Short Papers

A Weighted Principal Component Analysis and Its Application to Gene Expression Data

Joaquim F. Pinto da Costa, Hugo Alonso, and Luís Roque

Abstract—In this work, we introduce in the first part new developments in Principal Component Analysis (PCA) and in the second part a new method to select variables (genes in our application). Our focus is on problems where the values taken by each variable do not all have the same importance and where the data may be contaminated with noise and contain outliers, as is the case with microarray data. The usual PCA is not appropriate to deal with this kind of problems. In this context, we propose the use of a new correlation coefficient as an alternative to Pearson's. This leads to a so-called weighted PCA (WPCA). In order to illustrate the features of our WPCA and compare it with the usual PCA, we consider the problem of analyzing gene expression data sets. In the second part of this work, we propose a new PCA-based algorithm to iteratively select the most important genes in a microarray data set. We show that this algorithm produces better results when our WPCA is used instead of the usual PCA. Furthermore, by using Support Vector Machines, we show that it can compete with the Significance Analysis of Microarrays algorithm.

Index Terms—Correlation, principal component analysis, support vector machines, microarray data, gene selection.

1 Introduction

Principal Component Analysis (PCA) [18] is widely used in the analysis of high-dimensional data. There are, however, some applications where the usual PCA is not recommended because it gives the same importance to all observations and is sensitive to the presence of outliers and noise in the data. For instance, the larger absolute expression values in microarrays should be given higher importance as they relate to genes that are more "responsible" for the problem in analysis. Basically, the amount of expression of a gene indicates the approximate number of copies of mRNA of that gene which are produced inside the cell, and so it provides information about the gene function and contribution to the development of the related problem [4], [35], [13]. In addition, as pointed out in [32], a gene which has a lower expression level in one condition will typically be measured with relatively less precision in that condition, and so these expression values can be very noisy.

- *J.F. Pinto da Costa is with the Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal, and also with Centro de Matemática da Universidade do Porto (CMUP), Portugal. E-mail: jpcosta@fc.up.pt.*
- *H. Alonso is with the Faculdade de Economia e Gestão da Universidade Lusófona do Porto, Rua Augusto Rosa, 24, 4000-098 Porto, Portugal, with the Departamento de Matemática da Universidade de Aveiro, Campus de Santiago, 3810-193, Aveiro, Portugal, and also with the Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA), Universidade de Aveiro, Portugal. E-mail: hugo.alonso@ua.pt.*
- *L. Roque is with the Grupo de Investigação em Engenharia do Conhecimento e Apoio à Decisão (GECAD), Instituto Superior de Engenharia do Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal. E-mail: lar@isep.ipp.pt.*

Manuscript received 21 Nov. 2008; revised 30 Apr. 2009; accepted 14 July 2009; published online 19 July 2009.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2008-11-0203. Digital Object Identifier no. 10.1109/TCBB.2009.61.

In this paper, we propose:

- First, a new PCA to solve the aforementioned problems affecting the usual PCA. In order to cope with outliers and noise, we will use rankings instead of the original data. For instance, in microarray data, we will start by ranking the expression values (observations) in each gene (variable). Then, in order to give higher weight to the larger absolute expression values inside each gene, we will use a new weighted rank correlation coefficient instead of the usual Pearson's. This gives rise to a so-called weighted PCA (WPCA).
- Second, a new PCA-based algorithm to iteratively select the most important genes for discriminatory purposes in a microarray data set.

We will illustrate the application of our methods to microarray data because this kind of data possesses all the characteristics that we need to present the advances introduced in this work. We have searched for works that give higher importance to the larger expression values in PCA, but there are only a few. For instance, Jansen et al. [16] use an established method of weighted principal component analysis introduced in [20] to weigh the elements of metabolomics data in accordance with a priori information. The problem of robustness to outliers and noise is also of major importance (see, for instance, [14], [41]); however, the usual PCA and the weighted PCA of [20] do not cope with it. Finally, the PCA-based algorithms for selecting the most important genes in microarrays, like the one in [9], take into account the importance of each principal component in an isolated fashion, and highly correlated genes can be chosen as being important, which can add redundancy to the process of gene selection.

In this paper, we begin by introducing in Section 2 a novel weighted PCA based on a new rank correlation coefficient. Moreover, we illustrate its features and apply it to microarray data. Next, in Section 3, we propose a new PCA-based algorithm for selecting the most important variables for discriminatory purposes in a data set. We compare the classification results using Support Vector Machines [6] in microarray data sets for four different methods of choosing genes: our algorithm with our WPCA, our algorithm with the usual PCA, the popular Significance Analysis of Microarrays (SAM) supervised algorithm [37], [38], and an unsupervised algorithm called Pattern discovery via eigengenes [37], [38]. Furthermore, we explain the biological meaning of the relevant genes chosen. Finally, we end the paper in Section 4 with the main conclusions.

Before moving on, we note that although the application focused in this work concerns microarray experiments, a broader range of applications can be considered. As an example, we can look at problems containing preferences stated by humans or recommendations provided by decision support systems; naturally, the first preferences or recommendations are more important and accurate than the last ones. Another potential application of our methodology is when we have various stock trading support systems and we want to summarize the information given by them. In this case, what investors want is a grading of the stocks in question, which can be represented by a ranking. If we have various rankings corresponding to different support systems, we might want to summarize the information by using PCA; however, the stocks ranked higher are obviously more important than the last ones. Another problem that is usually handled as a ranking task is information retrieval [2]. Again, rank importance should be taken into account, although that rarely happens [17]. Similar remarks apply to recommender systems [5].

2 A New Weighted Version of PCA

In this section, we seek for a few linear combinations of the variables that account for most of the variation present in the data. This is

done by using Principal Component Analysis, introduced by Karl Pearson in 1901 and Hotelling in 1933 (see, for instance, [9], [11], [13], [18]). Let us denote by $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ the vector containing all measurements for the p variables (genes in our application). Thus, our data consist of n vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ in a space of p dimensions, where n is the number of samples. Mathematically, the PCA problem consists in finding a subspace of dimension K of the original space which maximizes the dispersion of the points projected onto that subspace. The solution to this optimization problem (see [13], [22]) is given by the eigenvectors corresponding to the K largest eigenvalues of the covariance matrix of the sample, $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \hat{\mu})(\mathbf{X}_i - \hat{\mu})^T$, where $\hat{\mu}$ is the mean vector of the sample. For various reasons, it is common to start by standardizing the data. This consists in subtracting from each observation the average of the variable in question and divide by the corresponding standard deviation times $\sqrt{n}$. With this initial standardization, the principal components obtained are linear combinations of the original variables, and the coefficients of these linear combinations are given by the elements of the eigenvectors of the usual correlation matrix based on Pearson's correlation coefficient r.

In this work, we introduce a weighted version of PCA (WPCA), where more importance is given to observations whose values are more important. We think this makes sense, for instance, with microarray data, given that, as explained in Section 1, the higher the absolute expression value the more probable is that the gene in question is related to the particular problem. To that end, this weighted PCA uses a new correlation coefficient that gives higher weights to observations that are considered to be more important. In addition, this correlation coefficient is not sensitive to the presence of outliers and noise in the data. A preliminary version of our work has been introduced in [26], where we used a different weighted rank correlation coefficient introduced in [28]. We now describe this previous coefficient.

2.1 Weighted Rank Correlation

In the usual PCA, the eigenvectors of the covariance matrix or the Pearson's correlation matrix (standardized data) contain the coefficients of the linear combinations of the original variables corresponding to the new variables (features and components). As is well known, the Pearson's correlation coefficient is very sensitive to the presence of outliers and noise. To overcome this, we will use the ranks of the observations. We must therefore start by ranking the observations in each variable from 1 (highest rank) to n (lowest rank). For the sake of simplicity, let us use the ranks directly rather than the values in the series, that is, R_i and Q_i to represent the ranks of two variables (genes in our application) corresponding to observation (sample) i. Now, if we calculate Pearson's correlation coefficient of the ranked data, we obtain the Spearman's rank correlation coefficient r_S, which is given by the expression

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(Q_i - \overline{Q})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(Q_i - \overline{Q})^2}},$$

where $\overline{R}$ and $\overline{Q}$ are the average ranks. However, for computational purposes, a more convenient expression which assumes there are no ties is

$$r_S = 1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2}{n^3 - n}.$$

It is clear from this rewritten form of r_S that the calculation of the distance between two ranks in Spearman's coefficient is given by

$$D_i^2 = (R_i - Q_i)^2,$$

which does not take rank importance into account, because if (R_i, Q_i) is, for instance, $(1, 3)$ or $(n-2, n)$, the contribution is the same. In [28], the following alternative distance measure is proposed:

$$\begin{aligned} WD_i^2 &= (R_i - Q_i)^2((n - R_i + 1) + (n - Q_i + 1)) \\ &= D_i^2(2n + 2 - R_i - Q_i). \end{aligned}$$

The first term of this product is D_i^2, exactly as in Spearman's coefficient, and represents the distance between R_i and Q_i; the second term is a linear weighting function which represents both the importance of R_i and Q_i. Taking this expression as the distance, the authors obtain the weighted rank measure of correlation

$$r_W = 1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2(2n + 2 - R_i - Q_i)}{n^4 + n^3 - n^2 - n}, \tag{1}$$

which yields values between -1 and $+1$. Some properties of the distribution of the statistic r_W, including its sample distribution, are analyzed in [28], [27]; in particular, the expected value of this statistic is zero when the two variables are independent, and its sampling distribution converges to the Gaussian when the sample size increases. A table of the most significant percentiles is also given. This coefficient r_W has nevertheless some drawbacks:

- it cannot be used when there are tied values;
- in some applications, a linear weighting function might not be enough; and
- we did not succeed in writing r_W as an inner product, which is a useful property for PCA as we will see.

We introduce now in this work the use of a new weighted rank correlation coefficient r_{W2}, whose expression has been given in [29], and propose to use it in PCA.

2.2 A New Weighted Rank Correlation Coefficient

In [28], the calculation of the distance between two ranks R_i and Q_i is given by $WD_i^2 = (R_i - Q_i)^2(2n + 2 - R_i - Q_i)$, where the second term of the product is a linear weighting function which represents the importance of R_i and Q_i. Now, we propose the distance measure

$$W_2D_i^2 = (R_i - Q_i)^2(2n + 2 - R_i - Q_i)^2,$$

which reflects more than WD_i^2 the higher importance of agreement on top ranks. It is common to define rank correlation coefficients, such as Spearman's, as a linear function of the distance between the two vectors of ranks [25]. In our case, this corresponds to define a coefficient of the form

$$A + B\sum_{i=1}^{n}(R_i - Q_i)^2(2n + 2 - R_i - Q_i)^2, \tag{2}$$

where the constants A and B are such that it takes values between -1 and $+1$. In order to find A and B, we will start by doing a specific data transformation and then compute the Pearson's coefficient on the transformed data. The expression obtained is exactly of the form (2), from where the constants A and B follow. The transformation consists in substituting the value of observation i in the first variable by the value of $R_i' = R_i(2n + 2 - R_i)$, where R_i is the rank of that observation. We do the same for the other variable, obtaining $Q_i' = Q_i(2n + 2 - Q_i)$.

It is easy to see that, in case of no ties, the average values are

$$\overline{R'} = \frac{1}{n}\sum_{i=1}^{n} R_i(2n + 2 - R_i) = \frac{(n+1)(4n+5)}{6}$$

and

$$\overline{Q'} = \frac{1}{n}\sum_{i=1}^{n} Q_i(2n + 2 - Q_i) = \frac{(n+1)(4n+5)}{6}.$$

We will now consider Pearson's correlation coefficient of these new values to get

TABLE 1
Data Sets Used in the Practical Experiments

Dataset	Samples	Genes
Embryonal tumours	60	7129
Global cancer map	144	16063
Leukemia	38	7129
NCI60	64	6830

TABLE 2
Genes Containing Outliers and Severe Outliers

Dataset	Genes	Outliers	Severe outliers
Embryonal tumours	7129	6672	3553
Global cancer map	16063	15893	13073
Leukemia	7129	4713	1504
NCI60	6830	4421	788

$$r_{W2} = \frac{\sum_{i=1}^{n}(R'_i - \overline{R'})(Q'_i - \overline{Q'})}{\sqrt{\sum_{i=1}^{n}(R'_i - \overline{R'})^2}\sqrt{\sum_{i=1}^{n}(Q'_i - \overline{Q'})^2}}. \quad (3)$$

When no ties occur in marginal ranks, we have

$$r_{W2} = 1 - \frac{90\sum_{i=1}^{n}(R_i - Q_i)^2(2n+2-R_i-Q_i)^2}{n(n-1)(n+1)(2n+1)(8n+11)}, \quad (4)$$

and so, we have found that the rank correlation coefficient that we were looking for (2) is the Pearson's correlation coefficient of the weighted ranks $R'_i = R_i(2n+2-R_i)$ and $Q'_i = Q_i(2n+2-Q_i)$ for $A = 1$ and $B = -\frac{90}{n(n-1)(n+1)(2n+1)(8n+11)}$.

From the expression above (4), it is clear that the maximum value of this new coefficient is attained when the two vectors of ranks are the same, that is, $R_i = Q_i,\ \forall i = 1, 2, \ldots, n$. It can also be seen that the minimum value of this coefficient is attained when the two vectors of ranks are inverted, that is, $Q_i = n + 1 - R_i$, $\forall i = 1, 2, \ldots, n$.

Now, we consider the effect of ties on the calculation of r_{W2}. First, we shall adopt the midrank method, i.e., we replace the rank numbers where ties exist by the average of these ranks. For instance, if we observe ties in the second, third, and fourth initial values, we replace those numbers by the average number $\frac{2+3+4}{3} = 3$; that is, $R_2 = R_3 = R_4 = 3$. We have found the modifications that the presence of ties has in (4). However, these modifications are quite complex and we think that in the presence of ties the easiest is to use Pearson's correlation coefficient of the weighted ranks, i.e., (3). When there are no ties, we can do the same or use, instead, (4).

It is clear from above that the computation of the new correlation coefficient is equivalent to do a data transformation to each variable as

$$R'_i = R_i(2n+2-R_i), \quad (5)$$

and then compute the Pearson's correlation coefficient. R_i represents the rank of each observation value; usually the smallest value has rank 1, the second smallest rank 2, and so on. However, depending on the problem, we can rank the observations differently. For instance, in microarray data, because we want to give higher weight to the larger absolute expression values and r_{W2} gives higher weights to the first ranks, we will therefore give rank 1 to the largest absolute value, rank 2 to the second largest, etc.

Finally, note that the weighted PCA can be done using any common software for PCA analysis as long as we start by transforming our data according to (5). If in our data set the number of variables is smaller than the number of observations, that is all that we have to do. Otherwise, it is common to use Singular Value Decomposition (SVD) or the NIPALS algorithm [39]. However, we propose in the Appendix, a simpler, yet efficient, way for computing the principal components whenever there are more variables than observations in the data.

2.3 Application of Our Method WPCA to Microarray Data

In bioinformatics, a common application of PCA is on the analysis of the high-dimensional microarray data. As pointed out by Bicciato et al. [3], molecular diagnostics based on microarray data present major challenges due to the overwhelming number of variables and the complex, multiclass nature of tumor samples. Thus, it is of paramount importance the development of both suitable automatic marker selection methods, like those based on PCA, that allow the identification of genes that are most likely to confer high classification accuracy of multiple tumor types, and suitable multiclass classification schemes. The works in [32] and [12] are also examples of the use of PCA for gene expression data analysis. Naturally, other types of data in bioinformatics can be analyzed by this powerful technique. For instance, Jansen et al. [16] use PCA to get a simplified view of metabolomics data. Scholz et al. [34] apply PCA, but also independent component analysis, to detect relevant information from spectra of total compositions of metabolites. Many more examples could be given to show the importance of PCA in these and other applications in bioinformatics.

In the following, we will apply our method to four microarray data sets, namely those described in Table 1.[1]

Our aim now is to compare results when we apply the usual PCA and our WPCA. We recall that in order to apply WPCA, we only need to consider as input data to a PCA software the original data transformed according to (5).

2.4 Robustness to Outliers and Noise

As discussed in Section 1, the presence of noise in microarray data is common and that motivated us to use ranks instead of the original data to cope with this problem. We will now design an experiment to show that, first of all outliers are very common in this kind of data and then we will use both the usual PCA and our weighted version to see how much outliers affect them. Let us designate by $\xi_{0.25}$ and $\xi_{0.75}$ the first and third quartiles of the expression values for a given gene. We will use a common criteria which considers that all observations outside the interval $[\xi_{0.25} - 1.5(\xi_{0.75} - \xi_{0.25}),\ \xi_{0.75} + 1.5(\xi_{0.75} - \xi_{0.25})]$ are considered outliers and those outside $[\xi_{0.25} - 3(\xi_{0.75} - \xi_{0.25}),\ \xi_{0.75} + 3(\xi_{0.75} - \xi_{0.25})]$ severe outliers. In the four data sets under our consideration, the number of genes containing outliers of the two types are as given in Table 2.

We see therefore that in the first two data sets more than 90 percent of the genes contain outliers and around 65 percent in the third and fourth data sets. Also, inside the genes containing outliers, we found that around 6 percent on average of the observations are outliers. Thus, as suspected, there are many outliers which we will show that affect the usual principal component analysis, given that it is based on Pearson's correlation coefficient which is very sensitive to outliers. We will now pick a small number of genes in order to illustrate the effect outliers have in the two correlations (Pearson's r and ours r_{W2}) and then compute the two principal component analyses before and after the removal of outliers to see how robust they are. We will use genes g19, g600, and g830 of the first data set, Embryonal Tumors, in our example which have three, five, and six outliers, respectively. The following tables contain the values of r and r_{W2} for these three genes before and after the removal of all outliers (see Table 3).

As it is clear from these tables, the effect of outliers in Pearson's correlation is dramatic; for instance, for the genes g19 and g830, before the removal of outliers the correlation was -0.5593 and after the removal of outliers it was 0.4696! The values of r_{W2} also change, because we have removed some observations; but, as it is clear, the differences before and after the removal are much smaller. To

1. The first three data sets can be found at http://www.lsi.us.es/~aguilar/datasets.html and the fourth one at http://genome-www.stanford.edu/nci60/.

TABLE 3
Values of r and r_{W2} for the Genes g19, g600, and g830

Before removal of all outliers			
Correlation	(g19, g600)	(g19, g830)	(g600, g830)
r	0.7113	-0.5593	-0.6883
r_{W2}	0.6198	0.5463	0.6563
After removal of all outliers			
Correlation	(g19, g600)	(g19, g830)	(g600, g830)
r	0.6597	0.4696	0.4984
r_{W2}	0.5395	0.4458	0.5953

TABLE 4
First Two Principal Components with and without Outliers

		r
PC1	Before	$0.0053 \times g19 + 0.9411 \times g600 - 0.3381 \times g830$
	After	$0.0042 \times g19 + 0.8740 \times g600 + 0.4859 \times g830$
PC2	Before	$0.0139 \times g19 + 0.1763 \times g600 + 0.9842 \times g830$
	After	$0.0023 \times g19 + 0.6691 \times g600 - 0.7432 \times g830$
		r_{W2}
PC1	Before	$0.5615 \times g19 + 0.5960 \times g600 + 0.5740 \times g830$
	After	$0.5468 \times g19 + 0.5933 \times g600 + 0.5908 \times g830$
PC2	Before	$-0.7708 \times g19 + 0.1247 \times g600 + 0.6248 \times g830$
	After	$-0.8320 \times g19 + 0.3661 \times g600 + 0.4169 \times g830$

finalize this section, we will now find the expression of the first two principal components (which account for around 90 percent of the total variation in all four cases), before and after the removal of outliers to analyze the differences (see Table 4).

As we suspected, large differences occur in the expression of the principal components when we use the usual PCA (correlation r); the importance (coefficient) given to the three genes changes sometimes drastically when we remove the outliers. As for our weighted PCA (correlation r_{W2}), the changes are comparatively small, as we expected.

Thus, the inclusion of the outliers, which are very important observations in this problem, has a dramatic effect as it can change completely the results: if we don't include the outliers and include only the other observations, the results can be opposite, so to say. This is not a desirable property of the usual correlation or PCA; we don't want to ignore the outliers because in this application they certainly represent important information about the problem; nevertheless, we would like that the difference in the final results was not so large. Our correlation r_{W2} is thus appropriate to this problem because it gives higher importance (weight) to the outlier observations which is very appropriate here; nevertheless, it is much more robust because including the outliers doesn't change dramatically neither the values of the correlation nor the principal components.

Up to now, we introduced a novel weighted PCA and showed its relevance in analyzing gene expression data. In the remainder of the paper, we will focus on choosing relevant genes for the diseases in analysis.

3 A New Method for Selecting Relevant Genes in Microarray Data

In the previous sections, we learned how to find the principal components, both for the usual and the weighted PCAs. Here, we propose a new PCA-based algorithm for selecting the most important genes for discriminatory purposes in a microarray data set, which is an important problem [37], [8], [7]. We compare the classification results using Support Vector Machines (SVMs) in the four microarray data sets above for four different methods of choosing genes: our algorithm with our WPCA, our algorithm with the usual PCA, the popular Significance Analysis of Microarrays (SAM) supervised algorithm [37], [38], and an unsupervised algorithm called Pattern discovery via eigengenes (here PDeig for short) [37], [38].

In any PCA, each of the principal components is a linear combination of all of the variables present in the data set; usually thousands of them. This makes it very difficult to interpret each principal component. Suppose that, for instance, the first principal component was $\sum a_i X_i$, where a_i are the coefficients in that component and X_i represents the values of variable i. It has been suggested [9] that restricting attention to those variables for which $|a_i| > c$, for some chosen cutoff value c, allowed us to focus on a small set of variables that might contain the most important information. However, this gives rise to four problems. First, as is usually known, the principal components have not all the same importance and so, should not be treated in the same way. Second, the variables which appear in all of the principal components (PCs) are the same and so we have to analyze all of the PCs at the same time and not separately. Third, in the case of supervised classification problems, we want our selection procedure to take into account the discriminant power of each gene. Fourth, many of the variables in the data set are usually highly correlated and in order to choose a good and small list of genes we should privilege uncorrelated genes; as Tibshirani et al. [37] pointed out, after a minimal list is found, one can always search for more genes that are highly correlated with the genes in that list. In our experiments using SVMs, we have found better discriminatory results by privileging uncorrelated genes. This is also found by Dudoit et al. [8, p. 85] in some of their experiments.

In order to solve the problems just mentioned, we introduce here a strategy for choosing the L most important original variables after a PCA (usual or weighted) has been performed on the data set. The number L must be chosen by the user. The first thing we have to do is to decide how many principal components to use. There are many ways to choose the number of principal components, K, and here we will choose as many as needed to have at least 90 percent of the information present in the data set. Thus, we can represent each of the K principal components by

$$\mathrm{PC}_k = \sum_{i=1}^{p} a_{ki} X_i, \quad k = 1, \ldots, K, \tag{6}$$

where X_i represents variable i and a_{ki} the coefficient given by the kth principal component to that variable. Let us designate by λ_k the importance of the kth component, which corresponds to an eigenvalue of a certain matrix, as seen above. We will now define the global importance of each variable X_i, $i = 1, 2, \ldots, p$, by the expression

$$\mathrm{GI}(X_i) \;=\; DP(X_i) \times \sum_{k=1}^{K} |a_{ki}| \lambda_k, \tag{7}$$

where

$$DP(X_i) = \frac{s^2_{\bar{X}_i}}{\sum_k s^2_{X_i^k}},$$

which is the ratio between the variance of the class centers and the sum of the variances within each class, allows us to take into account the discriminant power of the variable X_i (see the third problem above). Now, in order to find the L most important variables, we apply the next algorithm:

1. Choose the variable X_i which maximizes the global importance given by (7). This is the most important variable of all.
2. Now, for $l = 1, 2, \ldots, L-1$, do
   a. For each variable X_i not yet chosen, find the Pearson's correlation coefficient between X_i and the

TABLE 5
Support Vector Machines Error Results

Dataset	SVM kernel	WPCA	PCA	SAM	PDeig
Embryonal tumours	sigmoid	0.17	0.20	0.18	0.37
	radial basis	0.20	0.22	0.17	0.35
Global cancer map	sigmoid	0.47	0.51	0.62	0.68
	radial basis	0.42	0.45	0.61	0.68
Leukemia	sigmoid	0.00	0.00	0.03	0.20
	radial basis	0.00	0.00	0.03	0.18
NCI60	sigmoid	0.18	0.23	0.54	0.66
	radial basis	0.27	0.25	0.49	0.56

l variables which have already been chosen: $r(X_i, X_j)$, $j = 1, \ldots, l$. Let $r_{\max,i}$ be the maximum in absolute value of these l correlations.

b. Actualize the global importance of variable X_i by

$$\mathrm{GI}^{\star}(X_i) \;=\; \mathrm{GI}(X_i) \times (1 - r_{\max,i}). \tag{8}$$

c. Choose the variable X_i which maximizes the actualized global importance given by (8).

3.1 Support Vector Machines Classification: Results for Genes Chosen with SAM, PDeig, and with Our Method for WPCA and PCA

In this section, we will compare the genes selected by the weighted principal component analysis with those chosen by the usual PCA in terms of discriminatory power. Furthermore, we will show that our method of selecting genes is very competitive with the popular supervised method called SAM and better than an unsupervised method called PDeig. The data concerning the genes selected by the four methods are used as inputs to the SVMs software included in R [31], namely in the e1071 library. We used the default parameters in R for tuning the SVMs with the sigmoid and radial basis kernels, that is, for the sigmoid kernel $\mathcal{K}(u, v) = \tanh(\gamma u^T v + c)$, we set $\gamma = 1/(\text{data dimension})$ and $c = 0$, and for the radial basis kernel $\mathcal{K}(u, v) = \exp(-\gamma \|u - v\|^2)$, we took $\gamma = 1/(\text{data dimension})$.

Given that all four methods have parameters which influence the number of genes to choose, for comparison purposes we will use the same number of genes in the four cases. In the Embryonal tumours data set, SAM chose 16 genes by default and so we used the same number of genes in the WPCA and PCA methods. In the other three data sets, SAM chooses hundreds of genes by default; however, in our experiments, we found that the classification results with such a large number of genes are not significantly better that considering only 20 genes. This type of behavior was also observed in [8, p. 85] and pointed out in [1], where the authors conclude that the number of genes can be reduced greatly without increasing the prediction error.

Table 5 presents the mean classification error rate obtained with 10-fold cross-validation. It can be seen that in the first, second, and fourth data sets, WCPA is better than PCA and PDeig and very competitive with SAM. The third data set is very simple in what concerns discriminating between the classes and so three methods exhibit very good results, namely WPCA, PCA, and SAM, whereas the other method, PDeig, has a poorer performance.

3.2 Analysis of the Chosen Genes

In this section, we will study the biological relevance of some of the most important genes used in Section 3.1. We will restrict our attention to the genes chosen by WPCA, PCA, and SAM, given that the results obtained by PDeig are comparatively poor.

In the Embryonal tumors data set, the leukotriene C4 synthase $LTC4S$ (gene U50136rna1) and the neurotrophic tyrosine kinase, receptor, type 3 (TrkC) $NTRK3$ (gene S76475) are identified both by SAM and our WPCA-based algorithm. Mutations in $NTRK3$ have been associated with medulloblastomas, secretory breast carcinomas, and other cancers (see [30]). Furthermore, our method includes the high-mobility group AT-hook $1HMGA1$ (gene L17131rna1), the Sodium channel 2 mRNA (gene hBNaC2), and the alternatively spliced $ACCN2$ (gene U78180), which were not identified by SAM as being relevant; these genes are also referred to in [30].

In the Global cancer map data set, the ARHGDIB Rho GDP dissociation inhibitor (GDI) beta $ARHGDIB$ (gene L20688) is identified by both SAM and our WPCA-based algorithm. Moreover, our method includes the KLK3 kallikrein-related peptidase $3KLK3$ (gene X07730) and the vascular endothelial growth factor C $VEGFC$ (gene U43142), which were not identified by both SAM and the usual PCA-based algorithm. The first gene is in the Kallikreins subgroup of serine proteases, which have diverse physiological functions. Growing evidence suggests that many kallikreins are implicated in carcinogenesis and some have potential as novel cancer and other disease biomarkers (see http://www.ncbi.nlm.nih.gov/sites/entrez). In turn, the second gene is essential in lymph-node metastasis, presumably because enhanced metastatic potential including lymphangiogenesis induced by VEGF-C is vital in lymph-node metastasis of gastric cancer [21].

In the Leukemia data set, SAM and our WPCA-based algorithm identified six genes in common. Some of the remainder genes identified by our method include the cell division cycle 25 homolog A $CDC25A$ (gene M81933), $SMARCA4$ (gene D26156s), the interleukin 18 IL-18 (gene D49950), the myb myeloblastosis viral oncogene homolog (avian) MYB (gene U22376cds2s), the Non-SMC condensin I complex, subunit D2 $NCAPD2$ (gene D63880), and the CUG triplet repeat, RNA binding protein $1CUGBP1$ (gene U63289). $CDC25A$ is overexpressed in a variety of human malignancies [40]; inactivating mutations of the $SMARCA4$ gene, on chromosome $arm19p$, are present in several human cancer cell lines [23]; interleukin $18IL$-18 establishes a possible functional relationship between IL-18 and MMPs in myeloid leukemia; myb myeloblastosis viral oncogene homolog (avian) MYB is overexpressed in most human acute myeloid and lymphoid leukemias, and several studies using antisense oligonucleotides and dominant negative forms of MYB have shown that this gene activity is essential for continued proliferation of AML and CML cells; $CUGBP1$ is involved in the development of breast cancer and leukemia (see [24]). The three selection methods found homeobox A9 $HOXA9$ (gene U82759) as a relevant gene, which is in fact, important for leukemia identification [10].

In the NCI60 data set, the genes $CD53$ antigen and $DKK3$ dickkopf homolog 3 (Xenopus laevis) 2 were identified by all three methods. $CD53$ antigen interactions might contribute to cell survival in poorly vascularized regions of the tumor mass [42]. In turn, $DKK3$ can play a role in head and neck squamous cell carcinoma (HNSCC) carcinogenesis with unknown mechanism [19]. The genes laminin, alpha $3LAMA3$, and paxillin PXN were identified by both WPCA and PCA, but not by SAM. Down-regulation of Laminin-5 (LN5)-encoding genes ($LAMA3$, $LAMB3$, and $LAMC2$) has been reported in various human cancers [33]. Furthermore, the results in [33] demonstrate epigenetic inactivation of LN5-encoding genes in breast cancers and association of $LAMA3$ promoter methylation with increased tumor stage and tumor size. On the other hand, in lung cancer tissues [15], it has been established that there is an important role for paxillin PXN. Finally, our WPCA-based algorithm further identified the gene transducin-like enhancer of split 1 (E(sp1) homolog, Drosophila) $TLE1$. This gene was also consistently found by several independent groups to be an excellent discriminator between synovial sarcoma and other sarcomas, including histologically similar tumors such as malignant peripheral nerve sheath tumor [36].

4 Conclusions

In this paper, we introduced a new correlation coefficient that weighs observations according to their importance to the problem

in hand; moreover, this coefficient is robust to the presence of outliers and noise in the data.

We proposed the use of this new correlation coefficient on PCA, and concluded that its application to PCA is equivalent to carrying out a certain data transformation. This gave rise to a novel weighted PCA, WPCA, which is more robust than the usual PCA in microarray data sets.

We introduced also a new algorithm to select the most important variables in the original data set to which a PCA, usual or weighted, is applied. This PCA-based algorithm takes into account the global importance of each component and the discriminatory power of each variable and does not add redundancy by disabling the selection of new variables that are highly correlated with previously chosen ones. We verified in some microarray data sets that our PCA-based algorithm produces better results when our WPCA is used instead of the usual PCA. Furthermore, we showed that it can compete with the popular Significance Analysis of Microarrays (SAM) algorithm. The classifiers using the data corresponding to the genes chosen by these three methods were built using Support Vector Machines. The classification results were supported by the biological meaning of the relevant genes chosen.

APPENDIX A

COMPUTATION OF THE PRINCIPAL COMPONENTS IN THE CASE OF MORE VARIABLES THAN OBSERVATIONS

A.1 Computation of the Unweighted Principal Components

Let us start by designating by $\mathbb{X}$ the data matrix with n lines, corresponding to the n samples, and p columns, corresponding to the p variables. In the usual PCA, we must find the $p \times p$ matrix of Pearson's correlation coefficients. To do so, we start by standardizing the data as

$$X_{ij} \leftarrow \frac{X_{ij} - \overline{X}_j}{S_j\sqrt{n}}, \tag{9}$$

where $\overline{X}_j$ is the mean value of variable j and $S_j^2 = \frac{1}{n}\sum_{l=1}^{n}(X_{lj} - \overline{X}_j)^2$ the corresponding sample variance. The matrix of Pearson's correlations is then $\mathbb{X}^T\mathbb{X}$. The analysis that follows consists in finding the eigenvectors and eigenvalues of this matrix. However, because if we have a relatively large number of variables and much lesser samples, this matrix will have a huge dimension and will not have full rank. It is complicated and very time consuming to diagonalize such a matrix and we propose to find the eigenvectors and eigenvalues of the matrix $\mathbb{X}\mathbb{X}^T$ instead, and from these find the ones we want. If x is a unit eigenvector of $\mathbb{X}\mathbb{X}^T$ and λ the corresponding eigenvalue, then

$$\mathbb{X}\mathbb{X}^T x = \lambda x.$$

Premultiplying this equation by $\mathbb{X}^T$ gives

$$\mathbb{X}^T\mathbb{X}\mathbb{X}^T x = \lambda \mathbb{X}^T x,$$

which means that $\mathbb{X}^T x$ is an eigenvector of $\mathbb{X}^T\mathbb{X}$ with the same eigenvalue λ. To normalize this eigenvector, let us find its norm as

$$\|\mathbb{X}^T x\|^2 = (\mathbb{X}^T x)^T\mathbb{X}^T x = x^T\mathbb{X}\mathbb{X}^T x = x^T\lambda x = \lambda,$$

because the norm of x is 1. Hence, $\|\mathbb{X}^T x\| = \sqrt{\lambda}$. Therefore, we conclude that if x is an unit eigenvector of $\mathbb{X}\mathbb{X}^T$ and λ the corresponding eigenvalue, then $\frac{1}{\sqrt{\lambda}}\mathbb{X}^T x$ is a unit eigenvector of the matrix $\mathbb{X}^T\mathbb{X}$ with the same eigenvalue λ. This is a very useful result because it allows us to find the eigenvectors of a very huge matrix $\mathbb{X}^T\mathbb{X}$, by diagonalizing a much smaller matrix $\mathbb{X}\mathbb{X}^T$.

A.2 Computation of the Weighted Principal Components

To compute the weighted principal components, that is, using the correlation coefficient r_{W2}, we start by transforming our data according to (5). Then, similarly to the previous section, we standardize the transformed data as

$$R'_{ij} \leftarrow \frac{R'_{ij} - \overline{R'}_j}{S_{Rj}\sqrt{n}}, \tag{10}$$

where $\overline{R'}_j$ is the mean value of the weighted ranks corresponding to variable j and $S_{Rj}^2 = \frac{1}{n}\sum_{l=1}^{n}(R'_{lj} - \overline{R'}_j)^2$ the corresponding sample variance. Hence, if $\mathbb{X}'$ represents the data matrix corresponding to these transformations, the matrix of weighted correlation coefficients (r_{W2}) is $\mathbb{X}'^T\mathbb{X}'$. As before, if this is a huge matrix, in order to obtain its diagonalization, we proceed exactly as for the unweighted case, diagonalizing first $\mathbb{X}'\mathbb{X}'^T$. Hence, if x is a unit eigenvector of $\mathbb{X}'\mathbb{X}'^T$ and λ the corresponding eigenvalue, then $\frac{1}{\sqrt{\lambda}}\mathbb{X}'^T x$ is a unit eigenvector of the matrix $\mathbb{X}'^T\mathbb{X}'$ with the same eigenvalue λ.

ACKNOWLEDGMENTS

The second author (Hugo Alonso) would like to thank the Fundação para a Ciência e a Tecnologia for the financial support during the course of this project. The second author has also been partially supported by Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA), Universidade de Aveiro, Portugal.

REFERENCES

[1] C. Ambroise and G.J. McLachlan, "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data," *Proc. Nat'l Academy of Sciences USA*, vol. 99, pp. 6562-6566, 2002.

[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press, 1999.

[3] S. Bicciato, A. Luchini, and C. Di Bello, "PCA Disjoint Models for Multiclass Cancer Analysis Using Gene Expression Data," *Bioinformatics*, vol. 19, pp. 571-578, 2003.

[4] A. Brazma and J. Vilo, "Gene Expression Data Analysis," *FEBS Letters*, vol. 480, pp. 17-24, 2000.

[5] J. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Proc. 14th Conf. Uncertainty in Artificial Intelligence*, pp. 43-52, 1998.

[6] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining*, vol. 2, pp. 121-167, 1998.

[7] A.R. Dabney, "Classification of Microarray to Nearest Centroids," *Bioinformatics*, vol. 21, no. 22, pp. 4148-4154, 2005.

[8] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Am. Statistical Assoc.*, vol. 97, pp. 77-87, 2002.

[9] W.J. Ewens and G.R. Grant, *Statistical Methods in Bioinformatics—An Introduction*, second ed. Springer, 2005.

[10] J. Faber, A.V. Krivtsov, M.C. Stubbs, R. Wright, T.N. Davis, M. van den Heuvel-Eibrink, C.M. Zwaan, S.A. Kung, and A.L. Armstrong, "Hoxa9 Is Required for Survival in Human MLL-Rearranged Acute Leukemias," *Blood*, vol. 113, no. 11, pp. 2375-2385, 2009.

[11] *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, V.J. Carey, W. Huber, R.A. Irizarry, and S. Dudoit, eds. Springer, 2005.

[12] M. Girolami and R. Breitling, "Biologically Valid Linear Factor Models of Gene Expression," *Bioinformatics*, vol. 20, pp. 3021-3033, 2004.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2002.

[14] M. Hubert and S. Engelen, "Robust PCA and Classification in Biosciences," *Bioinformatics*, vol. 20, pp. 1728-1736, 2004.

[15] R. Jagadeeswaran, H. Surawska, S. Krishnaswamy, V. Janamanchi, A.C. Mackinnon, T.Y. Seiwert, S. Loganathan, R. Kanteti, T. Reichman, V. Nallasura, S. Schwartz, L. Faoro, Y.C. Wang, L. Girard, M.S. Tretiakova, S. Ahmed, O. Zumba, L. Soulii, V.P. Bindokas, L.L. Szeto, G.J. Gordon, R. Bueno, D. Sugarbaker, M.W. Lingen, M. Sattler, T. Krausz, W. Vigneswaran, V. Natarajan, J. Minna, E.E. Vokes, M.K. Ferguson, A.N. Husain, and R. Salgia, "Paxillin Is a Target for Somatic Mutations in Lung Cancer: Implications for Cell Growth and Invasion," *Cancer Research*, vol. 68, no. 1, pp. 132-142, 2008.

[16] J.J. Jansen, H.C. Hoefsloot, H.F. Boelens, J. van der Greef, and A.K. Smilde, "Analysis of Longitudinal Metabolomics Data," *Bioinformatics,* vol. 20, pp. 2438-2446, 2004.

[17] K. Järvelin and J. Kekäläinen, "IR Evaluation Methods for Retrieving Highly Relevant Documents," *Proc. ACM SIGIR '00,* N. Belkin, P. Ingwersen, and M.-K. Leong, eds, pp. 41-48, 2000.

[18] I.T. Jolliffe, *Principal Component Analysis,* second ed. Springer, 2002.

[19] N. Katase, M. Gunduz, L. Beder, E. Gunduz, M. Lefeuvre, O.F. Hatipoglu, S.S. Borkosky, R. Tamamura, S. Tominaga, N. Yamanaka, K. Shimizu, N. Nagai, and H. Nagatsuka, "Deletion at Dickkopf (dkk)-3 Locus (11p15.2) Is Related with Lower Lymph Node Metastasis and Better Prognosis in Head and Neck Squamous Cell Carcinomas," *Oncology Research,* vol. 17, no. 6, pp. 273-282, 2008.

[20] H.A.L. Kiers, "Weighted Least Squares Fitting Using Ordinary Least Squares Algorithm," *Psychometrika,* vol. 62, pp. 251-266, 1997.

[21] K. Kondo, T. Kaneko, M. Baba, and H. Konno, "VEGF-C and VEGF-A Synergistically Enhance Lymph Node Metastasis of Gastric Cancer," *Biological and Pharmaceutical Bull.,* vol. 30, no. 4, pp. 633-637, 2007.

[22] L. Lebart, A. Morineau, and J.P. Fénelon, *Traitement des Données Statistiques—Méthodes et Programmes.* 2e éd. Dunod/BORDAS, 1982.

[23] P.P. Medina, J. Carretero, M.F. Fraga, M. Esteller, D. Sidransky, and M. Sanchez-Cespedes, "Genetic and Epigenetic Screening for Gene Alterations of the Chromatin-Remodeling Factor, SMARCA4/BRG1, in Lung Tumors," *Genes, Chromosomes and Cancer,* vol. 41, no. 2, pp. 170-177, 2004.

[24] P.T. Nelson, D.A. Baldwin, L.M. Scearce, J.C. Oberholtzer, J.W. Tobias, and Z. Mourelatos, "Microarray-Based, High-Throughput Gene Expression Profiling of Micrornas," *Nature Methods,* vol. 1, pp. 155-161, 2004.

[25] D. Pestana and S. Velosa, *Introdução à Probabilidade e à Estatística,* vol. 1, 2a ed. Fundação Calouste Gulbenkian, 2006.

[26] J. Pinto da Costa, H. Alonso, L.A.C. Roque, and M.M. Oliveira, "Supervised and Unsupervised Selection of Genes in Microarray Data," *Proc. Workshop Statistics in Genomics and Proteomics,* vol. 27, pp. 65-74, 2006.

[27] J. Pinto da Costa and L. Roque, "Limit Distribution for the Weighted Rank Correlation Coefficient, r_w," *REVSTAT—Statistical J.,* vol. 4, no. 3, pp. 189-200, Nov. 2006.

[28] J. Pinto da Costa and C. Soares, "A Weighted Rank Measure of Correlation," *Australian and New Zealand J. Statistics,* vol. 47, no. 4, pp. 515-529, 2005.

[29] J. Pinto da Costa and C. Soares, "Rejoinder to Letter to the Editor from C. Genest and J.-F. Plante Concerning Pinto da Costa, J. & Soares, C. (2005) A Weighted Rank Measure of Correlation," *Australian and New Zealand J. Statistics,* vol. 49, no. 2, pp. 205-207, 2007.

[30] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J. Kim, L. Goumnerovak, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califanokk, G. Stolovitzkykk, D. Louis, J. Mesirov, E. Lander, and T. Golub, "Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression," *Nature Neuroscience,* vol. 415, pp. 436-442, 2002.

[31] R Development Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 2005.

[32] G. Sanguinetti, M. Milo, M. Rattray, and N.D. Lawrence, "Accounting for Probe-Level Noise in Principal Component Analysis of Microarray Data," *Bioinformatics,* vol. 21, pp. 3748-3754, 2005.

[33] U.G. Sathyanarayana, A. Padar, C.X. Huang, M. Suzuki, H. Shigematsu, B.N. Bekele, and A.F. Gazdar, "Aberrant Promoter Methylation and Silencing of Laminin-5-Encoding Genes in Breast Carcinoma," *Clinical Cancer Research,* vol. 9, no. 17, pp. 6389-6394, 2003.

[34] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, and J. Selbig, "Metabolite Fingerprinting: Detecting Biological Features by Independent Component Analysis," *Bioinformatics,* vol. 20, pp. 2447-2454, 2004.

[35] D. Slonim, T. Golub, P. Tamayo, J.P. Mesirov, and E.S. Lander, "Class Prediction and Discovery Using Gene Expression Data," *Proc. Int'l Conf. Research in Computational Molecular Biology (RECOMB),* pp. 263-272, 2000.

[36] J. Terry, T. Saito, S. Subramanian, C. Ruttan, C.R. Antonescu, J.R. Goldblum, E. Downs-Kelly, C.L. Corless, B.P. Rubin, M. van de Rijn, M. Ladanyi, and T.O. Nielsen, "TLE1 as a Diagnostic Immunohistochemical Marker for Synovial Sarcoma Emerging from Gene Expression Profiling Studies," *The Am. J. Surgical Pathology,* vol. 31, no. 2, pp. 240-246, 2007.

[37] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proc. Nat'l Academy of Sciences USA,* vol. 99, pp. 6567-6572, 2002.

[38] V. Tusher, R. Tibshirani, and C. Chu, "Significance Analysis of Microarrays Applied to Ionizing Radiation Response," *Proc. Nat'l Academy of Sciences USA,* vol. 98, pp. 5116-5121, 2001.

[39] H. Wold and E. Lyttkens, "Nonlinear Iterative Partial Least Squares (NIPALS) Estimation Procedures," *Proc. 37th Session Bull. Int'l Statistical Inst.,* pp. 1-15, 1969.

[40] X. Xu, H. Yamamoto, M. Sakon, M. Yasui, C.Y. Ngan, H. Fukunaga, T. Morita, M. Ogawa, H. Nagano, S. Nakamori, M. Sekimoto, N. Matsuura, and M. Monden, "Overexpression of cdc25a Phosphatase Is Associated with Hypergrowth Activity and Poor Prognosis of Human Hepatocellular Carcinomas," *Clinical Cancer Research,* vol. 9, no. 5, pp. 1764-1772, 2003.

[41] T. Yu and K.-C. Li, "Inference of Transcriptional Regulatory Network by Two-Stage Constrained Space Factor Analysis," *Bioinformatics,* vol. 21, pp. 4033-4038, 2005.

[42] M. Yunta and P.A. Lazo, "Apoptosis Protection and Survival Signal by the cd53 Tetraspanin Antigen," *Oncogene,* vol. 22, no. 8, pp. 1219-1224, 2003.