilitates a deeper understanding of problems at a systemic level. In order to interpret and appreciate the results obtained with the agglomerative clustering technique proposed in Subsection 2.2, we implemented two simple visualization tools.

In the first method, we first restrict the gene expression matrix to the probe sets belonging to the blocks union. We then rearrange its columns in order to emphasize both the blocks and the layered structure, and display its correlation matrix as an image: genes belonging to the same groups are drawn close to each other, and thick lines separate each block; in addition, within each module, the gene in the upper left corner has to be identified with the prototype or first layer gene, whereas the genes selected with increasing value of $\epsilon$ follow, separated by a thinner line.

In the second visualization, we project the genes appearing in the blocks union on the most representative *meta-patient*. Such meta-patient is identified with the 3D space spanned by the first three left eigenvectors of the normalized expression matrix $X_{ij}$ restricted to the genes which belong to the blocks union. Clearly highly corelated genes cluster together in the 3-dimensional space, while collinear genes present perfect overlap.

## 3 Results and Discussion

### 3.1 Toy Data

In order to test our approach in a controlled setting, we applied the two stage elastic net regularization, followed by the nesting-clustering technique on a toy example, where we exactly know which are the relevant or correlated features. The proposed toy problems are close to real data conditions, e. g. dependence on more than one variable and correlation, though in a lower dimensional setting. A set of $n = 30$ toy-patients are drawn from $\mathbb{R}^d$ with $d = 50$ in the following way:

$x_1, x_{11}, x_{21}, x_{31}, \ldots, x_{50}$ i.i.d. from $\{-0.5, 0.5\}$,
$x_{1+i} \sim s_i \cdot x_1 + \sigma_1 \varepsilon$ for $i = 1, \ldots, 9$,
$x_{11+i} \sim s_i \cdot x_{11} + \sigma_2 \varepsilon$ for $i = 1, \ldots, 9$,
$x_{21+i} \sim s_i \cdot x_{21} + \sigma_3 \varepsilon$ for $i = 1, \ldots, 9$.

where $\varepsilon \sim N(0, 1)$. A combination of features $x_1, x_{11}$ and $x_{21}$ separates the toy-patients in two classes (multivariate model) according to the rule:

$$\begin{aligned} \mathrm{P}\left[y = sign(X\beta + \varepsilon)\right] &= p \\ \mathrm{P}\left[y = -sign(X\beta + \varepsilon)\right] &= 1 - p \end{aligned}$$

where $\beta = (1, 1, 1, 0, \ldots, 0)$.

The family of toy problems described above considers three discriminating groups each containing 10 correlated features, while features $x_{31}, \ldots, x_{50}$ are uninformative. By applying our technique we aim at selecting and clustering such relevant blocks. We now examine the performance of our technique on three toy problems which differ in the correlation parameters $\sigma_1, \sigma_2, \sigma_3$, scaling factors $s_i, i = 1 \ldots, 9$ and Bayes risk $p$. We state beforehand that the selection step allows to achieve optimal prediction performance on all toy problems taken under consideration. Being interested in the second step, in the following we analyze in details the results on clustering and visualization.

#### 3.1.1 Toy Problem 1

In this toy problem, genes belonging to the same block, have comparable expressions ($s_i = 1,\ i = 1 \ldots, 9$), and high correlation with either $x_1, x_{11}$ or $x_{21}$ ($\sigma_1 = \sigma_2 = \sigma_3 \sim 0.1$). The two classes are perfectly separated ($p = 0$).

As shown in Figure 1(top), the first stage of the elastic net algorithm selects 4 minimal features instead of 3: $3, 19, 26$ and $24$. Note that the last two features belong to the same block and therefore are highly correlated; however, due to the non-perfect correlation, they are both selected at the first stage. By increasing the $\ell_2$ parameter all features from block 1 and 2 and 6 of the 10 features from block 3 are subsequently selected. Clearly from Figure 1(top), the nesting-clustering technique succeeds in correctly assigning each feature to its corresponding block. According to the data creation rule, Figure 1(bottom) shows that the clusters 1, 2 and 3 are far apart when projected on the 3D meta-patients, whereas block 4 clearly overlaps block 3.

#### 3.1.2 Toy Problem 2

We now add a small amount of noise to the classification problem by raising the error probability, $p$, to $0.01$. The other parameters are left unchanged as in Toy problem 1.

As in the previous case the first stage of the elastic net algorithm selects 4 minimal features instead of 3: $14, 7, 29$ and $23$, where, again features 29 and 23 are correlated. With higher $\epsilon$ almost all the features from the relevant blocks are recovered, and again the nesting-clustering technique succedes in correctly grouping the features. In Figure 2 the same behavior as in Toy Problem 1 is shown .

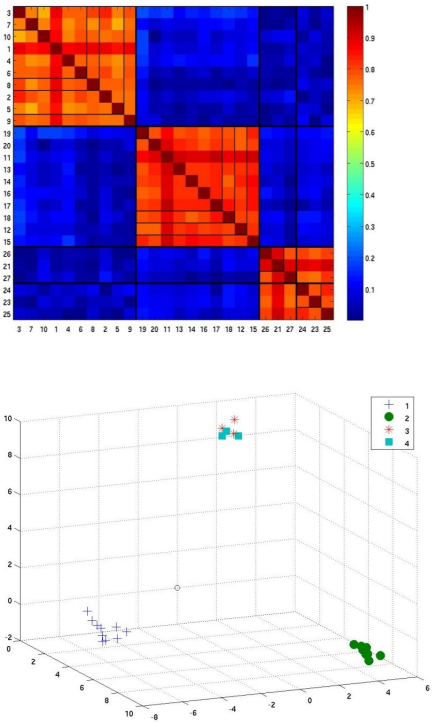

**Figure 1. Correlation and Clustering in Toy Problem 1 .**

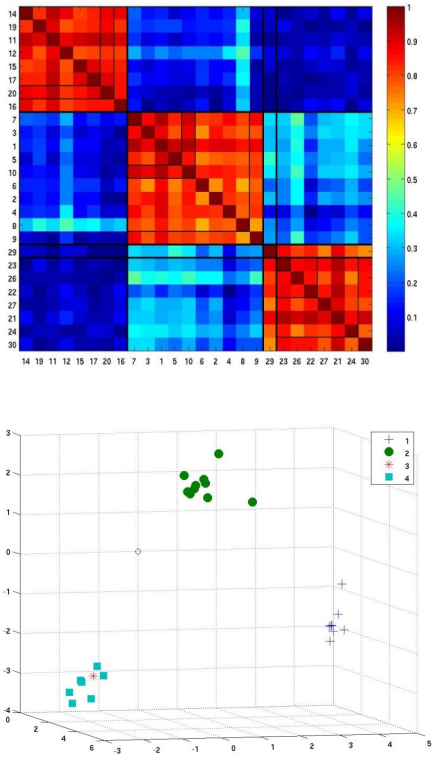

**Figure 2. Correlation and Clustering in Toy Problem 2.**

#### 3.1.3 Toy Problem 3

In the last toy problem, features in the three relevant blocks are rescaled replicates ($\sigma_1 = \sigma_2 = \sigma_3 = 0$) of either $x_1, x_{11}$ or $x_{21}$; this makes the features in the same block exactly correlated however the non-unit scaling factors ($s_i = 1 - i/10,\ i = 1 \ldots, 9$), make their expression values non comparable. As in the first problem, classes are perfectly separated ($\sigma_0 = 0$).

In this case, all the correlated features belonging to the same block bring the same amount of information, being their correlation equal to 1. In fact, the family of solutions

$$\begin{aligned}
\{\sum_{j=1}^{30} x_j \beta_j s.t. \beta_1 + \sum_{j=1}^{9} \beta_{1+j} s_j &= \beta_1^*, \\
\beta_{11} + \sum_{j=1}^{9} \beta_{11+j} s_j &= \beta_2^*, \\
\beta_{21} + \sum_{j=1}^{9} \beta_{21+j} s_j &= \beta_3^*\}
\end{aligned}$$

are exactly equivalent in terms of prediction performance. As a consequence, being $|s_i| < 1$, the $\ell_1$ penalty favours solution $x_1\beta_1^* + x_{11}\beta_2^* + x_{21}\beta_3^*$, as the one having lowest

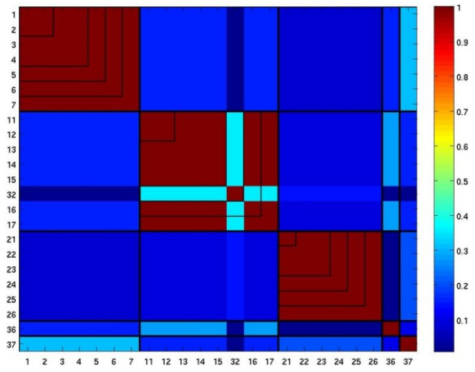

**Figure 3. Correlation on replicated Toy Problem 3**

$\ell_1$ norm. The trade-off between the $\ell_1$ and $\ell_2$ norms, has the effect of recovering more and more features, by increasing the $\ell_2$ parameter. As shown in Figure 3, at stage I the algorithm selects the 3 relevant features with highest expression, 1, 11 and 21, and two noisy features, 36, 37. The sec-

ond stage progressively includes almost all the correlated features according to their scaling factor, and the clustering algorithm correctly groups them.

### 3.2 Real data

We now analyze three well known microarray data sets.

**Prostate** A set of 51 normal prostate and 51 prostate cancer microarrays are represented in this dataset, proposed in [12]. Gene expressions are measured on the Affymetrix platform HU95, for a total of 12533 probe sets.

**Lung** There are 181 tissue samples among which 31 samples belong to Malignant pleural mesothelioma and 150 belong to lung adenocarcinoma. Each sample is described by 12533 genes. The used platform is Affymetrix U95A [6].

**Leukemia** In this experiment we used the well-known Leukemia dataset [5]. It consists of 72 gene expression microarrays of acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), on platform Affymetrix HuGeneFL, comprising the expression of 7129 probe sets.

We observe that the selection-classification step reaches 100% prediction accuracy on the Leukemia task, 99% on Lung cancer, data set and 96% on Prostate Cancer. Note that our performances are at least as good as and often better than those reported in the papers in which the data sets were introduced. As for the toy problems, we visualized our results with the tools described in Subsection 2.4. The well defined color patches in Figure 4, 5, 6 (top) and the clusters plotted in Figure 4, 5, 6 (bottom) clearly indicate that the proposed technique does detect a strong correlation pattern. Moreover in Table 1, 2, 3 we reported the number of genes present at the different $\epsilon$-layers of the top 10 modules.

### 3.3 Conclusion

In this paper we have proposed an effective methodology for identifying structured modules of correlated genes from the nested gene lists returned by the two-stage elastic net selection algorithm. The procedure has been successfully validated on synthetic data, and subsequently applied to benchmark microarray data sets. On toy problems, besides optimal classification performance, we identified the artificially generated patterns with high accuracy. On real data, while the obtained prediction performance is consistent with state-of-the-art results, the relevance of the layered clusters provided by our technique cannot be directly assessed. In fact, in unsupervised cluster analysis, despite a large number of heuristically motivated methods, there are

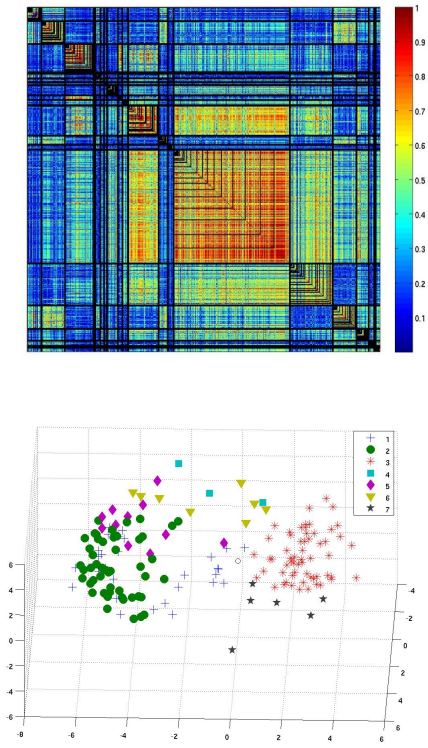

**Figure 4. Correlation and Clustering in Prostate data.**

no theoretically founded approaches capable to assess the goodness of the clusters. In order to accomplish such validation task in a indirect way, we have proposed two visualization tools which qualitatively confirm the methodology effectiveness also in the real data application. Concluding, the methodology we propose can be seen as a step further towards understanding the complex interactions among genes which are at the basis of system biology.

**Acknowledgments.**

This work has been partially supported by the FIRB project RBIN04PARL and by the EU Integrated Project Health-e-Child IST-2004-027749.

## References

[1] A. Barla, S. Mosci, L. Rosasco, and A. Verri. A method for robust variable selection with signicance assessment. In *Proc. of ESANN, European Symposium on Artificial Neural Networks*, 2008.

[2] C. De Mol, E. De Vito, and L. Rosasco. Elastic net regularization in learning theory. Technical report, DISI, 2008.

| $\epsilon$ | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $B_5$ | $B_6$ | $B_7$ | $B_8$ | $B_9$ | $B_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $4 \cdot 10^3$ | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| $3 \cdot 10^2$ | 2 | 4 | 2 | 1 | 2 | 2 | 1 | 3 | 3 | 1 |
| $2 \cdot 10^1$ | 7 | 9 | 5 | 2 | 4 | 2 | 1 | 8 | 5 | 3 |
| 1.7 | 12 | 28 | 21 | 3 | 8 | 4 | 6 | 17 | 5 | 7 |
| 13 | 27 | 40 | 48 | 3 | 10 | 7 | 6 | 21 | 8 | 9 |
| 100 | 31 | 49 | 60 | 3 | 11 | 8 | 6 | 22 | 10 | 11 |

**Table 1. Number of genes in the top 10 blocks, $B_1, \ldots, B_{10}$, at increasing values of $\epsilon$ for Prostate data.**

| $\epsilon$ | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $B_5$ | $B_6$ | $B_7$ | $B_8$ | $B_9$ | $B_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $4 \cdot 10^3$ | 1 | 2 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 1 |
| $3 \cdot 10^2$ | 1 | 11 | 2 | 2 | 2 | 4 | 2 | 1 | 2 | 1 |
| $2 \cdot 10^1$ | 1 | 27 | 4 | 4 | 3 | 6 | 3 | 8 | 5 | 7 |
| 1.7 | 9 | 61 | 15 | 6 | 6 | 13 | 10 | 19 | 29 | 11 |
| 13 | 30 | 108 | 32 | 15 | 12 | 30 | 21 | 34 | 73 | 33 |
| 100 | 59 | 140 | 68 | 24 | 20 | 57 | 33 | 78 | 103 | 53 |

**Table 2. Number of genes in the top 10 blocks, $B_1, \ldots, B_{10}$, at increasing values of $\epsilon$ for Lung data.**

| $\epsilon$ | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $B_5$ | $B_6$ | $B_7$ | $B_8$ | $B_9$ | $B_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $4 \cdot 10^3$ | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 |
| $3 \cdot 10^2$ | 1 | 3 | 4 | 1 | 2 | 2 | 3 | 2 | 1 | 1 |
| $2 \cdot 10^1$ | 2 | 5 | 9 | 1 | 3 | 6 | 4 | 6 | 2 | 1 |
| 1.7 | 3 | 8 | 25 | 1 | 7 | 23 | 11 | 11 | 23 | 2 |
| 13 | 10 | 28 | 101 | 4 | 46 | 77 | 22 | 19 | 71 | 10 |
| 100 | 28 | 36 | 239 | 10 | 91 | 151 | 38 | 25 | 131 | 26 |

**Table 3. Number of genes in the top 10 blocks, $B_1, \ldots, B_{10}$, at increasing values of $\epsilon$ for Leukemia data.**

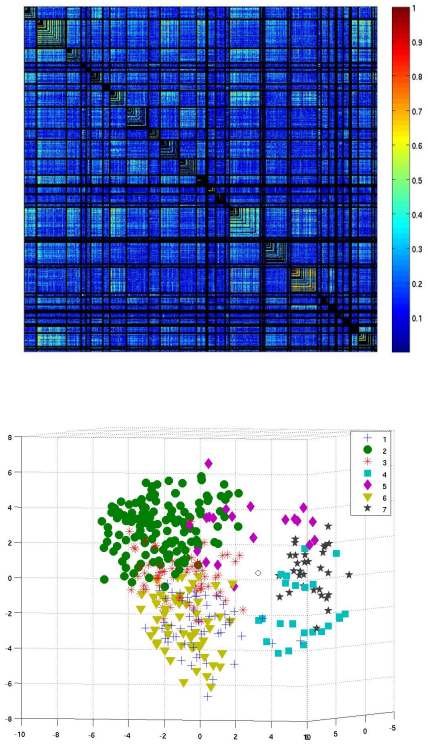

**Figure 5. Correlation and Clustering in Lung data.**

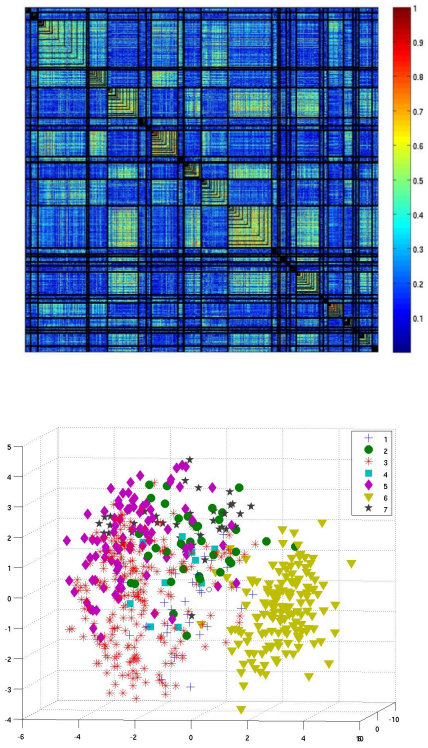

**Figure 6. Correlation and Clustering in Leukemia data.**


[3] C. De Mol, S. Mosci, M. Traskine, and A. Verri. A regularized method for selecting nested groups of relevant genes from microarray data. Technical report, DISI, 2007.

[4] A. Destrero, S. Mosci, C. De Mol, A. Verri, and F. Odone. Feature selection for high dimensional data. *Computational Management Science*, 2008.

[5] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. D. M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[6] R. Gordon, G. J.and Jensen, L.-L. Hsiao, S. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, 62:4963–4967, 2002.

[7] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[8] T. Hastie, J. Taylor, R. Tibshirani, and G. Walther. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29, 2007.

[9] S. Hilsenbeck, W. Friedrichs, P. O'Connell, R. Hansen, C. Osborne, and S. Fuqua. Statistical analysis of array expression data as applied to the problem tamoxifen resistance. *Journal of the National Cancer Institute*, 91:453–549, 1999.

[10] Y. Li, C. Campbell, and M. Tipping. Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, 18:1332–1339, 2002.

[11] S. Raychaudhuri, J. Stuart, and R. Altman. Principal components analysis to summarize microarray experiments: application to sporulation time series, 2000.

[12] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1, 2002.

[13] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 56:267–288, 1996.

[14] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 2005.