A method for robust variable selection with significance assessment

Annalisa Barla[1], Sofia Mosci[1 2], Lorenzo Rosasco[1], Alessandro Verri[1]

1- Universita degli studi di Genova - DISI
via Dodecaneso 35, Genova - Italy

2- Universita degli studi di Genova - DIFI
via Dodecaneso 33, Genova - Italy

Abstract.
Our goal is proposing an unbiased framework for gene expression analysis based on variable selection combined with a significance assessment step. We start by discussing the need of such a framework by illustrating the dramatic effect of a biased approach especially when the sample size is small. Then we describe our analysis protocol, based on two main ingredients. The first is a gene selection core based on *elastic net* regularization where we explicitly take into account regularization parameter tuning. The second is a general architecture to assess the statistical significance of the model via cross validation and permutation testing. Finally we challenge the system on real data experiments, and study its performance when changing variable selection algorithm or the dataset size.

1 Motivation

The ultimate goal of cancer research is the design of effective targeted therapies, which can be achieved only through accurate disease classification and molecular mechanisms understanding. A powerful approach to detect significant molecular alterations is provided by gene expression profiling. In this context a main goal, besides classification, is finding a gene signature, that is a panel of genes able to discriminate between two given classes, e.g. patients and control. Such an analysis encompasses, at least two steps, gene selection and model assessment. When dealing with high-throughput data the choice of a consistent selection algorithm is not sufficient to guarantee good results. It is therefore essential to introduce a robust methodology to select the significant variables not susceptible of selection bias [1] and to use valid statistical indicators to quantify and assess the significance of the results. With a simple example, we point out the risks of selection bias in gene expression analysis. If the gene selection phase is performed out of the validation loop a biased model is generated. This yields a perfect cross validation error on the given data, but does not guarantee generalization properties. As an example we performed the experiments illustrated in Fig.1 (left) on a real dataset[1], in two different set-ups. The curves represent least squares leave-one-out error at increasing number of genes, selected according to their t-score:

[1] see Results section for details on the dataset

in the dashed line case, gene ranking was based on the entire dataset, while in the other case the ranking phase was incorporated in a validation loop. An extreme case is shown in Fig.1 (right) where the same procedure was performed on a completely random dataset (where the values and the labels are randomly sampled from a uniform distribution); the biased method still achieves perfect accuracy, while the unbiased oscillates around chance.

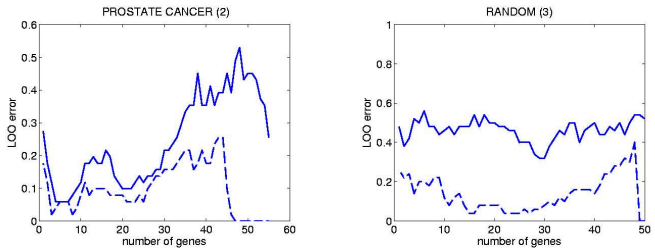


Fig. 1: Biased vs. unbiased schema on real and random dataset

2 Statistical Analysis Framework

We now present an unbiased model selection protocol which is able at the same time to identify the most relevant variables and to achieve a good prediction performance. Our protocol, shown in Fig.2 is based on two main steps: first, we select the relevant features and build the predictive model (internal loop - Sec. 2.1), then we assess its statistical robustness and significance within a complete validation framework (external loop - Sec. 2.2).

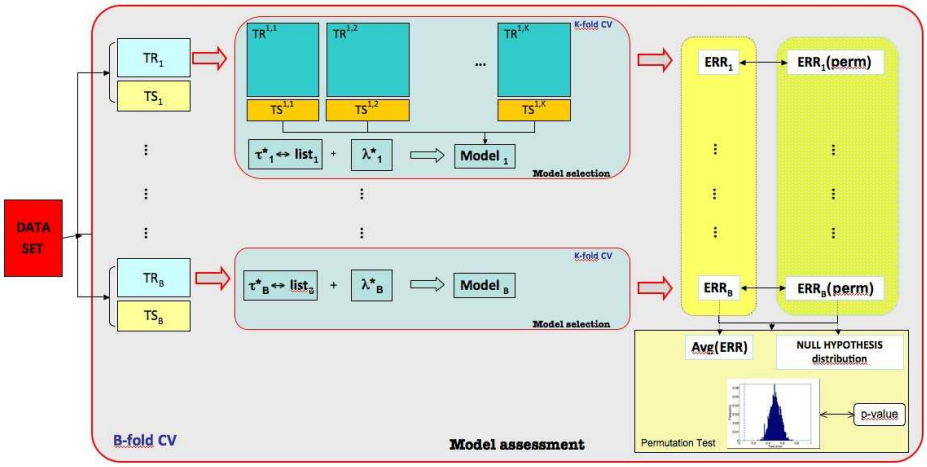


Fig. 2: The structure of a bias-selection aware framework.

2.1 Variable selection and classification algorithm

Before illustrating our algorithm, let us add few considerations on the desirable properties of a variable selection algorithm. A good algorithm should take into account at least linear interaction of multiple genes. Many approaches take into consideration one gene at the time and then rank them according to their fold-change, as in the analysis of differentially expressed genes, or to their prediction power, as with single variable classifiers [2, 3]. However, such methods discard relevant information concerning the combined effect of groups of two, three or even many genes together. Indeed in most cases a multivariate model is preferable. Another drawback of many variable selection algorithms is the rejection of part of the relevant genes due to redundancy. In many biological studies some of the input variables may be highly correlated with each other. As a consequence, when one variable is considered relevant to the problem, its correlated variables should be considered relevant as well. Finally, from the statistical viewpoint a minimal requirement is asymptotic consistency of the algorithm, ensuring that results will improve as the number of training samples increases, and eventually the best possible estimator is reached when enough data is available. Given the above premises we focused on the *elastic net* selection method originally presented by [4] and studied and used in [5, 6].

To describe such method we first fix some notation. Assume we are given a collection of n subjects, each represented by a p-dimensional vector $\mathbf{x}$ of gene expressions. Each sample is associated with a label $y \in \{-1, +1\}$, assigning it to a class (e.g. patient or control). The dataset is therefore represented by a $n \times p$ matrix $X \in \mathbb{R}^{n \times p}$, where $p >> n$ and $Y \in \mathbb{R}^n$ is the labels vector. We consider a linear model $f(\mathbf{x}) = \mathbf{x} \cdot \beta$ and our goal is is to find a sparse approximation $sign(\mathbf{x} \cdot \beta) \sim y$, where sparse means that many of the coefficients in β are exactly zero. Then elastic net regularization amounts to finding

$$\beta_{en} = \mathrm{argmin}_\beta \left[\|Y - X\beta\|_2^2 + \tau \left[\|\beta\|_1 + \epsilon \|\beta\|_2^2 \right] \right],$$

where the least square error is penalized with the ℓ^1 and ℓ^2 norm of the coefficient vector. This approach guarantees consistency of the estimator [5] and enforces the sparsity of the solution by the ℓ^1 term, while preserves correlation among input variables with the ℓ^2 term. Differently to [4] we estimate the minimizer of the functional using the simple *iterative soft thresholding* algorithm proposed in [5, 6]. Once the relevant features are selected we use regularized least squares (RLS) to estimate the classifier[2].

The parameter ϵ in the elastic net is fixed a priori and governs the amount of correlation we wish to take into account. The training for selection and classification requires the choice of the regularization parameters for *elastic net* and RLS denoted in the following with λ^* and τ^*, respectively. This is achieved via a cross validation incorporated in the inner loop of our procedure (see Fig.

[2]We use RLS since it is the same model as in EN, and it has been proved in [7] that RLS is consistent and converges to its Bayes estimator.

2). A similar data splitting technique is present in the outer loop to asses classification performance of the model, so we describe it in the next section.

2.2 Model Assessment

In this section we focus on the external loop, needed to verify the goodness of the estimated model both in terms of performance stability and significance. In order to obtain an unbiased estimate of the classification performance[1], this step must be carefully designed by holding out a blind test set. Since the available samples are very few compared to the number of variables, this step has to be performed on different subsamplings and its results averaged [8]. In our framework, given a dataset, we first split it in B subsplits. As shown in Fig. 2, for each dataset we evaluate the optimal regularization parameters τ_b^* and λ_b^* with $b = 1 \dots B$, by performing the inner loop described above on the training set TR_b and producing B models and lists of relevant genes. For each of the B models we now have a blind test set for model assessment. Prediction performance ERR_b of the b^{th} model can then be evaluated as the classification error on test set TS_b. In this way we can honestly estimate the average error $Avg(ERR)$ of our models.

We can further asses the significance of $Avg(ERR)$ via permutation testing [9]. In a randomized experimental design, objects or individuals are randomly assigned to an experimental group. In the case of binary classification, we randomly permute the labels of the available data so that patients of class A can belong to class B and vice-versa. As we do this, we produce B training sets where we destroyed on purpose the relation between genes and groups of patients. The RLS classifier is trained on the b^{th} bogus training set, restricted to the genes in $list_b$, and the error estimated on the true test set TS_b. As before we average over the B splits and repeat this step 1000 times, hence obtaining the null hypothesis distribution. Now we can say with which confidence the result obtained with the original training set differs from the random distribution. The probability of obtaining a result at least as extreme as $Avg(ERR)$ is referred to as p-value, which is computed as the cumulative sum on the random distribution (non-parametric test). To reject the null hypothesis the p-value must be compared with a fixed significance level, typically set to 5%,1% or 0.1%.

3 Results

We apply our schema to a prostate tumor dataset. We are given a cohort of 102 samples (tumor vs. normal tissue) analyzed by Affymetrix microrarray technology, platform U95Av2 [10]. In the context of the unbiased framework described above with $B = K = 10$, we compare the EN/RLS algorithm with an off-the-shelf approach based on t-score filtering and RLS. The error achieved is 7% in the former case while in the latter is 8%. This is shown in Fig. 3, in comparison with the null distribution provided by the permutation test; in both cases p-value is zero.

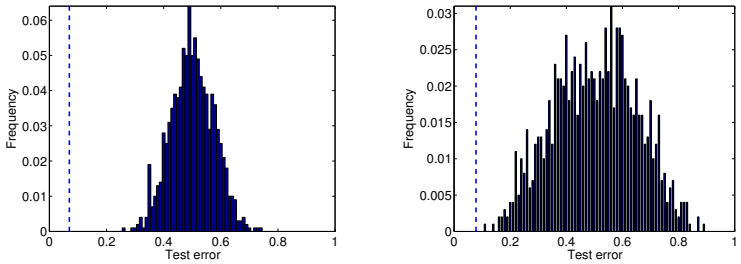


Fig. 3: Distribution obtained by 1000 permutations of training labels (p-value=0) in EN/RLS (left) and t-score/RLS algorithm (right). The dashed line represents the estimated error

In some cases the significance of an experiment is intrinsically weak due to the number of the samples. In fact, if the available data are not enough, the estimated test error could be interestingly low but still falling in the acceptance region of the null hypothesis distribution. To illustrate such a situation, we subsampled the prostate dataset by progressively decreasing the number of samples $n = 20, 30, 40, 50$ and performing the entire experiment for each set of data. As shown in Fig. 4 we observe the result of the permutation test for increasing number of samples. Notice that the distribution is more peaked as the number of data increases while the cross-validation test error (dashed line) decreases. To avoid bias, each population has been subsampled 10 times. For each dataset we repeat the schema and evaluate p-values. Fig. 5 represents the average p-value for the different populations.

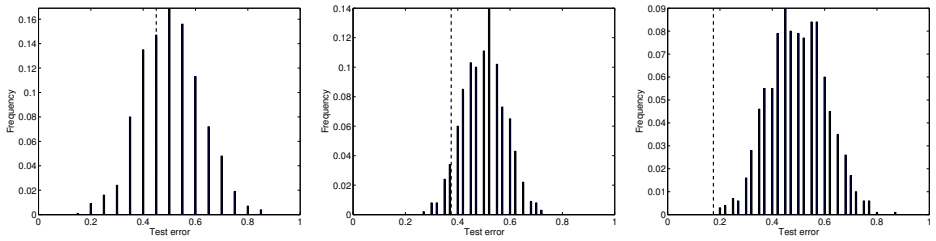
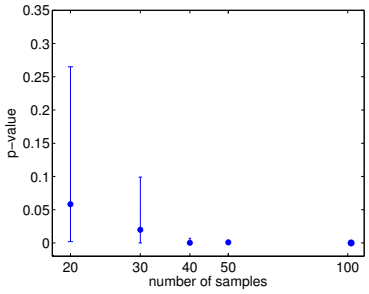


Fig. 4: Null-hypothesis distribution for $n = 20$ (left), $n = 30$ (center) and $n = 40$ (right); the dotted line represents the estimated cross validation test error.

4 Conclusions

In this paper we discuss a possible framework for a statistically significant analysis of microarray data. The core of such a framework is an embedded variable selection method, namely elastic net regularization, allowing for robust gene selection. A correct use of the latter requires at least two separate steps: parameter tuning (model selection) and performance estimates (model assessment). To avoid selection bias these two steps must be clearly distinct and carefully



	n		
α	20	30	40
5%	8/10	8/10	10/10
1%	3/10	8/10	10/10
0.1%	0/10	7/10	7/10

Fig. 5: Left: average p-value with $90-10$ percentile confidence interval. Right: number of datasets for which the null hypothesis is rejected for different significance levels

designed. Here we present a framework based on two nested loops: the internal one is responsible for model selection and is based on a cross validation strategy. The external loop is for model assessment and uses both validation estimates and permutation test. The importance of having an unbiased framework is supported by some simple but still informative experiments and the performance of the proposed procedure is illustrated on a benchmark microarray dataset.

References

[1] C. Ambroise and G.J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA.*, 99(10):6562–6566, 2002.

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

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

[4] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67, Part 2:301–320, 2005.

[5] C. De Mol, E. De Vito, and L. Rosasco. Sparse tikhonov regularization for variable selection and learning. Technical report, DISI, 2007.

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

[7] P.L. Bartlett, M.J. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, U.C. Berkeley, 2003.

[8] C. Furlanello, M. Serafini, S. Merler, and G. Jurman. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics*, 4(54):DOI:10.1186/1471–2105–4–54, 2003.

[9] S. Mukherjee, P. Golland, and D. Panchenko. Permutation test for classification. Technical Report AIM-2003-019, MIT, 2003.

[10] 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.