

A weighted principal component analysis and its application to microarray data

Joaquim F. Pinto da Costa^{a,1}, Hugo Alonso^b, Luís Roque^c

^a Department of Applied Mathematics, University of Porto Science School

^b Research Unit in Mathematics and Applications, University of Aveiro

^c Research Group GECAD, Polytechnic of Porto Engineering School

Abstract: In this work we apply a new method of Weighted Principal Component Analysis (WPCA) to microarray data, which are characterized by having a very large number of variables (genes) relative to the number of samples. As we will show, this type of data has usually many outliers and noise and in order to cope with this problem, a nonparametric correlation coefficient, which uses ranks instead of the original observations, will be used. In addition, given that inside each variable (gene) the larger absolute expression values are more “responsible” for the problem in analysis, and should thus have higher weight, a new nonparametric weighted correlation coefficient will be used.

Keywords: Correlation, Principal Component Analysis, Microarray Data

1. Introduction

The usual Principal Component Analysis (PCA) (Jolliffe, 2002), which is widely used in the analysis of high-dimensional data, is not recommended in certain types of data, like for instance microarray data, because of two main reasons. Firstly, the usual PCA is very sensitive to the presence of outliers and noise in the data, given that it is based on Pearson’s correlation coefficient; as we will show, noise and outliers are very common in microarrays. Secondly, it gives the same importance to all observations inside each variable (gene in our application). Well, 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, provides information about its function and contribution to the development of the related problem (Brazma and Vilo, 2000; Slonim *et al.*, 2000; Hastie *et al.*, 2002).

In order to cope with the outliers and noise problem, we will use rankings instead of the original data; we will start therefore by ranking the observations in each variable (gene). Furthermore, in order to give higher weight to the larger absolute expression values, we will use a new weighted rank correlation coefficient instead of the usual Pearson’s.

¹Address of correspondence: Departamento de Matemática Aplicada, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

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 (PCA). If we start by standardizing the data, as it is common, the principal components 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 which assigns different weights to different observations in each variable according to the problem in analysis. In our application, the higher absolute values are given more importance.

To overcome the effect that noise and outliers have we will use the ranks of the observations instead of the original data. Then, if we calculate the Pearson's correlation coefficient of the ranked data, we obtain the Spearman's rank correlation coefficient, r_S , which is given by the expression (in case of no ties) $r_S = 1 - \frac{6 \sum_{i=1}^n (R_i - Q_i)^2}{n^3 - n}$, where R_i and Q_i are rank values. It is clear from this 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 (Pinto da Costa and Soares, 2005; Pinto da Costa and Roque, 2006; Pinto da Costa and Soares, 2007) some weighted coefficients of correlation are proposed which take rank importance into account. We will use here the new weighted rank correlation coefficient,

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

whose expression has been given in (Pinto da Costa and Soares, 2007), and propose to use it in PCA. It is easy to see that the computation of this new correlation coefficient r_{W2} is equivalent to do a data transformation to each variable as

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

(the same for Q'_i) and then compute the Pearson's correlation coefficient between R'_i and Q'_i . R_i represents the rank of each observation value and because we want to give higher weight to the larger absolute expression values, we will therefore attribute rank 1 to the largest absolute value, rank 2 to the second largest, etc. The weighted PCA can then be done using any common software for PCA analysis as long as we start by transforming our data according to (2). Our weighted principal component analysis (WPCA) is therefore very easy to apply because it only requires an initial data transformation.

Given that in our application the number of variables (genes) compared to the number of observations can be huge, we will apply a known and simple (yet efficient) way for computing the principal components whenever there are more variables than observations in the data. To compute the weighted principal components, that is, using the correlation coefficient r_{W2} , we start by transforming our data according to equation (2). Then, we standardise the transformed data as $R'_{ij} \leftarrow \frac{R'_{ij} - \bar{R}'_j}{S_{Rj} \sqrt{n}}$, where $\bar{R}'_j$ is the mean value of the weighted ranks corresponding to variable j and $S_{Rj}^2 = \frac{1}{n} \sum_{l=1}^n (R'_{lj} - \bar{R}'_j)^2$ 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 $\mathbb{X}^T \mathbb{X}$. Because 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. In order to obtain its diagonalisation

we proceed by diagonalising first $\mathbb{X}\mathbb{X}^T$, which is a much smaller matrix. Then, 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 .

3. Application of our method, WPCA, to microarray data

In this section, we will use the microarray datasets ² described below in Table 1. All gene values 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, where $\xi_{0.25}$ and $\xi_{0.75}$ are the first and third quartiles of the expression values for a given gene.

Dataset	Samples	Genes	Outliers	Severe outliers
Embryonal tumours	60	7129	6672	3553
Global cancer map	144	16063	15893	13073
Leukemia	38	7129	4713	1504

Table 1: Number of samples, genes and outliers in three microarray datasets.

We see therefore that in the first two datasets more than 90% of the genes contain outliers and more than 65% in the third dataset. 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}). We will use genes g19, g600, g830 of the first dataset, Embryonal Tumours, in our example which have 3, 5 and 6 outliers, respectively.

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

Table 2: Values of r and r_{W2} before and after the removal of outliers.

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 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 finalise this section, we will now find the expression of the first two principal components (which account for around 90% of the total variation in all four cases), before and after the removal of outliers to analyse the differences:

²The datasets can be found at the url <http://www.lsi.us.es/~aguilar/datasets.html>.

		r	r_{W2}
PC1	Before	$0.005g19 + 0.941g600 - 0.338g830$	$0.562g19 + 0.596g600 + 0.574g830$
	After	$0.004g19 + 0.874g600 + 0.486g830$	$0.547g19 + 0.593g600 + 0.591g830$
PC2	Before	$0.014g19 + 0.176g600 + 0.984g830$	$-0.771g19 + 0.125g600 + 0.625g830$
	After	$0.002g19 + 0.669g600 - 0.743g830$	$-0.832g19 + 0.366g600 + 0.417g830$

Table 3: Expression of the principal components for the two situations.

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 represent certainly 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.

4. Conclusions

In this paper, we introduced a new correlation coefficient that weighs observations according to their importance to the problem in hands; 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 principal component analysis (PCA), and concluded that its application to PCA is equivalent to do a certain data transformation. This gave rise to a novel weighted PCA, WPCA, which is more robust than the usual PCA in microarray datasets.

References

- Brazma A. and Vilo J. (2000) Gene expression data analysis, *FEBS Letters*, 480, 17–24.
- Hastie T., Tibshirani R. and Friedman J. (2002) *The Elements of Statistical Learning*, Springer.
- Jolliffe I. (2002) *Principal Component Analysis*, Springer, New York, 2nd ed.
- Pinto da Costa J. and Roque L. (2006) Limit distribution for the weighted rank correlation coefficient, r_w , *REVSTAT - Statistical Journal*, 4, 3.
- Pinto da Costa J. and Soares C. (2005) A weighted rank measure of correlation, *Australian & New Zealand Journal of Statistics*, 47, 4, 515–529.
- Pinto da Costa J. and Soares C. (2007) 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 & New Zealand Journal of Statistics*, 49, 2, 205–207.
- Slonim D., Golub T., Tamayo P., Mesirov J. and Lander E. (2000) Class prediction and discovery using gene expression data, *RECOMB*, 263–272.