CSC345/M45:
Big Data & Machine Learning
(dimensionality reduction: PCA)

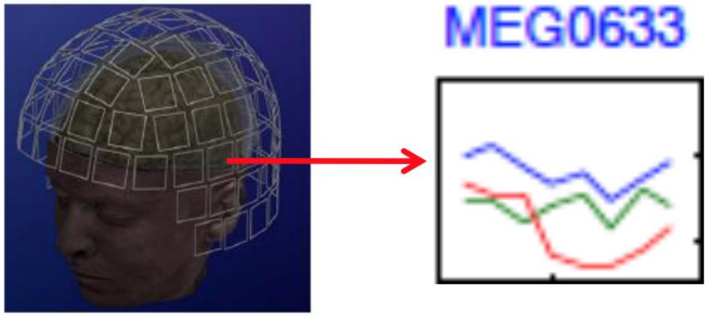
Sara.sharifzadeh@swansea.ac.uk

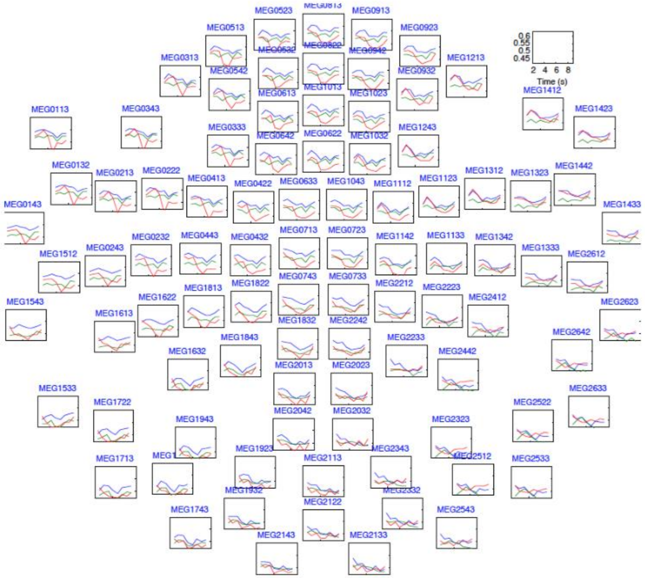
318 Computational Foundry, Bay Campus

Sliders adapted from Prof. Xianghua Xie slides.

Dimensionality Reduction

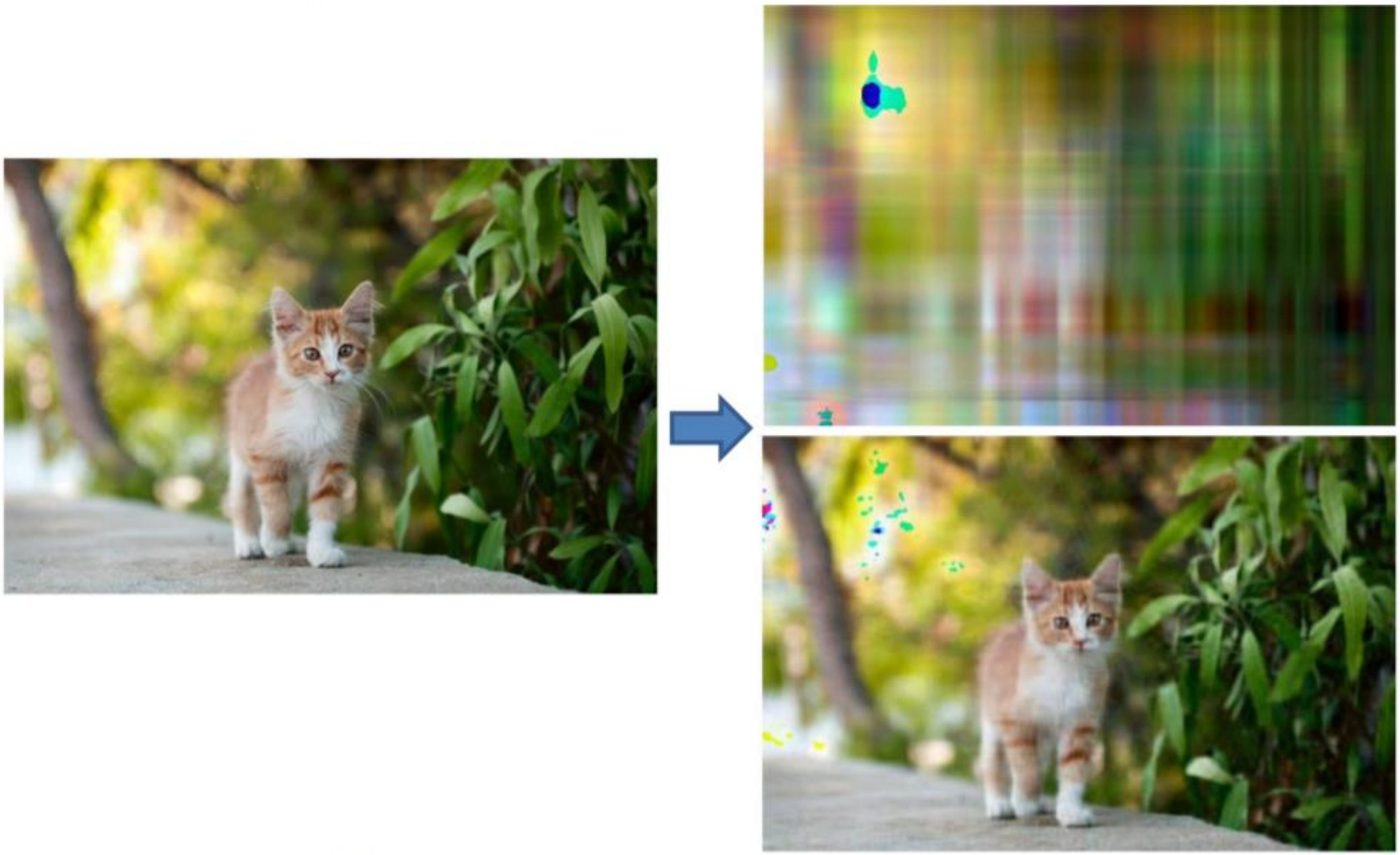
- Input data may have thousands or millions of dimensions
  - Amazon song example in our introduction lecture
  - Text/documents data
  - Gene expression data
  - MEG brain data
    - E.g. 120 locations x 500 time points





Dimensionality Reduction

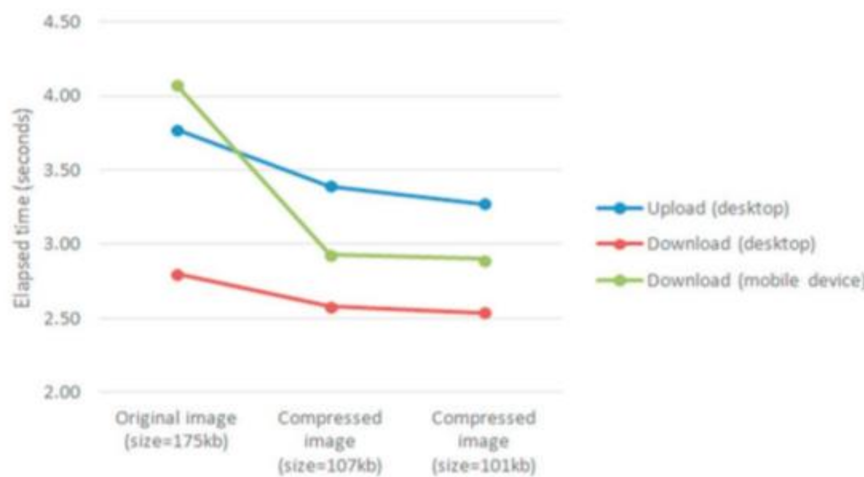
- Data compression: matrix factorization



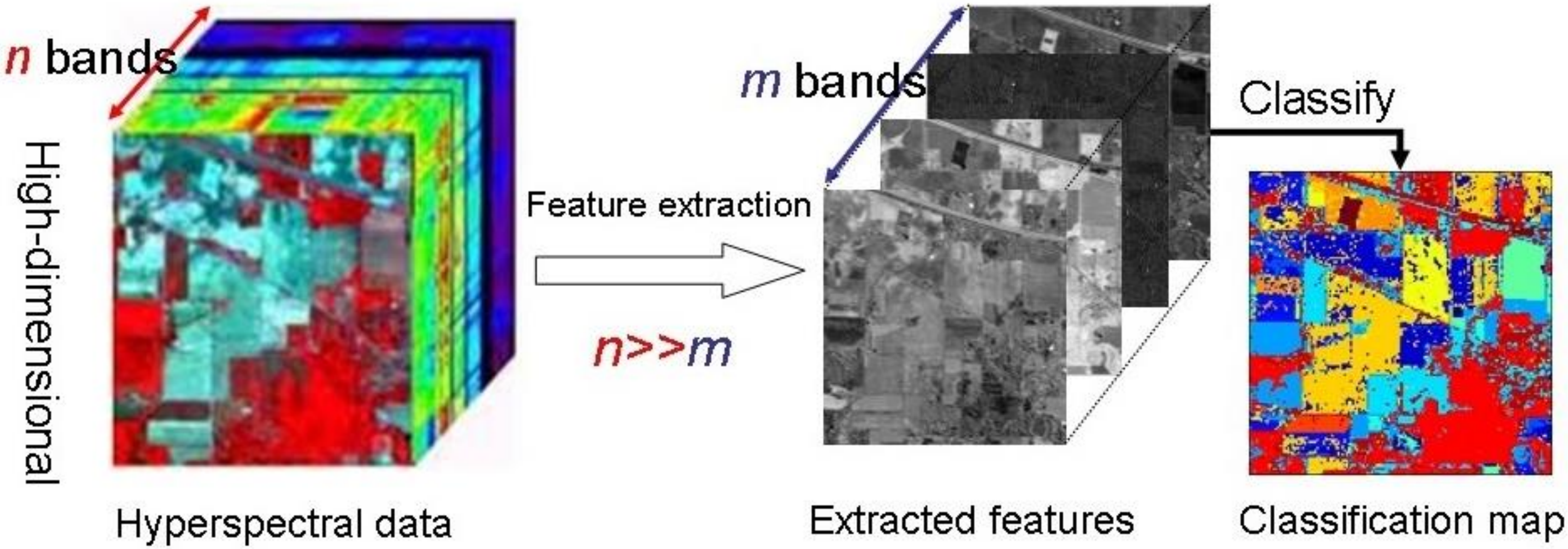
[http://www.aaronschlegel.com/image-compression-principalcomponent-analysis]

Dimensionality Reduction

- Data compression: keep main components



Dimensionality Reduction for Spectral imaging



https://telin.ugent.be/~wliao/Research.html

Dimensionality Reduction

Features (dimension)

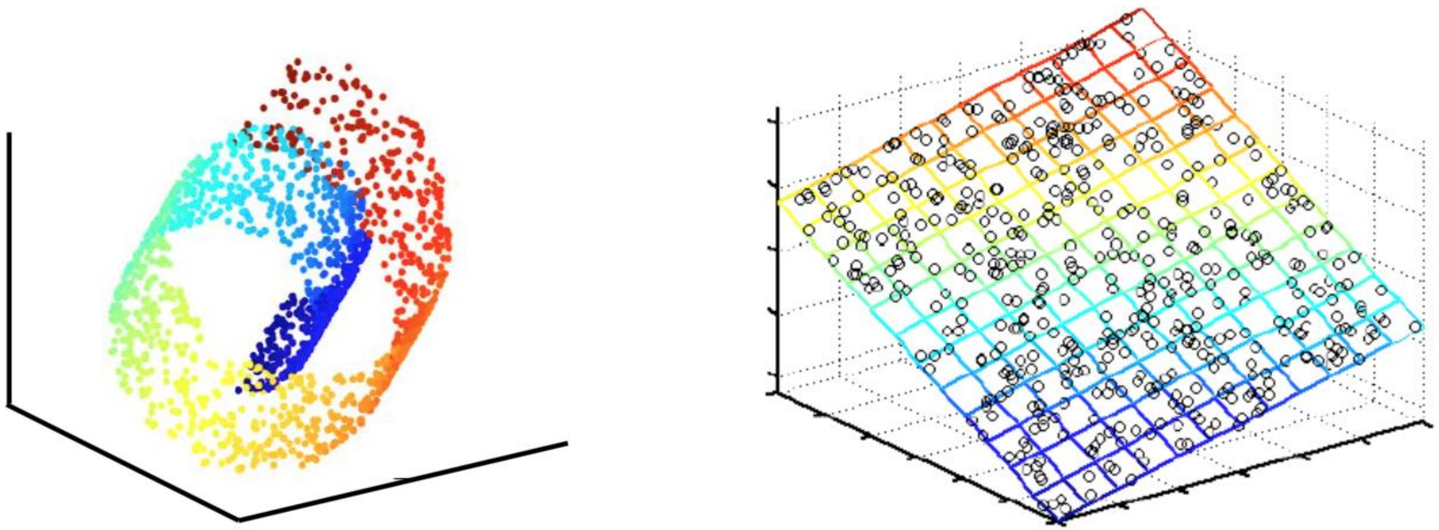
samples

x11	**X12**	**…**
x21	**x22**	**…**
…	**…**	**…**

- **Curse of dimensionality**
  - redundant features
    - e.g. not all words are useful in classifying documents: and, or, the, of, …
  - Data samples required **grows exponentially** with the increase of dimensionality
  - the efficiency of many algorithms depends on the number of dimensions
  - distance based similarity computations are at least linear to the number of dimensions
    - E.g. k-means, GMM
  - expensive to store for high dimensional data
  - indexing and retrieving data in high dimensional space

Dimensionality Reduction

- Why dimensionality reduction?
  - Reduce the dimensionality of the data while maintaining the meaningfulness of the data
  - Find a low-dimensional but useful representation of the data
  - Discover “intrinsic dimensionality” of the data
    - some high dimensional data is actually low dimensional in nature



An example of 3-D data is in fact 2-D

Principal Component Analysis

- Example

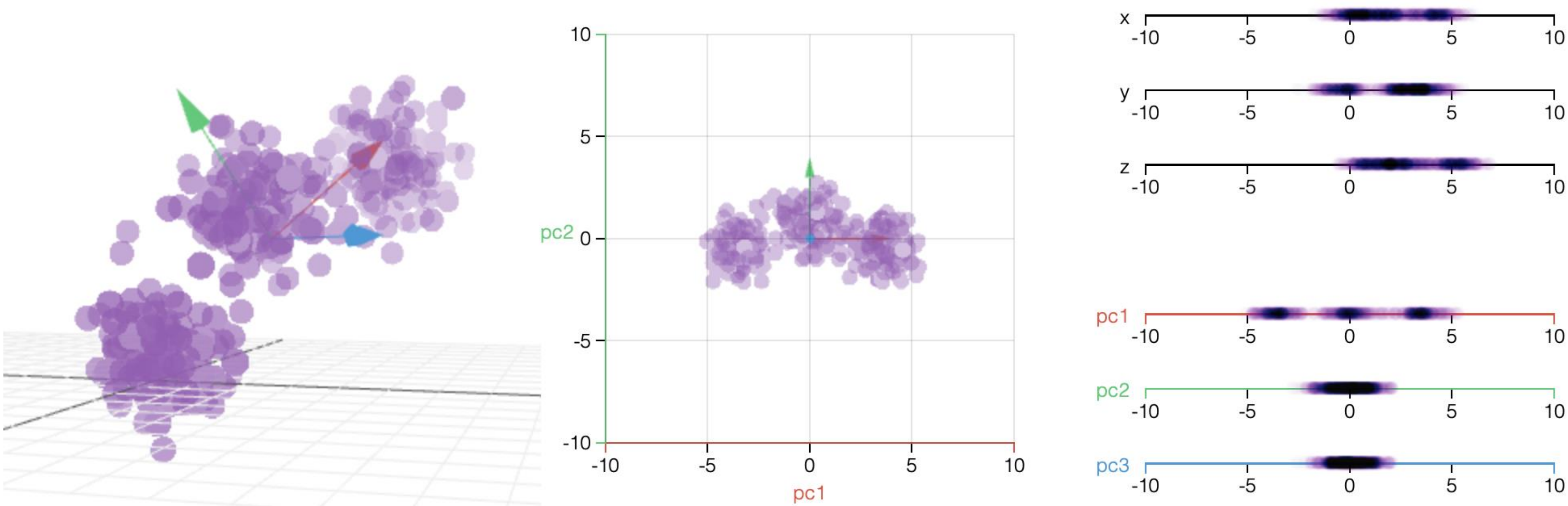
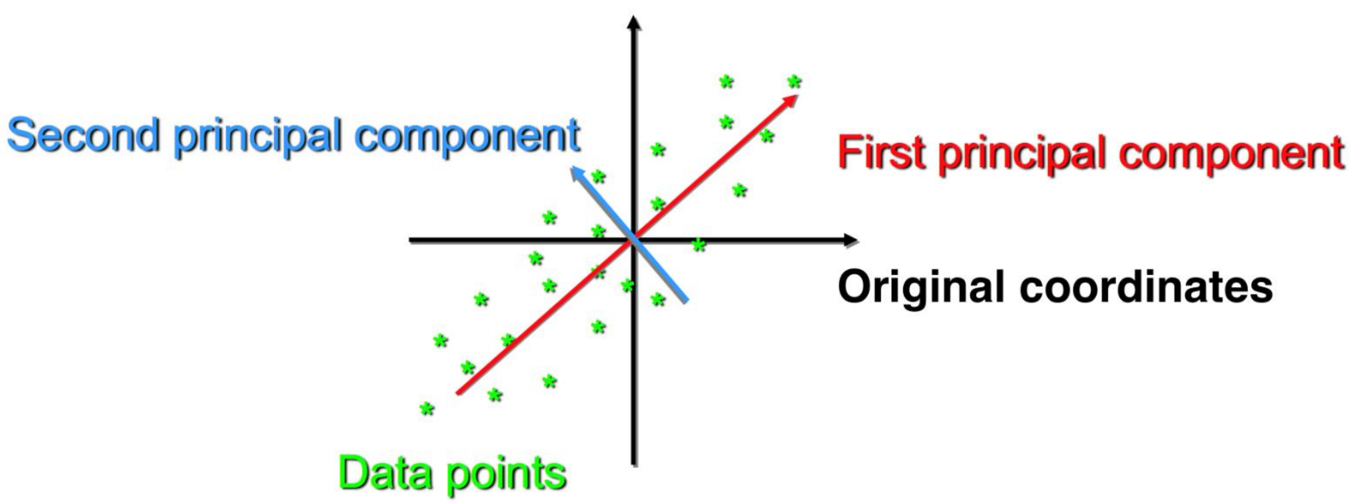


Image source: https://setosa.io/ev/principal-component-analysis/

Principal Component Analysis

- Principal component analysis (PCA)
  - a **linear method** used to reduce data dimensionality
  - reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the **variation** present in the data set.
  - achieved by transforming to a new set of variables, the principal components (PCs), which are **uncorrelated**, and which are ordered so that the **first few** retain most of the variation present in all of the original variables.

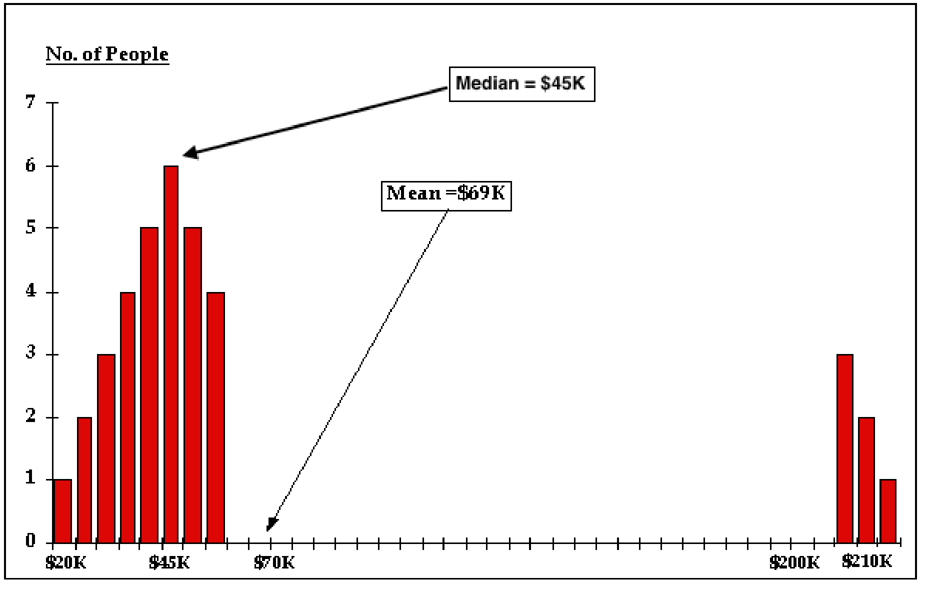


Mean and Median

- Mean: the average of all data values

$$\bar{x} = \frac{\sum x_i}{n}$$

  – n is the number of observations
- Median: is the value in the middle when the data items are sorted in either ascending or descending order
  – When the data has extreme values (outliers), median is often the preferred measure for location



Variance and Standard Deviation

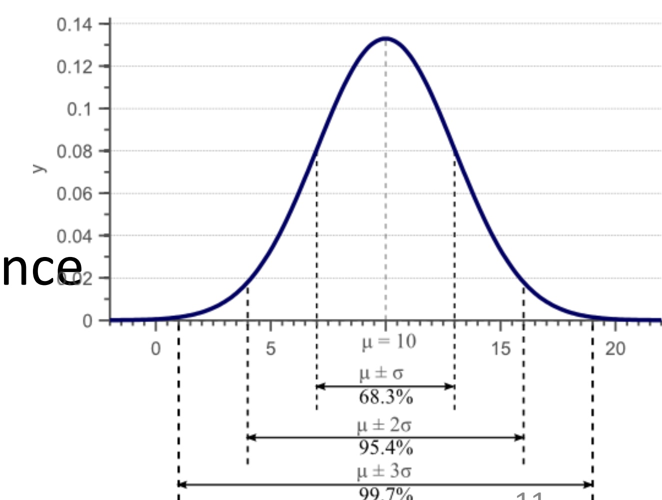
- Mean and Median are measures of location
- It is often desirable to consider measures of variability:
  - Variance & Standard deviation
- Variance
  - a measure of variability that utilises all data
  - average of the squared differences between data values and the means

$$var(X) = \sigma^2 = E[(X - \bar{X})^2], where\ E(.)\ denotes\ expected\ value, i.e. mean.$$

- Standard deviation
  - is the positive squared root of the variance
  - is measured in the same unit as the data,

  making it more easily interpreted than the variance

$$\sigma(X) = \sqrt{var(X)}$$



Variance and Covariance

- Recap, variance is defined as:

$$var(X) = \sigma^2 = E[(X - \bar{X})^2]$$

- The covariance between two (random) variables X_1 and X_2 is defined as:

$$Cov(X_1, X_2) = E[(X_1 - E(X_1))(X_2 - E(X_2))]$$

- The variance is a special covariance of a variable with itself:

$$Cov(X, X) = E[(X - E(X))(X - E(X))]$$

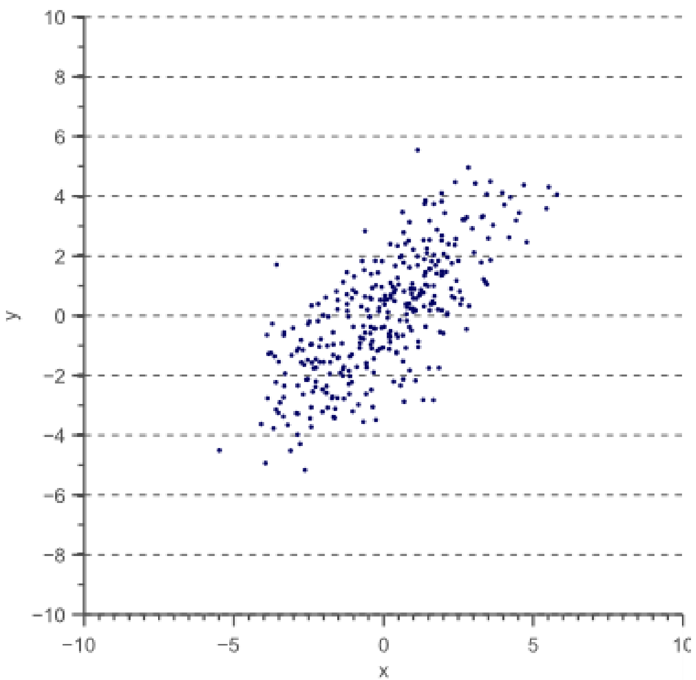
Variance and Covariance

- Zero-centred values
  - Subtract the mean (=E[X]) from observed variables
  - For zero-centred variables, the covariance simplifies to:

$$Cov(X_1, X_2) = E\left[\left(X_1 - E(X_1)\right)\left(X_2 - E(X_2)\right)\right] = E(X_1 X_2)$$

  - And variance simplifies to:

$$var(X) = \sigma^2 = E[X^2]$$



- Var(x): spread in horizontal
- Var(y): spread in vertical
- Cov(x,y): diagonal spread

Covariance

- Example: two dimensional data

	Hours(H)	*Mark(M)*
Data	9	39
	15	56
	25	93
	14	61
	10	50
	18	75
	0	32
	16	85
	5	42
	19	70
	16	66
	20	80
Totals	167	749
Averages	13.92	62.42

Covariance

- Example: two dimensional data

H	M	$(H_i - \bar{H})$	$(M_i - \bar{M})$	$(H_i - \bar{H})(M_i - \bar{M})$
9	39	-4.92	-23.42	115.23
15	56	1.08	-6.42	-6.93
25	93	11.08	30.58	338.83
14	61	0.08	-1.42	-0.11
10	50	-3.92	-12.42	48.69
18	75	4.08	12.58	51.33
0	32	-13.92	-30.42	423.45
16	85	2.08	22.58	46.97
5	42	-8.92	-20.42	182.15
19	70	5.08	7.58	38.51
16	66	2.08	3.58	7.45
20	80	6.08	17.58	106.89
Total				1149.89
Average				104.54

Covariance Matrix

- Covariance matrix for a 3-dimensional data:

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

- Covariance matrix for n-dimensional data:
  - The matrix is symmetrical about the main diagonal (top left to bottom right)
  - Along the main diagonal, the matrix contains the variance values

$$C^{n \times n} = (c_{i,j},\ c_{i,j} = cov(Dim_i, Dim_j))$$

$$cov(a,b) = cov(b,a)$$

Covariance Matrix

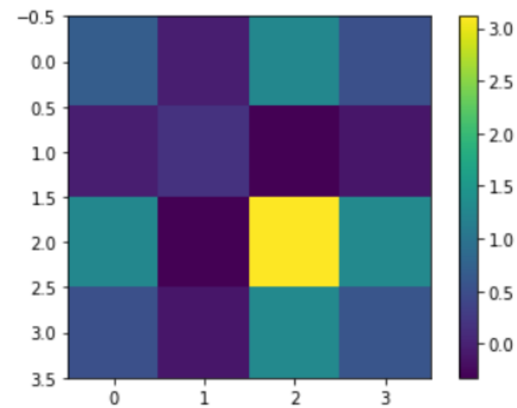
- Covariance matrix for a general d-dimensional data:

$$\sigma(x_k, x_k) = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ki} - \bar{x}_k)^2, k = 1, 2, \ldots, d$$

$$\sigma(x_m, x_k) = \frac{1}{n-1} \sum_{i=1}^{n} (x_{mi} - \bar{x}_m)(x_{ki} - \bar{x}_k) \quad \sigma(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\Sigma = \begin{bmatrix} \sigma(x_1, x_1) & \cdots & \sigma(x_1, x_d) \\ \vdots & \ddots & \vdots \\ \sigma(x_d, x_1) & \cdots & \sigma(x_d, x_d) \end{bmatrix} \in \mathbb{R}^{d \times d}$$

```
[[ 0.68569351 -0.03926846  1.27368233  0.5169038 ]
 [-0.03926846  0.18800403 -0.32171275 -0.11798121]
 [ 1.27368233 -0.32171275  3.11317942  1.29638747]
 [ 0.5169038  -0.11798121  1.29638747  0.58241432]]
```



The covariance matrix of the iris centered data

Covariance Matrix

- A quick way to compute: an example
- We have the following data set in 3D with each 2 samples

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 1 \end{bmatrix} = \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix}$$

- 1) Compute the average in each dimension

$$\bar{c} = \begin{bmatrix} 2 & 1.5 & 2 \end{bmatrix} = \begin{bmatrix} \bar{c_1} & \bar{c_2} & \bar{c_3} \end{bmatrix}$$

- 2) Each column' values subtract the averages $c_i = c_i - \bar{c_i}$

$$X = \begin{bmatrix} -1 & 0.5 & 1 \\ 1 & -0.5 & -1 \end{bmatrix}$$

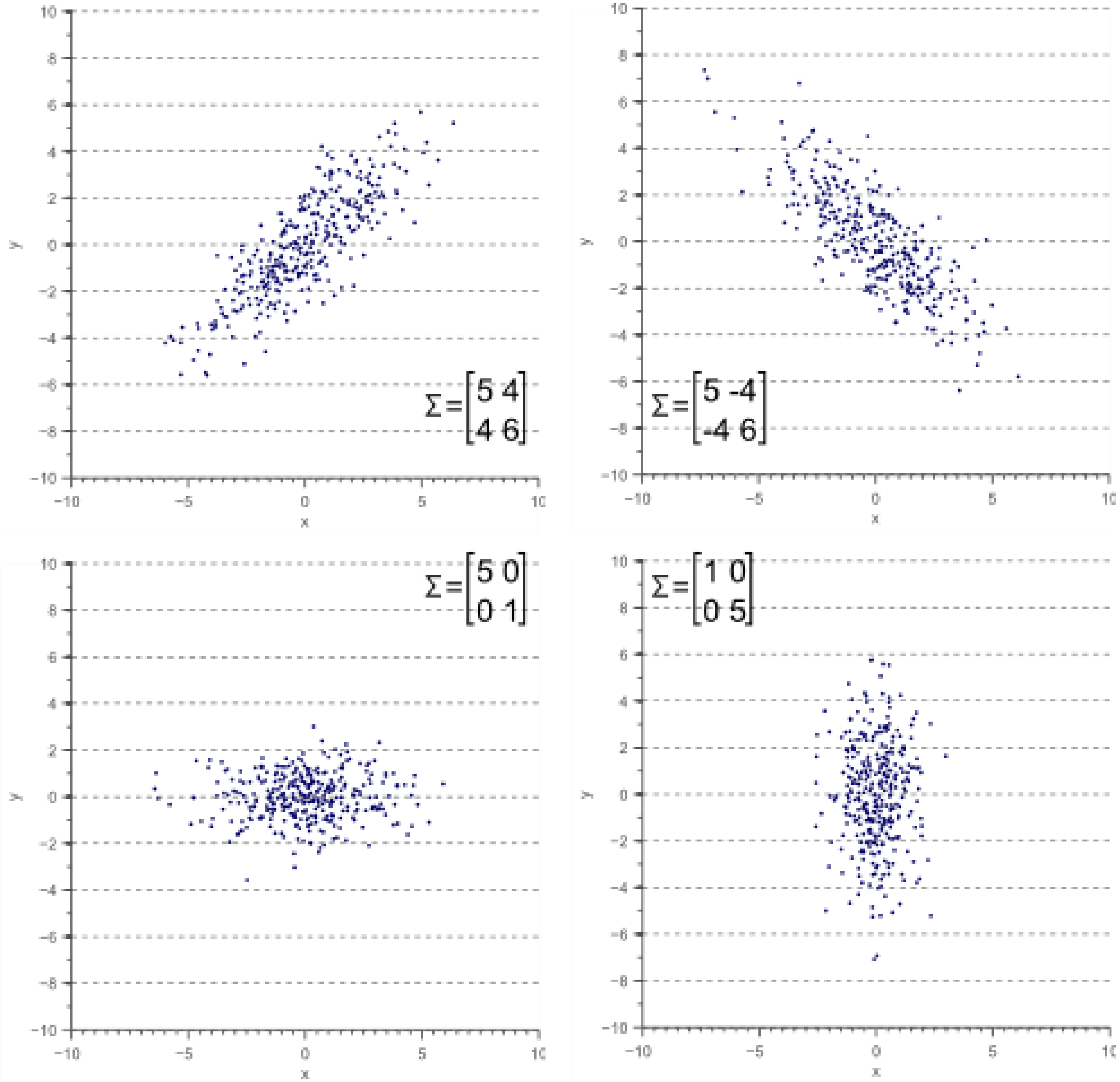
- 3) Compute the covariance matrix

Matrix form for covariance computation

$$\boxed{cov = \frac{1}{m-1} X^T X = \frac{1}{2-1} \begin{bmatrix} 2 & -1 & -2 \\ -1 & 0.5 & 1 \\ -2 & 1 & 2 \end{bmatrix}}$$

Covariance Matrix

- Examples

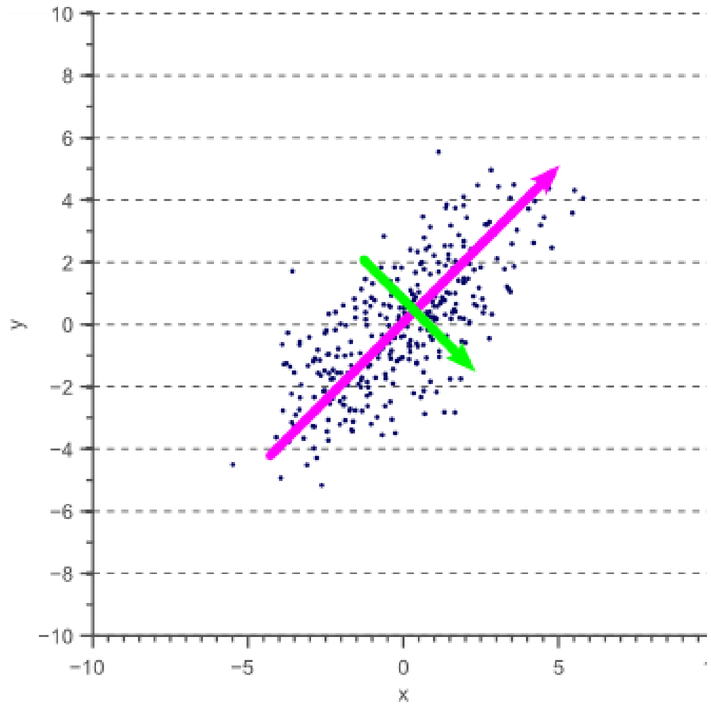


- The covariance matrix Σ defines the shape of the data.
- Diagonal spread is captured by covariance.
- Axis-aligned spread is captured by variance.

If cov(x,y)=0, we say x and y is uncorrelated or decorrelated.

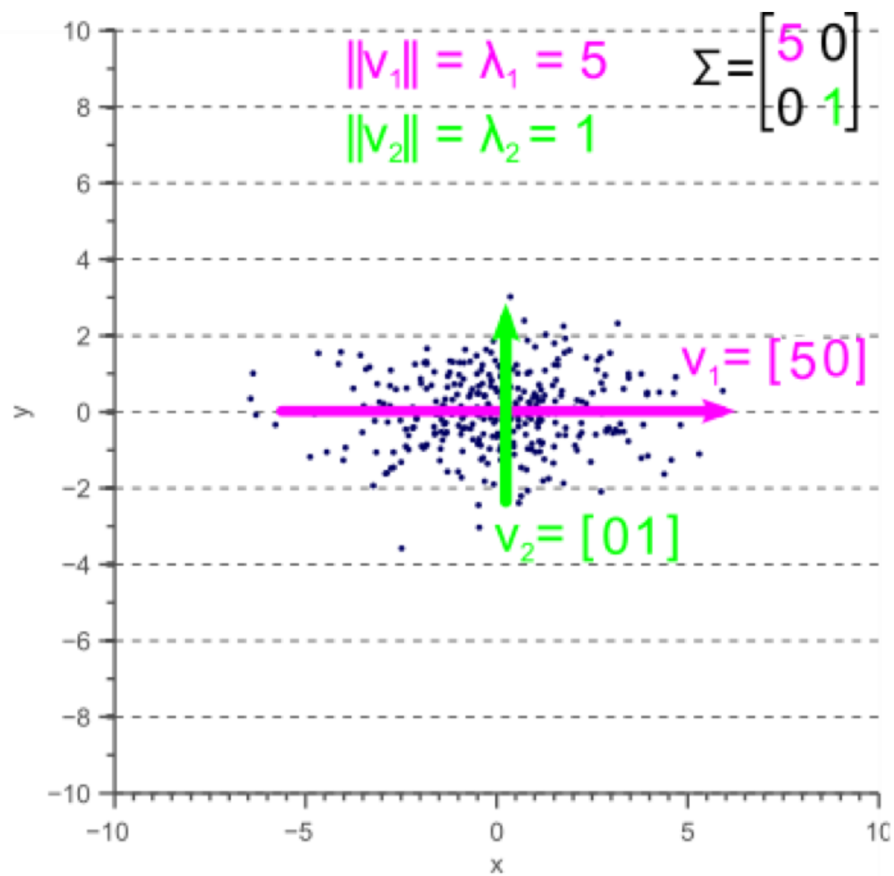
Eigenvectors and Eigenvalues for PCA

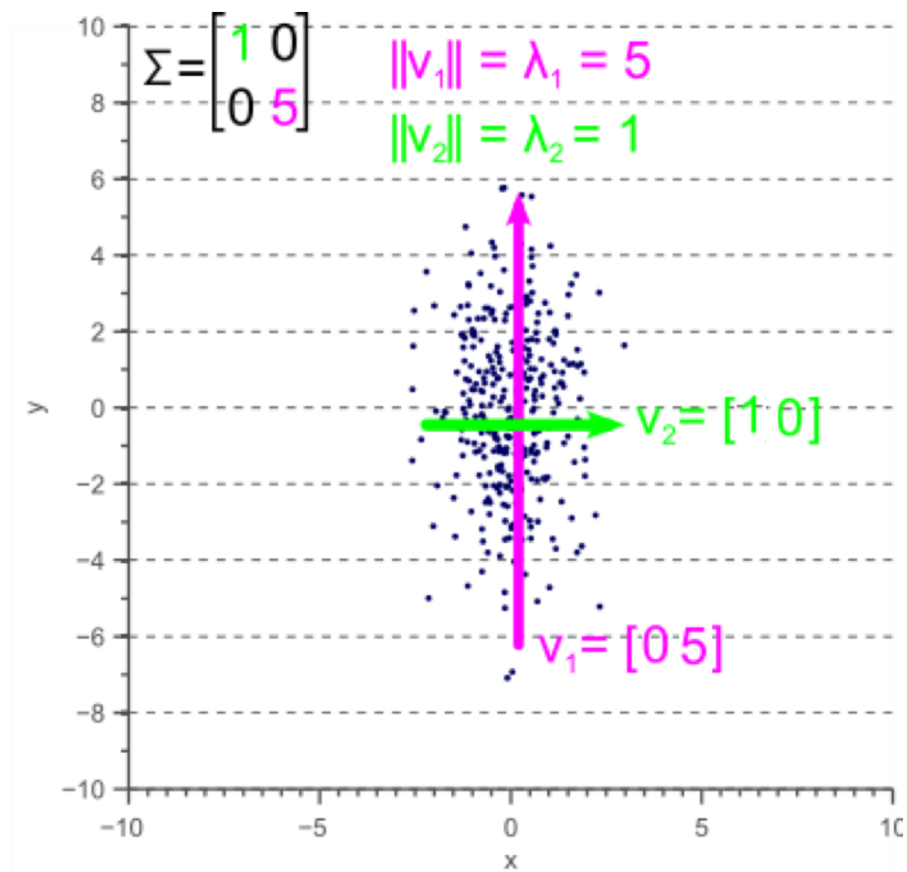
- Covariance matrix defines both the spread (variance), and the orientation (covariance) of the data
- The vector that points into the direction of the largest spread of the data is the **eigenvector** with the largest **eigenvalue**
- This eigenvalue equals the spread (variance) in this direction (defined by the corresponding eigenvector)



Eigenvectors and Eigenvalues

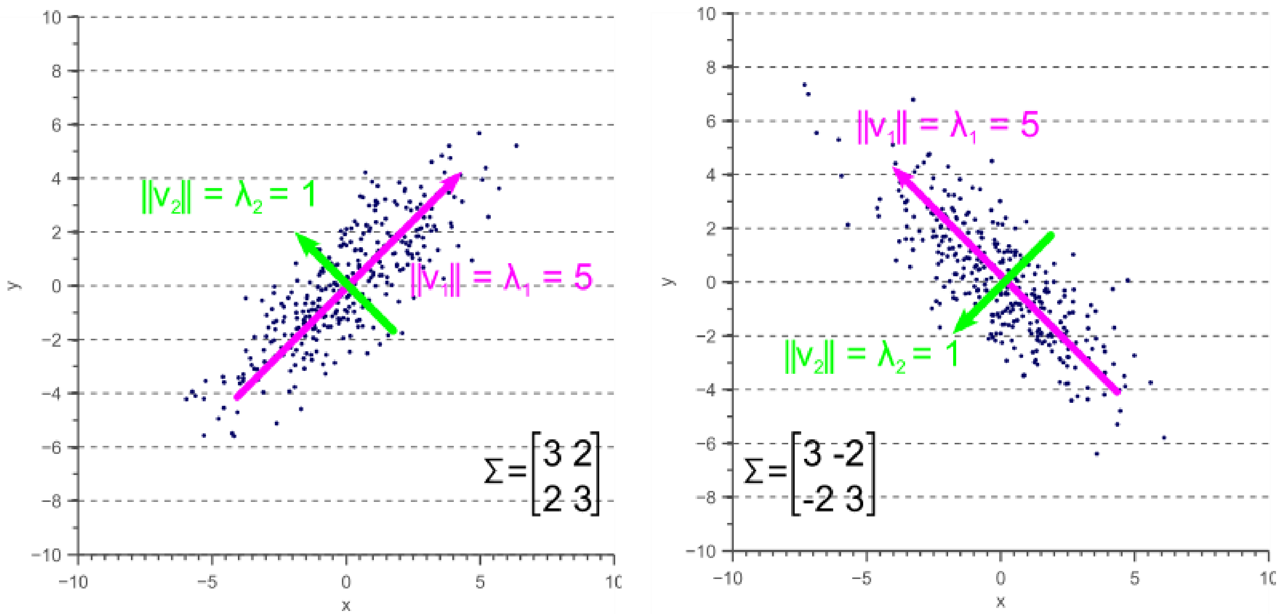
- If the covariance matrix of our data is a diagonal matrix, such that the covariances are zero, then this means that the variances must be equal to the eigenvalues λ





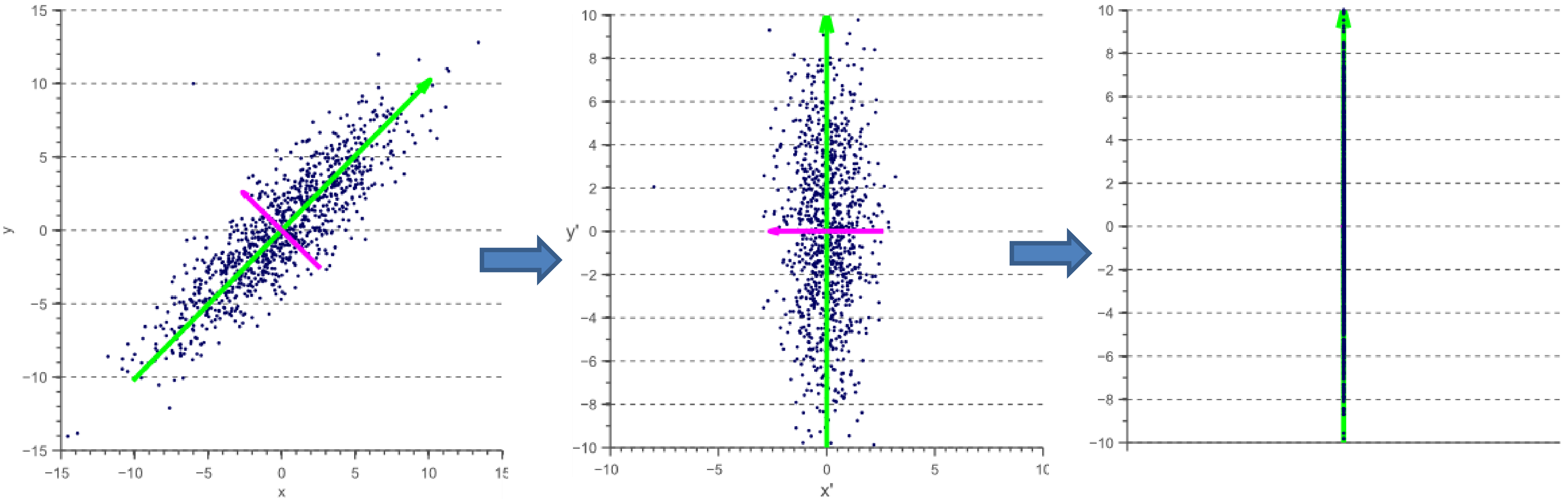
Eigenvectors and Eigenvalues

- If the covariance matrix is not diagonal, such that the covariances are not zero,
  - The eigenvalues still represent the variance magnitude in the direction of the largest spread of the data,
  - the variance components of the covariance matrix still represent the variance magnitude in the direction of the x-axis and y-axis.
  - But since the data is not axis aligned, these values are not the same anymore



Principal Component Analysis

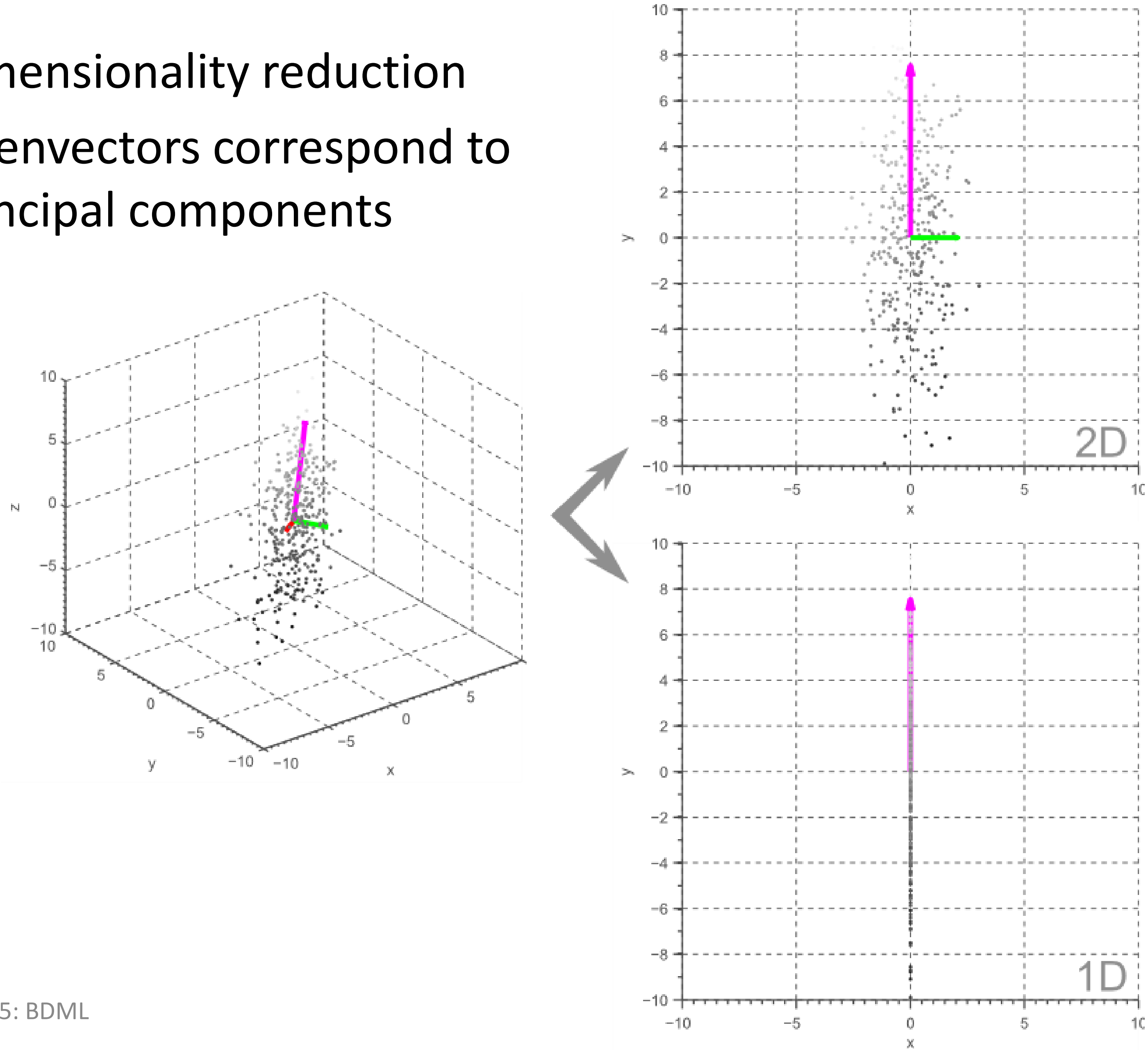
- PCA is a decorrelation method
  - Linearly transforms the data so that covariance values are all zeros
  - Retain the components with largest variances
  - Rid of components with small variances to achieve dimensionality reduction



$$\Sigma = \begin{bmatrix} 16.87 & 14.94 \\ 14.94 & 17.27 \end{bmatrix} \qquad \Sigma' = \begin{bmatrix} 1.06 & 0.0 \\ 0.0 & 16.0 \end{bmatrix} \qquad \text{variance} = 16.0$$

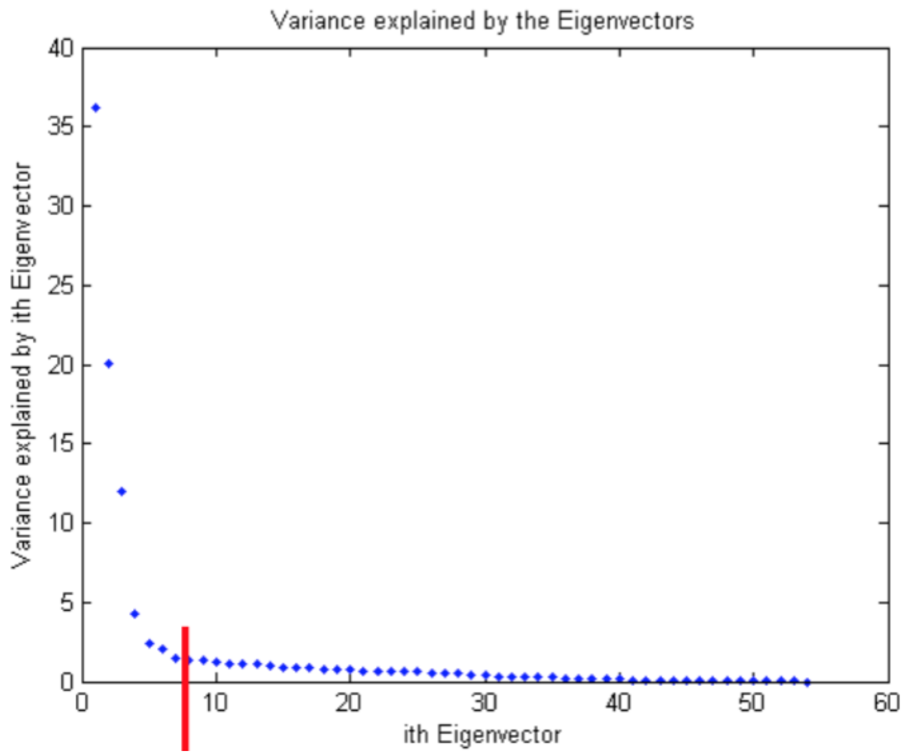
Principal Component Analysis

- Dimensionality reduction
- Eigenvectors correspond to principal components



Principal Component Analysis

- Dimensionality reduction
  - List the eigenvalues in descending order
  - Set a threshold and remove principal components that have small variances (small eigenvalues)
  - The data is then projected back with reduced dimensionality



How to compute Principal Component Analysis (PCA)

PCA is an **unsupervised** technique, there is no outcome variable (Y), let the data speak for itself (X)!

Step1. Considering data matrix $\boldsymbol{X_{N\times P}}$ and its square shape covariance $\boldsymbol{\Sigma_{P\times P}}$, the roots of the following characteristic equation gives the **eigenvalues (λ)** of $\boldsymbol{\Sigma}$:

$$\boldsymbol{det(\Sigma - \lambda I) = 0},$$

λ is a scaler and is called **eigenvalue** of Σ and I is the identity matrix.

Step2. For each of the eigenvalues, there is a corresponding **eigenvector**, $\boldsymbol{v_{P\times 1}}$ and can be found by solving:

$$\boldsymbol{\Sigma \mathrm{v} = \lambda \mathrm{v}}$$

This can also be illustrated in **Matrix form:** If **V=[v1,v2,…,vP]** is the (PxP) matrix of the eigenvectors , we have the matrix form equation:

$$\boldsymbol{\Sigma V = V \Lambda},$$

where $\Lambda_{P\times P} = \begin{bmatrix} \lambda_1 & \dots & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_P \end{bmatrix}$, is the diagonal matrix of the eigenvalues .

Eigen value decomposition (EVD)

$$\boldsymbol{\Sigma V} = \boldsymbol{V\Lambda} \quad \Rightarrow \quad \boldsymbol{V^T \Sigma V} = \boldsymbol{\Lambda}, \boldsymbol{\Sigma} = \boldsymbol{V^T \Lambda V}.$$

V is normalised and has unit magnitude and they are orthogonal, so that, $V^T V = V V^T = I$,therefore $V^T \Sigma V = \Lambda$ and $\boldsymbol{\Sigma} = \boldsymbol{V^T \Lambda V}$. This is **Eigen decomposition** of the matrix Σ .

Singular value decomposition (SVD)

Eigen decomposition of Σ is connected to the **singular value decomposition (SVD)** of the data matrix X:

$$\boldsymbol{X} = \boldsymbol{USV^T}$$

This is a standard decomposition in numerical analysis.

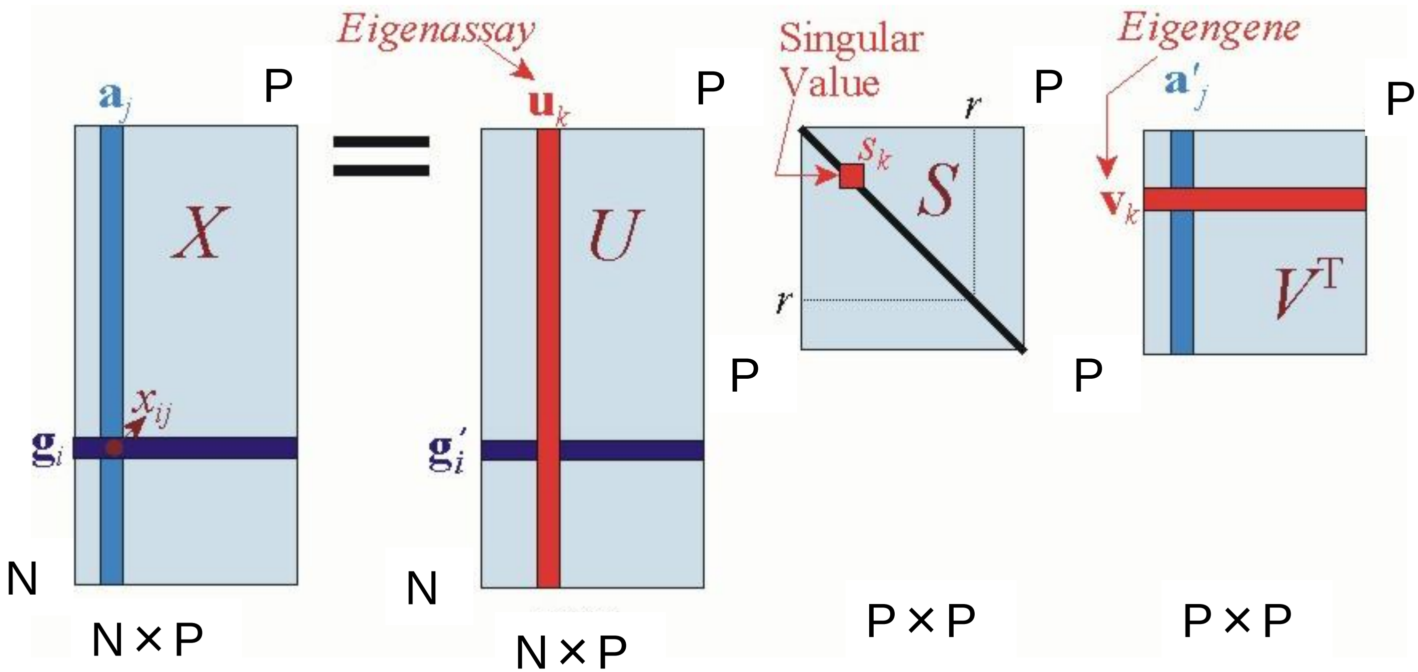
$V_{P\times P}$ is the orthogonal eigen vector matrix and the v_i are called the **right singular vectors**.

$U_{N\times P}$ is also orthogonal $U^T U = I$ and its columns u_i are called the **left singular vectors**.

$S_{P\times P}$ is a diagonal matrix, with diagonal elements s1 ≥ s2 ≥ · · · ≥ sp ≥ 0 known as the **singular values** and $s_i = \sqrt{\lambda_i}$.

Illustration of SVD for genetic data

- Singular Value Decomposition
  - For any matrix X: $X = USV^T$



Data X, one row per data point
Data is zero-centred

US gives coordinates of rows of X in the space of principle components

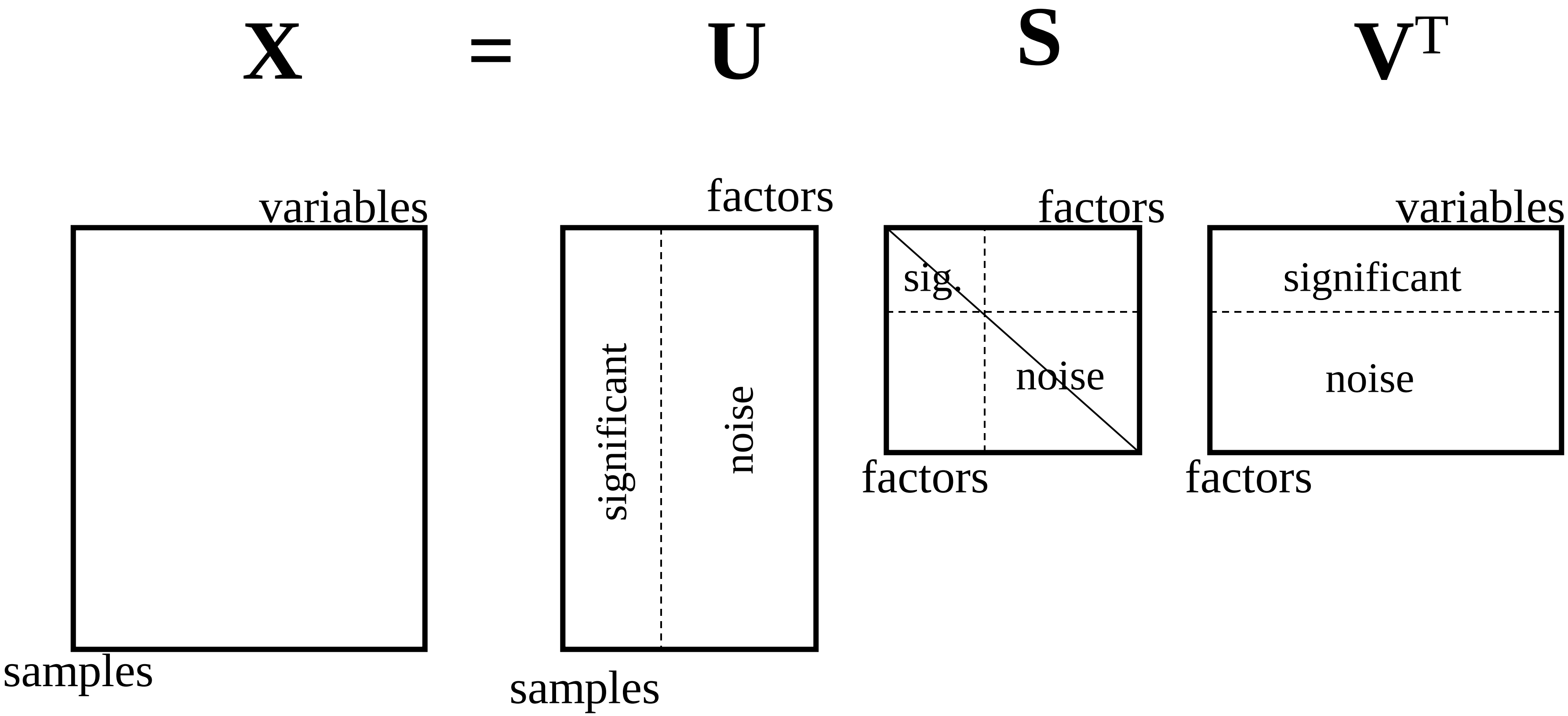
S is diagonal, $S_k > S_{k+1}$, S_k is k[th] largest eigenvalue

Rows of V^T are unit length eigenvectors

http://public.lanl.gov/mewall/kluwer2002.html

SVD interpretation

- PCA dimensionality reduction
  - Setting “noise” to zero to achieve reduced dimensionality



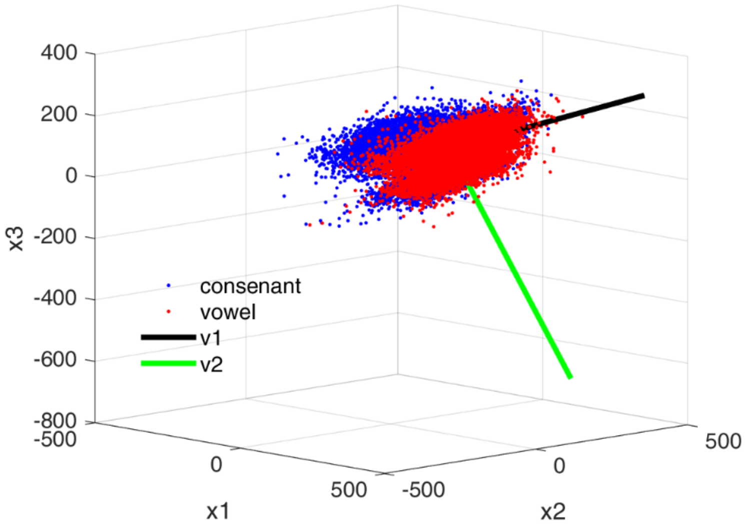
Data Projection :

- The matrix of eigenvectors **V** can be considered as a **linear transformation** which can transforms points from original coordinate system (x1,x2, … xP) into a new system (v1,v2, …, vp).
- The variables of the **transformed dataset** are **uncorrelated**.

- The **covariance matrix** of the data in the new coordinate system is **Λ** which has **zeros in all the off diagonal elements**.

- Then, **each λ_i** explains the **variance of data** along each orthogonal **direction vi.**
- **The directions are sorted based on their corresponding variance λ1>λ2>…> λP.**

$$\Lambda_{P\times P} = \begin{bmatrix} \lambda_1 & \dots & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_P \end{bmatrix} = \begin{bmatrix} \delta_1^2 & \dots & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \delta_P^2 \end{bmatrix}$$



Visualization of 3 phoneme features

Dimension reduction:

Considering the first **d** eigen values that explain most of the variations of data $\sum_{i=1}^{d} \lambda_i > \tau$, and their corresponding columns of V=[v1,…,vd], the dimensionality of data in the new orthogonal space can be reduced:

$$Z_{N\times d} = X_{\,N\times P} V_{P\times d} = U_{N\times P} S_{P\times d} \,.$$

The columns of **Z=US** are called the **principal components** of **X** .

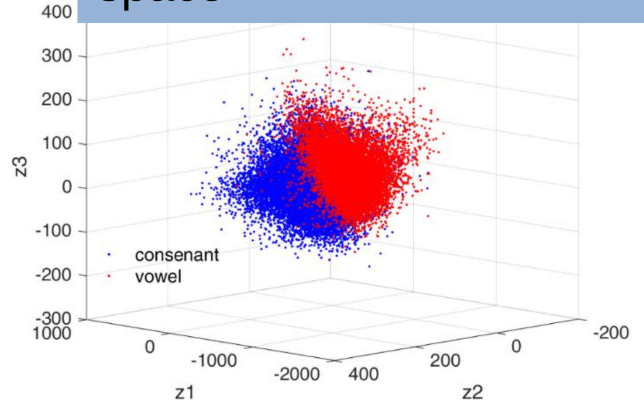
Z=[z1 z2,…,zd]=[x1 x2 ,…,xP]*[v1 v2,…,vd]

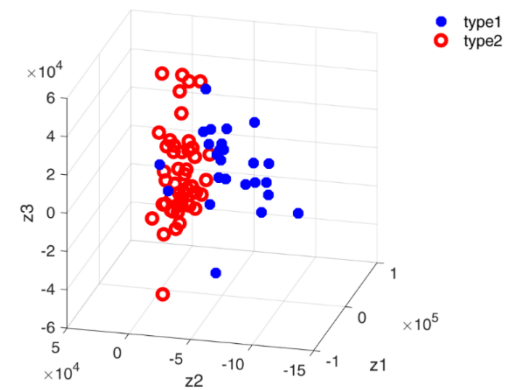
Instead of P number of variables, only d<P variables are available.

$\boldsymbol{\tau}$ is defined based on the maximum desired variations. For example:

$$\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{P} \lambda_i} > \mathbf{0.95} = \tau$$

Projection of 512 phoneme data into 3D orthogonal space





Projection of 7129 lukemia data into 3D orthogonal space

Linear transformation

- Z=[z1 z2,…,zd]=[x1 x2 ,…,xP]*[v1 v2,…,vd]

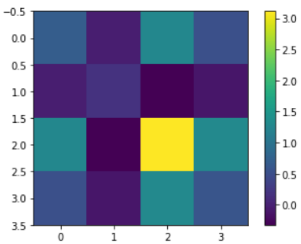
One Iris sample: x11=[-0.74 0.446 -2.358 -0.998]

Eigen matrix V is 4 × 4., [v1,v2,v3,v4]

Reduction to one:

```
[[ 0.68569351 -0.03926846  1.27368233  0.5169038 ]
 [-0.03926846  0.18800403 -0.32171275 -0.11798121]
 [ 1.27368233 -0.32171275  3.11317942  1.29638747]
 [ 0.5169038  -0.11798121  1.29638747  0.58241432]]
```

$$z11 = x_{11} \times v_1 = [-0.74 \quad 0.446 \quad -2.358 \quad -0.998] \begin{bmatrix} 0.36 \\ -0.08 \\ 0.856 \\ 0.358 \end{bmatrix}$$



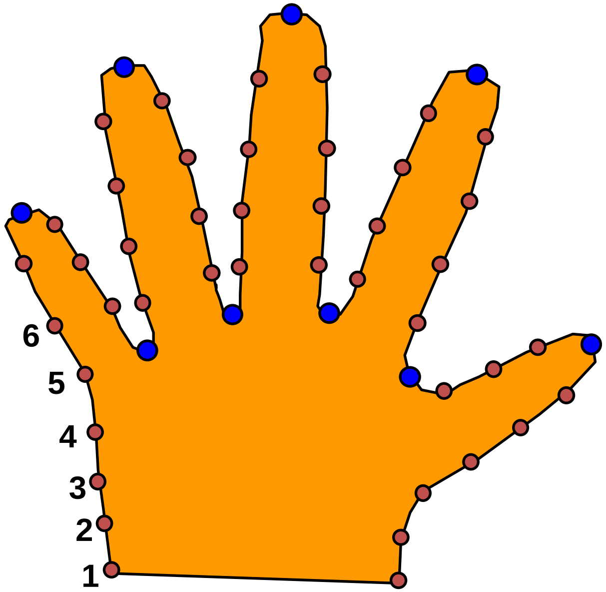
$$= 0.36(\textit{sepal length}) - 0.08(\textit{sepal width}) + 0.856(\textit{petal length}) + 0.358(\textit{petal width})$$

= -2.684

A linear combination of all original features is used to generate the transform feature z11

PCA Example

- Hand shape model
  - 72 points placed around boundary of hand
  - 18 hand outlines obtained by thresholding images of hand on a white background
  - Primary landmarks chosen at tips of fingers and joint between fingers
  - Other points placed equally between

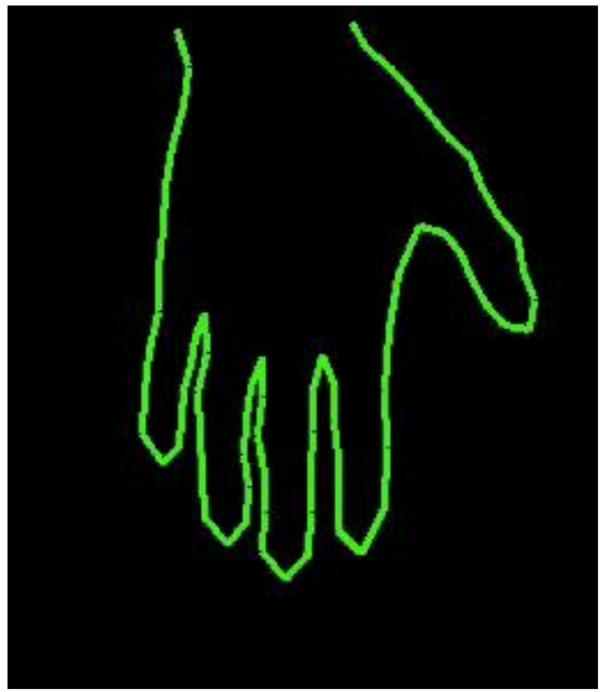


X (18 x 72)

PCA Example

- Hand Shape Model
  - varying shown by the largest three principal components

PC1



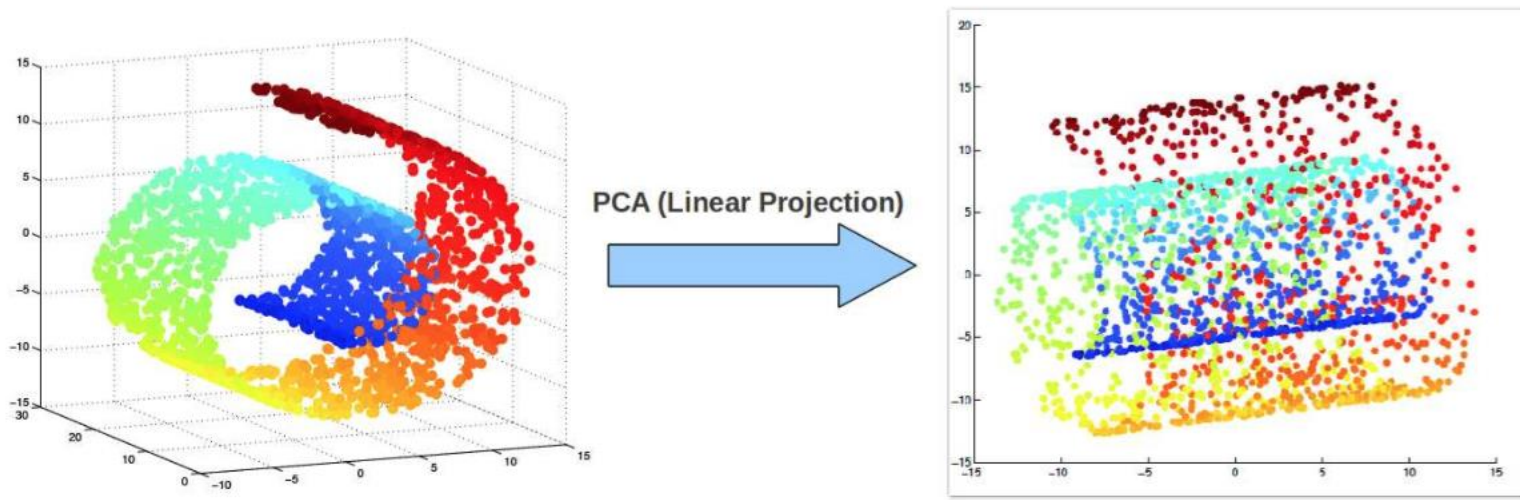
PC2

PC3

X (18 x 72) . V(72 x 3)

Principal Component Analysis

- Can not capture **intrinsic nonlinearity**
  - Because PCA uses linear projection
  - Methods, such as Kernel PCA, can be used to tackle nonlinearity



Example

- Consider the following matrix of 5 samples and 2 variables and compute the Eigen values and Eigen vectors based on EVD.

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{0} & \mathbf{-4} \\ \mathbf{0} & \mathbf{-2} \\ \mathbf{1} & \mathbf{-2} \\ \mathbf{3} & \mathbf{-1} \\ \mathbf{1} & \mathbf{-1} \end{bmatrix},$$

step1. centre the X: $Xc = X - \frac{1}{5-1}\sum_{i=1}^{5} X_i = \begin{bmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{bmatrix}$

step2. $Xc^T = \begin{bmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{bmatrix}$

Step3. Covariance matrix $\Sigma = \frac{1}{5-1} Xc^T Xc = \begin{bmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \end{bmatrix}$

Step 4. compute the Eigen values: $\boldsymbol{det(\Sigma - \lambda I) = 0}$

$$\det\left(\begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = 0 \rightarrow \begin{vmatrix} 1.5-\lambda & 1 \\ 1 & 1.5-\lambda \end{vmatrix} = 0$$

$$(1.5-\lambda)^2 - 1 = 0 \rightarrow \lambda^2 - 3\lambda + 1.25 = 0$$

$$\lambda_1 = 2.5 \quad , \lambda_2 = 0.5$$

Step 5. computing Eigen Vectors:

$$\boldsymbol{\Sigma V = V\Lambda} \rightarrow \begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \end{bmatrix} \begin{bmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{bmatrix} = \begin{bmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{bmatrix} . \begin{bmatrix} 2.5 & 0 \\ 0 & 0.5 \end{bmatrix} \rightarrow$$

$$\begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{bmatrix} \begin{bmatrix} 2.5 \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{bmatrix} \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}$$

$$\begin{bmatrix} 1.5v_{11} + v_{12} \\ v_{11} + 1.5v_{12} \end{bmatrix} = \begin{bmatrix} 2.5v_{11} \\ 2.5v_{12} \end{bmatrix} \rightarrow \begin{matrix} -v_{11} + v_{12} = 0 \\ v_{11} - v_{12} = 0 \end{matrix} \rightarrow v_{11} = v_{12} = 1$$

$$\begin{bmatrix} 1.5v_{21} & v_{22} \\ v_{21} & 1.5v_{22} \end{bmatrix} = \begin{bmatrix} 0.5v_{21} \\ 0.5v_{22} \end{bmatrix} \rightarrow \begin{matrix} v_{21} + v_{22} = 0 \\ v_{21} + v_{22} = 0 \end{matrix} \rightarrow v_{21} = -1, v_{22} = 1$$

$$V = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Step6. normalising Eigen vectors to unit length

$$V_1 = \frac{1}{\sqrt{1^2+1^2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.7 \end{bmatrix}, \; V_2 = \frac{1}{\sqrt{-1^2+1^2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.7 \\ 0.7 \end{bmatrix}$$

Example

- Consider the covariance Σ= $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, find the eigen values and eigen vectors.
- The characteristic equation is $\det(\Sigma - \lambda I) = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 3 - 4\lambda + \lambda^2 = 0$,
- $\lambda_1 = 1$ and $\lambda_2 = 3$.
- The eigen vector matrix is $V = [v_1 \quad v_2] = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$.
- We find each eigen vector using the corresponding eigen value:
- $\Sigma v_1 = \lambda_1 \, v_1 \rightarrow \begin{bmatrix} 2v_{11} + v_{21} \\ v_{11} + 2v_{21} \end{bmatrix} = 1 \begin{bmatrix} v_{11} \\ v_{21} \end{bmatrix} \rightarrow \begin{cases} v_{11} + v_{21} = 0 \\ v_{11} + v_{21} = 0 \end{cases} \rightarrow v_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$
- $\Sigma v_2 = \lambda_2 \, v_2 \rightarrow \begin{bmatrix} 2v_{12} + v_{22} \\ v_{12} + 2v_{22} \end{bmatrix} = 3 \begin{bmatrix} v_{12} \\ v_{22} \end{bmatrix} \rightarrow \begin{cases} -v_{12} + v_{22} = 0 \\ v_{12} - v_{22} = 0 \end{cases} \rightarrow v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
- In this example, the **normalized eigenvectors** are $v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

Quiz

- A 2D data contains 2 classes
  - Magenta and green lines indicate two different dimensionality reduction results
    - The lines are the resulting 1D axis
    - Which one is better for the purpose of classification?