

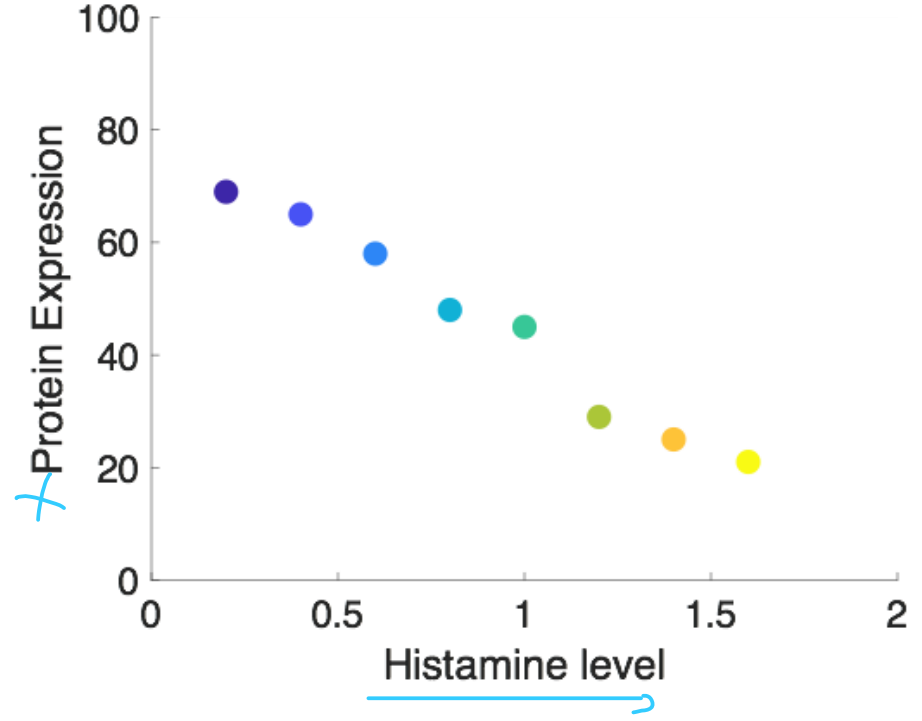
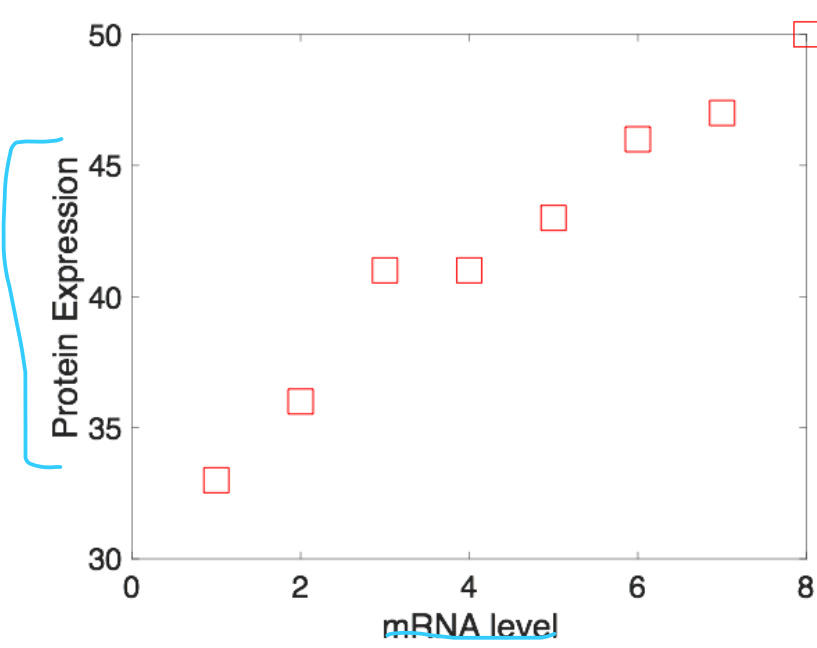
Week 14

Other numerical methods in biology

--

Scatter plot

Shows the relation between two variables

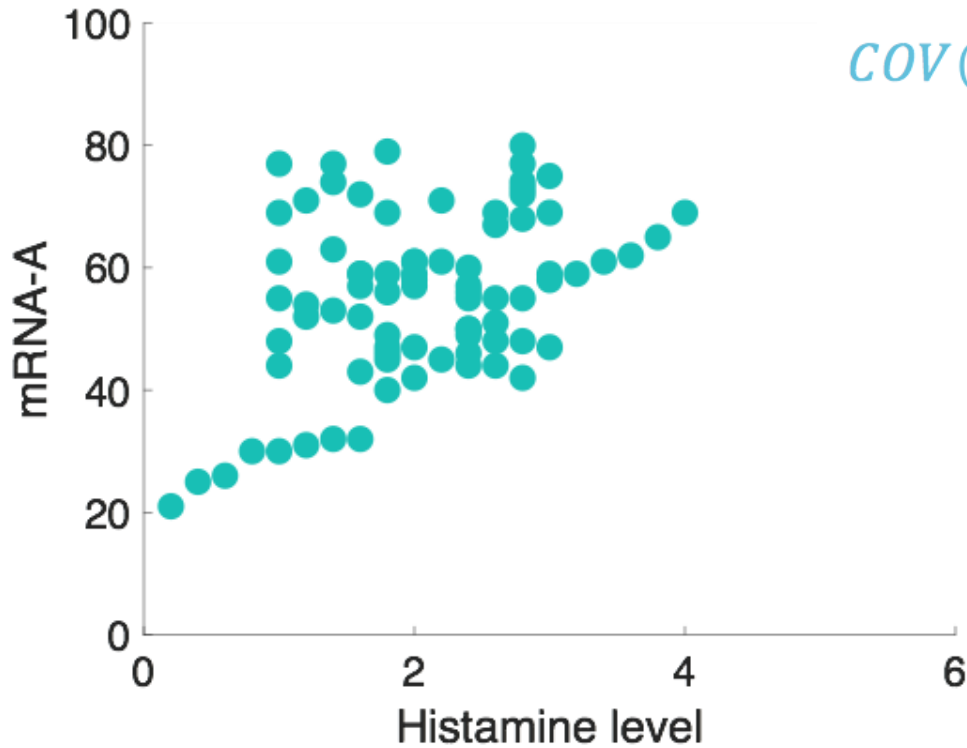


Can we quantitatively measure the strength of relationship between variables?

Linear regression is a form of regression in which one explanatory variable is used to predict the outcome of a response variable.

Covariance

Does Y get larges (smalleR) as Y increase?



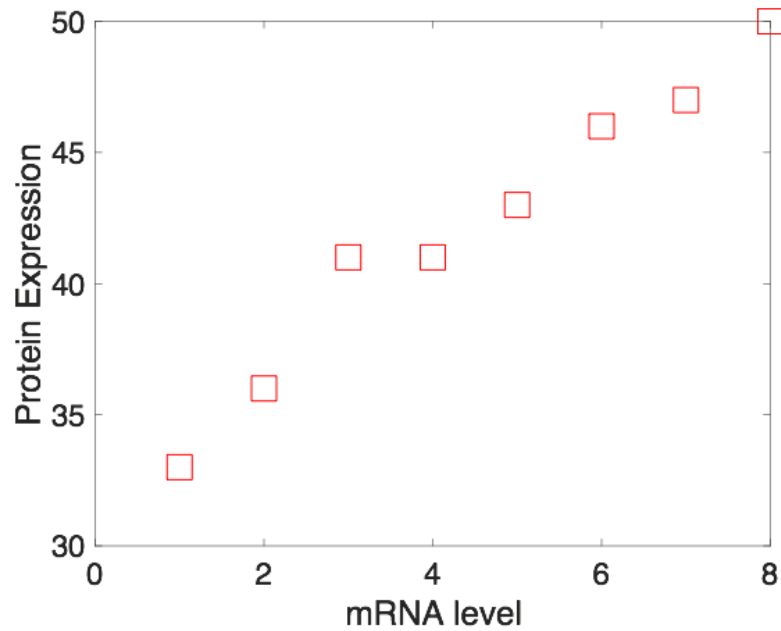
$$COV(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

mean of x= 2.0525
mean of y = 55.4125
Sx = 0.7916
Sy= 13.6537

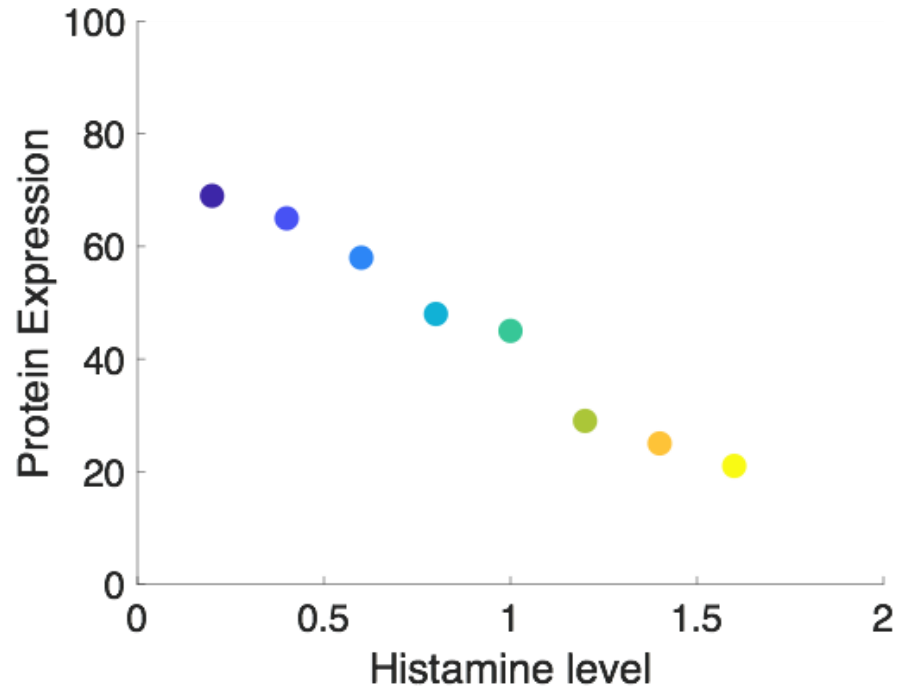
n=80

Covariance > 0 if X and Y variables gets larger

Covariance < 0 if X and Y variables moves opposite direction



$r=+0.9538$



$r=-0.987$

Correlation coefficient always lies between -1 to +1

Fitlm and polyfit functions

```
b = fitlm(hist',genetrial')
```

New to MATLAB? See resources for [Getting Started](#).

```
y ~ 1 + x1
```

Estimated Coefficients:

	<u>Estimate</u>	<u>SE</u>	<u>tStat</u>	<u>pValue</u>
(Intercept)	42.933	2.1767	19.724	4.544e-08
x1	3.2303	0.35081	9.2082	1.5659e-05

Number of observations: 10, Error degrees of freedom: 8

Root Mean Squared Error: 3.19

R-squared: 0.914, Adjusted R-Squared: 0.903

F-statistic vs. constant model: 84.8, p-value = 1.57e-05

```
fx >>
```

```
[co,S]=polyfit(hist,genetrial,1)
```

```
co =
```

```
3.2303 42.9333
```

```
S =
```

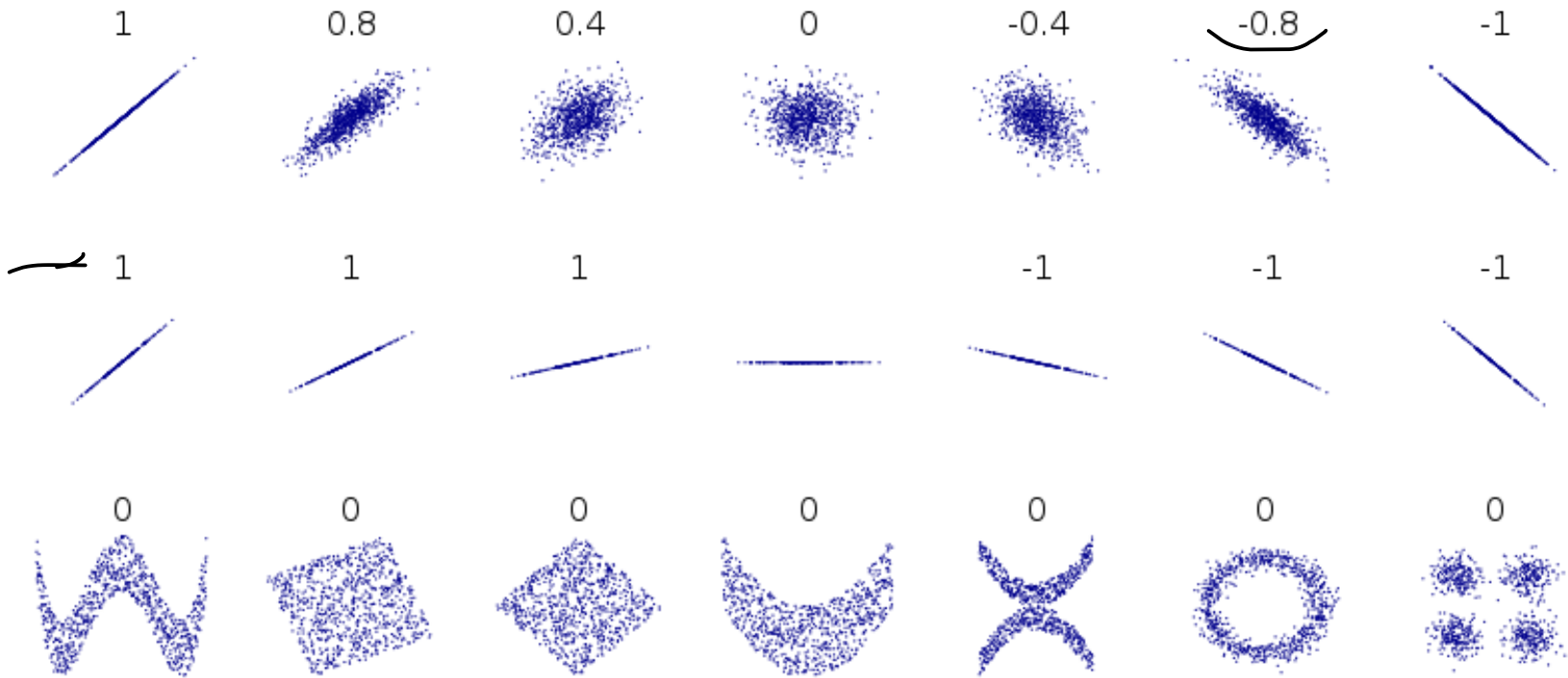
```
struct with fields:
```

```
R: [2x2 double]
```

```
df: 8
```

```
normr: 9.0124
```

Correlation sets



Remember that correlation coefficient is an indicator of the strength of a *linear* relationship between two variables, but its value generally does not completely characterize their relationship

r^2 IN REGRESSION

The square of the correlation, r^2 , is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x .

$$r^2 = \frac{\text{variance of predicted values } \hat{y}}{\text{variance of observed values } y}$$

Properties of r^2

$$0 \leq r^2 \leq 1$$

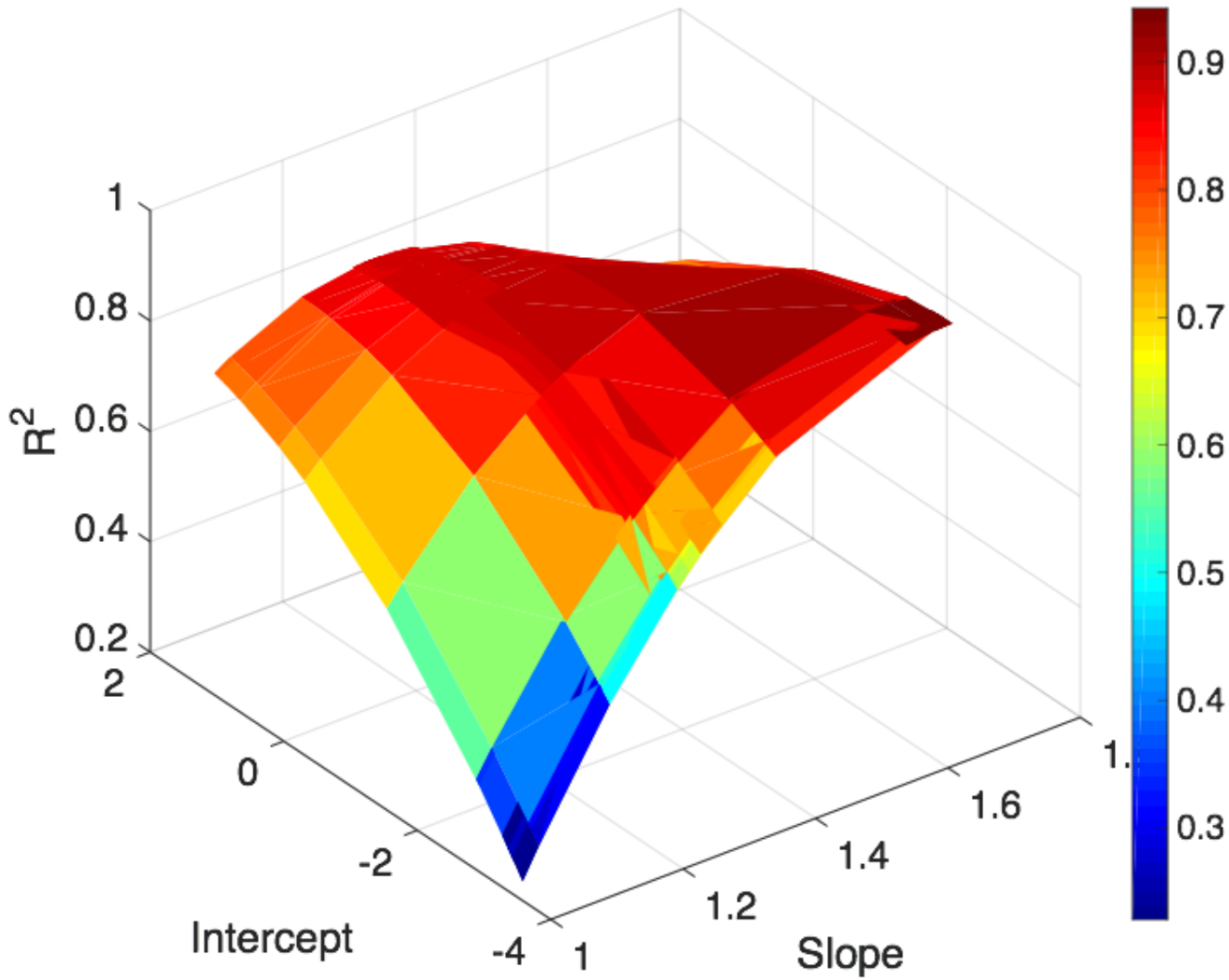
if $r^2 = 1$, it represents a straight line

if $r^2 = 0$, it indicates no correlation between y and x

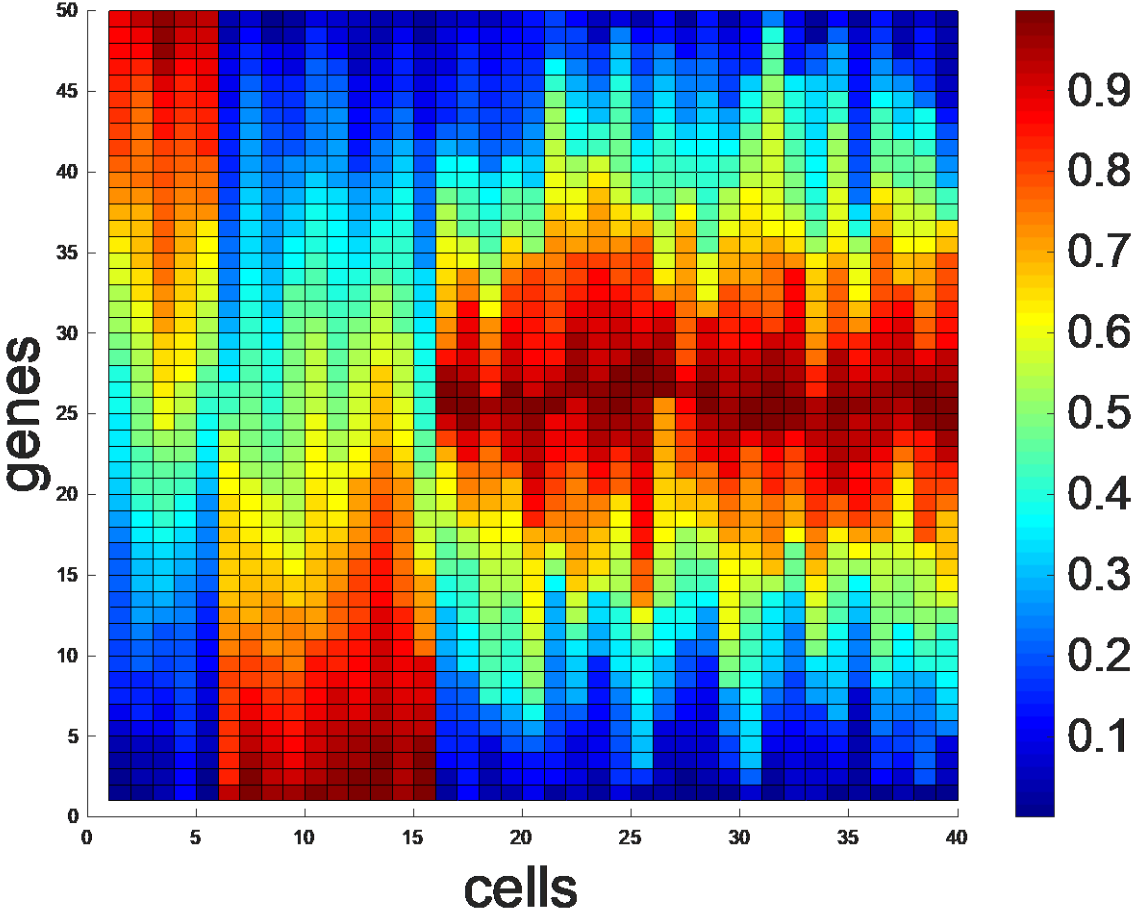
Larger the r^2 means higher correlation, but not always

R2 gets smaller by the size of

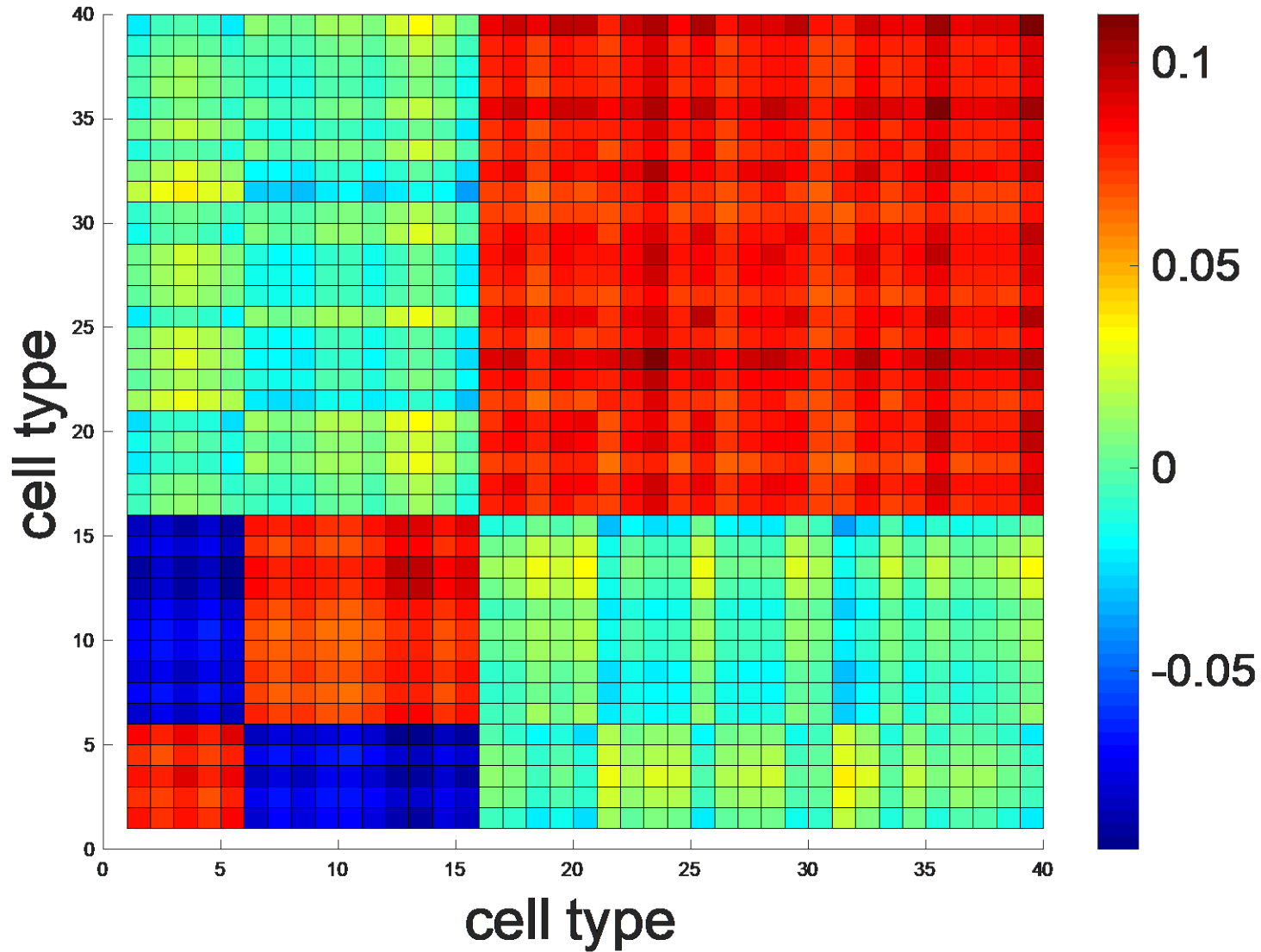
Slope = 1.63
Intercept = -3.35



Gene expression in different cells



What is the covariance between different cell types?



MULTIVARIATE REGRESSION

In linear regression, a single independent variable was present. A total of two variables. In multiple regression, y dependent variable (response variable) depends on a many explanatory independent variables.

Now we can define linear function as

$$Y = \text{constant (a)} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots + \beta_k x_n$$

It is also called as population regression equation.

y varies normally with a mean given by the population regression equation

MULTIVARIATE REGRESSION

- y - dependent variable or also called response variable
- $x_1, x_2, x_3 \dots, x_n$ are called independent variables

or explanatory variables.

- X values can either quantitative or categorical.

$$Y = \text{constant (a)} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots + \beta_k x_n$$

Examples of multivariate regression

2. Dependence of cancer risk to several genes (biology)
3. Dependence of home price to location, size, type etc. (home market)
4. Dependence of hormone levels to genes (health)
5. Dependence of reading score to mothers education, age, gender, family income etc. (social science)

In Matlab

```
mdl = fitlm(X,Y)
```

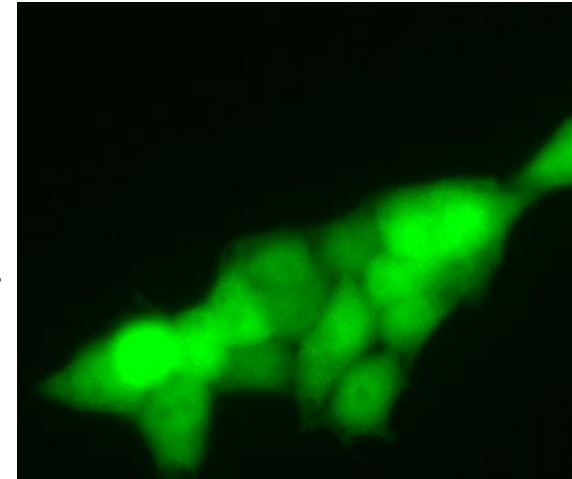
Dependence of cell growth to expression of geneX, geneY and geneZ

Linear regression model:

$$\underline{y} \sim 1 + x1 + x2 + x3$$

Estimated Coefficients:

	<u>Estimate</u>	<u>SE</u>	<u>tStat</u>	<u>pValue</u>
(Intercept)	47.153	26.499	1.7794	0.078342
x1	0.28602	0.069679	4.1048	8.4971e-05
x2	-0.0033967	0.0047938	-0.70856	0.48031
x3	-0.3098	0.071258	-4.3476	3.4254e-05



Number of observations: 100, Error degrees of freedom: 96

Root Mean Squared Error: 1.74

R-squared: 0.994, Adjusted R-Squared 0.993

F-statistic vs. constant model: 4.95e+03, p-value = 4.52e-105

>>

$$\text{Cell growth} = 47 + 0.28\text{geneX} - 0.003\text{geneY} - 0.30\text{geneZ}$$

Dependence of hormone levels to expression of geneX, geneY and geneZ

	1	2	3	4
1	120	140	249	
2	120	218	245	
3	123	124	244	
4	125	248	243	
5	128	186	241	
6	129	207	241	
7	130	190	240	
8	131	177	240	
9	132	172	238	
10	132	149	237	
11	133	162	237	
12	134	204	233	
13	136	215	232	
14	137	123	230	
15	138	166	229	
16	139	168	227	
17	140	135	227	
18	141	142	224	
19	141	177	221	
20	147	148	221	
21	147	167	221	
22	148	209	221	
23	153	221	220	
24	154	164	218	
25	155	122	216	
26	155	140	215	
27	156	157	215	

	1	2
1	2	
2	6	
3	7	
4	7	
5	8	
6	9	
7	11	
8	14	
9	18	
10	19	
11	21	
12	21	
13	21	
14	21	
15	22	
16	22	
17	22	
18	24	
19	26	
20	26	
21	27	
22	27	
23	27	
24	27	
25	27	
26	27	
27	28	

Lets predict cell growth

We conclude that geneX and gene Z contain useful information for predicting cell growth

Let's find the predicted cell growth for a sample with an 0.3 average in geneX and 0.6 in geneZ.

The explanatory variables are geneX and geneY. The predicted cell growth is

$$\text{Cell growth} = 47 + 0.28\text{geneX} - 0.003\text{geneY} - 0.30\text{geneZ}$$

$$\text{Cell growth} = 47 + 0.28\text{gene} - 0.3\text{geneZ}$$

$$\text{Cell growth} = 47 + 0.28(0.3) - 0.30(0.6)$$

Logistic Regression

NATIONAL CANCER INSTITUTE

GENETIC CHANGES AND CANCER

HOW GENETIC INFORMATION CREATES PROTEINS

DNA
DNA is a molecule in the cell nucleus that contains instructions for making proteins. It is made of four different bases: adenine (A), thymine (T), guanine (G), and cytosine (C). A segment of DNA that contains the information for making a protein is called a gene. In the process of **transcription**, DNA that makes up a gene is copied into a complementary molecule called messenger RNA (mRNA).

RNA
mRNA is also made of four bases: adenine (A), uracil (U), guanine (G), and cytosine (C). mRNA moves from the nucleus to the cytoplasm where it interacts with ribosomes, the protein factories of the cell. There, through a process called **translation**, mRNA is translated into amino acids. A sequence of three mRNA bases is called a **codon**, and each codon is translated into a specific amino acid. There are 20 different kinds of amino acids in humans.

PROTEIN
As an mRNA molecule is translated, a chain of amino acids is formed. The chain eventually folds into a three-dimensional protein. The shape of a protein determines its function. Proteins have millions of functions in cells.

Types of Genetic Mutations in Cancer

DNA alterations can affect the structure, function, and amount of the corresponding proteins. All of these effects can change a cell's behavior from normal to cancerous. For example, a genetic alteration can intensify or eliminate the protein's function, which could make cells divide uncontrollably. Many different kinds of genetic mutations are found in cancer cells, including missense, nonsense, and frameshift mutations and chromosome rearrangements.

MISSENSE MUTATION

Original	C T A	D C G	G P A	DNA
	(Leu)	(Trp)	(Val)	Amino Acids
Mutation	C T A	D C G	G P A	DNA
	(Leu)	(Trp)	(Val)	Amino Acids

A missense mutation is a change of a single DNA base that results in a change in the amino acid sequence. Sometimes a single amino acid change can greatly alter the protein's function.

NONSENSE MUTATION

Original	C T A	D C G	G P A	DNA
	(Leu)	(Trp)	(Val)	Amino Acids
Mutation	C T A	D C G	G P A	DNA
	(Leu)	(Trp)	(Val)	Amino Acids

A nonsense mutation is a change of a single DNA base that creates a "stop" codon, which terminates translation. The result is a shortened protein that may not function or that may have an abnormal function.

FRAMESHIFT MUTATION

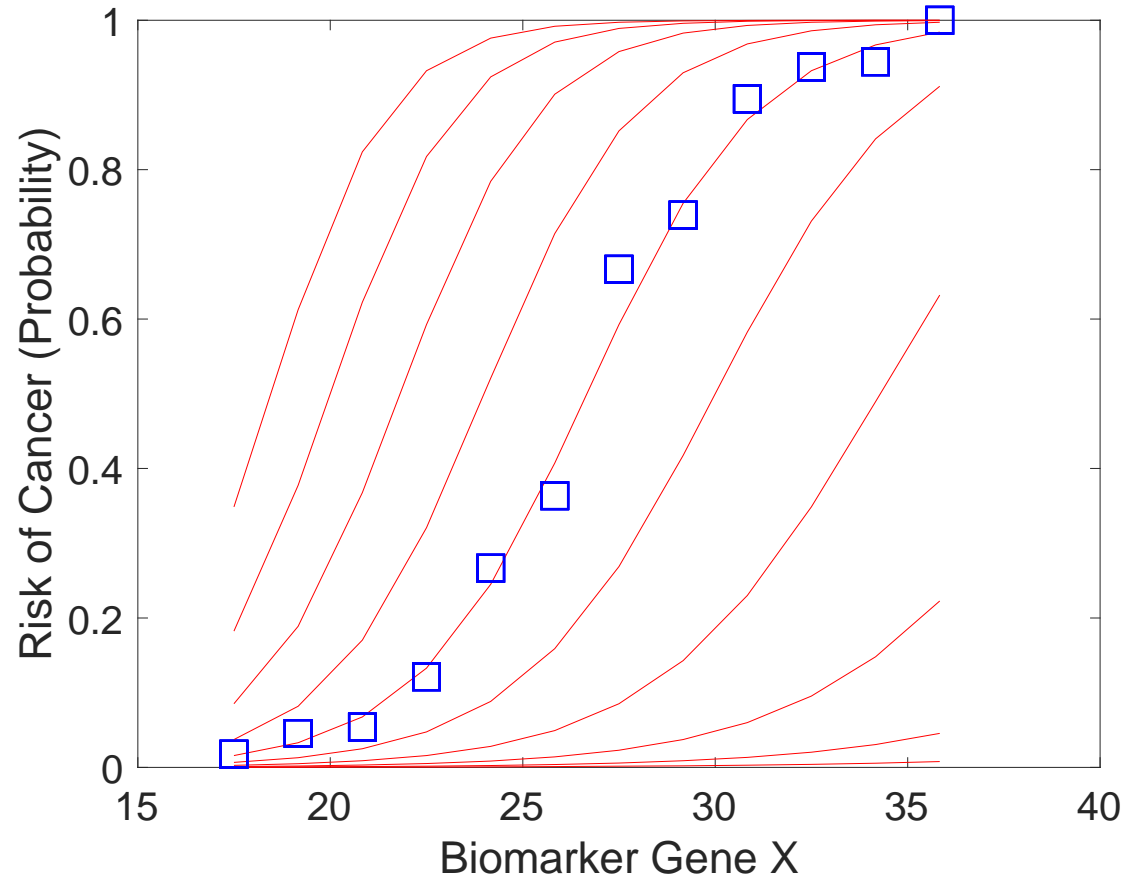
Original	C T A	D C G	G P A	DNA
	(Leu)	(Trp)	(Val)	Amino Acids
Mutation	C T A	D C G	G P A	DNA
	(Leu)	(Trp)	(Val)	Amino Acids

A frameshift mutation results from the addition or removal of DNA bases that shifts the DNA sequence and the corresponding amino acid sequence. The result is a protein whose sequence, structure, and function are very different from those of the original protein.

CHROMOSOME REARRANGEMENTS

DNA is wound tightly into structures called chromosomes. Chromosome rearrangements can occur when a piece of a chromosome breaks and is lost entirely (deletion), moves to a different chromosomal location (translocation), flips directions (inversions), or is repeated (duplication). These rearrangements can alter several genes at once. For example, they can generate fusion genes, in which parts of two separate genes are joined together. Proteins made from fusion genes sometimes cause cancer.

cancer.gov/genetics



What is logistic regression?

It is used to determine model parameters when dependent variables are binary rather than continuous

For example,
cell division, 0 or 1
Cancer diagnostic, cancer/not
Voting yes/no
Mortality alive/death
Product-marketing, sold/not sold
Arrived/delayed

The results of these data is not continuous as you have seen in multivariable linear regression

Logistic model can be used to make prediction for binary results

Logistic Regression

If a response variable such as yes/no or success/failure response variables., we cannot use linear regression models where it assumes a normal distribution.

Think about a cancer patient diagnosis whether a patient either have a cancer or not a cancer

One type of model that can be used is called **logistic regression**. We think in terms of a binomial model for the two possible values of the response variable and use one or more explanatory variables to explain the probability of success.

$$P(Y=1|\beta) = \frac{\exp(b(1)+b(2)x)}{1+\exp(b(1)+b(2)x)}$$

x= binary or cont

y= binary

b(1) and b(2) are coefficients

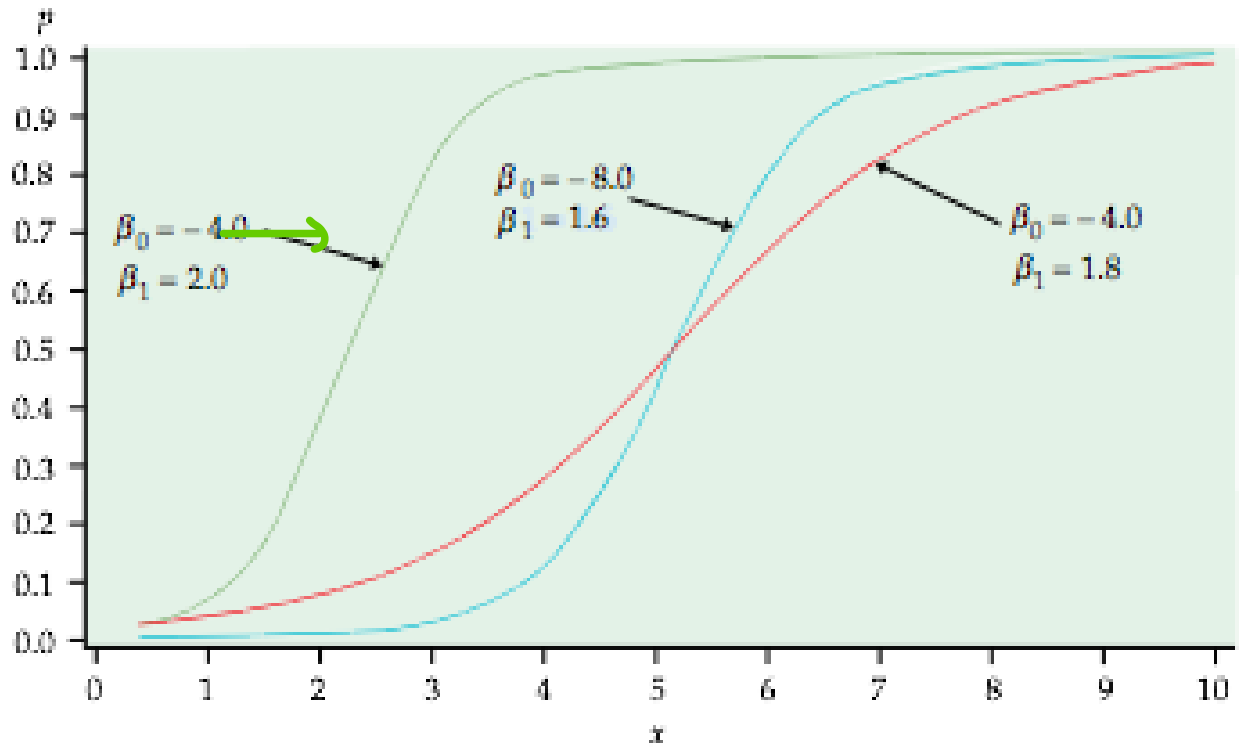
if y response variable is discrete

Y = P(Y=0) + P(Y=1)

Logistic function

it can be defined as

$f(x) = \frac{\exp(x)}{1 + \exp(x)}$

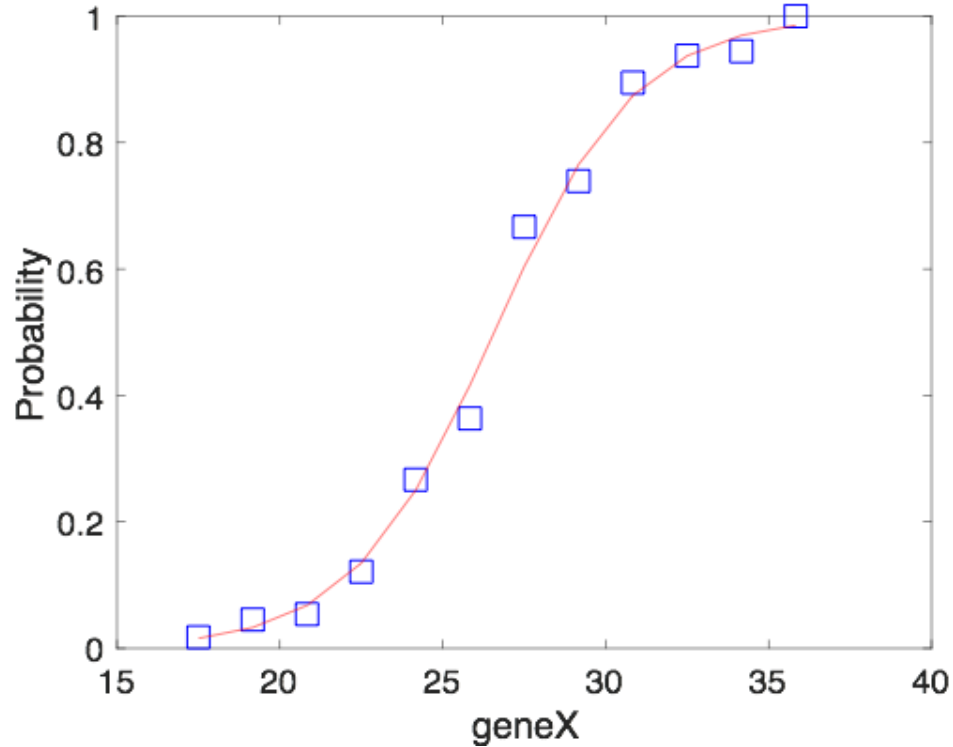


f(x) or y values always falls in range between 0 and 1

Solutions: Logistic regression

Logistic regression is the best model if response variable is binomial. Because it uses a fitting method that is appropriate for the binomial distribution.

Predicted proportions/probability values are present in the range from 0 to 1.



In matlab we use `glmfit` function to fit our data to a logistic model. This function returns coefficient estimates for a linear regression of the responses Y ($f(x)$) on the independent variable X

In Matlab,

```
%logistic regression
```

```
[logitCoef,dev,stats] = glmfit(geneX,[cancer  
tested],'binomial','logit');
```

```

geneX = [2180 2450 2640 2730 3100 3120 3320 3610 3800
% The number of patients tested at each levels (intervals)
tested = [57 44 37 33 30 22 21 23 19 16 18 21]';
% The number of cancer patients at each test
cancer = [1 2 2 4 8 8 14 17 17 15 17 21]';

```

	1	2	3	4	5
1	-12.6748				
2	0.3867				
3					
4					

```
%logistic regression
```

```

[logitCoef,dev,stats] = glmfit(geneX,[cancer tested],'binomial','logit');
logitFit = glmval(logitCoef,geneX,'logit');

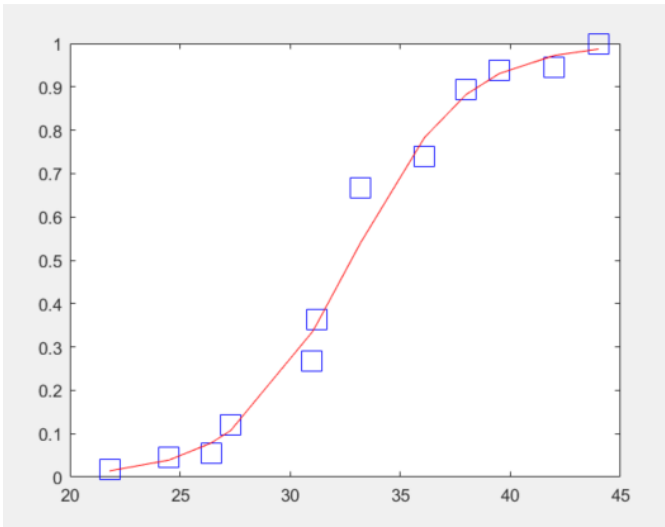
```

```

figure(3)
plot(geneX,proportion,'bs', geneX,logitFit,'r-','markersize',16);

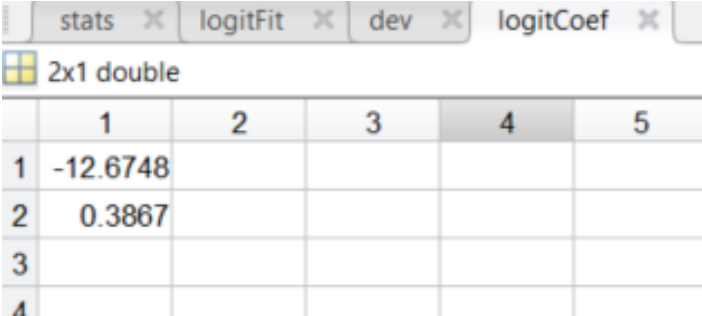
```

Glmval is uses to compute the predicted values for the model



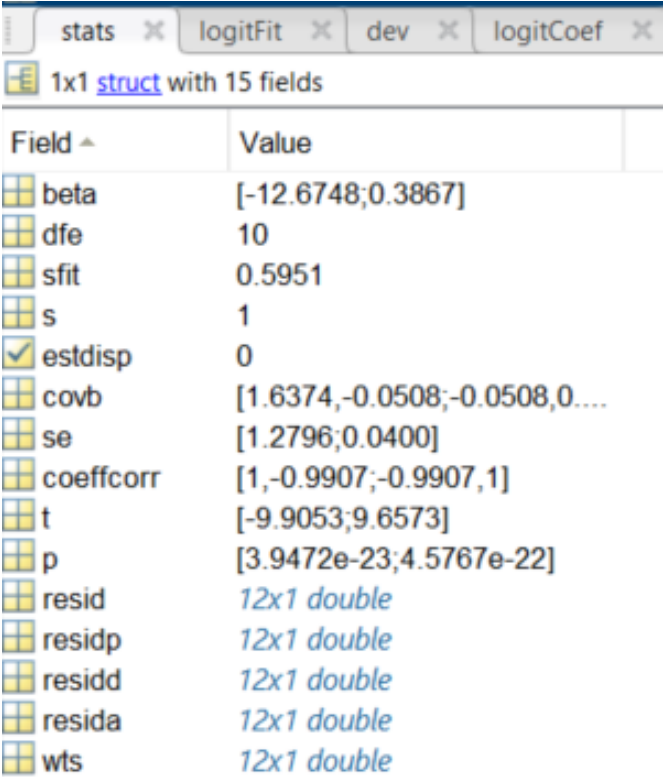
	1	2	3
1	0.0141		
2	0.0391		
3	0.0782		
4	0.1073		
5	0.3345		
6	0.3519		
7	0.5406		
8	0.7831		
9	0.8827		
10	0.9308		
11	0.9725		
12	0.9871		
13			
14			

glmfit: Logistic model coefficients



A MATLAB window titled 'logitCoef' showing a 2x1 double matrix. The matrix has 4 rows and 5 columns. The first row contains the values -12.6748, 0.3867, and three empty cells. The second row contains four empty cells. The third and fourth rows are also empty.

	1	2	3	4	5
1	-12.6748				
2	0.3867				
3					
4					



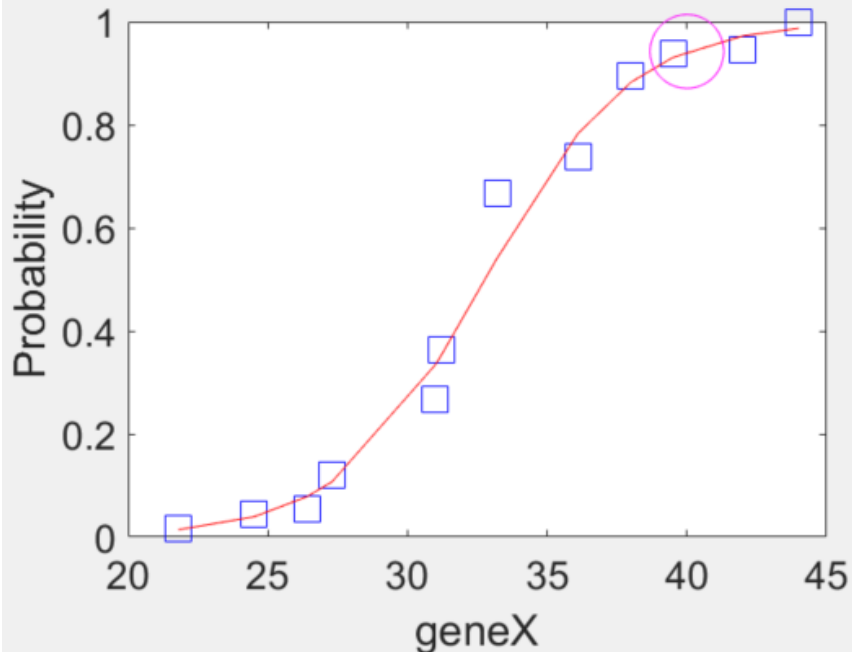
A MATLAB window titled 'logitFit' showing a 1x1 struct with 15 fields. The fields and their values are listed in a table below.

Field ^	Value
beta	[-12.6748;0.3867]
dfc	10
sfit	0.5951
s	1
<input checked="" type="checkbox"/> estdisp	0
covb	[1.6374,-0.0508;-0.0508,0....
se	[1.2796;0.0400]
coeffcorr	[1,-0.9907;-0.9907,1]
t	[-9.9053;9.6573]
p	[3.9472e-23;4.5767e-22]
resid	12x1 double
residp	12x1 double
residd	12x1 double
resida	12x1 double
wts	12x1 double

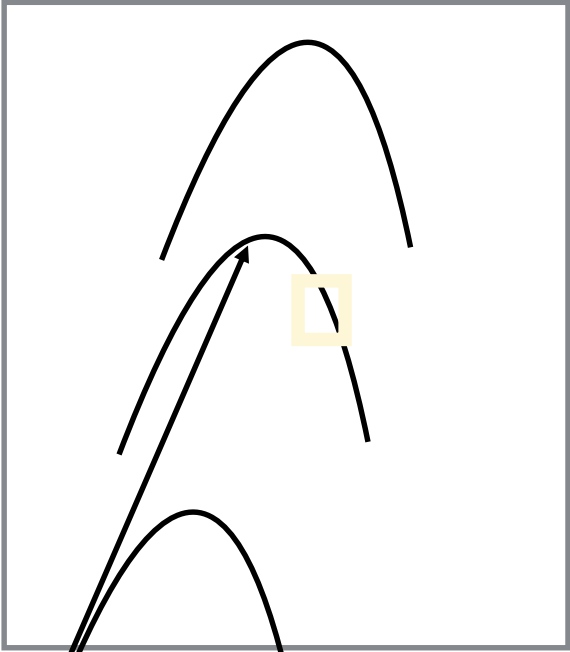
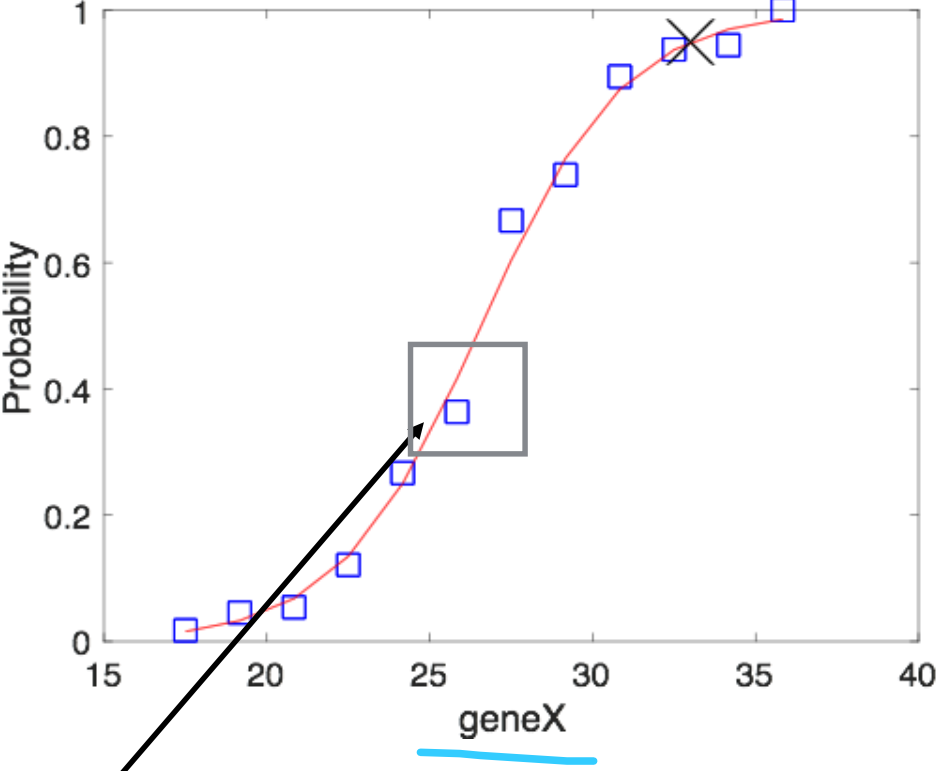
$$P(Y=1|\beta) = \frac{\exp(b(1)+b(2)x)}{1+\exp(b(1)+b(2)x)}$$

```
% prediction by using logistic model
% given that patient has an average RNA level from isolated cells
genepredict=40

% what is the risk of having cancer?
% model equation
cancerriskpro=exp(logitCoef(1)+genepredict*logitCoef(2))/(1+exp(logitCoef(1)+genepredict*logitCoef(2)))
% probability
disp(cancerriskpro)
figure(3)
plot(geneX,proportion,'bs', geneX,logitFit,'r-','markersize',16);
hold on
plot(genepredict,cancerriskpro,'mo','markersize',34);
xlabel('geneX');
ylabel('Probability');
set(gca,'fontsize',18)
```



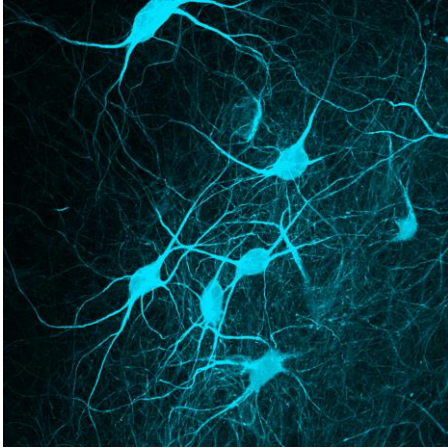
Coefficients are estimated by using a **maximum likelihood estimation method** where coefficients maximizes the prediction of observed values in the data



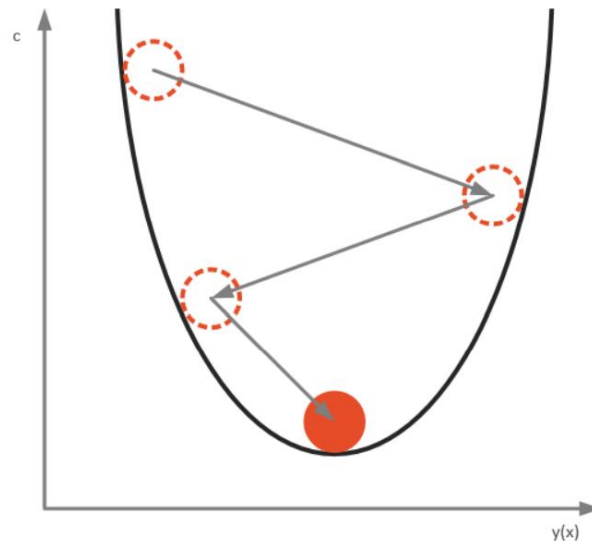
points on a line represents the highest points in the probability distribution

$$\log(\text{odds}) = b_0 + b_1x = -12.12 + 0.45x$$

Machine learning and Deep learning with Python



```
ents\Teaching\Python Programming\Codes and Notes\LectureX_deeplearning\handwritten_predictioncode.py
openfile.py x yourapplication.py x neuronalnetworks_example.py* x lecture1.py x lecture1_fall2021.py x handwritten_predictioncode.py* x
31
32
33 selnum=3
34
35 selD=data dev[:,selnum]
36 s=np.zeros((27,28))
37 jk=0
38 for i in range(27):
39     for j in range(28):
40         s[i,j]=selD[jk]
41         jk+=1
42
43 #
44
45 import pylab as plt
46
47 im = plt.imshow(s, cmap='hot')
48 plt.colorbar(im, orientation='horizontal')
49 plt.show()
50
51
52
53
54
55
56
57
```

A plot showing a handwritten digit '8' in white on a black background. The plot has x and y axes ranging from 0 to 25. To the right of the plot is a vertical toolbar with icons for various plot functions.

- There are many different machine learning models

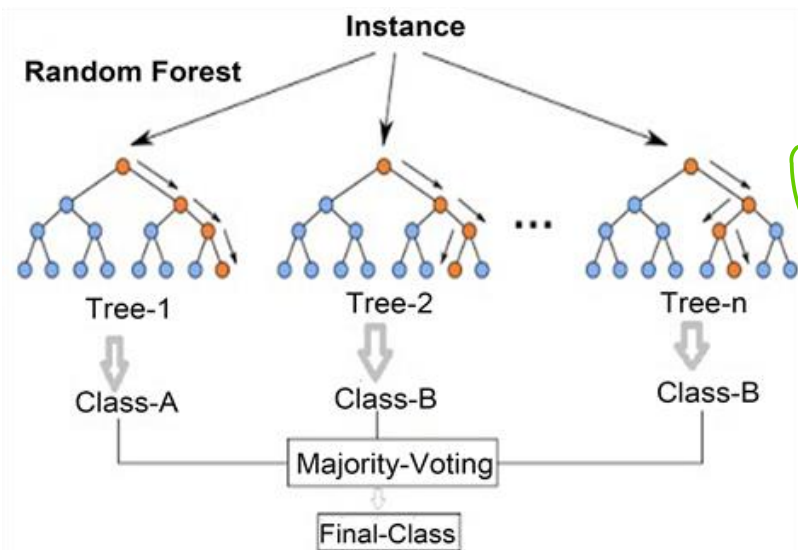
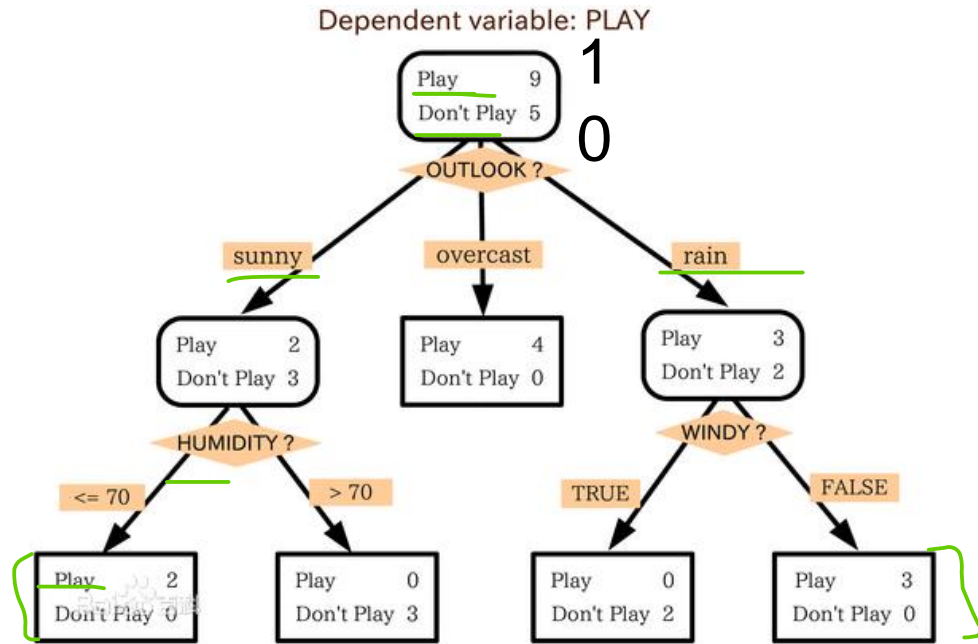
Including

Random Forest

Logistic Regression

Decision trees

Support Vector



	feature1	feature2	feature3
Person1	0	1	
Person2	1	0	
Person3	0		
.	1		
.			
.			
Personn			

Machine Learning

It is the learning process for understanding the data sets and use this knowledge to answer the questions.
Can be used to discover for new knowledge.

The goals are

- To improve the learning system and apply learning systems
- To perform the learning with these systems and train your model
- To apply the model and answer the questions

Our goal is to help you to understand the model selection within the machine learning that can be used to solve the real world problems

Machine Learning Types

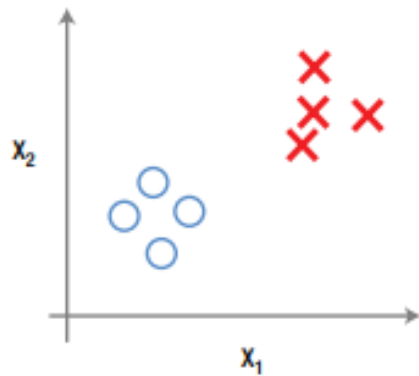
Supervised learning

Unsupervised learning

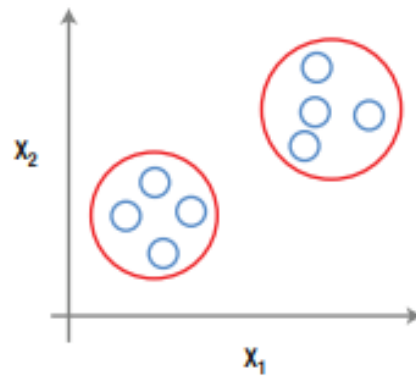
Semi-supervised learning

Reinforcement learning

Supervised learning



Unsupervised learning



supervised learning require supervision to train the model. This supervision is necessary for classification where we have labeled data on which we train the model to predict the labels of the unseen data.

Scikit_learn library



[Install](#) [User Guide](#) [API](#) [Examples](#) [Community](#) [More](#)

scikit-learn

Machine Learning in Python

Getting Started

Release Highlights for 1.3

GitHub

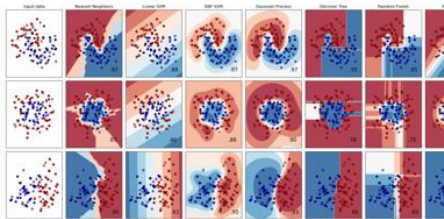
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: Gradient boosting, nearest neighbors, random forest, logistic regression, and more...



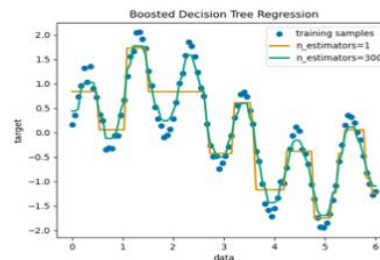
Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...



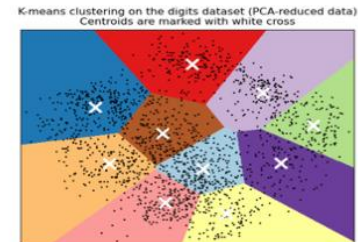
Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...



Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Model selection

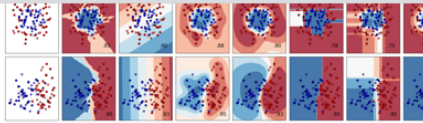
Comparing, validating and choosing parameters and models.

Preprocessing

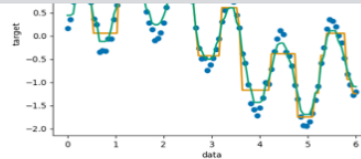
Feature extraction and normalization.

Scikit_learn library

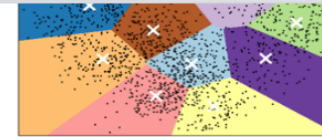
<https://scikit-learn.org/stable/index.html>



Examples



Examples

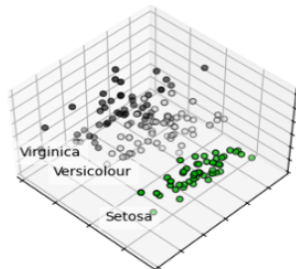


Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization, and more...

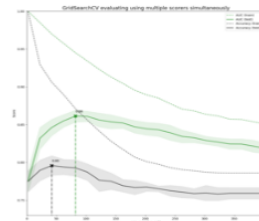


Examples

Model selection

Comparing, validating and choosing parameters and models.

Applications: Improved accuracy via parameter tuning
Algorithms: grid search, cross validation, metrics, and more...

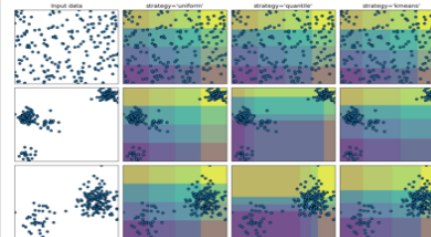


Examples

Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.
Algorithms: preprocessing, feature extraction, and more...



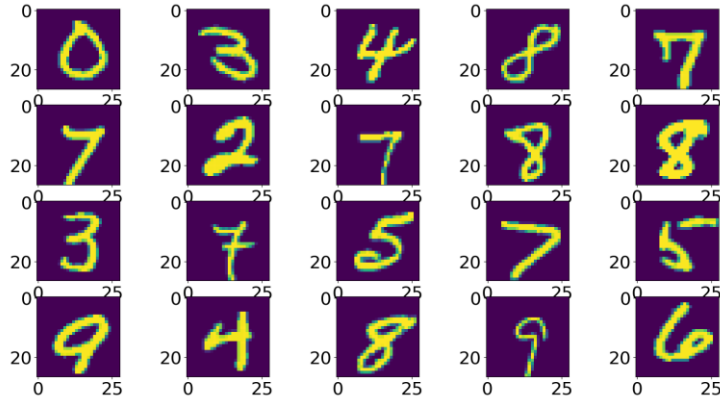
Examples

News

Community

Who uses scikit-learn?

Example 1: Reading handwritten digits with deep learning



```
selD=data_dev[:,selnum]
s=np.zeros((27,28))
jk=0
allnumbers=42000*['']
for k in range(1000):
    selD=data_dev[:,k]
    s=np.zeros((27,28))
    jk=0

    for i in range(27):
        for j in range(28):
            s[i,j]=selD[jk]
            jk+=1
            allnumbers[k]=s

import pylab as plt

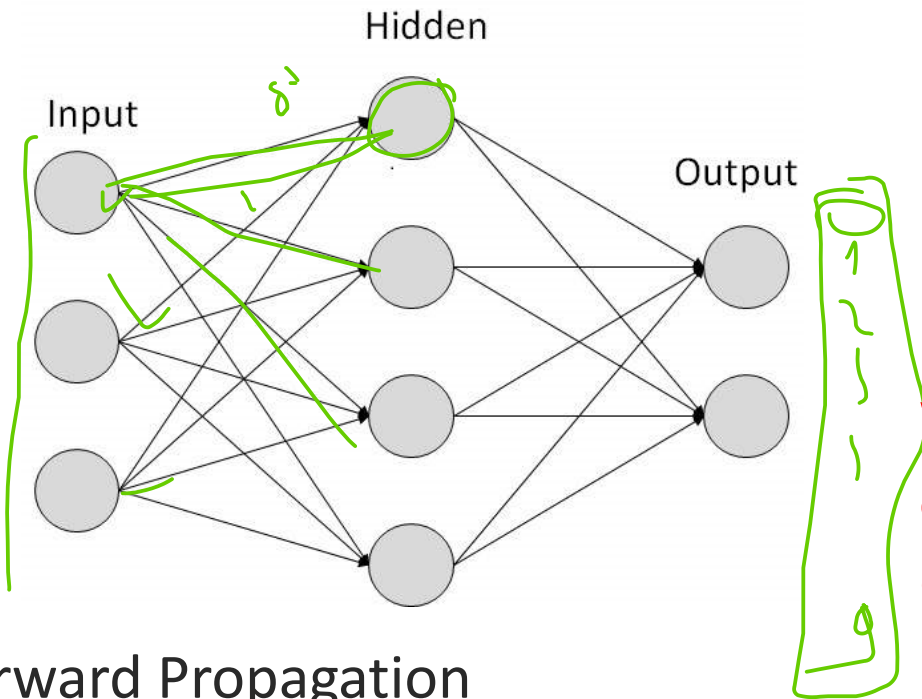
im = plt.imshow(s, cmap='gray')
plt.colorbar(im, orientation='vertical')
plt.show()
```

	0	1	2	3	4	5	6	7	
0	3	0	0	0	0	0	0	0	
1	9	0	0	0	0	0	0	0	0
2	7	0	0	0	0	0	0	0	0
3	6	0	0	0	0	0	0	0	0
4	4	0	0	0	0	0	0	0	0
5	6	0	0	0	0	0	0	0	0
6	3	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	4	0	0	0	0	0	0	0	0
9	3	0	0	0	0	0	0	0	0
10	8	0	0	0	0	0	0	0	0
11	2	0	0	0	0	0	0	0	0
12	4	0	0	0	0	0	0	0	0
13	8	0	0	0	0	0	0	0	0

Format Resize Background color

Save and Close Close

Deep learning with python



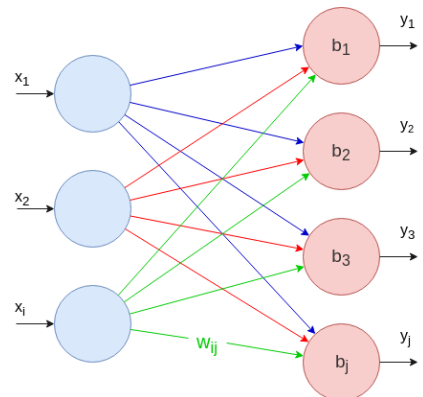
We find the most optimum weights for each output

Forward Propagation

The value of each output neuron can be calculated as the following :

$$y_j = b_j + \sum_i x_i w_{ij}$$

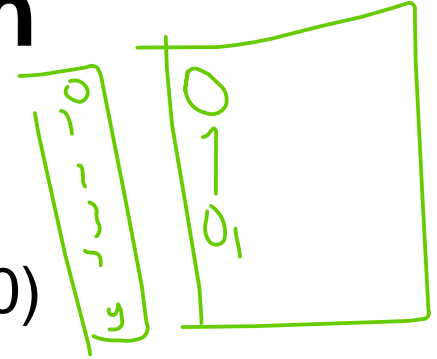
b is the bias



Training the data with a known values

output1, output2, output3,

output42=trainingdata(X_train, Y_train, 0.10, 500)

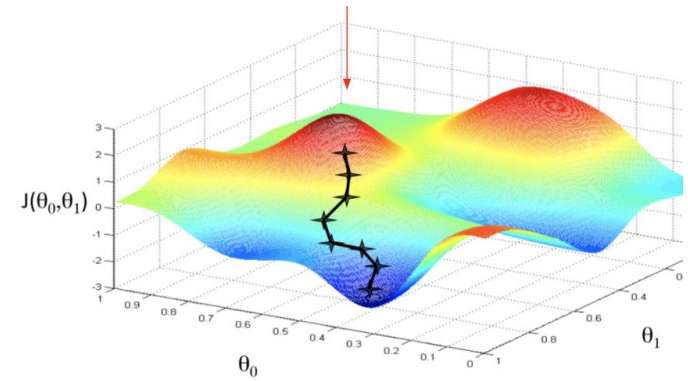


```
[0 7 3 ... 8 5 5] [0 3 3 ... 8 5 5]
0.8290487804878048
Iteration: 460
[0 7 3 ... 8 5 5] [0 3 3 ... 8 5 5]
0.831
Iteration: 470
[0 7 3 ... 8 5 5] [0 3 3 ... 8 5 5]
0.8328292682926829
Iteration: 480
[0 7 3 ... 8 5 5] [0 3 3 ... 8 5 5]
0.834780487804878
Iteration: 490
[0 7 3 ... 8 5 5] [0 3 3 ... 8 5 5]
0.8364878048780487
```

In [164]:

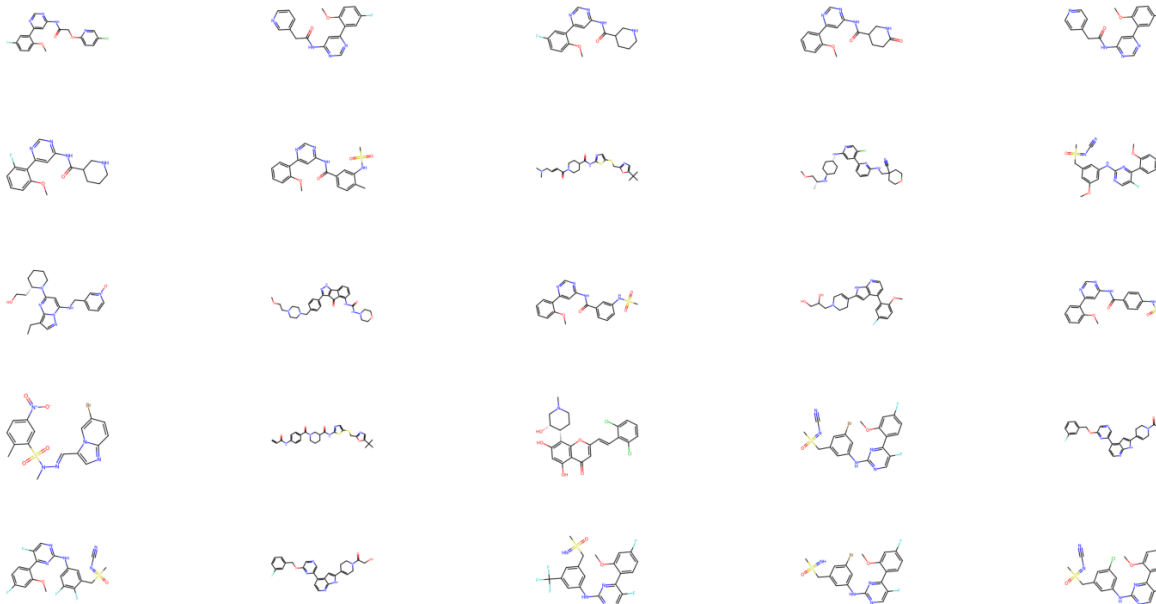
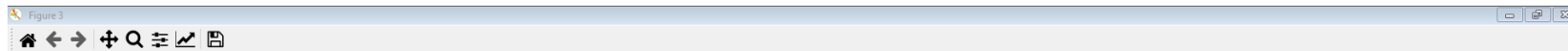
for i in range(80):

test_prediction(i, W1, b1, W2, b2)



```
Console 1/A X
True value of the digit: 0
Predicted number by the model: [3]
True value of the digit: 3
Predicted number by the model: [2]
True value of the digit: 2
Predicted number by the model: [8]
True value of the digit: 8
Predicted number by the model: [6]
True value of the digit: 6
Predicted number by the model: [6]
True value of the digit: 6
Predicted number by the model: [9]
True value of the digit: 8
Predicted number by the model: [2]
True value of the digit: 2
In [174]:
```

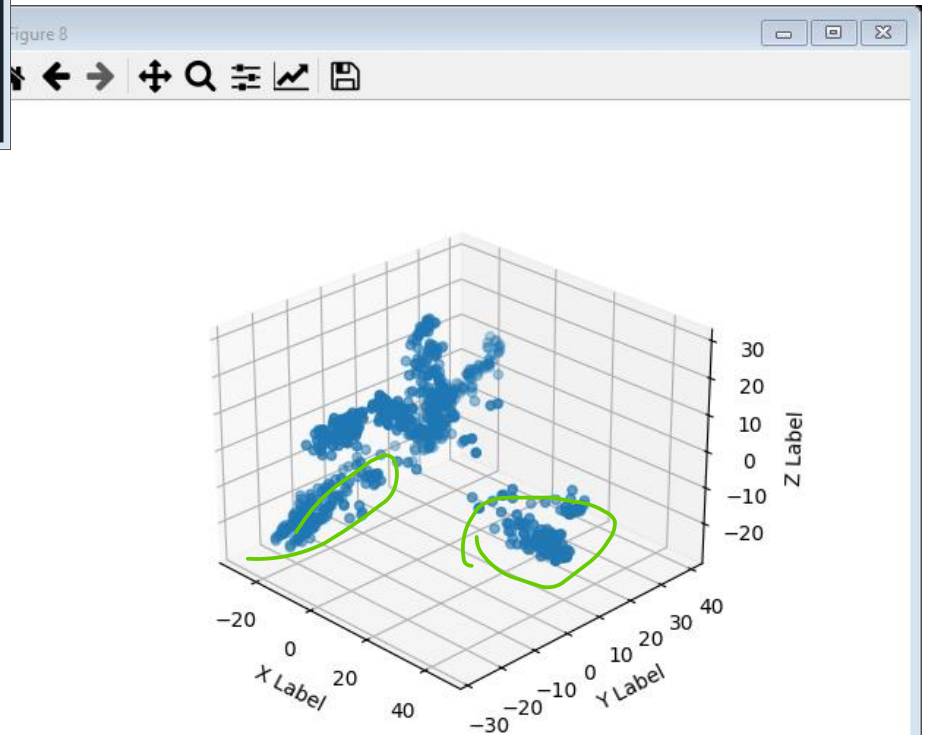
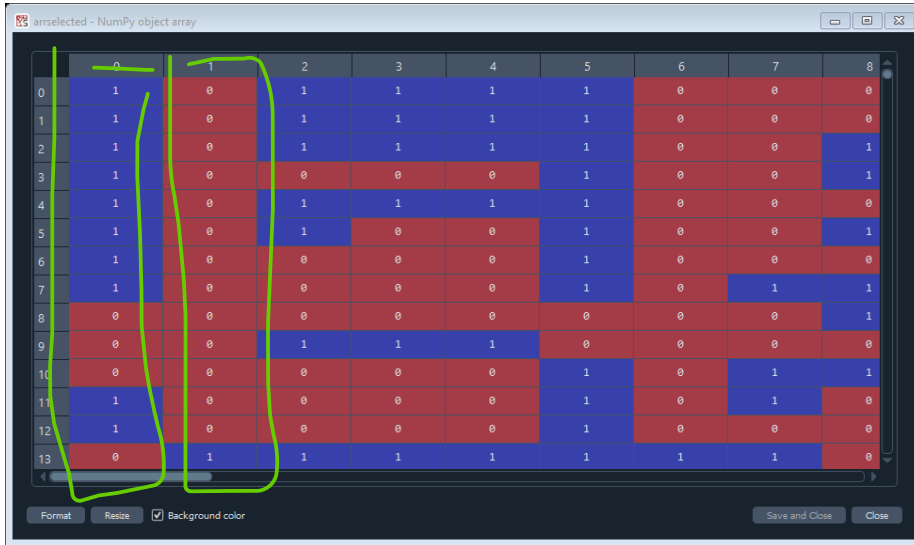
Example 2: Inhibitors to control cell motility



Unsupervised machine learning methods
Example: Principal component analysis

Principal Component Analysis :

It is an example of unsupervised machine learning



Principal Component Analysis :

It is a geometry based transformation of the numerical data

Mainly used

Dimensionality reduction

Higher dimensional data plotting in lower dimensional space

Data classification for machine learning algorithms

Unsupervised Learning

no supervision from the data are used while training the model.

We check if any clusters are present

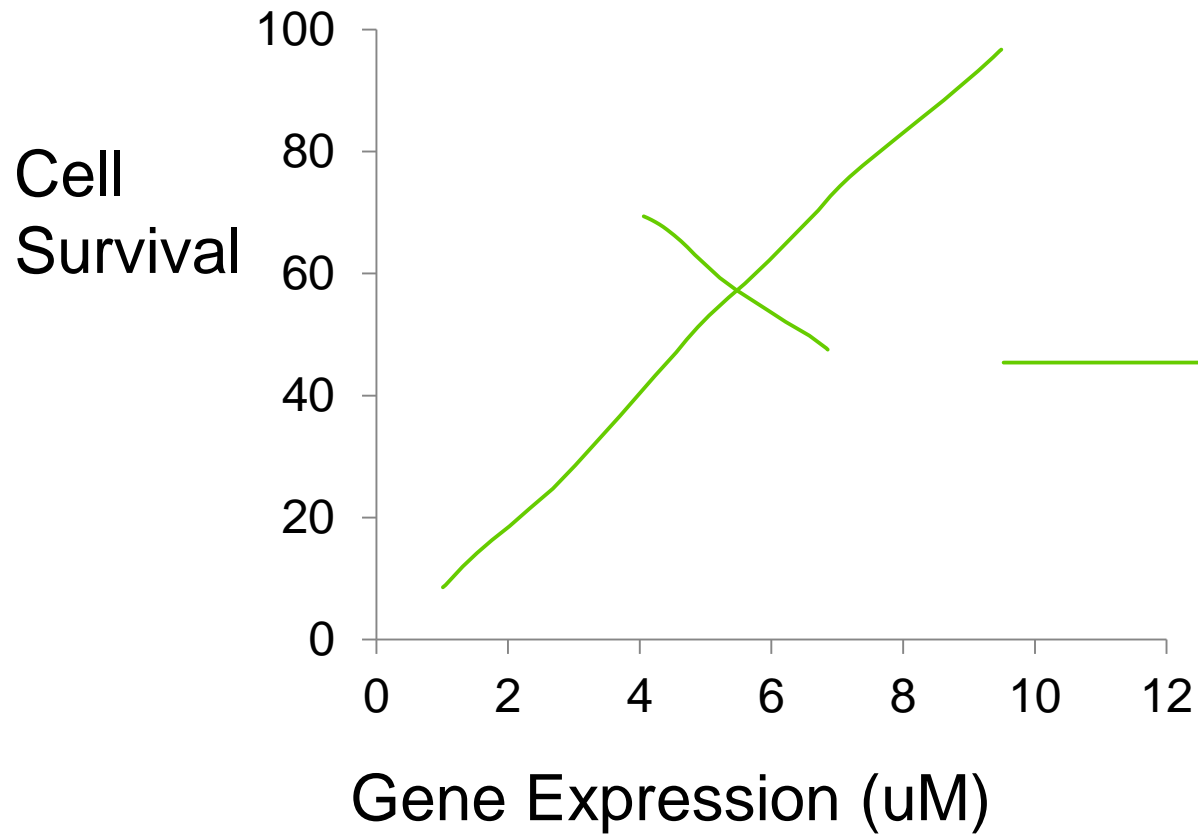
The discovered labels (for example with kmeans method) then become the basis for classifying any new unseen data.

Principal Component Analysis :

It is an example of unsupervised machine learnign

- PCA is a mathematical method to analyze complex and large data sets.
- Covariance can be considered to be a measure of how well correlated two variables are.

Lets start with a simple example:



Can we transform the coordinate system?

Statistics review

Mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

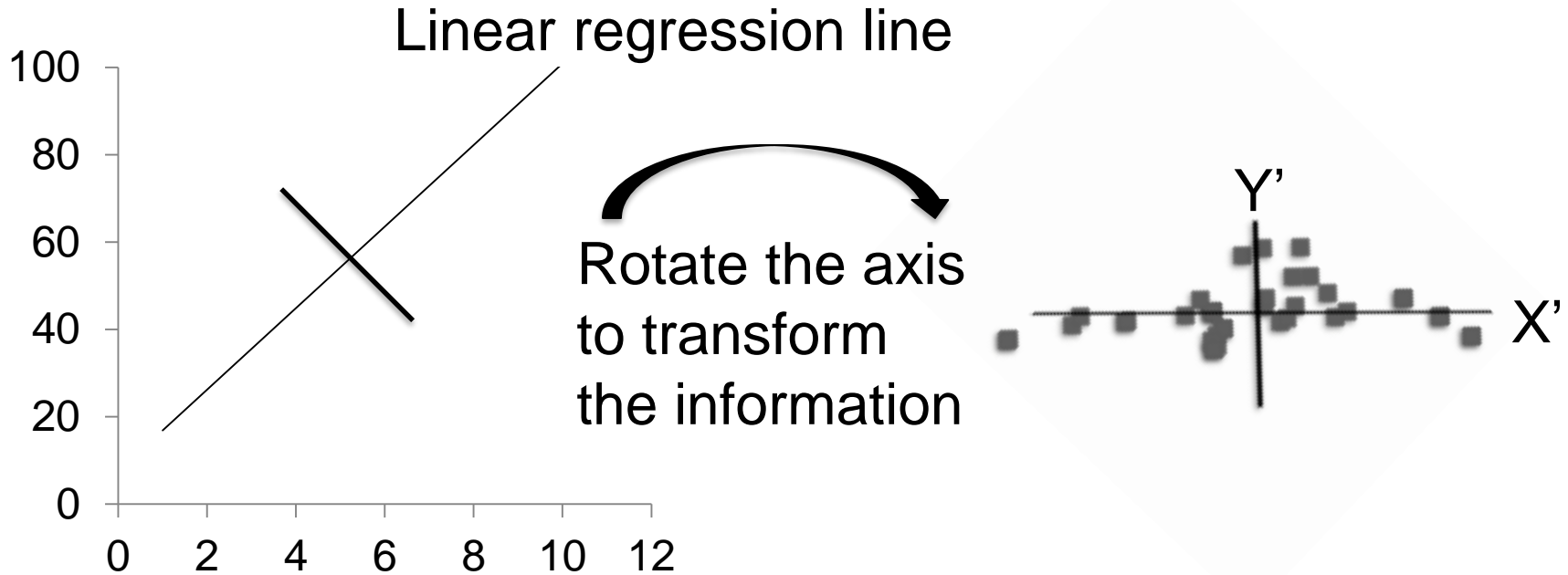
Variance : Determines the spread of only one variable. It measure 1 dimension and independent of other dimension.

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{var}(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{(n - 1)}$$

Covariance: Determines how two variables are related. It is always the measured between 2 dimensions.

$$\text{cov}(x, y) = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Is there any other way to represent the data?

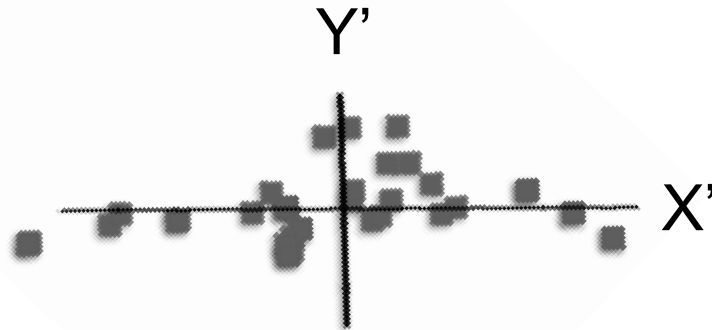


Find the regression line by linear fitting. It represents the largest variance in the data.

X' is the measure of the size.

Y'

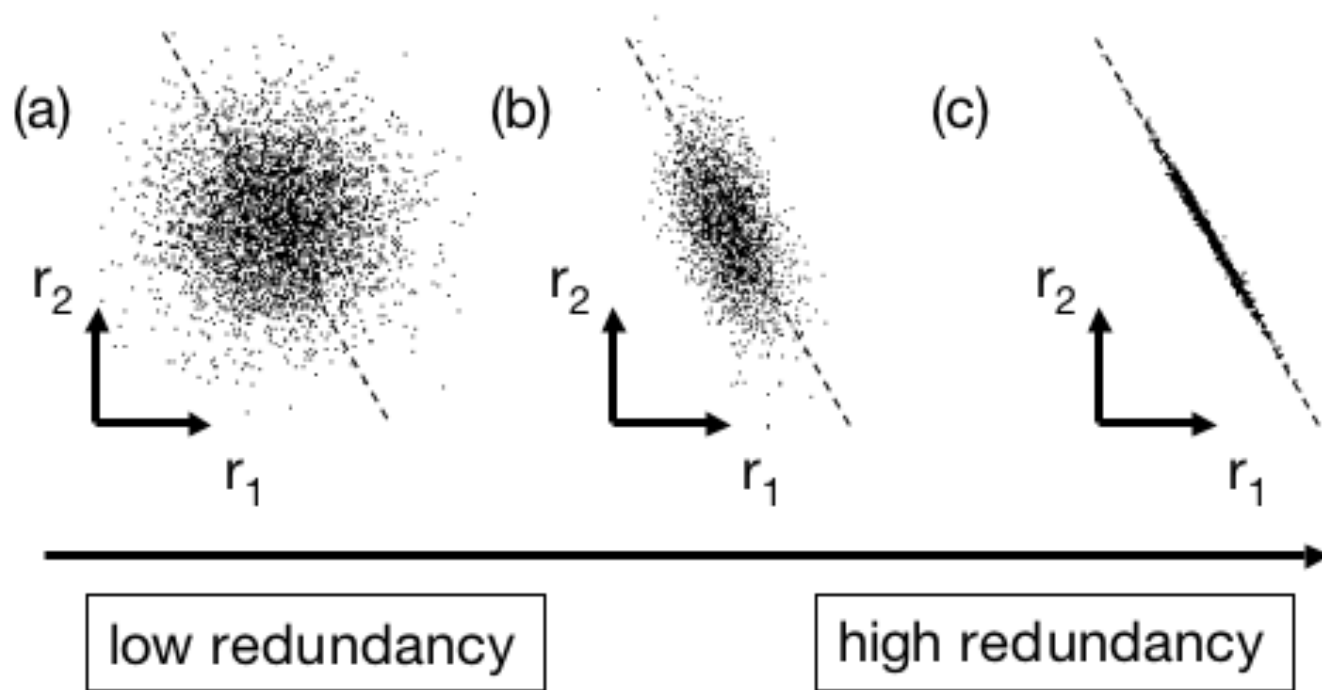
We can also determine the signal to noise ratio qualitatively by changing the coordinate system



X' is an indication of variance of signal σ^2_{signal}

Along the Y' axis, we observe the variance of noise σ^2_{noise}

$$\text{Signal-to-noise ratio (SNR)} = \sigma^2_{\text{signal}} / \sigma^2_{\text{noise}}$$

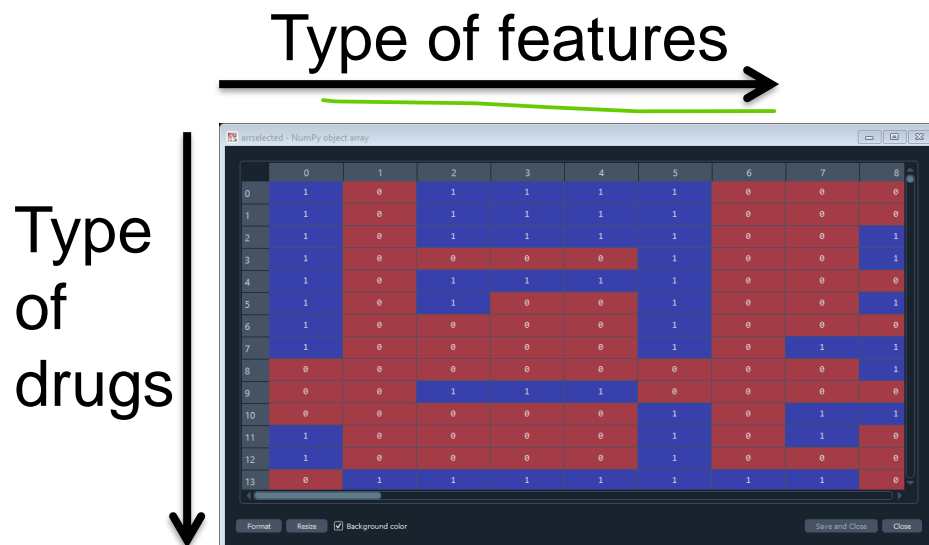


What if we are interested in many drugs? How to find the new basis?

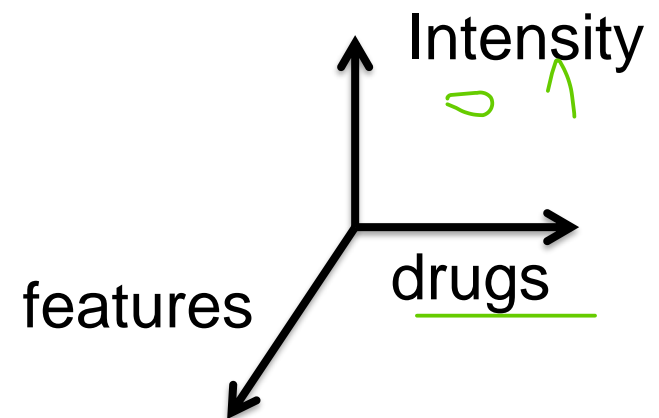
PCA is a tool that helps to find the relation of variables.

Which of the drugs are related?

What are the drugs that are active for target proteins but not for others?

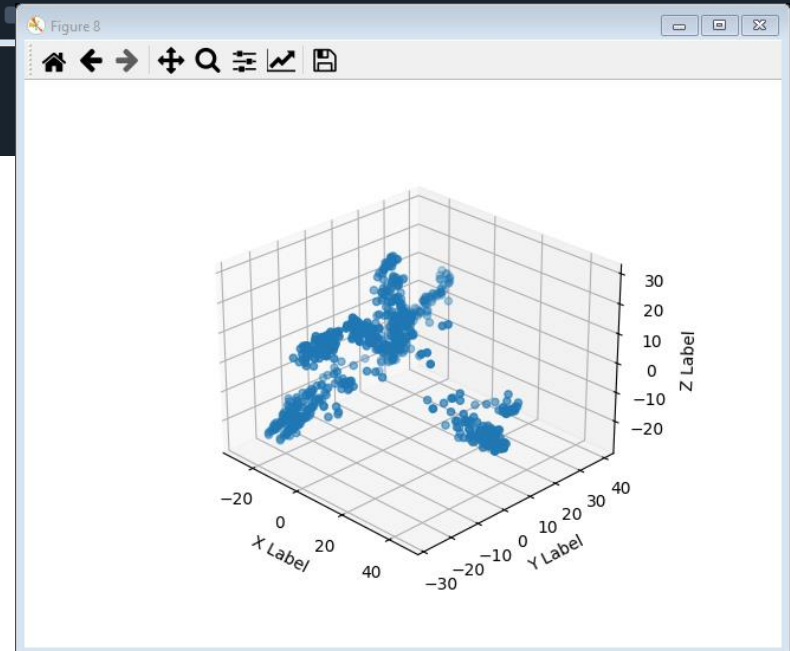
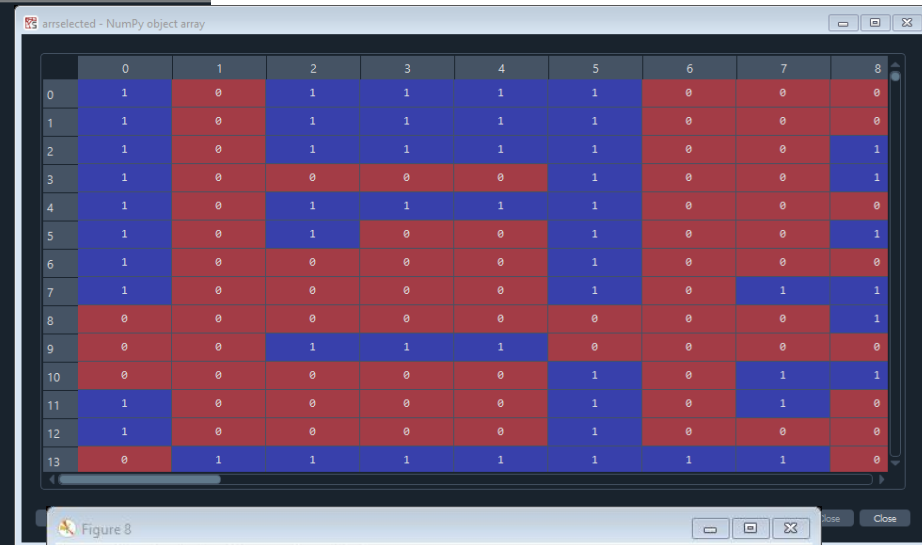


Colors define the intensity

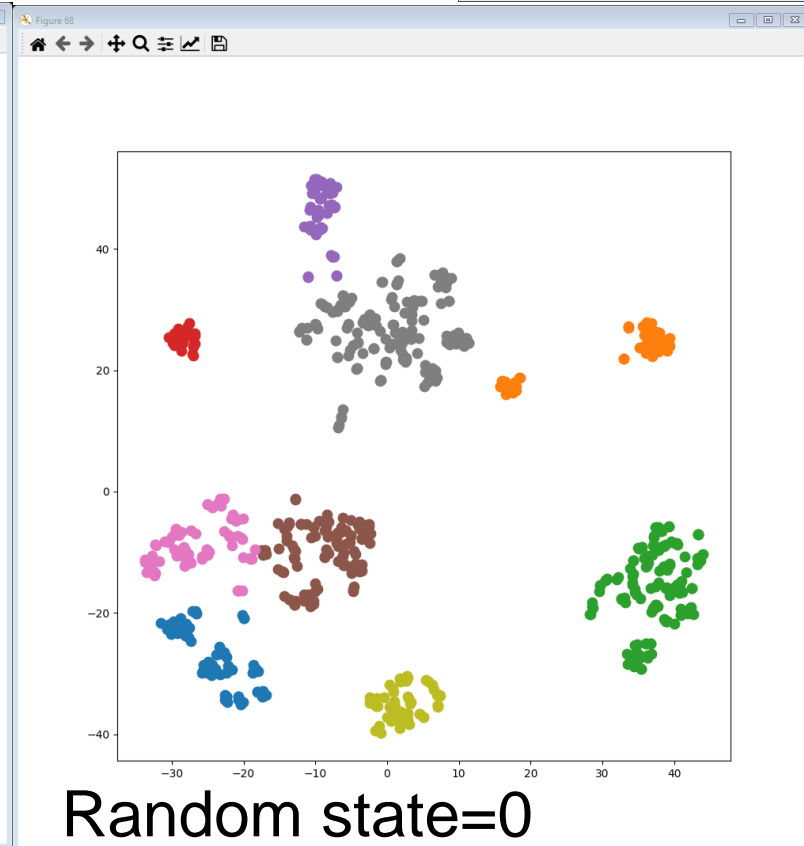
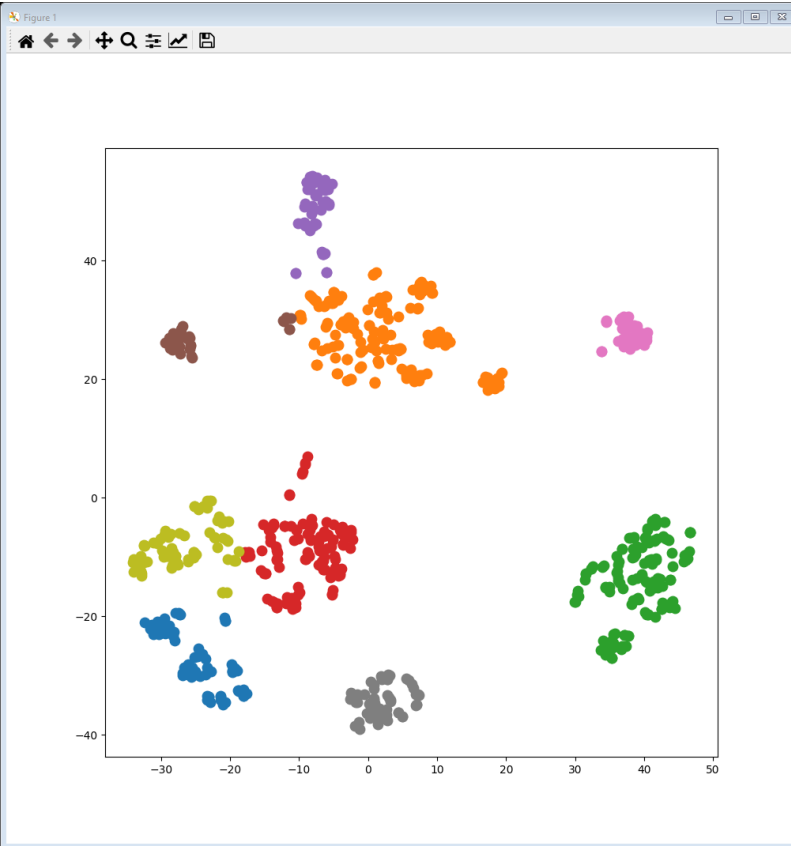
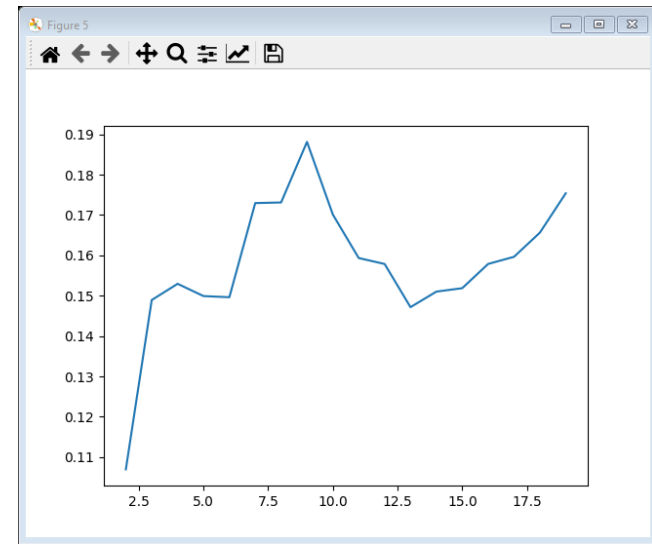


PCA for drug discovery

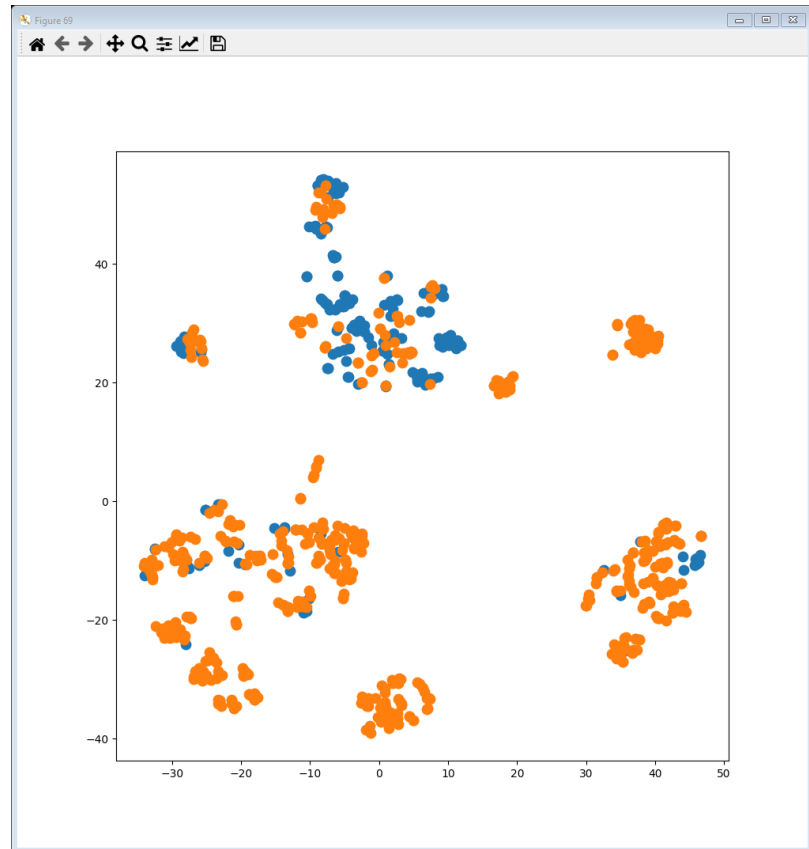
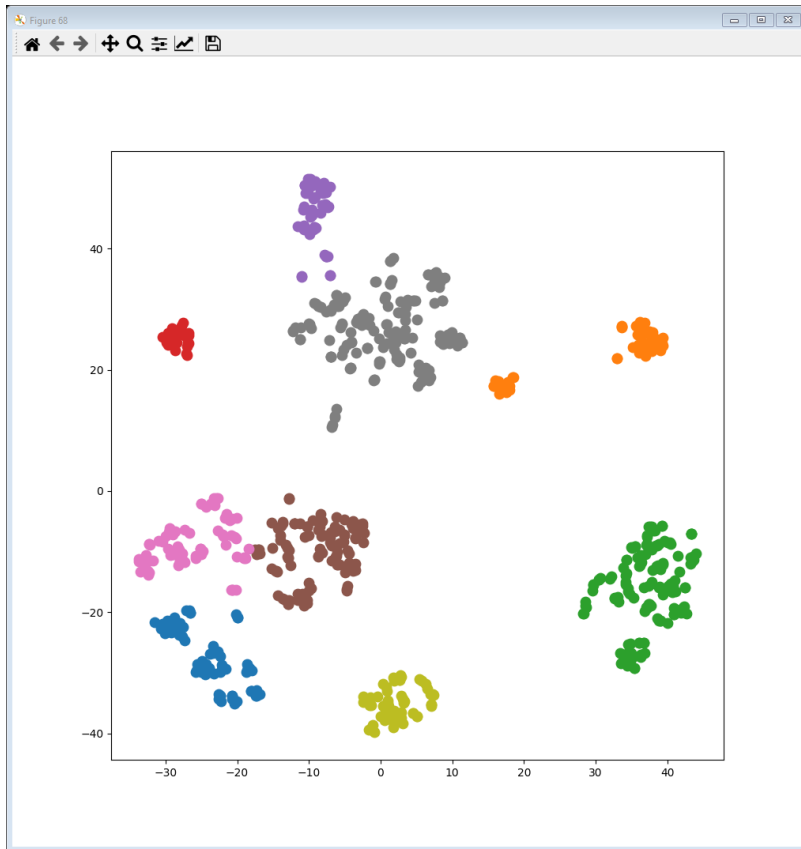
```
#%%  
  
pca = PCA(n_components=3)  
components = pca.fit_transform(arrselected)  
  
X=components  
  
fig = plt.figure(8)  
ax = fig.add_subplot(projection='3d')  
  
co=10  
  
ax.scatter(co*X[:1800,0], co*X[:1800,1], co*X[:1800,2])  
  
ax.set_xlabel('X Label')  
ax.set_ylabel('Y Label')  
ax.set_zlabel('Z Label')  
  
plt.show()
```

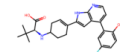
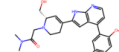
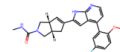
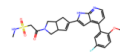
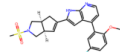
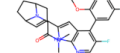
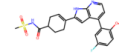
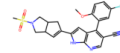
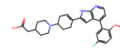
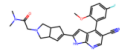
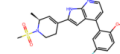
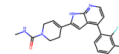
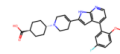
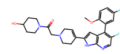
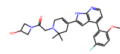
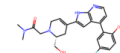
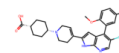
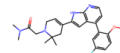
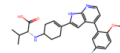
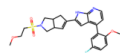
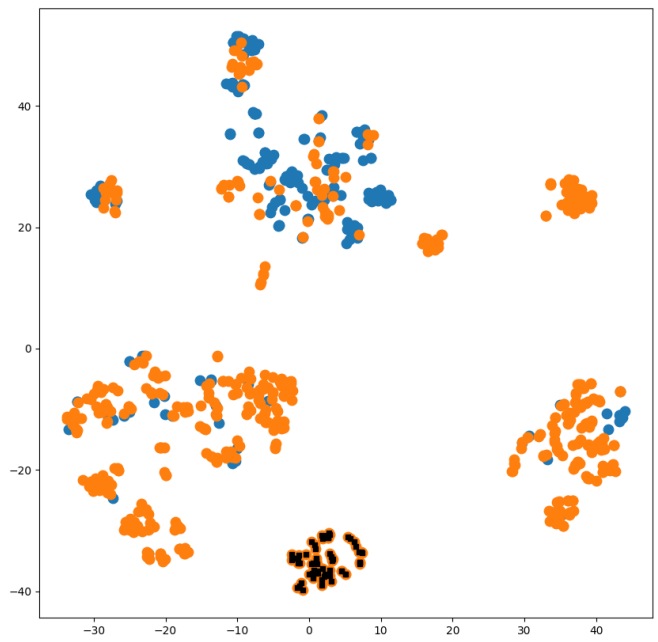


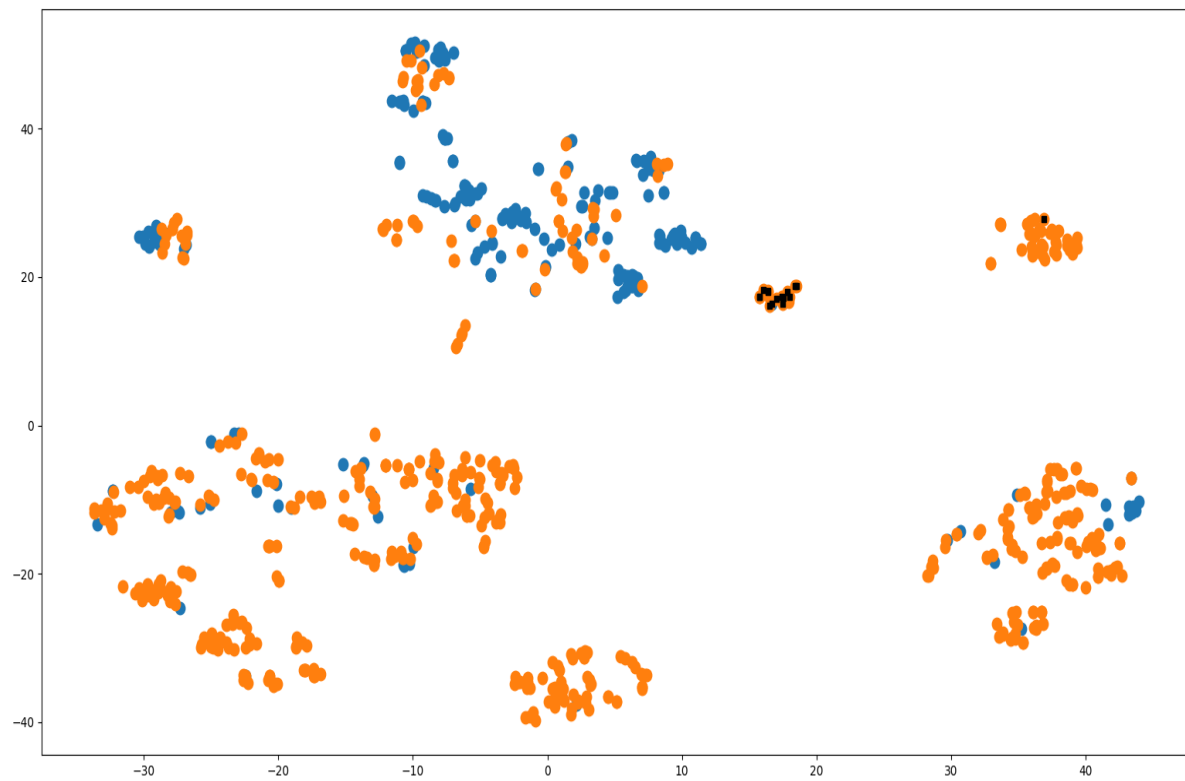
Silhouette score is used to determine the optimum number of clusters



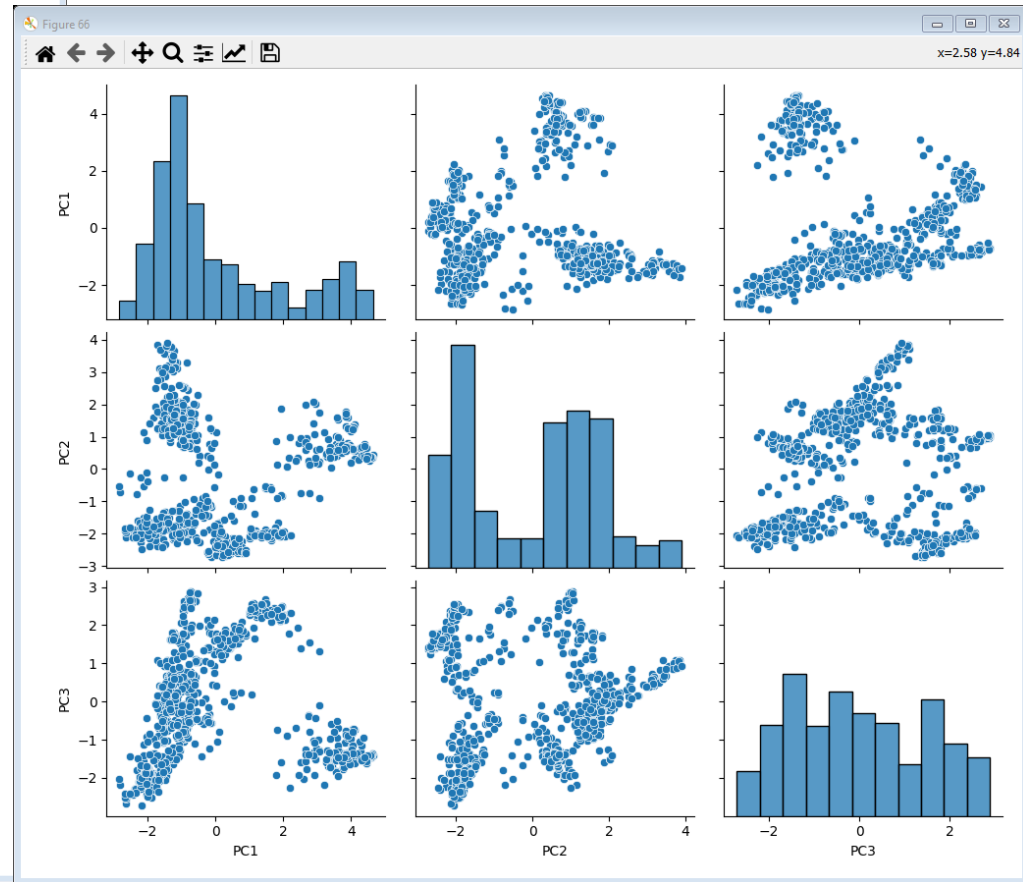
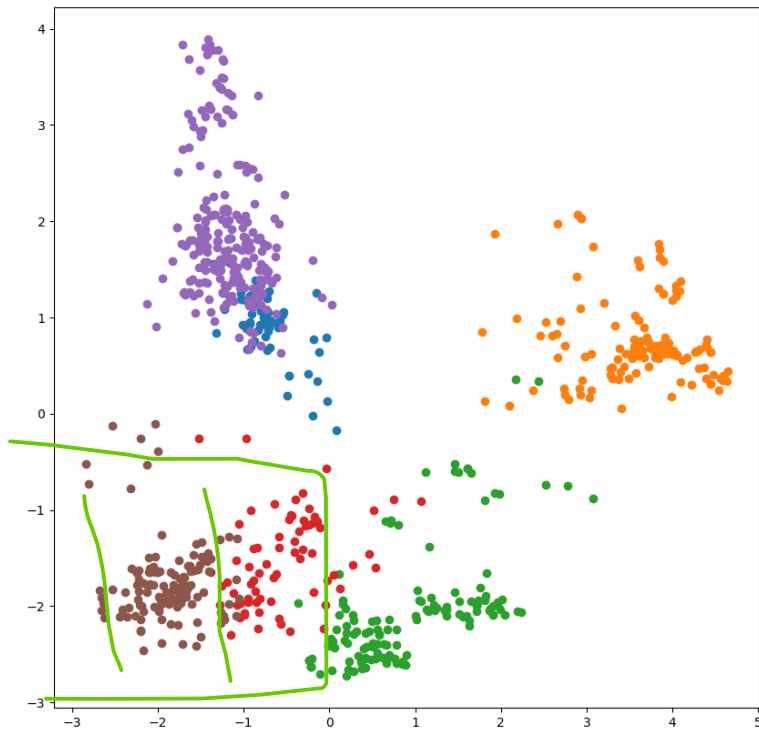
Finding the class of each compounds and compare it with inhibitor/inactive map



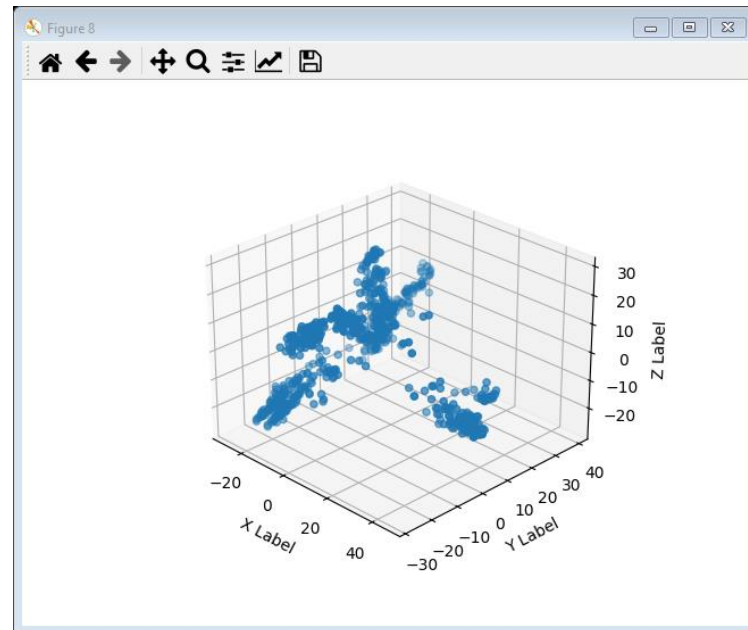




```
#####  
model = KMeans(n_clusters = 6, init = "k-means++")  
label = model.fit_predict(X)  
plt.figure(figsize=(10,10))  
uniq = np.unique(label)  
for i in uniq:  
    plt.scatter(components[label == i , 0] , components[label == i ,  
#plt.scatter(1*X[:1800,0], 1*X[:1800,1], marker="x", color='k')  
#This is done to find the centroid for each clusters.
```



Is there any clustering of drug molecules?

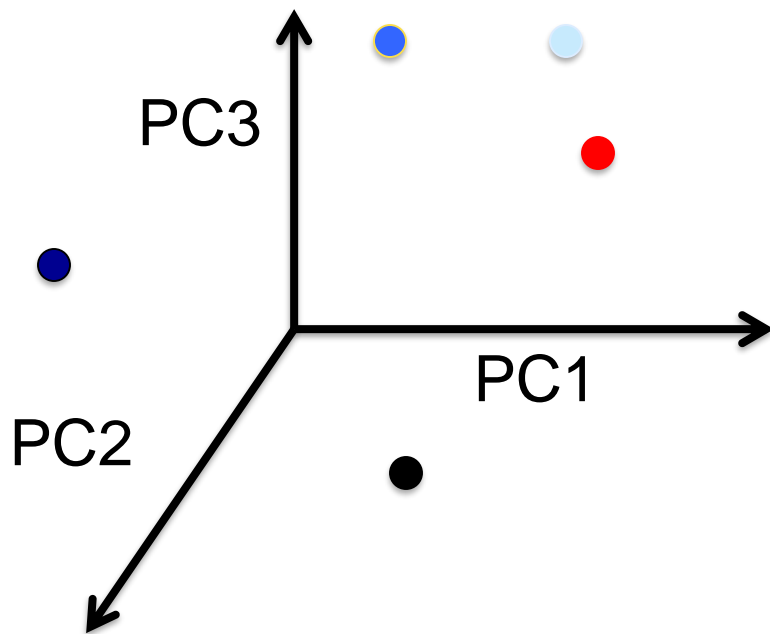


$$Y = P A$$

Transforming
information in
multivariable data set

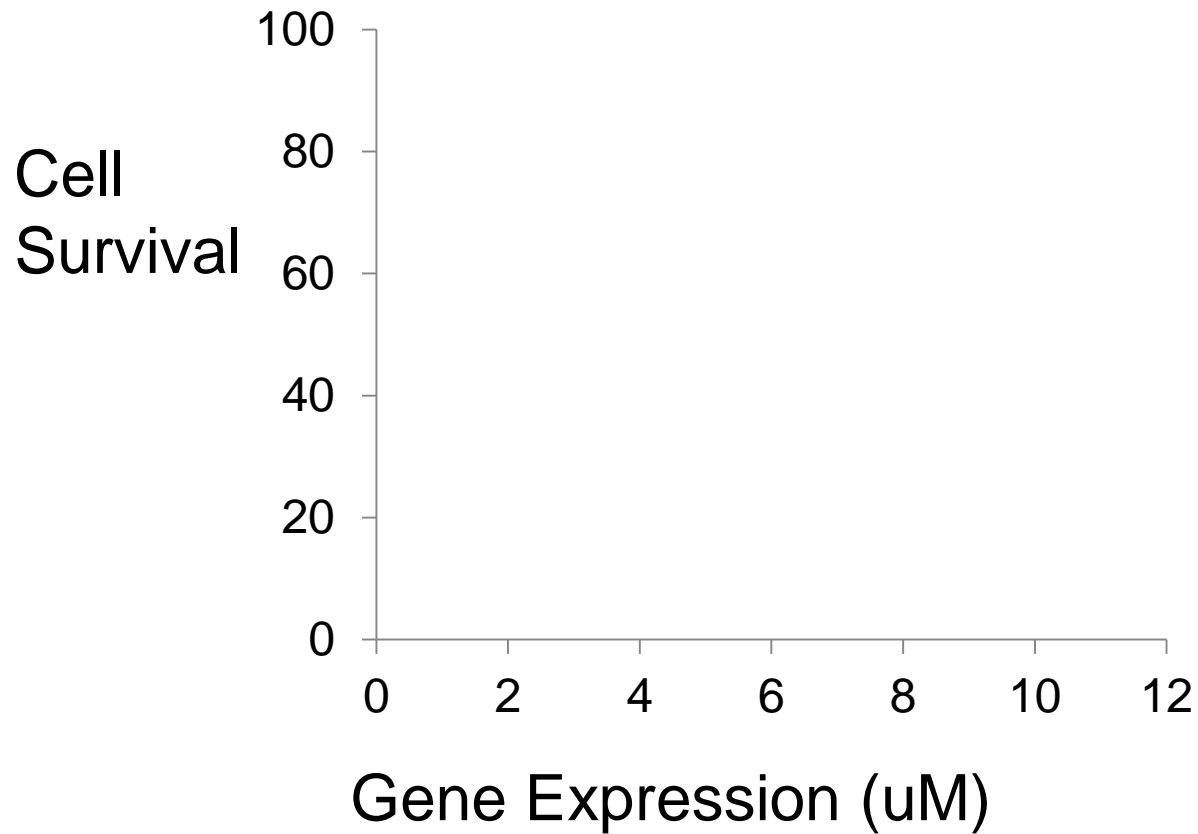
What information can be obtained in computational study of drug molecules?

1. Prediction of the class of drugs: The key drugs can be identified. The effectiveness of these drugs can be predicted.



2. Building large network of drugs: construct a graph that show the dependency of these drug molecules

Lets start with a simple example:



Can we transform the coordinate system?

Statistics review

Mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

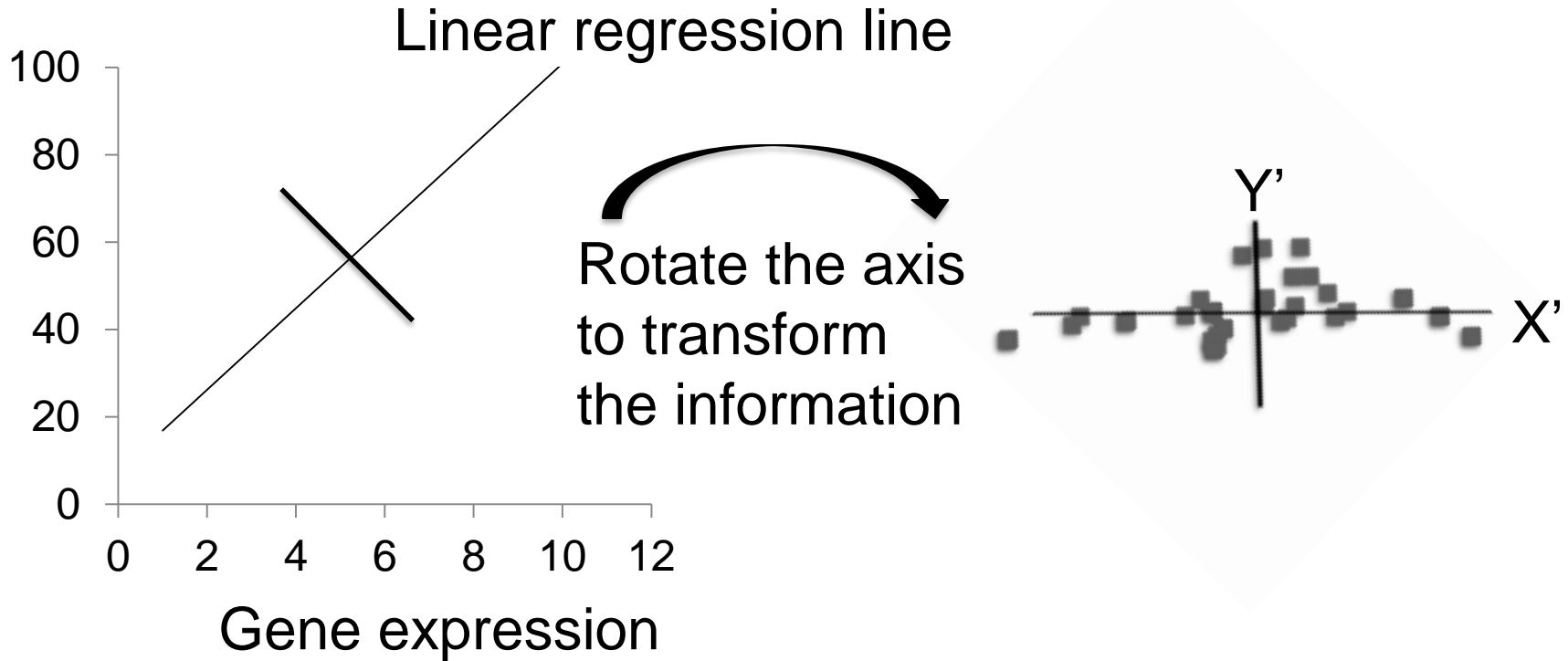
Variance : Determines the spread of only one variable. It measure 1 dimension and independent of other dimension.

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{var}(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{(n - 1)}$$

Covariance: Determines how two variables are related. It is always the measured between 2 dimensions.

$$\text{cov}(x, y) = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Is there any other way to represent the data?

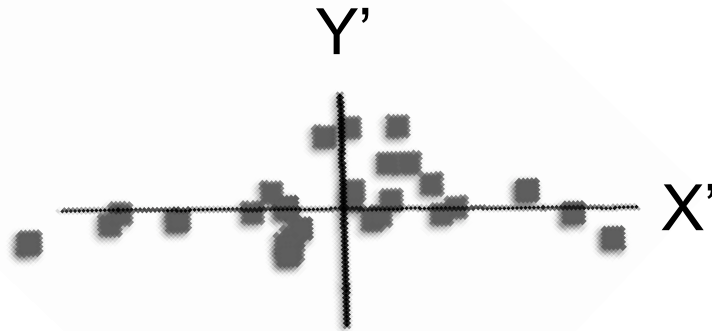


Find the regression line by linear fitting. It represents the largest variance in the data.

X' is the measure of the size.

Y' is the ratio of cell survival to gene expression

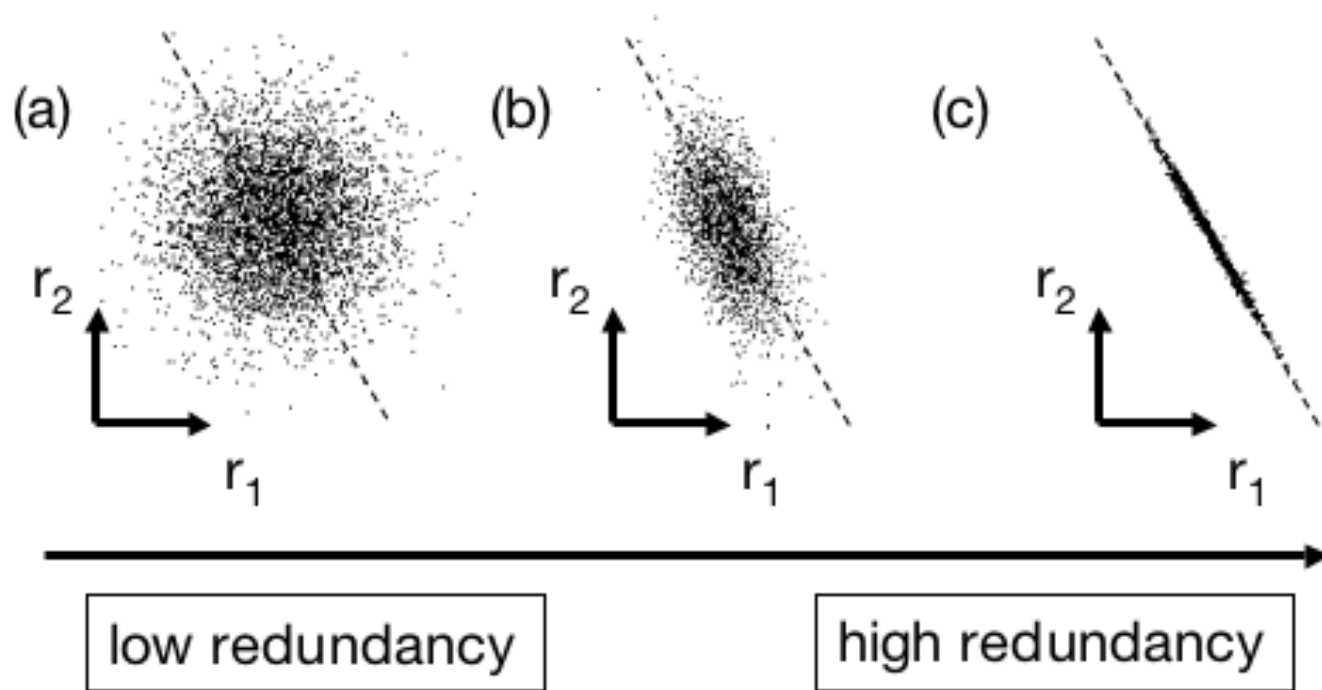
We can also determine the signal to noise ratio qualitatively by changing the coordinate system



X' is an indication of variance of signal σ^2_{signal}

Along the Y' axis, we observe the variance of noise σ^2_{noise}

$$\text{Signal-to-noise ratio (SNR)} = \sigma^2_{\text{signal}} / \sigma^2_{\text{noise}}$$



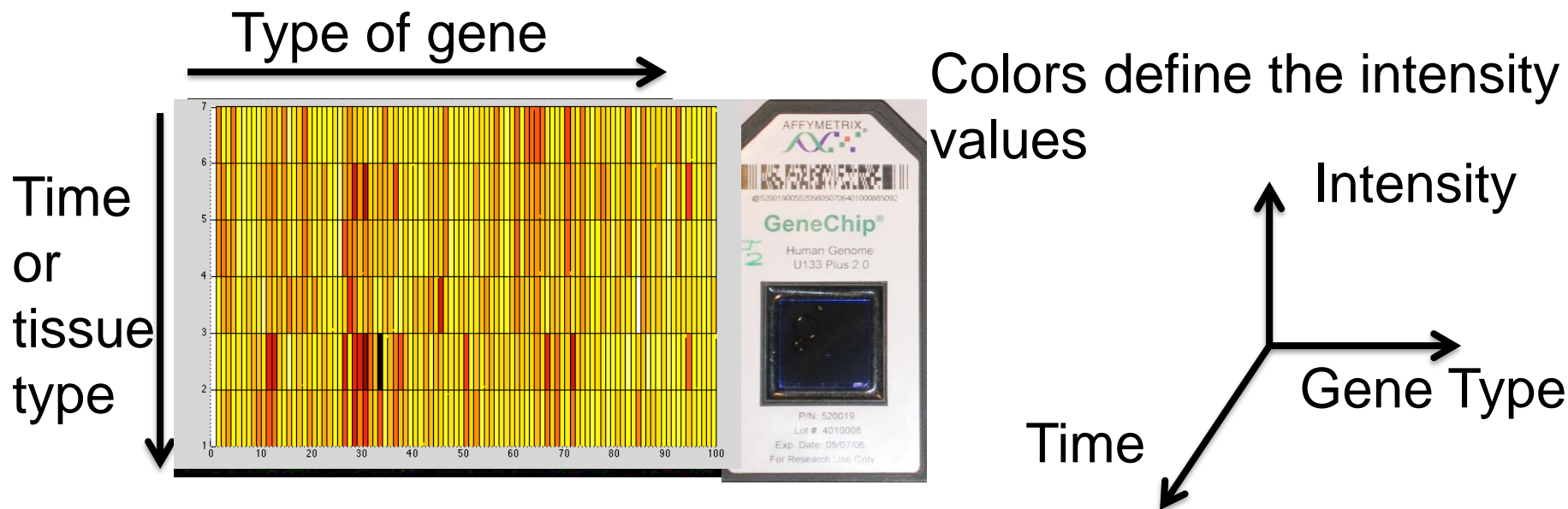
What if we are interested in many genes? How to find the new basis?

PCA is a tool that helps to find the relation of variables.

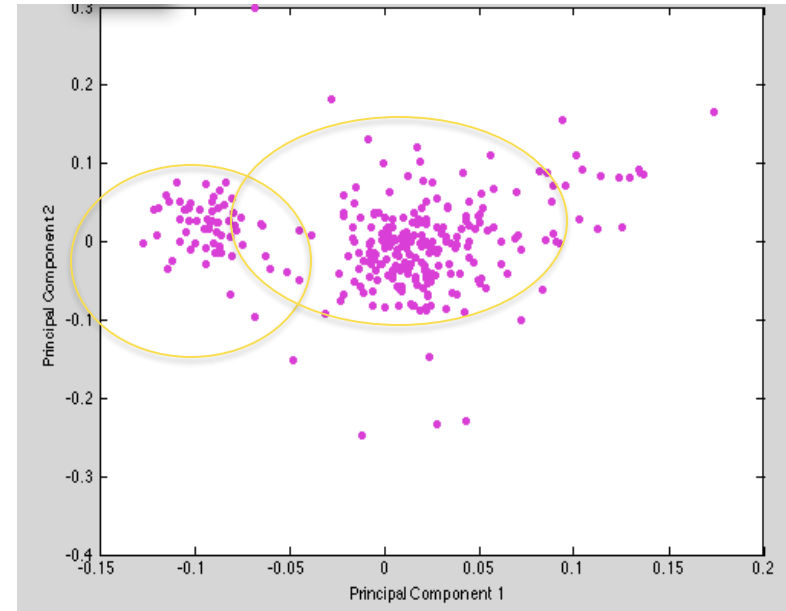
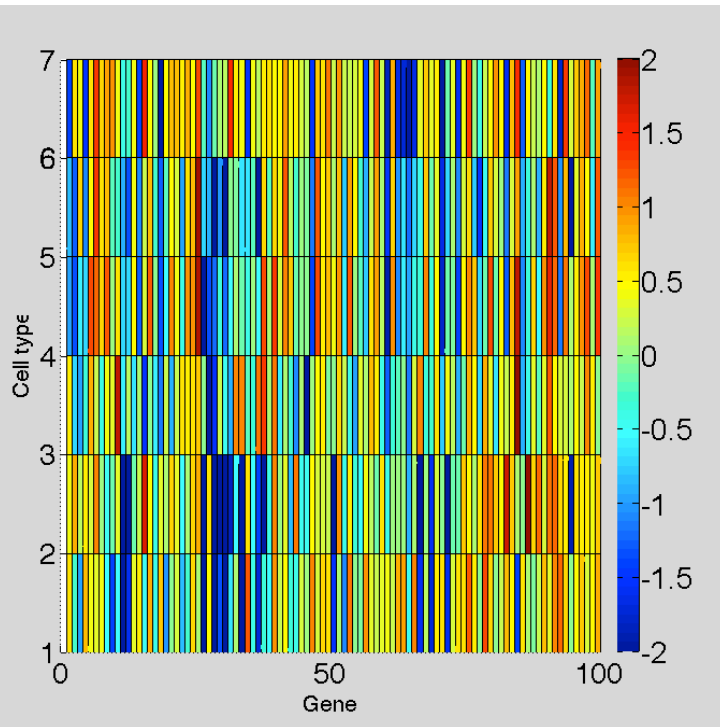
Which of the genes are expressed at the same time in cancer vs normal cells?

What are the genes that are expressed in cancer cells but not normal cells?

How the expression changes over time?



Is there any clustering of these genes?

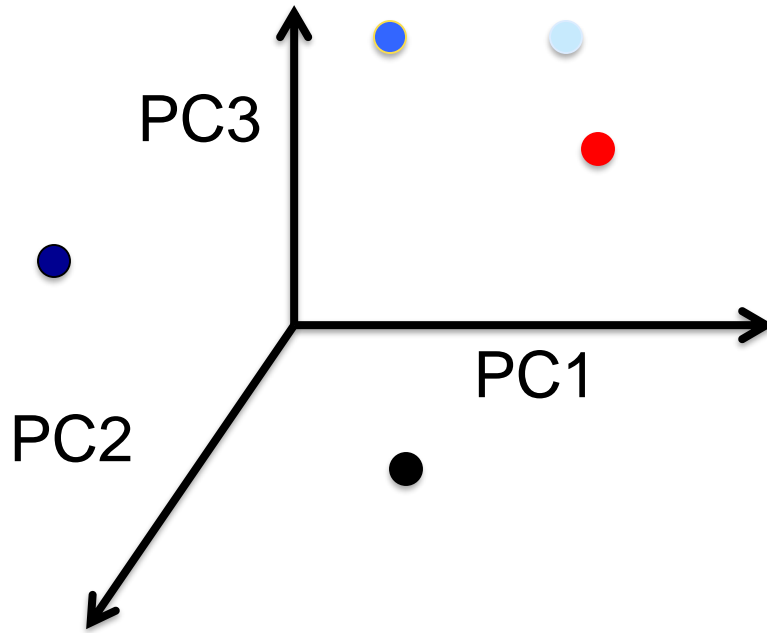


$$Y = P A$$

Transforming
information in
multivariable data set

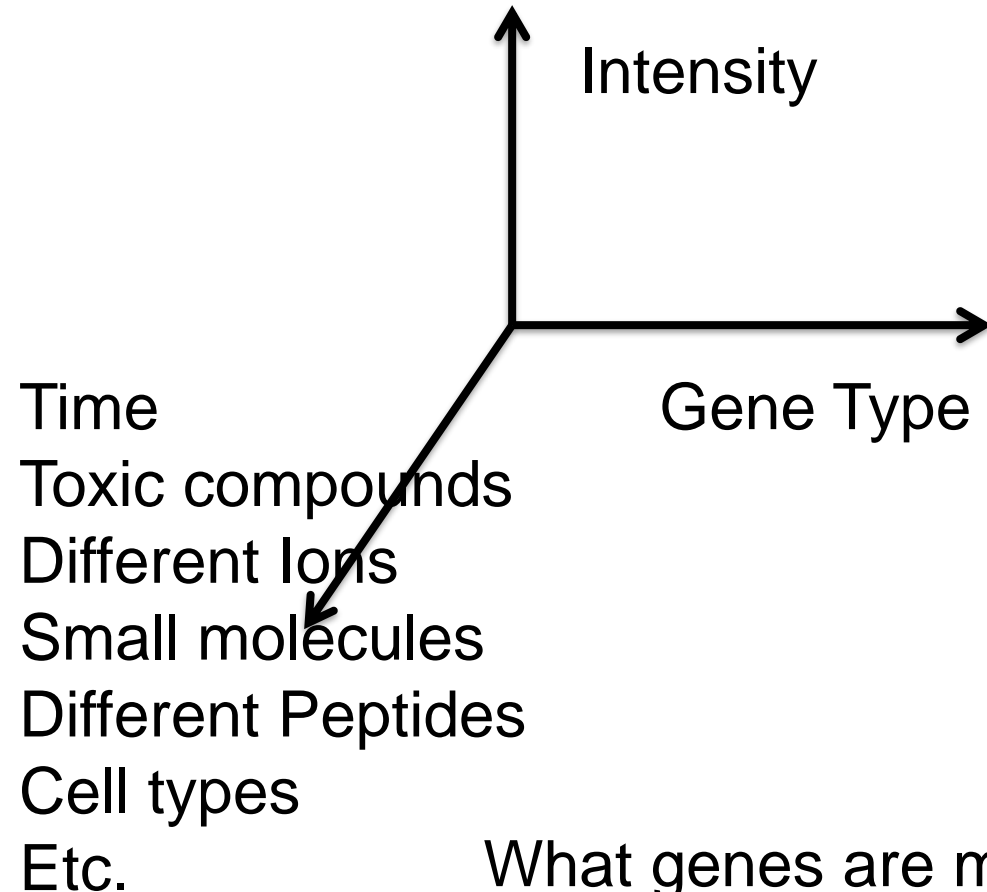
What information can be obtained in computational genetics?

1. Prediction of the class of these gene : The key genes can be identified. The expression of these genes can be used to *predict* the type (cancer or normal) of cell sample.



2. Building large network of genes: Complex data analysis is performed to learn the gene expression and construct a graph that show the dependency of expressed genes

Gene expression can also be probed for effect of toxic compounds, ions, different peptides etc.



What genes are most important? So the largest eigenvalues?

What is relative significance of these genes?

Can we identify the function of unknown genes?

Prove of Y (transformed matrix) solution by SVD

$$\mathbf{Y} = \mathbf{P}\mathbf{X}$$

Y = transformed matrix.

P = Principal Component

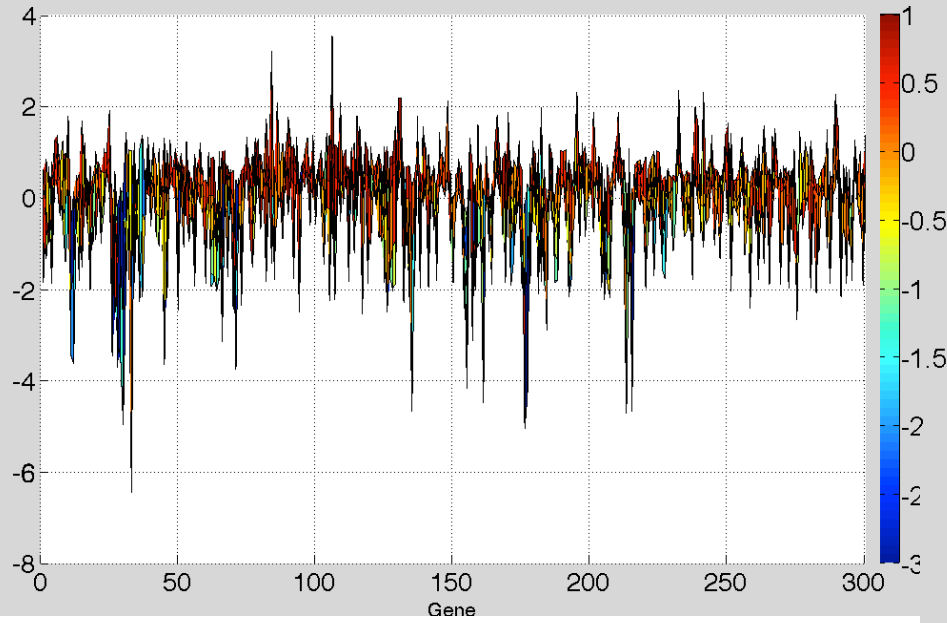
X = Data Set or Covariance of X

$$\mathbf{C}_Y = \frac{1}{n-1} \mathbf{Y}\mathbf{Y}^T = \frac{1}{n-1} (\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^T = \frac{1}{n-1} (\mathbf{P}\mathbf{X})(\mathbf{X}^T \mathbf{P}^T) = \frac{1}{n-1} \mathbf{P}(\mathbf{X}\mathbf{X}^T) \mathbf{P}^T$$

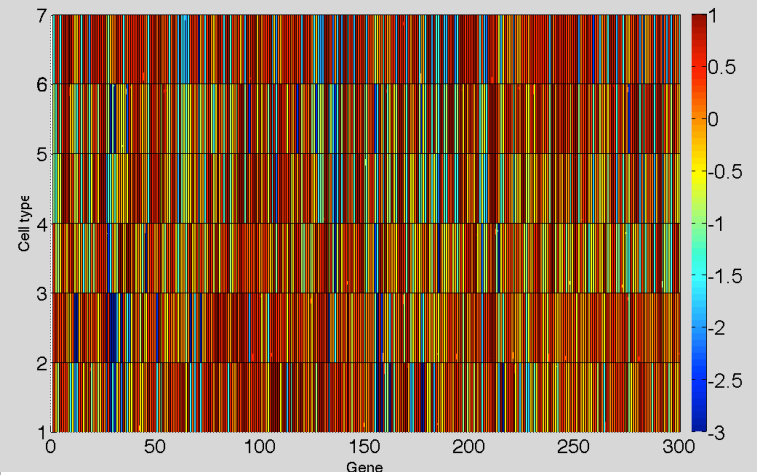
$$\text{i.e. } \mathbf{C}_Y = \frac{1}{n-1} \mathbf{P}\mathbf{S}\mathbf{P}^T \quad \text{where } \mathbf{S} = \mathbf{X}\mathbf{X}^T$$

Many iterations are needed to find S.

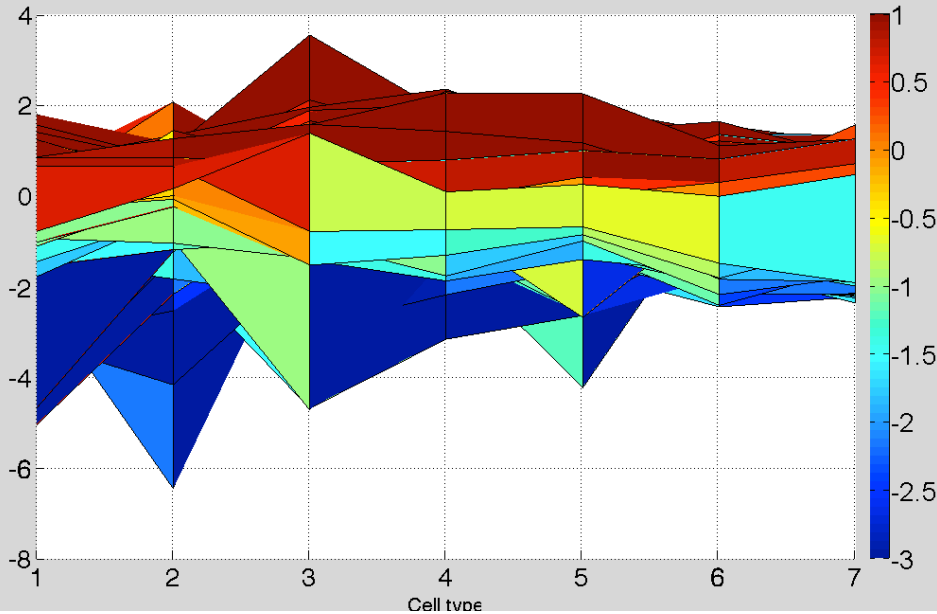
Multivariable Gene expression data



X-Z plot



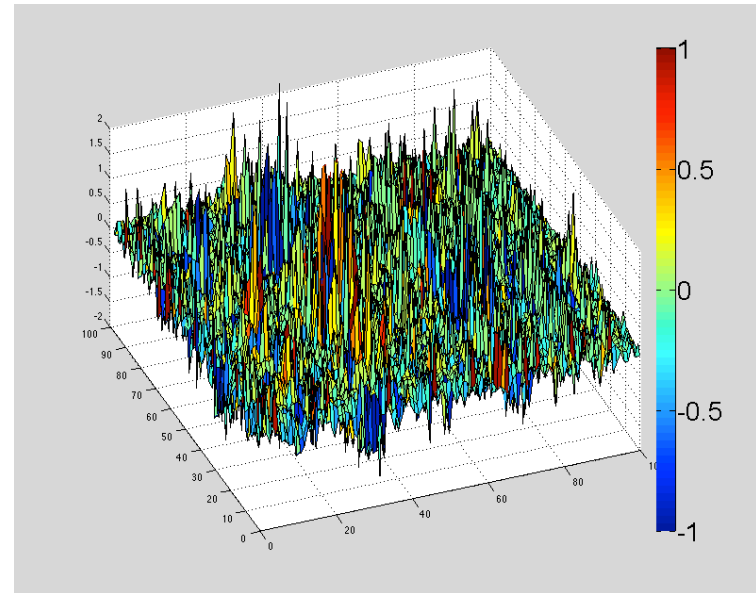
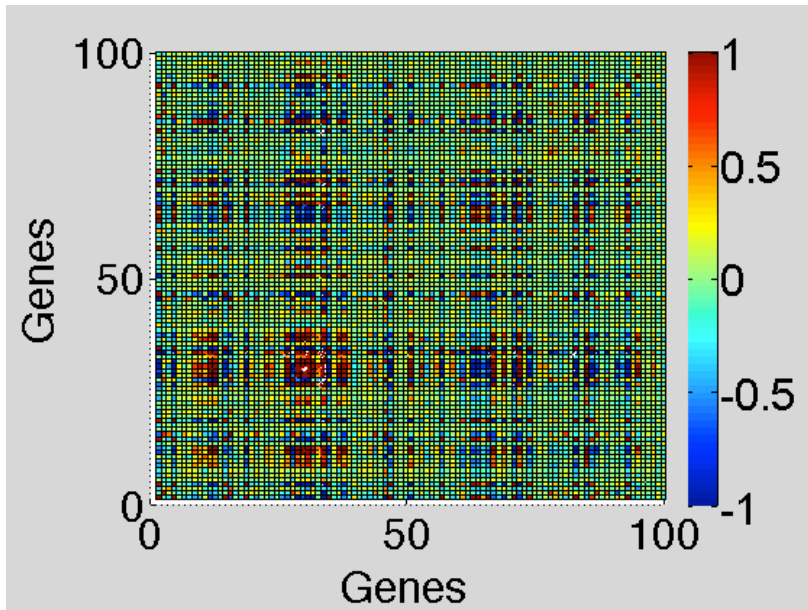
X-Y plot



Y-Z plot

Covariance matrix of multiple variables

Diagonal indicates the variance of a variable by itself



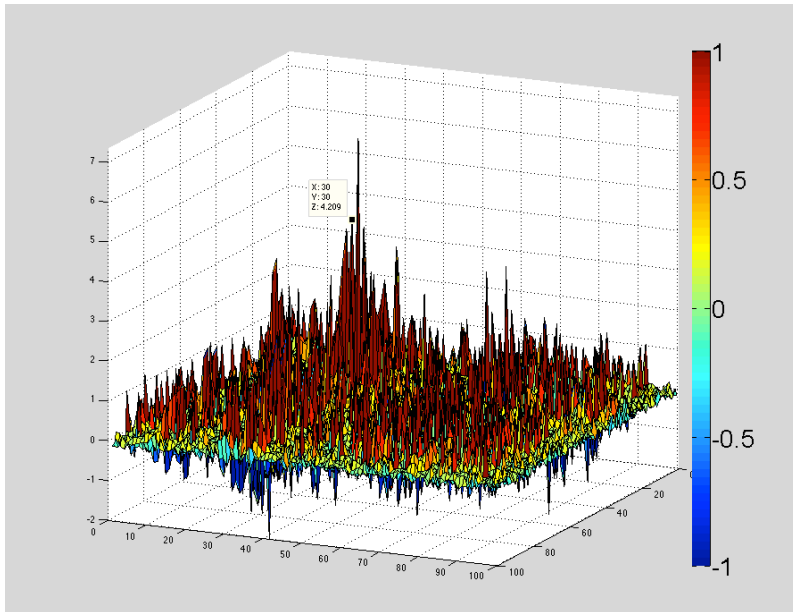
100 by 100 covariance matrix.

It is a symmetric matrix

$$\text{Cov}(\text{gene}) = \begin{matrix} & \begin{matrix} \text{gene 1} & \text{gene 2} & \text{gene 3} & \text{gene 4} \end{matrix} \\ \begin{matrix} \text{gene 1} \\ \text{gene 2} \\ \text{gene 3} \\ \text{gene 4} \end{matrix} & \begin{matrix} co(1,1) & co(1,2) & co(1,3) & co(1,4) \\ co(2,1) & co(2,2) & co(2,3) & co(2,4) \\ co(3,1) & co(3,2) & co(3,3) & co(3,4) \\ co(4,1) & co(4,2) & co(4,3) & co(4,4) \end{matrix} \end{matrix}$$

4by4

Covariance Matrix of 100 by 100 genes



We need a mathematical tool to find the vectors that demonstrate the largest variance in the covariance matrix.

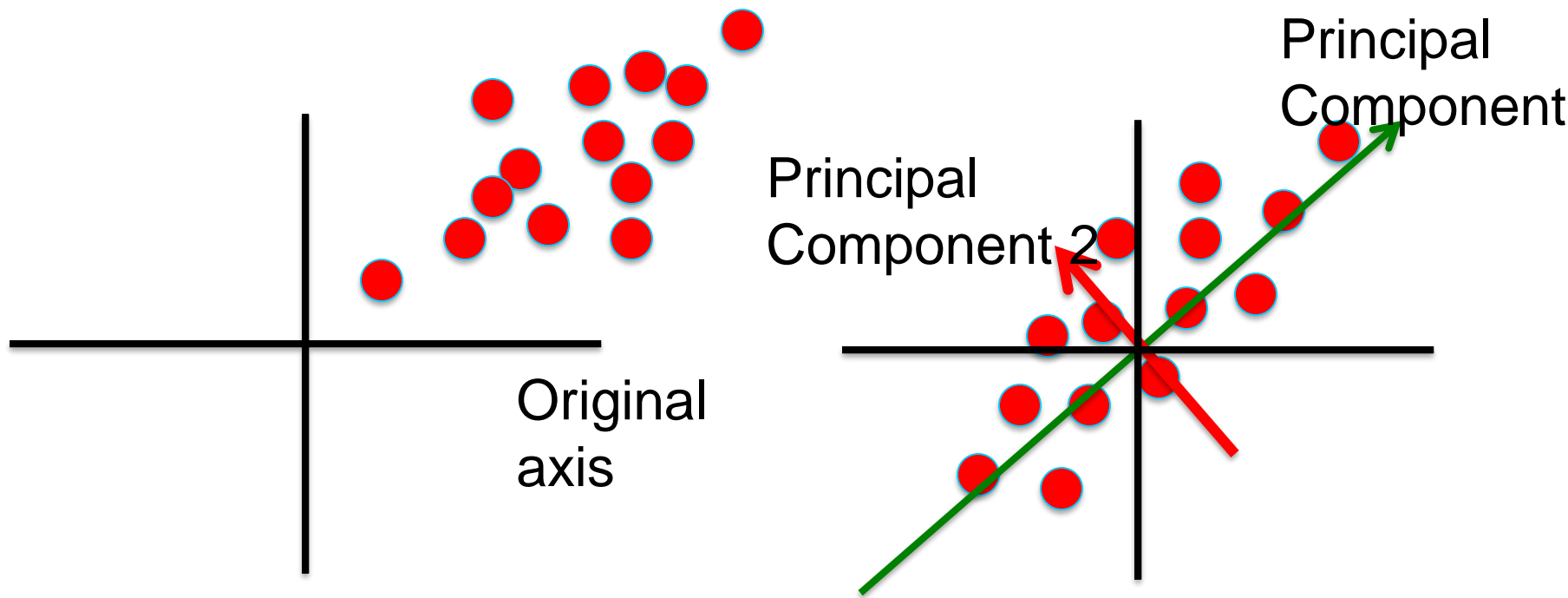
Remember many principal axes (number of variables) are present, but only a few of them describe the largest variance.

PCA by SVD

We can use SVD to perform PCA. We decompose A using SVD. PCA seeks a linear combination of variables such that the maximum variance is extracted from the variables.

$$A = USV$$

It approximates a high-dimensional data set with a lower-dimensional linear small set. It still contains most of the information in the large set.



Singular Value Decomposition (SVD analysis)

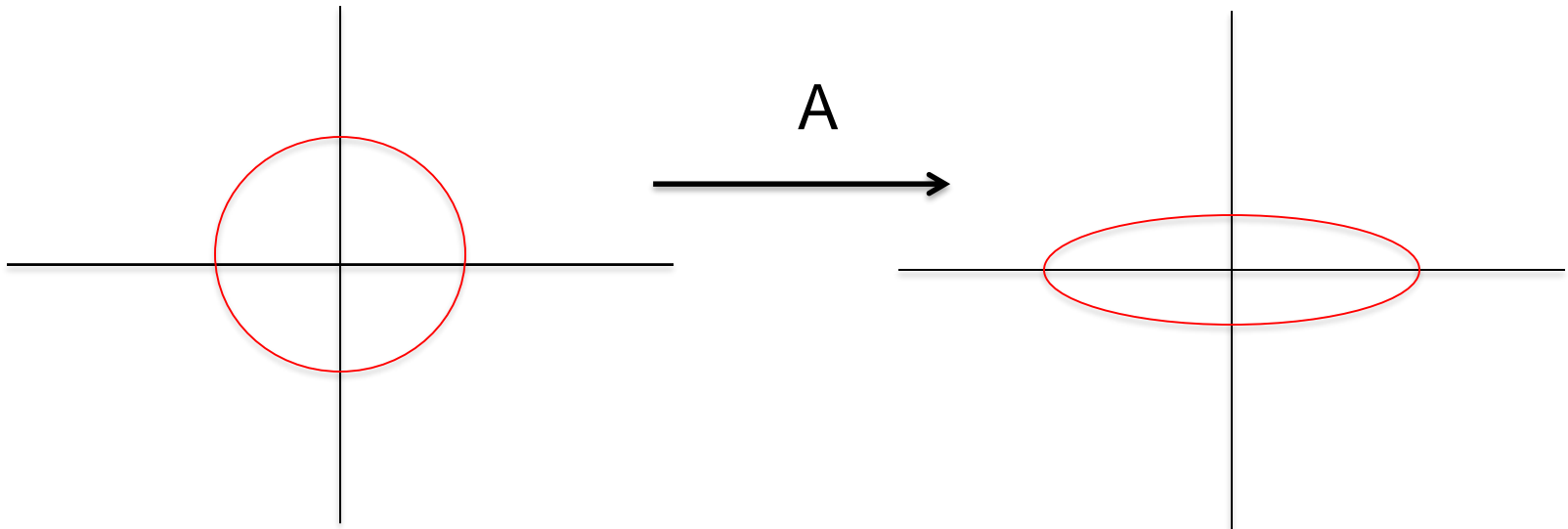
It is a mathematical matrix decomposition method or tool to analyze complex data and answer important questions.

It is used extensively for

It is very easy to use it and **rich information can be obtained from data.**

The main idea in PCA is to reduce the dimensionality of our data A by approximating A as a sum of rank matrices.

$$A_n = \sum_{i=1}^n \hat{a}_i u_i s_i v_i^T \quad \text{Rank matrix}$$



Singula value decomposition to calculate eigenvalue

$$Ab_i = \lambda_i x_i$$

$b_i, x_i = \text{Eigenvector}(\text{principal components})$

$\lambda_i = \text{Eigenvalue}$

The columns b_i and x_i of B and X are called the left and right eigenvectors respectively, and the diagonal elements λ_i of λ are called the singular values (eigenvalues).

Ab_i is in the direction of x_i

$$Ab_i = \lambda_i x_i$$

$$AV = US$$

Singular Value Decomposition (SVD) of a rectangular matrix A is a decomposition of the form

$$A = U S V^T$$

U and V are orthogonal matrices, and S is a diagonal matrix.

$$AV = US$$

$$AVV^T = USV^T$$

$$A = USV^T$$

U is $m \times n$ and orthonormal

S is $n \times n$ and diagonal

V is $n \times n$ and orthonormal

Singular value decomposition of A . SVD can be written always for A .

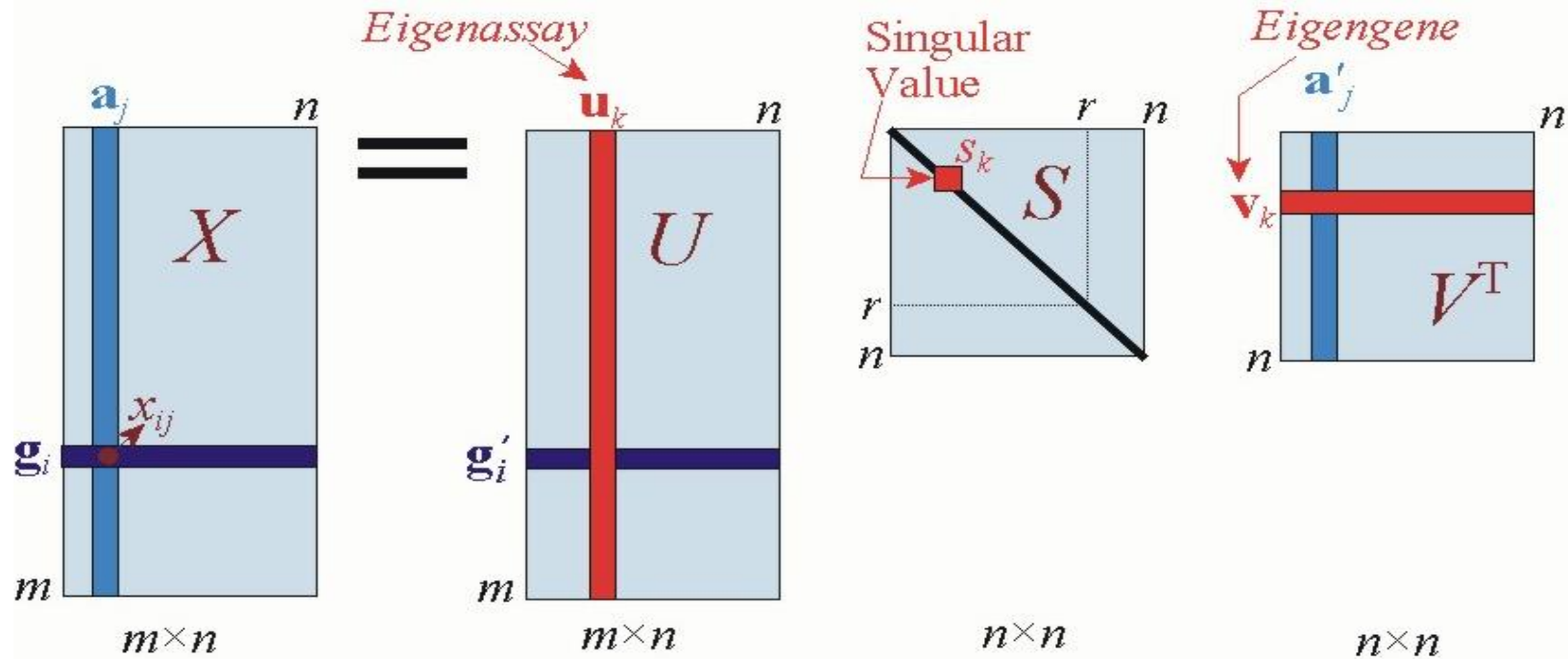
$$VV^T = I$$

$$S = \text{DIAG}(s_1, s_2, \dots, s_m)$$

$$s_i = \sqrt{\lambda_i}$$

Eigenvalues of AA^T or $A^T A$

PCA uses the SVD in its calculation



$$\mathbf{a}_j = \sum_{k=1}^r v_{jk} s_k \mathbf{u}_k, \quad j: 1, \dots, n$$

In PCA, we basically find eigenvalues and eigenvectors of covariance matrix.

$$C = AA^T/N$$

$$\sigma_a \sigma_b = 0$$

highly uncorr.

$$\sigma_a \sigma_b = \sigma_a^2$$

correlated

$$\begin{matrix} \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{matrix}
 \mathbf{A}
 \begin{matrix} \overset{\circ}{0} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \emptyset \end{matrix}
 =
 \begin{matrix} \mathbf{U} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{matrix}
 \begin{matrix} \overset{\circ}{0} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \emptyset \end{matrix}
 \begin{matrix} s_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & s_n \end{matrix}
 \begin{matrix} \overset{\circ}{0} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \emptyset \end{matrix}
 \mathbf{V}
 \begin{matrix} \overset{\circ}{0}^T \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \emptyset \end{matrix}$$

IMPORTANT:

s_i on the diagonal are called the singular values (eigenvalues) of A .

The columns of U represents the principal components (eigenvectors) of matrix A .

EIGENVALUES

Why eigenvalues are important?

Considered as characteristic tool of the matrix.

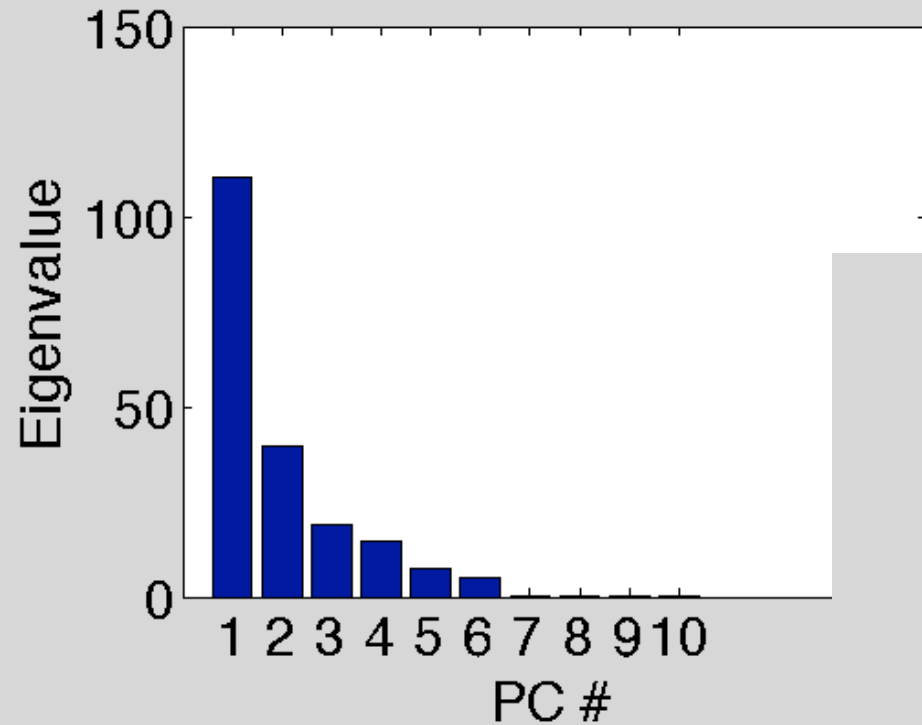
For example you tell if a large sets of genes are expressed at certain time but not the other

Briefly, the eigenvalue for a given factor measures the variance in all the variables which is accounted by that factor. Largest eigenvalues gives the principal axis where the variance is largest along the corresponding principal axis.

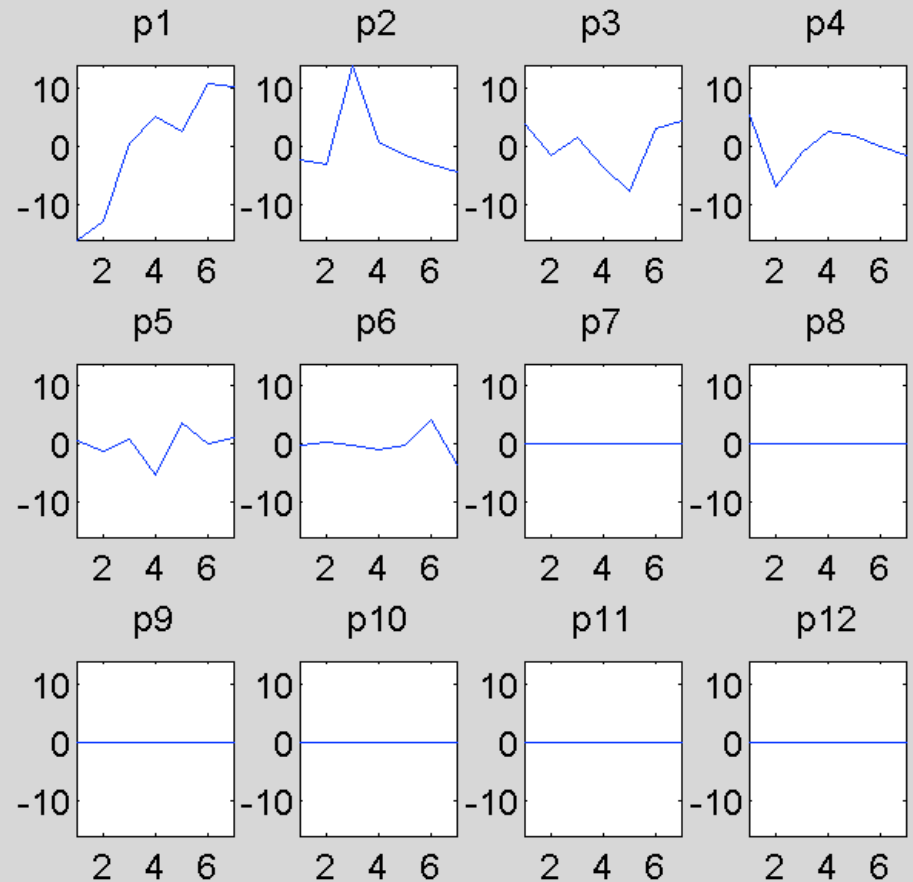
The ratio of eigenvalues :

It is extremely important. If a factor has a low eigenvalue, the variance in the variables can be explained less significantly by the eigenvalues.

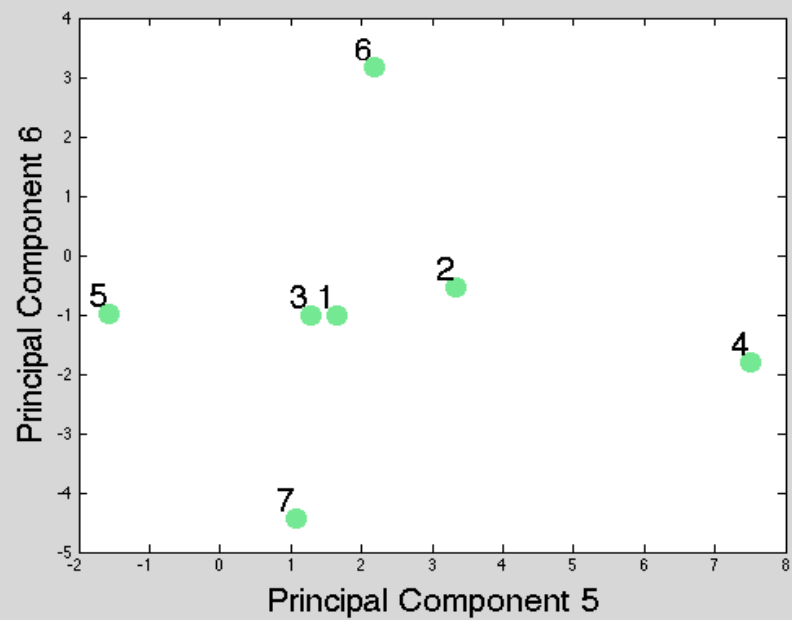
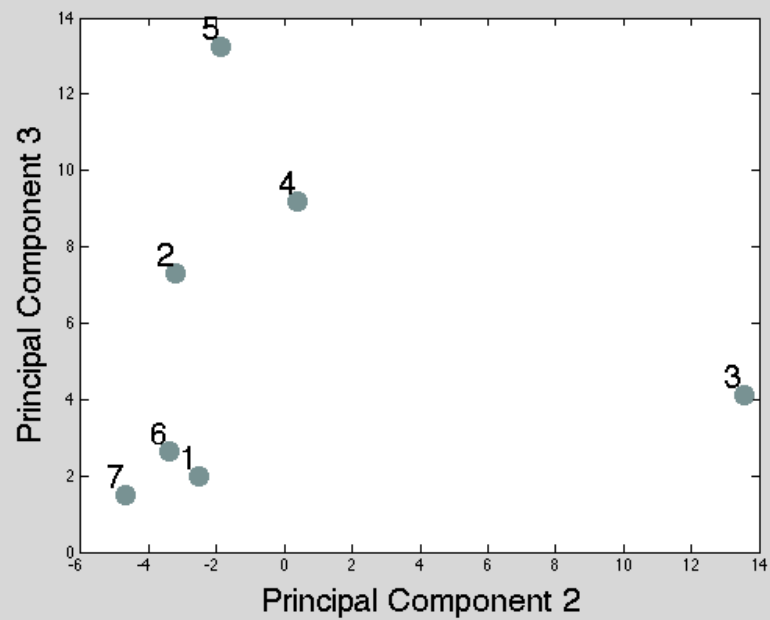
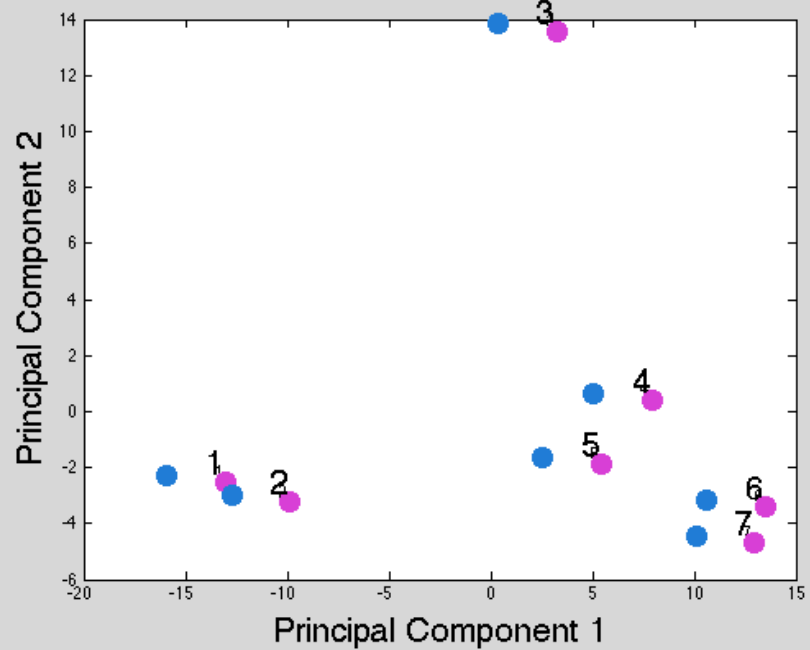
Eigenvalues of Covariance Matrix



First few P_c represent the most important features of data.

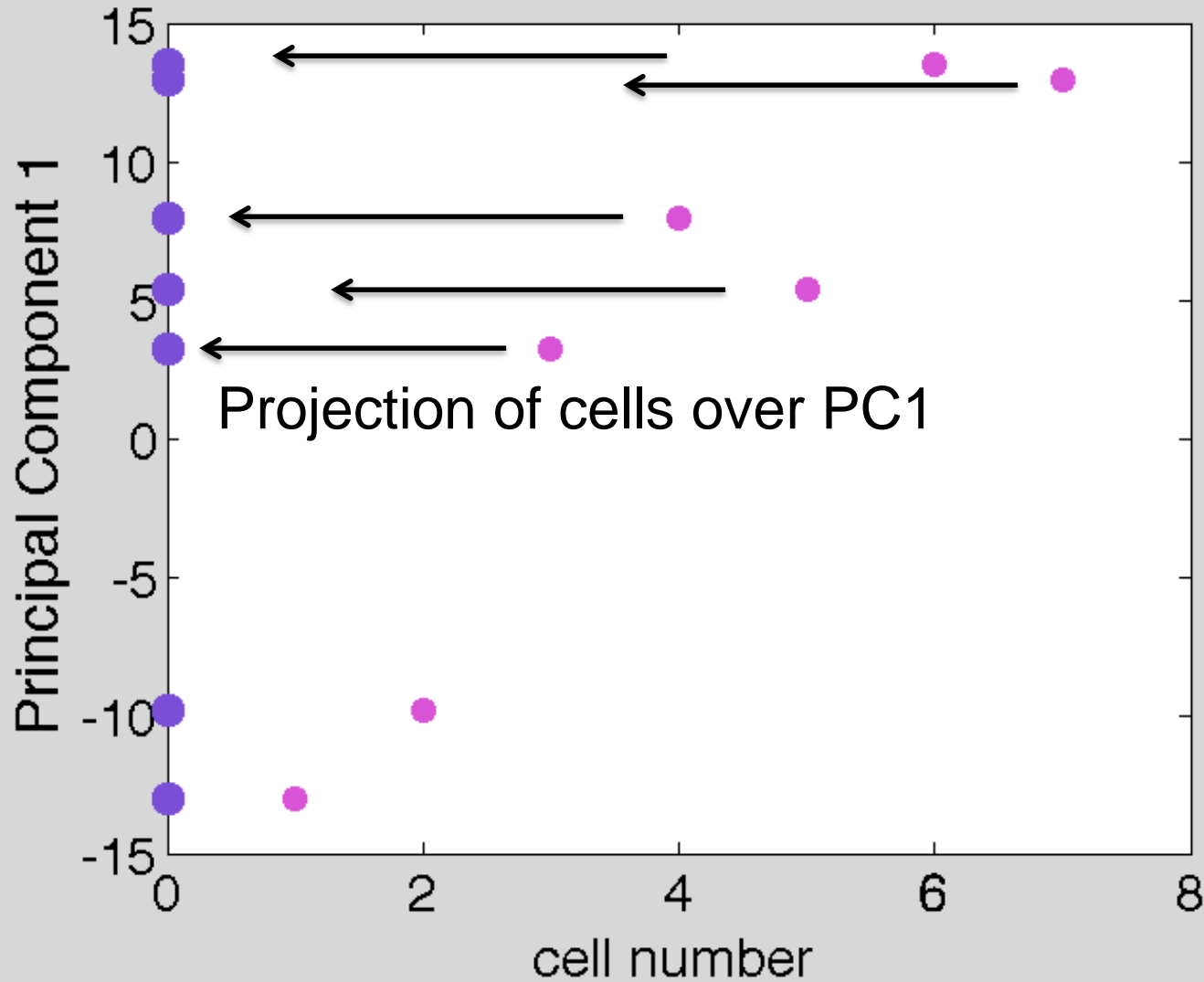


PC1 and PC2 represent the largest variance for cells

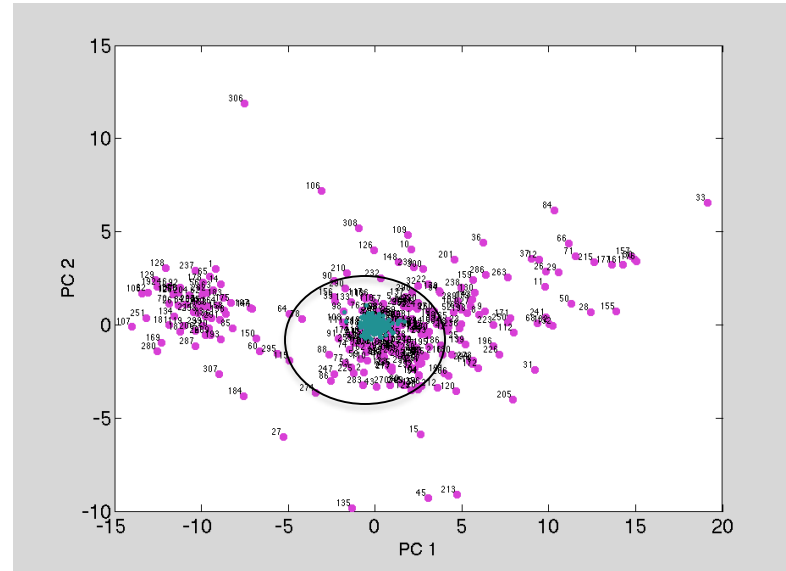
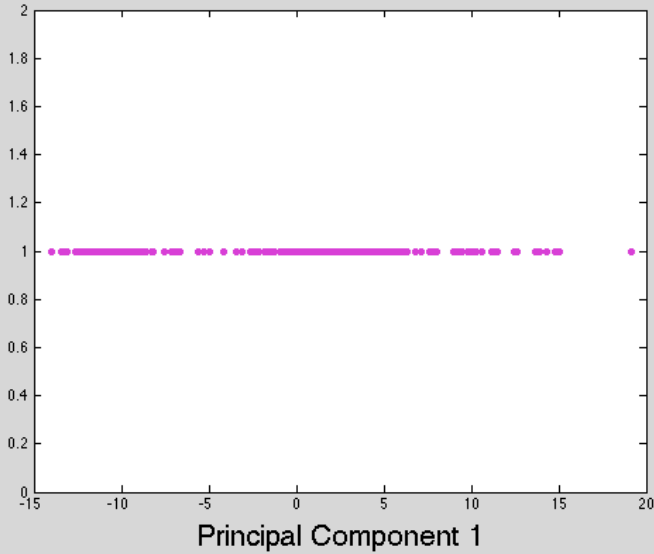


Classify Cells using only PC1

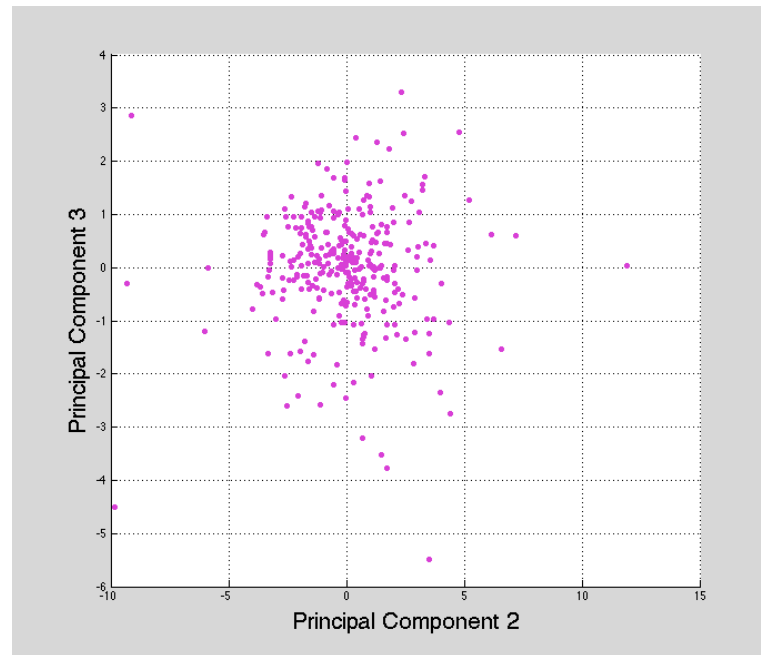
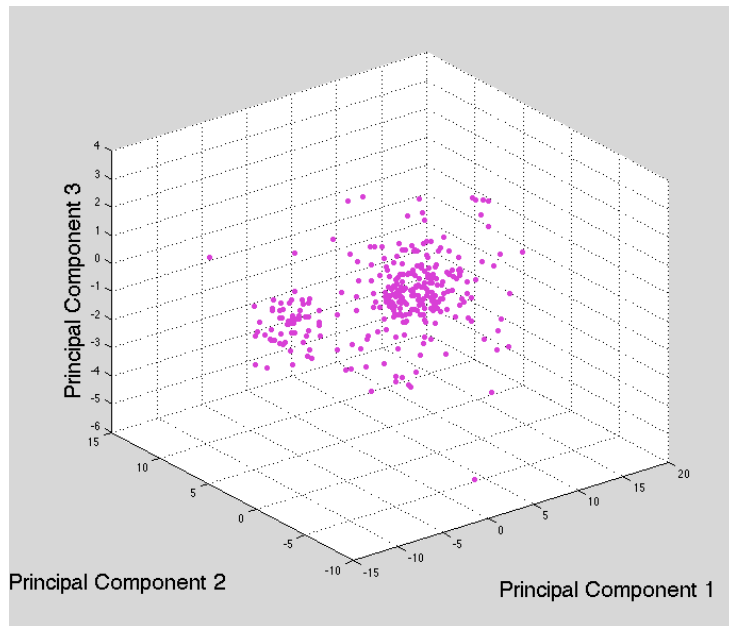
Only 1st PC

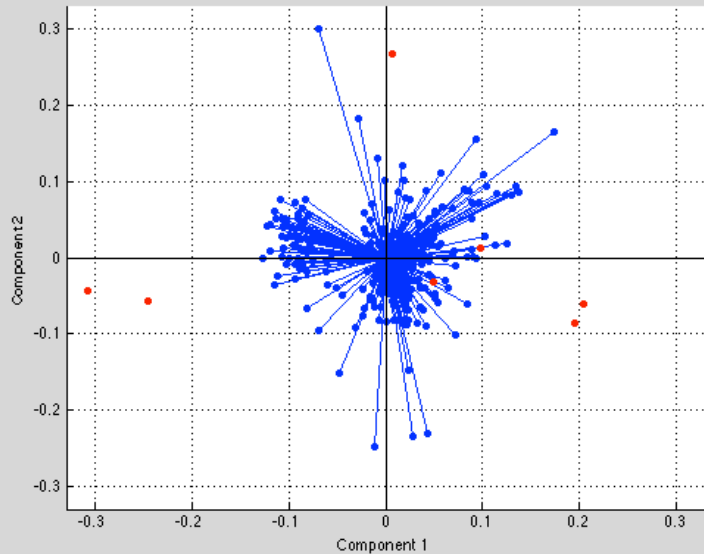
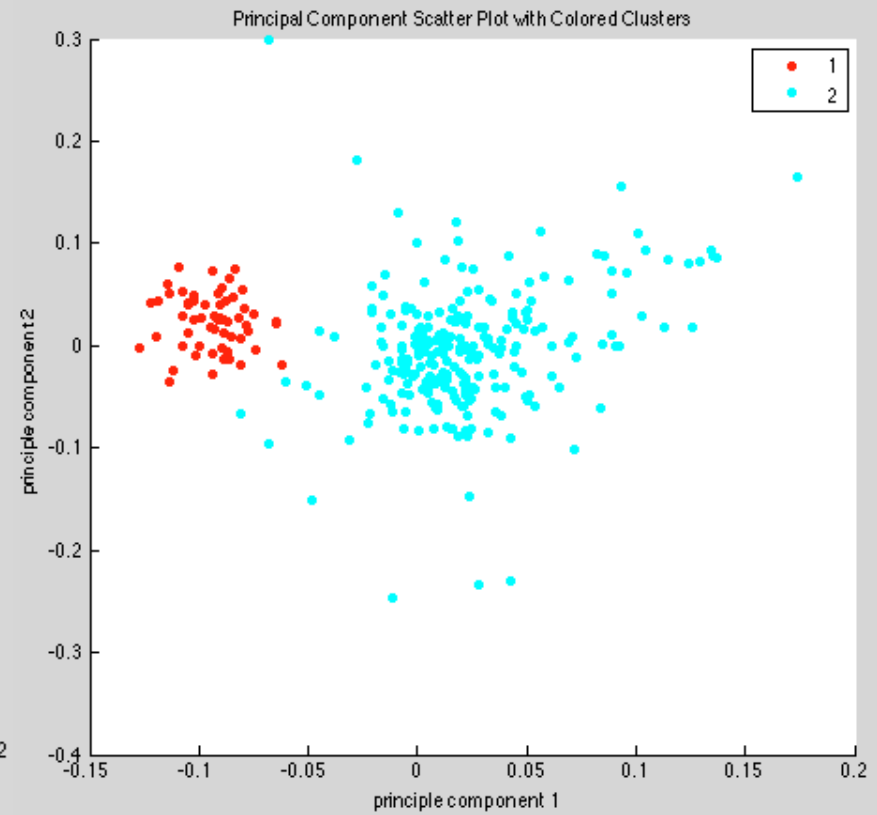
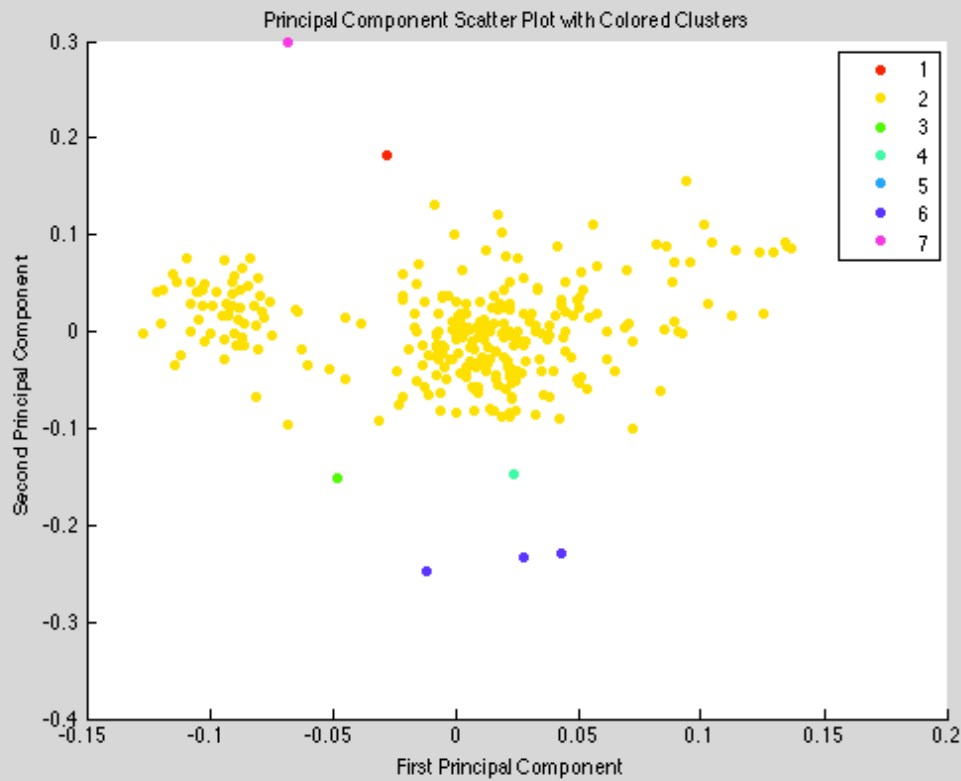


How many principal components do we need?



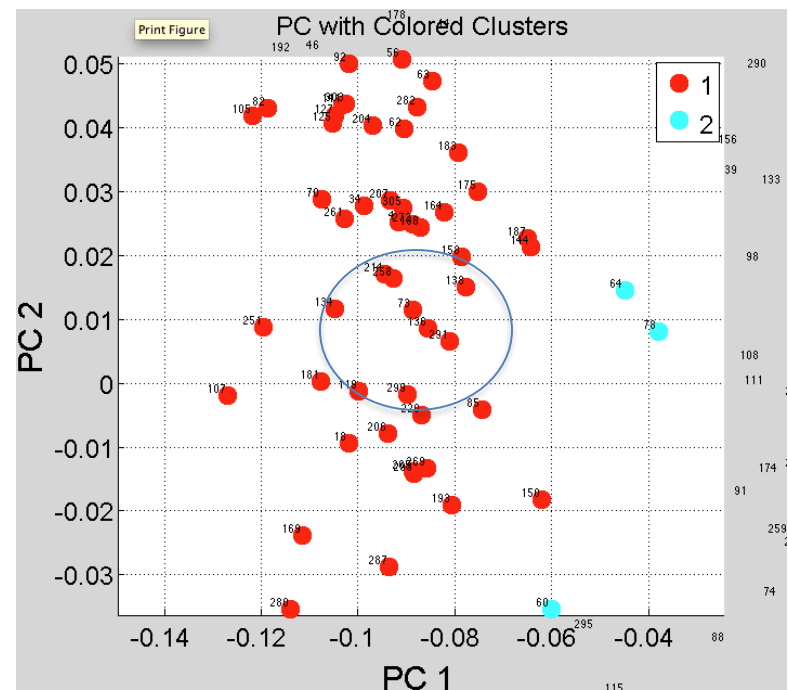
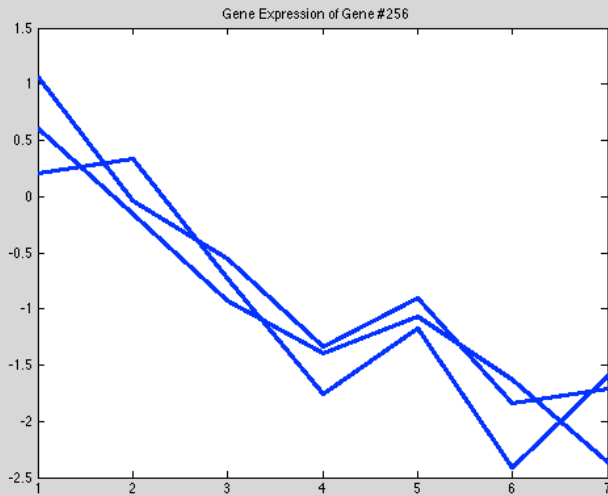
PC5
And
PC6



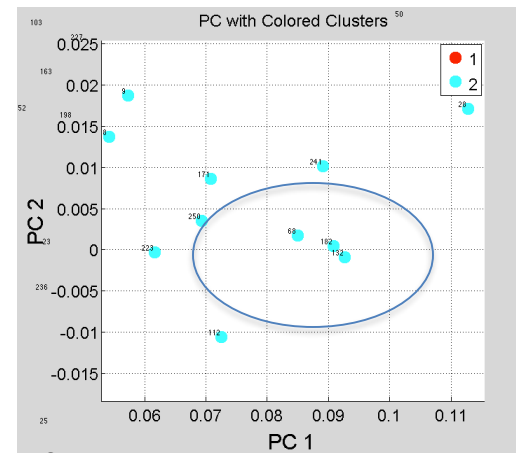
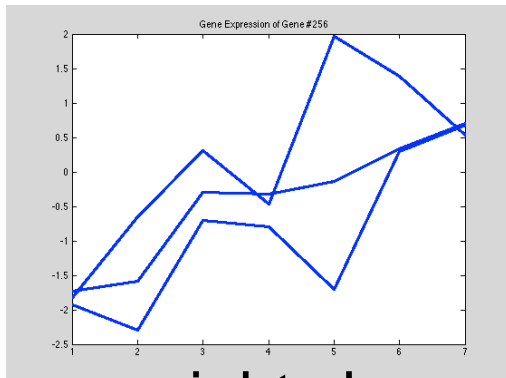


Here I checked the inner squared distance (minimum variance)

Genes 73,214, 258



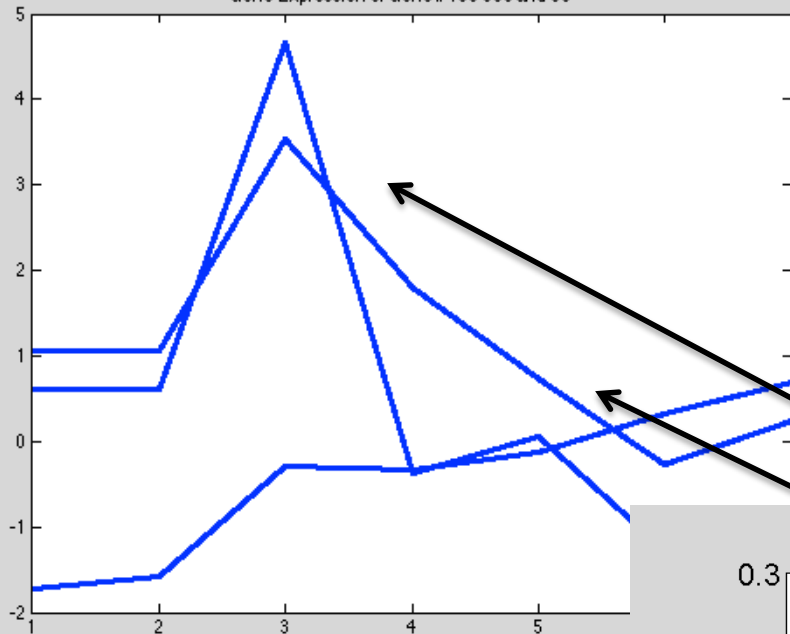
Genes 68, 182, 132



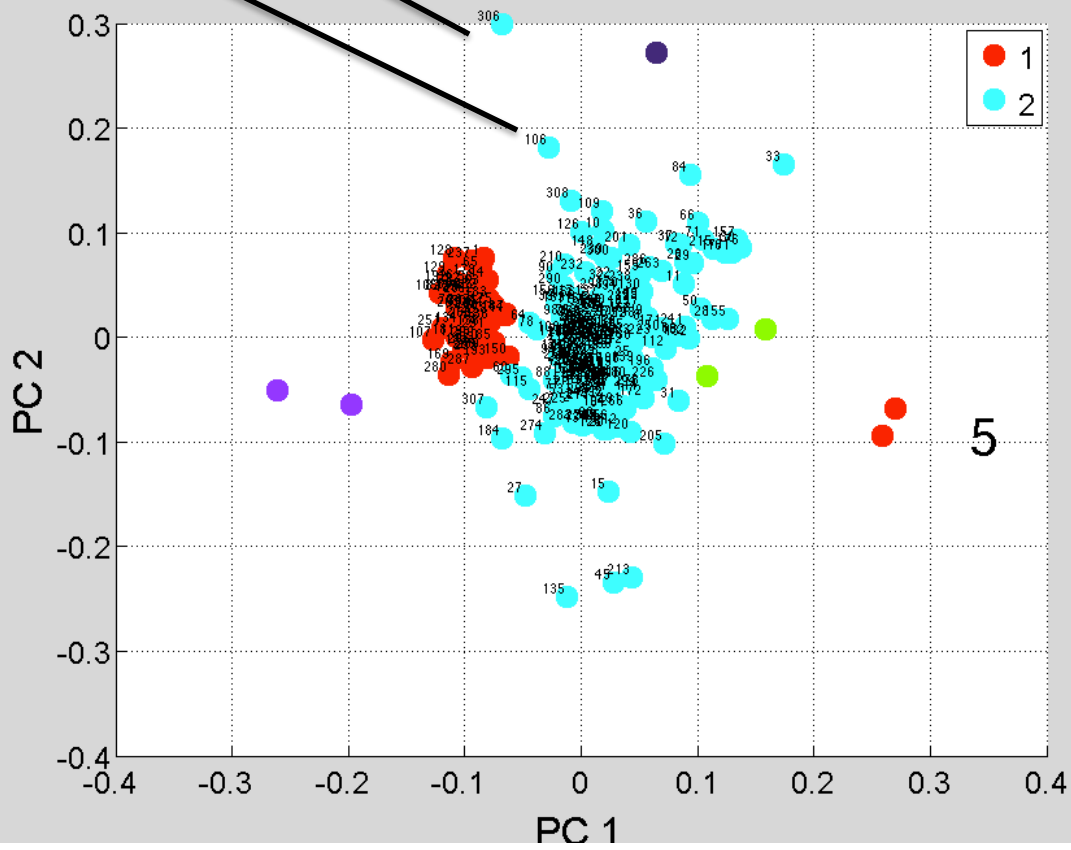
Genes might change the function of cells.

Or we may say that genes with same pattern may have a link in cells

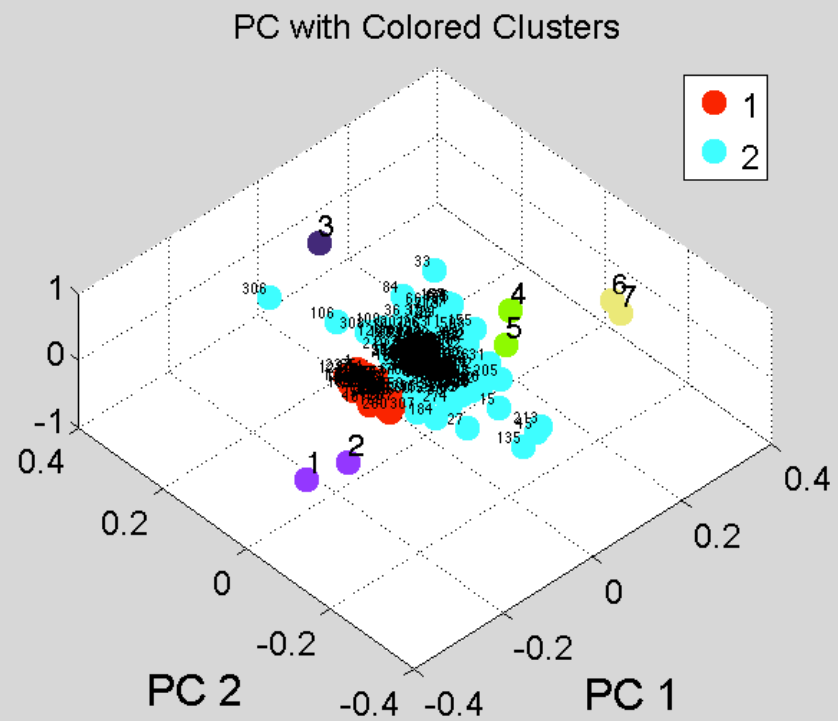
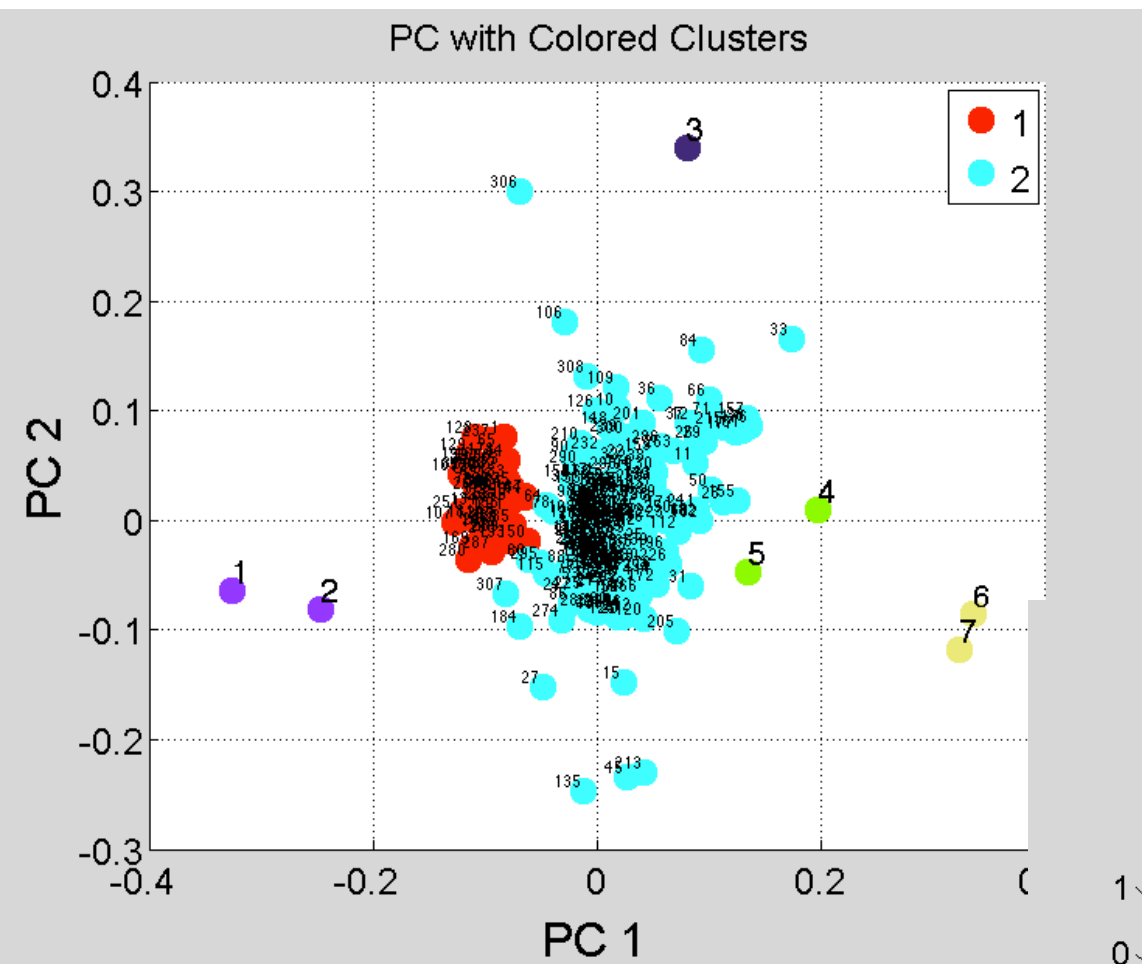
Gene Expression of Gene #106 306 and 68



PC with Colored Clusters

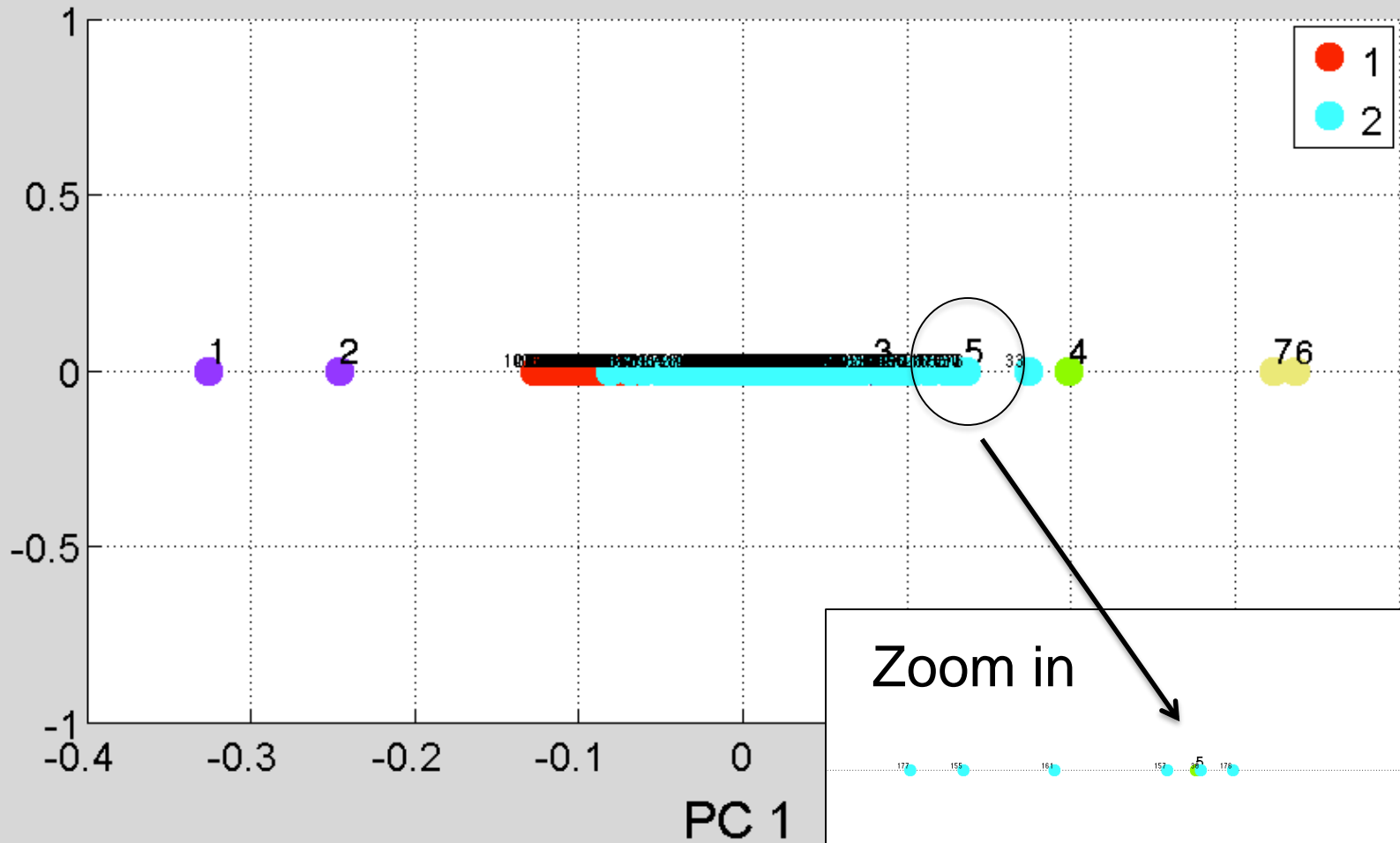


Clustering genes with cells



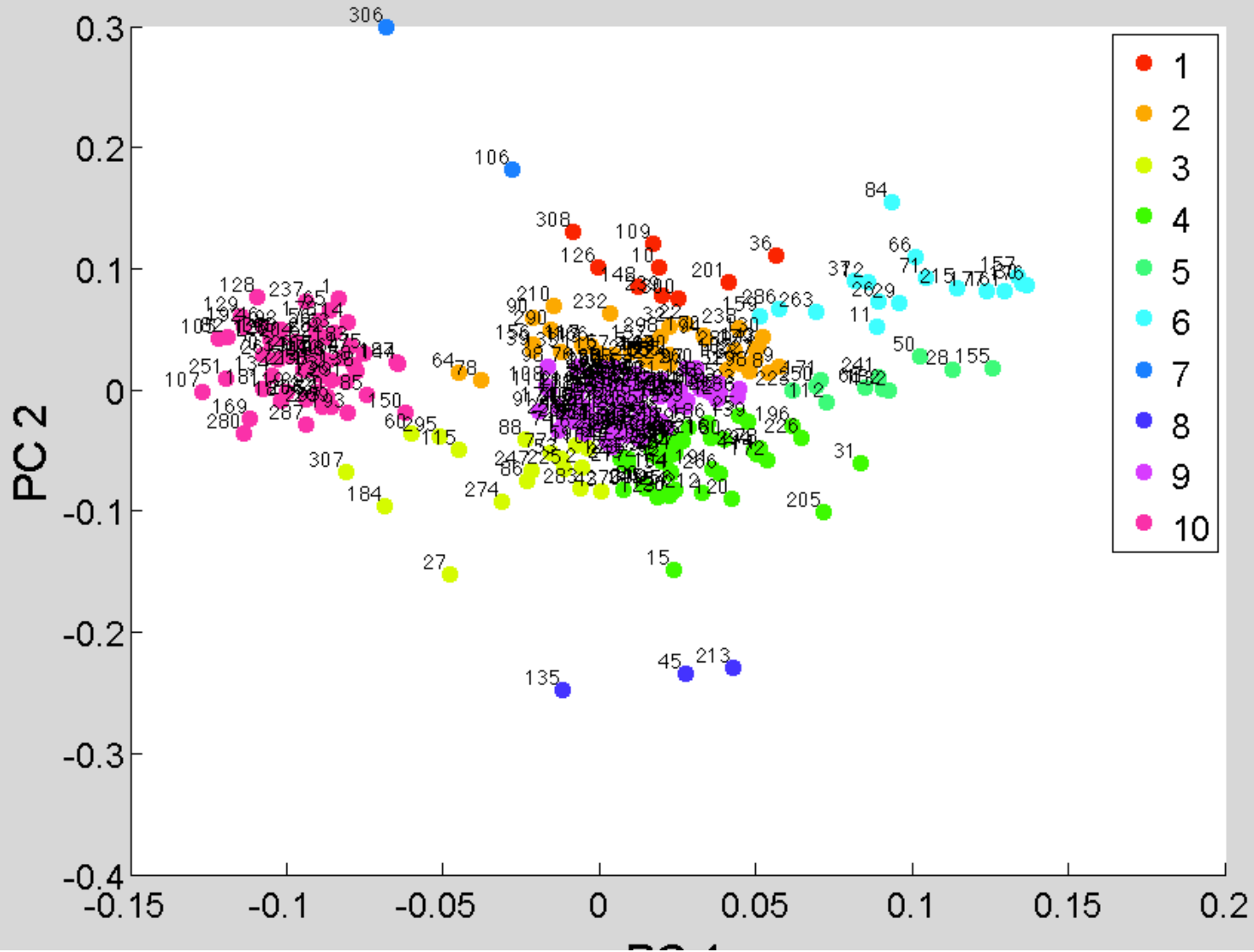
Change the perspective (plane) helps visually finding relations between cells and genes

PC with Colored Clusters



X-Z plot

PC Scatter Plot with Colored Clusters



Cell relation using dendrogram

1. Cells 6 and 7 are related
2. Cells 1 and 2 are related
3. Cell 6-7 and cells 1-2 are inversely proportional or they have a different levels of gene expression.

Are they different cells?

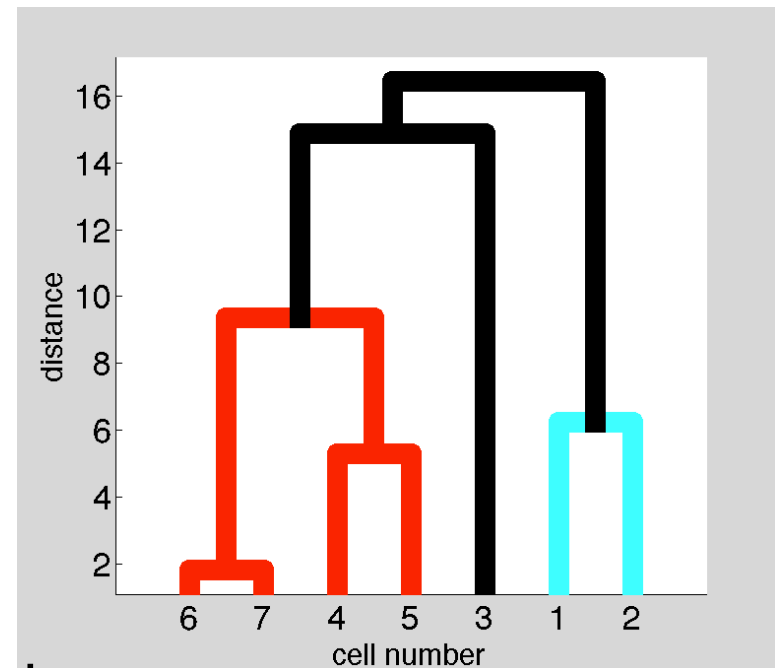
Can we say that

6 and 7 are abnormal cells?

Or

They may have a different function?

Their function can be addressed by looking what genes are expressed



1. Eigenvectors and eigenvalues always come in pairs.
2. Eigenvalues is the scaling factor of the vector.
3. Every matrix has SVD.
4. The eigenvalues can be determined and those values can be $S_1 \geq S_2 \geq S_3 \geq \dots S_n > 0$