CHAPTER 6:
Dimensionality Reduction
Why Reduce Dimensionality?

1. Reduces time complexity: Less computation
2. Reduces space complexity: Less parameters
3. Saves the cost of observing the feature
4. Simpler models are more robust on small datasets
5. More interpretable; simpler explanation
6. Data visualization (structure, groups, outliers, etc) if plotted in 2 or 3 dimensions
Feature Selection vs Extraction

- **Feature selection:** Choosing \( k < d \) important features, ignoring the remaining \( d - k \)
  - Subset selection algorithms
- **Feature extraction:** Project the original \( x_i, \ i = 1, \ldots, d \) dimensions to new \( k < d \) dimensions, \( z_j, \ j = 1, \ldots, k \)

  Principal components analysis (PCA), linear discriminant analysis (LDA), factor analysis (FA)
Subset Selection

- There are $2^d$ subsets of $d$ features
- Forward search: Add the best feature at each step
  - Set of features $F$ initially $\emptyset$.
  - At each iteration, find the best new feature
    \[ j = \arg\min_i E( F \cup x_i) \]
  - Add $x_j$ to $F$ if $E( F \cup x_j) < E( F )$

- Hill-climbing $O(d^2)$ algorithm
- Backward search: Start with all features and remove one at a time, if possible.
- Floating search (Add $k$, remove $l$)
Principal Components Analysis (PCA)

- Find a low-dimensional space such that when \( x \) is projected there, information loss is minimized.
- The projection of \( x \) on the direction of \( w \) is: \( z = w^T x \)
- Find \( w \) such that \( \text{var}(z) \) is maximized

\[
\text{var}(z) = \text{var}(w^T x) = \mathbb{E}[(w^T x - w^T \mu)^2] \\
= \mathbb{E}[(w^T x - w^T \mu)(w^T x - w^T \mu)] \\
= \mathbb{E}[w^T(x - \mu)(x - \mu)^T w] \\
= w^T \mathbb{E}[(x - \mu)(x - \mu)^T] w = w^T \Sigma w
\]

where \( \text{var}(x) = \mathbb{E}[(x - \mu)(x - \mu)^T] = \Sigma \)
Maximize \( \text{var}(z) \) subject to \( ||w||=1 \)

\[
\max_{w_1} w_1^T \Sigma w_1 - \alpha (w_1^T w_1 - 1)
\]

\( \Sigma w_1 = \alpha w_1 \) that is, \( w_1 \) is an eigenvector of \( \Sigma \)
Choose the one with the largest eigenvalue for \( \text{var}(z) \) to be max

Second principal component: Max \( \text{var}(z_2) \), s.t.,
\( ||w_2||=1 \) and orthogonal to \( w_1 \)

\[
\max_{w_2} w_2^T \Sigma w_2 - \alpha (w_2^T w_2 - 1) - \beta (w_2^T w_1 - 0)
\]

\( \Sigma w_2 = \alpha w_2 \) that is, \( w_2 \) is another eigenvector of \( \Sigma \)
and so on.
What PCA does

\[ z = W^T(x - m) \]

where the columns of \( W \) are the eigenvectors of \( \Sigma \), and \( m \) is sample mean

Centers the data at the origin and rotates the axes
How to choose $k$?

- Proportion of variance (PoV) explained

\[
\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_k + \cdots + \lambda_d}
\]

when $\lambda_i$ are sorted in descending order

- Typically, stop at PoV>0.9

- Scree graph plots of PoV vs $k$, stop at “elbow”
Factor Analysis

Find a small number of factors $z$, which when combined generate $x$:

$$x_i - \mu_i = \nu_{i1}z_1 + \nu_{i2}z_2 + \ldots + \nu_{ik}z_k + \epsilon_i$$

where $z_j$, $j = 1,...,k$ are the latent factors with

$E[z_j]=0$, $\text{Var}(z_j)=1$, $\text{Cov}(z_i, z_j)=0$, $i \neq j$,

$\epsilon_i$ are the noise sources

$E[\epsilon_i]=\psi_i$, $\text{Cov}(\epsilon_i, \epsilon_j)=0$, $i \neq j$, $\text{Cov}(\epsilon_i, z_j)=0$ ,

and $\nu_{ij}$ are the factor loadings
PCA vs FA

- **PCA** From \( x \) to \( z \)
  \[ z = W^T(x - \mu) \]

- **FA** From \( z \) to \( x \)
  \[ x - \mu = Vz + \varepsilon \]
**Factor Analysis**

- In FA, factors $z_j$ are stretched, rotated and translated to generate $x$
Multidimensional Scaling

- Given pairwise distances between $N$ points, $d_{ij}$, $i,j = 1, \ldots, N$
  - place them on a low-dimensional map such that the distances are preserved.
- $z = g(x | \theta)$
  - Find $\theta$ that min Sammon stress

\[
E(\theta | X) = \sum_{r,s} \frac{(\|z^r - z^s\| - \|x^r - x^s\|)^2}{\|x^r - x^s\|^2}
\]

\[
= \sum_{r,s} \frac{(\|g(x^r | \theta) - g(x^s | \theta)\| - \|x^r - x^s\|)^2}{\|x^r - x^s\|^2}
\]
Map of Europe by MDS

Map from CIA – The World Factbook: http://www.cia.gov/
Linear Discriminant Analysis

- Find a low-dimensional space such that when $x$ is projected, classes are well-separated.
- Find $w$ that maximizes

$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$

$$m_1 = \frac{\sum_t w^T x^t r^t}{\sum_t r^t} = w^T m_1$$

$$s_1^2 = \sum_t (w^T x^t - m_1)^2 r^t$$
- Between-class scatter

\[
(m_1 - m_2)^2 = (w^T m_1 - w^T m_2)^2 = w^T (m_1 - m_2)(m_1 - m_2)^T w = w^T S_B w
\]

- Within-class scatter

\[
s_1^2 = \sum_t (w^T x^t - m_1)^2 r^t = \sum_t w^T (x^t - m_1)(x^t - m_1)^T w r^t = w^T S_1 w
\]

where \( S_1 = \sum_t r^t (x^t - m_1)(x^t - m_1)^T \)

\[
s_1^2 + s_2^2 = w^T S_W w
\]

where \( S_W = S_1 + S_2 \)
Fisher’s Linear Discriminant

- Find \( \mathbf{w} \) that max

\[
J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} = \frac{|\mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)|^2}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}
\]

- LDA soln:

\[
\mathbf{w} = c \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)
\]

- Parametric soln:

\[
\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2)
\]

when \( p(\mathbf{x}|C_i) \sim \mathcal{N}(\mu_i, \Sigma) \)
$K > 2$ Classes

- **Within-class scatter:**
  
  $$S_W = \sum_{i=1}^{K} S_i$$
  
  $$S_i = \sum_{t} r_i^t (x^t - m_i) (x^t - m_i)^T$$

- **Between-class scatter:**

  $$S_B = \sum_{i=1}^{K} N_i (m_i - m) (m_i - m)^T$$

  $$m = \frac{1}{K} \sum_{i=1}^{K} m_i$$

- **Find $W$ that max**

  $$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|}$$

  The largest eigenvectors of $S_W^{-1} S_B$ maximum rank of $K - 1$