Machine Learning for Data Science (CS4786)
Lecture 2

Dimensionality Reduction
&
Principal Component Analysis

Course Webpage :

http://www.cs.cornell.edu/Courses/cs4786/2015sp/

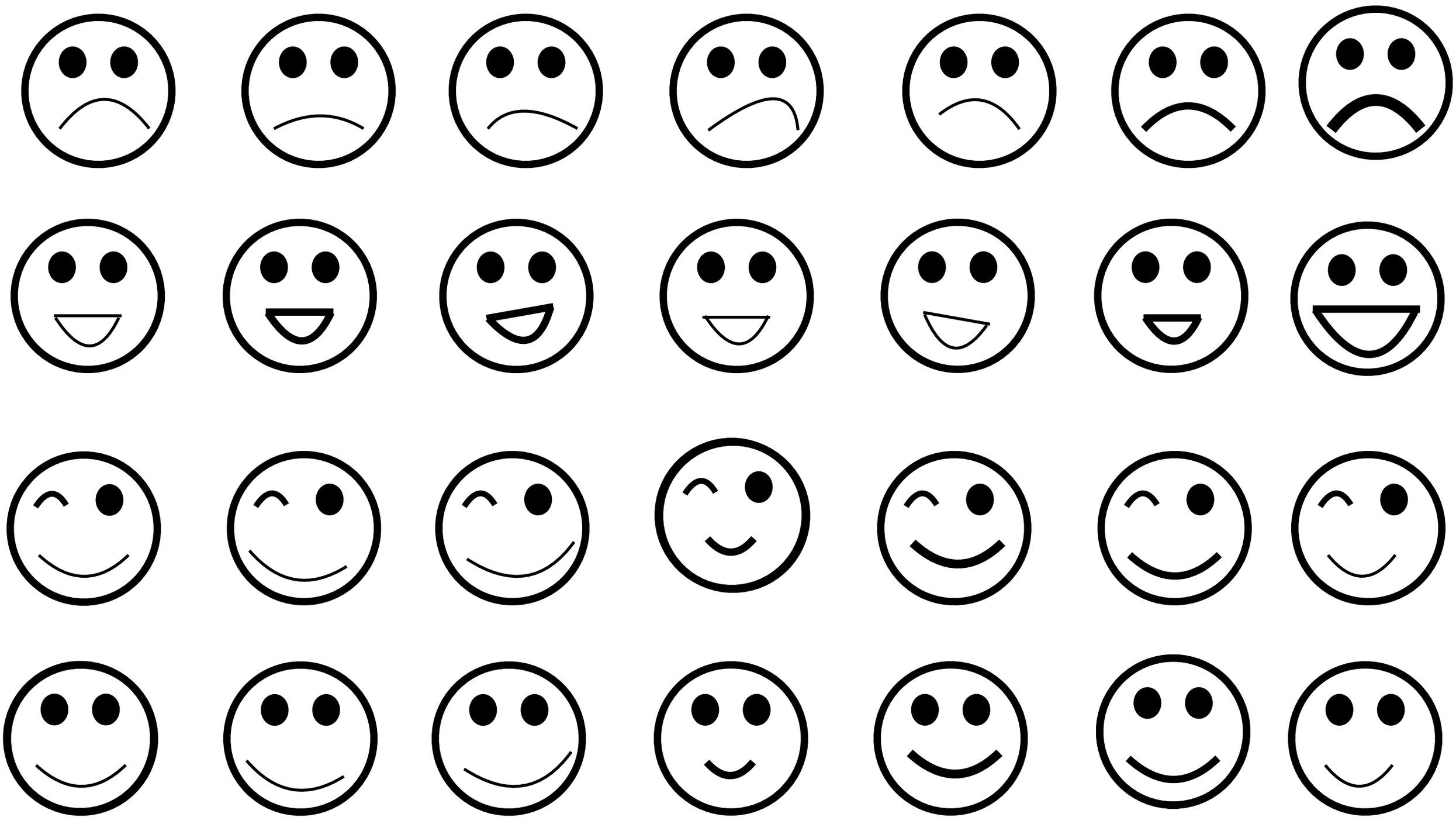
Announcements

- Diagnostic assignment due on 29th (Thursday) beginning of class
- Course webpage is the official source of all class related information
- Please make sure to add both Prof. Lee and Prof. Sridharan on all emails

Representing Data as Feature Vectors

- How do we represent data?
- Each data-point often represented as vector referred to as feature vector
- Eg. text document represented by vector in which each coordinate represents a word and value represents number of times the word occurred in the document
- Eg. Image represented as a vector where each coordinate represents a pixel and value represents the grayscale value of that pixel

Example: Images



Dimensionality Reduction

- You are provided with n data points each in $\mathbb{R}^d$
- Goal: Compress data into n, points in $\mathbb{R}^K$ where $K << d$
  - Retain as much information about the original data set
  - Retain desired properties of the original data set

Why dimensionality reduction?

- For computational ease
    - As input to supervised learning algorithm
        - Before clustering to remove redundant information and noise
- Data compression & Noise reduction
- Data visualization

Dimensionality Reduction

Given feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, compress the data points into low dimensional representation $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^K$ where $K << d$

Dimensionality Reduction

Desired properties:

1. Original data can be (approximately) reconstructed
2. Preserve distances between data points
3. "Relevant" information is preserved
4. Noise is reduced

Dim Reduction: Linear Transformation

- Pick a low dimensional subspace
- Project linearly to this subspace
- Subspace retains as much information

Dim Reduction: Linear Transformation

$$X = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1.1 & 2 & 3 & 4 \\ 3 & 2 & 3 & 4 \\ -1 & 2 & 3 & 4 \\ -0.2 & 2 & 3 & 4 \\ -2 & 2 & 3 & 4 \\ 1.4 & 2 & 3 & 4 \\ 1.4 & 2 & 3 & 4 \\ -0.1 & 2 & 3 & 4 \\ 0.5 & 2 & 3 & 4 \end{bmatrix}$$

Dim Reduction: Linear Transformation

$$X = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1.1 & 2 & 3 & 4 \\ 3 & 2 & 3 & 4 \\ -1 & 2 & 3 & 4 \\ -0.2 & 2 & 3 & 4 \\ -2 & 2 & 3 & 4 \\ 1.4 & 2 & 3 & 4 \\ 1.4 & 2 & 3 & 4 \\ -0.1 & 2 & 3 & 4 \\ 0.5 & 2 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1.1 \\ 3 \\ -1 \\ -0.2 \\ -2 \\ 1.4 \\ 1.4 \\ -0.1 \\ 0.5 \end{bmatrix} \times [1, 0, 0, 0] + \begin{bmatrix} 0 & 2 & 3 & 4 \\ 0 & 2 & 3 & 4 \\ 0 & 2 & 3 & 4 \\ 0 & 2 & 3 & 4 \\ 0 & 2 & 3 & 4 \\ 0 & 2 & 3 & 4 \\ 0 & 2 & 3 & 4 \\ 0 & 2 & 3 & 4 \\ 0 & 2 & 3 & 4 \\ 0 & 2 & 3 & 4 \end{bmatrix}$$

Orthonormal Projections

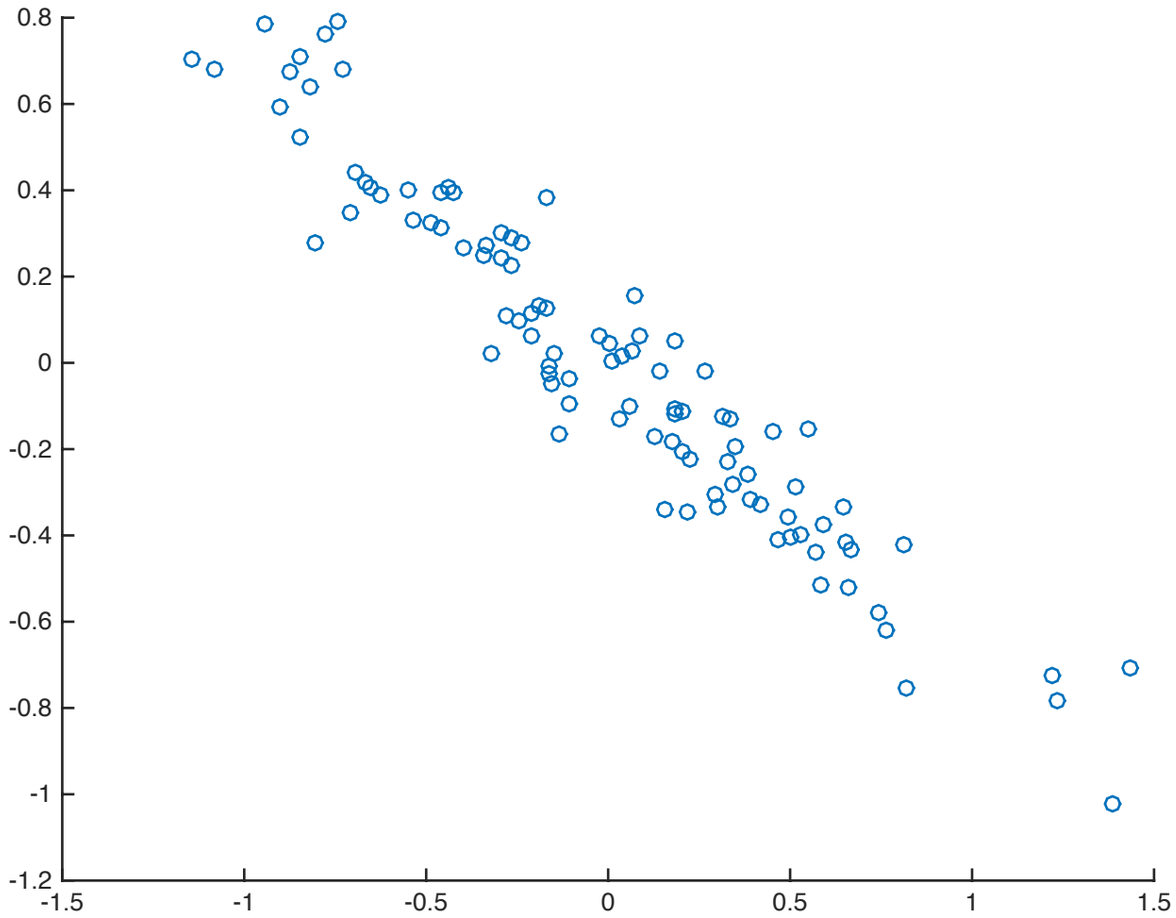
- (Centered) Data-points as linear combination of some orthonormal basis, i.e.

$$\mathbf{x}_t = \mu + \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j$$

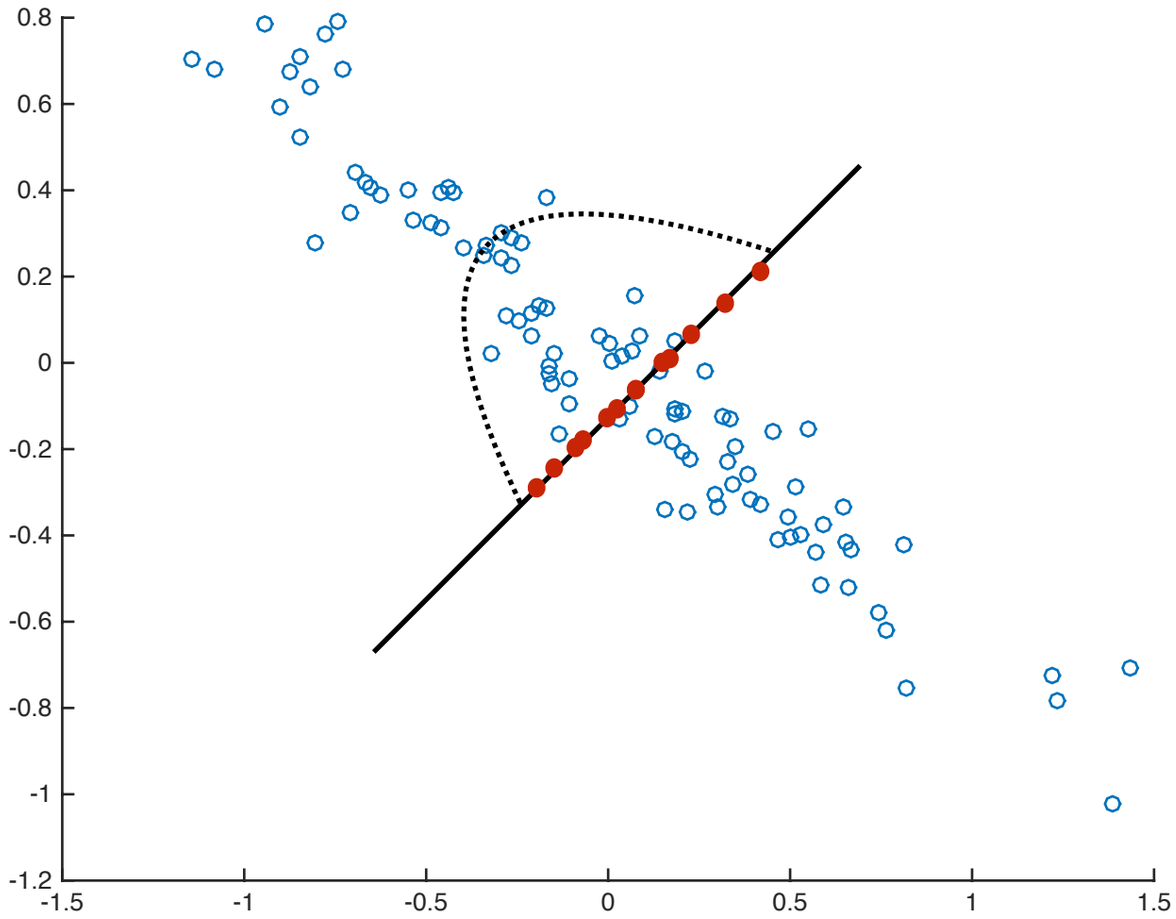
  where $\mathbf{w}_1, \ldots, \mathbf{w}_d \in \mathbb{R}^d$ are the orthonormal basis and $\mu = \frac{1}{n}\sum_{t=1}^{n} x_t$.
- Represent data as linear combination of just K orthonormal basis,

$$\hat{\mathbf{x}}_t = \mu + \sum_{j=1}^{K} \mathbf{y}_t[j]\mathbf{w}_j$$

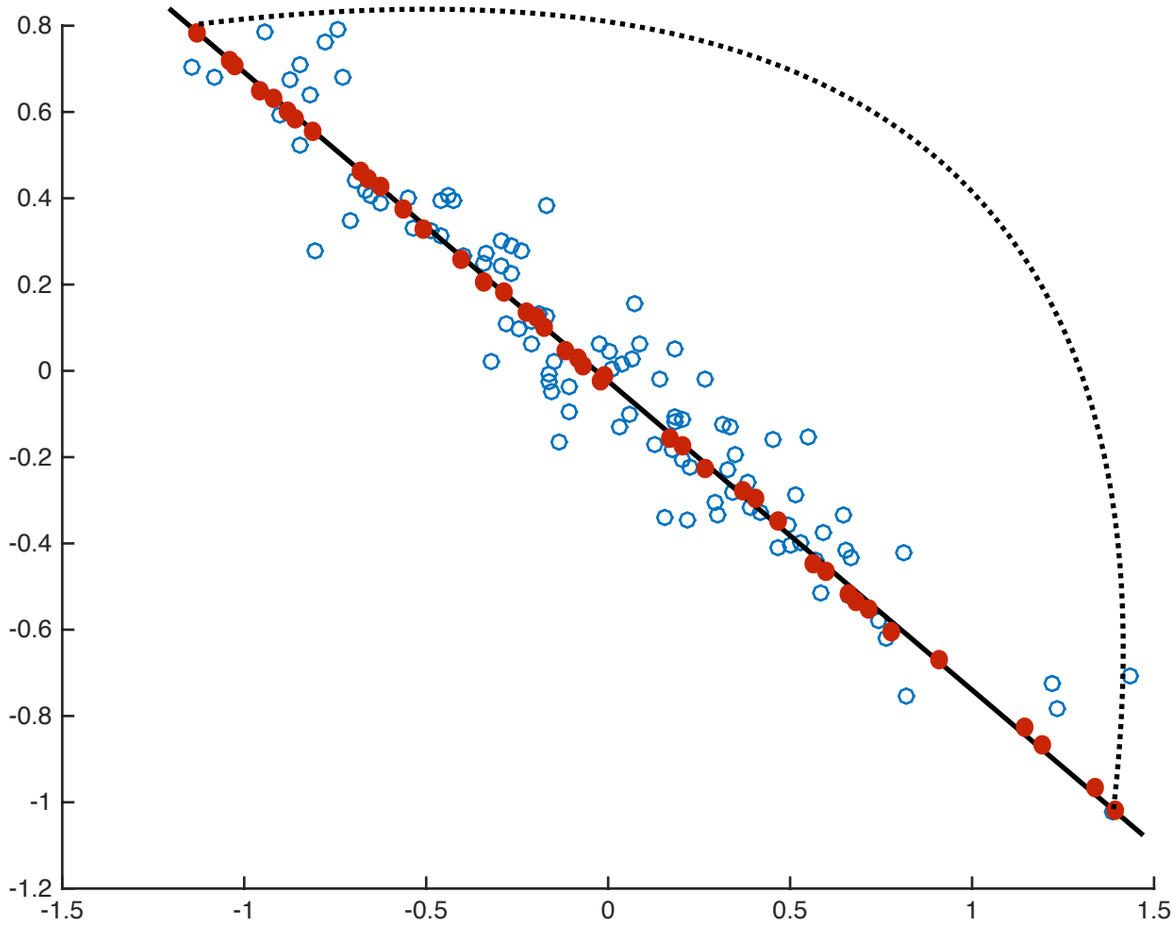
PCA: Variance Maximization



PCA: Variance Maximization



PCA: Variance Maximization



PCA: Variance Maximization

- Pick directions along which data varies the most

PCA: Variance Maximization

- Pick directions along which data varies the most
- First principal component:

$$\mathbf{w}_1 = \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n} \sum_{t=1}^{n} \mathbf{w}^\top \mathbf{x}_t \right)^2$$

PCA: Variance Maximization

- Pick directions along which data varies the most
- First principal component:

$$\mathbf{w}_1 = \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n} \sum_{t=1}^{n} \mathbf{w}^\top \mathbf{x}_t \right)^2$$

$$= \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}^\top (\mathbf{x}_t - \mu) \right)^2$$

PCA: Variance Maximization

- Pick directions along which data varies the most
- First principal component:

$$\mathbf{w}_1 = \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top \mathbf{x}_t\right)^2$$

$$= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top (\mathbf{x}_t - \mu)\right)^2$$

$$= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top (\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top \mathbf{w}$$

PCA: Variance Maximization

- Pick directions along which data varies the most
- First principal component:

$$\begin{aligned}
\mathbf{w}_1 &= \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top \mathbf{x}_t\right)^2 \\
&= \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top (\mathbf{x}_t - \mu)\right)^2 \\
&= \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top (\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top \mathbf{w} \\
&= \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \Sigma \mathbf{w}
\end{aligned}$$

Σ is the covariance matrix

PCA: Variance Maximization

- First principal component:

$$\mathbf{w}_1 = \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \Sigma \mathbf{w} \tag{1}$$

To solve the above maximization problem, we use Lagrange multipliers. Specifically there exists λ such that solution $\mathbf{w}_1$ is:

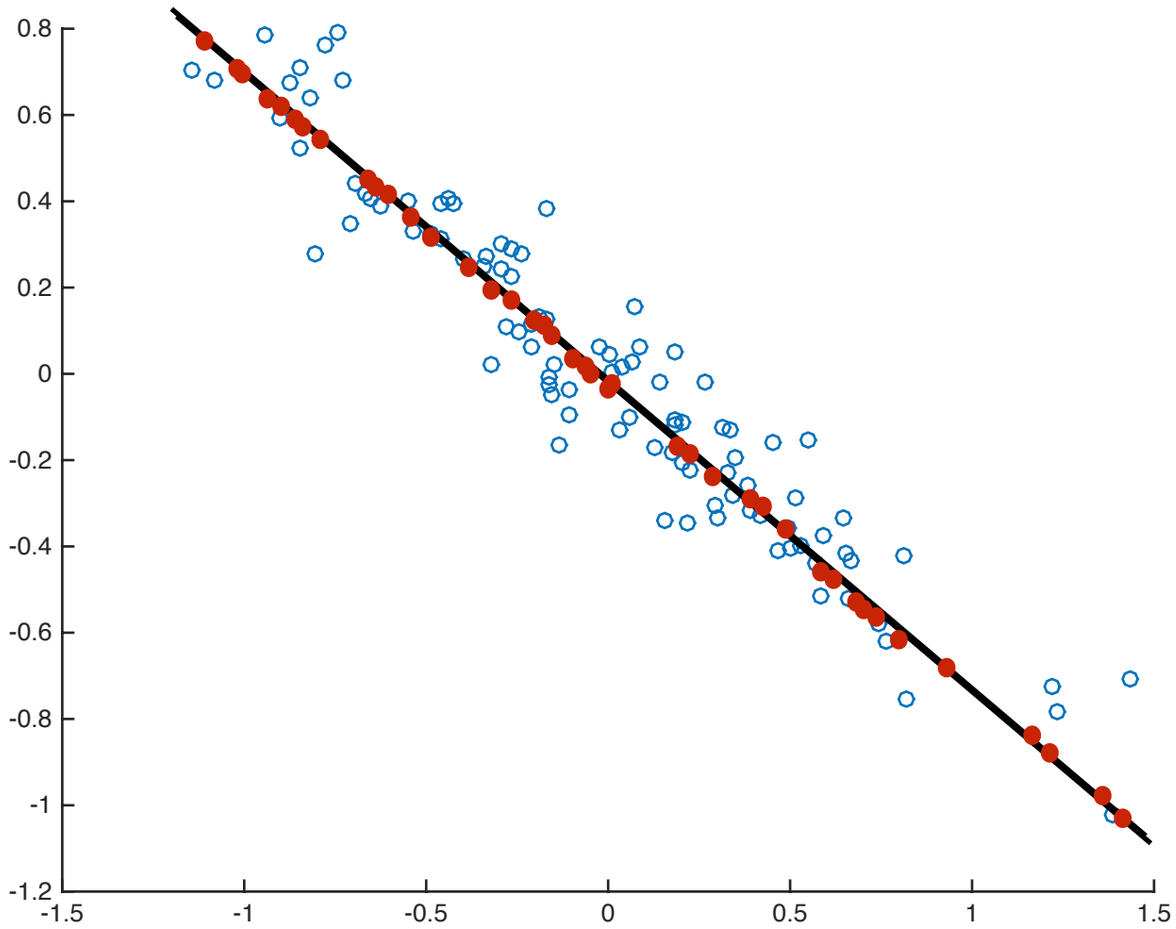
$$\mathbf{w}_1 = \arg\max_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} - \lambda \|\mathbf{w}\|_2^2$$

Taking derivate and equality to 0 we find that $\Sigma\mathbf{w} = \lambda\mathbf{w}$ (ie. eigenvector). Plugging this back into Eq. 1,

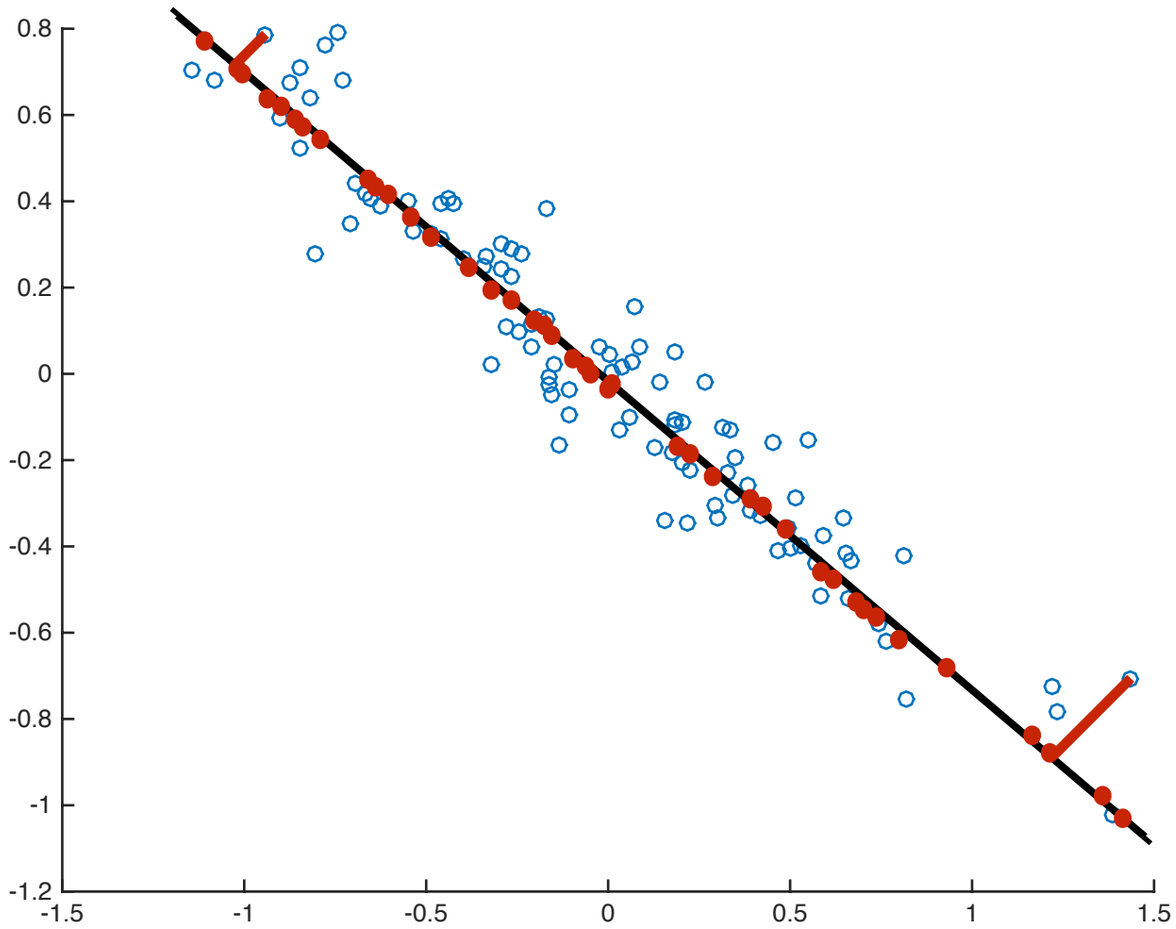
$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{w}^\top \Sigma \mathbf{w} = \mathbf{w}^\top(\lambda\mathbf{w}) = \lambda$$

Hence to maximize variance we pick direction with largest eigenvalue

PCA: Minimizing Reconstruction Error



PCA: Minimizing Reconstruction Error



PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \mathbf{x}_t \right\|_2^2$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \mathbf{x}_t \right\|_2^2$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j - \mu \right\|_2^2$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \mathbf{x}_t \right\|_2^2$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j - \mu \right\|_2^2$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} \mathbf{y}_t[j]\mathbf{w}_j \right\|_2^2$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \mathbf{x}_t \right\|_2^2$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j - \mu \right\|_2^2$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} \mathbf{y}_t[j]\mathbf{w}_j \right\|_2^2 \qquad \text{(note that } \mathbf{y}_t[j] = \mathbf{w}_j^\top(\mathbf{x}_t - \mu)\text{)}$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \mathbf{x}_t \right\|_2^2$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j - \mu \right\|_2^2$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} \mathbf{y}_t[j]\mathbf{w}_j \right\|_2^2 \qquad \text{(note that } \mathbf{y}_t[j] = \mathbf{w}_j^\top(\mathbf{x}_t - \mu)\text{)}$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} (\mathbf{w}_j^\top(\mathbf{x}_t - \mu))\mathbf{w}_j \right\|_2^2$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \mathbf{x}_t \right\|_2^2$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j - \mu \right\|_2^2$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} \mathbf{y}_t[j]\mathbf{w}_j \right\|_2^2 \qquad \text{(note that } \mathbf{y}_t[j] = \mathbf{w}_j^\top(\mathbf{x}_t - \mu)\text{)}$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} (\mathbf{w}_j^\top(\mathbf{x}_t - \mu))\mathbf{w}_j \right\|_2^2$$

$$= \sum_{t=1}^{n} \sum_{j=k+1}^{d} \left(\mathbf{w}_j^\top(\mathbf{x}_t - \mu) \right)^2$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \mathbf{x}_t \right\|_2^2$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j - \mu \right\|_2^2$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} \mathbf{y}_t[j]\mathbf{w}_j \right\|_2^2 \qquad \text{(note that } \mathbf{y}_t[j] = \mathbf{w}_j^\top(\mathbf{x}_t - \mu)\text{)}$$

$$= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} (\mathbf{w}_j^\top(\mathbf{x}_t - \mu))\mathbf{w}_j \right\|_2^2$$

$$= \sum_{t=1}^{n} \sum_{j=k+1}^{d} \left(\mathbf{w}_j^\top(\mathbf{x}_t - \mu) \right)^2 = \sum_{t=1}^{n} \sum_{j=k+1}^{d} \mathbf{w}_j^\top(\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top \mathbf{w}_j$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top(\mathbf{x}_t-\mu)(\mathbf{x}_t-\mu)^\top\mathbf{w}_j$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top(\mathbf{x}_t-\mu)(\mathbf{x}_t-\mu)^\top\mathbf{w}_j = \sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top(\mathbf{x}_t-\mu)(\mathbf{x}_t-\mu)^\top\mathbf{w}_j = \sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j$$

Minimize w.r.t. $\mathbf{w}$'s that are orthonormal,

$$\underset{\forall j,\ \|\mathbf{w}_j\|_2=1}{\operatorname{argmin}} \sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top(\mathbf{x}_t-\mu)(\mathbf{x}_t-\mu)^\top\mathbf{w}_j = \sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j$$

Minimize w.r.t. $\mathbf{w}$'s that are orthonormal,

$$\underset{\forall j,\ \|\mathbf{w}_j\|_2=1}{\operatorname{argmin}}\ \sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j$$

Using Lagrangian multipliers, there exists $\lambda_{k+1},\ldots,\lambda_d$ such that solution to above is given by:

$$\text{minimize}\ \sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j + \sum_{j=k+1}^{d}\lambda_j\|\mathbf{w}_j\|_2^2$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top(\mathbf{x}_t-\mu)(\mathbf{x}_t-\mu)^\top\mathbf{w}_j = \sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j$$

Minimize w.r.t. $\mathbf{w}$'s that are orthonormal,

$$\underset{\forall j,\ \|\mathbf{w}_j\|_2=1}{\operatorname{argmin}}\ \sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j$$

Using Lagrangian multipliers, there exists $\lambda_{k+1},\ldots,\lambda_d$ such that solution to above is given by:

$$\text{minimize}\ \sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j + \sum_{j=k+1}^{d}\lambda_j\|\mathbf{w}_j\|_2^2$$

Setting derivate to 0, $\Sigma\mathbf{w}_j = \lambda_j\mathbf{w}_j$. That is $\mathbf{w}_j$'s are eigenvectors and λ_j's are eigenvalues.

PCA: Minimizing Reconstruction Error

- Solution : $\mathbf{w}_j$'s are eigenvectors and λ_j's are corresponding eigenvalues

PCA: Minimizing Reconstruction Error

- Solution : $\mathbf{w}_j$'s are eigenvectors and λ_j's are corresponding eigenvalues
- Further, reconstruction error can be written as:

$$\underset{\mathbf{w}:\|\mathbf{w}_j\|_2=1}{\operatorname{argmin}} \sum_{j=k+1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j = \sum_{j=k+1}^{d} \lambda_j \mathbf{w}_j^\top \mathbf{w}_j = \sum_{j=k+1}^{d} \lambda_j$$

PCA: Minimizing Reconstruction Error

- Solution : $\mathbf{w}_j$'s are eigenvectors and λ_j's are corresponding eigenvalues
- Further, reconstruction error can be written as:

$$\underset{\mathbf{w}:\|\mathbf{w}_j\|_2=1}{\operatorname{argmin}} \sum_{j=k+1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j = \sum_{j=k+1}^{d} \lambda_j \mathbf{w}_j^\top \mathbf{w}_j = \sum_{j=k+1}^{d} \lambda_j$$

- Clearly to minimize reconstruction error, we need to minimize $\sum_{j=k+1}^{d} \lambda_j$. In other words we discard the $d - k$ directions that have the smallest eigenvalue

Principal Component Analysis

- Eigenvectors of the covariance matrix are the principal components

Principal Component Analysis

- Eigenvectors of the covariance matrix are the principal components
- Top K principal components are the eigenvectors with K largest eigenvalues

Principal Component Analysis

- Eigenvectors of the covariance matrix are the principal components

- Top K principal components are the eigenvectors with K largest eigenvalues

- Projection = Data $\times$ Top Keigenvectors

Principal Component Analysis

- Eigenvectors of the covariance matrix are the principal components
- Top K principal components are the eigenvectors with K largest eigenvalues
- Projection = Data $\times$ Top K eigenvectors
- Reconstruction = Projection $\times$ Transpose of top K eigenvectors

Principal Component Analysis

- Eigenvectors of the covariance matrix are the principal components

- Top K principal components are the eigenvectors with K largest eigenvalues

- Projection = Data $\times$ Top K eigenvectors

- Reconstruction = Projection $\times$ Transpose of top K eigenvectors

- Independently discovered by Pearson in 1901 and Hotelling in 1933.

Principal Component Analysis: Demo