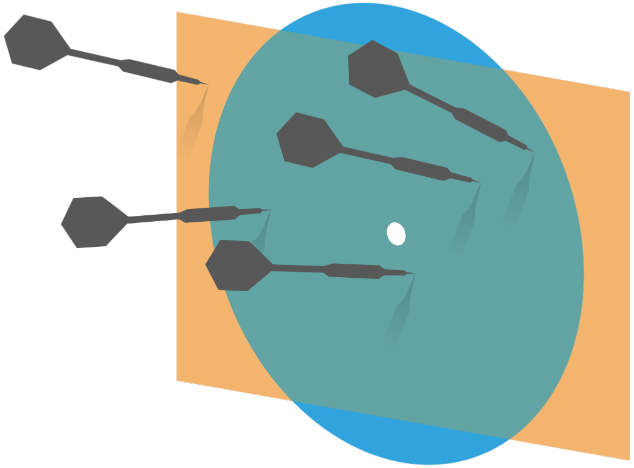
Probability and Statistics for Computer Science



Credit: wikipedia

Principal Component Analysis --- Exploring the data in less dimensions

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 03.30.2021

Last time

- Review of Bayesian inference
- Visualizing high dimensional data & Summarizing data
- The covariance matrix

Objectives

- Principal Component Analysis
- Examples of PCA

Examples: Immune Cell Data

- There are 38816 white blood immune cells from a mouse sample

$N = 38816$

$d \times N$

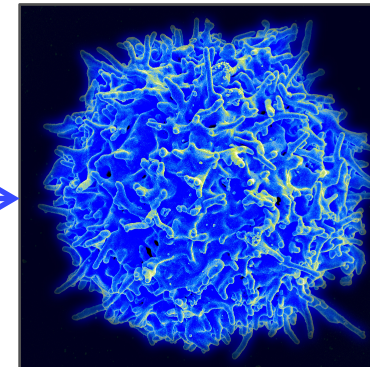
- Each immune cell has 40+ features/ components

- Four features are used as illustration.

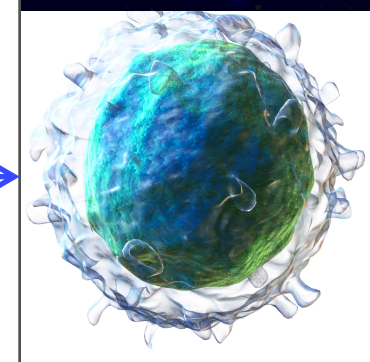
$d^* = 4$

- There are at least 3 cell types involved

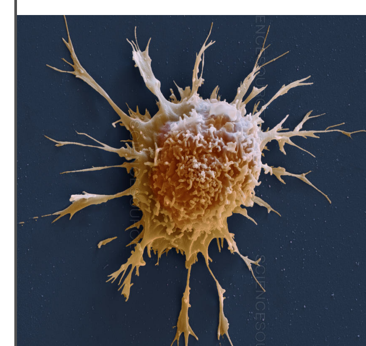
T cells



B cells

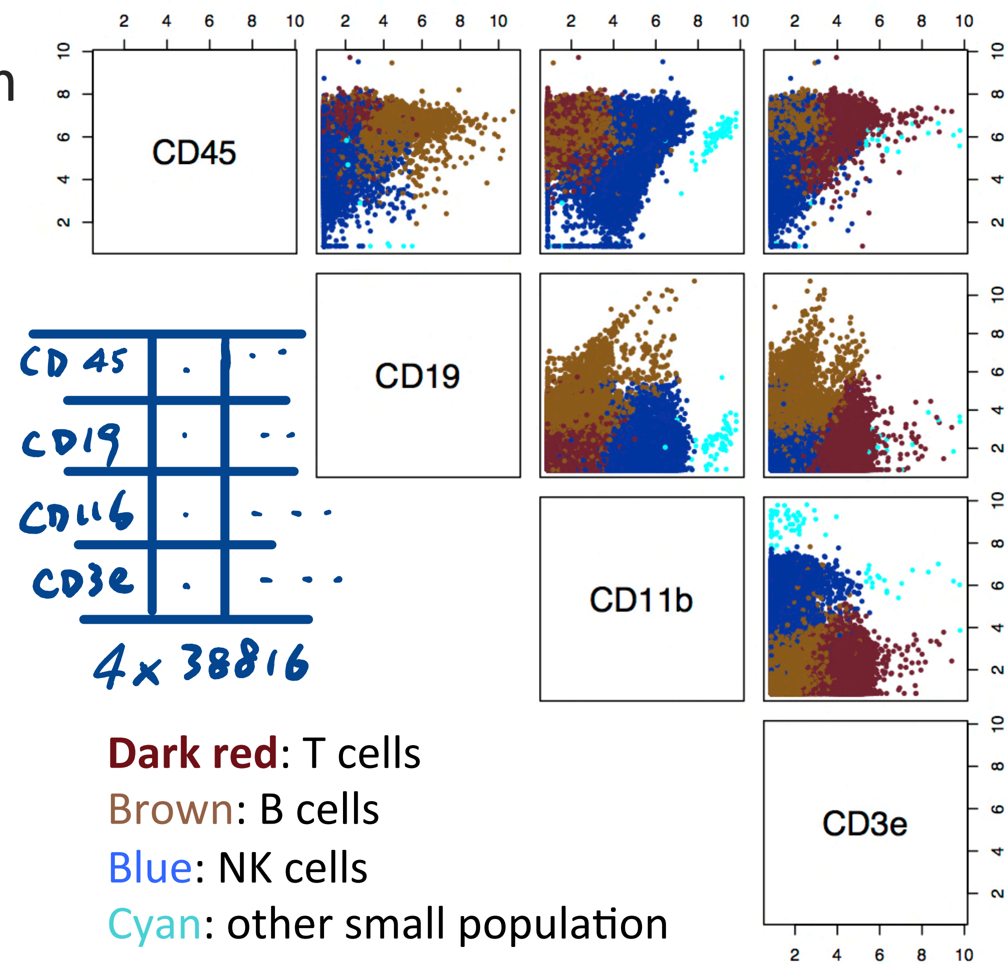


Natural killer cells



Scatter matrix of Immune Cells

- There are 38816 white blood immune cells from a mouse sample
- Each immune cell has 40+ features/ components
- Four features are used for the illustration.
- There are at least 3 cell types involved

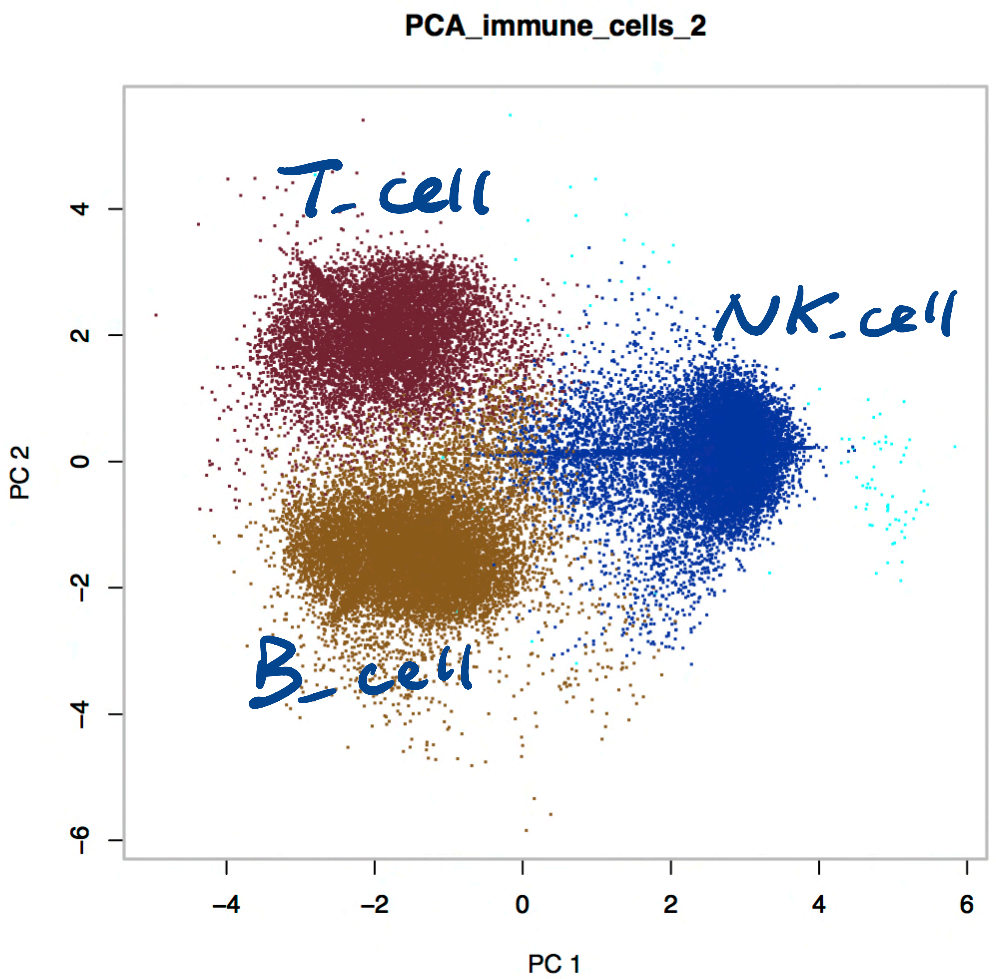


Dark red: T cells
Brown: B cells
Blue: NK cells
Cyan: other small population

PCA of Immune Cells' Data

> res1
$values Eigenvalues
[1] 4.7642829 2.1486896 1.3730662
0.4968255

Eigenvectors
$vectors
[,1] [,2] [,3] [,4]
[1,] 0.2476698 0.00801294 -0.6822740
0.6878210
[2,] 0.3389872 -0.72010997 -0.3691532
-0.4798492
[3,] -0.8298232 0.01550840 -0.5156117
-0.2128324
[4,] 0.3676152 0.69364033 -0.3638306
-0.5013477



Covariance matrix

Data set $\{\mathbf{x}\}$ 7×8

$d \times N$

$d = 7$ (# of features)

Covmat($\{\mathbf{x}\}$) 7×7

$cov(\{\mathbf{x}\}; 3, 5)$

	1	2	3	4	5	6	7	8
1	*	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*	*

	1	2	3	4	5	6	7
1	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*

Properties of Covariance matrix

$$cov(\{x\}; j, j) = var(\{x^{(j)}\})$$

Covmat($\{\mathbf{x}\}$) 7×7

- The diagonal elements of the covariance matrix are just variances of each jth components
- The off diagonals are covariance between different components

	1	2	3	4	5	6	7
1	σ_1^2	*	*	*	*	*	*
2	*	σ_2^2	*	*	*	*	*
3	*	*	σ_3^2	*	*	*	*
4	*	*	*	σ_4^2	*	*	*
5	*	*	*	*	σ_5^2	*	*
6	*	*	*	*	*	σ_6^2	*
7	*	*	*	*	*	*	σ_7^2

Properties of Covariance matrix

$$cov(\{x\}; j, k) = cov(\{x\}; k, j)$$

$\text{Covmat}(\{\mathbf{x}\})$ 7×7

	1	2	3	4	5	6	7
1	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*

- The covariance matrix is **symmetric**!
- And it's **positive semi-definite**, that is all $\lambda_i \geq 0$
- Covariance matrix is diagonalizable

Properties of Covariance matrix

- If we define X_c as the mean centered matrix for dataset {x}

$$Covmat(\{x\}) = \frac{X_c X_c^T}{N}$$

- The covariance matrix is a d×d matrix

$Covmat(\{\mathbf{x}\})$ 7×7

	1	2	3	4	5	6	7
1	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*

d =7

What is the correlation between the 2 components for the data m?

$$Covmat(\boldsymbol{m}) = \begin{bmatrix} 20 & 25 \\ 25 & 40 \end{bmatrix}$$

Corr(feature 1, feature 2)

$$\frac{25}{\sqrt{20}\sqrt{40}}$$

$$corr_{(x,y)} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Example: covariance matrix of a data set

(I) Mean centering

$$A_0 = \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix}$$

↓

$$A_1 = \begin{bmatrix} 2 & 1 & 0 & -1 & -2 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix}$$

(II) $A_2 = A_1 A_1^T$

Inner product of each pairs:

A_2 [1,1] = 10

A_2 [2,2] = 4

A_2 [1,2] = 0

(III)

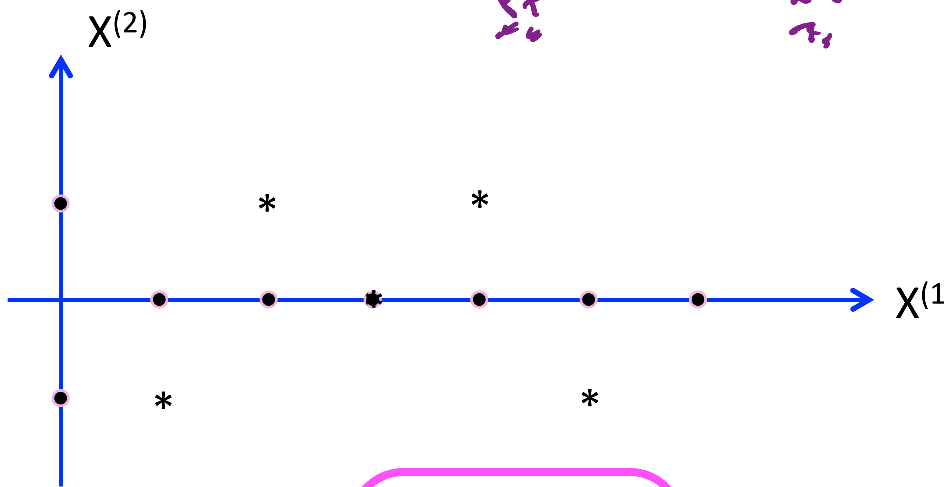
Divide the matrix with N – the number of data poits

$$\text{Covmat}(A_0) = \frac{1}{N} A_2 = \frac{1}{5} \begin{bmatrix} 10 & 0 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 0.8 \end{bmatrix}$$

$$\frac{X_c X_c^T}{N}$$

What do the data look like when Covmat({x}) is diagonal?

$$A_0 = \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix} \begin{matrix} x^{(1)} \\ x^{(2)} \end{matrix}$$



$$\text{Covmat}(A_0) = \frac{1}{N} A_2 = \frac{1}{5}\begin{bmatrix} 10 & 0 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 0.8 \end{bmatrix}$$

(σ_1^2 = 2, σ_2^2 = 0.8)

$\lambda_1 = \sigma_1^2$

$\lambda_2 = \sigma_2^2$

$$\text{covmat}(A_0) = \begin{bmatrix} 2 & 0 \\ 0 & 0.8 \end{bmatrix} = C_1$$

$$|C_1 - \lambda I| = 0$$

$$\begin{vmatrix} 2-\lambda & 0 \\ 0 & 0.8-\lambda \end{vmatrix} = 0$$

$$(2-\lambda)(0.8-\lambda) = 0$$

$$\Rightarrow \lambda_1 = 2, \quad \lambda_2 = 0.8$$

$$u_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad u_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$Au_1 = 2u_1$$

$$(A - 2I)u_1 = 0$$

$$\begin{bmatrix} 0 & 0 \\ 0 & -1.2 \end{bmatrix} u_1 = 0$$

$$u_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Linear transformation properties of mean and covariance matrix

- Linearly transforming the data set linearly transforms the mean

$$mean(\{A\mathbf{x}\}) = A\ mean(\{\mathbf{x}\})$$

- Linearly transforming the data set linearly changes the covariance matrix quadratically

$$Covmat(\{A\mathbf{x}\}) = A\ Covmat(\{\mathbf{x}\})A^T$$

if $\{x^*\} = \{U^T x\}$

$\searrow Covmat(\{x^*\}) = U^T Covmat(\{x\}) U = \Lambda$

Diagonalization

eig vect1 ↓   eig vect2 ↓

$$\underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 6 \end{bmatrix}}_{\Lambda} = \underbrace{\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}}_{X^{-1}} \underbrace{\begin{bmatrix} 1 & 5 \\ 0 & 6 \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_{X}$$

$$M = X \Lambda X^{-1}$$

$$\underbrace{\begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}}_{\Lambda} = \underbrace{\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}}_{U^T} \underbrace{\begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}}_{U}$$

"U^{-1}"

$$A = U \Lambda U^T$$

Diagonalization of a symmetric matrix

- If A is an **n×n** symmetric square matrix, the eigenvalues are real.
- If the eigenvalues are also distinct, their eigenvectors are orthogonal
- We can then scale the eigenvectors to unit length, and place them into an orthogonal matrix $U = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_n]$
- We can write the diagonal matrix $\Lambda = U^T A U$ such that the diagonal entries of Λ are $\lambda_1, \lambda_2 \ldots \lambda_n$ in that order.

Diagonalization example

- For

$$A = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$$

λ_i ? $\quad |A - \lambda I| = 0$

$$\begin{vmatrix} 5-\lambda & 3 \\ 3 & 5-\lambda \end{vmatrix} = 0 \Rightarrow \begin{cases} \lambda_1 = 2 \\ \lambda_2 = 8 \end{cases}$$

eigenvectors?

$\lambda_1 = 2 \quad A v_1 = 2 v_1$

$(A - 2I) v_1 = 0$

$$\begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix} v_1 = 0 \Rightarrow v_1 = c \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\Rightarrow u_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

↑ normalized eigenvector

$\lambda_2 = 8 \quad u_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$$U = [u_1 \ u_2] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$\Lambda = U^T A U \qquad \Lambda = \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix}$$

Diagonalization example

- For

$$A = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$$

λ_i ?

$$|A - \lambda I| = 0$$

$$\begin{vmatrix} 5-\lambda & 3 \\ 3 & 5-\lambda \end{vmatrix} = 0 \Rightarrow \begin{cases} \lambda_1 = 8 \\ \lambda_2 = 2 \end{cases}$$

eigenvectors?

$\lambda_1 = 8$

$$A v_1 = 8 v_1$$

$$(A - 8I) v_1 = 0$$

$$\begin{bmatrix} -3 & 3 \\ 3 & -3 \end{bmatrix} v_1 = 0 \Rightarrow v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\Rightarrow u_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

↑ normalized eigenvector

$\lambda_2 = 2$ $\quad u_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

$$U = [u_1 \;\; u_2]$$

$$= ?$$

$$\Lambda = U^T A U$$

$$\Lambda = ? \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}$$

Rotation Matrix

Def. $R^T = R^{-1}$

We can prove $U^T = U^{-1}$ if U is formed by normalized eigenvectors.

U^T & U are called orthonormal

$\Rightarrow$ U^T & U are rotation matrices.

If

$$U = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$u_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad u_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad u_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Dot prod.

$u_1^T \cdot u_2 = ?$ 0

$u_1^T \cdot u_3 = ?$ 0

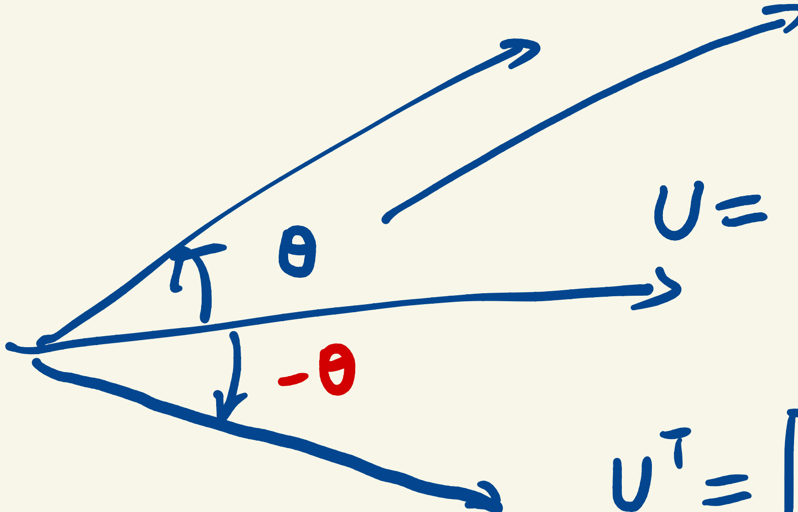
$u_2^T \cdot u_3 = ?$ 0

orthogonal (perpendicular) $\perp$

$\|u_1\| = ?$ 1 $\quad \|u_2\| = ?$ 1 $\quad \|u_3\| = ?$ 1

2D

$U \cdot x$



$U^T x$

$$U = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$U^T = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

$$U^T (U x) = I \cdot x$$

$$U^T = U^{-1} \Rightarrow U^T \cdot U = I$$

$$\det(U) = 1$$

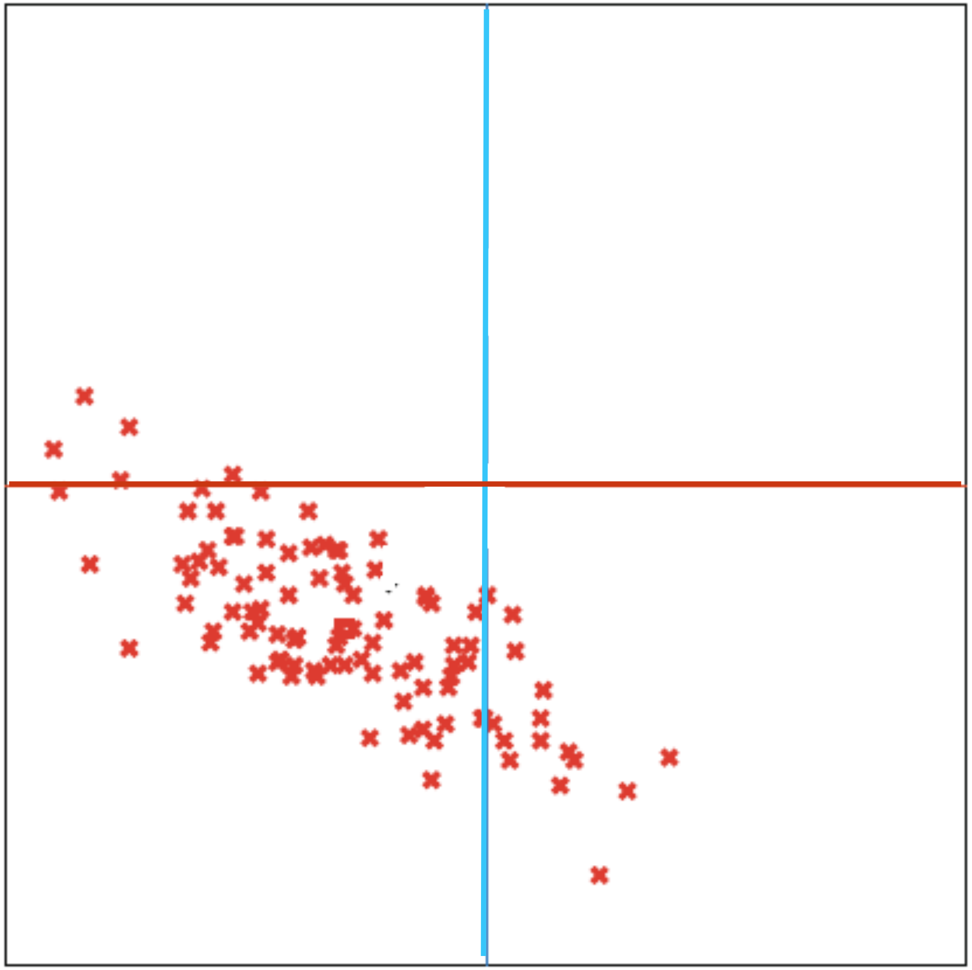
Q. Is this true?

Transforming a matrix with orthonormal matrix only rotates the data

A. Yes

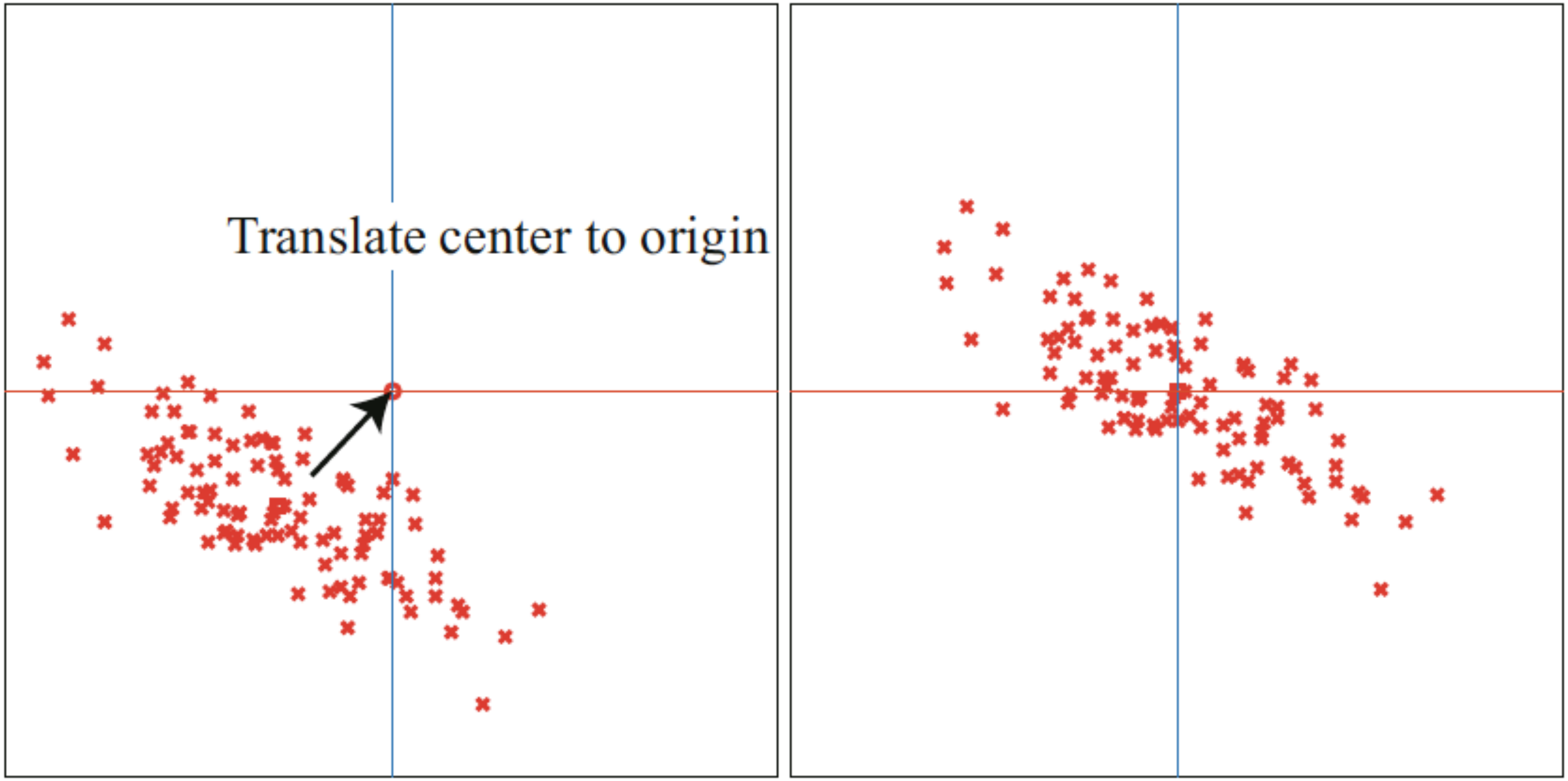
B. No

Dimension reduction from 2D to 1D



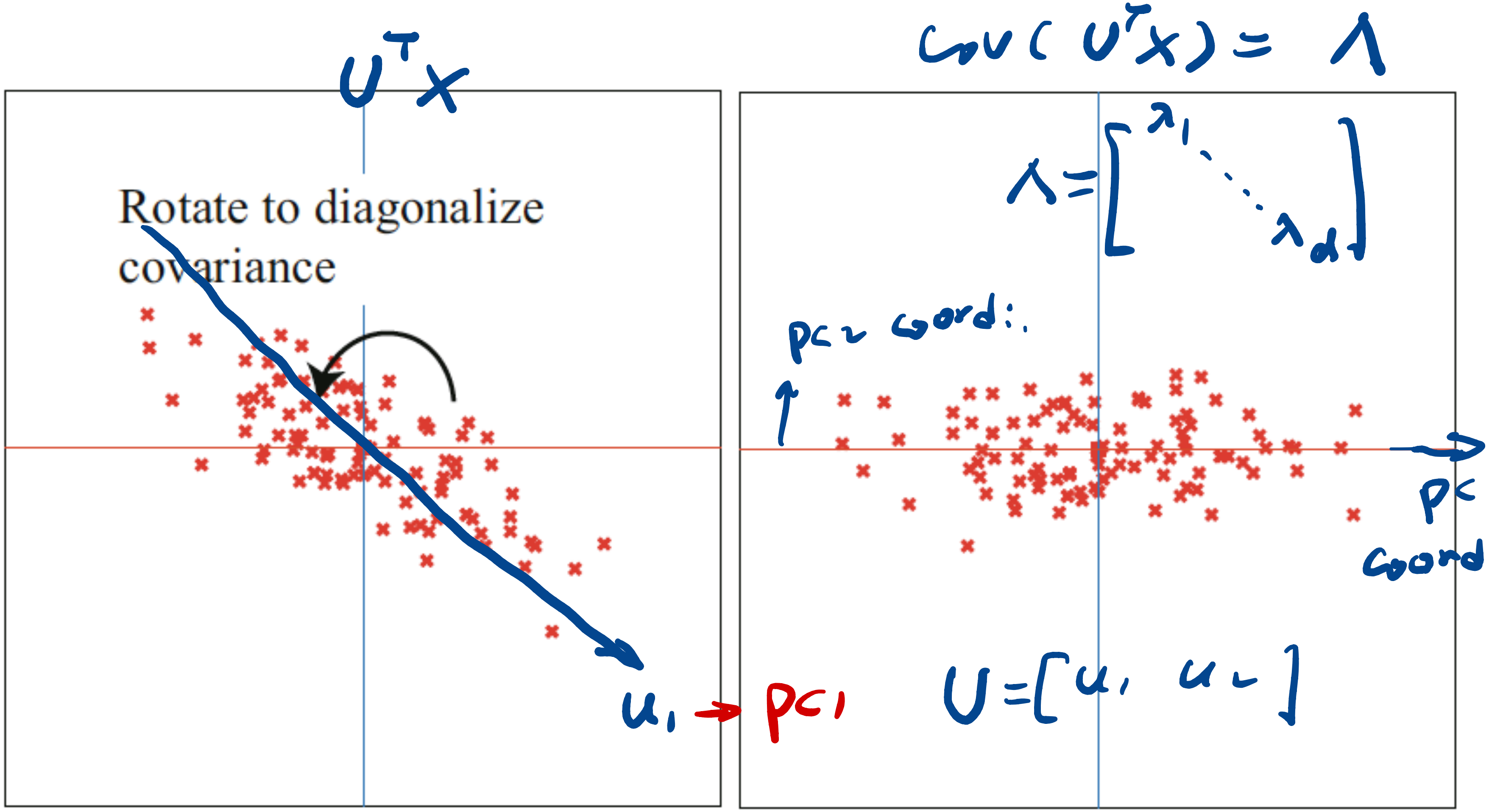
Credit: Prof. Forsyth

Step 1: subtract the mean



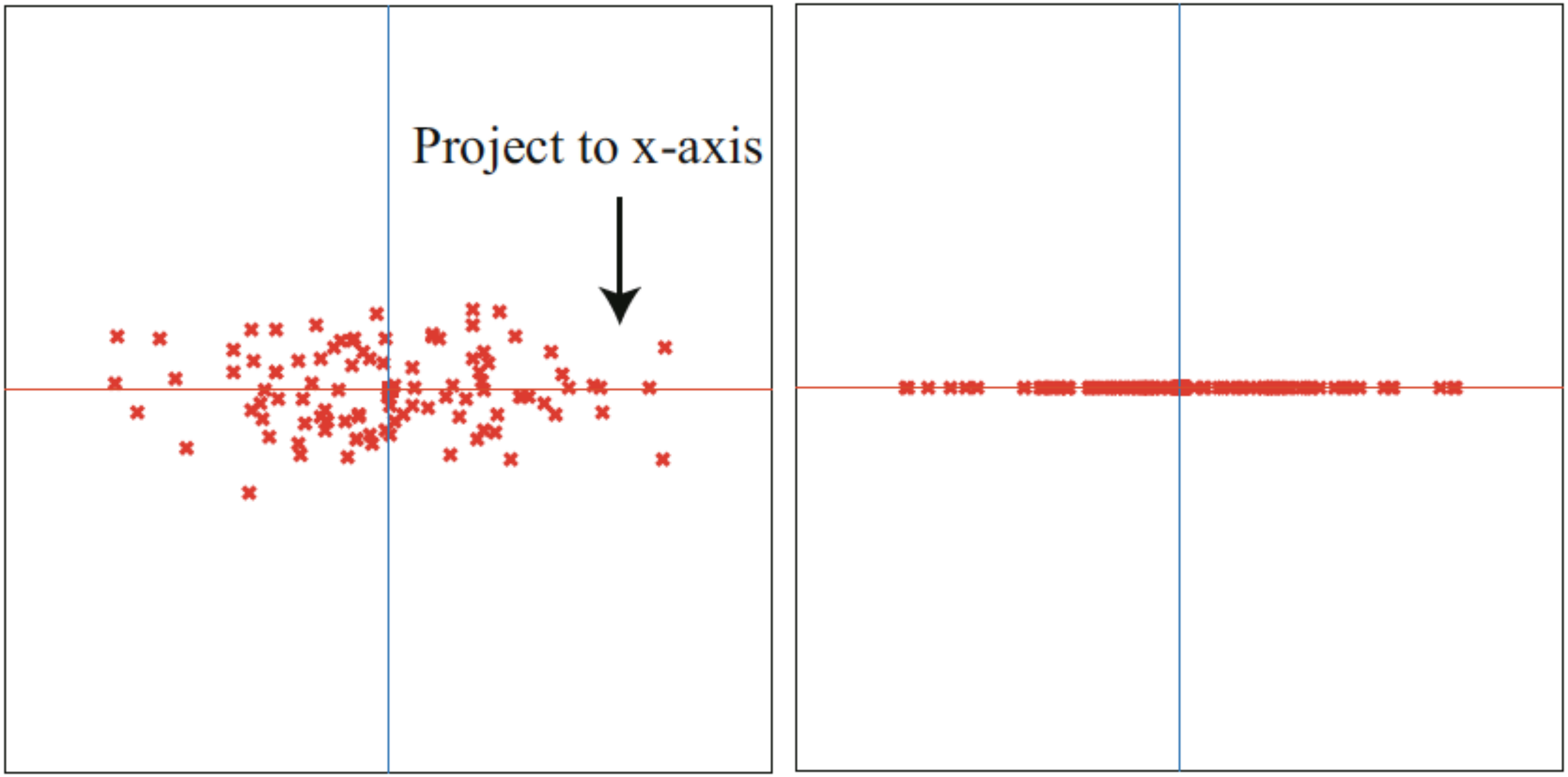
Credit: Prof. Forsyth

Step 2: Rotate so that the new data has diagonalized covariance matrix



Credit: Prof. Forsyth

Step 3: Drop component(s)



Credit: Prof. Forsyth

Principal Components

- The columns of U are the normalized eigenvectors of the Covmat({x}) and are called the **principal components** of the data {x}

Principal components analysis

- We reduce the dimensionality of dataset {**x**} represented by matrix $\boldsymbol{D}_{d \times n}$ from d to s (s < d).

- Step 1. define matrix $\boldsymbol{m}_{d \times n}$ such that $\boldsymbol{m} = \boldsymbol{D} - mean(\boldsymbol{D})$

- Step 2. define matrix $\boldsymbol{r}_{d \times n}$ such that $\boldsymbol{r}_i = \boldsymbol{U}^T \boldsymbol{m}_i$

  Where $\boldsymbol{U}^T$ satisfies $\boldsymbol{\Lambda} = \boldsymbol{U}^T \; Covmat(\{\boldsymbol{x}\})\boldsymbol{U}$, $\boldsymbol{\Lambda}$ is the diagonalization of $Covmat(\{\boldsymbol{x}\})$ with the eigenvalues sorted in decreasing order, $\boldsymbol{U}$ is the orthonormal eigenvectors' matrix

- Step 3. Define matrix $\boldsymbol{p}_{d \times n}$ such that $\boldsymbol{p}$ is $\boldsymbol{r}$ with the last d-s components of $\boldsymbol{r}$ made zero.

What happened to the mean?

- Step 1.

$$mean(\boldsymbol{m}) = mean(\boldsymbol{D} - mean(\boldsymbol{D})) = 0$$

- Step 2.

$$mean(\boldsymbol{r}) = \boldsymbol{U}^T mean(\boldsymbol{m}) = \boldsymbol{U}^T 0 = 0$$

- Step 3.

$$mean(\boldsymbol{p_i}) = mean(\boldsymbol{r}_i) = 0 \;\; while \; i \in 1 : s$$

$$mean(\boldsymbol{p}_i) = 0 \quad while \; i \in s + 1 : d$$

What happened to the covariances?

- Step 1.

$$Covmat(\boldsymbol{m}) = Covmat(\boldsymbol{D}) = Covmat(\{\boldsymbol{x}\})$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \phi \\ & \lambda_2 & & \\ & & \ddots & \\ \phi & & & \lambda_d \end{bmatrix}$$

- Step 2.

$$Covmat(\boldsymbol{r}) = \boldsymbol{U}^T Covmat(\boldsymbol{m}) \boldsymbol{U} = \boldsymbol{\Lambda}$$

- Step 3. $Covmat(\boldsymbol{p})$ is $\boldsymbol{\Lambda}$ with the last/smallest d-s diagonal terms turned to 0.

$U = [u_1 \; u_2 \; \cdots]$, $u_1 \downarrow \lambda_1$, $u_2 \downarrow \lambda_2$

$\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots$

$U^{-1} = U^T$ for symmetric diagonalizable matrix

Sample covariance matrix

- In many statistical programs, the sample covariance matrix is defined to be

$$Covmat(\boldsymbol{m}) = \frac{\boldsymbol{m}\ \boldsymbol{m}^T}{N-1}$$

- Similar to what happens to the unbiased standard deviation

PCA an example

- Step 1.

$$D = \begin{bmatrix} 3 & -4 & 7 & 1 & -4 & -3 \\ 7 & -6 & 8 & -1 & -1 & -7 \end{bmatrix} \Rightarrow mean(\boldsymbol{D}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{m} = \begin{bmatrix} 3 & -4 & 7 & 1 & -4 & -3 \\ 7 & -6 & 8 & -1 & -1 & -7 \end{bmatrix}$$

- Step 2.

$$Covmat(\boldsymbol{m}) = \begin{bmatrix} 20 & 25 \\ 25 & 40 \end{bmatrix} \Rightarrow \lambda_1 \simeq 57; \quad \lambda_2 \simeq 3$$

$$\Rightarrow \boldsymbol{U} = \begin{bmatrix} 0.5606288 & -0.8280672 \\ 0.8280672 & 0.5606288 \end{bmatrix} \qquad \boldsymbol{U^T} = \begin{bmatrix} 0.5606288 & 0.8280672 \\ -0.8280672 & 0.5606288 \end{bmatrix}$$

$$\Rightarrow \boldsymbol{r} = \boldsymbol{U}^T \boldsymbol{m} = \begin{bmatrix} 7.478 & -7.211 & 10.549 & -0.267 & -3.071 & -7.478 \\ 1.440 & -0.052 & -1.311 & -1.389 & 2.752 & -1.440 \end{bmatrix}$$

- Step 3. $\Rightarrow \boldsymbol{p} = \begin{bmatrix} 7.478 & -7.211 & 10.549 & -0.267 & -3.071 & -7.478 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$

What is this matrix for the previous example?

$$\boldsymbol{U}^T Covmat(\boldsymbol{m})\boldsymbol{U} = ?$$

$$\Lambda = \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix} = \begin{bmatrix} 57 & 0 \\ 0 & 3 \end{bmatrix}$$

$\lambda_1 \quad \lambda_2 \quad - \quad - \quad -$

What is this matrix for the previous example?

$$\boldsymbol{U}^T Covmat(\boldsymbol{m})\boldsymbol{U} = ?$$

$$\begin{bmatrix} 57 & 0 \\ 0 & 3 \end{bmatrix}$$

The Mean square error of the projection

- The mean square error is the sum of the smallest d-s eigenvalues in $\mathbf{\Lambda}$

$$\frac{1}{N-1}\sum_i \|r_i - p_i\|^2 = \frac{1}{N-1}\sum_i \sum_{j=s+1}^{d} (r_i^{(j)})^2 = \sum_{j=s+1}^{d} \sum_i \frac{1}{N-1}(r_i^{(j)})^2$$

$$= \sum_{j=s+1}^{d} var(r_i^{(j)})$$

$$= \sum_{j=s+1}^{d} \lambda_j$$

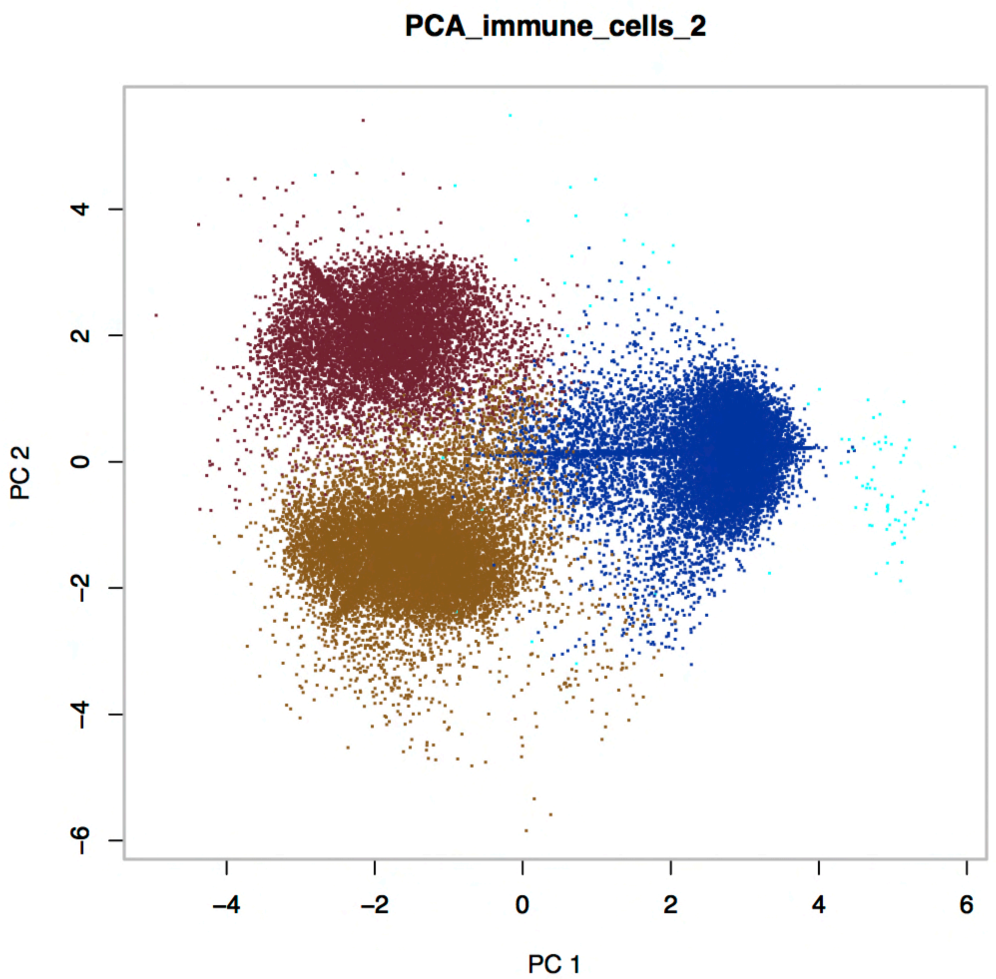
PCA of Immune Cells

```
> res1
$values
[1] 4.7642829 2.1486896 1.3730662
0.4968255
```

Eigenvalues

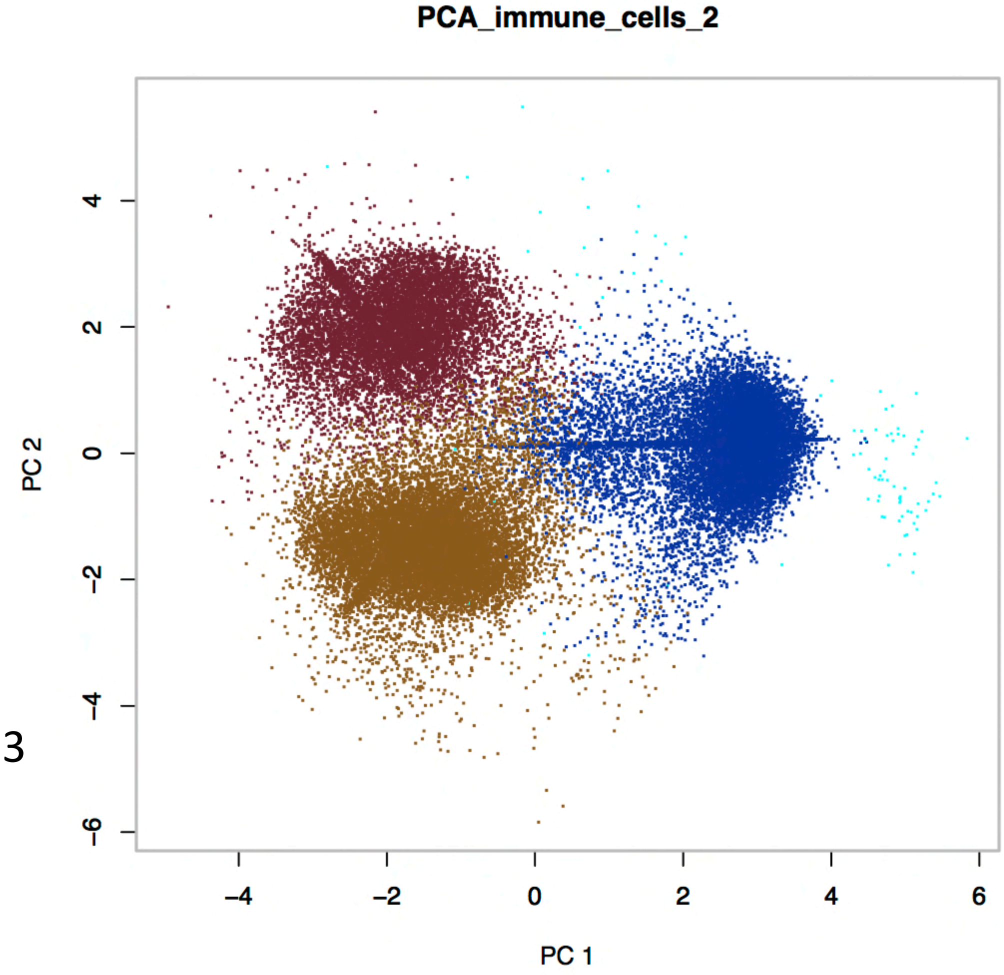
Eigenvectors

```
$vectors
          [,1]       [,2]       [,3]       [,4]
[1,]  0.2476698  0.00801294 -0.6822740
0.6878210
[2,]  0.3389872 -0.72010997 -0.3691532
-0.4798492
[3,] -0.8298232  0.01550840 -0.5156117
-0.2128324
[4,]  0.3676152  0.69364033 -0.3638306
-0.5013477
```



New coordinates in PCA

> head(new_coord_t)
PC1 PC2 PC3 PC4
1 3.6739228 0.1127233 -1.32744266
0.61005994
2 -0.9255199 -2.1016573 -0.80762548
-0.29104900
3 3.1150230 0.3526459 -0.83994064
0.46074556
4 3.1801414 0.5679807 -0.07097689
0.01539266
5 2.7972723 -0.1073053 -0.39168826
-0.03981390
6 3.3012610 0.1979659 0.17965423
0.45373049CD3e -0.3676152 0.69364033
-0.3638306 -0.5013477 [4,] 0.3676152
0.69364033 -0.3638306 -0.5013477



What is the percentage of variance that PC1 covers?

Given the eigenvalues: 4.7642829 2.1486896 1.3730662 0.4968255, what is the percentage that PC1 covers?

$\lambda_1 = \sigma_1^2$

A. 54%
B. 16%
C. 25%

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}$$

Assignments

- Read Chapter 10 of the textbook
- Week 10 module
- Next time: Intro to classification

Additional References

- Robert V. Hogg, Elliot A. Tanis and Dale L. Zimmerman. "Probability and Statistical Inference"
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics"

See you next time

See You!

$$\Lambda = U^T C_1 U \quad \Rightarrow \quad U = I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$