Principal component analysis

- Principal component: diections in feature space along which the original data are highly variable. In other words, these directions explain a lot about the data.
- These directions define lines and subspaces that are close to the clouds of data.
- Principal component analysis (PCA) refers to the process by which principal components are computerd.
- PCA provides a better method to visualize the n observations when p is large.
- Considered a unsupervised learning technique although it is widely used in both supervised and unsupervised frameworks.

- The question asked by PCA: Is there another basis which is a linear combination of the original basis that best represent our data set?
- First principal component of a set of features $x_1, x_2, \cdots, x_p$ (columns of design matrix): Z_1

$$Z_1 = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p$$

- In matrix form the relationship can be written as:

$$X\alpha = Z$$

- α is a matrix $\in \mathcal{R}^{p \times k}$ that transform X into Y.
- Geometrically α rotate and stretch X into Y.
- The columns of α denotes by: $\alpha_i \in \mathcal{R}^p, i = 1, \cdots, k$ are a set of new basis vectors for expressing rows of X.

- Matrix Y is the "new" matrix data $\in \mathcal{R}^{n \times k}$.
- Each row of Y is the representation of the original row of X on new basis α.
- Each entry in a row of Y is the result of the corresponding row of X with a column of α.
- How do we find a good basis?
- What is the best way to reexpress X?

- Noise: is in any data set. If there is too much noise, no technique can work.
- A common measure for noise: signal to noise ratio (SNR) or ratio of variance:

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

- $SNR \gg 1$ indicates high precision data, a low SNR is strongly contaminated data.
- Goal: Find the new basis along which the SNR is highest (the variance of signal is highest along the basis).

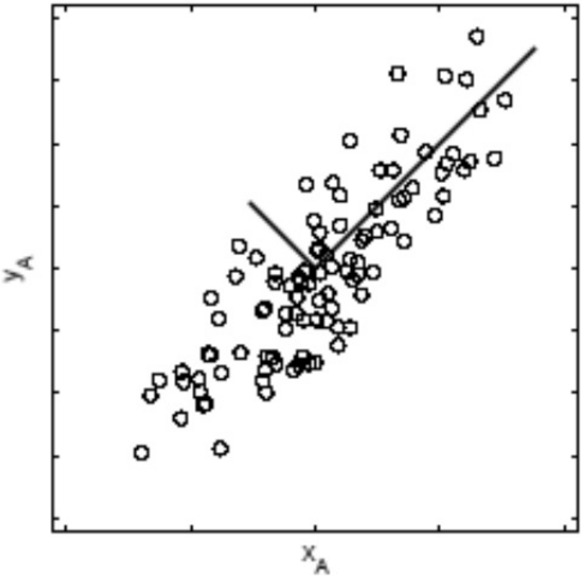

Figure: http://dai.fmph.uniba.sk/courses/ml/sl/PCA.pdf

- Redundancy: Multiple attributes could measure similar dynamic information.
- If two attributes are un-correlated , the two recordings have no redundancy. (i.e a student height and GPA).
- Two attributes could be strongly related : one can be used to express the other.
- Covariance matrix: $S_X = \frac{1}{p-1} X^T X$.
- The ijth element of S_X is the dot product between the column vector of ith measurement type (attribute i(and the column vector of j th attribute.

- Diagonal elements of a covariance matrix measures the variance of a particular attributes.
- Off diagonal eleemnts of a covariance matrix are the covariance betwen the two attributes.
- Goal: Our transformed data has low redundancy measured by covariance (off diagonals), and high SNR, measured by variance (diagonal).
- PCA: new basis vectors are orthonormal.
- Directions with largest variance are the most important or principal components.

- PCA selects a normalized direction in p-dimensional space along which the variance in X is maximized: Z_1
- Find another direction along which variance is maximized and perpendicular to Z_1: Z_2
- Continue until k directions are selected. The resulting set of basis are called principal components.
- The variance associated with each components quantify how principal they are.
- Solved using linear algebra.

- $X\alpha = Y$. The goal is S_Y is "diagonalized".

$$S_y = Y^T Y = (X\alpha)^T (X\alpha) = \alpha^T X^T X\alpha.$$

- Matrix S_X is symmetric and it can be diagonalized by an orthogornal matrix of its eigenvectors.
- $A = EDE^T$ where D is a diagonal matrix of eigen values, E is matrix of eigen vectors.
- Choose basis α to be the eigen vectors of S_X then S_Y will be diagonalized as D, the diagonal matrix of eigenvalues of S_X.

- PCA Algorithm
  1. Normalize the attributes.
  2. Construct covariance matrix.
  3. Find eigenvalues and eigenvectors of covariance matrix.
  4. Select k largest eigenvalues and corresponding eigenvectors as new basis.
  5. Use the new basis to transform data X.

- Design matrix X, each column corresponds to an attribute. Normalize as:

$$\bar{x^j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}, s_j = \sqrt{\frac{\sum_{i=1}^{m}\langle x^j, x^j\rangle}{n-1}}$$

$$x^j = \frac{x^j - \bar{x}^j}{s_j}$$

- Construct the covariance matrix: $\Sigma = \frac{1}{n-1}X^T X$.
- Compute the eigen values λ and eigen vectors of Σ.

- Eigenvector with higest eigenvalue is the principal component of the data set.
- Choose only the biggest k eigenvalues and eigenvectors.
- The proportion of variance explained by k principal components are: $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$
- New basis α are the set of eigen vectors.
- $Newdata = X * \alpha$

- Matlab example.