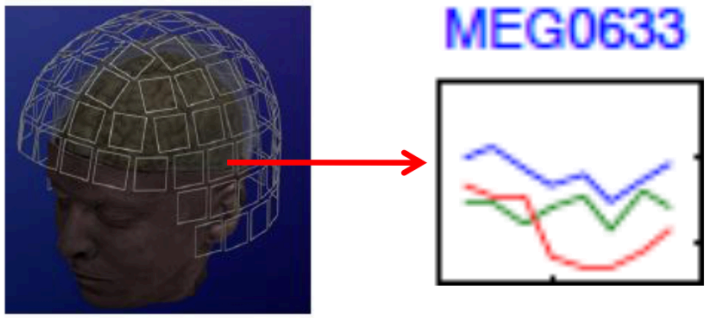
CSC345/M45:
Big Data & Machine Learning
(dimensionality reduction: PCA)

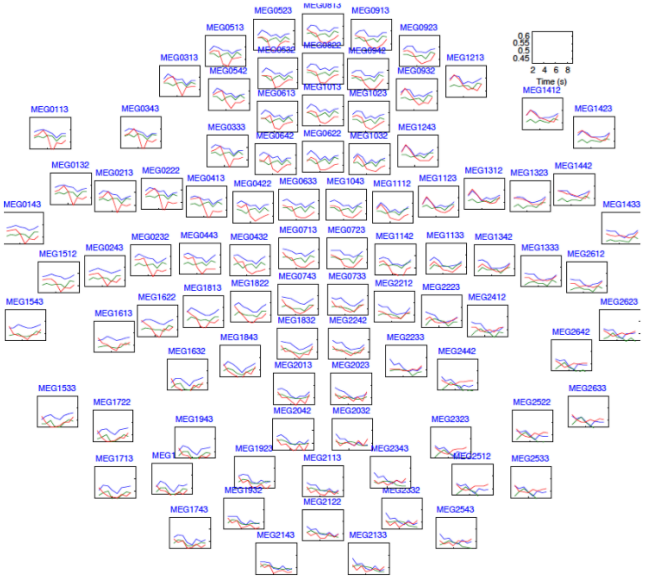
Prof. Xianghua Xie

x.xie@swansea.ac.uk
http://csvision.swan.ac.uk
224 Computational Foundry, Bay Campus

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



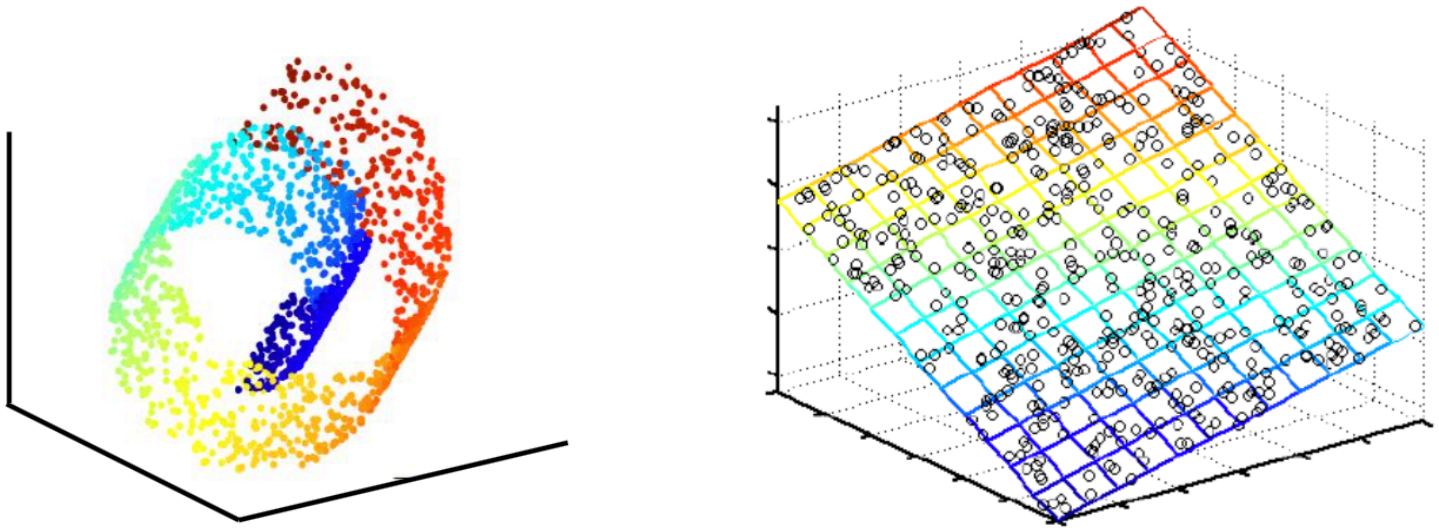


Dimensionality Reduction

- Curse of dimensionality
  - redundant features
    - e.g. not all words are useful in classifying documents: and, or, the, of, …
  - 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

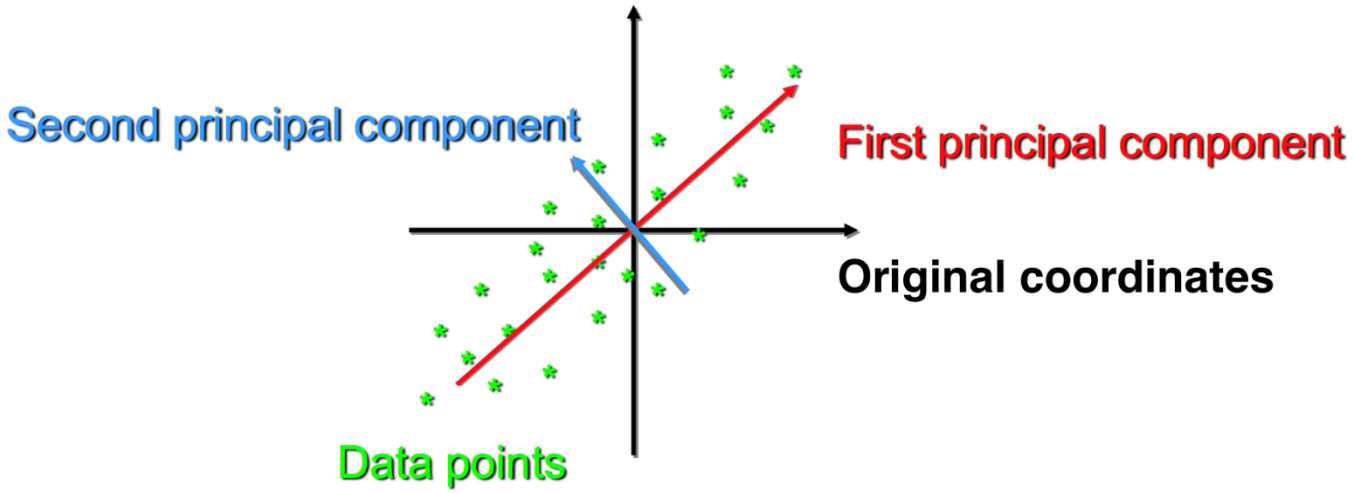
- Why dimensionality reduction?
  - Reduce the dimensionality of the data while maintaining the meaningfulness of the data
  - Find a low-dimensional but useful representation of the data
  - Discover "intrinsic dimensionality" of the data
    - some high dimensional data is actually low dimensional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



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

Principal Component Analysis

- 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.

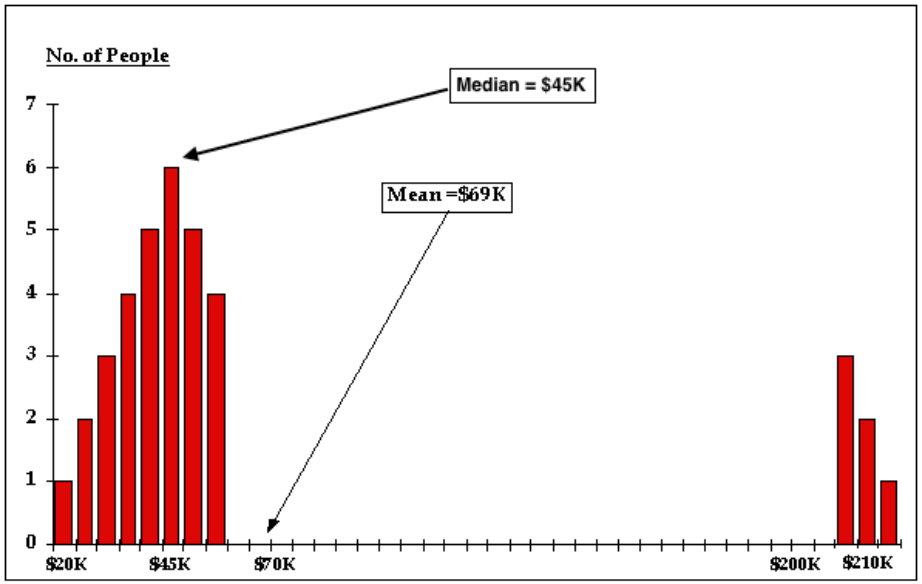


Mean and Median

- 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

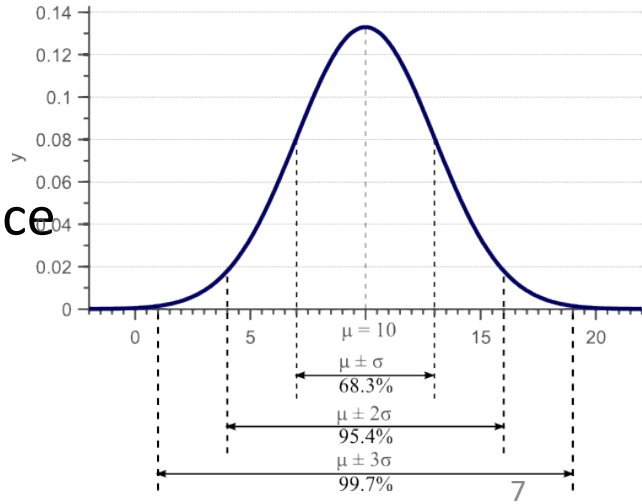
- Mean and Median are measures of location
- It is often desirable to consider measures of variability:
  - Variance & Standard deviation
- Variance
  - a measure of variability that utilises all data
  - average of the squared differences between data values and the means

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

- 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:

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

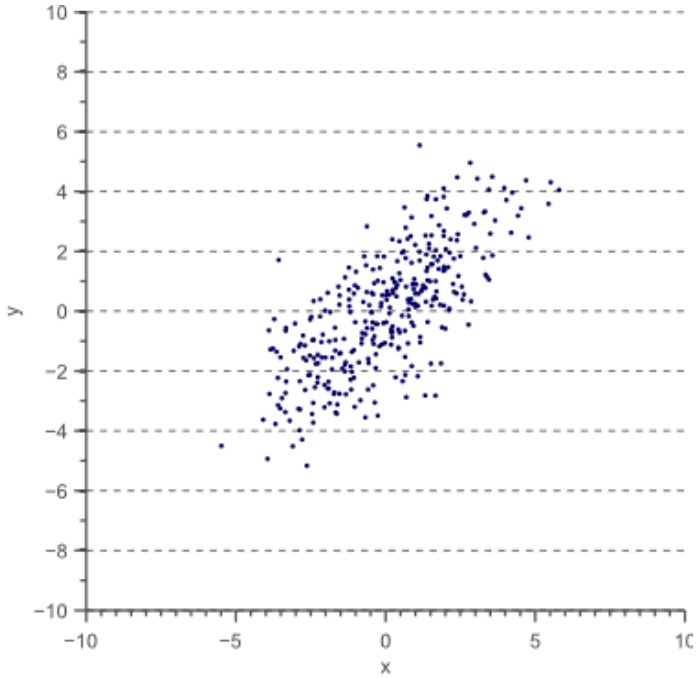
Variance and Covariance

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

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

  - 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:

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

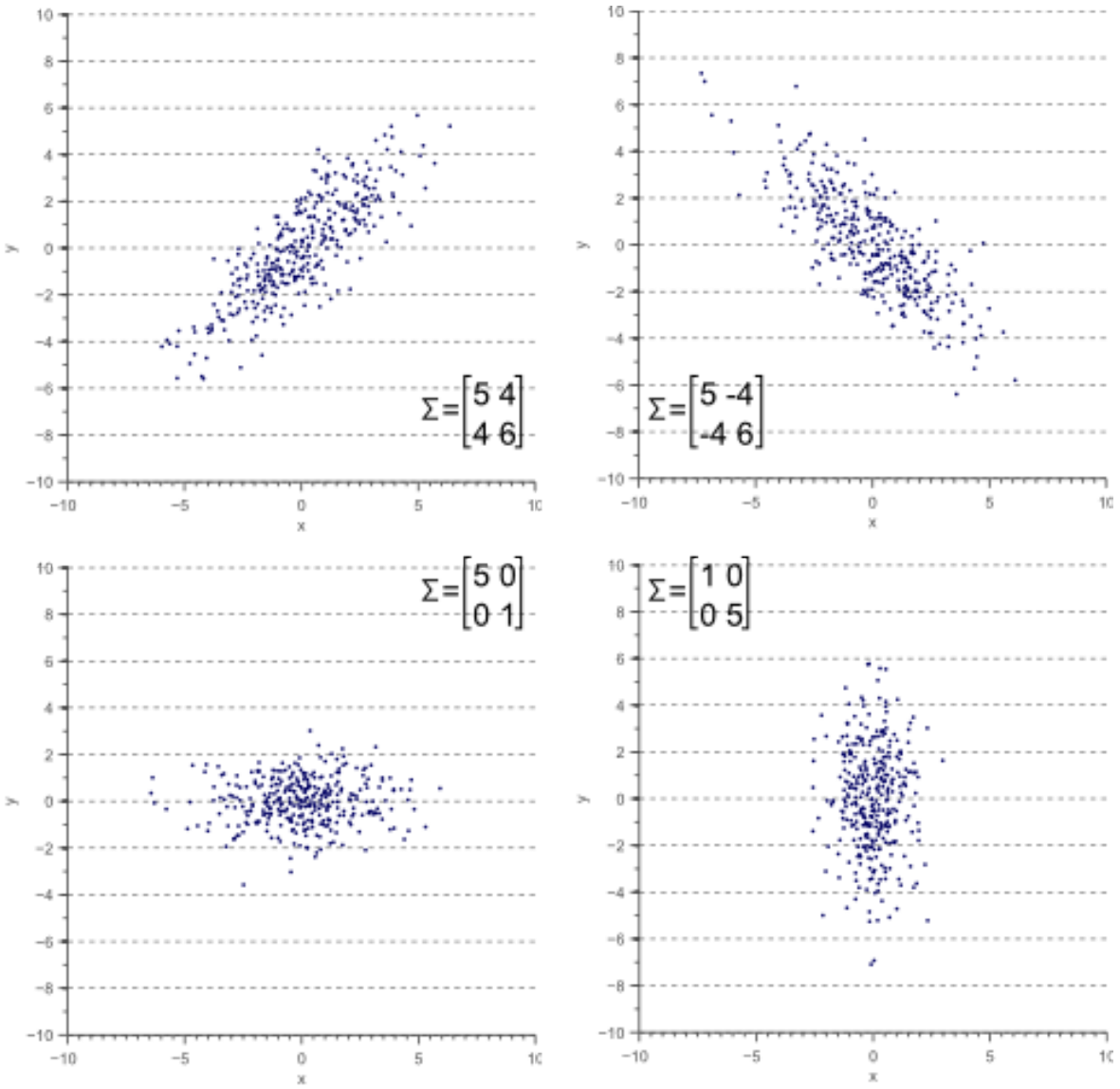
- Covariance matrix for n-dimensional data:
  - The matrix is symmetrical about the main diagonal (top left to bottom right)
  - Along the main diagonal, the matrix contains the variance values

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

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

Covariance Matrix

- Examples

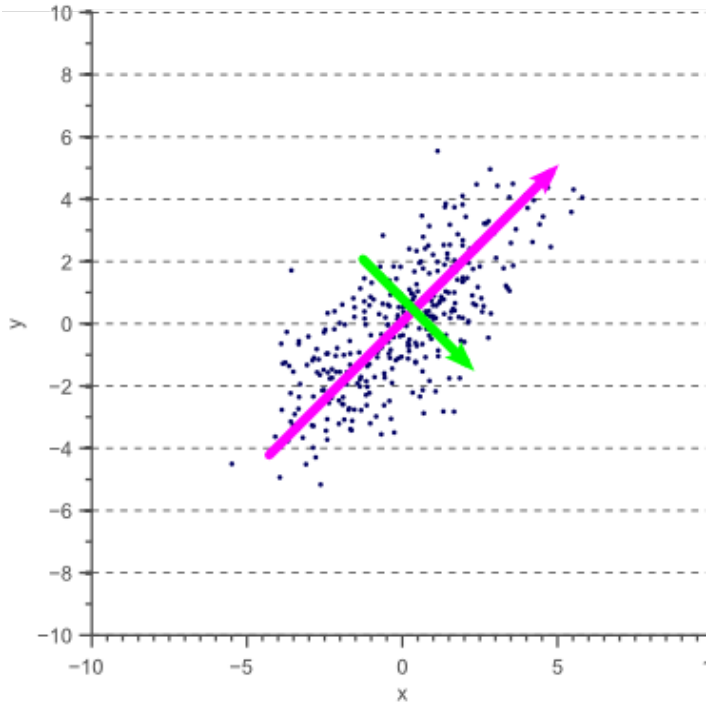


- The covariance matrix Σ defines the shape of the data.
- Diagonal spread is captured by covariance.
- Axis-aligned spread is captured by variance.

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

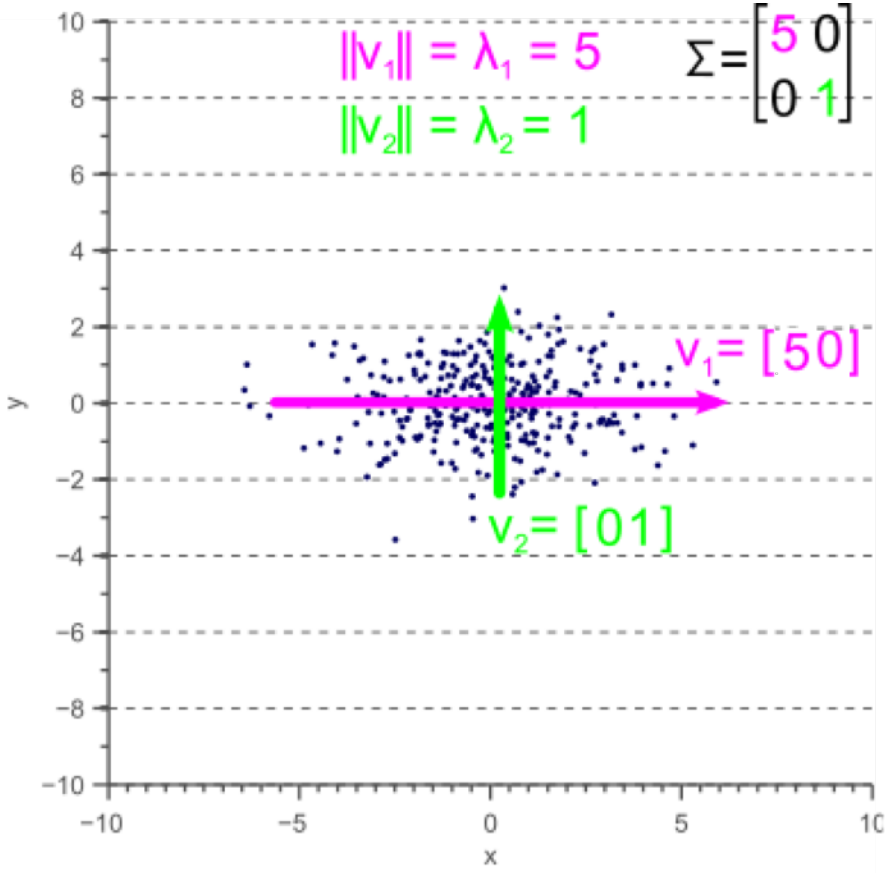
Eigenvectors and Eigenvalues

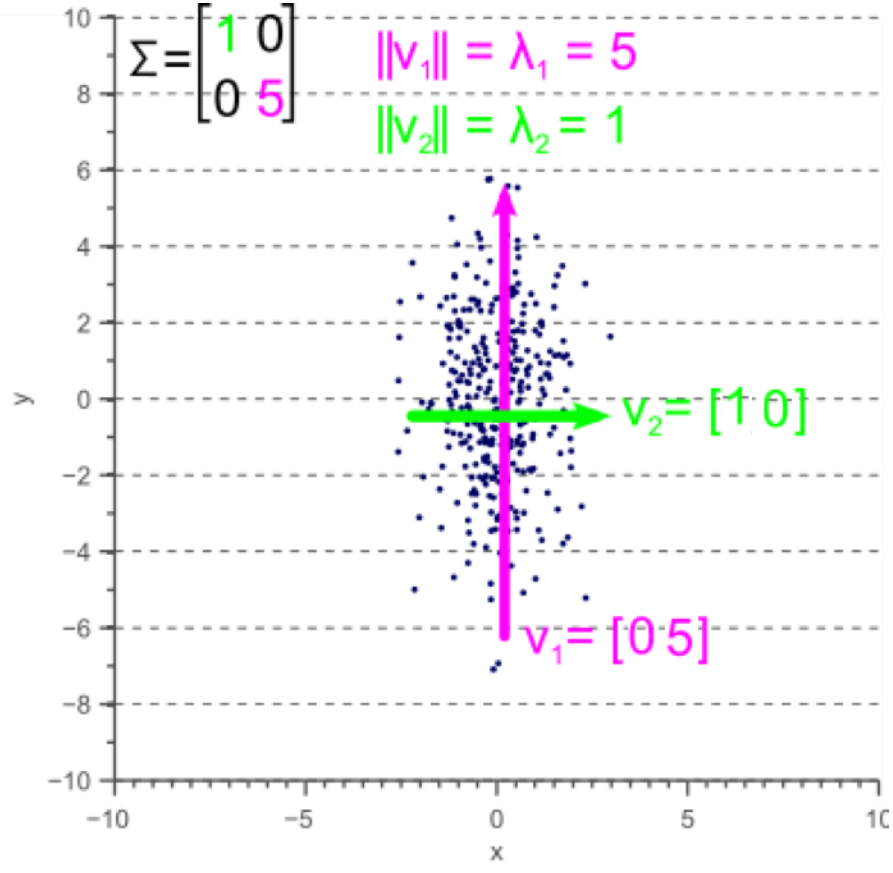
- Covariance matrix defines both the spread (variance), and the orientation (covariance) of the data
- The vector that points into the direction of the largest spread of the data is the **eigenvector** with the largest **eigenvalue**
- This eigenvalue equals the spread (variance) in this direction (defined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



Eigenvectors and Eigenvalues

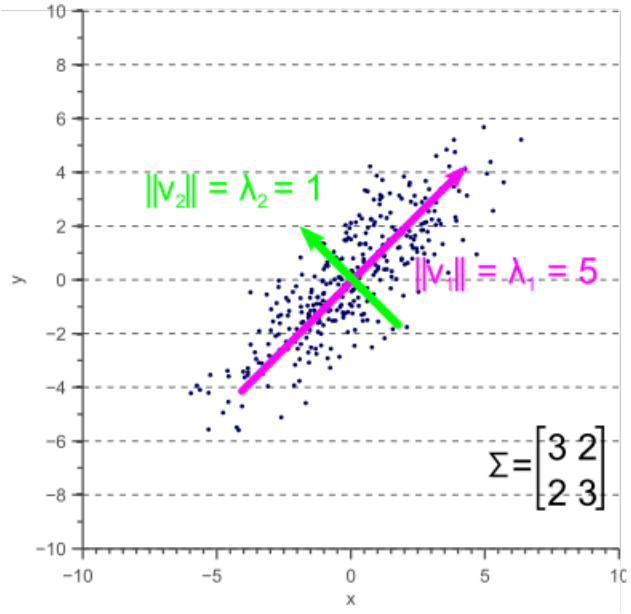
- If the covariance matrix of our data is a diagonal matrix, such that the covariances are zero, then this means that the variances must be equal to the eigenvalues λ

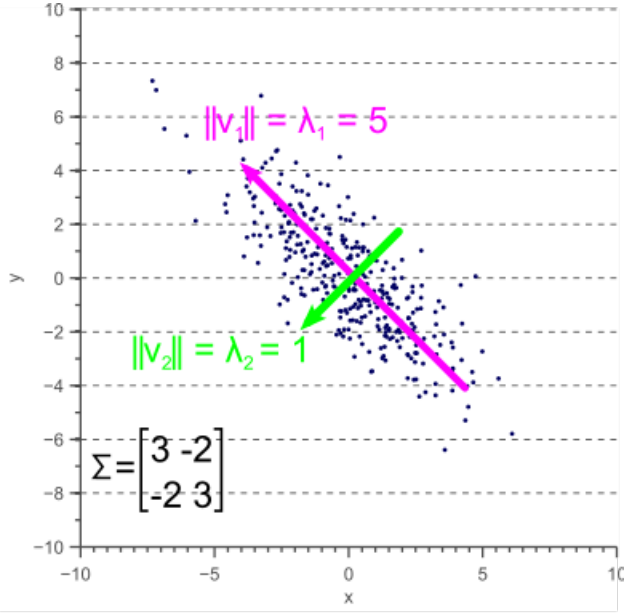




Eigenvectors and Eigenvalues

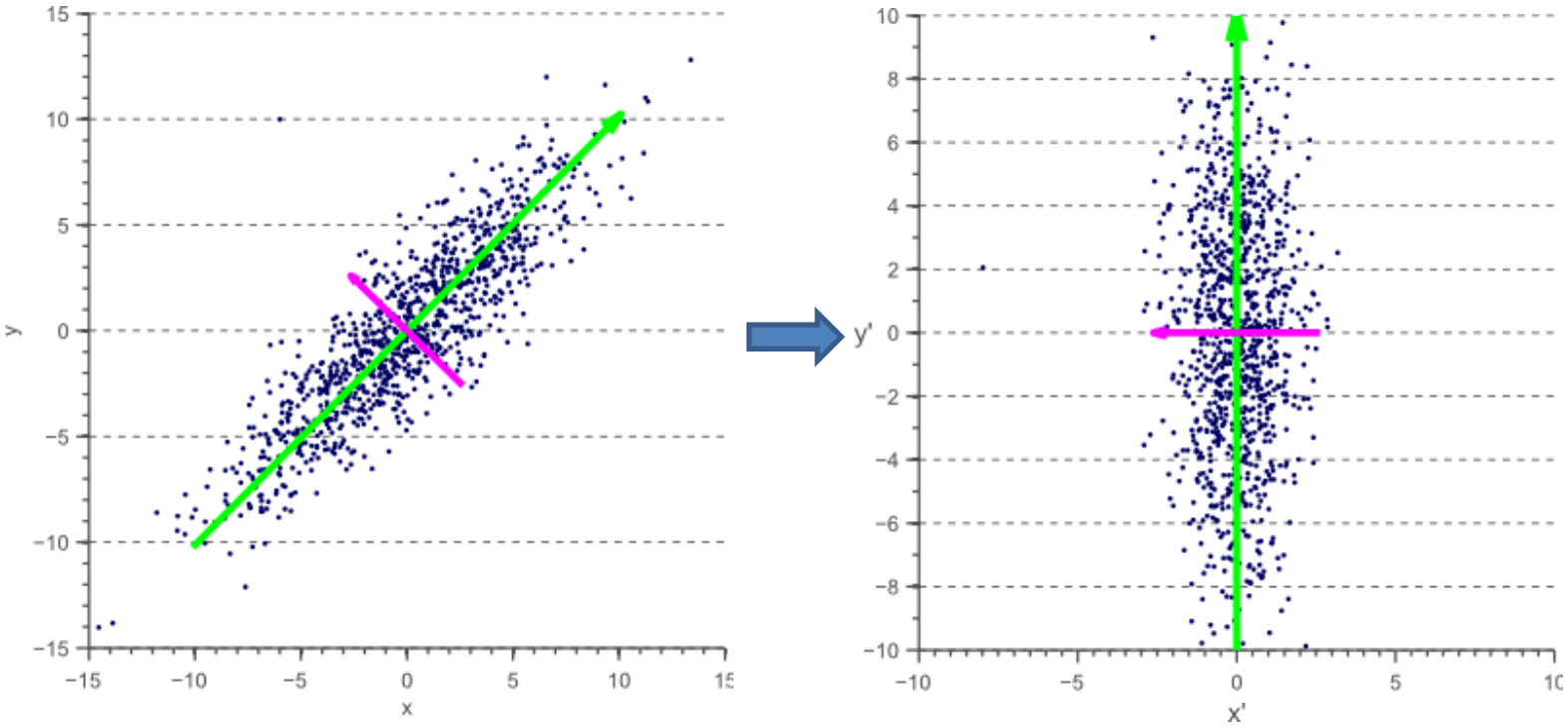
- If the covariance matrix is not diagonal, such that the covariances are not zero,
  - The eigenvalues still represent the variance magnitude in the direction of the largest spread of the data,
  - the variance components of the covariance matrix still represent the variance magnitude in the direction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





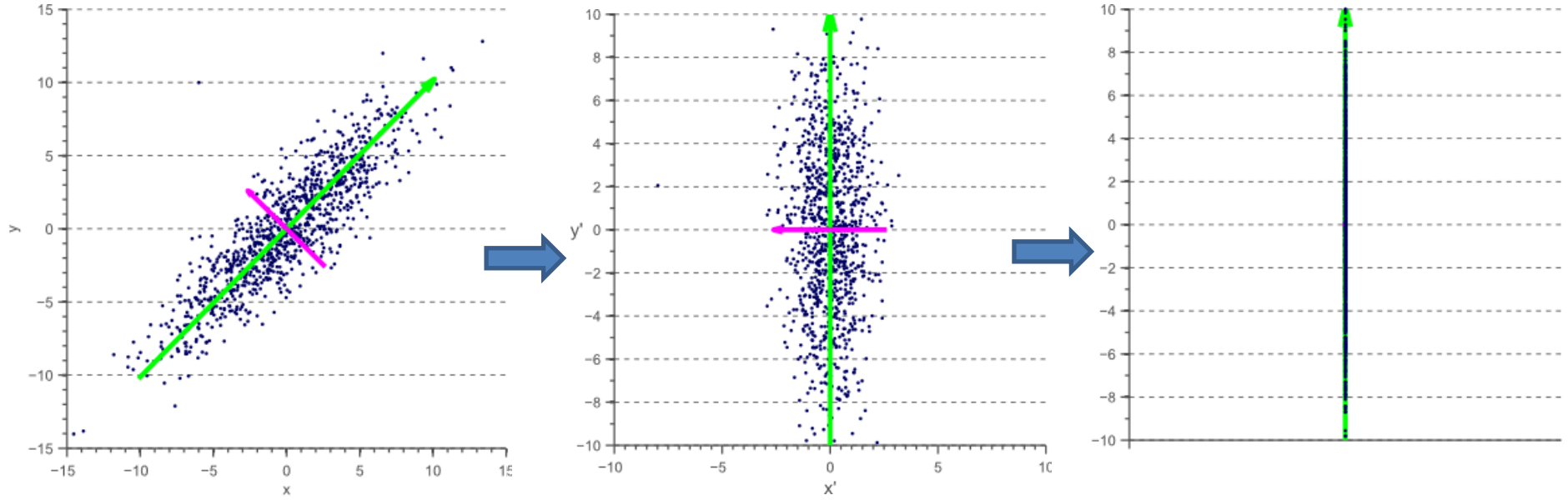
Principal Component Analysis

- 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



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



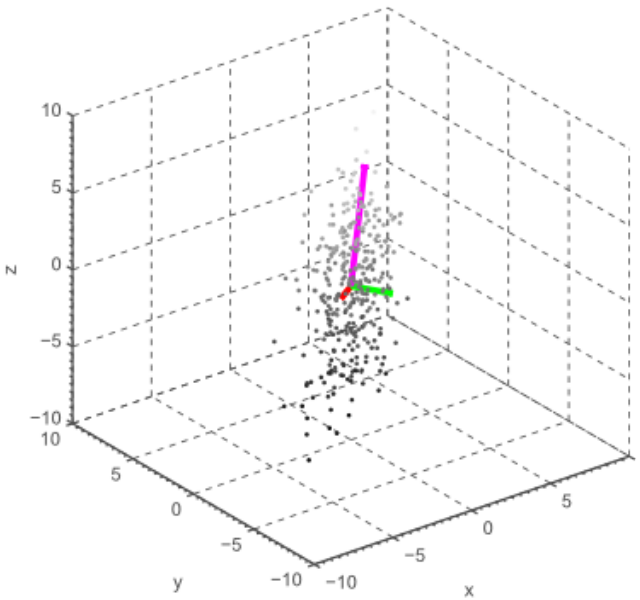
$$\Sigma = \begin{bmatrix} 16.87 & 14.94 \\ 14.94 & 17.27 \end{bmatrix}$$

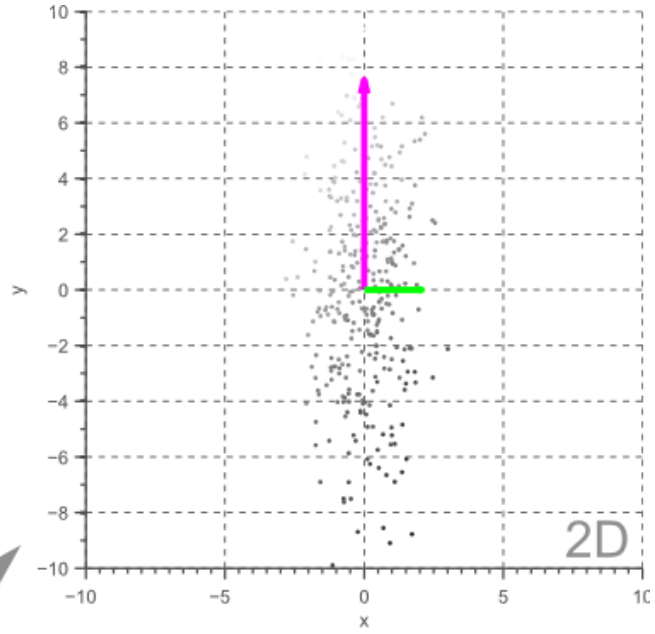
$$\Sigma' = \begin{bmatrix} 1.06 & 0.0 \\ 0.0 & 16.0 \end{bmatrix}$$

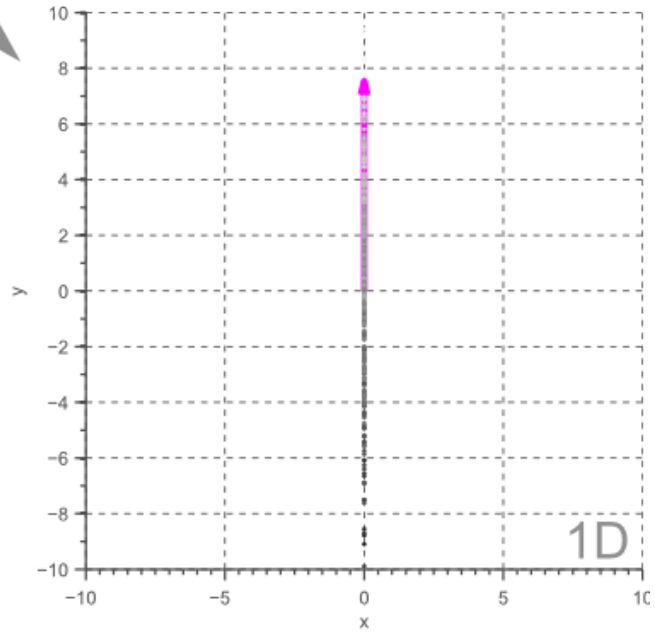
variance = 16.0

Principal Component Analysis

- Dimensionality reduction
- Eigenvectors correspond to principal components

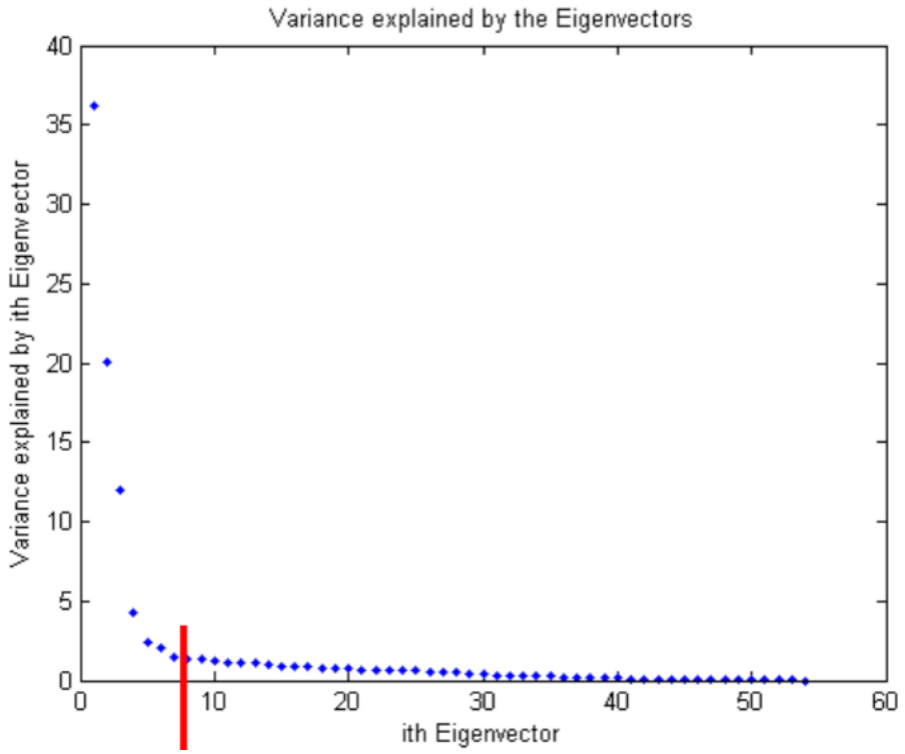






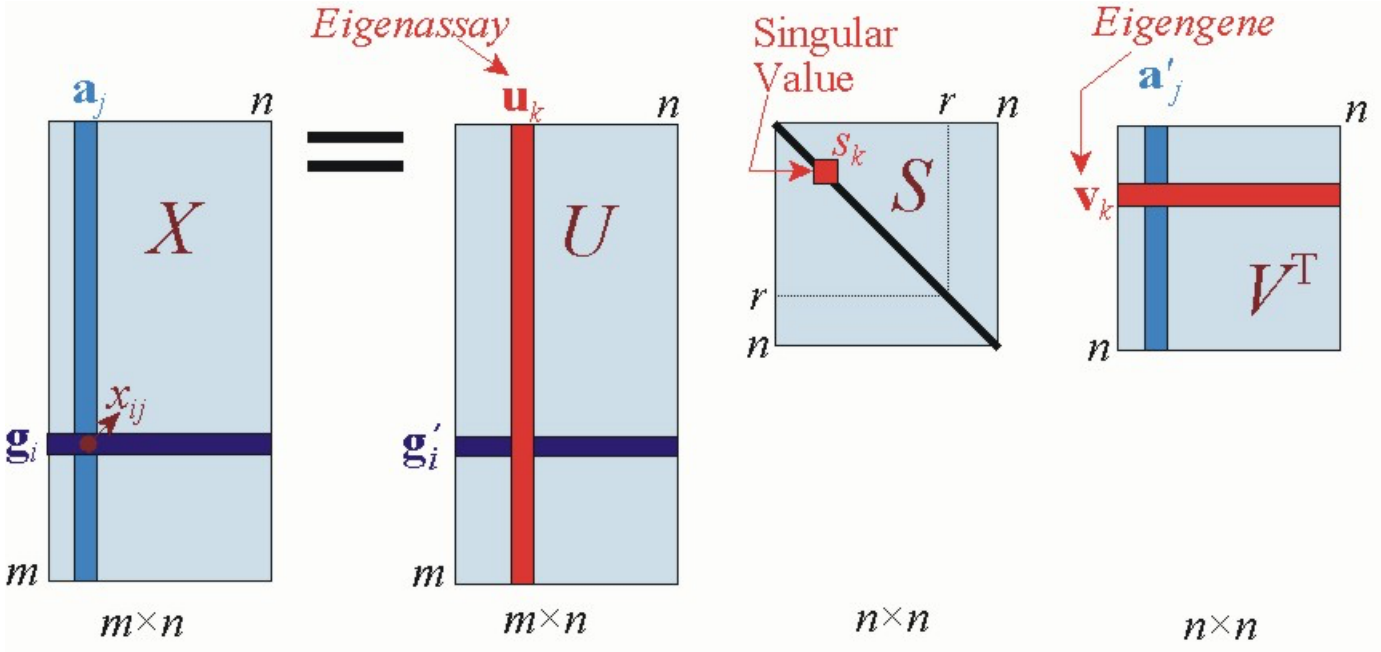
Principal Component Analysis

- 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



SVD and PCA

- 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

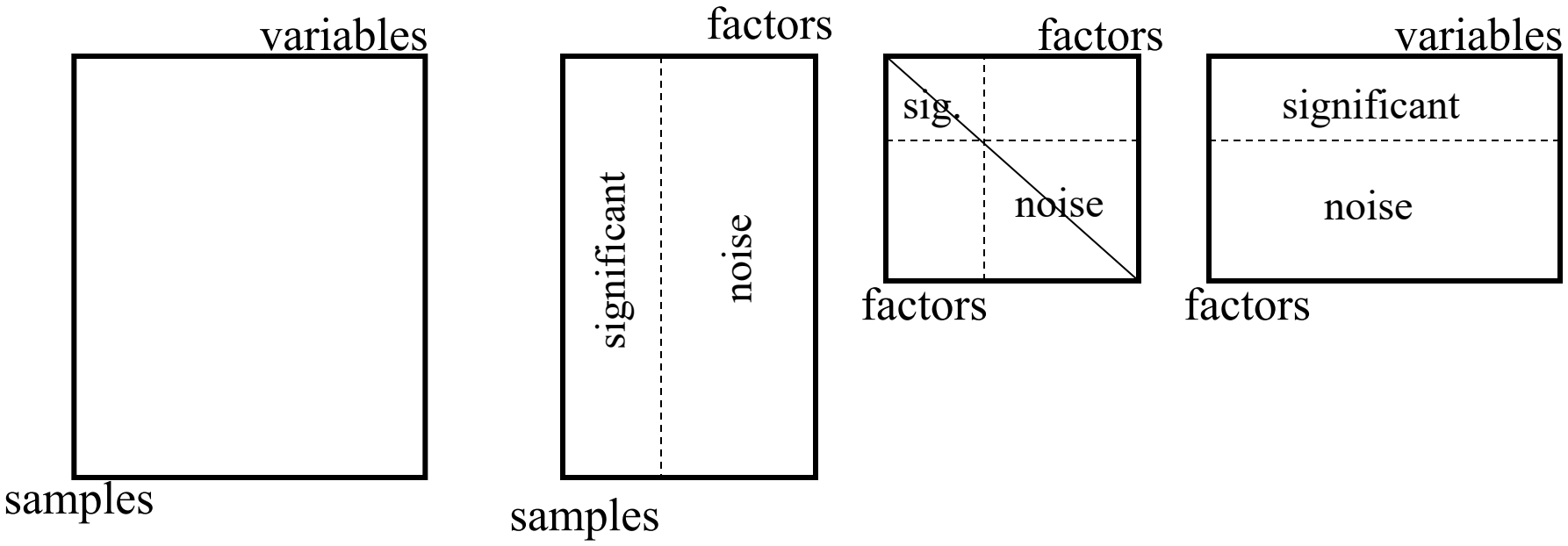
S is diagonal, $S_k > S_{k+1}$, S_k is k^{th} largest eigenvalue

Rows of V^T are unit length eigenvectors

SVD and PCA

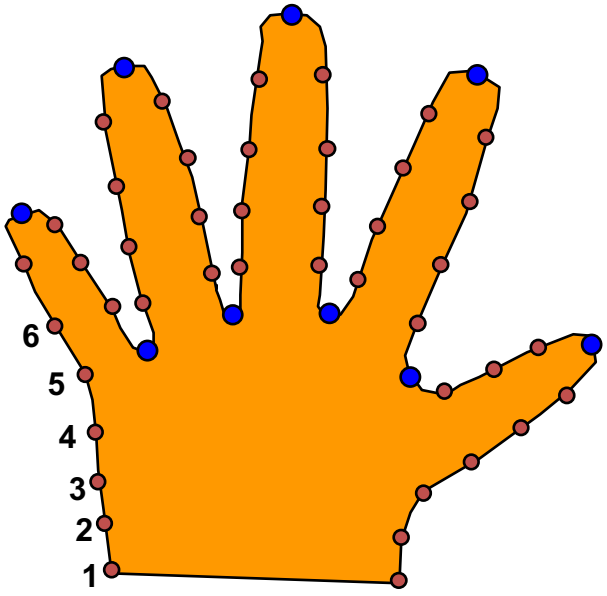
- PCA dimensionality reduction
  - Setting “noise” to zero to achieve reduced dimensionality

$$\mathbf{X} = \mathbf{U}\ \mathbf{S}\ \mathbf{V}^T$$



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



PCA Example

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

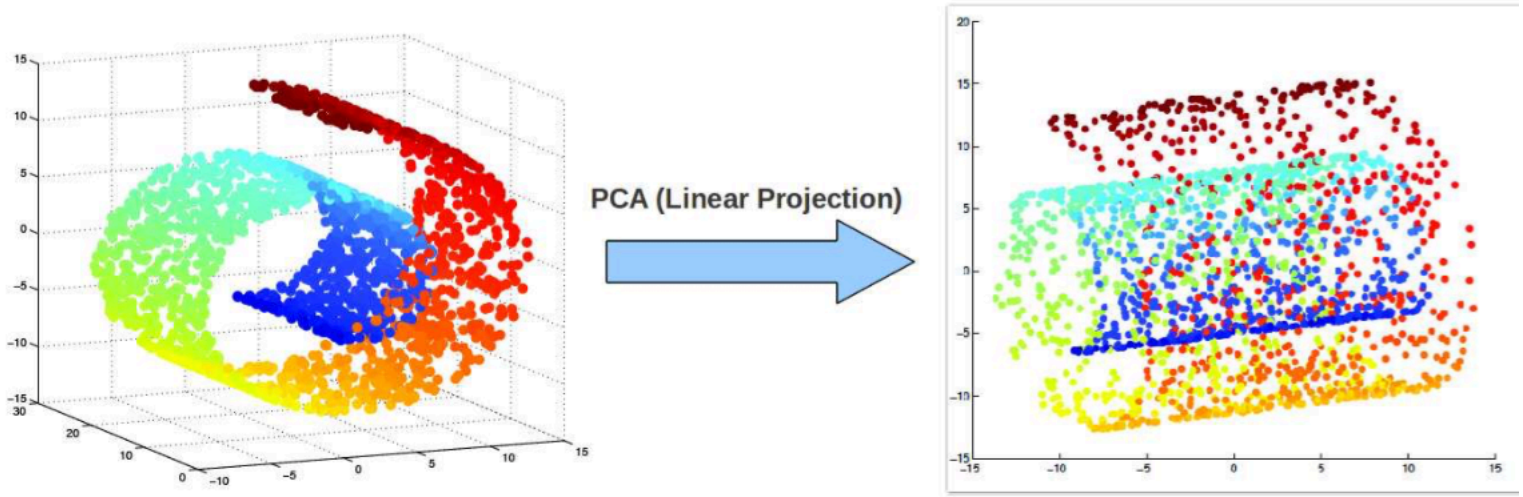
PC1

PC2

PC3

Principal Component Analysis

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



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?