Principal Component Analysis

Jake Sauter

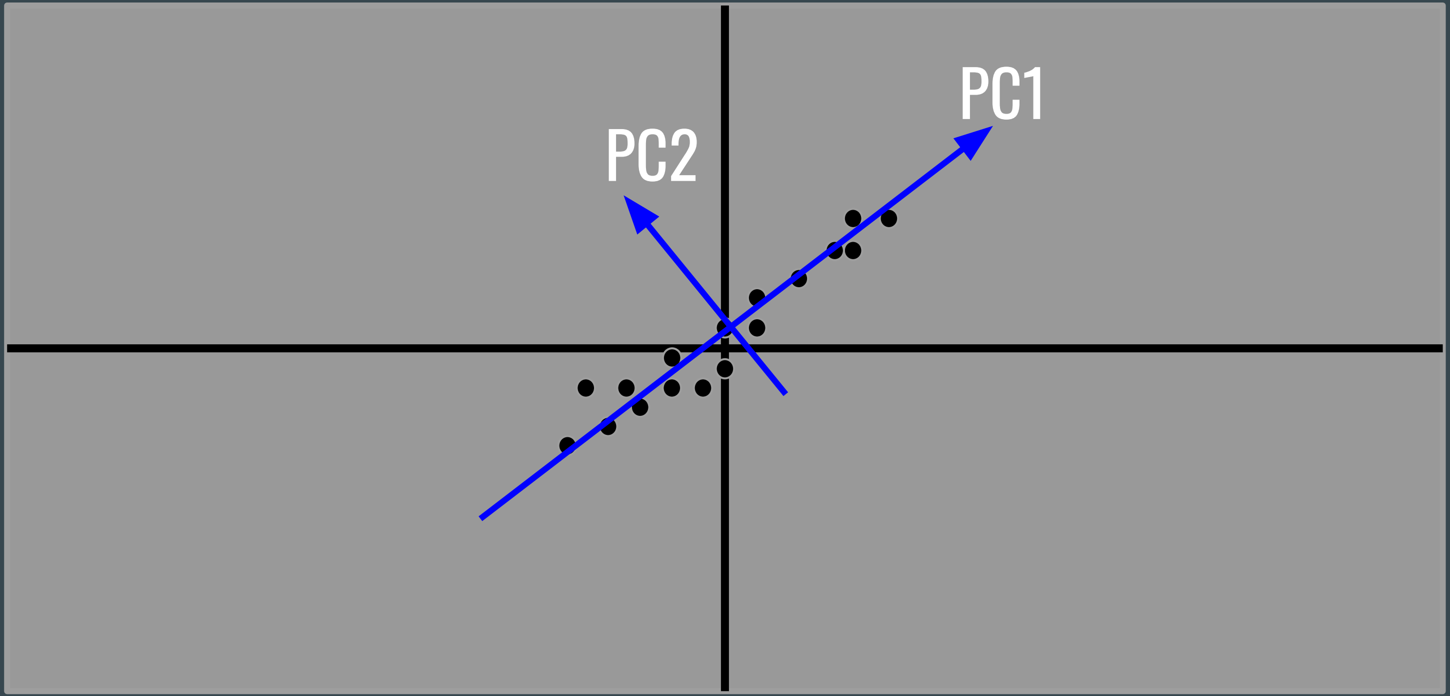
Motivation

- In some data sets, especially in DNA microarrays, there can be thousands of features per sample
- It is difficult to classify samples and isolate which of these features is most important in such a high dimensional feature space
- A form of dimensionality reduction would be very helpful, in which the dimensions of the data set are reduced, while the majority of variance and the relation of samples in the space is preserved
- Principal Component Analysis (PCA) is a form of dimensionality reduction

Principal Components

- The result of PCA is a list of principal components, or the most varying directions of the data set, such that all principal components are orthogonal
- A principal component can be realized as a weight of values for each feature
- Principle components are named in order of significance, being how much variability of the data they make up (PC1 , PC2 , ... , PCN)
- PCs can aid in data visualization, by plotting the data points on the new axes of the principal components

Visualization



Background - Covariance Matrix

- A covariance matrix of a data set is a matrix that expresses how each feature varies with every other feature
  - As such it captures the shape of the data set

$$\begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \dots & cov(x_1, x_n) \\ cov(x_1, x_2) & cov(x_2, x_2) & \dots & cov(x_2, x_n) \\ . & & & . \\ . & . & & . \\ . & & . & . \\ cov(x_1, x_n) & \dots & & cov(x_n, x_n) \end{bmatrix}$$

PCA - How it Works

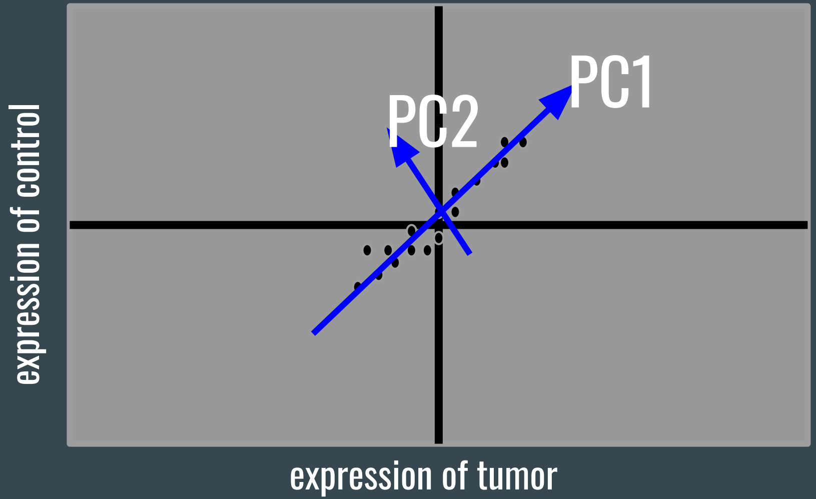
- PCA is a procedure that uses an orthogonal transformation to reduce dimensions, and this is done by finding the eigenvectors of the covariance matrix of the data
- If *A* is the covariance matrix of the data, then the eigenvector z_1 and eigenvalue λ_1 would form the transformation

$$Az_1 = \lambda_1 z_1$$

- The eigenvalue with the largest absolute value will indicate that the data have the largest variance along its vector

Caution

- In Microarray analysis and other fields, variance along one axis may be expected
- In DNA Microarrays, this expected variance comes along in the form of expression level, it is known that different genes will express at different levels, but we are interested in the **ratio** of the expression levels in comparative analysis



PCA in R

- Two functions are available for PCA in R, being **prcomp()** and **princomp()**
  - Difference in how PCA is calculated
- **Princomp()** follows the method that we have described
- **Prcomp()** uses a process called singular value decomposition (svp), with very little difference in output, and is a preferred computational method

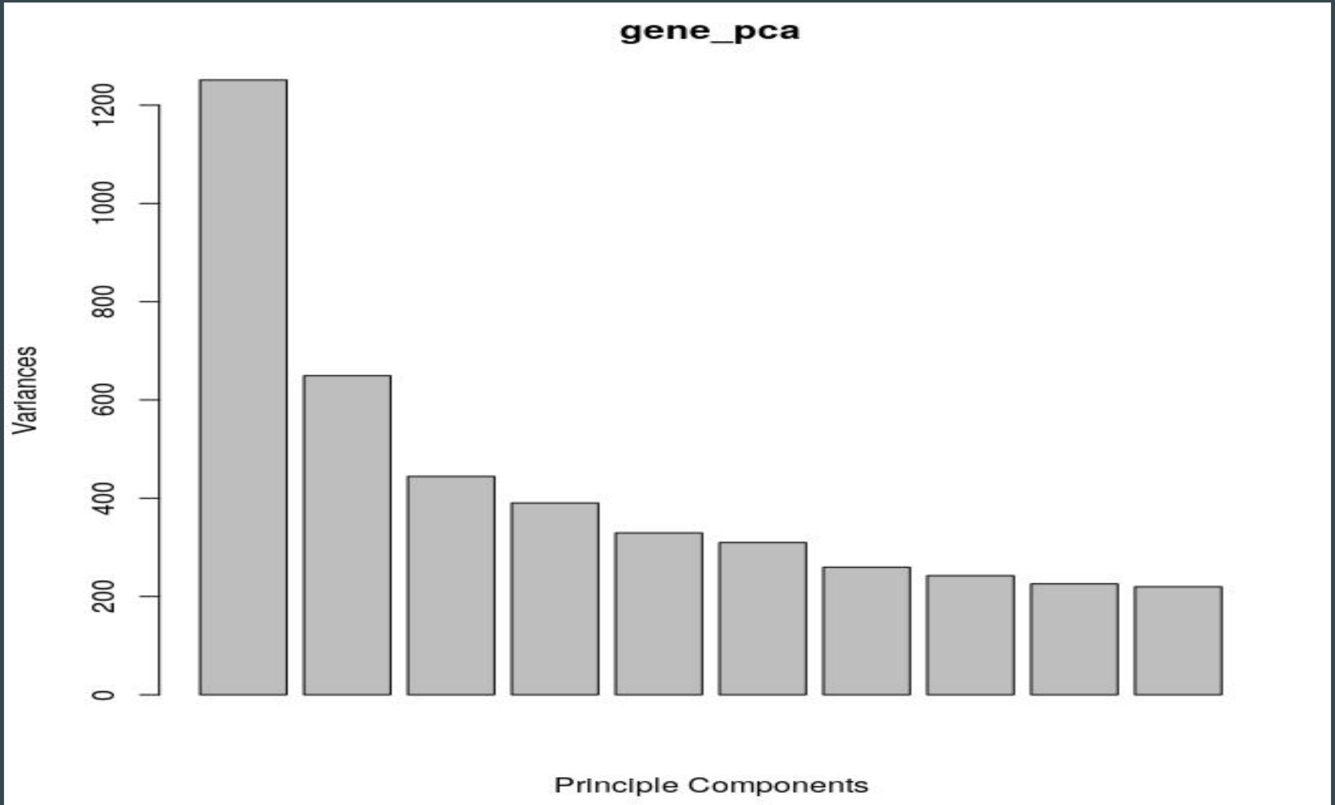
PCA in R

- **Prcomp()** options:
  - center: 0-center the data
  - scale: scale the data to have unit variance
- Usage:

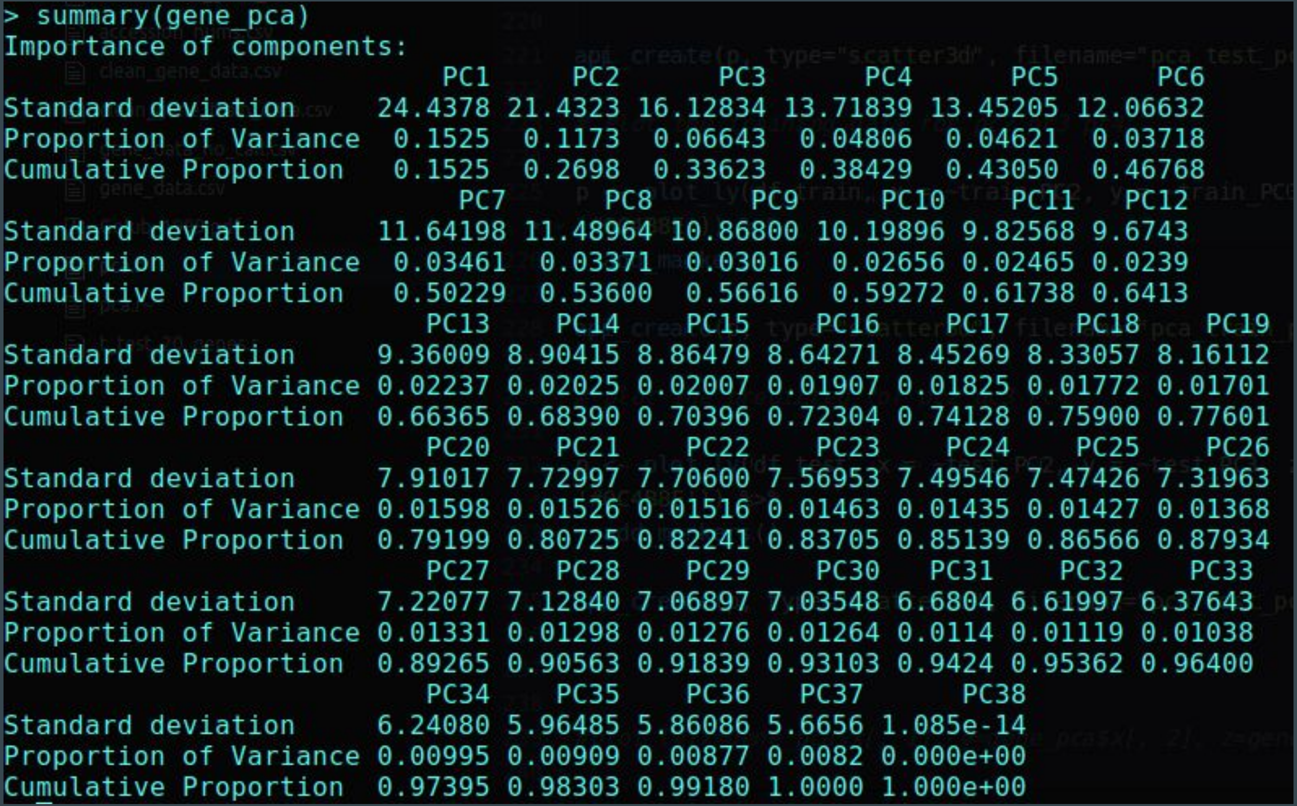
```
gene_pca = prcomp(cleaned_data, scale=T, center=T)
```

where cleaned_data is a matrix of samples x features

PCA Results


gene_pca
Variances
0
200
400
600
800
1000
1200
Principle Components

PCA Results



```
> summary(gene_pca)
Importance of components:
                           PC1     PC2      PC3      PC4      PC5      PC6
Standard deviation     24.4378 21.4323 16.12834 13.71839 13.45205 12.06632
Proportion of Variance  0.1525  0.1173  0.06643  0.04806  0.04621  0.03718
Cumulative Proportion   0.1525  0.2698  0.33623  0.38429  0.43050  0.46768
                            PC7      PC8      PC9     PC10    PC11   PC12
Standard deviation     11.64198 11.48964 10.86800 10.19896 9.82568 9.6743
Proportion of Variance  0.03461  0.03371  0.03016  0.02656 0.02465 0.0239
Cumulative Proportion   0.50229  0.53600  0.56616  0.59272 0.61738 0.6413
                          PC13    PC14    PC15    PC16    PC17    PC18    PC19
Standard deviation     9.36009 8.90415 8.86479 8.64271 8.45269 8.33057 8.16112
Proportion of Variance 0.02237 0.02025 0.02007 0.01907 0.01825 0.01772 0.01701
Cumulative Proportion  0.66365 0.68390 0.70396 0.72304 0.74128 0.75900 0.77601
                          PC20    PC21    PC22    PC23    PC24    PC25    PC26
Standard deviation     7.91017 7.72997 7.70600 7.56953 7.49546 7.47426 7.31963
Proportion of Variance 0.01598 0.01526 0.01516 0.01463 0.01435 0.01427 0.01368
Cumulative Proportion  0.79199 0.80725 0.82241 0.83705 0.85139 0.86566 0.87934
                          PC27    PC28    PC29    PC30   PC31    PC32    PC33
Standard deviation     7.22077 7.12840 7.06897 7.03548 6.6804 6.61997 6.37643
Proportion of Variance 0.01331 0.01298 0.01276 0.01264 0.0114 0.01119 0.01038
Cumulative Proportion  0.89265 0.90563 0.91839 0.93103 0.9424 0.95362 0.96400
                          PC34    PC35    PC36   PC37      PC38
Standard deviation     6.24080 5.96485 5.86086 5.6656 1.085e-14
Proportion of Variance 0.00995 0.00909 0.00877 0.0082 0.000e+00
Cumulative Proportion  0.97395 0.98303 0.99180 1.0000 1.000e+00
```

Plotly Library

- **Plotly** is a free online tool for plotting
- They have developed an R library to allow easy interfacing
- Setup is quick and easy! All one has to do is make an account and set username and password environmental variables in R
- Then a plot can be created through the **plot_ly()** command and posted to an account through the **api_create()** command

```
    p <- plot_ly(df, x = ~PC2, y = ~PC3, z = ~PC4, color = ~group, colors =
c('#BF382A', '#0C4B8E')) %>% add_markers()

    api_create(p, type="scatter3d", filename="scatter3d-my_8")
```

Plotly Output

- The plotting of training data on PC1, PC2 and PC3
- The plotting of training data on PC2, PC3, and PC4
- The plotting of testing data on PC1, PC2 and PC3 generated from training data
- The plotting of testing data on PC2, PC3 and PC4 generated from training data

References

[1] Drăghici Sorin. Statistics and Data Analysis for Microarrays: Using R and Bioconductor. Chapman and Hall, 2012.