# Variable Selection for Multifactorial Genomic Data

Sonia Tarazona[a,b], Sonia Prado-López[c], Joaquín Dopazo[a], Alberto Ferrer[b], Ana Conesa[a,*]

[a] *Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain*
[b] *Department of Applied Statistics, Operations Research and Quality, Universidad Politécnica de Valencia, Valencia, Spain*
[c] *Cellular Reprogramming Laboratory, Centro de Investigación Príncipe Felipe, Valencia, Spain*

**Abstract**

Dimension reduction techniques are used to explore genomic data. Due to the large number of variables (genes) included in this kind of studies, variable selection methods are needed to identify the most responsive genes in order to get a better interpretation of the results or to conduct more specific experiments. These methods should be consistent with the amount of signal in the data. For this purpose, we introduce a novel selection strategy called minAS and also adapt other existing strategies, such us Gamma approximation, resampling techniques, etc. All of them are based on studying the distribution of statistics measuring the importance of the variables in the model. These strategies have been applied to the ASCA-genes analysis framework and more generally to dimension reduction techniques as PCA. The performance of the different strategies was evaluated using simulated data. The best performing methods were then applied on an experimental dataset containing the transcriptomic profiles of human embryonic stem cells cultured under different oxygen concentrations. The ability of the methods to extract relevant biological information from the data is discussed.

*Keywords:* Gene Expression, Multifactorial Data, Principal Component Analysis, Variable Selection

*Corresponding author

## 1. Introduction

High-throughput genomic and transcriptomic experiments generate data for a high amount of variables (e.g. genes) on a much lower number of individuals (samples). Common approaches to explore this kind of data are clustering methods such as hierarchical or KNN clustering [1, 2], and dimensionality reduction techniques such as Principal Component Analysis (PCA). PCA is frequently used in transcriptomic data to group samples, identify associated genes or to spot those genes or samples behaving completely different from the rest [3, 4]. In simple case-control studies, the methodology is able to provide biologically interpretable results. However, more complex experimental designs can also be found in transcriptome research, that include factors such as time effect, treatment, tissue, strain, etc., at different levels, giving rise to high-dimensional multifactorial datasets. For these multifactorial experiments, other dimension reduction techniques exist that tackle the analysis of the data in a more efficient way and achieve a better interpretation of the results. Some examples are Tucker3 [5] or PARAFAC [6], which have been successfully applied to the analysis of genomic data [7]. Another interesting approach is ASCA (ANOVA-Simultaneous Component Analysis) [8], adapted to genomic data in the ASCA-genes software [9]. ASCA-genes is a powerful tool to extract targeted signals from noisy data in complex experimental setups using a combination of ANOVA-like data decomposition and PCA.

In many cases, though, descriptive analysis is not the only goal of the experiment, but also the identification of responsive (or activated) genes, since they give the clue to the molecular biology interpretation of transcriptional regulation. When facing the issue of variable selection within the framework of dimension reduction techniques, there exist some rules of thumb such as considering that a variable is important if its loading absolute values are higher than a certain threshold. However, this is a rather arbitrary way of selecting variables. More sophisticated variable selection methods can be found in the literature, especially for PCA. Jolliffe [10, 11] used the absolute value of PCA loadings to measure the contribution of the original variables to the model and selected as many of these variables as the number of selected latent variables in order to retain the maximum variance of the data. McCabe [12] recommended four different criteria to select what he called principal variables and then evaluated all possible subsets of original variables to find the one optimizing the pursued criterion. Krzanowski [13] combined PCA

with Procrustes analysis to select those variables preserving the multivariate data structure, and used a Procrustes criterion to quantify the similarity of compared structures. Since exploring all the subsets of $q$ variables (being $q$ the number of variables to be selected) might be very computationally expensive, he included a backward procedure to discard variables. Guo *et al.* [14] improved the search of the best subset in the latter method by applying a genetic algorithm to avoid exhaustive searching. Westad *et al.* [15] used Student's $t$-tests based on loadings and their estimated standard uncertainties to calculate the significance on each variable for each component. Principal Feature Analysis [16] is based on taking PCA loadings and clustering them using K-Means algorithm. The number of clusters must be equal or greater than the number of PCs. In each cluster, the closest variable to the mean of the cluster is selected (principal feature). Finally, variable selection can be carried out by applying Sparse Principal Component Analysis [17]. Sparse PCA generates linear combinations of the data variables explaining a maximum amount of variance in the data while having only a limited number of nonzero coefficients.

The purpose of most of these methods is reducing the number of variables to achieve a better interpretation of the principal components. Several of them are unfeasible in the context of genomic data due to the large number of variables (genes) or inappropriate due to the low signal to noise ratio that characterizes these data. Another drawback is that the majority of these approaches need to set in advance the number of variables to be selected (or removed), which is generally an unwanted constraint when trying to identify responsive genes. In this work, we compile several selection strategies that avoid this constraint and compare these methods when being applied to the analysis of multifactorial genomic data. This issue had been previously addressed by our group in the adaptation of ASCA [8] to genomic experiments, the ASCA-genes tool [9]. ASCA-genes was shown to be an effective approach for the analysis of complex datasets and the gene selection strategy presented in that work was proven to give good results with signal rich transcriptomic datasets. Here, we extend that study and consider a vast array of signal to noise conditions together with different selection strategies to provide a comprehensive understanding of the behavior of complex transcriptomic designs.

This work proposes two novel approaches, minAS and Gamma approximation, for variable selection in the context of multifactorial gene expression experiments. We use the ASCA-genes framework for treating the multifactorial nature of the data. However, the gene selection strategies we propose

rely on the probability distribution of PCA statistics and can be applied together with other dimension reduction techniques. Both minAS and Gamma methods in combination with ASCA-genes have been implemented in the web suite for Serial Gene Expression Analysis: SEA (http://sea.bioinfo.cipf.es/) [18], which is freely available for the scientific community.

## 2. Methods

Gene expression in high-throughput experiments has been traditionally measured by means of microarrays. New technologies to quantify the gene expression, such as RNA-seq, are now emerging. However, they are still expensive and the analysis of complex experimental setups giving rise to multifactorial genomic data are very rarely addressed with these new sequencing techniques. Therefore, all the methods presented in this work have been validated on microarray data and the simulation studies also mimic the behavior of this kind of data.

Let $\mathbf{X}_0$ be the gene expression matrix, with dimensions $M \times N$, where $N$ is the number of variables (e.g. genes) and $M$ is the number of observations (biological samples). If samples have been taken according to a certain experimental design, including one or more different factors such as treatment, tissue, time, etc. with different levels and different number of replicates in each level, we are dealing with multifactorial datasets. The experimental setup must be taken into account when choosing the appropriate dimension reduction technique in order to better extract the information contained in the data. ASCA model was used because it tackles the problem of complex experimental designs and efficiently separates signal from noise to achieve an optimal interpretation of the results in terms of experimental factors effects [9].

### *2.1. The ASCA-genes framework*

To present the ASCA-genes methodology, we consider the specific case of an experiment with two factors. In the context of genomic experimental designs, one of the factors is usually time (say, for example, factor $a$). The other factor $b$ indicates the experimental group, such as treatment or tissue. ASCA separates the different sources of variation in the data (as ANOVA does) by decomposing the mean centered data matrix $\mathbf{X}$, resulting from subtracting the column means to $\mathbf{X}_0$, into different submatrices as in (1). Each

one of these submatrices contains the estimated effects associated to a determined experimental factor, for example factor $a$, factor $b$, the interaction $ab$ between them or the residuals $abg$ (see [9] for further details). The estimation of these effects depends on the nature of the factors (between or within subjects, random or fixed effects, etc.). In this work, we have considered the most simple case of ANOVA-like decomposition: fixed effect factors between subjects. Independence of measurements holds when expression values along time are independent, as frequently happens in genomics because they correspond to different biological samples.

$$\mathbf{X} = \mathbf{X}_a + \mathbf{X}_b + \mathbf{X}_{ab} + \mathbf{X}_{abg} \tag{1}$$

As the goal in time-course experiments is usually to detect gene expression profile changes between experimental groups (factor $b$), in this study, the interaction effect has been joined to factor $b$ effect and analyzed in one submodel as it is shown in (2). More information on whether or not joining submodels $b$ and $ab$ can be found in [8, 9]:

$$\mathbf{X} = \mathbf{X}_a + \mathbf{X}_{b+ab} + \mathbf{X}_{abg} \tag{2}$$

For the remainder of this work, ASCA submodels in (2) will be named as "submodel $a$" and "submodel $b + ab$", respectively.

PCA is applied to each one of the submatrices (Simultaneous Component Analysis) to reveal major expression patterns associated to the experimental factors and to identify relevant experimental conditions. At this point, dimensionality reduction is undertaken by selecting for each submodel $k_x$ principal components (for $x$=$a, b + ab, abg$). The resulting ASCA-model is given in (3):

$$\mathbf{X} = \mathbf{T}_a\mathbf{P}_a^t + \mathbf{T}_{b+ab}\mathbf{P}_{b+ab}^t + \mathbf{T}_{abg}\mathbf{P}_{abg}^t + \mathbf{E} \tag{3}$$

where, the scores of each submodel are given by the $M \times k_x$ matrices indicated by $\mathbf{T}_a$, $\mathbf{T}_{b+ab}$, $\mathbf{T}_{abg}$, and the submodel loadings are given by the $N \times k_x$ matrices $\mathbf{P}_a$, $\mathbf{P}_{b+ab}$, $\mathbf{P}_{abg}$, where $\mathbf{P}_x^t\mathbf{P}_x$=$\mathbf{I}$ for $x$=$a$, $b + ab$ or $abg$. $\mathbf{E}$ is a matrix in which the residuals of all submodels of ASCA-model are collected: $\mathbf{E} = \mathbf{E}_a + \mathbf{E}_{b+ab} + \mathbf{E}_{abg}$, where $\mathbf{E}_x = \mathbf{X}_x - \mathbf{T}_x\mathbf{P}_x^T$ for $x = a$, $b + ab$ or $abg$. The extension of this model to more than two experimental factors is straightforward.

Once the major variability patterns have been identified and assuming that the model is biologically meaningful, next step is to select genes whose

expression is affected by the experimental factors. When considering the expression of an individual gene, this might follow the general model, change but with a different pattern or simply present a flat profile. Two statistics are proposed to characterize the behavior of genes within each submodel: the leverage and the Squared Prediction Error (SPE).

The leverage measures the importance of a variable (gene) in the PCA model. Leverage values for all the genes in the submodel $x$ can be computed from loadings matrix according to (4) (see [19]):

$$\mathbf{h}_x = diag[\mathbf{P}_x\mathbf{P}_x^t]; \quad x = a, b + ab \tag{4}$$

The SPE associated to a particular gene is a measure of the goodness of fit of the model for that specific gene. Genes not following the general structure of the model will have high SPE. SPE values can be computed from residuals matrix in each submodel ($\mathbf{E}_x = \mathbf{X}_x$ - $\mathbf{T}_x\mathbf{P}_x^t$) according to (5):

$$\mathbf{SPE}_x = diag[\mathbf{E}_x^t\mathbf{E}_x]; \quad x = a, b + ab \tag{5}$$

In general, "interesting" (or regulated) genes will be those showing high leverage (i.e. genes following the major expression trends) or high SPE values (i.e. having odd but distinct behaviors). Genes with low leverage and low SPE will be regarded as not affected by the experiment (for an extensive explanation on the interpretation of leverage and SPE values, we refer the reader to the original ASCA-genes paper from [9]). To decide which genes should be classified as "interesting", a threshold must be established for both leverage and SPE in such a way that those genes presenting SPE or leverage higher than this threshold will be selected. Nueda and co-workers calculated SPE threshold by using Box's approximation [20] for SPE distribution. Leverage threshold was obtained by resampling techniques [21]. However, they observed that these selection strategies presented a good performance when the signal to noise ratio in the dataset was high, but they were not so effective for data with low signal to noise ratio. Hence, in this work, we introduce other selection methods and compare them with the ones in ASCA-genes under a much wider variety of biological scenarios. Both simulated and real datasets will be used to evaluate the performance of the proposed selection methods. All of these methods have been implemented in the statistical language R and are available from the authors.

*2.2. Variable selection strategies*

Once the dimension reduction model has been established, the goal is often finding the variables with higher contribution in the model. In our case, the most "regulated" genes. The variable selection strategies we propose in this paper always consist of three steps: first, choosing an appropriate statistic to measure the importance of the variables in the model (leverage and SPE in this study); second, estimating the probability distribution of this statistic (in a parametric or non-parametric way) and, finally, establishing the threshold to separate "interesting" from "uninteresting" variables (genes). As in ASCA-genes, our proposals are focused on studying the univariate distribution of both SPE and leverage statistics, although most of the methods we present are valid for other statistics or even other multivariate methods.

It should be noted that SPE and leverage statistics can be computed for each gene in each of the different ASCA submodels $a$, $b + ab$ and $abg$. Gene selection is therefore possible for each of these submodels independently. In this work we have chosen to evaluate the gene selection coming from both $a$ and $b+ab$ submodels as these capture the gene expression changes of interest in the proposed scenario, namely, the time associated changes (submodel $a$) and the time-experimental factor interaction (submodel $b + ab$). Depending on the aim of the experiment, all or only specific submodels might be relevant for the study, and selection will have to be based on the SPE and/or leverage statistics of the corresponding submodels. Thus, interpretation of the gene selection has always to be done on the light of the ASCA submodels considered.

Generally and because of the nature of expression data, most genes present low SPE or low leverage values. Hence, it is expected that these statistics follow a mixture distribution of, at least, two populations. The biggest population is that of "uninteresting" genes (with statistic values closer to zero). The other(s) population(s) corresponds to "interesting" genes (those with higher values in the statistic). As our aim is to separate "interesting" from "uninteresting" genes, the mixture model can be written as in (6):

$$f(x) = p_0 f_0(x) + p_1 f_1(x) \tag{6}$$

where, $x$ is the value of either SPE or leverage for a particular gene, $p_0$ is the proportion of "uninteresting" genes (a priori unknown), $f_0(x)$ is the null probability density function (i.e. probability density function for

"uninteresting" genes), and $p_1$ and $f_1(x)$ are, respectively, the proportion of "interesting" genes and their probability density function.

Two different approaches can be used to establish the threshold for SPE or leverage values. The first one consists of estimating the "uninteresting" genes distribution (null distribution) and using a percentile of this estimated distribution as the threshold. The methods compared in this work that follow this first approach are: Box's method [20], Jackson & Mudholkar's method [22], Gamma method and resampling techniques [21]. In the first three, the null distribution is estimated in a parametric way, while resampling is considered a non-parametric technique. In the second approach, an approximation is obtained for the mixture distribution and the threshold is taken as the value which best separates the two components of the mixture. Many authors have focused on the parametric estimation of the mixture components distribution (see, for example, Efron's work at [23] or [24]). But we observed that, due to the huge difference between the sizes of both populations, it was very difficult to estimate parametrically the probability distribution of each component. Therefore, only a non-parametric approach is introduced here, which is called minAS (MINimum Algorithmic Selection).

**Box's method.** Assuming that errors from a PCA model follow approximately a multivariate normal distribution and given that SPE is a quadratic form of the error associated with a particular variable, Box [20] showed that SPE distribution could be estimated by a weighted $\chi^2$-distribution ($g\chi^2_h$). In ASCA-genes [9], this distribution was used to calculate the (1-$\alpha$)% confidence SPE threshold for each PCA submodel. Parameters $g$ and $h$ are estimated by the matching moments method and the following expression is obtained for SPE threshold at $\alpha$ level of significance, where $m$ is the sample mean and $v$ is the sample variance:

$$SPE_\alpha = \frac{v}{2m}\chi^2_{\frac{2m^2}{v}}(\alpha) \tag{7}$$

**Jackson & Mudholkar's method.** Jackson and Mudholkar [22] found another approximation for SPE distribution in PCA models, by using the residuals matrix (**E**). Then, for PCA coming from each ASCA submodel, SPE threshold at $\alpha$ level of significance can be computed as follows:

$$SPE_\alpha = \theta_1[1 - \frac{\theta_2 h(1-h)}{\theta_1^2} + \frac{z_\alpha(2\theta_2 h^2)^{1/2}}{\theta_1}]^{1/h} \tag{8}$$

where $\mathbf{V}=\frac{\mathbf{E'E}}{N-1}$, being $N$ the number of variables (genes) in the model;

$\theta_i$=trace($\mathbf{V}^i$), for $i$=1,2,3; and $h$=1-$\frac{2\theta_1\theta_3}{3\theta_2^2}$.

**Gamma method.** We propose using the gamma distribution to approximate SPE or leverage null probability density functions because it has more flexibility than the distributions described above to suit different density curves. Given the statistic values for each gene in each submodel, shape and scale parameters for the gamma distribution have been estimated by maximum likelihood [25]. The corresponding thresholds for either SPE or leverage are then the percentile (1-$\alpha$)% of the estimated gamma distribution.

**Resampling techniques.** Resampling methods are non-parametric procedures to determine the statistical significance of a result, sampling repeatedly within the same data. An empirical distribution is generated for an statistic under the null hypothesis by taking the original data, randomly shuffling them numerous times and computing the statistic value for each of the permuted datasets. The way of permuting the data depends on the null hypothesis to be tested [21].

In ASCA-genes [9], a permutation method was used to define the threshold of leverage. In the present work, we study the performance of permutation techniques to obtain the confidence thresholds not only for leverage but also for SPE. We also compare their permutation strategy with our proposal. Both strategies are described below.

Given the $M \times N$ data matrix $\mathbf{X}$, the two permutation strategies are:

Strategy 1.- As implemented in ASCA-genes, $K$ row permutations of matrix $\mathbf{X}$ are generated, destroying the structure of the experimental design. In this case, the null hypothesis to be tested is that experimental conditions do not affect gene expression, i.e. all genes have a flat profile across conditions.

Strategy 2.- The null hypothesis to test in this strategy is that all genes are equally responsive. If this is true, all the genes would have the same contribution in the PCA model and the residual errors would be also similar. Hence, the novel permutation strategy we propose in this work consists of performing $K$ column permutations. Moreover, the permutation of values in the columns is different for each row so that the structure in the data (associations among genes, and among genes and experimental conditions) is totally broken.

In this work, the number of permutations $K$ was set to 1000. Once the permuted matrices have been generated, an ASCA model is fitted to each

Table 1: Methods to calculate SPE or leverage threshold by resampling techniques

| *Method* | *Permutation strategy* | *Threshold computation* |
|---|---|---|
| 1 | 1 - Permuting conditions | Option (a) - For each gene |
| 2 | 2 - Permuting genes | Option (a) - For each gene |
| 3 | 2 - Permuting genes | Option (b) - Globally |

one of them. SPE and leverage values are then obtained for each gene in each permutation to generate the reference distribution. The threshold can be calculated from this reference distribution in two ways:
Option (a).- First, the (1-$\alpha$)% percentile of the $K$ statistic values for each gene is computed and the threshold is obtained as the (1-$\alpha$)% percentile of the $N$ gene percentiles. This is the option implemented in ASCA-genes.
Option (b).- We propose using the (1-$\alpha$)% percentiles of the $K$x$N$ statistic values obtained from the $K$ permutations and $N$ genes.
The three resampling methods to be compared in this work are combinations of permutation strategies 1 and 2 and options (a) and (b) to compute thresholds. They are described in Table 1.

**minAS.** We introduce in this work the minAS method. This algorithmic approach consists of estimating empirically the mixture density function for either the SPE or the leverage and then computing the first local minimum closest to the "uninteresting" genes probability density curve. The SPE or leverage value in which this minimum is reached is taken as the threshold that separates both distributions. The minAS strategy assumes that the mixture distribution in (6) for SPE or leverage is, at least, bimodal. The intrinsic nature of genomic data makes this assumption hold in general. However, it is not always possible to visualize this bimodality in histograms, due to the large difference between sizes of both populations.

To estimate the mixture density curve, a kernel density estimator (KDE) [26] was used (provided by *density* function from the R library *stats*). A KDE is a sophisticated version of histograms that produces smoothed density curves. The KDE depends on a smoothing parameter called *bandwidth*, which is equivalent to the bin width of histograms. To estimate the curve in a given point, all the observations are weighted by a continuous function (kernel) instead of considering only the observations falling into the bin as histograms do. The default option for kernel in the R "density" function is the Gaussian distribution. It is well known that KDE goodness of fit relies

more on bandwidth than in the kernel choice. There are different rules of thumb to compute the optimum bandwidth. For instance, [27] takes into account the dispersion in the data and the sample size to compute bandwidth for KDE with Gaussian kernel. It is implemented as the default option in *density* function ("nrd0").

The minAS algorithm allows users to choose the kernel and the method to calculate bandwidth (offering the same options than *density* function), as well as the number of points for which the density is fitted. The smoothing of the KDE is determined by the bandwidth computed by the chosen method. To increase or decrease this smoothing, the value of the coefficient *adjust* (which defaults to 1) can be increased or decreased, respectively. In the appendix, we provide an study of the influence of the parameter *adjust* in the performance of the method. If several kernel functions or methods to calculate the bandwidth are chosen, minAS selects the mixture estimation that best fits the data according to one of the two implemented options: "max" and "mean". As the true density function is unknown, cumulative distribution functions computed from the KDE are compared with the empirical cumulative distribution function derived from SPE or leverage values. In order to compare them, the difference between the empirical distribution and the KDE cumulative distribution is computed for each value. Then, in the case of "max" option, the maximum of these differences (Kolmogorov-Smirnov distance) is taken. For "mean" option, the mean of all these differences is obtained. The KDE with the smallest maximum (or mean) difference is selected.

Once the best KDE has been obtained, minAS computes the minima of this curve. By default, the first local minimum after the highest peak is taken as the cutoff value to separate the two populations, i.e. "interesting" from "uninteresting" genes. However, minAS users can also set the maximum number of minima to be computed, calculate all of them or provide the interval where the minimum has to be found. A plot is provided in which all the computed minima are represented over the mixture distribution. Then, if more than one minimum is found, users may decide to reduce the number of selected genes by choosing a more restrictive threshold.

## 3. Results

The variable selection methods described above were firstly evaluated on simulated data along several comparative studies. According to the results of these comparisons, the best strategies were determined and applied on

an experimental dataset. Additionally, performance was evaluated using the biological information associated with the selected genes.

### *3.1. Simulated data*

Simulation studies have been conducted, on the one hand, to compare the performance of the proposed variable selection methods and, on the other hand, to see which methods are preferred under certain biological scenarios or which ones are less affected by the biological characteristics of the data. In order to measure the performance of the variable selection methods on the simulated data, we have chosen the Matthews Correlation Coefficient [28], which is considered a good performance measure since it can be used even if the sizes of the sets are very different- as it happens in genomic contexts- and it takes into account all types of classification errors.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FP)}} \tag{9}$$

where $TP$ is the number of True Positives, $TN$ the number of True Negatives, $FP$ the number of False Positives and $FN$ the number of False Negatives. Unless otherwise stated, MCC is calculated from the selection made using both SPE and leverage values.

To simplify the interpretation of results, for each simulated dataset, only two factors (e.g. time and experimental group) have been considered: the time factor consisted of three time points and the number of experimental groups was also three. Four replicates have been generated for each experimental group at each time point. The description of the simulation algorithm can be found in the Supplementary Material. Two different simulation experiments were conducted. The first experiment was used to compare different options in each selection method and determine a good range for parameter values. Next, a global comparison of the best methods combinations was carried out on the second simulation experiment to obtain a more precise benchmarking of the selection approaches.

**Simulation experiment 1.** The biological scenarios to be simulated for this first experiment were defined by the values of the following parameters:

- Number of genes in the dataset ($N$): 3000, 15000 or 30000.
- Percentage of differentially expressed genes (responsive or signal genes) with regard to the total number of genes ($\%deg$): 1%, 5% or 15%.

- Number of gene classes (*class*): 5, 10 or 25. Genes in the same class have the same expression time pattern under the same experimental group.
- Level of noise in the data (*noise*): 10% or 30%.

These parameters define 54 different biological scenarios and 10 datasets were generated for each one of them.

As already mentioned in Section 2.1, variable selection strategies implemented in ASCA-genes were Box's method for SPE and resampling techniques for leverage. However, these approaches were not efficient in separating "interesting" from "uninteresting" genes in large datasets scenarios (unpublished results). So in this work, a complete study was designed to determine under which biological scenarios these selection strategies failed and to compare the three different resampling options to calculate leverage threshold (see Table 1) at significance levels 0.01 and 0.05. The Box's method was maintained to compute SPE threshold, as in the ASCA-genes paper.

Hence, the ASCA model was obtained for each of the 540 simulated datasets and these variable selection strategies were applied. The Matthews Correlation Coefficient (MCC) was obtained in each case and the results were analyzed by means of an ANOVA model with repeated measures [29] to evaluate the effect of the biological factors indicated above, the resampling strategy ("leverage method") and the significance level ($\alpha$) on MCC values. An ANOVA with repeated measures was used because the variable selection methods were applied to SPE and leverage values obtained from the same simulated datasets, so the measurements were not independent in this sense. The ANOVA results indicated that factors with a significant effect on MCC ($p$-value$<0.002$) were: leverage method, significance level, number of signal gene classes (*class*) and percentage of signal genes (%*deg*). The noise level and the number of genes had no statistically significant influence on MCC ($p$-value$>0.6$). Post-hoc tests showed that the best MCC results ($p$-value$<0.001$) were obtained for leverage method 3, i.e. permuting genes and computing threshold as a global percentile; $\alpha$=0.01; low number of signal genes classes and medium signal genes percentage (5%). Further details on this analysis may be found in Supplementary Material. We also observed that for $\alpha$=0.01, the real False Positive Rate (FPR) obtained with any of the resampling methods was similar to the significance level, but when setting $\alpha$ to 0.05, FPR reached 80% in some cases. Classification failures were mainly due to the strategy used to calculate SPE threshold (Box's method).

In the second study on these simulated datasets, Box's method was compared to the other SPE parametric methods: Jackson & Mudholkar's and Gamma. In this case, twelve different significance levels were evaluated, varying from 0.001 to 0.1. No leverage thresholds were calculated, so gene selection was based only on SPE values. Consequently, MCC results can be used to compare SPE methods, but not as a measure of the global performance of the methods. It can be observed in Fig. 1 that when significance level is around 0.03, all the three methods present a similar performance. For the rest of significance levels, Box's method produces much worse results than the other two, which behave similarly.

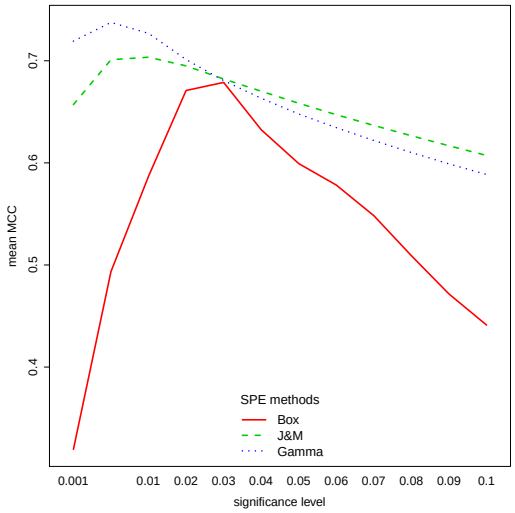


Figure 1: SPE selection methods performance (measured by MCC) according to significance level.

An ANOVA model with repeated measures showed that SPE method, significance level and all the biological factors had a statistically significant effect on MCC ($p$-value$<$0.001), except the number of genes ($p$-value$>$0.7). From post-hoc tests, it was deduced that SPE methods were significantly different ($p$-value$<$0.008), being Jackson & Mudholkar and Gamma the ones presenting better results. For significance levels between 1% and 3% the best MCC results were obtained ($p$-value$<$0.001). No statistically significant differences were observed between 5 or 10 signal genes classes ($p$-value$>$0.3), but significantly better results were obtained when number of classes was 25 ($p$-value$<$0.001), maybe because when so many different patterns are present

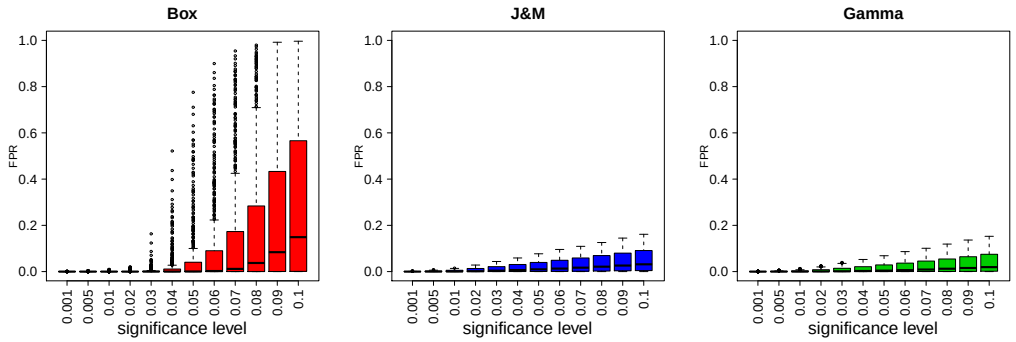


Figure 2: FPR according to significance level for each one of the SPE methods studied.

in the data, there are more genes badly explained by the model and hence those genes have a high SPE value. The best MCC results were obtained when responsive genes percentage was 5%, followed by 15% and lastly 1% ($p$-value$<$0.001). Finally, MCC was higher when noise level was 30% ($p$-value$<$0.001).

Joining the results of this study, we determined the most convenient significance levels for each method to obtain the best MCC value whichever the signal genes percentage and the number of signal gene classes were. The recommendations were $\alpha$=0.03 for Box's method, 0.005$<\alpha<$0.02 for Jackson & Mudholkar's method and $\alpha$=0.01 for Gamma approximation (see Fig. 1).

Again, the significance level was compared to the False Positive Rate (FPR) obtained for each method. Fig. 2 presents these results and shows that in Gamma and Jackson & Mudholkar's methods this relation was preserved, while this did not happen for Box's method.

Once the methods estimating the null distribution were compared, we included minAS method in the study (always taking the first local minimum after the highest peak as the threshold for both SPE and leverage). To see if minAS selection was satisfactory enough to continue studying the method in depth, it was compared to the combinations of methods evaluated in the first study (Box's method for SPE and resampling techniques for leverage). In this preliminary comparison, default options in R "density" function (Gaussian kernel and "nrd0" method to compute bandwidth) were used. As it can be seen in Fig. 3, MCC obtained from minAS was, in general, higher than MCC obtained with the other methods.

In addition, using the same simulated datasets, the default options in minAS (Gaussian kernel and "nrd0" bandwidth computing method) were

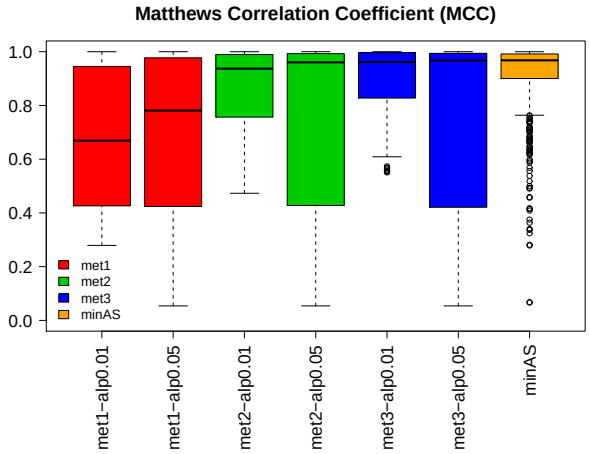


Figure 3: MCC obtained by applying the three resampling methods in Table 1 for $\alpha$=0.01 and $\alpha$=0.05 and minAS method.

compared to the best estimators according to minAS options "max" and "mean" (see 2.2). Fig. 4 shows that minAS resulted in better MCC scores when using the default KDE than with the KDE producing the minimum maximum or minimum mean distance to the empirical data distribution. The reason is that the other kernels or methods to compute bandwidth tended to generate infra-smoothed curves with too many local minima. In these cases, the selection by the first local minimum increased the number of false positives. Therefore, default "density" options were used when applying the minAS procedure hereinafter.

The influence of biological parameters defining the scenarios on MCC results for minAS method was also analyzed using an ANOVA model. All the parameters had a statistically significant effect on MCC ($p$-value<0.01), especially the number of genes, the number of classes and the signal genes percentage, as well as the interactions between them. It was observed that the greater the number of classes and the percentage of signal genes, the better MCC results minAS produced, no matter the number of genes. As number of genes and signal percentage increased, MCC was less dependent on the number of classes. Boxplots describing these results can be seen in Supplementary Material. Hence, as general guidelines, we recommend using minAS for datasets with a high number of variables because otherwise the goodness of fit of KDE is not guaranteed and the multimodality is more

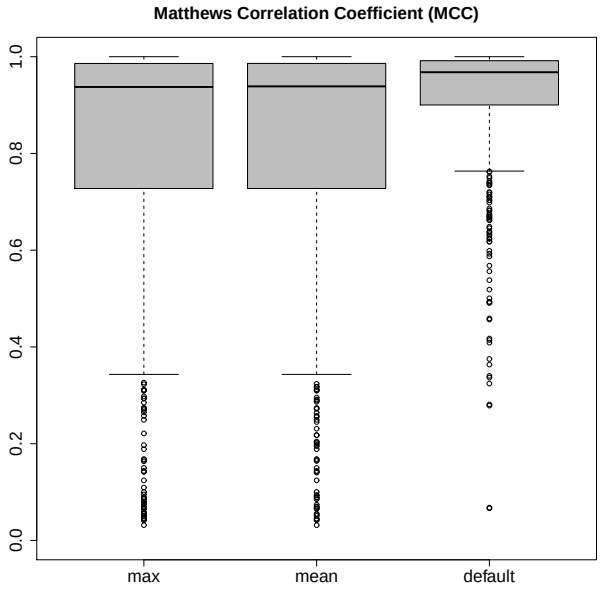


Figure 4: MCC obtained by applying minAS for both SPE and leverage on datasets from simulation experiment 1, considering three options: "max" criterion, "mean" criterion and default options.

dependent on the value of the smoothing parameter. The method can be applied to datasets with a thousand variables approximately, but results show that the best performance is obtained for more than 15000 variables. A Gaussian kernel and the method "nrd0" to compute bandwidth have been proved to offer the best minAS performance. Furthermore, increasing the parameter "adjust" to get a more smoothed KDE produces even better results (see Supplementary Material), although this parameter was not changed in any of the simulation experiments we performed.

**Simulation experiment 2.** To conclude the evaluation of variable selection methods on simulated data, a new simulation experiment was designed in order to compare simultaneously all of the previously described methods for computing SPE and leverage thresholds. In this last comparison, other biological scenarios were simulated taking into account the results obtained in the previous studies. The level of noise was not included as a parameter in these simulations because it had, in general, very little influence on MCC results, so it was set to 20%. The values for the rest of biological parameters were:

Table 2: Selection methods combinations included in global comparison.

| *Combination* | *SPE method* | *Leverage method* |
|---|---|---|
| 1 | Box - $\alpha$=0.03 | Permut2b - $\alpha$=0.01 |
| 2 | Box - $\alpha$=0.03 | minAS |
| 3 | Box - $\alpha$=0.03 | Gamma - $\alpha$=0.01 |
| 4 | J&M - $\alpha$=0.01 | Permut2b - $\alpha$=0.01 |
| 5 | J&M - $\alpha$=0.01 | minAS |
| 6 | J&M - $\alpha$=0.01 | Gamma - $\alpha$=0.01 |
| 7 | Gamma - $\alpha$=0.01 | Permut2b - $\alpha$=0.01 |
| 8 | Gamma - $\alpha$=0.01 | minAS |
| 9 | Gamma - $\alpha$=0.01 | Gamma - $\alpha$=0.01 |
| 10 | minAS | Permut2b - $\alpha$=0.01 |
| 11 | minAS | minAS |
| 12 | minAS | Gamma - $\alpha$=0.01 |
| 13 | Permut2b - $\alpha$=0.01 | Permut2b - $\alpha$=0.01 |
| 14 | Permut2b - $\alpha$=0.01 | minAS |
| 15 | Permut2b - $\alpha$=0.01 | Gamma - $\alpha$=0.01 |

- Number of genes in the dataset: 5000 or 20000.
- Percentage of responsive genes: 3% or 10%.
- Number of gene classes: 5 or 25.

For each one of the 8 possible scenarios, again 10 datasets were generated. SPE selection methodologies to be compared in this analysis were Box's method, Jackson & Mudholkar's (J&M), Gamma, minAS and resampling using permutation strategy 2 (genes permutation) and option (b) to compute threshold by global percentile (Permut2b). Regarding to leverage, we compared resampling method (Permut2b), Gamma approximation and minAS method. The resulting combinations of all these methods are shown in Table 2. The significance level that produced the best results in the previous studies was chosen.

Fig. 5 shows 95% confidence intervals for mean MCC produced by each one of the methods. It can be deduced from this plot the overall good performance of the methods, since all of them got a mean MCC higher than 0.9. However, the ANOVA model with repeated measures showed a statistical significant difference among them ($p$-value<0.001). The worst results were

obtained for those combinations in which resampling techniques were used to compute SPE threshold. Box's method for SPE is not recommended for its high standard deviation. The Gamma approximation for leverage worked excellently. Considering both MCC mean and standard deviation, the best combinations were number 6 (J&M+Gamma), number 8 (Gamma+minAS) and number 9 (Gamma+Gamma). The ANOVA model also showed that the number of genes and the number of signal genes classes had no significant effect on mean MCC value ($p$-value=0.137 and $p$-value=0.353, respectively). However, signal genes percentage significantly affected MCC value ($p$-value<0.001), as well as the interaction between signal genes percentage and method combination ($p$-value<0.02). In general, the higher signal genes percentage, the higher mean MCC. Combination 9 (Gamma+Gamma) did not result in big differences in mean MCC for the different percentages of signal genes. However, some combinations including minAS, for example numbers 10, 11 and 12, worked much better when the percentage of signal genes was higher. Additionally, interaction plots illustrating the methods combinations performance can be found in the Supplementary Material.

In all the simulation studies, the bandwidth was computed following the Silverman's rule ("nrd0" option). To check what happened if bandwidth was modified with the "adjust" coefficient, minAS was applied to the 80 simulated datasets in simulation experiment 2, using the default options in "density" and varying the coefficient "adjust" from 0.5 (i.e., half the bandwidth obtained by "nrd0" method) to 5 (i.e., 5 times the bandwidth obtained by "nrd0" method), as it is described in Supplementary Material. Interestingly, minAS performance improves for "adjust" values higher than one, that is, when the estimated density curve is more smoothed (see Supplementary Material).

To summarize, minAS and Gamma approximation (with $\alpha$=0.01) behaved slightly better than the rest of the studied methods. Furthermore, Gamma method presented less differences in MCC value for different signal genes percentages, while minAS had a better performance when this percentage was higher.

### *3.2. Experimental data: Hypoxia*

Once the benchmarking with simulated data was completed, the methods producing the best results were applied on an experimental dataset and evaluated for their ability to select genes that led to outstanding biological information. The Hypoxia gene expression data in [30] was used for this

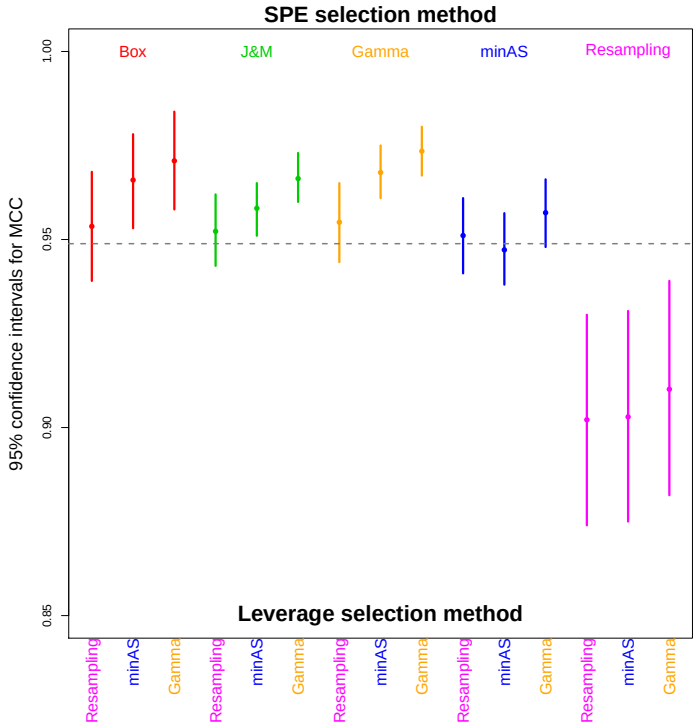


Figure 5: 95% confidence intervals for mean MCC according to methods combination. Horizontal dashed-line corresponds to overall average MCC.

biological validation. This dataset collects the transcriptomic profile of human embryonic stem cells cultured under different oxygen concentrations. The oxygen conditions were: normoxia (21% of oxygen) and hypoxia (5% or 1% of oxygen). Gene expression for 30826 genes was measured in several time points using Agilent microarrays. An ASCA model was fit to the data. Factor $a$ is the time (0 hours, 12 hours, 24 hours, 5 days and 10 days) and factor $b$ was used for the oxygen level (21%, 5% and 1%). Oxygen level and interaction effects were joined together in the model (as in Equation 2). Two principal components were selected in each submodel ($a$ and $b$+$ab$), which explained 83.2% of the variability in submodel $a$ and 71.9% in submodel $b$+$ab$. Model analysis showed different gene behaviors for each oxygen level, differentiating clearly normoxia from hypoxia conditions, and time points 12-24 hours from 5-10 days (results not shown). In order to compute SPE and leverage thresholds, several combinations of selection methods showing the best performance in the previous simulation studies were used: Jackson &

Table 3: Number of genes selected by the studied combinations of methods in hypoxia dataset.

| *Combination* | *SPE method* | *Leverage method* | *Sub-model a* | *Sub-model b+ab* | *Total* |
|---|---|---|---|---|---|
| 5 | J&M | minAS | 1347 | 1827 | 2668 |
| 6 | J&M | Gamma | 1182 | 1919 | 2618 |
| 8 | Gamma | minAS | 1287 | 1076 | 2034 |
| 9 | Gamma | Gamma | 1122 | 1176 | 1976 |
| 11 | minAS | minAS | 1862 | 1309 | 2705 |
| 12 | minAS | Gamma | 1706 | 1405 | 2649 |

Mudholkar's SPE method ($\alpha$=0.01), Gamma approximation ($\alpha$=0.01) and minAS method. Table 3 shows the number of genes selected by each one of these combinations and Fig. 6 shows the histograms and distributions fitted for SPE and leverage values in each submodel, as well as the thresholds obtained by the selection methods.

To evaluate the validity of the different variable selection methods, selected genes lists were investigated to see whether the biological information they contained was relevant for the study. Hence, for each one of the selected genes sets, we carried out a functional enrichment (FE) analysis by means of FatiGO tool, included in Babelomics suite [31], using Gene Ontology (GO) gene function annotation to compare selected versus non-selected genes. FE is a established methodology to interpret and evaluate transcriptomic data, that assesses whether specific cellular functions (in this case, GO terms) are overrepresented within the set of significant genes . Significant enriched GO terms for the selected genes sets were visualized with the Blast2GO software [32], that allowed to color them depending on the number of selection methods by which they had been detected. This kind of graph enabled us to evaluate which of the selected genes sets contributed more to the biological interpretation of the experimental results (see graphs in Supplementary Material).

In general, all tested methodologies generated gene selections enriched in a number of GO terms that represent key general processes of the hypoxia treatment. These were, among others, "developmental process", "metabolic process", "response to stimulus", "transcription factor activity", "chemokine

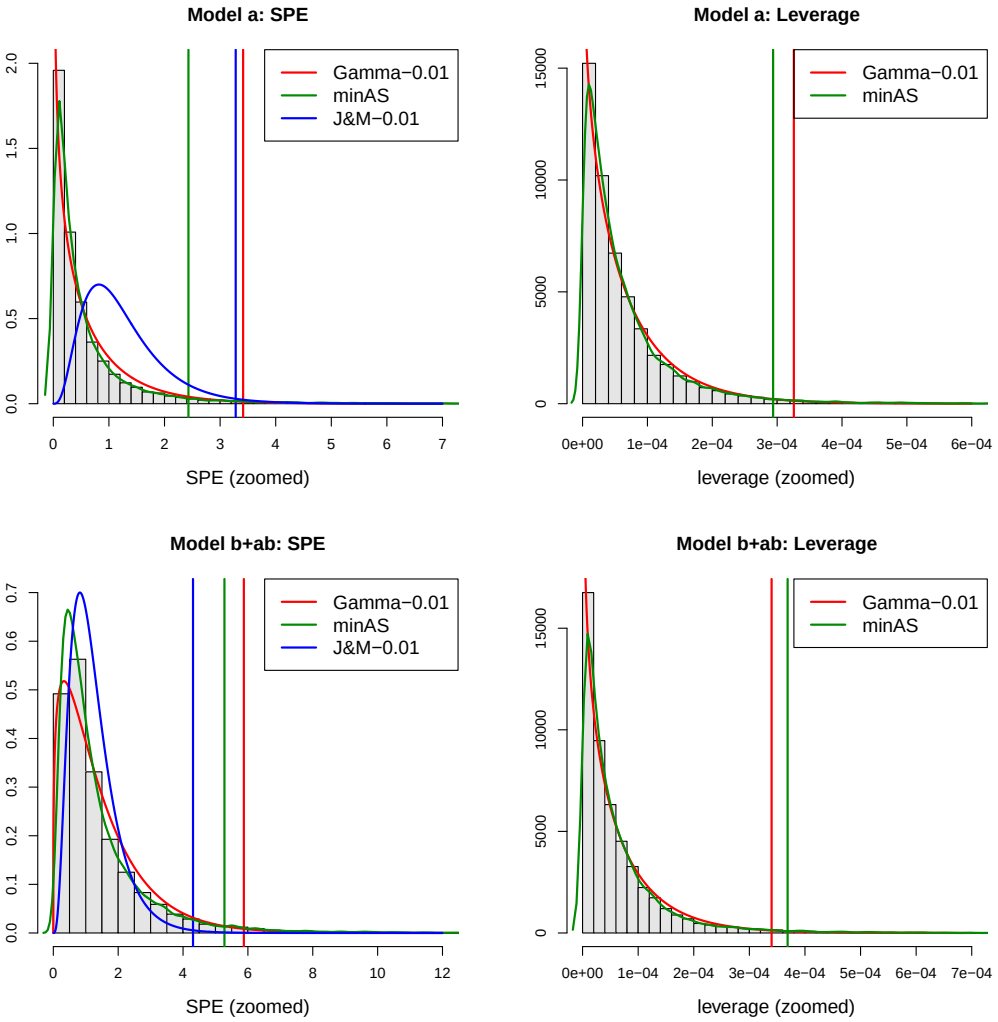


Figure 6: Histograms for SPE and leverage in each submodel. Curves represent the distributions fitted by variable selection methods applied. Vertical lines are the thresholds computed from these distributions. The X-axis have been zoomed for better visualization and therefore they do not show the full range of SPE and leverage values.

receptor binding", "lipid transport activity", "immune system process", "intrinsic to plasma membrane", "organ morphogenesis", "angiogenesis", "response to wounding" and "humoral immune response". "Organ morphogenesis" and "angiogenesis" refer to the establishment of the circulatory system in mammals, one of the first events during the embryo development [33, 34]; while the metabolism of lipids has also been postulated to play an important role in the embryo differentiation [35]. Also "ectoderm development" and "epidermis development" were functions identified by most of the methods, and are directly related to the differentiation process analyzed in this experiment. Additionally, some specific processes were only revealed by some of the selection methods. For example, combinations 5 and 6 (both using J&M method for SPE selection) highlighted the "central nervous system development" (associated to normoxia) and "sensory organ development" or "chemokine receptor binding" (both related to hypoxia). Combinations 11 and 12 (SPE selection by minAS) discovered metabolic processes such as "hormone metabolic process", "hexose metabolic process" and "glucose metabolic process". Combinations 9 and 12 (leverage selection by Gamma) found "extracellular matrix part" GO term, which plays a fundamental role in regulating remodeling processes in embryo development and is also involved in repair processes, inflammation and tumor invasion [36].

In summary, most of the biological information is shared by all the compared methods combinations, but not all of them contribute equally to improve our biological knowledge about the gene products dynamics in this context. Each one of the combinations leads to the extraction of some particular biological functions than the rest of the methods cannot detect or, at least, not in such degree of specificity.

## 4. Conclusion

In this work, we have presented and compared several strategies to select the most relevant genes in multivariate models applied to the analysis of complex genomic data. The starting point of this contribution is the adoption of a multivariate dimension reduction strategy, commonly used in data exploration for the identification of important genes. In comparison to univariate methods that carry out gene-wise analysis, the multivariate approach exploits the coordinated nature of gene expression and avoids the application of multiple testing corrections that seriously diminishes statistical power in genomic research. In these scenarios, two additional factors are also impor-

tant. Firstly, the high-dimensionality of the feature space, that results in data structures where the number of variables can be two or three orders of magnitude the number of observations. And second, the low signal to noise ratio of the measurements. This implies that traditional multivariate feature selection methods are generally not applicable. The basic contribution in this paper is that the variable selection choices we propose always involve studying the distribution of the statistics used to measure the importance of the variables. Hence, the threshold for these statistics is set according to the shape of the distribution rather than selecting a fixed percentage of the total number of variables, which is a common and rather arbitrary practice in this kind of analysis. Our main concern in this study was to identify methodologies that will generally work well in different scenarios of dataset size, diversity of gene expression signals and levels of noise, since these are normally given features not fixed by the experimentalist. We have also tried to gather selection methods with an easy implementation and comprehension, as we understand that variable selection is only a small part of a genomic study and researchers may need quick but consistent solutions. In this work, some methods were taken from the literature (Box, Jackson & Mudholkar) or adapted to be used in this context (resampling), while others are novel proposals (minAS, Gamma approximation). These variable selection methods were first compared on simulated datasets to evaluate which ones presented the best performance and to quantify the influence of some biological data features on the goodness of the selection. In general, Gamma and minAS methods showed the best behavior for both SPE and leverage thresholds computation, as well as Jackson & Mudholkar's method for SPE. It was also seen that the higher the percentage of signal genes or the number of genes are, the better minAS performance is, while Gamma approximation is not significantly affected by these biological parameters, being therefore a more robust methodology. However, modifying minAS default options (such as increasing the smoothing parameter) proved to improve the performance of this method. The application of these three approaches on a real experimental dataset verified their usefulness to select relevant genes. In all cases, relevant biological conclusions could be obtained on the gene selection provided by the different methods, although specific biological functions were differentially uncovered by each approach. Interestingly, the major differences in gene selection and functional enrichment were the result of the method choice for the SPE statistic, while leverage seemed to be more robust for the statistical model applied. This result is interesting as the SPE measures the

deviation of each gene from the general multivariate model. Differentially expressed genes that follow a minority expression pattern tend to have high SPE values [9]. Our results indicate that selection on this part of the signal is also biologically relevant.

It should be outlined that the conclusions of this work are based on the simulation studies performed and might not be valid outside the biological scenarios analyzed. However, since the simulation algorithm was carefully designed to mimic real datasets and a vast variety of scenarios was considered (comprising more than 600 datasets), we believe that the results are generally valid for most multifactorial gene expression experiments.

Finally, we have focused on multifactorial designs because the variable selection issue has not been sufficiently developed for these complex experimental setups. The ASCA-genes framework was chosen to model these data, since it is considered a suitable methodology for the analysis of genomic datasets with such experimental designs. However, as the proposed variable selection methods are based on modeling the distribution of multivariate statistics, they are generally applicable to different dimension reduction techniques and kind of data by changing the statistic measuring the importance of the variables in the model. In fact, we have successfully applied our methods in other contexts, as for example in [37], where minAS was used for selecting variables from genomic and metabolomic data in Tucker3 and N-PLS models.

The minAS and Gamma variable selection methods applied to ASCA-genes analysis have been implemented in the web suite for Serial Expression Analysis, SEA (http://sea.bioinfo.cipf.es/) [18], and is freely available to the scientific community.

## References


[1] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, Molecular Biology of the Cell 9 (12) (1998) 3273–3297.

[2] A. V. Lukashin, R. Fuchs, Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters, Bioinformatics 17 (5) (2001) 405–414.

[3] S. Raychaudhuri, J. M. Stuart, R. B. Altman, Principal components analysis to summarize microarray experiments: application to sporulation time series, Pac Symp Biocomput. (2000) 455–466.

[4] J. Dai, L. Lieu, D. Rocke, Dimension reduction for classification with gene expression microarray data, Statistical applications in genetics and molecular biology 5 (1) (2006) 6.

[5] L. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika 31 (1966) 279–311.

[6] R. A. Harshman, Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis, UCLA Working Papers in Phonetics 16 (1970) 1–84.

[7] B. Yener, E. Acar, P. Aguis, K. Bennett, S. Vandenberg, G. Plopper, Multiway modeling and analysis in stem cell systems biology, BMC Systems Biology 2 (1) (2008) 63.

[8] A. K. Smilde, J. J. Jansen, H. C. J. Hoefsloot, R.-J. A. N. Lamers, J. van der Greef, M. E. Timmerman, ANOVA-Simultaneous Component Analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21 (2005) 3043–3048.

[9] M. J. Nueda, A. Conesa, J. A. Westerhuis, H. C. J. Hoefsloot, A. K. Smilde, M. Talón, A. Ferrer, Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA, Bioinformatics 23 (2007) 1792–1800.

[10] I. T. Jolliffe, Discarding variables in a Principal Component Analysis. I: Artificial data, Journal of the Royal Statistical Society. Series C (Applied Statistics) 21 (2) (1972) 160–173.

[11] I. T. Jolliffe, Discarding variables in a Principal Component Analysis. II: Real data, Journal of the Royal Statistical Society. Series C (Applied Statistics) 22 (1) (1973) 21–31.

[12] G. P. McCabe, Principal Variables, Technometrics 26 (2) (1984) 137–144.

[13] W. J. Krzanowski, Selection of variables to preserve multivariate data structure, using Principal Components, Journal of the Royal Statistical Society. Series C (Applied Statistics) 36 (1) (1987) 22–33.

[14] Q. Guo, W. Wu, D. L. Massart, C. Boucon, S. de Jong, Feature selection in Principal Component Analysis of analytical data, Chemometrics and Intelligent Laboratory Systems 61 (2002) 123–132.

[15] F. Westad, M. Hersleth, P. Lea, H. Martens, Variable selection in PCA in sensory descriptive and consumer data, Food Quality and Preference 14 (2003) 463–472.

[16] Y. Lu, I. Cohen, X. S. Zhou, Q. Tian, Feature selection using Principal Feature Analysis, in: ACM Multimedia Conference, 2007.

[17] R. Luss, A. d'Aspremont, Clustering and feature selection using Sparse Principal Component Analysis, Optimization and Engineering (2008) 1573–2924.

[18] M. Nueda, J. Carbonell, I. Medina, J. Dopazo, A. Conesa, Serial Expression Analysis: a web tool for the analysis of serial gene expression data, Nucleic Acids Research 38 (suppl 2) (2010) W239.

[19] H. Martens, T. Naes, Multivariate calibration, John Wiley & Sons, Ltd. Chichester, 1989.

[20] G. E. P. Box, Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification, The Annals of Mathematical Statistics 25 (3) (1954) 484–498.

[21] E. Edgington, Randomization Tests, 1980.

[22] J. E. Jackson, G. S. Mudholkar, Control procedures for residuals associated with Principal Component Analysis, Technometrics 21 (3) (1979) 341–349.

[23] B. Efron, R. Tibshirani, J. Storey, V. Tusher, Empirical bayes analysis of a microarray experiment, Journal of the American Statistical Association 96 (456) (2001) 1151–1160.

[24] B. Efron, Large-scale simultaneous hypothesis testing, Journal of the American Statistical Association 99 (465) (2004) 96–104.

[25] S. C. Choi, R. Wette, Maximum likelihood estimation of the parameters of the gamma distribution and their bias, Technometrics 11 (4) (1969) 683–690.

[26] M. Rosenblatt, Remarks on some non-parametric estimates of a density function, The Annals of Mathematical Statistics 27 (1956) 832–837.

[27] B. Silverman, Density estimation for statistics and data analysis, Chapman & Hall/CRC, 1986.

[28] B. W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta 405 (1975) 442–451.

[29] R. S. Barcikowski, R. R. Robey, Decisions in single group repeated measures analysis: Statistical tests and three computer packages, The American Statistician 38 (2) (1984) 148–150.

[30] S. Prado-López, A. Conesa, A. Armiñán, M. Martínez-Losa, C. Escobedo-Lucea, C. Gandia, S. Tarazona, D. Melguizo, D. Blesa, D. Montaner, S. Sanz-González, P. Sepúlveda, S. Götz, J. E. O'Connor, R. Moreno, J. Dopazo, D. J. Burks, M. Stojkovic, Hypoxia promotes efficient differentiation of human embryonic stem cells to functional endothelium, Stem Cells.

[31] F. Al-Shahrour, P. Minguez, J. Tarraga, D. Montaner, E. Alloza, J. M. Vaquerizas, L. Conde, C. Blaschke, J. Vera, J. Dopazo, BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments, Nucl. Acids Res. 34 (suppl_2) (2006) W472–476.

[32] A. Conesa, S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (18) (2005) 3674–3676.

[33] P. Carmeliet, Angiogenesis in life, disease and medicine, Nature 438 (2005) 932–936.

[34] J. C. Kovacic, J. Moore, A. Herbert, D. Ma, M. Boehm, R. M. Graham, Endothelial progenitor cells, angioblasts, and angiogenesisold terms reconsidered from a current perspective, Trends Cardiovasc. Med. 18 (2) (2008) 45–51.

[35] Y. A. Hannun, L. M. Obeid, Principles of bioactive lipid signalling: lessons from sphingolipids, Nature Reviews Molecular Cell Biology 9 (2008) 139–150.

[36] S. C. Huang, B. C. Sheu, W. C. Chang, C. Y. Cheng, P. H. Wang, S. Lin, Extracellular matrix proteases - cytokine regulation role in cancer and pregnancy., Front Biosci. 14 (2009) 1571–1588.

[37] A. Conesa, J. Prats-Montalbán, S. Tarazona, M. Nueda, A. Ferrer, A multiway approach to data integration in systems biology based on Tucker3 and N-PLS, Chemometrics and Intelligent Laboratory Systems 104 (1) (2010) 101–111.