MONASH University

ETC3250/5250: Introduction to Machine Learning

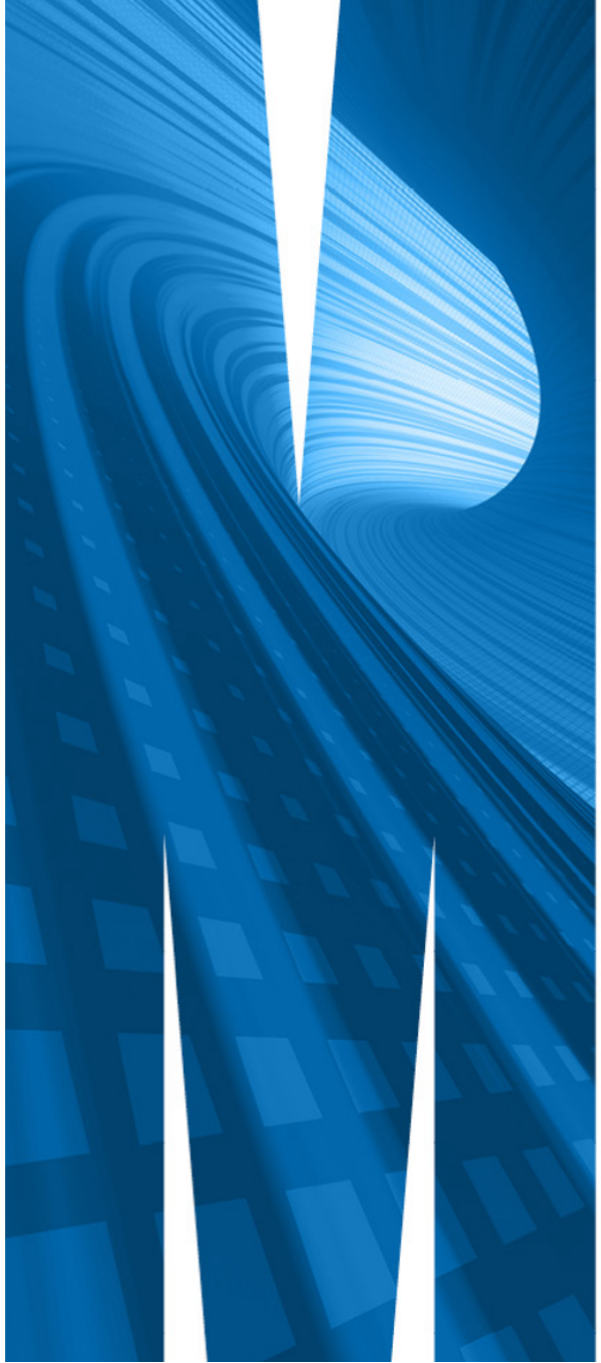
Categorical response: Discriminant analysis

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 4a



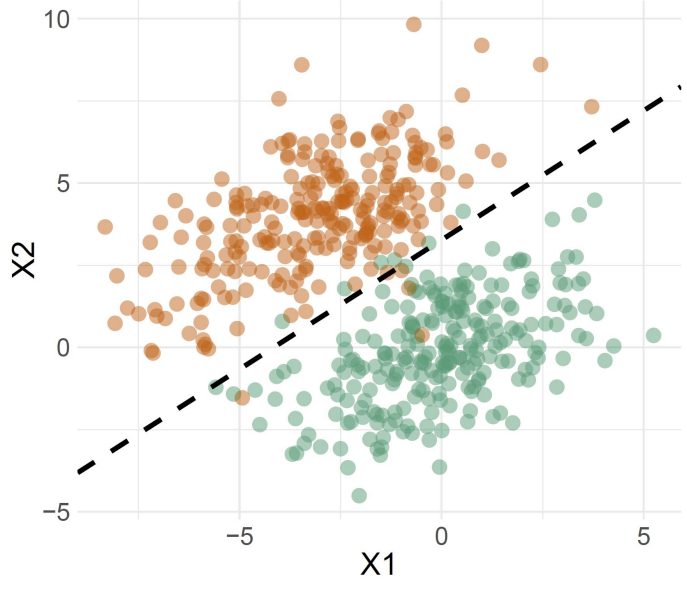
Linear Discriminant Analysis

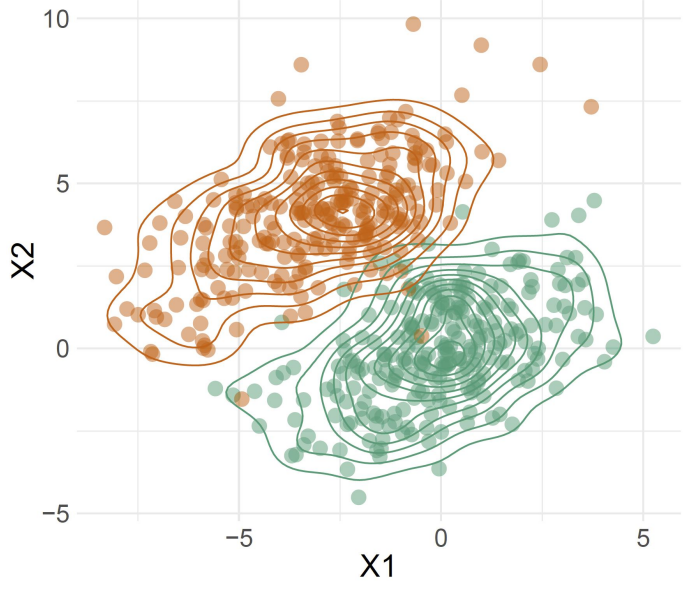
Logistic regression involves directly modeling $P(Y = k|X = x)$ using the logistic function. Rounding the probabilities produces class predictions, in two class problems; selecting the class with the highest probability produces class predictions in multi-class problems.

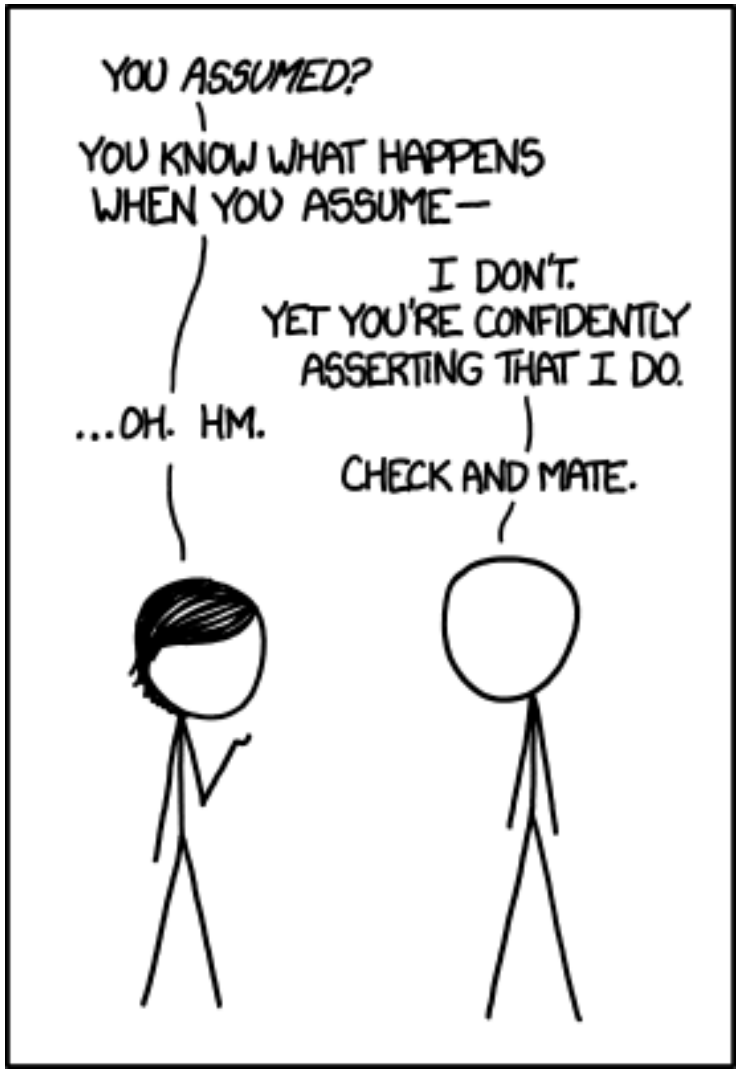
Another approach for building a classification model is linear discriminant analysis. This involves assuming the distribution of the predictors is a multivariate normal, with the same variance-covariance matrix, separately for each class.

Compare the pair

Logistic Regression	Linear Discriminant Analysis
Goal - directly estimate $P(Y\|X)$ (*the dashed line*)	**Goal** - estimate $P(X\|Y)$ (*the contours*) to then deduce $P(Y\|X)$
Assumptions - no assumptions on predictor space	**Assumptions** - predictors are normally distributed







Source: https://xkcd.com

Assumptions are critical in LDA

- All samples come from normal populations
- All the groups have the same variance-covariance matrix

Bayes Theorem

Let $f_k(x)$ be the density function for predictor x for class k. If f is small, the probability that x belongs to class k is small, and conversely if f is large.

Bayes theorem (for K classes) states:

$$P(Y = k|X = x) = p_k(x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_k f_k(x)}$$

where $\pi_k = P(Y = k)$ is the prior probability that the observation comes from class k.

LDA with $p = 1$ predictors

We assume $f_k(x)$ is univariate Normal (Gaussian):

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

where μ_k and σ_k^2 are the mean and variance parameters for the kth class. Further assume that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$; then the conditional probabilities are

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K}\pi_l \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

LDA with $p = 1$ predictors

The Bayes classifier is assign new observation $X = x_0$ to the class with the highest $p_k(x_0)$. A simplification of $p_k(x_0)$ yields the discriminant functions:

$$\delta_k(x_0) = x_0 \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$$

and the rule Bayes classifier will assign x_0 to the class with the largest value.

LDA with $p = 1$ predictors

If $K = 2$ and $\pi_1 = \pi_2$, we assign x_0 to class 1 if

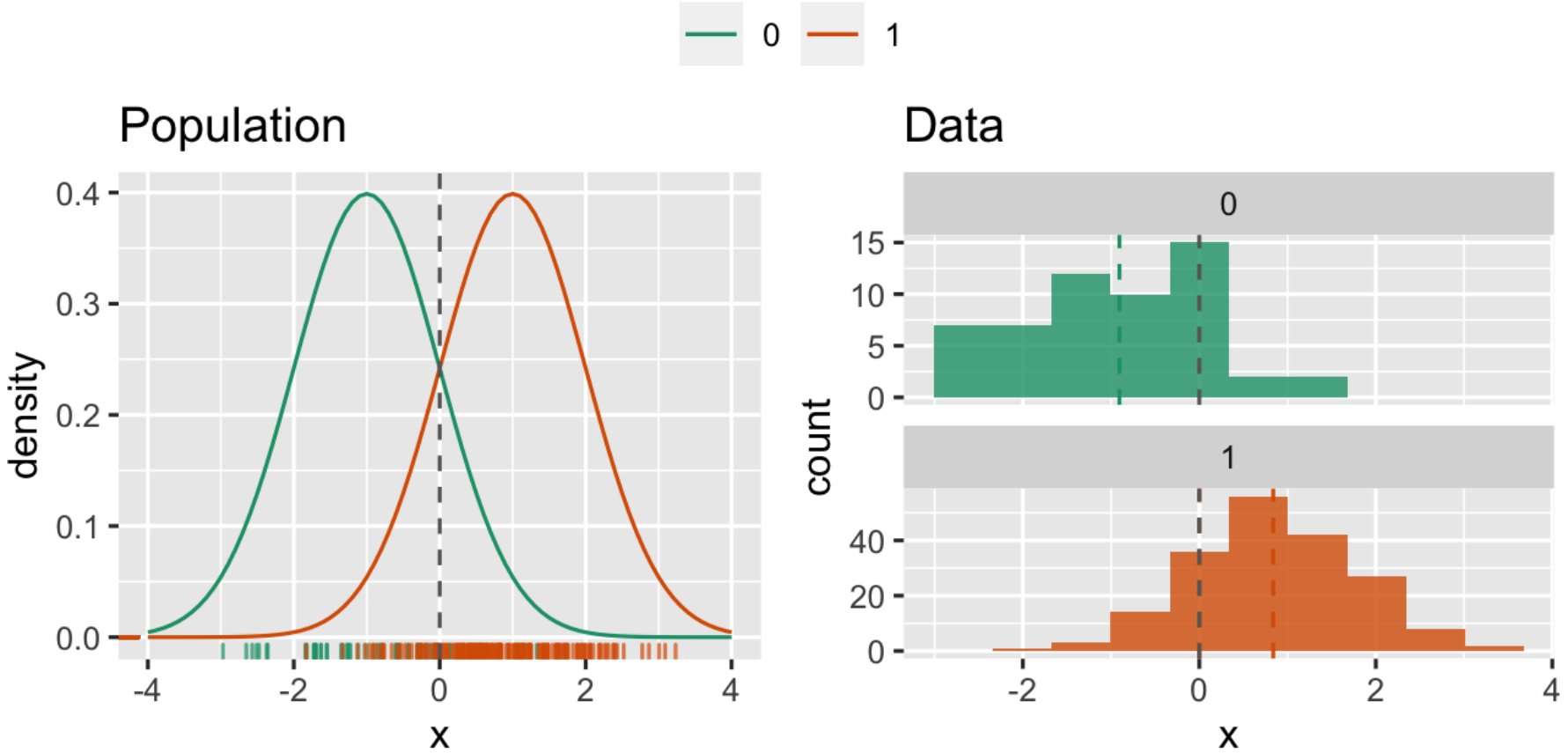
$$\delta_1(x_0) > \delta_2(x_0)$$

$$x_0 \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \log(\pi) > x_0 \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \log(\pi)$$

which simplifies to $x_0 > \frac{\mu_1+\mu_2}{2}$.

This is estimated on the data with $x_0 > \frac{\bar{x}_1+\bar{x}_2}{2}$.

LDA with $p = 1$ predictors



Multivariate LDA

To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution with $E[X] = \mu$ and $\mathrm{Cov}(X) = \Sigma$, we write $X \sim N(\mu, \Sigma)$.

The multivariate normal density function is:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}$$

with x, μ are p-dimensional vectors, Σ is a $p \times p$ variance-covariance matrix.

Multivariate LDA

The discriminant functions are:

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log(\pi_k)$$

and Bayes classifier is assign a new observation x_0 to the class with the highest $\delta_k(x_0)$.

When $K = 2$ and $\pi_1 = \pi_2$ this reduces to

Assign observation x_0 to class 1 if

$$x_0^T \underbrace{\Sigma^{-1}(\mu_1 - \mu_2)}_{dimension\ reduction} > \frac{1}{2}(\mu_1 + \mu_2)^T \underbrace{\Sigma^{-1}(\mu_1 - \mu_2)}_{dimension\ reduction}$$

Class 1 and 2 need to be mapped to the classes in the your data. The class "to the right" on the reduced dimension will correspond to class 1 in this equation.

Dimension reduction

Dimension reduction via LDA

Discriminant space: a benefit of LDA is that it provides a low-dimensional projection of the p-dimensional space, where the groups are the most separated. For $K = 2$, this is

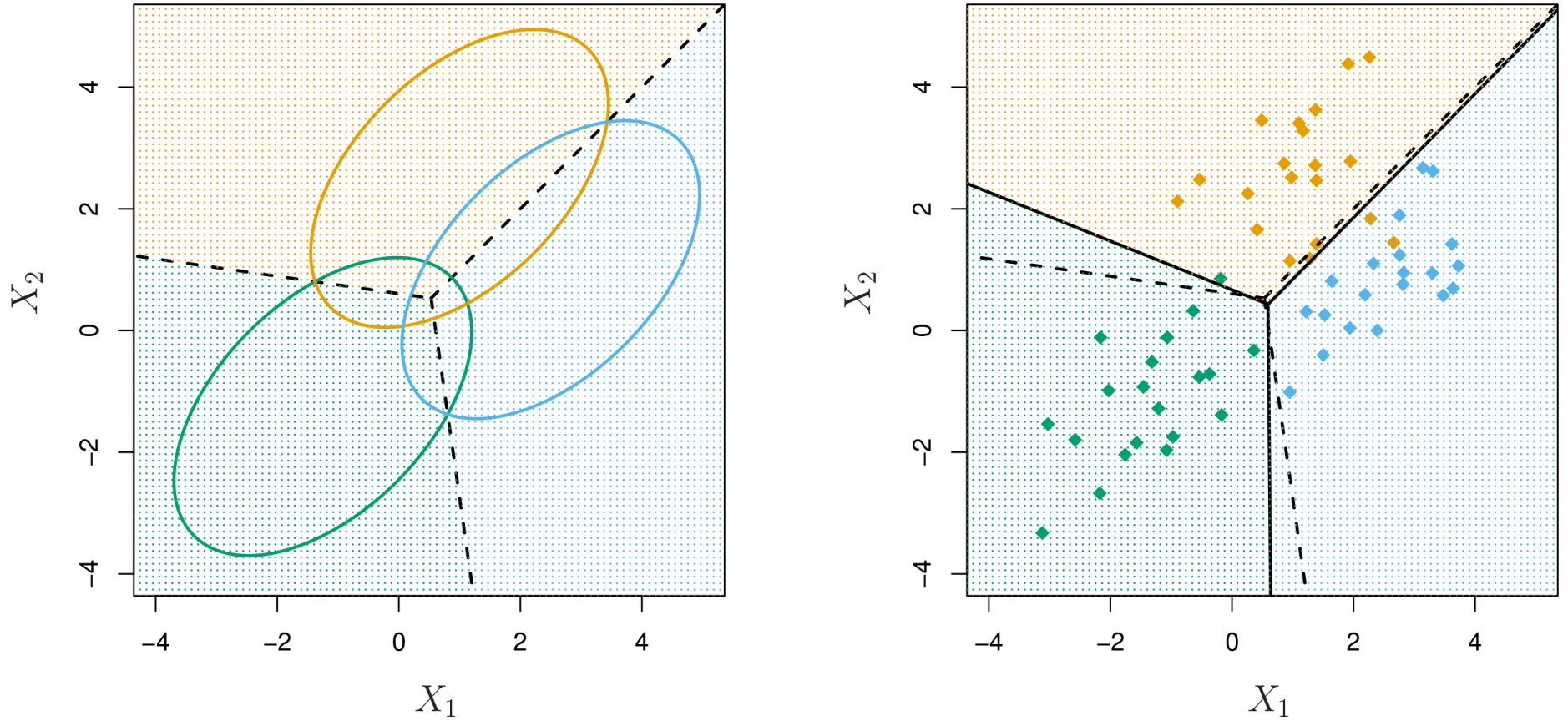
$$\Sigma^{-1}(\mu_1 - \mu_2)$$

This corresponds to the biggest separation between means relative to the variance-covariance.

For $K > 2$, the discriminant space is found be taking an eigen-decomposition of $\Sigma^{-1}\Sigma_B$, where $\Sigma_B = \frac{1}{K}\sum_{i=1}^{K}(\mu_i - \mu)(\mu_i - \mu)^T$

Discriminant space

The dashed lines are the Bayes decision boundaries. Ellipses that contain 95% of the probability for each of the three classes are shown. Solid line corresponds to the class boundaries from the LDA model fit to the sample.



(Chapter4/4.6.pdf)

Discriminant space: using sample statistics

Discriminant space: is the low-dimensional space where the class means are the furthest apart relative to the common variance-covariance.

The discriminant space is provided by the eigenvectors after making an eigen-decomposition of $\hat{\Sigma}^{-1}\hat{\Sigma}_B$, where

$$\hat{\Sigma}_B = \frac{1}{K}\sum_{i=1}^{K}(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \quad \text{and} \quad \hat{\Sigma} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}(x_i - \bar{x}_k)(x_i - \bar{x}_k)^T$$

Mahalanobis distance

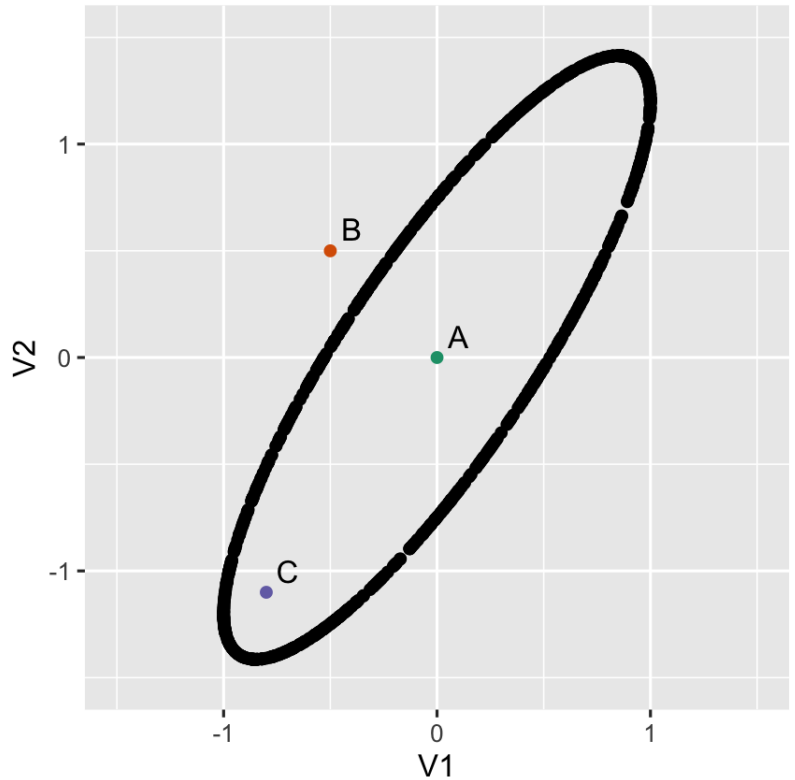
For two p-dimensional vectors, Euclidean distance is

$$d(x, y) = \sqrt{(x - y)^T (x - y)}$$

and Mahalanobs distance is

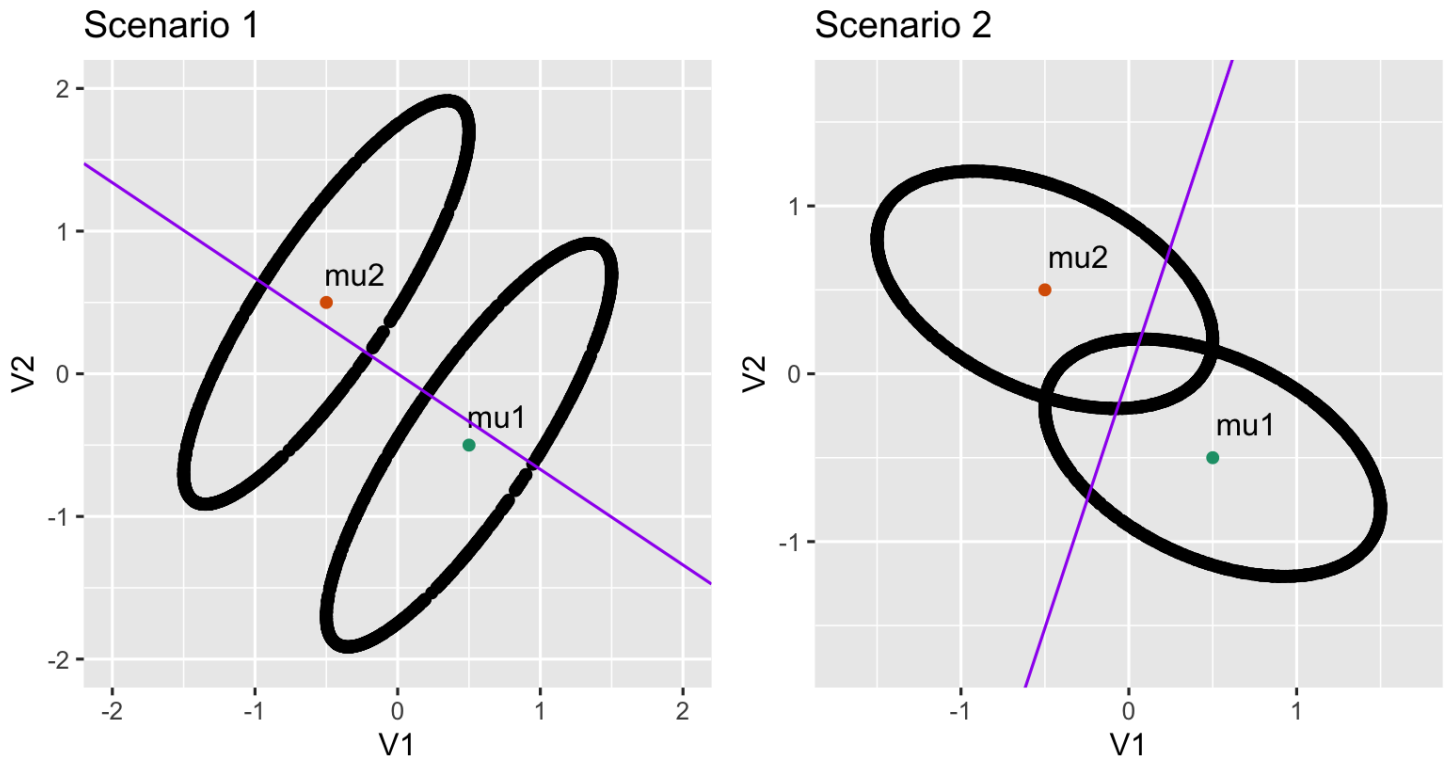
$$d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

Which points are closest according to Euclidean distance? Which points are closest relative to the variance-covariance?



Discriminant space

Both means the same. Two different variance-covariance matrices. Discriminant space depends on the variance-covariance matrix.



Quadratic Discriminant Analysis

If the groups have different variance-covariance matrices, but still come from a normal distribution

Quadratic DA (QDA)

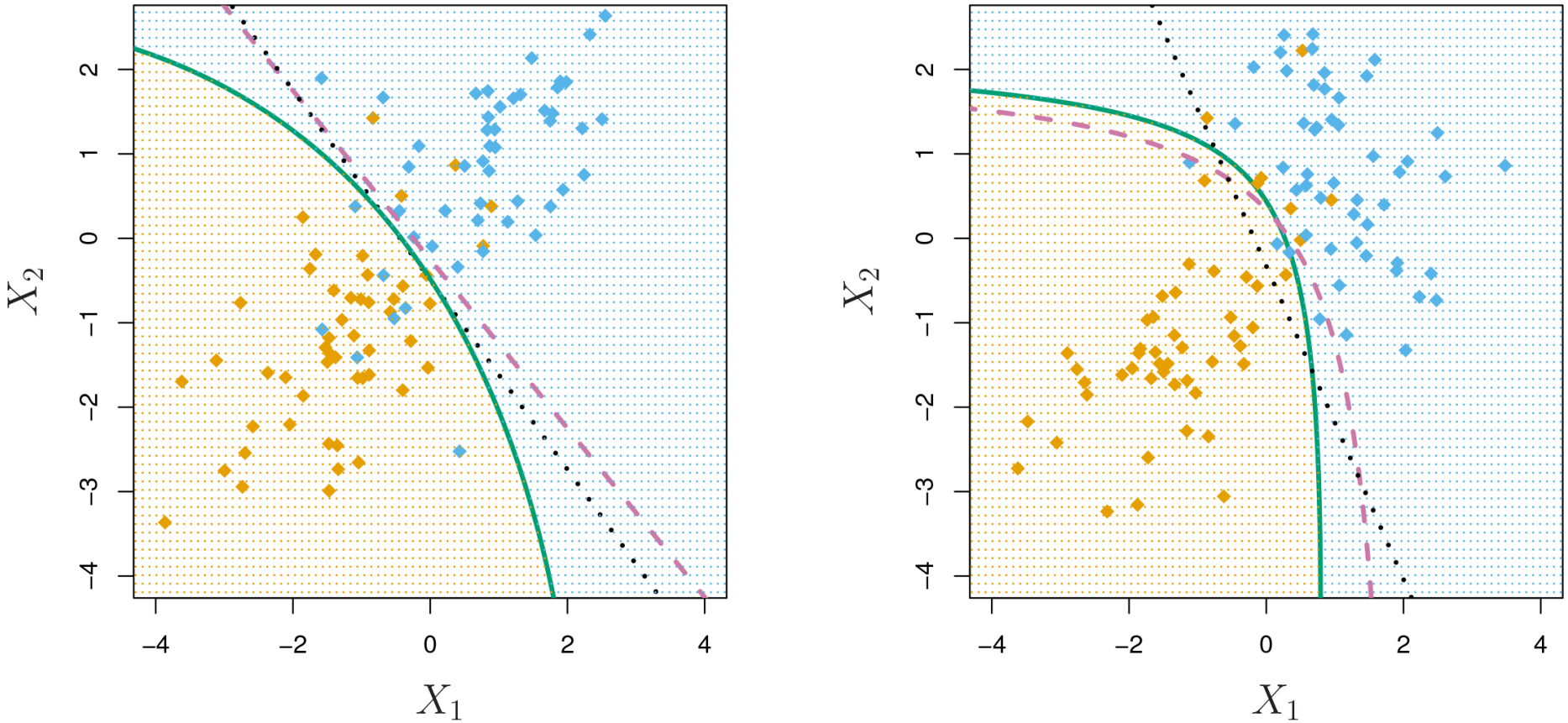
If the variance-covariance matrices for the groups are NOT EQUAL, then the discriminant functions are:

$$\delta_k(x) = x^T \Sigma_k^{-1} x + x^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \log |\Sigma_k| + \log(\pi_k)$$

where Σ_k is the population variance-covariance for class k, estimated by the sample variance-covariance S_k, and μ_k is the population mean vector for class k, estimated by the sample mean $\bar{x}_k$.

Quadratic DA (QDA)

A quadratic boundary is obtained by relaxing the assumption of equal variance-covariance, and assume that $\Sigma_k \neq \Sigma_l, \quad k \neq l, k, l = 1, \ldots, K$

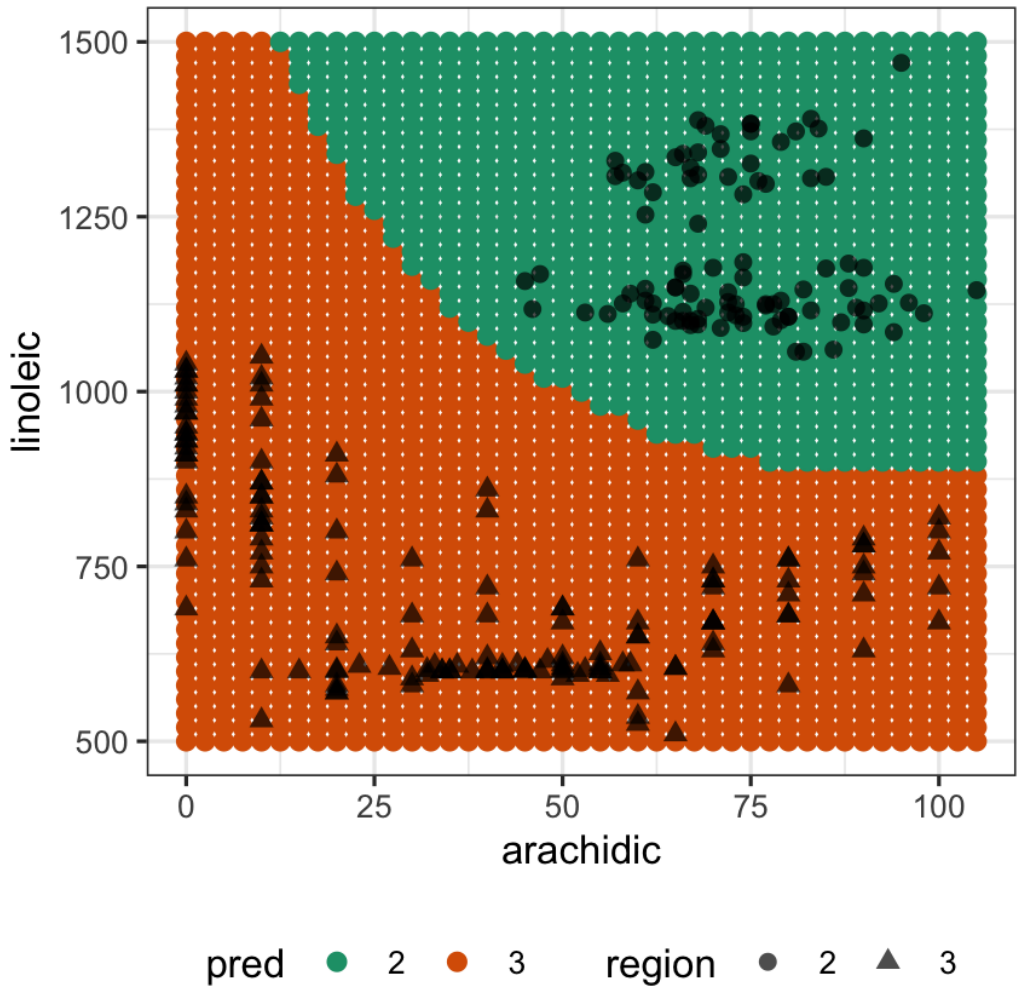


true, LDA, QDA.

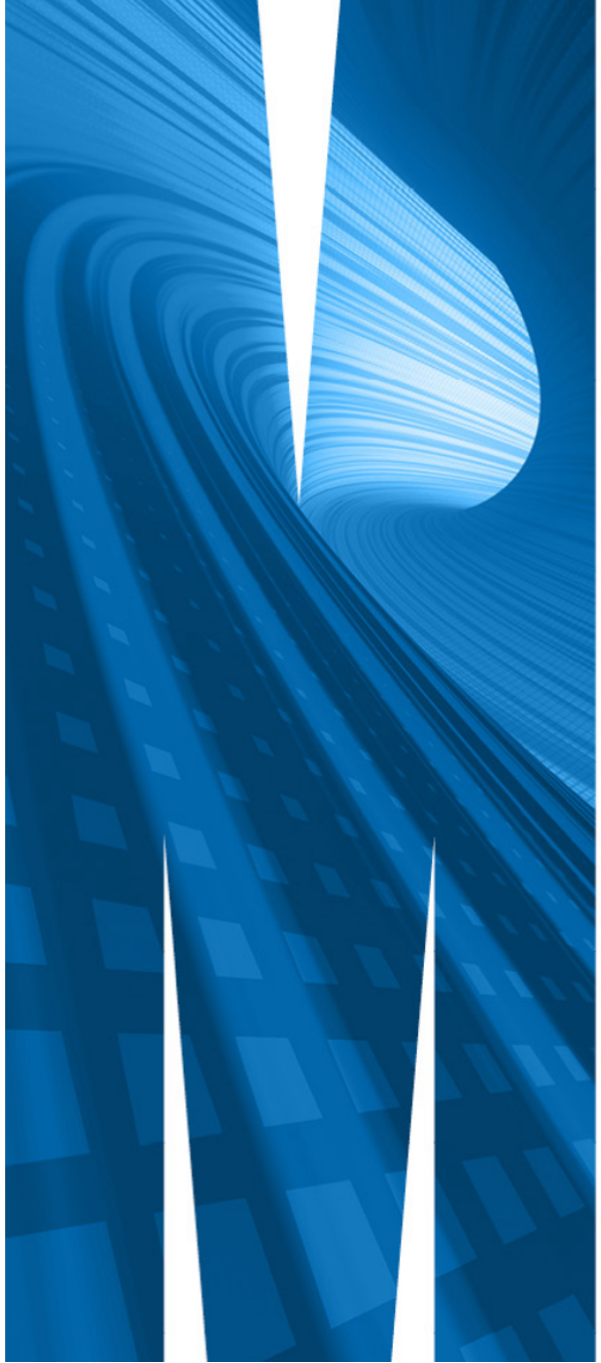
(Chapter4/4.9.pdf)

QDA: Olive oils example

Even if the population is NOT normally distributed, QDA might do reasonably. On this data, region 3 has a "banana-shaped" variance-covariance, and region 2 has two separate clusters. The quadratic boundary though does well to carve the space into neat sections dividing the two regions.



MONASH University



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 4a