Canonical Correlation Analysis (CCA)

Course Webpage:
http://www.cs.cornell.edu/Courses/cs4786/2016fa/
Announcement

• We are grading HW0 and you will be added to cms by monday

• HW1 will be posted tonight on webpage (homework tab)

• HW1 on CCA and PCA (due in a week)
Assume points are centered. Which of the following are equal to the covariance matrix?

A. $\Sigma = \frac{1}{n} \sum_{t=1}^{n} x_t^\top x_t$

B. $\Sigma = \frac{1}{n} \sum_{t=1}^{n} x_t x_t^\top$

C. $\Sigma = XX^\top$

D. $\Sigma = XX^\top$
Example: Students in classroom
Maximize Spread

Minimize Reconstruction Error
**Principal Component Analysis**

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

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

3. \[ Y = (X - \mu) \times W \]

Eigenvectors of the covariance matrix are the principal components. Top K principal components are the eigenvectors with K largest eigenvalues. Independently discovered by Pearson in 1901 and Hotelling in 1933.

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

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

\[ Y = (X - \mu) \times W \]
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, $\Sigma = (X - \mu)^\top (X - \mu) = VD^2V$
- Hence, matrix $V$ is the same as matrix $W$ got from eigen decomposition of $\Sigma$, eigenvalues are diagonal elements of $D^2$
- Alternative algorithm:
  \[ [U, V] = \text{SVD}(X - \mu, K) \quad W = V \]
When to use PCA?

- When data naturally lies in a low dimensional linear subspace
- To minimize reconstruction error
- Find directions where data is maximally spread
Canonical Correlation Analysis

Age
Gender
Angle
Canonical Correlation Analysis

- Age
- Gender
- Angle
Data comes in pairs \((x_1, x'_1), \ldots, (x_n, x'_n)\) where \(x_t\)'s are \(d\) dimensional and \(x'_t\)'s are \(d'\) dimensional.

Goal: Compress say view one into \(y_1, \ldots, y_n\), that are \(K\) dimensional vectors

- Retain information redundant between the two views
- Eliminate "noise" specific to only one of the views
Audio might have background sounds uncorrelated with video

Video might have lighting changes uncorrelated with audio

Redundant information between two views: the speech
Method A and Method B are both equally good feature extraction techniques.

Concatenating the two features blindly yields large dimensional feature vector with redundancy.

Applying techniques like CCA extracts the key information between the two methods.

Removes extra unwanted information.
How do we get the right direction? (say $K = 1$)

- Age
- Gender
- Angle
WHICH DIRECTION TO PICK?

View I

View II
Which Direction to Pick?

PCA direction
Which Direction to Pick?

Direction has large covariance
How do we pick the right direction to project to?
Say $\mathbf{w}_1$ and $\mathbf{v}_1$ are the directions we choose to project in views 1 and 2 respectively we want these directions to maximize,

$$
\frac{1}{n} \sum_{t=1}^{n} \left( y_t[1] - \frac{1}{n} \sum_{t=1}^{n} y_t[1] \right) \cdot \left( y'_t[1] - \frac{1}{n} \sum_{t=1}^{n} y'_t[1] \right)
$$

where $y_t[1] = \mathbf{w}_1^\top \mathbf{x}_t$ and $y'_t[1] = \mathbf{v}_1^\top \mathbf{x}'_t$
Maximizing Correlation Coefficient

Say \( w_1 \) and \( v_1 \) are the directions we choose to project in views 1 and 2 respectively we want these directions to maximize,

\[
\frac{1}{n} \sum_{t=1}^{n} \left( y_t[1] - \frac{1}{n} \sum_{t=1}^{n} y_t[1] \right) \cdot \left( y'_t[1] - \frac{1}{n} \sum_{t=1}^{n} y'_t[1] \right)
\]

s.t. \( \frac{1}{n} \sum_{t=1}^{n} (y_t[1] - \frac{1}{n} \sum_{t=1}^{n} y_t[1])^2 = \frac{1}{n} \sum_{t=1}^{n} (y'_t[1] - \frac{1}{n} \sum_{t=1}^{n} y'_t[1]) = 1 \)

where \( y_t[1] = w_1^\top x_t \) and \( y'_t[1] = v_1^\top x'_t \)
What is the problem with the above?
Why not Maximize Covariance

Say \[ \frac{1}{n} \sum_{t=1}^{n} x_t[2] \cdot x_t'[2] > 0 \]

Scaling up this coordinate we can blow up covariance
Say $w_1$ and $v_1$ are the directions we choose to project in views 1 and 2 respectively we want these directions to maximize,

$$\frac{\frac{1}{n} \sum_{t=1}^{n} (y_t[1] - \frac{1}{n} \sum_{t=1}^{n} y_t[1]) \cdot (y'_t[1] - \frac{1}{n} \sum_{t=1}^{n} y'_t[1])}{\sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t[1] - \frac{1}{n} \sum_{t=1}^{n} y_t[1])^2} \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y'_t[1] - \frac{1}{n} \sum_{t=1}^{n} y'_t[1])}}$$
Basic Idea of CCA

- Normalize variance in chosen direction to be constant (say 1)
- Then maximize covariance
- This is same as maximizing “correlation coefficient”
Covariance $Vs$ Correlation

- **Covariance** $(A, B) = \mathbb{E}[(A - \mathbb{E}[A]) \cdot (B - \mathbb{E}[B])]$

  Depends on the scale of $A$ and $B$. If $B$ is rescaled, covariance shifts.

- **Correlation** $(A, B) = \frac{\mathbb{E}[(A - \mathbb{E}[A]) \cdot (B - \mathbb{E}[B])]}{\sqrt{\text{Var}(A)} \sqrt{\text{Var}(B)}}$

  Scale free.
**Maximizing Correlation Coefficient**

Say \( \mathbf{w}_1 \) and \( \mathbf{v}_1 \) are the directions we choose to project in views 1 and 2 respectively we want these directions to maximize,

\[
\frac{1}{n} \sum_{t=1}^{n} \left( \mathbf{y}_t[1] - \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_t[1] \right) \cdot \left( \mathbf{y}_t'[1] - \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_t'[1] \right)
\]

s.t. \( \frac{1}{n} \sum_{t=1}^{n} \left( \mathbf{y}_t[1] - \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_t[1] \right)^2 = \frac{1}{n} \sum_{t=1}^{n} \left( \mathbf{y}_t'[1] - \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_t'[1] \right)^2 = 1 \)

where \( \mathbf{y}_t[1] = \mathbf{w}_1^\top \mathbf{x}_t \) and \( \mathbf{y}_t'[1] = \mathbf{v}_1^\top \mathbf{x}_t' \)
Hence we want to solve for projection vectors $\mathbf{w}_1$ and $\mathbf{v}_1$ that

maximize $\frac{1}{n} \sum_{t=1}^{n} \mathbf{w}_1^\top (\mathbf{x}_t - \mu) \cdot \mathbf{v}_1^\top (\mathbf{x}_t' - \mu')$

subject to $\frac{1}{n} \sum_{t=1}^{n} (\mathbf{w}_1^\top (\mathbf{x}_t - \mu))^2 = \frac{1}{n} \sum_{t=1}^{n} (\mathbf{v}_1^\top (\mathbf{x}_t' - \mu'))^2 = 1$

where $\mu = \frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t$ and $\mu' = \frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t'$
Hence we want to solve for projection vectors $w_1$ and $v_1$ that

$$\text{maximize } \frac{1}{n} \sum_{t=1}^{n} w_1^\top (x_t - \mu) \cdot v_1^\top (x_t' - \mu')$$

subject to

$$\frac{1}{n} \sum_{t=1}^{n} (w_1^\top (x_t - \mu))^2 = \frac{1}{n} \sum_{t=1}^{n} (v_1^\top (x_t' - \mu'))^2 = 1$$

where $\mu = \frac{1}{n} \sum_{t=1}^{n} x_t$ and $\mu' = \frac{1}{n} \sum_{t=1}^{n} x'_t$
Hence we want to solve for projection vectors $\mathbf{w}_1$ and $\mathbf{v}_1$ that maximize $\mathbf{w}_1^\top \Sigma_{1,2} \mathbf{v}_1$

subject to $\mathbf{w}_1^\top \Sigma_{1,1} \mathbf{w}_1 = \mathbf{v}_1^\top \Sigma_{2,2} \mathbf{v}_1 = 1$
$W_1 = \text{eigs}(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, K)$

$W_2 = \text{eigs}(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}, K)$
1. \( X = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \)

2. \( \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \text{cov}(X) \)

3. \( W_1 = \text{eigs}(\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}, K) \)

4. \( Y_1 = (X - \mu_1) \times W_1 \)