Principal Component Analysis

Course Webpage:
http://www.cs.cornell.edu/Courses/cs4786/2016sp/
Announcements

- Waitlist size currently about 55 :(
Given feature vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, compress the data points into low dimensional representation $y_1, \ldots, y_n \in \mathbb{R}^K$ where $K \ll d$. 

$$X \overset{n}{\leftarrow} x_1^\top \cdots x_n^\top \overset{n}{\rightarrow} Y \overset{n}{\leftarrow} y_1^\top \cdots y_n^\top \overset{n}{\rightarrow} Y$$
Given feature vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, compress the data points into low dimensional representation $y_1, \ldots, y_n \in \mathbb{R}^K$ where $K << d$.
PCA: Variance Maximization

First principal direction = Top eigen vector
Principal Component Analysis

1. \[ \Sigma = \text{cov}(X) \]

2. \[ W = \text{eigs}(\Sigma, K) \]

3. \[ Y = (X - \mu) \times W \]
\[ y_2[1] = x_1^\top w = \|x_2\| \cos(\angle x w) \]
Think of $W_1, \ldots, W_K$ as coordinate system for PCA

$y$ values provide coefficients in this system

Without loss of generality, $W_1, \ldots, W_K$ can be orthonormal, i.e. $W_i \perp W_j \& \|W_i\| = 1$.

Reconstruction:

$$\hat{x}_t = y_t^T W^T + \mu$$
How do we find the remaining components?

- We are looking for orthogonal directions.
- Start with the $d$-dimensional space.
- While we haven't yet found $K$ directions, find the first principal component direction.
- Remove this direction and consider data points in the remaining subspace after projecting to the first component.

This solution is given by $W = \text{Top}_K$ eigenvectors of $\Sigma$. 
How do we find the remaining components?

We are looking for orthogonal directions.

This solution is given by

\[ W = \text{Top}_K \text{eigenvectors of } \mathbf{\Sigma} \]
How do we find the remaining components?

We are looking for orthogonal directions.

Start with the $d$ dimensional space

While we haven’t yet found $K$ directions,
  - Find first principal component direction
  - Remove this direction and consider data points in the remaining subspace after projecting to first component

End

This solution is given by $W = \text{Top } K \text{ eigenvectors of } \Sigma$
PCA: Variance Maximization

Covariance matrix:

\[
\Sigma = \frac{1}{n} \sum_{t=1}^{n} (x_t - \mu)(x_t - \mu)^\top
\]

- It's a \( d \times d \) matrix, \( \Sigma[i, j] \) measures “covariance” of features \( i \) and \( j \)
- Recall \( \text{cov}(A, B) = \mathbb{E}[(A - \mathbb{E}[A])(B - \mathbb{E}[B])] \)
- Alternatively,

\[
\Sigma[i, j] = \frac{1}{n} \begin{bmatrix}
    x_1[i] - \mu[i] \\
    \vdots \\
    x_n[i] - \mu[i]
\end{bmatrix}^\top
\begin{bmatrix}
    x_1[j] - \mu[j] \\
    \vdots \\
    x_n[j] - \mu[j]
\end{bmatrix}
\]

Inner products measure similarity.
Goal: find the basis that minimizes reconstruction error,

\[
\sum_{t=1}^{n} \| \hat{x}_t - x_t \|_2^2 = \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} y_t[j] w_j + \mu - x_t \right\|_2^2
\]

\[
= \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} y_t[j] w_j + \mu - \sum_{j=1}^{d} y_t[j] w_j - \mu \right\|_2^2
\]

\[
= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} y_t[j] w_j \right\|_2^2 \quad \text{(note that } y_t[j] = w_j^\top (x_t - \mu))
\]

\[
= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} (w_j^\top (x_t - \mu)) w_j \right\|_2^2
\]

\[
= \sum_{t=1}^{n} \sum_{j=k+1}^{d} \left( w_j^\top (x_t - \mu) \right)^2
\]
Goal: find the basis that minimizes reconstruction error,

\[
\sum_{t=1}^{n} \| \hat{x}_t - x_t \|^2 = \sum_{t=1}^{n} \left\| \sum_{j=1}^{k} y_t[j] w_j + \mu - x_t \right\|^2
\]

\[
= \sum_{t=1}^{n} \left\| \sum_{j=1}^{d} y_t[j] w_j + \mu - \sum_{j=1}^{d} y_t[j] w_j - \mu \right\|^2
\]

\[
= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} y_t[j] w_j \right\|^2 \quad \text{(note that } y_t[j] = w_j^\top (x_t - \mu) )
\]

\[
= \sum_{t=1}^{n} \left\| \sum_{j=k+1}^{d} (w_j^\top (x_t - \mu)) w_j \right\|^2
\]

\[
= \sum_{t=1}^{n} \sum_{j=k+1}^{d} \left( w_j^\top (x_t - \mu) \right)^2 = \sum_{t=1}^{n} \sum_{j=k+1}^{d} w_j^\top (x_t - \mu)(x_t - \mu)^\top w_j
\]
PCA: Minimizing Reconstruction Error

Goal: find the basis that minimizes reconstruction error,

\[
\frac{1}{n} \sum_{t=1}^{n} \| \hat{x}_t - x_t \|^2 = \frac{1}{n} \sum_{t=1}^{n} \sum_{j=k+1}^{d} w_j^\top (x_t - \mu)(x_t - \mu)^\top w_j = \sum_{j=k+1}^{d} w_j^\top \Sigma w_j
\]

Minimize w.r.t. \( w \)'s that are orthonormal,

\[
\argmin \sum_{j=k+1}^{d} w_j^\top \Sigma w_j \quad \text{s.t.} \quad \forall j, \|w_j\|_2 = 1
\]

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

\[
\text{minimize} \sum_{t=1}^{n} \sum_{j=k+1}^{d} w_j^\top \Sigma w_j + \sum_{j=k+1}^{d} \lambda_j \|w_j\|_2^2
\]
Goal: find the basis that minimizes reconstruction error,

\[
\frac{1}{n} \sum_{t=1}^{n} \| \hat{x}_t - x_t \|_2^2 = \frac{1}{n} \sum_{t=1}^{n} \sum_{j=k+1}^{d} w_j^\top (x_t - \mu)(x_t - \mu)^\top w_j = \sum_{j=k+1}^{d} w_j^\top \Sigma w_j
\]

Minimize w.r.t. \( w \)'s that are orthonormal,

\[
\arg\min_{\forall j, \|w_j\|_2 = 1} \sum_{j=k+1}^{d} w_j^\top \Sigma w_j
\]

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

\[
\minimize \sum_{t=1}^{n} \sum_{j=k+1}^{d} w_j^\top \Sigma w_j + \sum_{j=k+1}^{d} \lambda_j \|w_j\|_2^2
\]

Setting derivate to 0, \( \Sigma w_j = \lambda_j w_j \). That is \( w \)'s are eigenvectors and \( \lambda \)'s are eigenvalues.
Solution: $w_j$’s are eigenvectors and $\lambda_j$’s are corresponding eigenvalues

Further, reconstruction error can be written as:

$$\begin{align*}
\arg\min_{w: \|w_j\|_2=1} \sum_{j=k+1}^{d} w_j^\top \Sigma w_j &= \sum_{j=k+1}^{d} \lambda_j w_j^\top w_j = \sum_{j=k+1}^{d} \lambda_j
\end{align*}$$

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

1. \[ \Sigma = \text{cov}(X) \]

2. \[ W = \text{eigs}(\Sigma, K) \]

3. \[ Y = (X - \mu) \times W \]
4. \hat{X} = Y \times W^T + \mu
When $d >> n$

- If $d >> n$ then $\Sigma$ is large
- But we only need top $K$ eigen vectors.
- Idea: use SVD

$$X - \mu = UDV^T$$

Then note that, $\Sigma = (X - \mu)(X - \mu)^T = UD^2U$

- Hence, matrix $U$ is the same as matrix $W$ got from eigen decomposition of $\Sigma$, eigenvalues are diagonal elements of $D^2$
- Alternative algorithm:

$$W = \text{SVD}(X - \mu, K)$$