Dimensionality Reduction
&
Principal Component Analysis

Course Webpage:
http://www.cs.cornell.edu/Courses/cs4786/2016fa/
How do we represent data?

Each data-point often represented as vector referred to as feature vector.
Example: Images

\[ d = K^2 \]
**Example: Text (Bag of Words)**

**Documents:**

- car
- Chomsky
- corpus
- emissions
- engine
- hood
- make
- model
- noun
- parsing
- tagging
- tires
- truck
- trunk
- wonderful

Matrix:

<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>0</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>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</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>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>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
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 \( K << d \).

\[
X_{n \times d} \xrightarrow{\text{Dimensionality Reduction}} Y_{n \times K}
\]
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 $K << d$.
Flowers

Iris-Setosa

Iris-versicolor

Iris-virginica
Principal Component Analysis: Demo
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
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

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

\[ x_1 \cdots x_n \]
\[ y_1 \cdots y_n \]

\[ X \times d W = Y \]

\[ y_i^\top = x_i^\top W \]
Prelude: reducing to 1 dimension

\[ y_1 = w^\top x_1 = \|x_1\| \cos (\angle w x_1) \]
PCA: Variance Maximization

First principal component:

\[
\mathbf{w}_1 = \arg \max_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} = 1
\]

subject to

\[
\mathbf{w}^\top \mathbf{1} = 1
\]

\[
\mathbf{w}^\top (\mathbf{x}_t - \mu) = \mathbf{w}^\top \mathbf{1} \mathbf{w} = \mathbf{w}^\top \mathbf{w} = 1
\]

Writing down Lagrangian and optimizing, we have

\[
\mathbf{w}_1 = \mathbf{\Sigma}^{-1/2} \mathbf{e}_1
\]
PCA: Variance Maximization

Pick directions along which data varies the most

First principal component:

\[ w_1 = \arg \max_{w : \|w\|^2 = 1} \mathbb{E}_{t=1}^{n} \|w^\top x_t - \mu\|^2 \]

\[ \mathbb{E}_{t=1}^{n} \|w^\top (x_t - \mu)\|^2 = \arg \max_{w : \|w\|^2 = 1} w^\top \mathbb{C} w \]

\[ \mathbb{C} \text{ is the covariance matrix} \]

Writing down Lagrangian and optimizing,

\[ w \mathbb{C} w = \]
PCA: VARIANCE MAXIMIZATION

Pick directions along which data varies the most

First principal component:

\[ w^1 = \arg \max_{w : \|w\|^2 = 1} \sum_{t=1}^{n} \left( w^T x_t - \mu \right)^2 \]

Writing down Lagrangian and optimizing,

\[ w = \mathbf{w}_1 \]
PCA: Variance Maximization

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

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

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

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

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

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

where \( \Sigma \) is the covariance matrix and \( \mu = \frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t \)
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 $j$

$$\Sigma[i, j] = \frac{1}{n} \sum_{t=1}^{n} (x_t[i] - \mu[i])(x_t[j] - \mu[j])$$
First principal component:

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

The solution to the above optimization problem is \( w_1 \) is the top Eigen vector of matrix \( \Sigma \)

Hence in “matlab”,

\[
S = \text{Cov}(X) \\
[W, E] = \text{eigs}(S, 1) \\
Y = W \times X
\]
Principal Component Analysis: Demo
What do we do when $K > 1$?
Prelude: reducing to 1 dimension

\[ \hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4 \rightarrow x_1, x_2, x_3, x_4 \]
Think of $\mathbf{w}_1, \ldots, \mathbf{w}_K$ as coordinate system for PCA (in a $K$ dimensional subspace)

$\mathbf{y}$ values provide coefficients in this system

Without loss of generality, $\mathbf{w}_1, \ldots, \mathbf{w}_K$ can be orthonormal, i.e. $\mathbf{w}_i \perp \mathbf{w}_j$ & $\|\mathbf{w}_i\| = 1$.

Reconstruction:

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

If we take all $\mathbf{w}_1, \ldots, \mathbf{w}_d$, then $\mathbf{x}_t = \sum_{j=1}^{d} y_t[j] \mathbf{w}_j$. To reduce dimensionality we only consider first $K$ vectors of the basis
How do we find the $K$ components?

We are looking for orthogonal directions that maximize total spread in each direction.

Find orthonormal $W$ that maximizes

$$\sum_{j=1}^{K} \frac{1}{n} \sum_{t=1}^{n} \left( y_t[j] - \frac{1}{n} \sum_{t=1}^{n} y_t[j] \right)^2 = \sum_{j=1}^{K} \frac{1}{n} \sum_{t=1}^{n} \left( w_j^T \left( x_t - \frac{1}{n} \sum_{t=1}^{n} x_t \right) \right)^2$$

$$= \sum_{j=1}^{K} w_j^T \Sigma w_j$$

This solutions is given by $W = \text{Top } K \text{ eigenvectors of } \Sigma$
**Principal Component Analysis**

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

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

3. \( Y = X \times W \)

 Independently discovered by Pearson in 1901 and Hotelling in 1933.
Principal Component Analysis: Demo