CS 559 Machine Learning

Lecture 5: Principal Component Analysis

Ping Wang
Department of Computer Science
Stevens Institute of Technology

1870
STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Dimensionality Reduction

- Many dimensions are often interdependent (correlated).
  - Solution 1: reduce the dimensionality of problems;
  - Solution 2: transform interdependent coordinates into significant and independent ones.
- PCA transforms the original input space into a lower dimensional space, by constructing dimensions that are linear combinations of the given features.

Eigenvalues and Eigenvectors

- For an $n \times n$ square matrix A, e is an eigenvector with eigenvalue λ if:

$$Ae = \lambda e$$

or

$$(A - \lambda I)e = 0$$

- If $(A - \lambda I)$ is invertible, the only solution is $e = 0$ (trivial).

Eigenvalues and Eigenvectors

- For non-trivial solutions

$$\det(A - \lambda I) = 0$$

- Solutions are not unique because if e is an eigenvector ae is also an eigenvector.

Eigenvalues and Eigenvectors: Example

- For a 2×2 matrix:

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

- Given

$$\det(A - \lambda I) = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0$$

- We get:

$$a_{11}a_{22} - a_{12}a_{21} - (a_{11} + a_{22})\lambda + \lambda^2 = 0$$
$$1 \cdot 4 - 2 \cdot 2 - (1 + 4)\lambda + \lambda^2 = 0$$
$$\Rightarrow \lambda = 0\ and\ \lambda = 5$$

Eigenvalues and Eigenvectors: Example

- The eigenvector for the first eigenvalue $\lambda = 0$ is:

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x + 2y \\ 2x + 4y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- One solution for both equations is $x = 2;\ y = -1$.

- The eigenvector for the first eigenvalue $\lambda = 5$ is:

$$\begin{bmatrix} -4 & 2 \\ 2 & -1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -4x + 2y \\ 2x - y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- One solution for both equations is $x = 1;\ y = 2$.

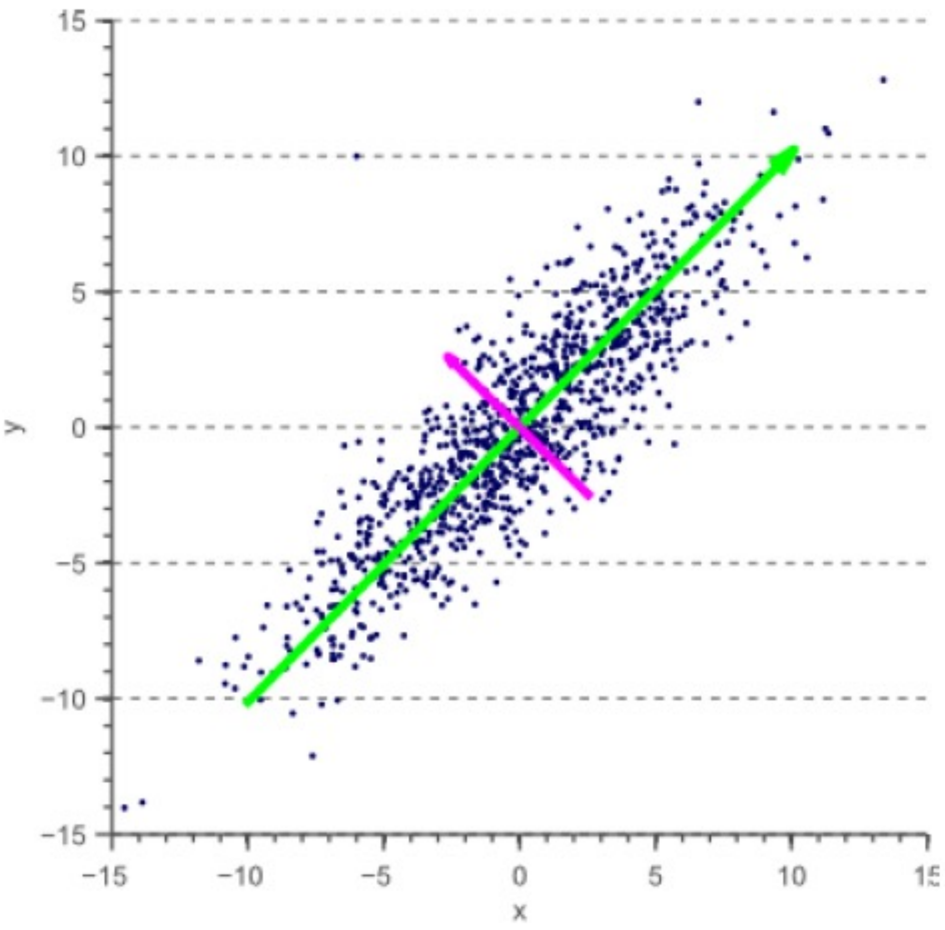
Eigenvalues and Eigenvectors: Properties

- The product of the eigenvalues is the determinant of A: det(A)
- The sum of the eigenvalues = trace(A)
- The eigenvectors are pairwise orthogonal

Principal Component Analysis

- PCA transforms the original input space into a lower dimensional space, by constructing dimensions that are linear combinations of the given features;
- The objective is to consider independent dimensions along which data have largest variance (i.e., greatest variability).
  - The first principal component accounts for as much of the variability in the data as possible;
  - Each succeeding component (orthogonal to the previous ones) accounts for as much of the remaining variability as possible.

Illustration of Principal Component Analysis



PCA rotates the data (centered at the origin) in a way that the maximum variability is visible (i.e., aligned with the axes.)

Green: first principal component of a two-dimensional dataset;
Pink: second principal component

Principal Component Analysis

- So: PCA finds n linearly transformed components, $s_1, s_2, \dots, s_n$, so that they explain the maximum amount of variance;
- We can define PCA in an intuitive way using a recursive formulation.

Principal Component Analysis

- Suppose data are first centered at the origin (i.e., their mean is 0)
- We define the direction of the first principal component, say, w_1 as follows:

$$w_1 = \arg \max_{\|w\|=1} E[(w^T x)^2]$$

  where w_1 is of the same dimension d as the input x.

- Thus: the first principal component is the projection on the direction along which the variance of the projection is maximized.

Principal Component Analysis

- Having determined the first $k-1$ principal components, the k^{th} principal component is determined as the principal component of the data residual:

$$w_k = \arg\max_{\|w\|=1} E\left[w^T(x - \sum_{i=1}^{k-1} w_i w_i^T x)\right]^2$$

- The transformed components are then given by:

$$s_i = w_i^T x$$

PCA: How to compute the principal components

- Let w be the direction of the first principal component, with $\|w\| = 1$.
- $s_i = w^T x_i$ is the projection of x_i along w, and we can consider it as the new coordinate in the 1D subspace of the first principal component.
- $\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i = \frac{1}{N}\sum_{i=1}^{N} w^T x_i$
- Variance of data along w:

$$\frac{1}{N}\sum_{i=1}^{N}(s_i - \bar{s})^2 = \frac{1}{N}\sum_{i=1}^{N}\left(w^T x_i - \frac{1}{N}\sum_{j=1}^{N} w^T x_j\right)^2$$

- The variance of data along direction w can be represented with the sample covariance matrix: $w^T \Sigma w$. How?

PCA: How to compute the principal components

$$\frac{1}{N}\sum_{i=1}^{N}(s_i-\bar{s})^2=\frac{1}{N}\sum_{i=1}^{N}\left(w^Tx_i-\frac{1}{N}\sum_{j=1}^{N}w^Tx_j\right)^2$$

$$=\frac{1}{N}\sum_{i=1}^{N}\left[w^T(x_i-\frac{1}{N}\sum_{j=1}^{N}x_j)\right]^2$$

$$=\frac{1}{N}\sum_{i=1}^{N}[w^T(x_i-\bar{x})]^2$$

$$=\frac{1}{N}\sum_{i=1}^{N}[w^T(x_i-\bar{x})(x_i-\bar{x})^Tw]$$

$$=w^T\boxed{\frac{1}{N}\left\{\sum_{i=1}^{N}[(x_i-\bar{x})(x_i-\bar{x})^T]\right\}}w=w^T\Sigma w$$

Sample covariance matrix

Unbiased estimation of the covariance:

$$var(X)=\frac{1}{N-1}\sum_{n=1}^{N}(X_n-\bar{X})^2$$

$$Cov(X,Y)=\frac{1}{N-1}\sum_{n=1}^{N}(X_n-\bar{X})(Y_n-\bar{Y})$$

PCA: How to compute the principal components

- Our objective is to find w such that $w = \arg\max_{w} w^T \Sigma w$ with constraint $w^T w = 1$.
- By introducing one Lagrange multiplier λ, we obtain the unconstrained optimization problem:

$$w = \arg\max_{w}[w^T \Sigma w - \lambda(w^T w - 1)]$$

- Take the derivative with respect to w and set to 0, we can get

$$2\Sigma w - 2\lambda w = 0$$

$$\Rightarrow \Sigma w = \lambda w$$

- Therefore, the problem can be reduced to an eigenvalue problem.

PCA: How to compute the principal components

The solution w is the eigenvector of Σ corresponding to the largest eigenvalue λ:

$$\Sigma w = \lambda w$$

The eigenvector associated with the largest eigenvalue corresponds to the first principal component; the eigenvector associated with the second largest eigenvalue corresponds to the second principal component; and so on...

PCA: In Practice

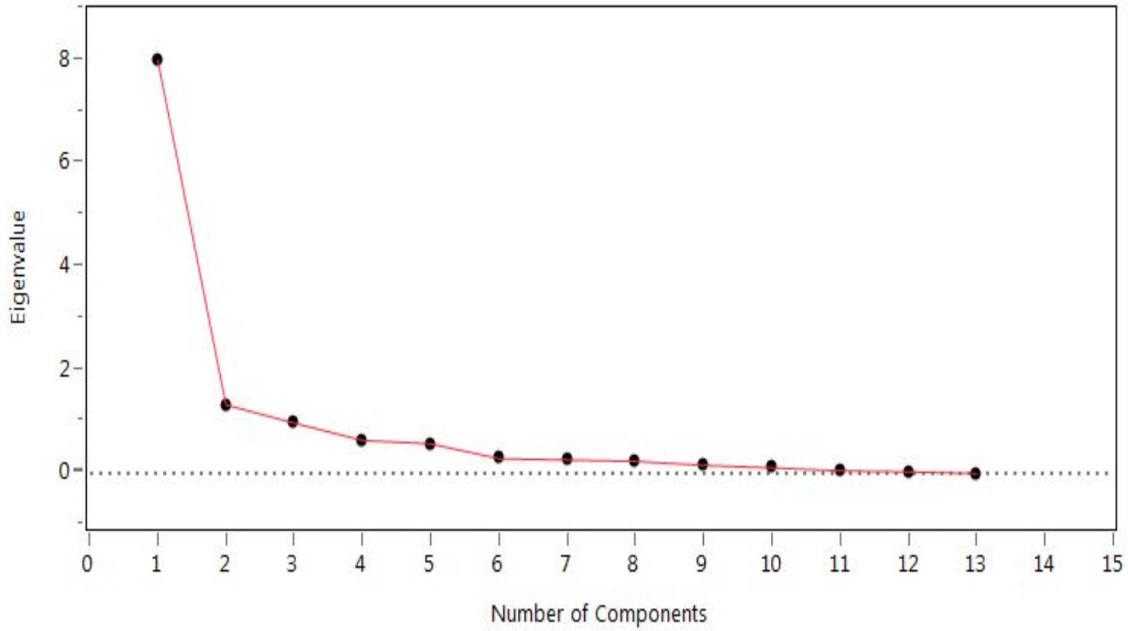
- The basic goal of PCA is to reduce the dimensionality of the data. Thus, one usually chooses:

$$n \ll d$$

- But how do we select the number of components?

Determining the number of components

- Plot the eigenvalues: each eigenvalue is related to the amount of variation explained by the corresponding axis (eigenvector);
- If the points on the graph tend to level out (show an "elbow" shape), these eigenvalues that are usually close enough to zero can be ignored;
- In general: Limit the variance accounted for.



Determining the number of components

- $x_i \in \mathbb{R}^d, i = 1, \ldots, N$
- $w_1, w_2, \ldots, w_d$: d eigenvectors (principal component directions)
- $\|w_i\| = 1$ (the w_is are orthogonal vectors)
- Representation of x_i in eigenvector space:

$$z_i = (w_1^T x_i)w_1 + (w_2^T x_i)w_2 + \ldots + \left(w_d^T x_i\right)w_d$$

- Suppose we retrain the first k principal components:

$$z_i^k = (w_1^T x_i)w_1 + (w_2^T x_i)w_2 + \ldots + \left(w_k^T x_i\right)w_k$$

- Then,

$$z_i - z_i^k = \left(w_{k+1}^T x_i\right)w_{k+1} + \cdots + \left(w_d^T x_i\right)w_d$$

Determining the number of components

$$(\mathrm{z}_i - \mathrm{z}_i^k)^T(\mathrm{z}_i - \mathrm{z}_i^k) =$$

$$\left[(\mathrm{w}_{k+1}^T \mathrm{x}_i)\mathrm{w}_{k+1} + \ldots + (\mathrm{w}_d^T \mathrm{x}_i)\mathrm{w}_d\right]^T \left[(\mathrm{w}_{k+1}^T \mathrm{x}_i)\mathrm{w}_{k+1} + \ldots + (\mathrm{w}_d^T \mathrm{x}_i)\mathrm{w}_d\right]$$

$$= \mathrm{w}_{k+1}^T(\mathrm{w}_{k+1}^T \mathrm{x}_i)^2\mathrm{w}_{k+1} + \ldots + \mathrm{w}_d^T(\mathrm{w}_d^T \mathrm{x}_i)^2\mathrm{w}_d$$

(note $\mathrm{w}_i^T \mathrm{w}_j = 0 \forall i \neq j$ since w_i and w_j are orthogonal vectors)

$$= (\mathrm{w}_{k+1}^T \mathrm{x}_i)^2\mathrm{w}_{k+1}^T\mathrm{w}_{k+1} + \ldots + (\mathrm{w}_d^T \mathrm{x}_i)^2\mathrm{w}_d^T\mathrm{w}_d$$

$$= (\mathrm{w}_{k+1}^T \mathrm{x}_i)^2 + \ldots + (\mathrm{w}_d^T \mathrm{x}_i)^2$$

$$= (\mathrm{w}_{k+1}^T \mathrm{x}_i)(\mathrm{x}_i^T \mathrm{w}_{k+1}) + \ldots + (\mathrm{w}_d^T \mathrm{x}_i)(\mathrm{x}_i)^T\mathrm{w}_d$$

$$= \mathrm{w}_{k+1}^T(\mathrm{x}_i\mathrm{x}_i^T)\mathrm{w}_{k+1} + \ldots + \mathrm{w}_d^T(\mathrm{x}_i\mathrm{x}_i^T)\mathrm{w}_d$$

Determining the number of components

Mean Square Error

$$\frac{1}{N}\sum_{i=1}^{N}(\mathrm{z}_i - \mathrm{z}_i^k)^T(\mathrm{z}_i - \mathrm{z}_i^k) =$$

$$\frac{1}{N}\sum_{i=1}^{N}[\mathrm{w}_{k+1}^T(\mathrm{x}_i\mathrm{x}_i^T)\mathrm{w}_{k+1} + \ldots + \mathrm{w}_d^T(\mathrm{x}_i\mathrm{x}_i^T)\mathrm{w}_d]$$

$$= \mathrm{w}_{k+1}^T[\frac{1}{N}\sum_{i=1}^{N}(\mathrm{x}_i\mathrm{x}_i^T)]\mathrm{w}_{k+1} + \ldots + \mathrm{w}_d^T[\frac{1}{N}\sum_{i=1}^{N}(\mathrm{x}_i\mathrm{x}_i^T)]\mathrm{w}_d$$

$$= \mathrm{w}_{k+1}^T\Sigma\mathrm{w}_{k+1} + \ldots + \mathrm{w}_d^T\Sigma\mathrm{w}_d$$

(Note: $\Sigma\mathrm{w}_{k+1} = \lambda_{k+1}\mathrm{w}_{k+1}, \ldots, \Sigma\mathrm{w}_d = \lambda_d\mathrm{w}_d$)

$$= \mathrm{w}_{k+1}^T\lambda_{k+1}\mathrm{w}_{k+1} + \mathrm{w}_d^T\lambda_d\mathrm{w}_d$$

$$= \lambda_{k+1} + \ldots + \lambda_d$$

Determining the number of components

$$\frac{1}{N}\sum_{i=1}^{N}\left(z_i - z_i^k\right)^T\left(z_i - z_i^k\right) = \lambda_{k+1} + \cdots + \lambda_d$$

- The mean square error of the truncated representation is equal to the sum of the remaining eigenvalues.
- In general: choose k so that 90-95% of the variance of the data is captured, which is defined as the cumulative explained variance or cumulative energy content $(\lambda_1 + \cdots + \lambda_k)/(\lambda_1 + \cdots + \lambda_d)$.

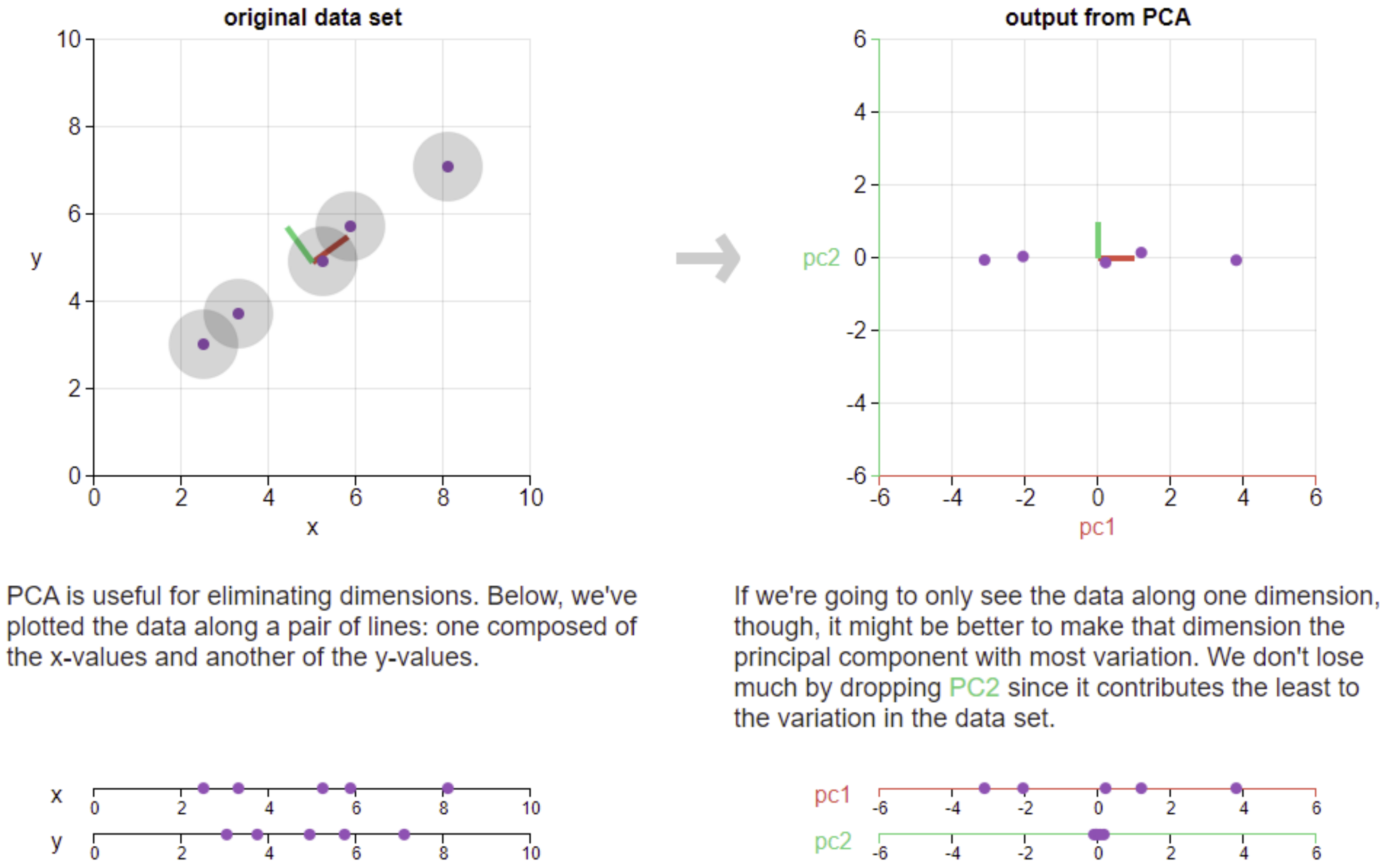
PCA: Steps

- Input: data matrix
- Output: a set of transformed coordinates
- Steps:
  1. Calculate the empirical mean, and get the covariance matrix
  2. Find the eigenvectors and eigenvalues of the covariance matrix
  3. Rearrange the eigenvectors and eigenvalues
  4. Compute the cumulative explained variance for each eigenvector
  5. Select a subset of the eigenvectors as basis vectors

PCA: Advantages

- Optimal linear dimensionality reduction technique in the mean-square sense;
- Reduce the curse-of-dimensionality;
- Computational overhead of subsequent processing stages is reduced;
- Noise may be reduced;
- A projection into a subspace of a very low dimensionality, e.g. two dimensions, is useful for visualizing the data.

PCA: Illustration Example 1



PCA is useful for eliminating dimensions. Below, we've plotted the data along a pair of lines: one composed of the x-values and another of the y-values.

If we're going to only see the data along one dimension, though, it might be better to make that dimension the principal component with most variation. We don't lose much by dropping PC2 since it contributes the least to the variation in the data set.

Illustration example with two dimensions. Source: https://setosa.io/ev/principal-component-analysis/

PCA: Illustration Example 2

Try it interactively through the link below to better understand PCA.

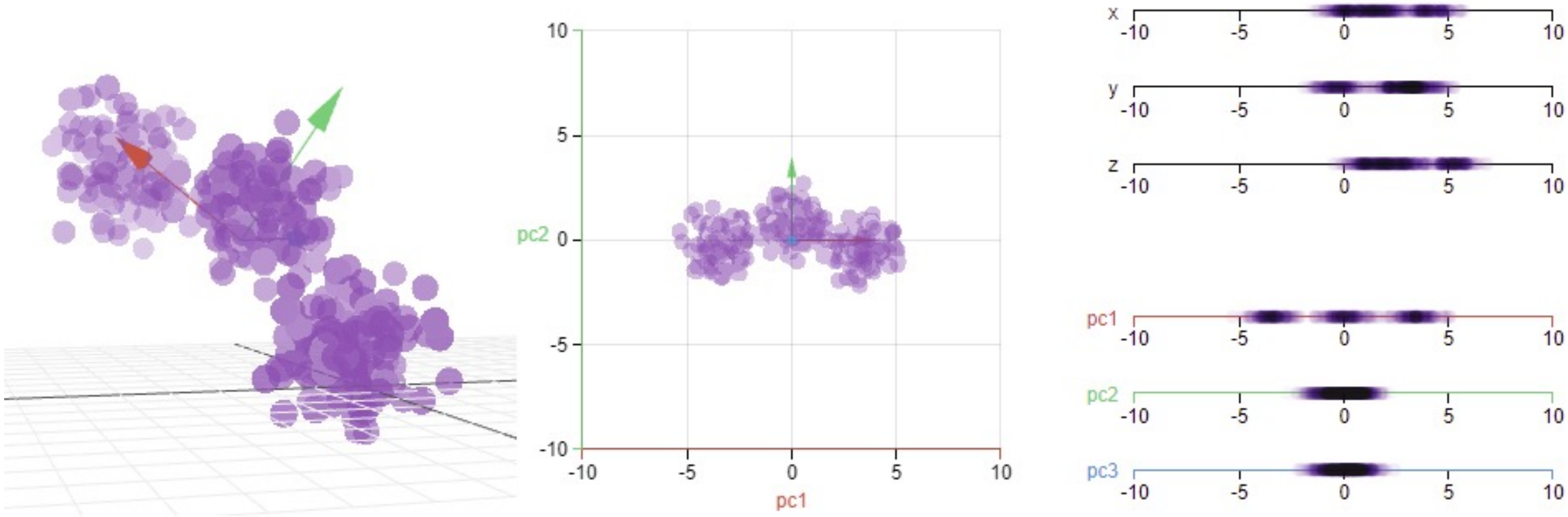


Illustration example with three dimensions. Source: https://setosa.io/ev/principal-component-analysis/

Summary of Today’s Lecture

- Principal Component Analysis
  - Motivation
  - How to compute
  - How to determine the number of components