**Goals of LDA**

1. Linear discriminant analysis has two goals:
   - **Discrimination**: use the information in a training set to construct a classifier that maximizes separation of the classes.
   - **Classification**: given a new observation, use the classifier to predict the class of that observation.

2. Sometimes don't care about interpreting the classifier.
   - Just want a black box that can predict well.

3. Sometimes interested in interpreting the classifier in order to understand how the variables relate to the classes.
   - Make inferences about the classes.

**Ways to LDA**

1. Gaussian LDA

2. Fisher's linear discriminant

**Set-up**

1. Population is partitioned into $k$ unordered classes ($k \geq 2$)
   - $\Pi_0, \Pi_1, \ldots, \Pi_{k-1}$.

2. $f_k(x)$ on $\mathbb{R}^p$ is a probability density associated with population $\Pi_k$.
   - $X$: $p$-variate random vector (of features) is presumably relevant variables.
   - Each observation belongs to exactly one of $k$ possible classes.

3. A classifier $\Phi$ is a mapping from the domain of $X$ to $\mathbb{R}_k$.
   - How? divide $\mathbb{R}^p$ into disjoint regions, $R_0, \ldots, R_{k-1}$.
   - $x \in \Theta_{R_k}$ $\Rightarrow$ allocate $x$ into population $k$.
   - $P_k \cap R_{k'} = \emptyset \quad k \neq k'$
   - $\bigcup_{k=0}^{k-1} R_k = \mathbb{R}^p$.
4. **Training data**

\[ X = \begin{bmatrix}
X_0 \\
\vdots \\
X_{K-1}
\end{bmatrix} \rightarrow n_{\text{exp}} \times p \]

\[ n_{\text{exp}} \rightarrow \sum_{k=0}^{K-1} n_k \]

\( X_k \) (n_{exp} \times p \text{ matrix}) represents a sample of \( n_k \) individuals from the population \( \Pi_k \).

5. Use the training data set to construct a classifier.

6. Given future observation, \( X_{t+1} \), allocate it to one of these unlabeled \( K \) groups.

There may be no clear distinction between \( K \rightarrow \text{classification rules cannot usually provide an error free method of assignment} \).

\( \Rightarrow \) A good classifier should result in few misclassification.

**Loss Function**

(cost associated w/ misclassification)

1. How much should we pay for a mistake?

2. In some problems, one type of mistake can be more serious than another.

\( \text{ex: medical diagnosis - unnecessary vs risky} \)

\( \text{failing to diagnose a potentially fatal illness vs concluding that a disease is present when it is not} \).

3. \( L(i,j) = \begin{cases} 
0 & \text{if } i=j \\
W_{ij} & \text{if } i \neq j
\end{cases} \)

\[ \begin{array}{cccc}
\text{Action} & 0 & 1 & \ldots & K-1 \\
\hline
0 & 0 & W_{01} & \ldots & W_{0,K-1} \\
1 & \vdots & \ddots & \ddots & \vdots \\
K-1 & & & & 0
\end{array} \]
A loss function representing the cost or loss incurred when an observation is allocated to \( \pi_i \) when in fact it comes from \( \pi_j \).

\[
\begin{align*}
L(i,j) &= \begin{cases} 
0 & \text{if } i = j \\
W_{01} & \text{if } i = 0, j = 1 \\
W_{10} & \text{if } i = 1, j = 0
\end{cases}
\end{align*}
\]

If \( W_{01} = W_{10} = 1 \), then it is called zero-one loss.

\[\text{Ex 2: } K \geq 2 \quad \text{Many possibilities up to } K \cdot (K-1) \text{ possible distinct values.} \]

Simplest is 0/1 loss.

\[
L(i,j) = \begin{cases} 
0 & \text{if } i = j \\
1 & \text{if } i \neq j
\end{cases}
\]

From the binary case (\( K = 2 \))

Expected cost of misclassification = \( W_{01} \cdot P(X \text{ comes from } \pi_0 \text{ and is classified into } \pi_1) + W_{10} \cdot P(X \text{ comes from } \pi_1 \text{ and is classified into } \pi_0) \)

\[= W_{01} \cdot P(X \in \pi_0 | \pi_1) \cdot P(\pi_0) + W_{10} \cdot P(X \in \pi_1 | \pi_0) \cdot P(\pi_1) \]

1. Determine the region, \( \pi_0 \) and \( \pi_1 \), that minimize the risk.

2. If \( \frac{\sum_{i \neq 0} W_{01} P(X \in \pi_0 | \pi_i) P(\pi_0)}{\sum_{i \neq 0} W_{10} P(X \in \pi_1 | \pi_i) P(\pi_1)} > 1 \) classify \( X^* \) into \( \pi_0 \)

For \( K > 2 \), classify \( X \) as \( \pi_k \) if

\[
\frac{\sum_{i \neq 0} W_{ki} P(X \in \pi_k | \pi_i) P(\pi_i)}{\sum_{i \neq 0} W_{kj} P(X \in \pi_j | \pi_i) P(\pi_i)}
\]

\[< \frac{1}{K-1} \sum_{i \neq 0} W_{ij} P(X \in \pi_j | \pi_i) P(\pi_i) \quad \text{for any } j = 1, \ldots, K-1\]
When 0-1 loss is assumed, minimizing the expected cost of misclassification is equivalent to minimizing the total probability of misclassification.

**How to construct a classifier? (Misclassification costs are equal)**

Bayes rule classifier \( g(x) = \arg \max_k f_k(x) \), \( k = 0, \ldots, K-1 \).

Choose the class w/ maximal posterior probability \( f_k(x) \).

Details: Suppose the pdf of the \( k \)-th population \( f_k(x | t_k) \).

Then \( g(x) = \arg \max_k p(T_k | x) \)

- \( = \arg \max_k f(x | t_k) p(T_k) \)
- class cond. class prior prob.

Two approaches

1. Assume a model for the conditional prob. of populations given \( x \) and estimate.
   - EX) Logistic regression
   - \( K = 2 \):
     - \( \logit(p(T_0 | x)) = \beta_0 + \beta^T x \)

2. Assume a model for the population conditional distribution (s) and estimate.
   - EX) Gaussian LDA \( p(x | t_k) \)

**Gaussian Linear Discriminant Analysis**

Assume population conditional distributions are \( \text{MVN} \) w/ equal variance.

\( x | t_k \sim \text{MVN}(\mu_k, \Sigma) \).

Class-specific
Diagnosing breast cancer

- Breast cancer is the 2nd largest cause of cancer deaths among women.
- Three methods of diagnosis are currently available: mammography, fine needle aspirate (FNA), and surgical biopsy.
- Surgical biopsy is the most accurate, but it is invasive, time consuming, and expensive.
Medical

\[ k = 2, \quad k = 0, \quad \frac{1}{\text{appendicitis}} \]

- A patient enters the emergency room with severe stomach pains and symptoms consistent with both food poisoning and appendicitis.
- A decision has to be made as to which illness is more likely for that patient.
- Diagnosis leads to very different treatments. Incorrect diagnosis could be fatal.
Credit

$k = 0, 1, 2$

- A bank knows from past experience that there are good customers and bad customers.
- A new customer asks for a loan.
- The bank has to decide whether or not to give the loan.
Wisconsin Diagnostic Breast Cancer Data

- A computer imaging system has been developed with the goal of developing a procedure that diagnoses FNAs with high accuracy.
- A small-gauge needle is used to extract a fluid sample (FNA) from a patient's breast lump or mass.
- FNA placed on glass slide, stained, and imaged with video microscope
Wisconsin Diagnostic Breast Cancer Data

- 30 features computed from the image
- Data set consists of 569 cases, 212 malignant, 357 benign
- Goal: develop a classification rule to separate the malignant lumps from the benign lumps

\( k = 2 \)

\( \neq 4 \quad \neq 2 \)

\( Y \quad Y \)

\( o \quad o \)

\( o \quad a \)

\( 1 \quad 1 \)
$x$ is marginally a mixture of Gaussians w/ mixing weights $p(T_k)$

$$g(x) = \arg \max_k f(x | T_k) \frac{p(T_k)}{p(T \kappa)}$$

$$= \arg \max_k \exp \left( -\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) \right) p(T_k)$$

$$= \arg \max_k -\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) + \log p_k$$

**Note:**

1. If population priors are equal, $p_k = p$ for all $k$

$$g(x) = \arg \max_k \left[ -\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) \right]$$

That is, the population whose mean is closest to $x$ in Mahalanobis distance.

$$g(x) = \arg \min_k \frac{1}{2} (x - \hat{\mu}_k)^T \Sigma^{-1} (x - \hat{\mu}_k) - \log \hat{p}_k$$

- Potentially $2K + 1$ quantities to estimate.

\[
\Sigma, \mu_0, \ldots, \mu_{K-1}, p_0, \ldots, p_{K-1}
\]

**MLE of Gaussian LDA**

$$\hat{g}(x) = \arg \min_k \frac{1}{2} \left( \hat{\Sigma}^{-1/2} (x - \hat{\mu}_k) \right)^2 - \log \hat{p}_k$$

where

\[
\hat{\Sigma} = \frac{1}{n} \sum_{k=0}^{K-1} \sum_{i \in \text{pop. k}} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T
\]

\[
\hat{\mu}_k = \bar{x}_k = \frac{1}{n_k} \sum_{i \in \text{pop. k}} x_i
\]

\[
\hat{p}_k = \frac{n_k}{n}
\]

Why is it called linear?

Allocate $x^*$ into pop. $T_k$ if

$$p_k \cdot f(x^* | T_k) = \max_j p_j \cdot f(x^* | T_j)$$
\[
\log(\Pr_k) = \frac{1}{2} (\mathbf{x}^* - \mu_k)^T \Sigma^{-1} (\mathbf{x}^* - \mu_k)
\]

\[
\alpha \log(\Pr_k) = \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \mu_k^T \Sigma^{-1} \mathbf{x}^* = \text{linear combination in } \mathbf{x}^*
\]

= discriminant score

**Evaluation of Classification Functions**

Use misclassification probabilities. \(k = 2, k = 0, 1\)

TPM (Total probability of misclassification)

\[
= P_0 P(\mathbf{x} \in R_1 | \Theta_0) + P_1 P(\mathbf{x} \in R_0 | \Theta_1)
\]

measure how the classification function will perform in the future.

The optimum error rate is the smallest value of the TPM.

\(\text{OER}\)

The OER is the error rate for the minimum TPM classification rule.

\(p_k = 1, \quad w_{01} = w_{40} = 1\)

\(R_0: \quad (\mu_0 - \mu_1)^T \Sigma^{-1} \mathbf{x}^* - \frac{1}{2} (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 + \mu_1) \geq 0\)

allocate \(\mathbf{x}^*\) into \(\Theta_0\) if

\(R_1: \quad (\mu_0 - \mu_1)^T \Sigma^{-1} \mathbf{x}^* - \frac{1}{2} (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 + \mu_1) < 0\)

\[\gamma = (\mu_0 - \mu_1)^T \Sigma^{-1} \mathbf{x}^* = \mathbf{a}^T \mathbf{x}\]

\[\Theta\]
\[ R_0: \quad a^T x \geq \frac{1}{2} (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 + \mu_1) \]
\[ R_1: \quad a^T x < \frac{1}{2} (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 + \mu_1) \]

\[
\begin{align*}
\mu_0^Y &= a^T \mu_0 \\
\sigma_0^2 &= a^T \Sigma a = (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1) \\
\mu_1^Y &= a^T \mu_1 \\
\sigma_1^2 &= (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1)
\end{align*}
\]

\[ \text{TPM} = \frac{1}{2} P(\text{misclassifying a } \pi_0 \text{ obs as } \pi_1) + \frac{1}{2} P(\text{misclassifying a } \pi_1 \text{ obs as } \pi_0) \]

\[ \Theta = P(Y \leq \frac{1}{2} (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 + \mu_1)) \quad \text{when } Y \sim N(\mu_Y^Y, \sigma_Y^2) \]

\[ = P\left( \frac{Y - \mu_Y^Y}{\sigma_Y} < \frac{1}{2} \frac{(\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 + \mu_1)}{(\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1)} \right) \]

\[ = \Phi(-\frac{\Delta}{2}) \]

\[ \Xi = P(Y > \frac{1}{2} (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 + \mu_1)) \]

\[ = P(\frac{Z}{\sigma_Y} > \frac{\Delta}{2}) = \Phi(-\frac{\Delta}{2}) \]

\[ \text{OER} = \text{minimum } \text{TRM} = \frac{1}{2} \Phi(-\frac{\Delta}{2}) + \frac{1}{2} \Phi(-\frac{\Delta}{2}) = \Phi(\frac{\Delta}{2}) \]

Sample classification functions:

\[ \text{actual error rate} = \int_{R_0} f(x | \pi_0) \ dx + \int_{R_1} f(x | \pi_1) \ dx \]

\[ \hat{R}_0 \quad \text{Unknown} \]

\[ P_0 \text{ function for } \pi_0 \]

\[ P_1 \text{ function for } \pi_1 \]
Apparent Error Rate (APER): fraction of observations in the training set that are misclassified by the sample classification

\[
P_{\text{Err}} = \frac{n_0 + n_1}{n_0 + n_1}
\]

\[
\begin{array}{c|cc}
\text{Predicted} & 0 & 1 \\
\hline
\text{True} & n_{00} & n_{01} & \cdots & n_{10} & n_{11} \\
0 & n_{00} & n_{01} & \cdots & n_{01} & n_{01} \\
1 & n_{10} & n_{11} & \cdots & n_{10} & n_{11}
\end{array}
\]

loss function: underestimate the true error rate \( \Rightarrow \) cross-validation

\[
\Sigma_0 = \Sigma_1 = \cdots = \Sigma_{k-1} = \Sigma, \\
\text{allowing} \ \Sigma_k
\]

\[
\mathbf{X}_k \sim \mathcal{N}(\mu_k, \Sigma_k)
\]

rule: allocate \( x^* \) to \( \mathbb{M}_k \) if

\[
\log P_{\mathbb{M}_k}(x^* | \mathbb{M}_k) \propto \log P_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x^* - \mu_k)^T \Sigma_k^{-1} (x^* - \mu_k)
\]

\[
= \max_j \left\{ \log P_j - \frac{1}{2} \log |\Sigma_j| - \frac{1}{2} (x^* - \mu_j)^T \Sigma_j^{-1} (x^* - \mu_j) \right\}
\]

discriminant score

\[
(= \text{quadratic in } x^*)
\]

\[
\hat{p}_j = \frac{n_j}{n}
\]

\[
\hat{\mu}_j = \bar{x}_j
\]

\[
\hat{\Sigma}_j = S_j
\]
Example: Fisher's iris data

- \( n_0 = n_1 = n_2 = 50 \) plants of \( K = 3 \) species: iris setosa, iris versicolor, and iris virginica.

- \( X = (x_{sepal \ length}, x_{sepal \ width}, x_{petal \ length}, x_{petal \ width}) \)
  \( p = 4 \)

- \( k = 0, 1, 2 \) (\( K = 3 \))

Goal:

1. Find the linear function \( y = a^T X \), or a few univariate variable linear combinations of \( x \), \( a_1^T X \), \( a_2^T X \), .... which maximizes separation.

2. Combine dimension reduction w/ classification.

3. Find projections that "explain" separation of classes.

Derivation of Fisher's LDA (No assumption on the distribution but common covariance matrix)

Law of Total Variance

\[
\text{Var}(X) = \text{Var}(E(X|\pi)) + E(\text{Var}(X|\pi))
\]

\[
= \Sigma_B + \Sigma_w
\]

In words, total var. = var( class cond. means ) + mean( class cond. variances )
Variance of Linear Combinations

\[
\text{Var}(Y) = \text{Var}(a^T X) = \text{Var}(E(a^T X | \pi)) + E(\text{Var}(a^T X | \pi))
\]

\[
= a^T \text{Var}(E(X | \pi)) a + a^T E(\text{Var}(a^T X | \pi)) a
\]

\[
= a^T \Sigma_B a + a^T \Sigma_W a
\]

1. overall mean : \( \overline{\mu} = \sum_{k=0}^{K-1} p_k \cdot \mu_k \)

2. between class : \( \Sigma_B = \sum_{k=0}^{K-1} p_k \cdot (\mu_k - \overline{\mu})(\mu_k - \overline{\mu})^T \)

3. within class : \( \Sigma_W = \sum_{k=0}^{K-1} p_k \cdot \text{Var}(X | \pi_k) \)

**Idea** : Find a linear combination that maximizes separation.

\[
\text{max}_a \frac{\text{Var}(E(a^T X | \pi))}{E(\text{Var}(a^T X | \pi))} = \text{max}_a \frac{a^T \Sigma_B a}{a^T \Sigma_W a}
\]

common variability within groups
Reparameterize \( a = \Sigma_w^{-1/2} b \)

\[
\max_a \frac{a^T \Sigma_B a}{a^T \Sigma_w a} = \max_\Sigma_w \frac{b^T \Sigma_w^{1/2} \Sigma_B \Sigma_w^{-1/2} b}{\Sigma_w^{1/2} b} = \max_u \frac{u^T \Sigma_w^{1/2} \Sigma_B \Sigma_w^{-1/2} u}{|u|^2} = 1
\]

since \( a \) can be rescaled w/o affecting the ratio.

The ratio is maximized w/ \( u = v_1 \) where

\( v_1 \) is the eigenvector of \( \Sigma_w^{-1/2} \Sigma_B \Sigma_w^{-1/2} \) associated w/ the largest eigenvalue.

\( \Rightarrow a \) is the standardized vector proportional to \( \Sigma_w^{-1/2} v_1 \)

**Why?** \( \Sigma_w^{1/2} \Sigma_B \Sigma_w^{1/2} \) and \( \Sigma_w^{-1/2} \Sigma_B \Sigma_w^{-1/2} \) have the same eigenvalues

and \( a = \Sigma_w^{-1/2} v_1 \) is an eigenvector of \( \Sigma_w^{-1} \Sigma_B \)

---

**Discriminant Score** \( a^T x^* \)

**Rule:** allocate \( x^* \) into population \( k \)

if \( |a^T x^* - a^T \mu_k| < |a^T x^* - a^T \mu_j| \), \( j = 0, \ldots, k-1 \)

\( j \neq k \)
Sample between groups matrix:
\[ \hat{\Sigma}_B = \sum_{k=0}^{K-1} n_k (\bar{X}_k - \bar{X})(\bar{X}_k - \bar{X})^T \]

Sample within group matrix:
\[ \hat{\Sigma}_W = \sum_{k=1}^{K} \sum_{i \in k} (X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)^T \]

\[ \frac{\hat{\Sigma}_W}{\hat{\Sigma}_B} \] has eigenvalues \( \hat{\lambda}_1, \ldots, \hat{\lambda}_S \) where \( S = \min(K-1,p) \) none zero.

\( \hat{V}_1, \ldots, \hat{V}_S \) are the corresponding eigenvectors (scaled such that \( \| \hat{V}_l \| = 1 \)).

The linear combination, \( \hat{\alpha}_i^T x \) is called the sample i-th discriminant.

\[ \hat{\alpha}_i^T x : i-th \ discriminant, \quad i=1,\ldots,S \]

Classification Rule

allocate \( x^* \) into \( \Pi_k \) if
\[ |\hat{\alpha}_i^T x^* - \hat{\alpha}_i^T \bar{X}_k| < |\hat{\alpha}_j^T x^* - \hat{\alpha}_j^T \bar{X}_k| \]
\[ j=0,\ldots,K-1 \]
\[ j \neq k \]

If we use \( r (\leq S) \) discriminants,

allocate \( x^* \) into \( \Pi_k \) if
\[ \sum_{i=1}^{r} |\hat{\alpha}_i^T x^* - \hat{\alpha}_i^T \bar{X}_k| \leq \sum_{j=0}^{r} |\hat{\alpha}_j^T x^* - \hat{\alpha}_j^T \bar{X}_k| \]
\[ j=0,\ldots,K-1 \]
\[ j \neq k \]

\[ = (y - \bar{y}_k)^T (y - \bar{y}_k) \]

\( p=2 \), Fisher's discriminant function is a special case of Gaussian LDF.

\( \Rightarrow p \geq 3 \), not the same.

See MKB 11.5