Warsaw, 22.06.2021

Edge-group sparse PCA for network guided data analysis. Final project report for *Complex Biological Systems Modelling* at University of Warsaw

Author: Jakub Piotr Otręba

1. Introduction

Principal components analysis is one of the statistical methods of factor analysis. A data set consisting of N observations, each of which includes K variables, can be interpreted as a cloud of N points in a K-dimensional space. The aim of PCA is to rotate the coordinate system in such a way as to maximize first the variance of the first coordinate, then the variance of the second coordinate and so on. The coordinate values thus transformed are called the loads of the generated factors (principal components). In this way, a new observation space is constructed in which most of the variability is explained by the initial factors.

PCA is often used to reduce the size of a statistical dataset by discarding the last factors. It is also possible to look for a substantive interpretation of the factors, depending on the type of data, which allows a better understanding of the nature of the data, although it can be difficult with a larger number of variables under study. In signal processing, PCA is used, for example, for signal compression.

PCA can be based on either a correlation matrix or a covariance matrix formed from the input set. The algorithm in both versions is otherwise identical, but the results obtained are different. When using a covariance matrix, the variables in the input set with the largest variance have the largest impact on the result, which may be desirable if the variables represent comparable quantities, e.g., percentage changes in prices of different shares. The use of a correlation matrix, on the other hand, corresponds to an initial normalization of the input set so that each variable has an identical variance in the input, which may be desirable if the values of the variables are not comparable.

PCA and its variants can capture the linear relationship of variables to best explain the latent patterns of samples. However, the non-sparse principal component (PC) loadings by PCA employ all gene variables and lead to limited biological interpretability. Thus, variable selection is needed for gene expression analysis to select a small number of important genes. Recently, several studies have focused on developing sparse PCA models to encourage sparsity of PC loadings to extract gene modules with a limited number of genes for better interpretation (Wenwen Min, Juan Liu, Shihua Zhang, Edge-group sparse PCA for network-guided high dimensional data analysis, *Bioinformatics*, Volume 34, Issue 20, 15 October 2018, Pages 3479–3487).

For example, LASSO regularized, and elastic net regularized sparse PCA models have been proposed respectively. However, these sparse PCA models can not accurately control the sparse level of PC loadings. Thus, the sparse PCA model with L_0-penalty has been proposed to solve this issue. Several studies have developed different algorithms to solve these sparse PCA models. A kind of commonly used methods employ regularized SVD framework to solve sparse PCA models (Wenwen Min, Juan Liu, Shihua Zhang, Edge-group sparse PCA for network-guided high dimensional data analysis, *Bioinformatics*, Volume 34, Issue 20, 15 October 2018, Pages 3479–3487).

In my project, I used the code written by the researchers in the aforementioned scientific paper to test and compare three PCA methods: EPSCA, SPCA and classical PCA.

2. Materials and methods

2.1. Simulation study

In first, simulation study I generated two PC loadings by the following formulas:

- ***u1*** = [1, -1, 0.7, 0.1, -0.5, 0, 0, 0, 0, 0]
- ***u2*** = [0, 0, 0, 0, 0, 0.1, -2, -0.5, 0.3, 0.1]

And then I generated two PCs by the following formulas:

- ***v1*** = rnorm(100)
- ***v2*** = rnorm(100)

After these steps, the expression matrix ***X*** was created, along with the gene interactions network, which looked as follows:

- ***G*** = {(1,2),(1,3),(1,5),(2,3),(2,4),(3,4),(4,5),(6,7),(6,10),(7,8),(8,9),(8,10),(9,10)}

The top two identified PC1 and PC2 loadings by PCA, SPCA and ESPCA (Table 1.) are visible at the bottom of this paper in the Plots and Tables section. I then tested ESPCA and compared it with PCA and SPCA through artificially generated data. This data was constructed in such a way that the first five variables were associated with PC1 loading and the next five with PC2 loading. As can be seen in Table 1., ESPCA correctly extracted all five variables associated with PC1 loading and PC2 loading respectively. The SPCA appears to bypass the fourth variable for PC1 loading and the sixth variable for PC2 loading. As we can see, the absolute value of the signal of the fourth variable is smaller than that of the seventh variable. Because of this the SPCA does not find the fourth variable and mistakenly takes the seventh variable. However, the fourth variable is connected to the other three important variables in the G-connection network, so it is a more important variable than the seventh variable, which the ESCPA seems to catch, but the SPCA does not.

2.2. Biological applications

I used gene expression data from an article by Myron G. Best and colleagues (Best 2015) as biological data. The researchers' aim in this study was to distinguish platelets from healthy donors from those from cancer patients. To prepare the gene interaction network, I downloaded the ranks file (.rnk), which I used as input to GSEA. Gene Set Enrichment Analysis attempts to answer the question: can we summarize changes in gene expression as a list of altered pathways? Another input file was a gene set database file (.gmt), which in this case was the human s4et database culled from various sources, generated by the Bader lab. To do all this work, I used a library *fgsea,* which is a R package. After obtaining the gene interactions network, which is an input needed for ESPCA analysis, I was ready to compare these three methods again.

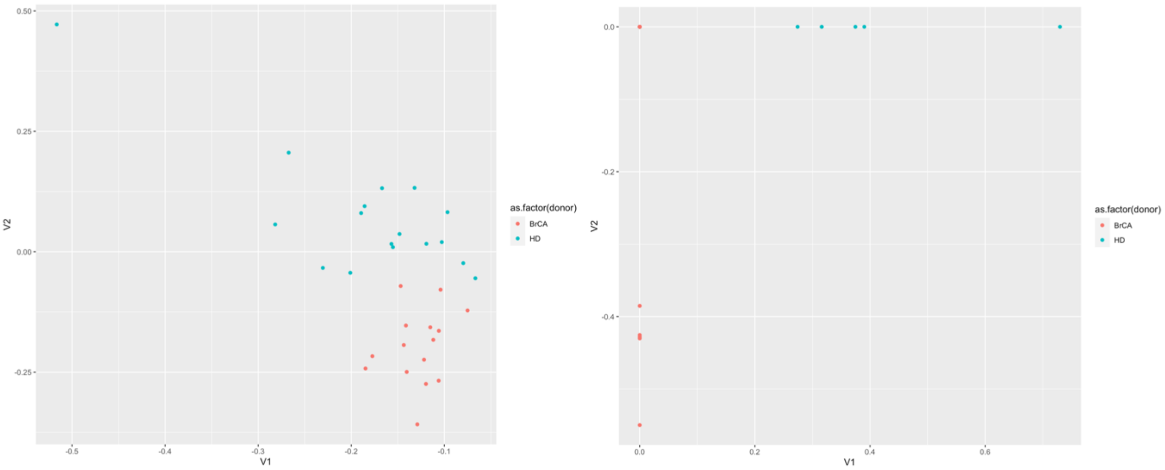
3. Results

I plotted two dimensional plots for all three methods of principal component analysis. Plots 1 and 2 show comparison between PCA and SPCA approach. On Plot 1 it is visible, that PCA managed to cluster cancer and healthy donors quit well. On Plot 2 SPCA did an excellent job in truly sparse analysis, ensuring the maximum sparsity in variables. Plot 3 shows how ESPCA did in my analysis. The ability to incorporate prior network knowledge allowed ESPCA to do sparse PCA as well as to choose the representation of very similar variables i.e., the cancer ones. ESPCA chose only one point coming from cancer data, which let us all acknowledge that cancer cells gene expression is more like each other than the level of expression among cells coming from healthy donors. The percent of variability explained by each PC is shown on Plot 3 and comparison between the first two of them is on Plot 4. Plot 5 shows a comparison of all three methods in terms of explaining the variability in the data. The black horizontal lines indicate the levels for ESPCA, making it easier to read the plot, which shows that it was ESPCA that fitted the data best.

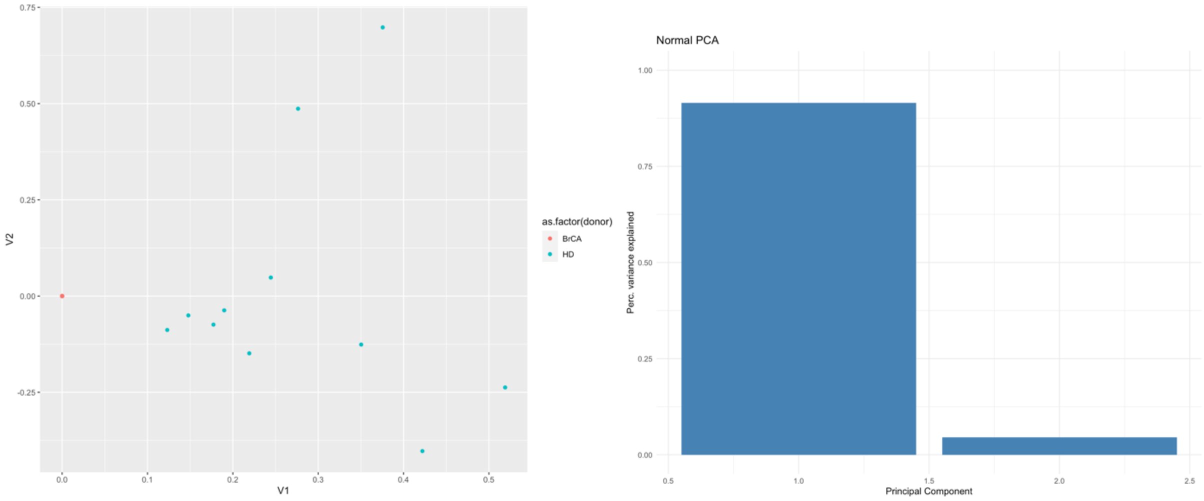
4. Discussion and conclusion

On one hand, sparse PCA is a typical unsupervised learning method for dimension reduction and feature selection. On the other hand, network-based methods for analysis have been employed to extract gene biomarkers models (Wenwen Min, Juan Liu, Shihua Zhang, Edge-group sparse PCA for network-guided high dimensional data analysis, *Bioinformatics*, Volume 34, Issue 20, 15 October 2018, Pages 3479–3487). One of the advantages of ESCPA is to ensure the obtained PCs are orthogonal. ESPCA is an 'offspring' of sparse PCA and the network-based methods for analysis to extract gene biomarkers. In my analysis, ESCPA did slightly better than SPCA and even better than normal PCA but it is not possible to state that ESCPA is the best of these three methods. In my analysis I did not change the hyperparameters and the dataset was one of smaller rather than larger ones. A larger dataset and playing with the hyperparameters would certainly increase the chances of answering the question: how much better ESPCA really is.

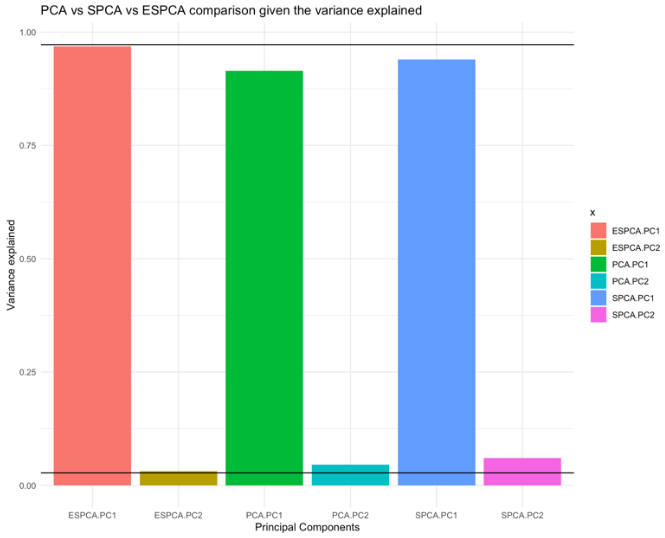
5. Plots and Tables



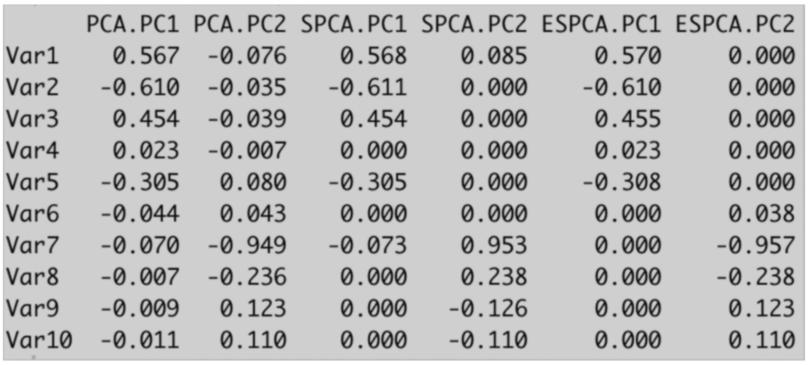
Plots 1 and 2



Plots 3 and 4



Plot 5



```
      PCA.PC1 PCA.PC2 SPCA.PC1 SPCA.PC2 ESPCA.PC1 ESPCA.PC2
Var1    0.567  -0.076    0.568    0.085     0.570     0.000
Var2   -0.610  -0.035   -0.611    0.000    -0.610     0.000
Var3    0.454  -0.039    0.454    0.000     0.455     0.000
Var4    0.023  -0.007    0.000    0.000     0.023     0.000
Var5   -0.305   0.080   -0.305    0.000    -0.308     0.000
Var6   -0.044   0.043    0.000    0.000     0.000     0.038
Var7   -0.070  -0.949   -0.073    0.953     0.000    -0.957
Var8   -0.007  -0.236    0.000    0.238     0.000    -0.238
Var9   -0.009   0.123    0.000   -0.126     0.000     0.123
Var10  -0.011   0.110    0.000   -0.110     0.000     0.110
```

Table 1