Discriminant Analysis

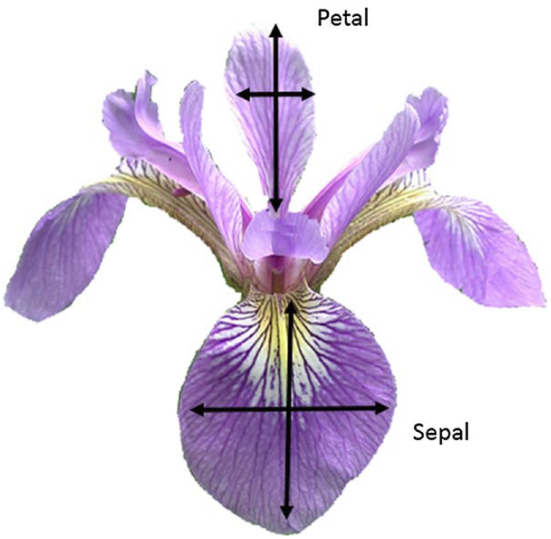
Ngoc Hoang Luong

University of Information Technology (UIT), VNU-HCM

May 27, 2024

Motivation

The **Iris dataset** is a collection of 150 labeled examples of Iris flowers, 50 of each type, described by 4 features: sepal length, sepal width, petal length, petal width.

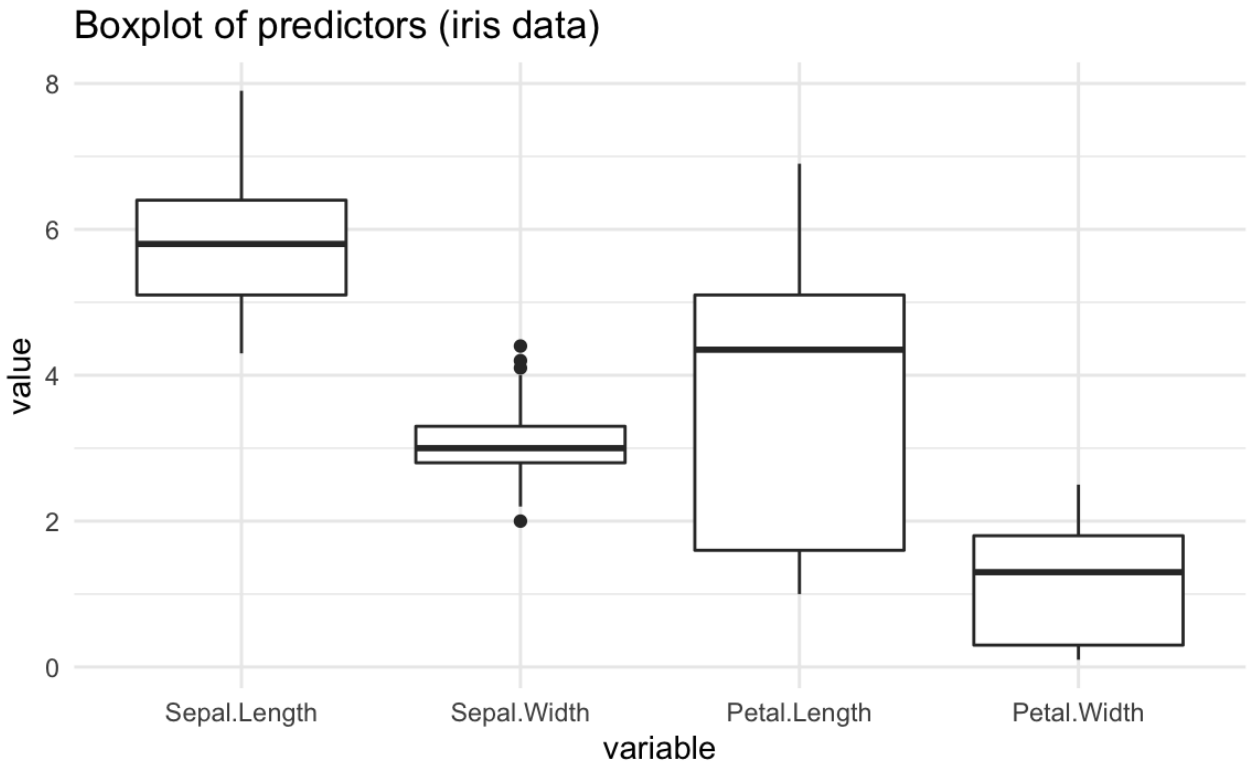


index	sl	sw	pl	pw	label
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
	...				
50	7.0	3.2	4.7	1.4	Versicolor
	...				
149	5.9	3.0	5.1	1.8	Virginica

Figure: A subset of the Iris design matrix.

Motivation

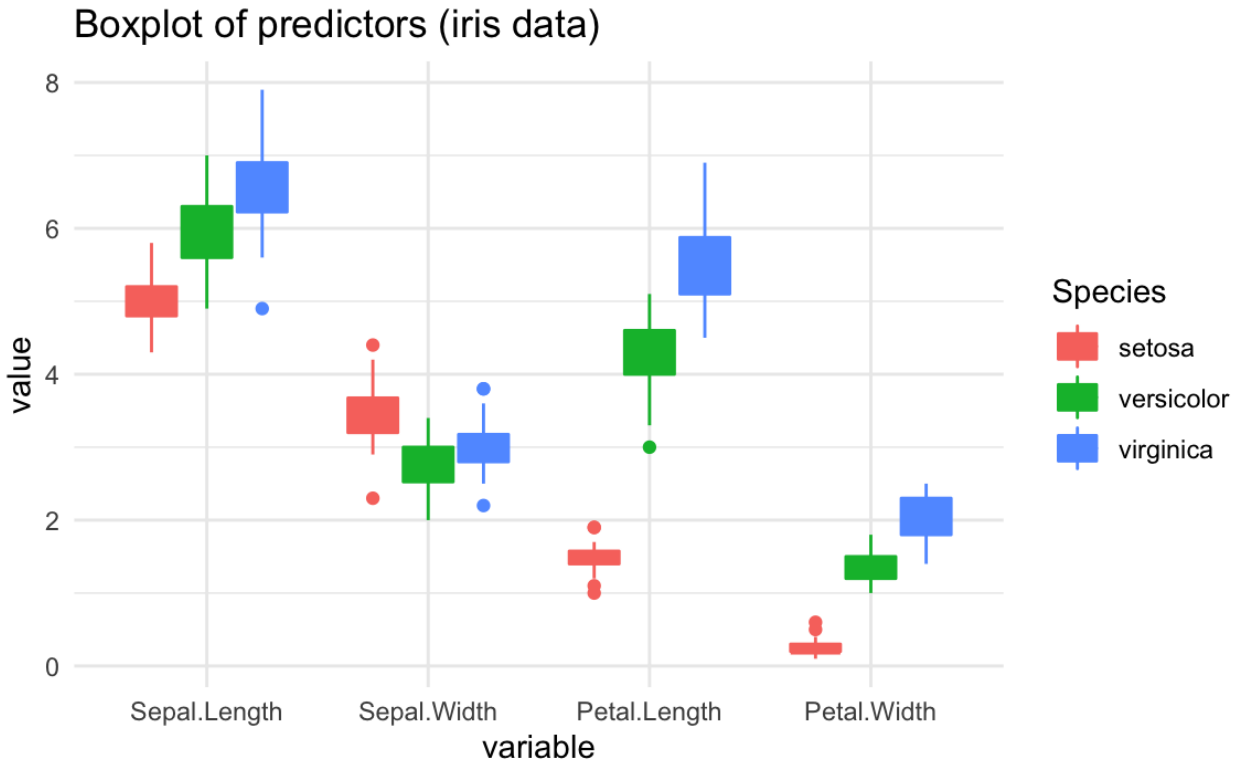
- Let's look at the distribution of the four input variables without making distinction between species of iris flowers:



- All predictors have different ranges, as well as different types of distributions.

Motivation

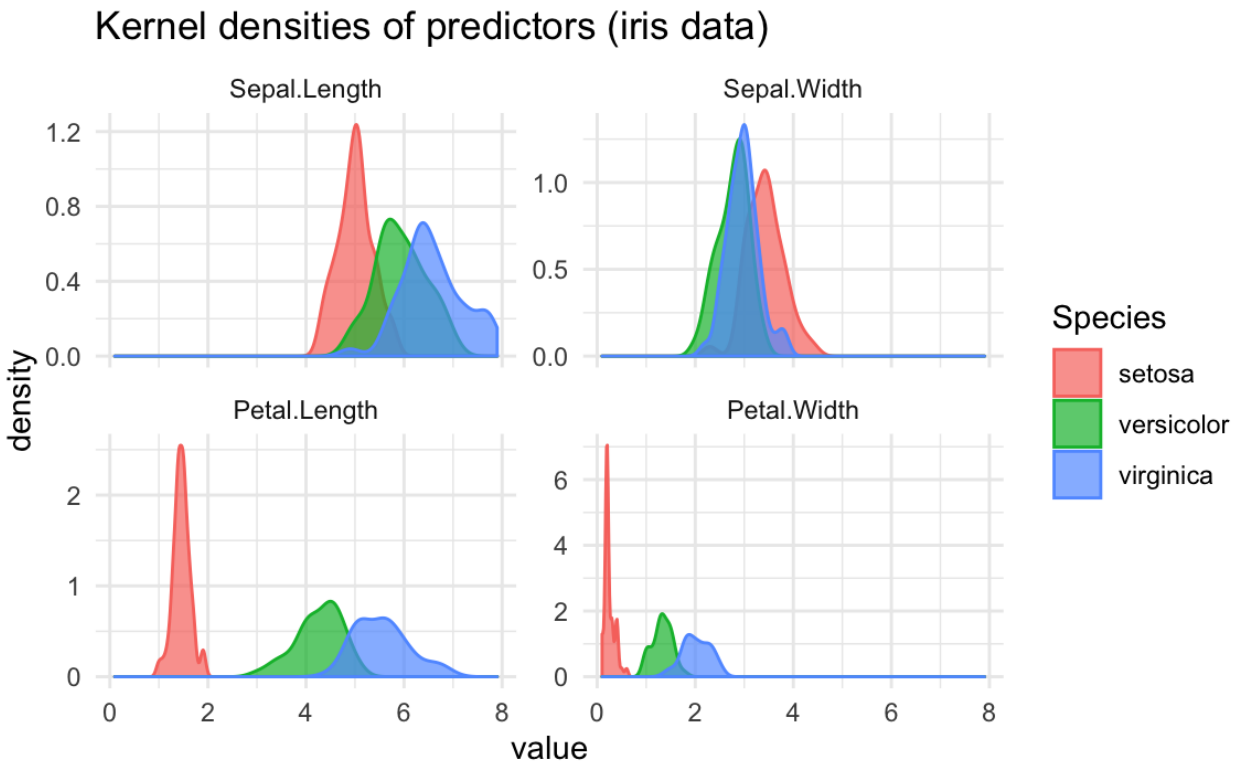
- The boxplot of some predictors are fairly different between iris species.



- The boxplots of petal length and petal width are different between species.
- In contrast, predictors like sepal length and sepal width haves similar boxplot between species.

Motivation

- The same differences can be seen for density curves.



Distinguishing Species

- Which predictor provides the "best" distinction between species?
- In classification problem, the response variable Y provides a group or class structure to the data.
- We expect that the predictors will help us to discriminate between one class and the others.
- The general idea is to look for systematic differences among classes.

Distinguishing Species

- Let's consider a single predictor X and a categorical response Y measured on n individuals.
- Assume there are K classes. Let C_k represents the k-th class in Y. Let n_k be the number of observations in class C_k. Then

$$n = n_1 + n_2 + \ldots + n_K = \sum_{k=1}^{K} n_k$$

- The global mean value of X is:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Each class k will have its means $\bar{x}_k$:

$$\bar{x}_k = \frac{1}{n_k} \sum_{i \in C_k} x_{ik}$$

Distinguishing Species

- A measure of global dispersion in X is the total sum of squares:

$$\text{tss} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Each class k has its own sum-of-squares ss_k:

$$\text{ss}_k = \sum_{i \in C_k} (x_{ik} - \bar{x}_k)^2$$

- If there is no group difference in X, then the group means $\bar{x}_k$ should be similar. If there is really a difference, it is likely that one or more of the mean values will differ.
- A useful measure to compare differences among the k means is the deviation from the overall mean: $\bar{x}_k - \bar{x}$

Distinguishing Species

- To summarize the deviations of each mean to overall mean, the between-group sum of squares is:

$$\text{bss} = \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2$$

- To summarize group variances, the within-group sum of squares is:

$$\text{wss} = \sum_{i=1}^{K} \text{ss}_k = \sum_{i=1}^{K} \sum_{i \in C_k} (x_{ik} - \bar{x}_k)^2$$

- We have three types of sums of squares:

$$\text{total} \quad \text{tss} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\text{between} \quad \text{bss} = \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2$$

$$\text{within} \quad \text{wss} = \sum_{k=1}^{K} \sum_{i \in C} n_k (x_{ik} - \bar{x}_k)^2$$

Sum of Squares Decomposition

- We want to decompose the squared deviations $(x_i - \bar{x})^2$ in terms of the class structure:

$$\begin{aligned} x_i - \bar{x} &= x_i - \bar{x}_k + \bar{x}_k - x \\ &= (\bar{x}_k - \bar{x}) + (x_i - \bar{x}_k) \end{aligned}$$

- We can decompose tss in terms of bss and wss as follows:

$$\underbrace{\sum_{k=1}^{K} \sum_{i \in C_k} (x_{ik} - \bar{x})^2}_{tss} = \underbrace{\sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2}_{bss} + \underbrace{\sum_{k=1}^{K} \sum_{i \in C_k} (x_{ik} - \bar{x}_k)^2}_{wss}$$

- In summary:

$$\text{tss} = \text{bss} + \text{wss}$$

Correlation Ratio

- The correlation ratio is a measure of the relationship between the dispersion within groups and the dispersion across all individuals.
- Correlation ration η^2 (proposed by Karl Pearson) is:

$$\eta^2(X, Y) = \frac{bss}{tss}$$

- η^2 takes values between 0 and 1.
- $\eta^2 = 0$ is the special case of no dispersion among the means of the different groups.
- $\eta^2 = 1$ refers to no dispersion within the respective groups.

F-Ratio

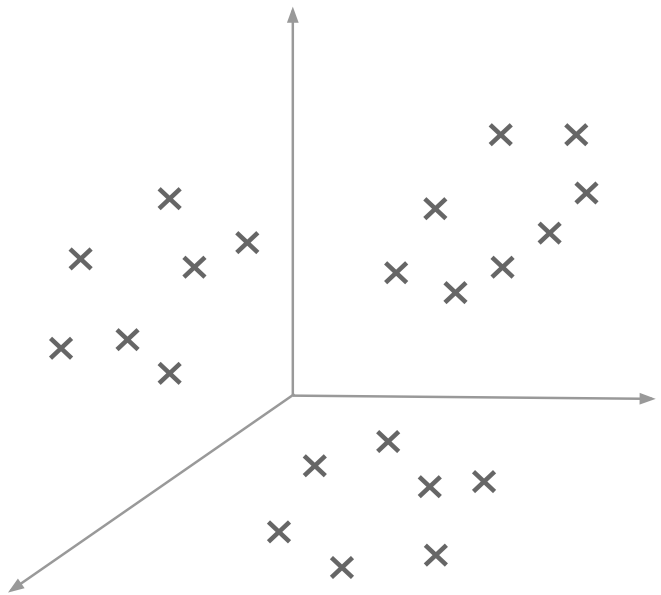
- The F-ratio (proposed by R.A. Fisher) is:

$$F = \frac{bss/(K-1)}{wss(n-K)}$$

- The larger the value of both ratios, the more variability there is between groups than within groups.

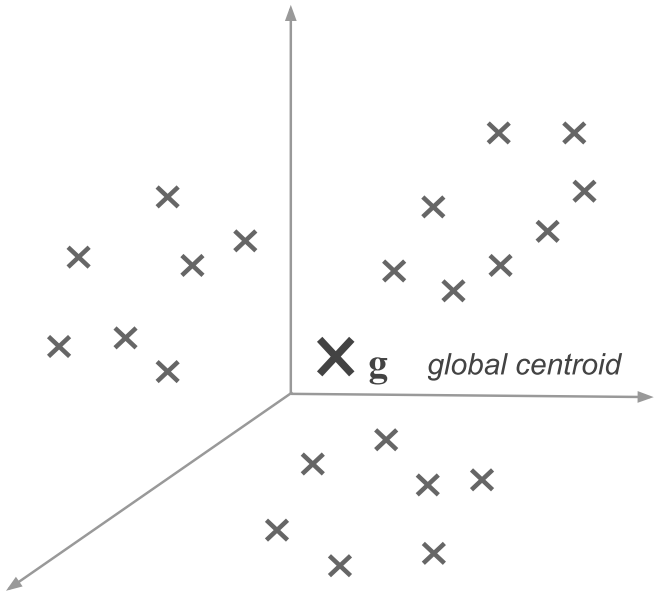
Geometric Perspective

- Assume that the objects form a cloud of points in p-dimensional space.



Geometric Perspective

- Consider the average individual $\mathbf{g}$, known as the global centroid (i.e., the center of gravity of the cloud of points):



- The global centroid is the point of averages which consists of the point formed with all the variable means: $\mathbf{g} = [x_1, x_2, \ldots, x_p]$ where

$$x_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

Geometric Perspective

- If all variables are mean-centered, the centroid is the origin $\mathbf{g} = [0, 0, \ldots, 0]$.
- Taking the global centroid as a point of reference, we compute the amount of spread or dispersion in the data.
- Assuming the centered features, a matrix of total dispersion is given by the Total Sums of Squares (TSS):

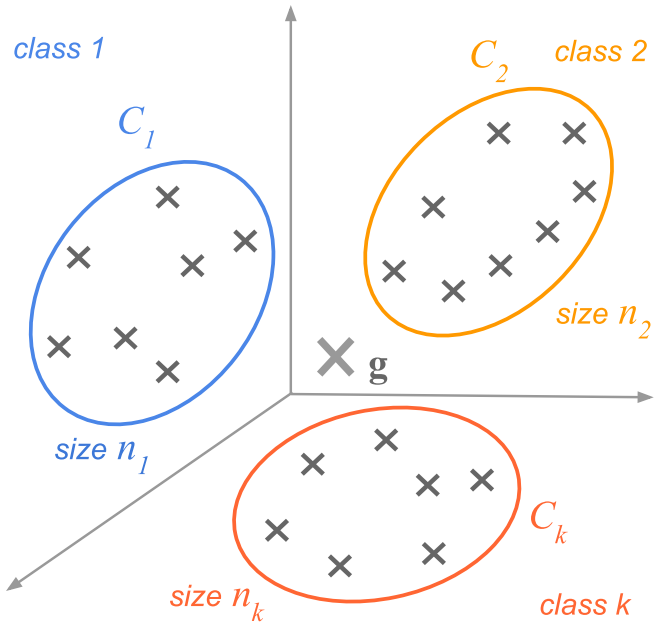
$$\text{TSS} = \mathbf{X}^\top \mathbf{X}$$

- Alternatively, we can get the sample variance-covariance matrix $\mathbf{V}$:

$$\mathbf{V} = \frac{1}{n-1} \mathbf{X}^\top \mathbf{X}$$

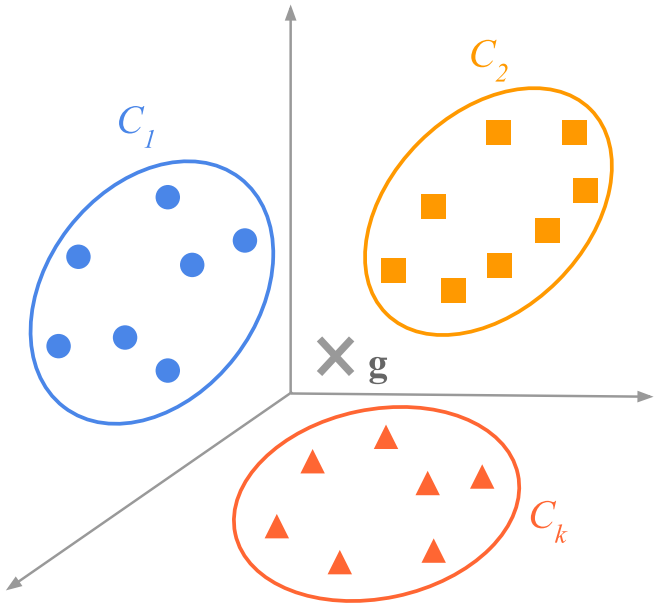
Geometric Perspective

- n_k: the number of observations in the k-th class.
- x_{ijk}: the i-th observation, the j-th variable, in the k-th class.
- x_{ik}: the i-th observation in the k-th class.
- x_{jk}: the j-th variable in the k-th class.

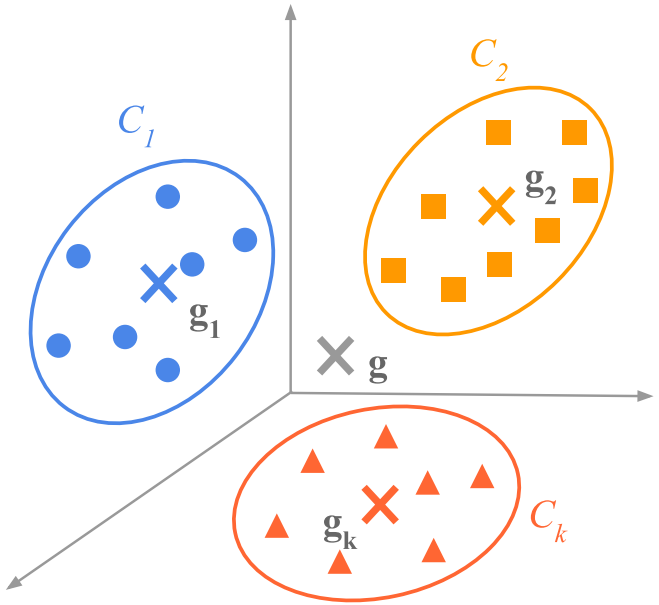


Geometric Perspective

Each class is denoted as C_k and is formed by n_k individual.



Geometric Perspective

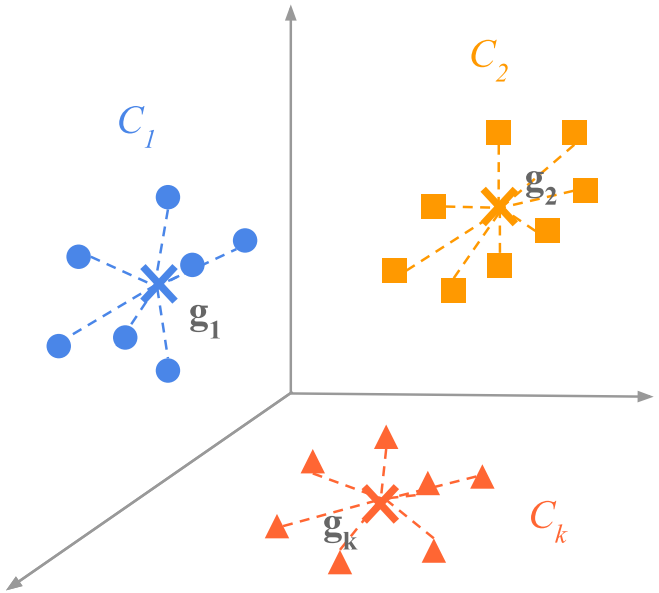


- We can look at the local or class centroids (one per class).
- The class centroid $\mathbf{g}_k$ is the point of averages for those observations in class k: $\mathbf{g}_k = [\bar{x}_{1k}, \bar{x}_{2k}, \ldots, \bar{x}_{pk}]$ where:

$$\bar{x}_{jk} = \frac{1}{n_k} \sum_{i \in C_k} x_{ij}$$

Geometric Perspective

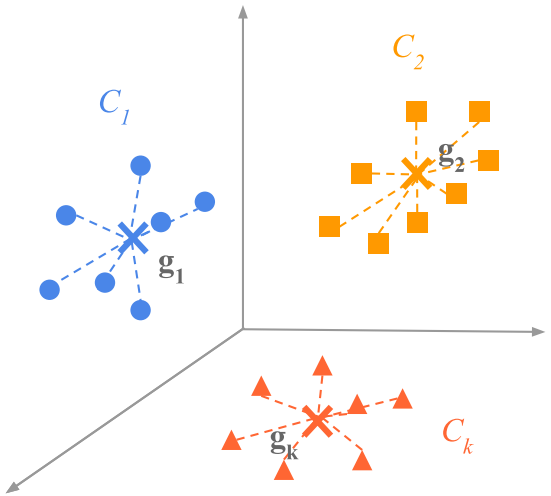
- We consider the dispersion within the clouds.



- Each group will have an associated spread or dispersion matrix given by a Class Sums of Squares (CSS): $CSS_k = \mathbf{X}^\top \mathbf{X}$
- There is an associated variance matrix $\mathbf{W}_k$ for each class:

$$\mathbf{W}_k = \frac{1}{n_k - 1} \mathbf{X}_k^\top \mathbf{X}_k$$

Geometric Perspective



- We can combine the class dispersion to obtain a Within-class Sums of Squares (WSS) matrix:

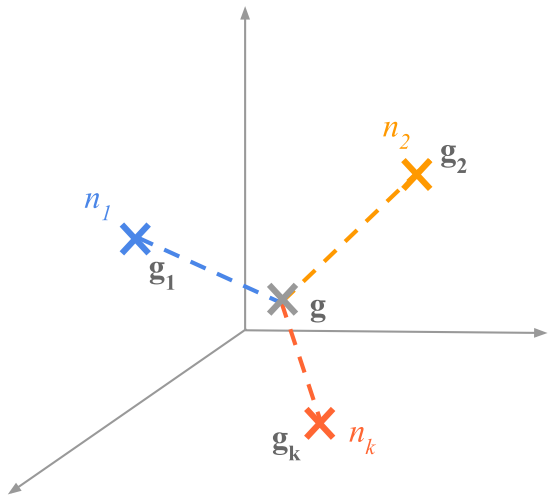
$$WSS = \sum_{k=1}^{K} CSS_k = \sum_{k=1}^{K} \mathbf{X}_k^\top \mathbf{X}_k$$

- We can combine the class variances $\mathbf{W}_k$ as a weighted average to get the Within-class variance matrix $\mathbf{W}$:

$$\mathbf{W} = \sum_{k=1}^{K} \frac{n_k - 1}{n - 1} \mathbf{W}_k$$

Geometric Perspective

- If we just focus on the centroids

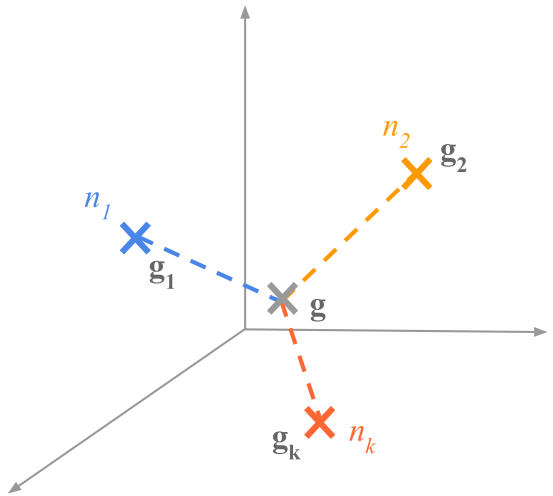


- The global centroid $\mathbf{g}$ can be expressed as a weighted average of the group centroids

$$\mathbf{g} = \frac{n_1}{n}\mathbf{g}_1 + \frac{n_2}{n}\mathbf{g}_2 + \cdots + \frac{n_K}{n}\mathbf{g}_K = \sum_{k=1}^{K} \frac{n_k}{n}\mathbf{g}_k$$

Geometric Perspective

- If we just focus on the centroids



- The matrix of dispersion Between Sum of Squares (BSS):

$$\text{BSS} = \sum_{k=1}^{K} n_k (\mathbf{g}_k - \mathbf{g})(\mathbf{g}_k - \mathbf{g})^\top$$

- The Between-Class Variance Matrix $\mathbf{B}$:

$$\mathbf{B} = \sum_{k=1}^{K} \frac{n_k}{n-1} (\mathbf{g}_k - \mathbf{g})(\mathbf{g}_k - \mathbf{g})^\top$$

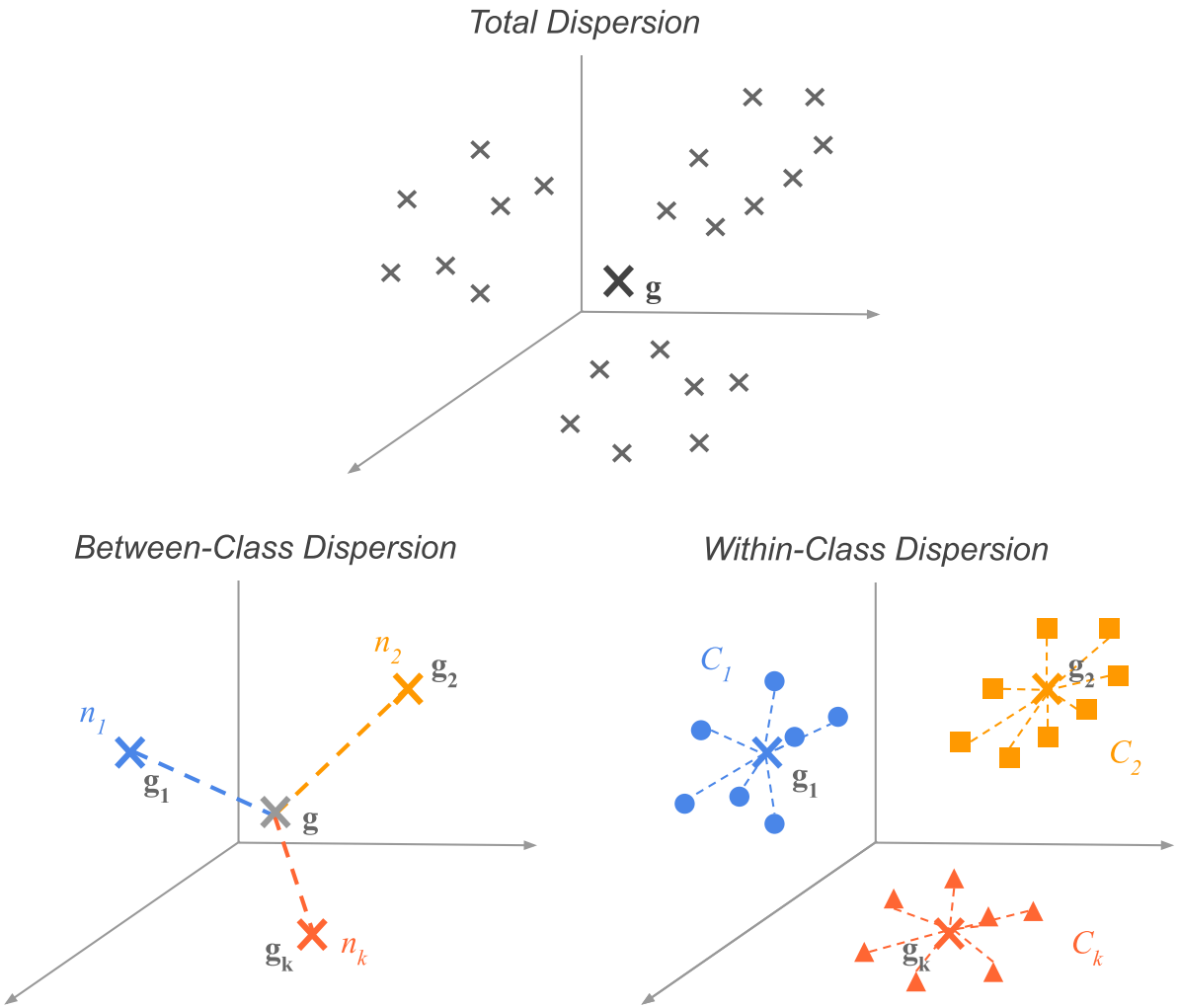
Geometric Perspective

- We have three types of Sum of Squares matrices:
  1. TSS: Total Sum of Squares
  2. WSS: Within-class Sum of Squares
  3. BSS: Between-class Sum of Squares
- We have three types of variance matrices:
  1. $\mathbf{V}$: Total variance
  2. $\mathbf{W}$: Within-class variance
  3. $\mathbf{B}$: Between-class variance

Dispersion Decomposition

- It can be shown, for both sums of squares and variances, the total dispersion (i.e. TSS or $\mathbf{V}$) can be decomposed as:
  - TSS = BSS + WSS
  - $\mathbf{V} = \mathbf{B} + \mathbf{W}$
- Let $\mathbf{X}$ be the $n \times p$ mean-centered matrix of predictors and $\mathbf{Y}$ be the $n \times K$ matrix of classes.
  - TSS = $\mathbf{X}^\top \mathbf{X}$
  - BSS = $\mathbf{X}^\top \mathbf{Y} (\mathbf{Y}^\top \mathbf{Y})^{-1} \mathbf{Y}^\top \mathbf{X}$
  - WSS = $\mathbf{X}^\top (\mathbf{I} - \mathbf{Y} (\mathbf{Y}^\top \mathbf{Y})^{-1} \mathbf{Y}^\top) \mathbf{X}$

Dispersion Decomposition

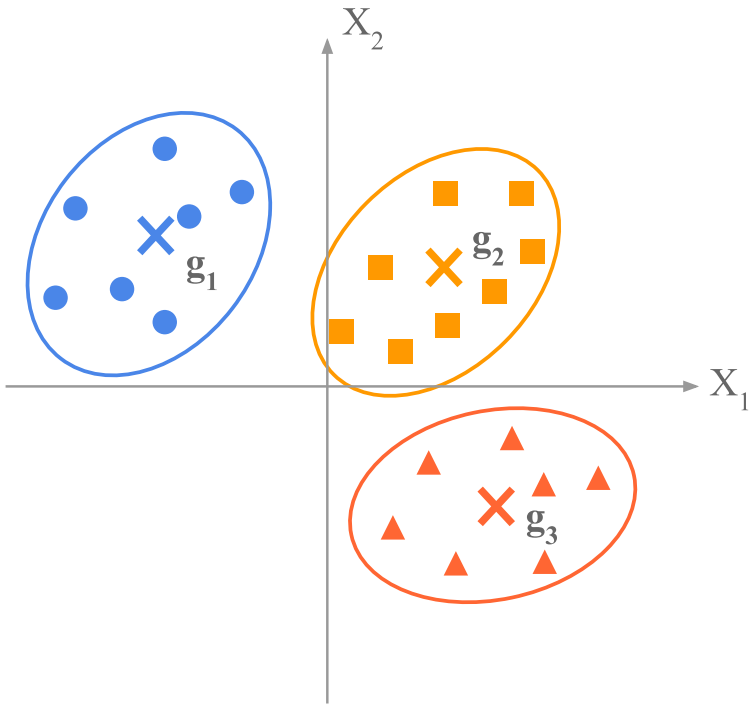


Canonical Discriminant Analysis (CDA)

- **Canonical Discriminant Analysis (CDA)** is a special case of Linear Discriminant Analysis (LDA).
- CDA has two aspects:
  - Semin-supervised: How to find a representation of the objects which provides the best separation between classes?
  - Supervised: How to find the rules for assigning a class to a given object?

CDA: Semi-Supervised Aspect

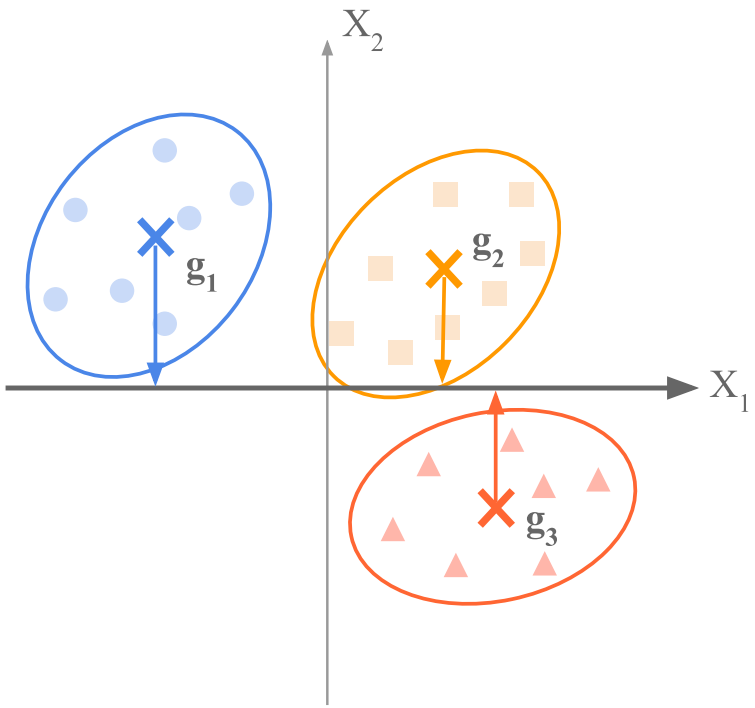
- We formulate the classification problem of CDA in a geometric way.
- For example, let's consider three classes in a 2D space as below.



- From an exploratory/descriptive perspective, we look for a good low dimensional representation that separates the three classes.

CDA: Semi-Supervised Aspect

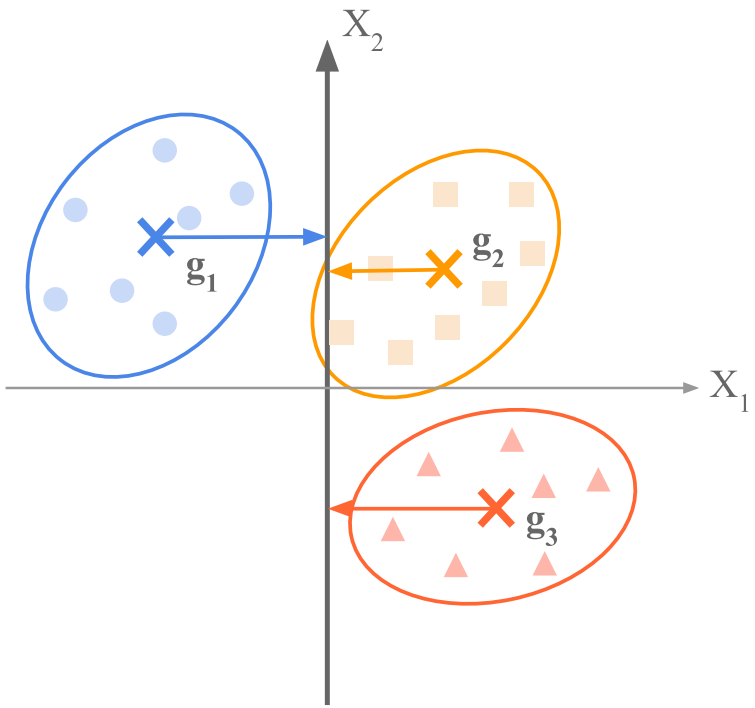
- We consider the axis associated with the predictor X_1, i.e., the horizontal axis.



- If we project the individuals on the X_1 axis, class 1 is separated from classes 2 and 3.
- However, class 2 is largely confounded with class 3.

CDA: Semi-Supervised Aspect

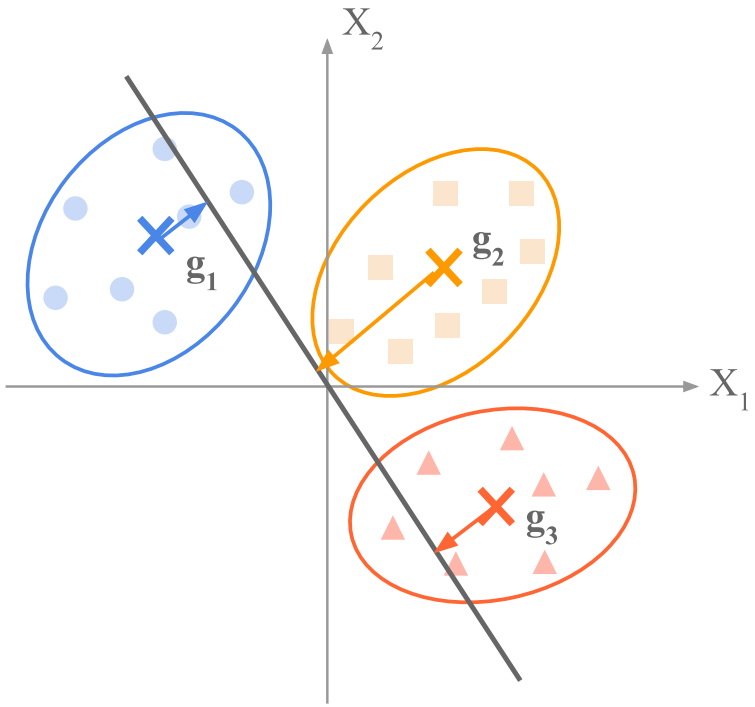
- We consider the axis associated with the predictor X_2, i.e., the vertical axis.



- If we project the individuals on the X_2 axis, class 3 is separated from classes 1 and 2. However, class 1 is largely confounded with class 2.
- Is there an axis that "best" separates the three clouds?

CDA: Semi-Supervised Aspect

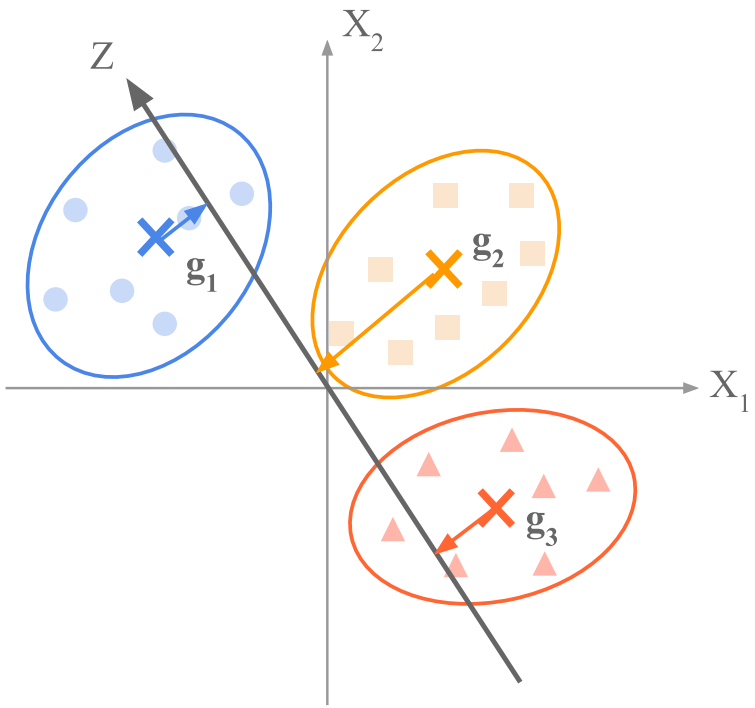
- We look for an optimal representation in the sense of finding an axis that best separates the clouds:



- The exploratory aspect of CDA involves seeking a low dimensional representation in which the class of objects are well separated.

Looking for a discriminant axis

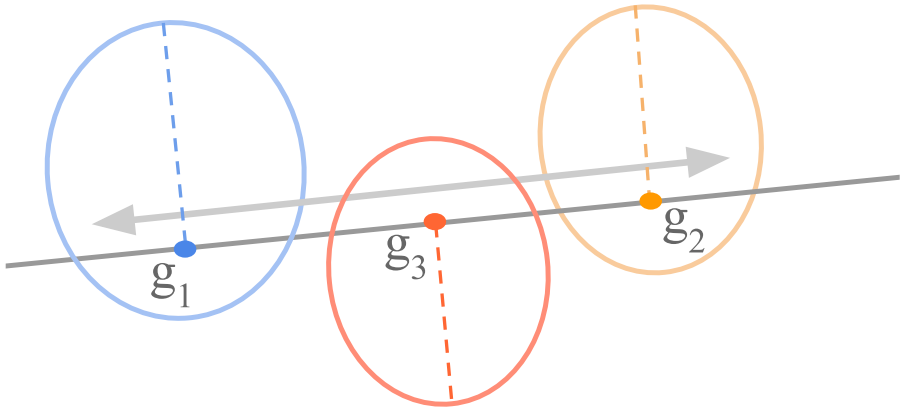
- We look for an axis Δ_u, spanned by some vector $\mathbf{u}$, separates all three groups in an adequate way.



- We look for a linear combination of the predictors $\mathbf{z} = \mathbf{X}\mathbf{u}$ that *ideally* could achieve the following goals:
  - Minimize within-class dispersion (wss): $\min\{\mathbf{u}^\top \mathbf{W}\mathbf{u}\}$
  - Maximize between-class dispersion (bss): $\max\{\mathbf{u}^\top \mathbf{B}\mathbf{u}\}$

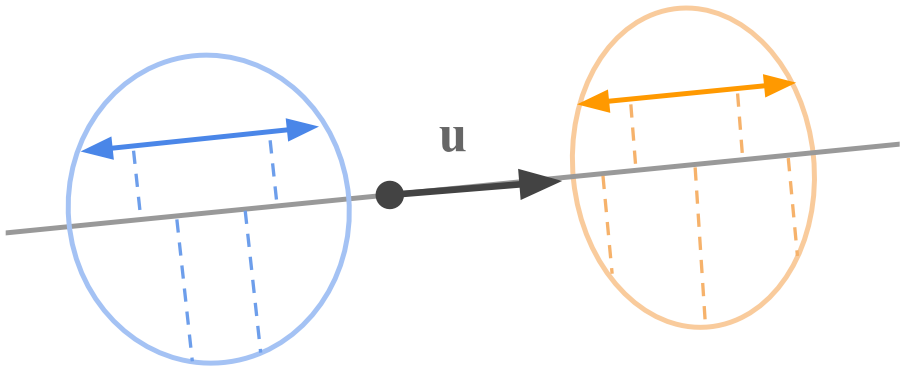
Looking for a discriminant axis

- We want to have $\mathbf{u}$ such that between class dispersion is maximized.
- This is a situation in which the class centroids are well separated:



Looking for a discriminant axis

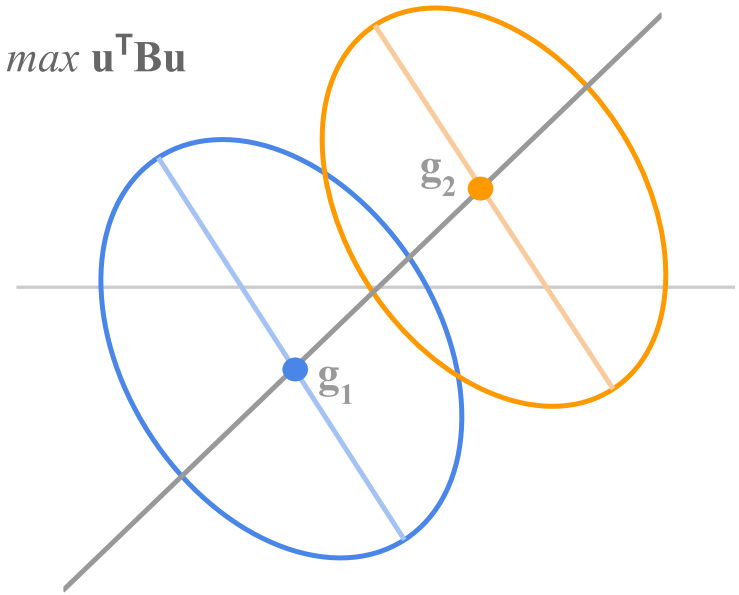
- We also want to have $\mathbf{u}$ such that the within-class dispersion is minimized.
- This implies having classes in which, on average, the "inner" variation is small (i.e., concentrated local dispersion).



- It is generally impossible to find an axis Δ_u, spanned by $\mathbf{u}$, which simultaneously maximizes the between-group variance and minimizes the within-group variance.

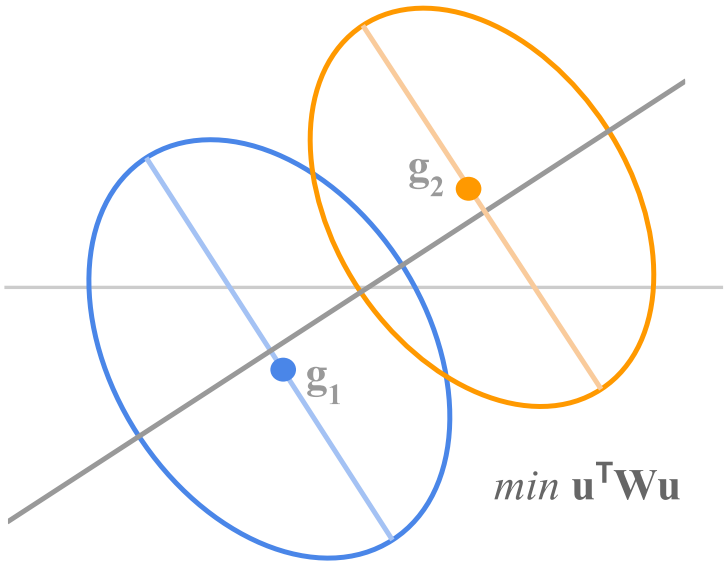
Looking for a discriminant axis

- For example, we have two classes as below. Maximize the between-class separation involves choosing $\mathbf{u}_a$ parallel to the segment linking the centroids.
- The direction of $\mathbf{u}$ crosses the centroids.



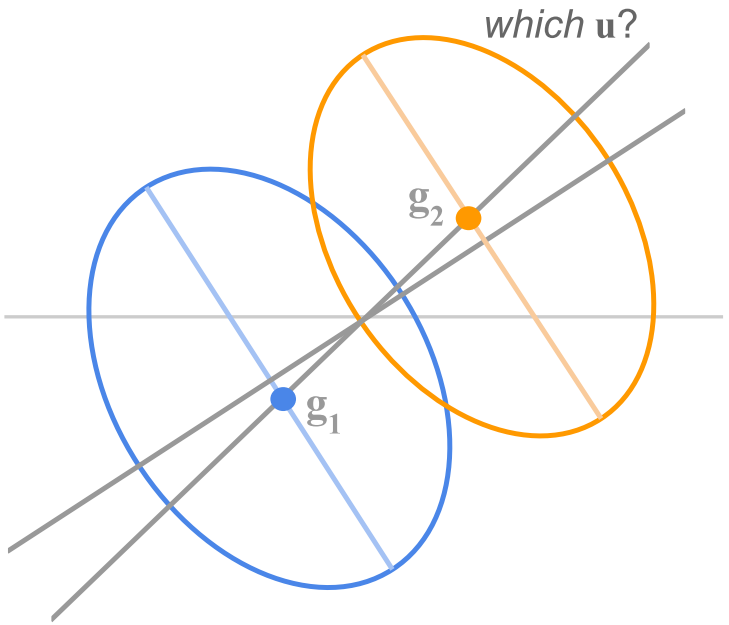
Looking for a discriminant axis

- Minimizing the within-class separation involves finding $\mathbf{u}_b$ perpendicular to the principal axis of the ellipses.
- This type of $\mathbf{u}$ does not necessarily crosses through the centroids.



Looking for a discriminant axis

- In general, we end up with two possibilities $\mathbf{u}_a \neq \mathbf{u}_b$



Looking for a discriminant axis - A compromise criterion

- It is generally impossible to minimize $\mathbf{u}^\top \mathbf{W}\mathbf{u}$ and maximize $\mathbf{u}^\top \mathbf{B}\mathbf{u}$ simultaneously. We look for a compromise.
- Let's consider the variance decomposition.

$$\mathbf{V} = \mathbf{W} + \mathbf{B}$$

$$\mathbf{u}^\top \mathbf{V}\mathbf{u} = \underbrace{\mathbf{u}^\top \mathbf{W}\mathbf{u}}_{minimize} + \underbrace{\mathbf{u}^\top \mathbf{B}\mathbf{u}}_{maximize}$$

- We have two options:

$$\max\left\{\frac{\mathbf{u}^\top \mathbf{B}\mathbf{u}}{\mathbf{u}^\top \mathbf{V}\mathbf{u}}\right\} \qquad \max\left\{\frac{\mathbf{u}^\top \mathbf{B}\mathbf{u}}{\mathbf{u}^\top \mathbf{W}\mathbf{u}}\right\}$$

  which corresponds to the correlation ratio η^2 and the F-ratio:

$$\eta^2 \longleftrightarrow \frac{\mathbf{u}^\top \mathbf{B}\mathbf{u}}{\mathbf{u}^\top \mathbf{V}\mathbf{u}} \qquad F \longleftrightarrow \frac{\mathbf{u}^\top \mathbf{B}\mathbf{u}}{\mathbf{u}^\top \mathbf{W}\mathbf{u}}$$

- Both are equivalent.

Correlation Ratio Criterion

- Let's look for $\mathbf{u}$ such that $\max\left\{\frac{\mathbf{u}^\top\mathbf{B}\mathbf{u}}{\mathbf{u}^\top\mathbf{V}\mathbf{u}}\right\}$.
- For convenience, we use a normalizing restriction $\mathbf{u}^\top\mathbf{V}\mathbf{u} = 1$.

$$\max\left\{\frac{\mathbf{u}^\top\mathbf{B}\mathbf{u}}{\mathbf{u}^\top\mathbf{V}\mathbf{u}}\right\} \Longleftrightarrow \max\{\mathbf{u}^\top\mathbf{B}\mathbf{u}\} \text{ s.t. } \mathbf{u}^\top\mathbf{V}\mathbf{u} = 1$$

- Using the method of Lagrangian multiplier:

$$L(\mathbf{u}) = \mathbf{u}^\top\mathbf{B}\mathbf{u} - \lambda(\mathbf{u}^\top\mathbf{V}\mathbf{u} - 1)$$

- Deriving w.r.t. $\mathbf{u}$ and equating to zero:

$$\frac{\partial L(\mathbf{u})}{\partial \mathbf{u}} = 2\mathbf{B}\mathbf{u} - 2\lambda\mathbf{V}\mathbf{u} = \mathbf{0}$$

$$\mathbf{B}\mathbf{u} = \lambda\mathbf{V}\mathbf{u}$$

$$\mathbf{V}^{-1}\mathbf{B}\mathbf{u} = \lambda\mathbf{u} \qquad \text{if } \mathbf{V} \text{ is invertible.}$$

- The optimal vector $\mathbf{u}$ is eigenvector of $\mathbf{V}^{-1}\mathbf{B}$.

F-Ratio Criterion

- Similarly, let's look for $\mathbf{u}$ such that:

$$\max\left\{\frac{\mathbf{u}^\top\mathbf{B}\mathbf{u}}{\mathbf{u}^\top\mathbf{W}\mathbf{u}}\right\} \Longleftrightarrow \max\{\mathbf{u}^\top\mathbf{B}\mathbf{u}\} \text{ s.t. } \mathbf{u}^\top\mathbf{W}\mathbf{u} = 1$$

- Using the same Lagrangian procedure with a multiplier ρ, we have:

$$\mathbf{B}\mathbf{u} = \rho\mathbf{W}\mathbf{u}$$

and if $\mathbf{W}$ is invertible, then it can be shown that $\mathbf{u}$ is also eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ associated with eigenvalue ρ:

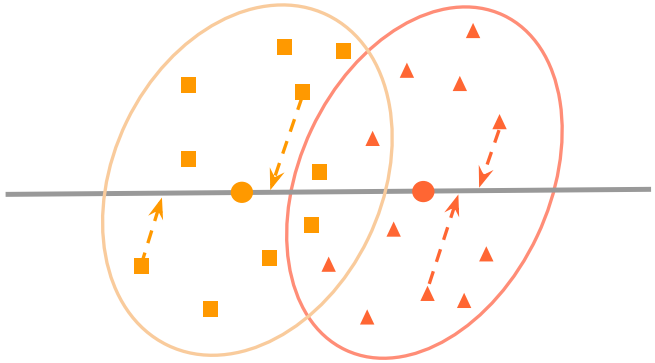
$$\mathbf{W}^{-1}\mathbf{B}\mathbf{u} = \rho\mathbf{u}$$

- The relationship between the eigenvalues λ and ρ is:

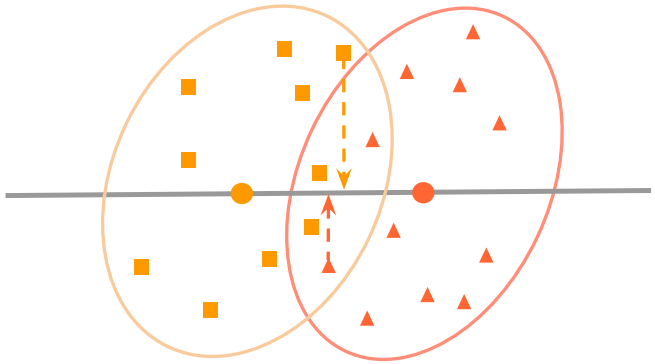
$$\rho = \frac{\lambda}{1-\lambda}$$

A Special PCA

- The vector $\mathbf{u}$ is the axis from the PCA on the clouds of centroids $\mathbf{g}_1, \ldots, \mathbf{g}_K$. The points are projected obliquely, not orthogonally.



- Without this obliqueness ($\mathbf{V}^{-1}$ and $\mathbf{W}^{-1}$), this would be a simple PCA performed on the centroids: the classes would be less well-separated because of an orthogonal projection:



CDA: Supervised Aspect

- How do we use CDA for classification purposes?
- This involves establishing a decision rule that lets us predict the class of an object.

Distance behind CDA

- Consider two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$. The inner product of $\mathbf{a}$ and $\mathbf{b}$ is $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \mathbf{b} = \mathbf{a}^\top \mathbf{I}_p \mathbf{b}$. where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix.
- For a symmetric $\mathbf{M}$ (called the **metric matrix**), we define the inner product of $\mathbf{a}$ and $\mathbf{b}$, under metric $\mathbf{M}$, to be:

$$\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{M}} = \mathbf{a}^\top \mathbf{M} \mathbf{b}$$

- The squared Euclidean distance between two vectors $\mathbf{x}_i$ and $\mathbf{x}_l$:

$$\begin{aligned} d^2(i,l) &= \langle \mathbf{x}_i - \mathbf{x}_l, \mathbf{x}_i - \mathbf{x}_l \rangle \\ &= (\mathbf{x}_i - \mathbf{x}_l)^\top (\mathbf{x}_i - \mathbf{x}_l) \\ &= (\mathbf{x}_i - \mathbf{x}_l)^\top \mathbf{I}_p (\mathbf{x}_i - \mathbf{x}_l) \end{aligned}$$

- Similarly, we replace $\mathbf{I}_p$ with any metric matrix $\mathbf{M}$ to obtain a generalized distance metric:

$$d^2_{\mathbf{M}}(i,l) = (\mathbf{x}_i - \mathbf{x}_l)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_l)$$

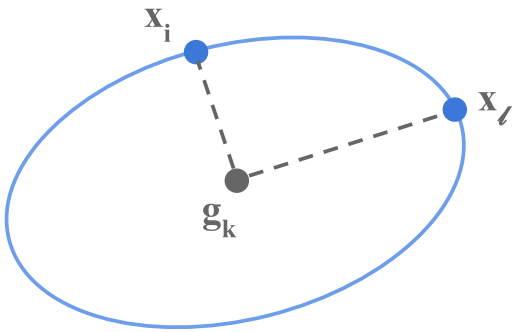
Distance behind CDA

- The classification rule used in CDA consists of assigning each individual $\mathbf{x}_i$ to the class C_k for which the distance to the centroid $\mathbf{g}_k$ is minimal.
- But we don't use the Euclidean distance, but the **Mahalanobis** distance. This is based on the Mahalanobis metric matrix $\mathbf{W}^{-1}$. The formula of the (squared) distance is:

$$\text{Mahalanobis:} \quad d^2(\mathbf{x}_i, \mathbf{g}_k) = (\mathbf{x}_i - \mathbf{g}_k)^\top \mathbf{W}^{-1} (\mathbf{x}_i - \mathbf{g}_k)$$

- The Mahalanobis distance measures the (squared) distance of a point $\mathbf{x}_i$ to a centroid $\mathbf{g})k$ by taking into account the correlational structure of the variables.

Distance behind CDA



Euclidean distance

$d^2(\mathbf{x}_i, \mathbf{g}_k) \neq d^2(\mathbf{x}_\ell, \mathbf{g}_k)$

$\mathbf{x}_i$ and $\mathbf{x}_\ell$ are at different distances to centroid $\mathbf{g}_k$

Mahalanobis distance

$d^2(\mathbf{x}_i, \mathbf{g}_k) = d^2(\mathbf{x}_\ell, \mathbf{g}_k)$

$\mathbf{x}_i$ and $\mathbf{x}_\ell$ are equidistant to centroid $\mathbf{g}_k$

- The distance of a point $\mathbf{x}_i$ to the centroid $\mathbf{g}_k$ depends on how spread out is the clouds of points in class k.
- If two points i and l are on the same density ellipsoid, they are equidistant to the centroid:

$$d^2_{\mathbf{W}^{-1}}(\mathbf{x}_i, \mathbf{g}_k) = d^2_{\mathbf{W}^{-1}}(\mathbf{x}_l, \mathbf{g}_k)$$

CDA Classifier

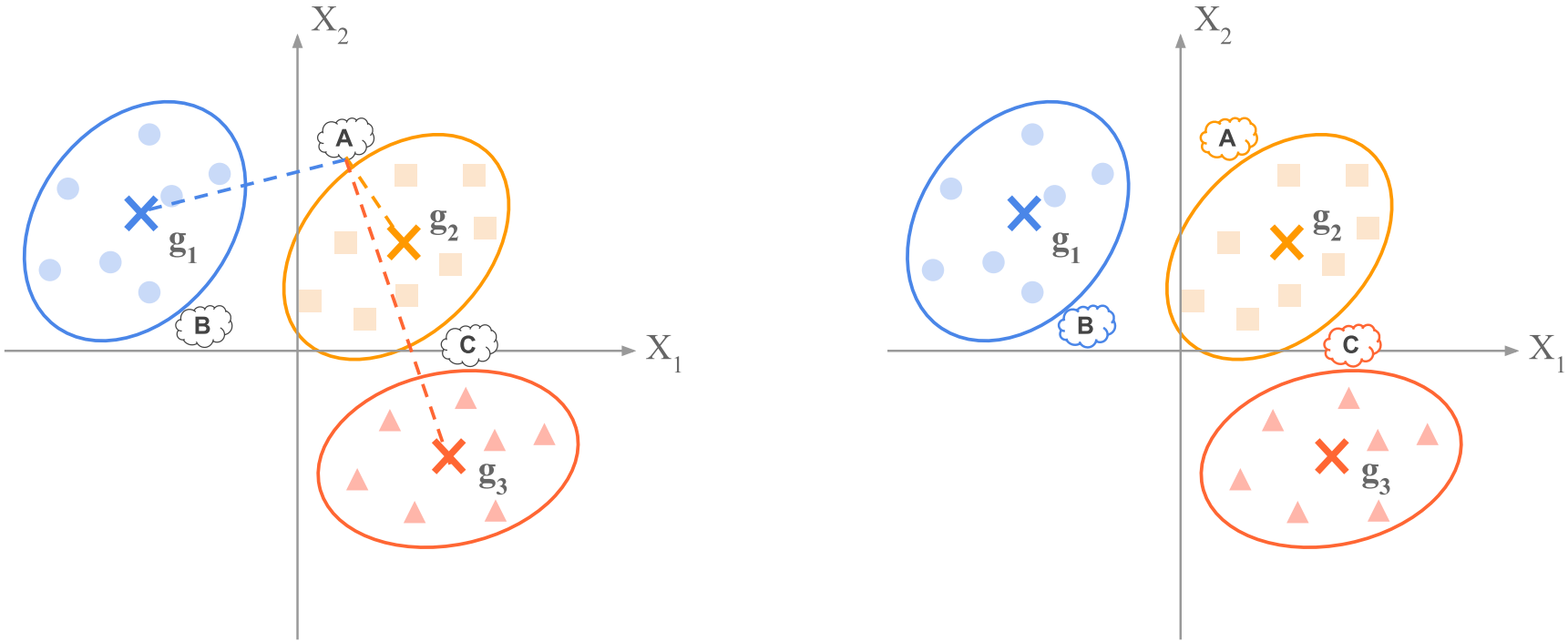
- The classification rule behind CDA is that we should assign an object $\mathbf{x}_i$ to the class it is nearest to, using the $\mathbf{W}^{-1}$ metric matrix to calculate the distance of the object from the centroid of the group.
- The Mahalanobis distance of $\mathbf{x}_i$ to centroid $\mathbf{g}_k$:

$$\begin{aligned} d^2(\mathbf{x}_i, \mathbf{g}_k) &= (\mathbf{x}_i - \mathbf{g}_k)^\top \mathbf{W}^{-1} (\mathbf{x}_i - \mathbf{g}_k) \\ &= \underbrace{\mathbf{x}_i^\top \mathbf{W}^{-1} \mathbf{x}_i}_{constant} - \underbrace{2\mathbf{g}_k^\top \mathbf{W}^{-1} \mathbf{x}_i}_{\text{depends on } k} + \underbrace{\mathbf{g}_k^\top \mathbf{W}^{-1} \mathbf{g}_k}_{\text{depends on } k} \end{aligned}$$

- Minimizing $d^2(\mathbf{x}_i, \mathbf{g}_k)$ is equivalent to maximizing:

$$2\mathbf{g}_k^\top \mathbf{W}^{-1} \mathbf{x}_i - \mathbf{g}_k^\top \mathbf{W}^{-1} \mathbf{g}_k$$

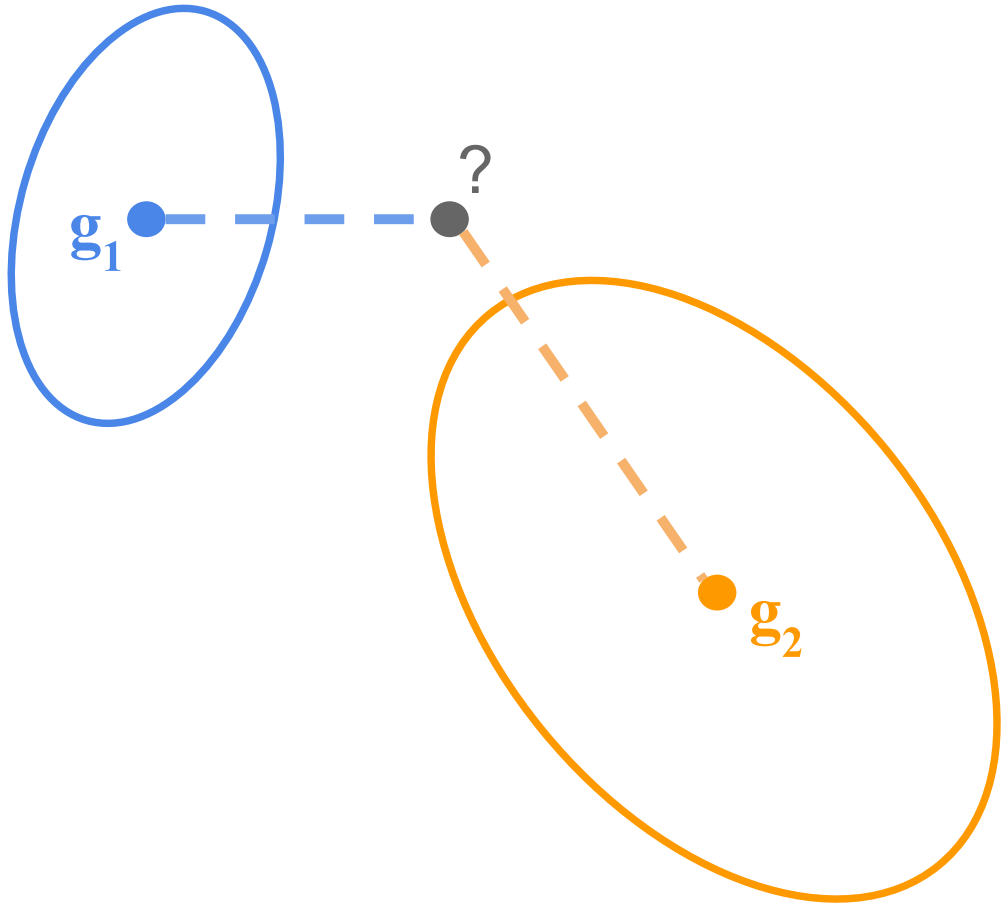
CDA Classifier



- For example, we have three unclassified individuals A, B, and C.
- We compute the Mahalanobis distances of the unclassified objects to the three centroids.
- With individual A, for example, we have to compute $d^2(\mathbf{x}_A, \mathbf{g}_1), d^2(\mathbf{x}_A, \mathbf{g}_3), d^2(\mathbf{x}_A, \mathbf{g}_2)$ and select class C_2 for which the Mahalanobis distance is minimal.

Limitations of CDA Classifier

- CDA geometric rule should not be used if the two classes have different *a priori* probabilities of variances.



Probabilistic Discriminant Analysis

- Suppose we have a response with K classes $C_1, C_2, \ldots, C_K$ and we have p predictors. The Bayes classifier gives classification rule:

  assign $\mathbf{x}_i$ to the class for which $P(y_i = k \mid \mathbf{x}_i)$ is the largest

- The rule is optimal because it minimizes the classification error.

$$\underbrace{P(y_i = k \mid \mathbf{x})}_{\text{posterior}} = \frac{\overbrace{P(\mathbf{x}_i \mid y_i = k)}^{\text{likelihood}} \overbrace{P(y_i = k)}^{\text{prior}}}{P(\mathbf{x}_i)}$$

where the denominator is obtained as:

$$P(\mathbf{x}_i) = \sum_{k=1}^{K} P(y_i = k) P(\mathbf{x}_i \mid y_i = k)$$

Probabilistic Discriminant Analysis

$$\underbrace{P(y_i = k \mid \mathbf{x})}_{\text{posterior}} = \frac{\overbrace{P(\mathbf{x}_i \mid y_i = k)}^{\text{likelihood}} \overbrace{P(y_i = k)}^{\text{prior}}}{P(\mathbf{x}_i)}$$

- Changing the notations, let:
  - $P(y = k) = \pi_k$: the prior probability of class k.
  - $P(X = \mathbf{x} \mid y = k) = f_k(\mathbf{x})$: the class-conditional density for inputs X in class k.
- Thus, posterior probability (the conditional probability of the response given the inputs) is:

$$P(y = k \mid X = \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_{k=1}^{K} f_k(\mathbf{x})\pi_k}$$

The denominator is the same for all classes $k = 1, \ldots, K$.

Normal Distributions

- **We assume that our class-conditional probabilities follow Gaussian distributions**
- For univariate data $(p = 1)$, the class-conditional density has a Normal distribution $x|C_k \sim N(\mu_k, \sigma_k)$:

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2\right\}$$

- For multivariate data $(p > 1)$, $\mathbf{x}|C_k \sim N(\mu_k, \mathbf{\Sigma}_k)$

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)^\top \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\mu_k)\right\}$$

where $\mathbf{\Sigma}_k$ and μ_k are the covariance matrix and the centroid of class C_k, the exponent is the Mahalanobis distance between $\mathbf{x}$ and μ_k:

$$d^2(\mathbf{x}, \mu_k) = (\mathbf{x}-\mu_k)^\top \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\mu_k)$$

Estimating Parameters of Normal Distributions

- To use the previous equations, we need to estimate: prior probabilities $\hat{\pi}_k$, mean-vectors $\hat{\mu}_k$, variance-covariance matrix $\hat{\mathbf{\Sigma}}_k$
- Estimating π_k is straightforward:

$$\hat{\pi}_k = \frac{n_k}{n}$$

where $n_k = |C_k|$ is the size of class k and n is the total number of data points.

- For the mean vectors $\hat{\mu}_k$ we can use the centroid of class k: $\hat{\mu}_k = \mathbf{g}_k$
- For the variance-covariance matrices $\hat{\mathbf{\Sigma}}_k$, we can use the within-variance matrix:

$$\hat{\mathbf{\Sigma}}_k = \frac{1}{n-1}\mathbf{X}_k^\top \mathbf{X}_k$$

where $\mathbf{X}_k$ is the mean-centered data matrix for objects of class k.

Discriminant Functions

- Given all the estimations, we can now find an estimate for the posterior probability $P(y_i = k \mid \mathbf{x}_i)$:

$$\hat{P}(y_i = k \mid \mathbf{x}) = \frac{\hat{\pi}_k \hat{f}_k(\mathbf{x})}{\sum_{k=1}^{K} \hat{\pi}_k \hat{f}_k(\mathbf{x})}$$

- Because the denominators remains constant across classes, we can focus on the numerator.

$$\ln\left[P(y_i = k \mid \mathbf{x})\right] \propto \ln\left(\hat{\pi}_k\right) + \ln\left[\hat{f}_k(\mathbf{x})\right]$$

- Consider the Multivariate Normal pdf, we have:

$$\ln\left[\hat{f}_k(\mathbf{x})\right] = \ln\left[(2\pi)^{-p/2}|\hat{\mathbf{\Sigma}}_k|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)^\top \hat{\mathbf{\Sigma}}_k^{-1}(\mathbf{x}-\mu_k)\right\}\right]$$

$$\longrightarrow -\frac{1}{2}\ln\left(|\hat{\mathbf{\Sigma}}_k|\right) - \frac{1}{2}\left[\mathbf{x}^\top \hat{\mathbf{\Sigma}}_k^{-1}\mathbf{x} - 2\mathbf{x}^\top \hat{\mathbf{\Sigma}}_k^{-1}\hat{\mu}_k + \hat{\mu}_k^\top \hat{\mathbf{\Sigma}}_k^{-1}\hat{\mu}_k\right]$$

Discriminant Functions

- Substitute the previous expression into $\ln\left[P(y_i = k \mid \mathbf{x})\right]$, we have the discriminant functions:

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\ln\left(|\hat{\mathbf{\Sigma}}_k|\right) - \frac{1}{2}\left[\mathbf{x}^\top \hat{\mathbf{\Sigma}}_k^{-1}\mathbf{x} - 2\mathbf{x}^\top \hat{\mathbf{\Sigma}}_k^{-1}\hat{\mu}_k + \hat{\mu}_k^\top \hat{\mathbf{\Sigma}}_k^{-1}\hat{\mu}_k\right] + \ln\left(\hat{\pi}_k\right)$$

- Our classification rule is then:

assign $\mathbf{x}$ to the class for which $\delta_k(\mathbf{x})$ is the largest

Quadratic Discriminant Analysis (QDA)

- We classify the order of each term in $\delta_k(\mathbf{x})$ with respect to $\mathbf{x}$:

$$\delta_k(\mathbf{x}) = \underbrace{-\frac{1}{2}\ln\left(|\hat{\mathbf{\Sigma}}_k|\right)}_{\text{constant}} - \frac{1}{2}\Bigg[\underbrace{\mathbf{x}^\top\hat{\mathbf{\Sigma}}_k^{-1}\mathbf{x}}_{\text{quadratic}} - \underbrace{2\mathbf{x}^\top\hat{\mathbf{\Sigma}}_k^{-1}\hat{\mu}_k}_{\text{linear}} + \underbrace{\hat{\mu}_k^\top\hat{\mathbf{\Sigma}}_k^{-1}\hat{\mu}_k}_{\text{constant}}\Bigg] + \underbrace{\ln\left(\hat{\pi}_k\right)}_{\text{constant}}$$

- We have a quadratic function of $\mathbf{x}$. This leads us to Quadratic Discriminant Analysis (QDA).
- Having a quadratic discriminant function causes the decision boundaries in QDA to be quadratic surfaces.

Linear Discriminant Analysis (LDA)

- Assume we have all covariance matrices in $\delta_k(\mathbf{x})$ are the same:

$$\hat{\mathbf{\Sigma}}_1 = \hat{\mathbf{\Sigma}}_2 = \ldots = \hat{\mathbf{\Sigma}}_K = \hat{\mathbf{\Sigma}}$$

- The discriminant function $\delta_k(\mathbf{x})$ becomes:

$$\underbrace{-\frac{1}{2}\ln\left(|\hat{\mathbf{\Sigma}}|\right)}_{\text{no } k \text{ dependency}} -\frac{1}{2}\Bigg[\underbrace{\mathbf{x}^\top\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}}_{\text{no } k \text{ dependency}} - \underbrace{2\mathbf{x}^\top\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_k}_{\text{linear}} + \underbrace{\hat{\mu}_k^\top\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_k}_{\text{constant}}\Bigg] + \underbrace{\ln\left(\hat{\pi}_k\right)}_{\text{constant}}$$

- Ignoring the terms that do not depend on k, we obtain:

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\Bigg[\underbrace{-2\mathbf{x}^\top\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_k}_{\text{linear}} + \underbrace{\hat{\mu}_k^\top\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_k}_{\text{constant}}\Bigg] + \underbrace{\ln\left(\hat{\pi}_k\right)}_{\text{constant}}$$

- The discriminant function now is linear w.r.t. $\mathbf{x}$, we have Linear Discriminant Analysis (LDA).

Canonical Discriminant Analysis (CDA)

- We again assume that all covariance matrices are the same across classes, as well as the prior probabilities:

$$\hat{\boldsymbol{\Sigma}}_1 = \hat{\boldsymbol{\Sigma}}_2 = \ldots = \hat{\boldsymbol{\Sigma}}_K = \hat{\boldsymbol{\Sigma}} \quad \text{and} \quad \hat{\pi}_1 = \hat{\pi}_2 = \ldots = \hat{\pi}_K = \hat{\pi}$$

- The discriminant function $\delta_k(\mathbf{x})$ becomes:

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\left[\underbrace{-2\mathbf{x}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_k}_{\text{linear}} + \underbrace{\hat{\mu}_k^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_k}_{\text{constant}}\right] + \underbrace{\ln(\hat{\pi})}_{\text{no } k \text{ dependency}}$$

- After ignoring the k-independent terms, we can have:

$$\delta_k(\mathbf{x}) = \underbrace{\mathbf{x}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_k}_{\text{linear}} - \frac{1}{2}\underbrace{\hat{\mu}_k^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_k}_{\text{constant}}$$

- This is the distance between $\mathbf{x}$ and the centroid of C_k using the within-class variance matrix as a metric matrix. The above expression is Canonical Discriminant Analysis (CDA).

Naive Bayes

- After assuming equal covariance matrices, equal priors, we can also assume that $\hat{\mathbf{\Sigma}}$ is diagonal:

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} Var(X_1) & 0 & \dots & 0 \\ 0 & Var(X_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Var(X_p) \end{pmatrix}$$

- The predictors are uncorrelated.
- The discriminant function with diagonal $\hat{\mathbf{\Sigma}}$:

$$\delta_k(\mathbf{x}) = \underbrace{\mathbf{x}^\top \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_k}_{\text{linear}} - \frac{1}{2} \underbrace{\hat{\mu}_k^\top \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_k}_{\text{constant}}$$

- This is Naive Bayes.

Comparisons

Method	Covariance Matrices	Priors	Bayes Rule
QDA	Unequal across classes	Unequal across classes	quadratic
LDA	Equal across classes	Unequal across classes	linear
CDA	Equal across classes,	Equal across classes	linear
Naive Bayes	Equal across classes, Diagonal	Equal across classes	linear