Omics Central

Amrit Singh

2020-01-24

Contents

Chapter 1

Rationale

This project was developed in order to create a resource warehouse for researchers analyze omics datasets of various types such as transcriptomics, proteomcs, metabolomics. I expect this resource to grow as others contribute to it. Think of it as an awesome-resource github repo but in a bookdown format. However, since this book is meant as documentation to the omics central web application, adding new methods will require pull requests to the omics central web app repos (omics-central-frontend, omics-central-backend and omics-central-docker) and bookdown repos (omics-central-learn and omics-central-contribute).

The purpose of this book is not to copy, paste other works but to link works by different authors in one place and explain concepts through the lens of an omics researcher.

Chapter 2

Introduction

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 2. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```r
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

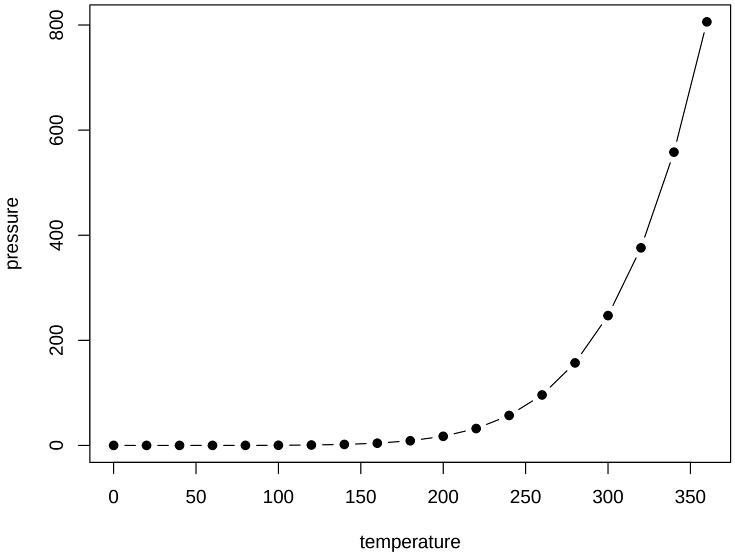


Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

```r
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2019) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

Chapter 3

Data-types

3.1 Microarrays

3.2 RNA sequencing

3.3 Nanostring

3.4 Biocrates

3.5 Multiple Reaction Monitoring

Chapter 4

Exploratory Data Analysis

4.1 Principal Component Analysis

//TODO insert video of EDA using Omics Central here

4.1.1 Method

```
## select dataset
X <- X.trainList$clinical

## constants
n <- length(hosp_3months) # number of samples
k <- 5 # number of PCs
p <- ncol(X) # number of variables

## run Principal Component Analysis
pca <- prcomp(X, scale. = TRUE)
```

> prcomp() performs Principal Component Analysis (PCA) on the clinical dataset X (n observations (samples) x p variables).

4.1.1.1 What is PCA?

- method to turn a dataset with correlated variables into another dataset with linearly uncorrelated variables called principal components (PCs).

4.1.1.2 Why is PCA useful?

- The first few PCs capture most of the variability in the data.
- PCA can be used to visualize clustering patterns (samples or variables) in the data, determine relationships between samples (see Principal Component plot), between variables (see Correlation circle), between samples and variables (see Biplot).
- PCA is also useful in determining the influece of covariates, both techincal (*e.g.* batch effects) or biological (*e.g.* sex).

4.1.1.3 What is a principal component (PC)?

- a PC is a weighted average of the original predictors, $\mathbf{PC}_i = \mathbf{X}\mathbf{v}_i$, where $\mathbf{X}$ is a centered matrix and *i=1,…,n*.

4.1.1.4 What do the vector of weights $\mathbf{v}_i$ do?

- v_i maximizes the variance; $\mathbf{X^TX}$ and are called eigenvectors, weights or loadings.

4.1.1.5 How do I compute the vector of weights, v_i?

- apply a factorization method called singular value decomposition (SVD). SVD decomposes a matrix X into a product of 3 matrices, $\mathbf{UDV^T}$; $\mathbf{X}_{np}$ = $\mathbf{U}_{nxp}$ x $\mathbf{D}$~*pxp*~ x $\mathbf{V^T}_{pp}$ or $\mathbf{X^TX} = \mathbf{VD^2V^T}$.
- The columns of $\mathbf{V}$ are the weights/loadings for each principal component.
- $\mathbf{D}$ is a diagnoal matrix where entry $\mathbf{D}_{i,i}$ is the standard deviation of the *ith* principal component (PC).
- Only the first k PCs are needed to capture the majority of the variation in the high dimensional dataset ($n << p$ and $k << p$); $\mathbf{X}_{nk} = \mathbf{U}_{nxk}$ x $\mathbf{D}_{pxk}$ x $\mathbf{V^T}_{nk}$ such that $\mathbf{X}_{nk} \approx \mathbf{X}_{np}$.

4.1.1.6 Why scale the data before applying PCA?

- The clinical variables are on different unit scales (*e.g.* Age (years) *vs.* Ejection fraction (%)). Scaling makes the mean of each variable zero and the standard deviation one.

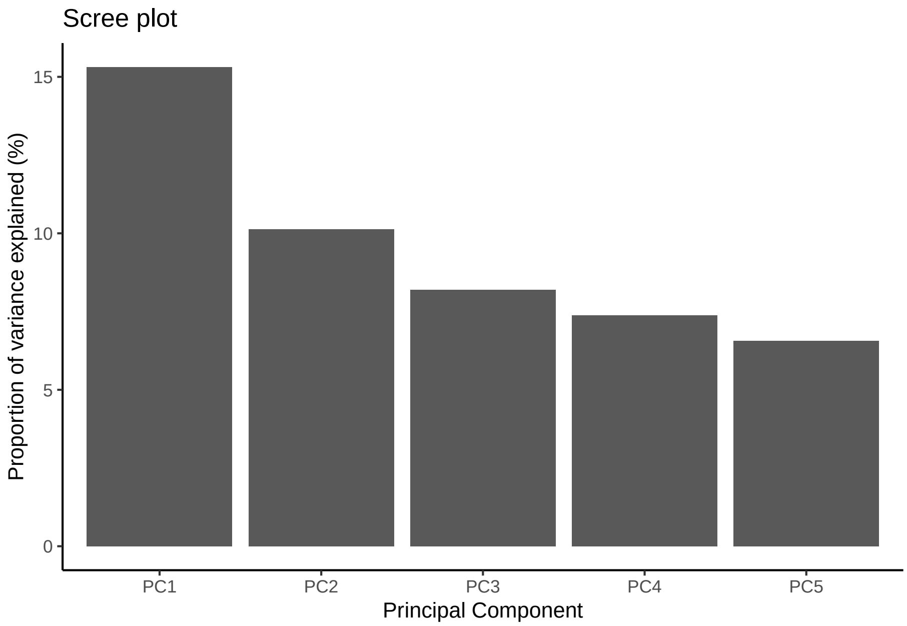
References
1. page 64-66 from ESL: https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print10.pdf
2. Wikipedia: https://en.wikipedia.org/wiki/Principal_component_analysis

4.1.2 Visualizations

4.1.2.1 Scree plot

- determine the proportion of variation explained by each principal component.

```
propVar <- (100 * pca$sdev^2/sum(pca$sdev^2))[1:k]
data.frame(var = propVar,
  comp = factor(paste0("PC", 1:k), paste0("PC", 1:k))) %>%
  ggplot(aes(comp, var)) +
  geom_bar(stat = "identity") +
  xlab("Principal Component") +
  ylab("Proportion of variance explained (%)") +
  ggtitle("Scree plot") +
  theme_classic()
```

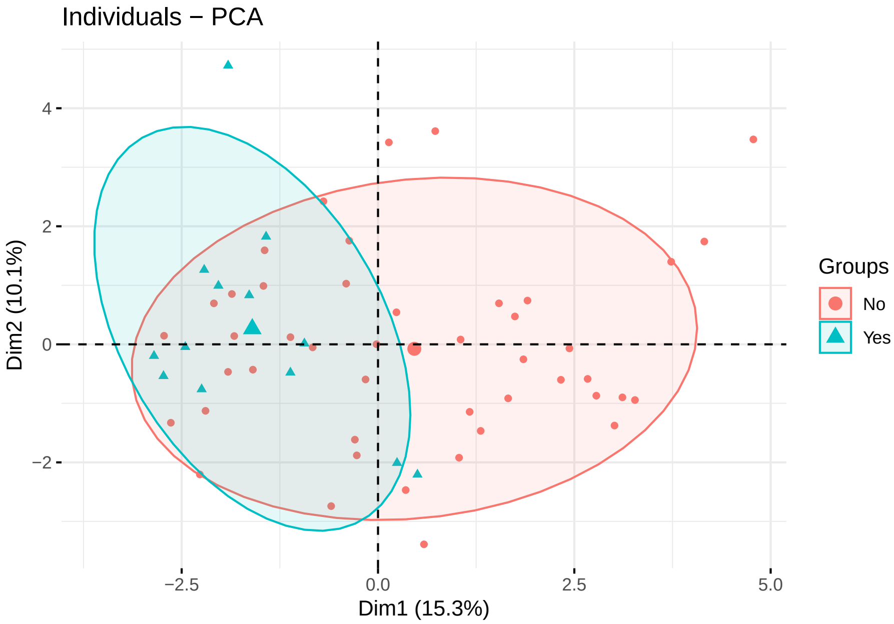


The barplot depicts the proportion of variation that is captured by the first five PCs; the first PC captures ~15.3% of the variability in the dataset consisting of 26 variables.

4.1.2.2 Component plot

- visualize the clustering of the samples and identify any clustering with respect to covariates of interest.

```r
## plot pca$x
fviz_pca_ind(pca, label="none", habillage=hosp_3months,
  addEllipses=TRUE, ellipse.level=0.80)
```



The scatter plot above is a 2D depiction of a 26 (# of clinical variables) dimensional dataset. PC1 and PC2 together capture 25.4% of the variability in the clinical dataset. Some separation between the groups of interest can be observed.

4.1.2.3 Correlation Circle

- determine relationship between variables (based on the correlation between each variable and PCs).
- the angle (θ) between two vectors determines the correlation between the two variables:

- θ=0: postive correlation (corr=1)

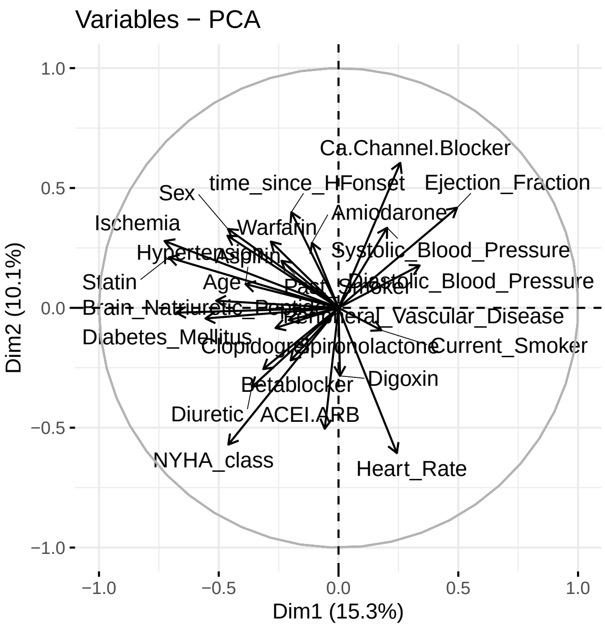
- 0<θ<90: postive correlation

- θ=90: zero correlation

- 90<θ<180: negative correlation

- θ=180: negative correlation (corr=-1)

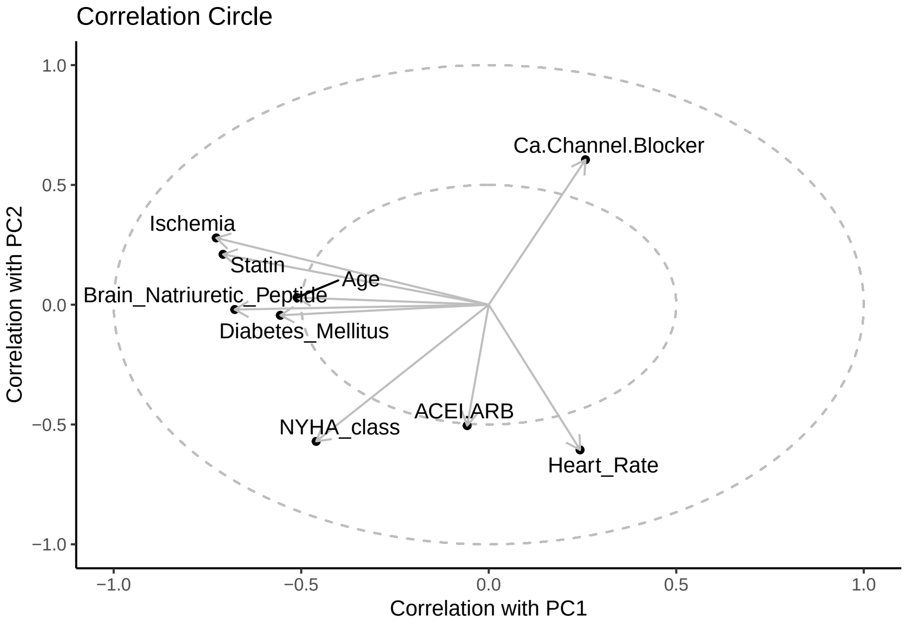
```
fviz_pca_var(pca, repel = TRUE)
```



4.1.2.4 Correlation Circle (with a cut-off)

```
# compute correlation between variables and PCs
corr <- t(cor(pca$x[, 1:2], X))

# correlation circle
cor_cutoff <- 0.5
corr %>%
  as.data.frame() %>%
  mutate(label = rownames(.)) %>%
  filter(abs(PC1) > cor_cutoff | abs(PC2) > cor_cutoff) %>%
  ggplot() +
  geom_circle(aes(x0 = 0, y0 = 0, r = 1), color = "gray", linetype="dashed") +
  geom_circle(aes(x0 = 0, y0 = 0, r = cor_cutoff), color = "gray", linetype="dashed") +
  geom_point(aes(x = PC1, y = PC2)) +
  geom_segment(aes(x = 0, y = 0, xend = PC1, yend = PC2), color = "gray", arrow = arrow(length =
  geom_text_repel(aes(x = PC1, y = PC2, label = label)) +
  xlab("Correlation with PC1") +
  ylab("Correlation with PC2") +
  theme_classic() +
  ggtitle("Correlation Circle")
```



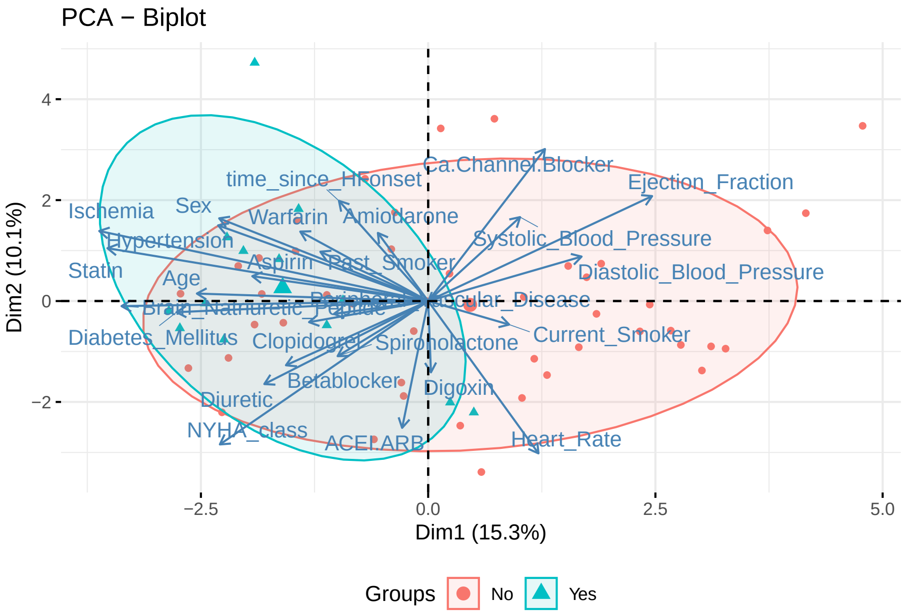
The above plot only displays the variables if they have a correlation greater than 0.5 with either PC1 or PC2. Ischemia and Statins are positively correlated suggesting that patients with ischemia are likely to be on statins. BNP (Brain Natriuretic Peptide) is positively correlated with age and negatively correlated with Heart Rate.

References
1. Figure 1 from BioData Mining volume 5, Article number: 19 (2012)
2. plotVar(): mixOmics R-library 3. fviz_pca_var(): factoextra R-library

4.1.2.5 Biplot

- superimpose the principal components with loadings vectors.

```
fviz_pca_biplot(pca,
      geom = "point",
      col.ind = "cos2",
      habillage=hosp_3months,
      repel = TRUE,
      addEllipses=TRUE, ellipse.level=0.80) +
      theme(legend.position="bottom")
```



Each arrow can be thought of as an axis. For example, BNP points to the left which means that patients on the left (PC1 $<$ 0) have lower BNP levels than patients on the right (PC1 $>$ 0). Patients at the center (PC=1) have an average BNP level. Note that this aligns well with the hospitalization status; *ie.* patients on the left are more likely to be hospitalized as compared to patients on the right.

References
1. ggbiplot(): https://github.com/vqv/ggbiplot
2. Biplot: https://stackoverflow.com/questions/6578355/plotting-pca-biplot-with-ggplot2
3. biplot(): K. R. Gabriel (1971). The biplot graphical display of matrices with application to principal component analysis. Biometrika, 58, 453–467. doi: 10.2307/2334381.
4. fviz_pca_biplot(): http://www.sthda.com/english/wiki/fviz-pca-quick-principal-component-analysis-data-visualization-r-software-and-data-mining

4.2 K-Means

4.3 Hierarchical clustering

4.4 Sample plots

4.4.1 Sample correlation heatmap

4.4.2 Sample histograms

4.5 Variable plots

Chapter 5

References

1. BioData Mining volume 5, Article number: 19 (2012)

2. PH525x series: http://genomicsclass.github.io/book/
3. mixOmics: https://mixomicsteam.github.io/Bookdown

Chapter 6

Batch Correction

6.1 ComBat

6.2 Surrogate Variable Analysis

6.3 Model adjustment

6.4 References

1. Batch effect simluations: http://jtleek.com/svaseq/simulateData.html
2. Surrogate Variable Analysis: https://bioconductor.org/packages/release/bioc/vignettes/sva/inst/doc/sva.pdf

Chapter 7

Differential Expression Analysis

7.1 Ordinary Least Squares

7.2 LInear Models for MicroArrays and RNA-Seq

7.2.1 Robust LIMMA

7.2.2 LIMMA VOOM

7.3 Significance Analysis for Microarrays (SAM)

7.4 cell-specific Analysis for Microarrays (csSAM)

Chapter 8

Network Analysis

8.1 DINGO

8.2 WGCNA

8.3 PANDA

8.4 BioNetStat

Chapter 9

Data Integration

9.1 Supervised

9.1.1 DIABLO (SGCCDA)

9.1.2 Ensemble of glmnet classifiers

9.1.3 DIABLO2 (sMB-PLSDA)

9.2 Unsupervised

9.2.1 PANDA

9.2.2 MOFA

9.2.3 JIVE

9.2.4 SNF

Chapter 10

Biological Enrichment

10.1 Enrichr

10.2 SEAR

10.3 CAMERA

10.4 Network-based Gene Set Analysis

Chapter 11

Literature Mining

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2019). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.16.