Principal Components Analysis: Theory and Application to Gene Expression Data Analysis

Hristo Todorov\(^1,2\), David Fournier\(^1\), Susanne Gerber\(^1\,*\)

\(^1\)Computational Systems Genetics, Faculty of Biology, Institute for Developmental Biology and Neurobiology (IDN) and Center for Computational Sciences in Mainz, Johannes Gutenberg-University Mainz, 55128, Germany
\(^2\)Fresenius Kabi Deutschland GmbH, Else-Kröner-Str. 1, 61352, Bad Homburg, Germany

* Correspondence: Staudinger Weg 9, Mainz, 55128, Germany; sugerber@uni-mainz.de

Received 2017-10-27; Accepted 2018-01-12

ABSTRACT

Advances in computational power have enabled research to generate significant amounts of data related to complex biological problems. Consequently, applying appropriate data analysis techniques has become paramount to tackle this complexity. However, theoretical understanding of statistical methods is necessary to ensure that the correct method is used and that sound inferences are made based on the analysis. In this article, we elaborate on the theory behind principal components analysis (PCA), which has become a favoured multivariate statistical tool in the field of omics-data analysis. We discuss the necessary prerequisites and steps to produce statistically valid results and provide guidelines for interpreting the output. Using PCA on gene expression data from a mouse experiment, we demonstrate that the main distinctive pattern in the data is associated with the transgenic mouse line and is not related to the mouse gender. A weaker association of the pattern with the genotype was also identified.

KEYWORDS

Principal components analysis; gene expression; exploratory data analysis

INTRODUCTION

Research in the field of computational biology often requires measuring a huge number of variables simultaneously. This makes exploratory data analysis challenging since visualization techniques are optimal only in two- or three-dimensional space. On the other hand, if variables are analyzed individually, important associations may be ignored or some calculations may be redundant due to overlapping variance [1].

Principal components analysis (PCA) is a statistical technique applied to complex data sets aiming at reducing the dimensionality of the data while simultaneously retaining the maximum amount of variance. Dimensionality reduction is achieved by creating a set of new variables called principal components (PC) which are linear combinations of the original variables. Usually, a small number of components are sufficient to capture most of the variability of the data set. Exploratory data analysis can then be applied to a subset of principal components rather than analyzing the larger number of initial variables. Herewith, the complexity of the analysis can be reduced.

Furthermore, PCA can be applied to investigating the relationship among the original variables. Principal components identify subsets of variables which are correlated with each other, thereby possibly uncovering meaningful patterns in the original data. Such patterns might easily be overlooked if multivariate techniques are not used, since a large number of variables usually prohibits the systematic investigation of all possible pairwise interactions or interactions of higher order, respectively. PCA may also facilitate uncovering underlying processes in the data set. In this context, principal components can be perceived as latent variables which cannot be directly measured. Latent processes are considered to be responsible for the correlations between the observed variables [1].

BASIC PRINCIPLES OF PCA

The input for PCA is a data matrix \(X\), in which the columns represent different variables and the rows correspond to values measured on the variables. For simplicity, consider the example in Table 1 which includes measurements on two variables. 100 data points (corresponding to different individuals) for each of the variables were generated using a bivariate normal distribution with a mean of 30 and 15, respectively, variances 10 and 2, respectively, and a Pearson correlation coefficient equal to 0.67 using the \texttt{mvrnorm()} R function. The maximal number of components in PCA is equal to the number of variables or the number of observations, whichever is smaller. Therefore, a maximum of two components can be extracted using the data in Table 1. The principal components are linear combinations of the original variables, i.e., a weighted sum of the input variables:

\[
PC1 = w_{11}Var1 + w_{12}Var2
\]

\[
PC2 = w_{21}Var1 + w_{22}Var2
\]

where \(w_i\) is a weighting coefficient. Importantly, the first principal component (PC1) points in the direction of the largest variance in the data set (Figure 1a). PC2 is orthogonal to PC1 and shows the direction of the

Publisher: Kernel Press. Copyright © (2018) the Author(s). This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Table 1: Simulated data set including two variables. 100 data points were sampled from a random bivariate normal distribution with means 30 and 15, variances 10 and 2 and a Pearson correlation coefficient equal to 0.67 for var1 and var2 with the help of the `mvrnorm()` R function. Scores for 10 random subjects are shown.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Var1</th>
<th>Var2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35.19</td>
<td>16.45</td>
</tr>
<tr>
<td>2</td>
<td>32.79</td>
<td>16.2</td>
</tr>
<tr>
<td>3</td>
<td>30.82</td>
<td>15.15</td>
</tr>
<tr>
<td>4</td>
<td>25.96</td>
<td>13.42</td>
</tr>
<tr>
<td>5</td>
<td>30.09</td>
<td>15.5</td>
</tr>
<tr>
<td>6</td>
<td>28.09</td>
<td>15.5</td>
</tr>
<tr>
<td>7</td>
<td>26.83</td>
<td>13.89</td>
</tr>
<tr>
<td>8</td>
<td>32.79</td>
<td>16.2</td>
</tr>
<tr>
<td>9</td>
<td>29.31</td>
<td>13.62</td>
</tr>
<tr>
<td>10</td>
<td>34.2</td>
<td>15.16</td>
</tr>
</tbody>
</table>

second greatest variance, and so forth, until the maximal number of components is reached. As PC1 usually captures the majority of variance, the original (potentially high-dimensional) data can be projected on a subset of a few essential principal components only, while higher dimensions can be discarded without a major loss of information. In our simplified example (see Figure 1b, 1c), PC1 already captured 83.5% of the original variance, therefore PC2 can be neglected. Methods for selecting the appropriate number of components are discussed below. Furthermore, the R script used to generate the tables and figures is provided with the article (Supplementary File 1).

ASSUMPTIONS OF PCA

Several prerequisites need to be met to produce meaningful results with PCA. Firstly, the input data need to be continuous, real-valued variables measured on an interval or ratio scale because standard PCA investigates patterns of covariance/correlation, which only makes sense for such variables. Appropriate methods for discrete variables measured on an interval scale such as integers or categorical variables include among others correspondence analysis [2], multiple correspondence analysis [2, 3] or non-metric multidimensional scaling [2]. However, these techniques will not be discussed in the context of this article. Secondly, covariance/correlation measures require that the relationship between each pair of variables is linear. If non-linear relationships are detected, appropriate data transformation techniques (e.g. logarithmic transformation) should be considered. Screening for outliers should be performed prior to the analysis, since outliers can affect the size of the covariance/correlation and thus distort the results. Finally, a sufficiently large sample size is needed to obtain more accurate estimates for the covariance/correlation population parameters, which leads to more robust PCA results [1].

Figure 1: Basic principles of PCA. (a) Scatter plot of the standardized variables shown in Table 1. Variables were standardized by subtracting the respective mean from each value and dividing the result by the standard deviation. The red arrow represents the direction of the largest variance of the multivariate Gaussian distribution derived from the data. The blue arrow captures the direction of the second largest variance orthogonal to the first vector. (b) A scatter plot of the two variables projected on the first and second principal components. The amount of variance accounted for by each component is given in brackets. (c) Dimension reduction in this example can be achieved by retaining only the first principal component. Since PC1 captures 83.5% of the variance in the original data, discarding PC2 does not lead to a major loss of information.
EXTRACTION OF PRINCIPAL COMPONENTS

The main step of PCA is the computation of the weighting coefficients that are needed to create the linear combinations of the original variables. The different possible approaches are all based on the principles of low-rank matrix factorization, i.e., on decomposing a matrix into a product of matrices of a smaller rank/dimension. A common approach uses the covariance matrix $S$ of the original variables to extract the principal components. $S$ is a symmetric matrix whose off-diagonal elements correspond to the covariance between pairs of variables in a data set. The entries in the main diagonal represent the variance of a variable. In a procedure called eigendecomposition, $S$ can be represented in the following way:

$$S = V S L V^T$$

with $V_S$ being the matrix of the normalized eigenvectors $v$ of $S$. $L_S$ is a diagonal matrix containing the corresponding eigenvalues and $V_S^T$ the transpose of $V_S$. Note that the eigenvector $v$ of a square matrix $M$ has the following property:

$$Mv = \lambda v$$

where $\lambda$ is the corresponding eigenvalue. The eigenvectors of $S$ are orthogonal to each other and the eigenvalues are non-negative.

The eigenvectors of the covariance matrix $S$ are then used as weighting coefficients to calculate the principal component scores:

$$F = X V_S$$

where $X$ is the matrix of mean-centered data and $F$ is a matrix whose columns contain the principal components. From a statistical point of view each principal component can be regarded as a new variable which can be used for statistical tests or data visualization. Therefore, every subject from the original data set has a specific value measured on each principal component, also called a principal component score. The first eigenvector $v_1$, corresponding to the largest eigenvalue $\lambda_1$, is used to calculate the first principal component. This procedure produces principal components with a mean equal to 0 and a variance equal to the corresponding eigenvalue. The sum of all eigenvalues is equal to the total variance in the original data matrix $X$. Since the first eigenvalue is the largest, the first principal component accounts for the biggest amount of variability in the original data.

An alternative approach of calculating the principal components is to use the eigenvectors of the correlation matrix $R$ as weighting coefficients. Recall that $R$ is a symmetric matrix whose off-diagonal elements represent the Pearson correlation coefficients between pairs of variables in a data set. The entries in the main diagonal are all equal to 1 because they correspond to the correlation of a variable with itself. This strategy is recommended when input variables are measured on different scales (otherwise variables with a larger scale dominate the size of the covariance). Component scores are calculated by multiplying the standardized data matrix $X$ with the matrix $V_R$ containing the eigenvectors of the correlation matrix $R$.

Finally, extraction of principal components may be facilitated by performing singular value decomposition of the original data matrix $X$:

$$X = U D V^T$$

where $U$ is a matrix of the so-called left singular vectors, $D$ is a diagonal matrix of singular values and $V$ contains the right side singular vectors [4]. Singular value decomposition is performed either on the mean centered (each column has a mean of 0) or the standardized data matrix, respectively, where each column has a mean of 0 and a variance of 1. If the mean-centered matrix is used, the matrix $V$ (i.e., the right singular vectors) is equal to the eigenvectors of the covariance matrix $S$. If, however, the standardized data matrix is used, then the right singular vectors correspond to the eigenvectors of the correlation matrix $R$. In both cases, the right singular vectors are used as weighting coefficients to calculate the principal component scores from the original variables.

In order to show a more detailed example of PCA, we extended the data shown in Table 1 with two additional variables. The new variables were sampled from a bivariate normal distribution with means 110 and 135, variances 20 and 33 and Pearson correlation coefficient 0.856. Since the scale of the variables was not specified, PCA was performed by singular value decomposition on the standardized data using the `prcomp()` R function. This produces identical results to performing eigendecomposition of the correlation matrix because `prcomp()` automatically rescales the singular values to the eigenvalues and reports the square root of the eigenvalues as standard deviation of the principal components. However, singular value decomposition is preferred due to numerical stability. The eigenvectors and the corresponding eigenvalues from our analysis on the simulated data set are shown in Table 2. This example demonstrates important characteristics of PCA, namely that the maximal number of components is equal to the number of original variables and that the first component accounts for the largest variance in the data.

<table>
<thead>
<tr>
<th>Eigenvectors</th>
<th>$V_1$</th>
<th>$V_2$</th>
<th>$V_3$</th>
<th>$V_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.3</td>
<td>0.64</td>
<td>-0.7</td>
<td>-0.12</td>
</tr>
<tr>
<td></td>
<td>-0.28</td>
<td>0.65</td>
<td>0.71</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>-0.66</td>
<td>-0.25</td>
<td>-0.07</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>-0.63</td>
<td>-0.33</td>
<td>0.09</td>
<td>-0.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eigenvalues (% variance captured)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.90</td>
</tr>
<tr>
<td>1.63</td>
</tr>
<tr>
<td>0.33</td>
</tr>
<tr>
<td>0.14</td>
</tr>
</tbody>
</table>

(48%) (41%) (8%) (3%)
PRINCIPAL COMPONENTS SELECTION

A major issue in PCA is to decide on how many components to retain in order to still keep a sufficient amount of variance but, at the same time, to achieve a substantial reduction in dimensionality. One possibility is to define a threshold (e.g. a certain percentage of the original variability) and keep as many principal components as necessary to exceed this threshold. Another option is to keep only principal components with a corresponding eigenvalue equal to or greater than 1, the so called Kaiser criterion. This, however, is only applicable if PCA is performed by eigendecomposition of $R$ or singular value decomposition of the standardized data matrix $X$. In this case, each original variable has a variance of 1. Therefore, only components which account for more variability than a single variable are meaningful and should be retained.

The probably most popular approach is the Scree plot [5], where the principal components are plotted on the x-axis in descending order against their corresponding eigenvalues. This leads to a decreasing function showing the variance explained by each PC. This plot often shows a clear crease (the so-called "elbow") separating the 'most important' components from the 'least important' components. All components to the right of the break point can be discarded. The disadvantage of this method is the visual inspection of the Scree plot - a subjective way to identify the correct number of principal components. Furthermore, in some practical applications, it might be difficult to determine the cut-off point where the slope of the line which goes through the eigenvalues changes the most.

A more sophisticated technique is parallel analysis [6]. Here, PCA is performed on a simulated data set with as many variables and cases as there are in the original data set. Averaged eigenvalues from the simulated data are compared to the eigenvalues obtained from the real data. Components from the real data with eigenvalues lower than the eigenvalues for the simulated data are discarded. Parallel analysis offers a more objective way to assess the appropriate number of components to keep. Therefore, it can be more useful than the Scree test in many real-world applications of PCA.

We applied the Scree test and parallel analysis on our example of PCA on the small simulated data set (Figure 2). Both techniques indicated that two components are meaningful, therefore the third and fourth principal components can be discarded.

INTERPRETATION OF PRINCIPAL COMPONENTS

Principal components are interpreted based on the original variables which “load” on them. Loadings correspond to correlations or covariances between the original variables and principal components. Variable loadings are stored in a loading matrix, $A$, which is produced by multiplying the matrix of the eigenvectors with a diagonal matrix containing the square root of the corresponding eigenvalues:

$$ A = V \sqrt{\Lambda} $$

The entries in $A$ are dependent on the technique used for extracting the components. If extraction is based on singular value decomposition of the matrix of mean-centered data or on eigendecomposition of the covariance matrix, then unstandardized loadings represent the covariance between mean-centered variables and standardized component scores. However, if eigendecomposition of the correlation matrix is performed, then standardized loadings are produced. These loadings represent correlations between the original variables and component scores. Standardized loadings are easier to interpret, since they always range from -1 to 1 and are independent of the scale used. A threshold is usually defined and only variables with loadings above this threshold are considered. A common rule of thumb suggests to only consider standardized loadings exceeding 0.45 (since this corresponds to 20% shared variance between the original variable and the principal component) [1].

The loading matrix for the PCA performed on our simulated data set with four variables and 100 cases is shown in Table 3. Since we used singular value decomposition of the standardized data matrix $X$, loadings represent correlations between principal components and original variables. PCA managed to capture the structural pattern of high correlation between var1 and var2 and between var3 and var4, respectively, in our simulated data.

VISUALIZATION OF RESULTS

PCA results are usually graphically represented by two- or three-dimensional dot plots of scores on the first few principal components. Each point represents
Table 3: Standardized loading matrix. The loading matrix is based on our PCA analysis of simulated data including 4 variables and 100 cases (see Table 1 and Table 2). Loadings on the retained first two PC (see Figure 2) represent correlations between the original variables and the principal components. Loadings below 0.45 were set to 0 to ease interpretation. The loading matrix reveals a structural pattern in the original data. Var3 and var4 both load significantly on PC1 whereas var1 and var2 load significantly on PC2, thus indicating an association between var1 and var2 on the one hand and between var3 and var4 on the other.

<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1</th>
<th>PC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var1</td>
<td>0.82</td>
<td>0.00</td>
</tr>
<tr>
<td>Var2</td>
<td>0.83</td>
<td>0.00</td>
</tr>
<tr>
<td>Var3</td>
<td>-0.91</td>
<td>0.00</td>
</tr>
<tr>
<td>Var4</td>
<td>-0.87</td>
<td>0.00</td>
</tr>
</tbody>
</table>

AN EXAMPLE OF PCA ON GENE EXPRESSION DATA

In order to show a practical example of PCA applied to genomic data, we analyzed gene expression data from brain tissue samples obtained from a transgenic mouse model and wild-type controls (GEO identifier: GSE47029). The study by Maung and colleagues tested whether genetic ablation of the CCR5 chemokine receptor impacts brain damage induced by HIV envelope protein gp120 in two transgenic mouse lines [7]. The input data matrix included 225 samples and 33503 protein features.

Gene expression levels have been shown to be heteroscedastic, meaning that their variance changes depending on mean expression level. Usually, genes with lower expression levels are associated with higher variability across different measurements [8]. The Poisson distribution which is often applied to modelling count data tends to underestimate the variance in gene expression data [8]. In contrast, the negative binomial (NB) distribution has been shown to provide a good fit when modelling differential gene expression [9].

The NB model assumes that an observation has a population mean \( \mu \) and variance \( \sigma^2 = \mu + \phi \mu^2 \), where \( \phi \) is the so-called dispersion parameter [9]. To account for the heteroscedasticity of the input data, we estimated the dispersion parameter for each gene with the help of the DESeq R package. Afterwards, we applied the varianceStabilizingTransformation() function to the input data in order to produce a data matrix in which expression levels are homoscedastic. Finally, we extracted the principal components from the transformed data using the plotPCA() function. The plotPCA() function internally calls on prcomp(), meaning that PCA was performed by singular value decomposition of the matrix of mean centered, variance stabilized gene expression data. With default settings plotPCA() extracts the components using only the top 500 most variable genes. Therefore, the input data matrix for the example shown here consisted of 225 rows (samples) and 500 columns (genes).

PCA identified two distinct clusters in the data separated based on scores on the first principal component (Figure 3). Different colouring schemes were applied to investigate if the separation was associated with a specific factor. This analysis revealed that the two clusters correspond to mouse line 1 and 2 (Figure 3a), suggesting an overall differential gene expression between the two transgenic mouse lines. Maung et al., who conducted the study the data were obtained from, reported an increased gp120 RNA expression in animals from line 2 compared to line 1. Furthermore, an up-regulation associated with gp120 was observed for six genes (CCR5, CCL2, CCL3, CCL4, CXCL10 and C4b) in mouse line 2 relative to line 1. These findings are in agreement with the two distinct clusters corresponding to both mouse lines revealed by our PCA (Figure 3a).

No differences could be determined based on gender (Figure 3b), however a certain discrimination associated with genotype was present based on scores on the second principal component (Figure 3c). Notably, Maung and colleagues identified a core common set of genes for the gp120+ genotype in mouse line 1 and line 2 which were differentially expressed in the presence or absence of the CCR5 receptor. The multivariate genotype pattern which we observed within both clusters corresponding to mouse line 1 and 2 is very similar and might therefore mainly relate to the core set of genes identified by Maung et al. A detailed inspection of the loading matrix could provide even more insights into which specific genes are differentially expressed based on genotype.

PC1 and PC2 together only accounted for 24.03% of the total variance which means that additional components are important and may be associated with multivariate patterns. However, our goal here is not to provide a complete analysis of the gene expression data set, but more to demonstrate how PCA can be applied to this type of data for exploratory purposes. Since a multivariate pattern already emerged based on PC1 and PC2, we did not further evaluate the remaining components.

The R script used to perform the PCA analysis on gene expression data is included in Supplementary File 2.

ACKNOWLEDGEMENTS

The work of SG and DF was funded by the Center for Computational Sciences in Mainz (CSM). The work of HT was funded by Fresenius Kabi Deutschland GmbH.
Figure 3: PCA analysis on gene expression data. PCA was performed on transformed gene expression data from a mouse model. PCA identified two distinct clusters in the data separated along the first principal component. Different colouring schemes revealed that clustering was associated with the mouse line (a). Gender was not associated with a pattern in the data (b), but genotype revealed a gradient of discrimination along the second principal component (c). The amount of variance in percent accounted for by each PC is included in brackets.

AUTHOR CONTRIBUTIONS

HT drafted the manuscript and performed the analysis on the simulated data. DF performed the analysis on the gene expression data and reviewed the manuscript. SG initiated the study, contributed to interpreting the results and editing the manuscript. All authors read and approved the final manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

SUPPLEMENTARY DATA

High resolution figure files, together with supplementary items listed below, are available at Genomics and Computational Biology online.

Supplementary File 1. R script for simulated data; File: simdataPCA.R. This file includes the R script used to generate the simulated data and perform the example PCA.

Supplementary File 2. R script for gene expression data; File: expressionPCA.R. This file includes the R script used to process the gene expression data and perform PCA.

ABBREVIATIONS

F : female
GEO : gene expression omnibus

HIV : human immunodeficiency virus
M : male
NB : negative binomial
PCA : principal components analysis
PC : principal component
WT : wild type

gp120 : envelope glycoprotein gp120

REFERENCES

8. Anders S, McCarthy D, Chen Y, Okoniewski M, Smyth G, Huber W,