Principal Components Analysis

Authors: Dian Zhao, Letian Zhao, Yuhan Yin, Yiwei Zhou.

Lecture Content

1. PCA and its principles
2. Linear Supervised Method
3. Non-Linear Embeddings for Visualization
4. PCA and t-SNE
5. Bonferroni's Theorem

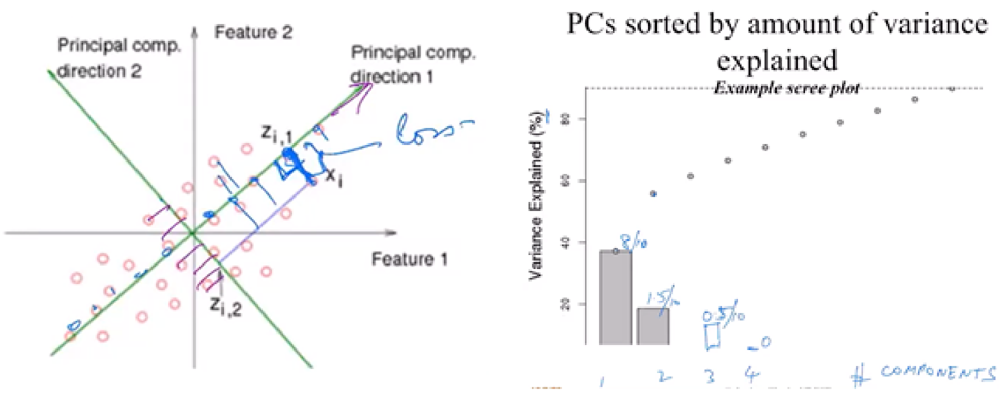
PCA and its principles:

- Principal Component Analysis (PCA) - pay attention to the spelling of "Principal"
- **PCA** is a standard technique for visualizing high dimensional data and for data pre-processing. PCA finds the best "subspace" based on eigen-decomposition of data covariance matrix to capture the most data variance possible while reducing the data dimensionality. This will also incur a loss that is represented by the perpendicular distance of each data point to the PCA directions during projection.

Sequential optimality of PCA:

- Every dimension PCA uses to capture the data variance is the optimal choice that minimizes the loss.

Scree Plot: visualization of how efficient the eigenvalues are in capturing the total variance:



Prof. Ghosh's class examples on PCA and the corresponding scree plot

Extensive reading on PCA and eigen algebra:

- The eigenvectors and eigenvalues of a covariance (or correlation) matrix is what we need to understand firstly before the PCA: Eigenvectors determine the directions of the new feature space and eigenvalues are the magnitude of data variance in the PCA dimensions.

- The classic approach to PCA is to perform the eigendecomposition on the covariance matrix, in which each element is the covariance between the two corresponding features. Recap the following equation to calculate the covariance:

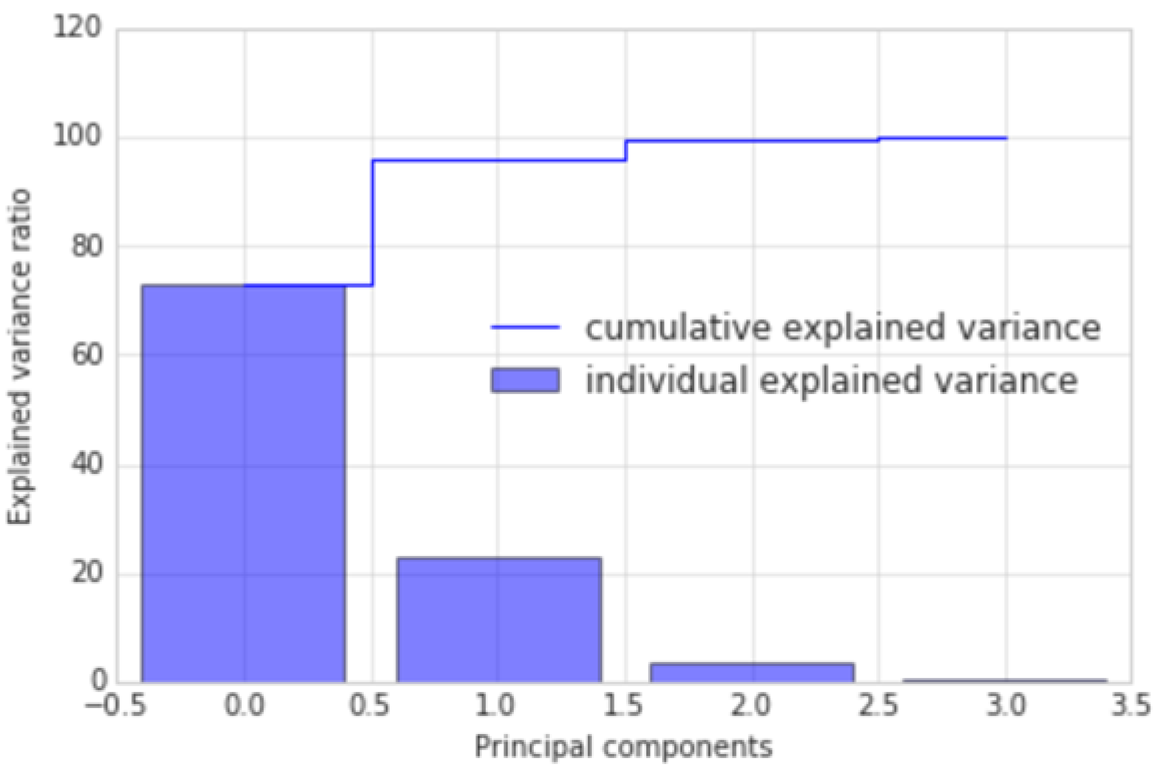
$$\sigma_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right).$$

- Covariance matrix:

$$\Sigma = \frac{1}{n-1}\left(\left(\mathbf{X} - \bar{\mathbf{x}}\right)^T \left(\mathbf{X} - \bar{\mathbf{x}}\right)\right)$$

where $\bar{\mathbf{x}}$ is the mean vector $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} x_i.$

- After getting the eigenvectors and eigenvalues of the above matrix, we will inspect and drop the eigenvectors with the lowest eigenvalues because they incorporate the **least** information regarding data variance.
- After sorting the eigenpairs, we should determine the number of principal components to be chosen for the new feature subspace. A useful measure is the **explained variance**, which can be calculated from the eigenvalues. It tells us how much information (variance) can be attributed to each of the principal components and is similar to the scree plot. We expect the first few eigenvalues capture about **80%-90%** explained variance.

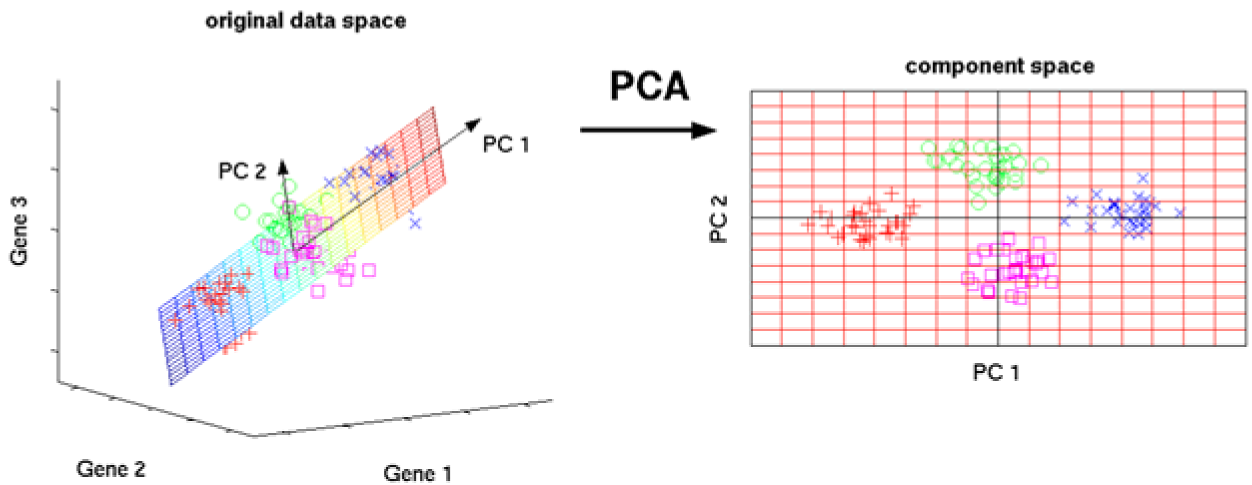


Scree plot with the explained variance
Sample codes can be found through the website [1] in the reference list

- In our class example, we see a drastic decrease in eigenvalues as more principal components are selected. The first two principal component dimensions have dominated the analysis and captured 95% of the total variance. Intuitively, the number of dimensions we choose is much less needed than the total variance in percentage they could capture.

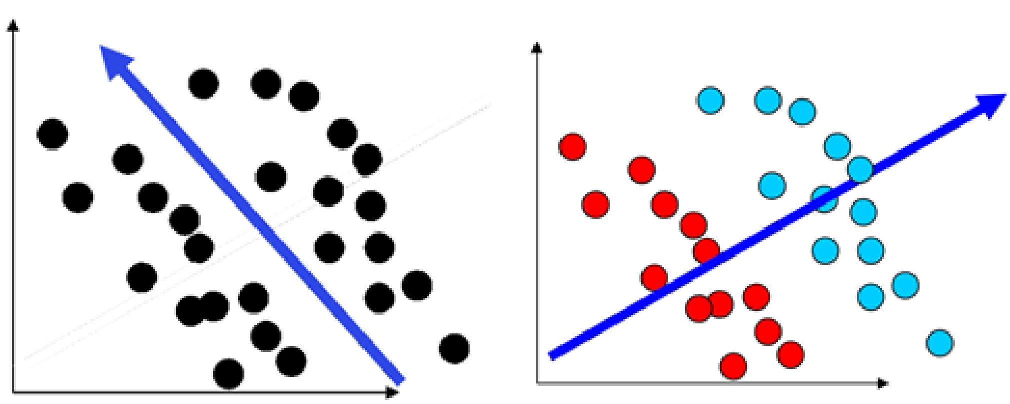
Reducing dimensionality

Naturally occurring data may have high dimensionality while its subspace is much lower. PCA is used to visualize these data by reducing the dimensionality. It rotates the original data space such that the axes of the new coordinate system point into the directions of *highest variance* of the data. We can identify the two-dimensional plane that optimally describes the highest variance of the three-dimensional data below.



Linear Supervised Method

- PCA is unsupervised learning and is not the best solution for classification.
- Ignoring colors, the PCA direction should be like the left picture. But when we consider color classification, the direction should be like the right picture. Red points are projected to a small range and blue points are also projected to a small range on the arrow.
- A special Linear Supervised Method is Fisher's Linear Discriminant (FLD) which finds the projection direction that best separates the two classes.
- Multiple discriminant analysis (MDA) extends LDA to multiple classes.

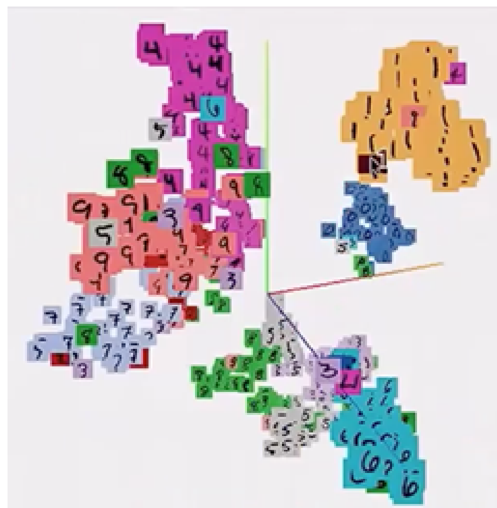


Non-Linear Embeddings for Visualization

- Manifold is a topological space with the property that each point has a neighborhood that is homeomorphic to the Euclidean space of dimension n. It captures the intrinsic dimensionality of data in a nonlinear fashion.
- The earth is a three dimensional space, but its intrinsic dimensionality is two. In other words, a two dimensional manifold embedded in three dimension space.
- A coil is intrinsically one dimension, but it's embedded in three dimension space.
- Application: capture handwritten digits using a two dimensional manifold.

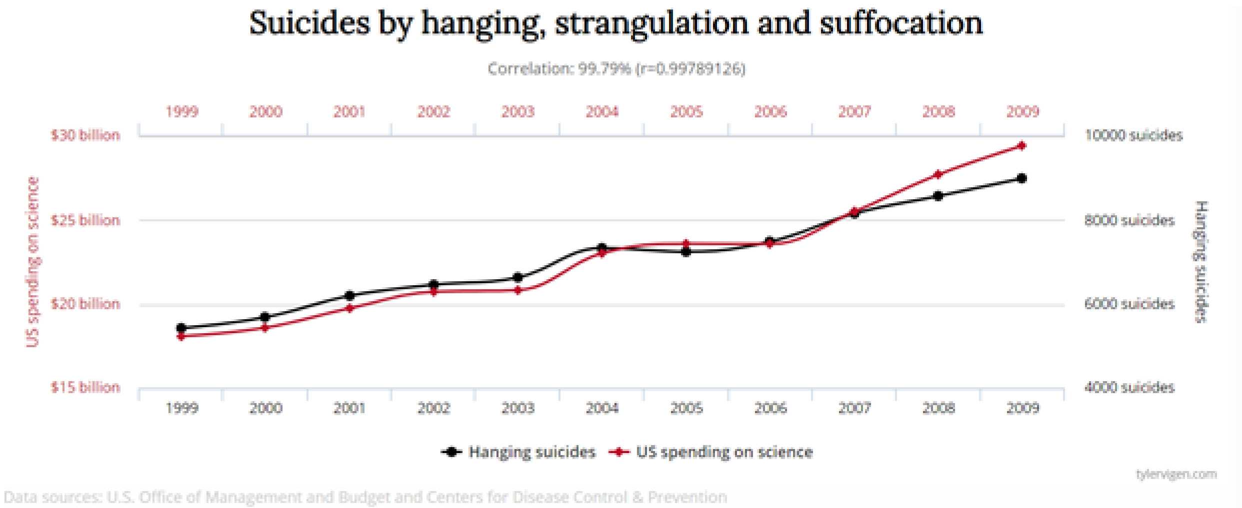
PCA and t-SNE

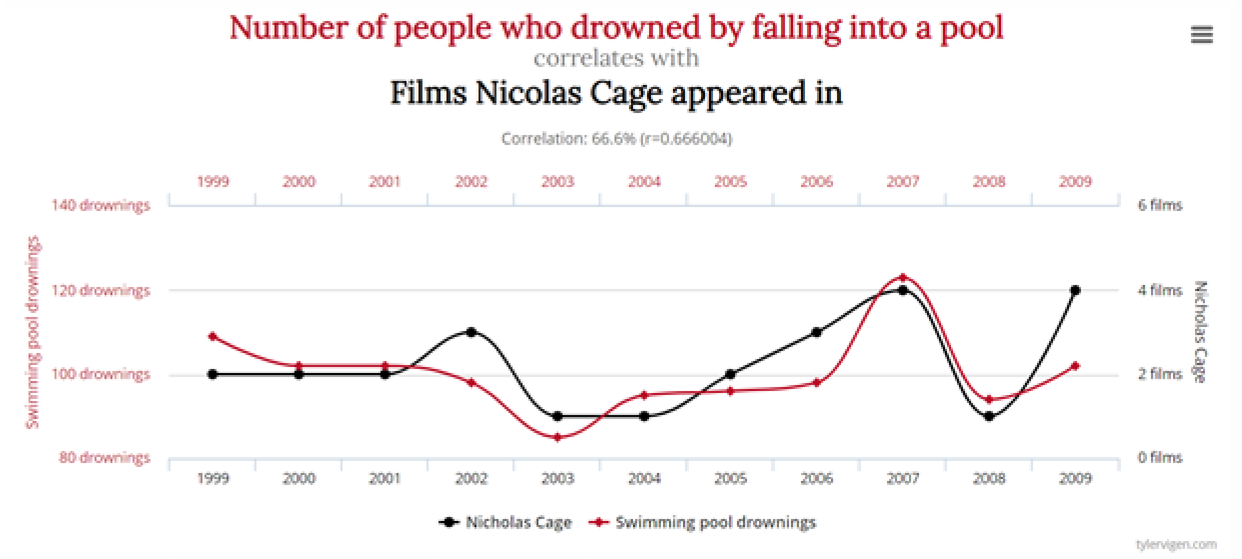
- Project 784 (28*28 images) dimensions to 3 dimensions
- Left picture is PCA, right one is t-distributed stochastic neighbor embedding (t-SNE)



Bonferroni’s Theorem:

If there are too many possible conclusions to draw, some will be true for purely statistical reasons, with no physical validity -> correlation does not suggest causation。





www.tylervigen.com

https://www.youtube.com/watch?v=jbkSRLYSojo

Hans Rosling Ted Talk

- The best stats you've ever seen | Hans Rosling: https://www.youtube.com/watch?v=hVimVzgtD6w (https://www.youtube.com/watch?v=hVimVzgtD6w)
- How not to be ignorant about the world | Hans and Ola Rosling: https://www.youtube.com/watch?v=Sm5xF-UYgdg (https://www.youtube.com/watch?v=Sm5xF-UYgdg)
- Hans Rosling: Global population growth, box by box: https://www.youtube.com/watch?v=fTznEIZRkLg (https://www.youtube.com/watch?v=fTznEIZRkLg)
- Religions and babies | Hans Rosling: https://www.youtube.com/watch?v=ezVk1ahRF78 (https://www.youtube.com/watch?v=ezVk1ahRF78)

Reference and further readings:

1. https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html#1---eigendecomposition---computing-eigenvectors-and-eigenvalues (https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html#1---eigendecomposition---computing-eigenvectors-and-eigenvalues)
2. https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues (https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues)