Machine Learning for Data Science (CS4786)
Lecture 2

Dimensionality Reduction
&
Principal Component Analysis

Course Webpage:
http://www.cs.cornell.edu/Courses/cs4786/2016sp/
Diagnostic assignment due on 4th Feb (Thursday) beginning of class

Course webpage is the official source of all class related information

You will be added to CMS once you return Assignment 0 with your net-id on it.
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
Example: Images

vectorize
Example: Images

\[ d = K^2 \]

Vectorize
**Example: Text (Bag of Words)**

*Documents:*

- Car
- Engine
- Hood
- Tires
- Truck
- Trunk

- Car
- Emissions
- Hood
- Make
- Model
- Trunk

- Chomsky
- Corpus
- Noun
- Parsing
- Tagging
- Wonderful
**Example: Text (Bag of Words)**

**Documents:**

- Car
- Chomsky
- Corpus
- Emissions
- Engine
- Hood
- Make
- Model
- Noun
- Parsing
- Tagging
- Tires
- Truck
- Trunk
- Wonderful

<table>
<thead>
<tr>
<th></th>
<th>car</th>
<th>Chomsky</th>
<th>corpus</th>
<th>emissions</th>
<th>engine</th>
<th>hood</th>
<th>make</th>
<th>model</th>
<th>noun</th>
<th>parsing</th>
<th>tagging</th>
<th>tires</th>
<th>truck</th>
<th>truck</th>
<th>wonderful</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
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
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 $K << d$. 
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 $K \ll d$. 
Dimensionality Reduction

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 $K \ll d$. 

$n \times d$
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 $K \ll d$. 
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 $K \ll d$. 

\[
\begin{bmatrix}
\mathbf{x}_1 \\
\vdots \\
\mathbf{x}_n
\end{bmatrix} \xrightarrow{} 
\begin{bmatrix}
\mathbf{y}_1 \\
\vdots \\
\mathbf{y}_n
\end{bmatrix}
\]
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

\[ X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \]
Dim Reduction: Linear Transformation

\[ X \times dW = K \]
\[ X \times dW = Y \]
**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**

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

\[
X = \begin{bmatrix}
1 & 2 & 3 & 4 \\
1.1 & 2 & 3 & 4 \\
3 & 2 & 3 & 4 \\
-1 & 2 & 3 & 4 \\
-0.2 & 2 & 3 & 4 \\
-2 & 2 & 3 & 4 \\
1.4 & 2 & 3 & 4 \\
1.4 & 2 & 3 & 4 \\
-0.1 & 2 & 3 & 4 \\
0.5 & 2 & 3 & 4
\end{bmatrix} = \begin{bmatrix}
1 \\
1.1 \\
3 \\
-1 \\
-0.2 \\
-2 \\
1.4 \\
1.4 \\
-0.1 \\
0.5
\end{bmatrix} \times [1, 0, 0, 0] + \begin{bmatrix}
0 & 2 & 3 & 4 \\
0 & 2 & 3 & 4 \\
0 & 2 & 3 & 4 \\
0 & 2 & 3 & 4 \\
0 & 2 & 3 & 4 \\
0 & 2 & 3 & 4 \\
0 & 2 & 3 & 4 \\
0 & 2 & 3 & 4 \\
0 & 2 & 3 & 4 \\
0 & 2 & 3 & 4
\end{bmatrix}
\]
**Dim Reduction: Linear Transformation**

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

\[
Y
\]
Compressing these data points…
… is same as compressing these.
Centering Data

\[ -\mu \]
Centering Data

\[ n \frac{X - \mu}{\sigma} \]
Centering Data

\[ n \frac{X - \mu}{\delta} \quad \text{and} \quad n \frac{Y}{K} \]
Centering Data

\[ n(X - \mu) \times dW = nY \]

Where:
- \( n \) is the number of observations.
- \( X - \mu \) represents the centered data.
- \( d \) is the number of dimensions.
- \( W \) is the weight matrix.
- \( K \) is the kernel function.
PCA: Variance Maximization

Pick directions along which data varies the most

First principal component:
\[ w_1 = \arg \max_w w : \sum_{t=1}^{n} (w x_t - \mu)^2 = \sum_{t=1}^{n} \|w x_t - \mu\|^2 = \arg \max_w \sum_{t=1}^{n} \|w^T (x_t - \mu)\|^2 \]

\[ w \triangleq \frac{1}{n} \sum_{t=1}^{n} x_t \]

Writing down Lagrangian and optimizing,
\[ w \triangleq \frac{1}{n} \sum_{t=1}^{n} x_t \]
**PCA: Variance Maximization**

Pick directions along which data varies the most.

The first principal component is given by:

\[ w_1 = \arg \max_{w : \|w\|^2 = 1} \sum_{t=1}^{n} \langle w, x_t - \mu \rangle^2 = \arg \max_{w : \|w\|^2 = 1} \sum_{t=1}^{n} w^T (x_t - \mu) (x_t - \mu)^T w \]

where \( \mu \) is the mean of the data. Writing down the Lagrangian and optimizing gives:

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

This is the covariance matrix.

The diagram shows a scatter plot of data points with the first principal component indicated by the red line.
**PCA: Variance Maximization**

Pick directions along which data varies the most

First principal component:

\[
\mathbf{w}_1 = \arg \max_{\mathbf{w}} \mathbf{w} : \mathbf{w} \mathbf{w}^T = 1 \\
\mathbf{w}^T \mathbf{x}_t = 1/n \sum_{t=1}^n \mathbf{w}^T (\mathbf{x}_t - \mu) \mathbf{w}^T = \arg \max_{\mathbf{w}} \mathbf{w} : \mathbf{w}^T \mathbf{w} = 1 \\
\mathbf{w}^T \mathbf{x}_t = \mathbf{w}^T \mathbf{w} = \mathbf{w}^T \mathbf{\Sigma} \mathbf{w} = 1
\]

\[
\mathbf{\Sigma} = \mathbf{w}_1 \mathbf{w}_1^T
\]
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:

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

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

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

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

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

\[
= \arg \max_{\mathbf{w}: \|\mathbf{w}\|_2=1} \mathbf{w}^\top \Sigma \mathbf{w}
\]

\(\Sigma\) is the covariance matrix
PCA: Variance Maximization

Covariance matrix:

$$\Sigma = \frac{1}{n} \sum_{t=1}^{n} (x_t - \mu)(x_t - \mu)^T$$

Inner products measure similarity.
Covariance matrix:

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

- It's a \( d \times d \) matrix, \( \Sigma[i, j] \) measures "covariance" of features \( i \) and \( j \)
PCA: Variance Maximization

Covariance matrix:

$$\Sigma = \frac{1}{n} \sum_{t=1}^{n} (x_t - \mu)(x_t - \mu)^T$$

- It's a $d \times d$ matrix, $\Sigma[i, j]$ measures "covariance" of features $i$ and $j$
- Recall $\text{cov}(A, B) = \mathbb{E}[(A - \mathbb{E}[A])(B - \mathbb{E}[B])]$
PCA: Variance Maximization

Covariance matrix:

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

- 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,

\[
\Sigma[i, j] = \frac{1}{n} \begin{bmatrix}
    x_1[j] - \mu[j] \\
    x_2[j] - \mu[j] \\
    \vdots \\
    x_n[j] - \mu[j]
\end{bmatrix}^T
\begin{bmatrix}
    x_1[j] - \mu[j] \\
    x_2[j] - \mu[j] \\
    \vdots \\
    x_n[j] - \mu[j]
\end{bmatrix}
\]

Inner products measure similarity.
PCA: Variance Maximization

First principal component:

\[ w_1 = \arg \max_{w: \|w\|_2=1} w^\top \Sigma w \]  \hspace{1cm} (1)

To solve the above maximization problem, we use Lagrange multipliers. Specifically, there exists such that the solution \( w_1 \) is:

\[ w_1 = \arg \max_{w} w^\top \Sigma w - \|w\|_2^2 \]

Taking the derivative and setting it to zero, we find that \( \Sigma w = \lambda w \) (i.e., eigenvector). Plugging this back into Eq. 1,

\[ \sum_{i=1}^{n} x_i = w_1^\top \sum w = w_1^\top (w_1) = \lambda \]

Hence to maximize variance, we pick the direction with the largest eigenvalue.
First principal component:

\[ w_1 = \arg \max_{w: \|w\|_2 = 1} w^T \Sigma w \]  

(1)

To solve the above maximization problem, we use Lagrange multipliers. Specifically there exists \( \lambda \) such that solution \( w_1 \) is:

\[ w_1 = \arg \max_{w} w^T \Sigma w - \lambda \|w\|^2_2 \]
First principal component:

\[ w_1 = \arg \max_{w: \|w\|_2 = 1} w^T \Sigma w \quad (1) \]

To solve the above maximization problem, we use Lagrange multipliers. Specifically there exists \( \lambda \) such that solution \( w_1 \) is:

\[ w_1 = \arg \max_w w^T \Sigma w - \lambda \|w\|_2^2 \]

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

\[ \frac{1}{n} \sum_{t=1}^{n} w^T \Sigma w = w^T (\lambda w) = \lambda \]

Hence to maximize variance we pick direction with largest eigenvalue
Eigenvectors of the covariance matrix are the principal components
Eigenvectors of the covariance matrix are the principal components

Top $K$ principal components are the eigenvectors with $K$ largest eigenvalues
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
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.
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