Multivariate Analysis (slides 8)

- Today we consider Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA).
- These are used if it is assumed that there exists a set of k groups within the data and that there is a subset of the data that is labelled, *i.e.*, whose group membership is known.
- Discriminant analysis refers to a set of 'supervised' statistical techniques where the class information is used to help reveal the structure of the data.
- This structure then allows the 'classification' of future observations.

Discriminant Analysis

- We want to be able to use knowledge of labelled data (*i.e.*, those whose group membership is known) in order to classify the group membership of unlabelled data.

- We previously considered the k-nearest neighbours technique for this problem.

- We shall now consider the alternative approaches of:
  - LDA (linear discriminant analysis)
  - QDA (quadratic discriminant analysis)

LDA & QDA

- Unlike k-Nearest Neighbours (and all the other techniques so far covered), both LDA and QDA assume the use of a distribution over the data.
- Once we introduce distributions (and parameters of those distributions), we can start to quantify uncertainty over the structure of the data.
- As far as classification is concerned, this means that we can start to talk about the probability of group assignment.
- The distinction between a point that is assigned a probability of 0.51 to one group and 0.49 to another, against a point that is assigned a probability of 0.99 to one group and 0.01 to another, can be quite important.

Multivariate Normal Distribution

- Let $\mathbf{x}^T = (x_1, x_2, ..., x_m)$, where $x_1, x_2, ..., x_m$ are random variables.
- The Multi-Variate Normal (MVN) distribution has two parameters:
  - Mean μ, an m-dimensional vector.
  - Covariance matrix $\mathbf{\Sigma}$, with dimension $m \times m$.
- A vector $\mathbf{x}$ is said to follow a MVN distribution, denoted $\mathbf{x} \sim MVN(\mu, \mathbf{\Sigma})$, if it has the following probability density function:

$$f(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{m}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)\right]$$

- Here $|\mathbf{\Sigma}|$ is used to denote the determinant of $\mathbf{\Sigma}$.

Multivariate Normal Distribution

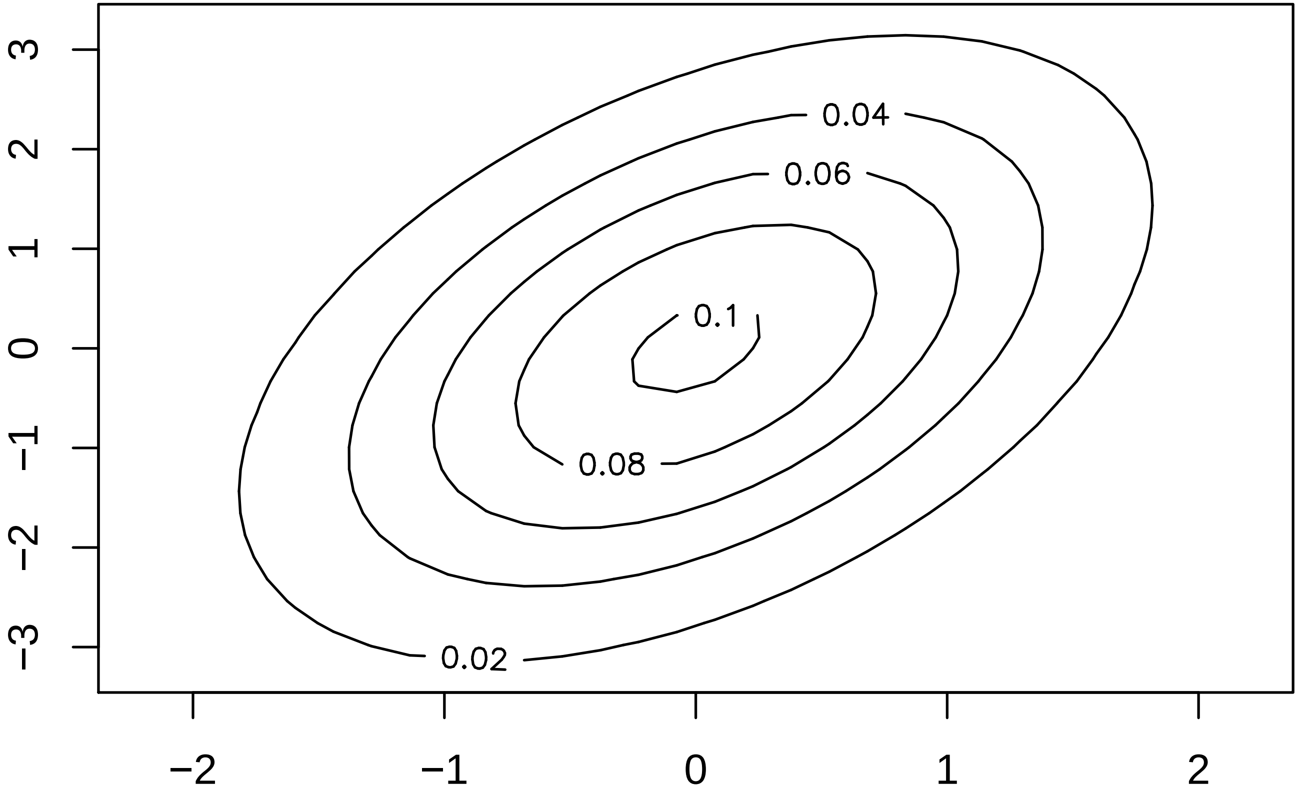
- The MVN distribution is very useful when modelling multivariate data.
- Notice:

$$\{\mathbf{x} : f(\mathbf{x}|\mu, \mathbf{\Sigma}) > C\} = \left\{\mathbf{x} : (\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) < -2 \log \left[C (2\pi)^{\frac{m}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}\right]\right\}$$

- This corresponds to an m-dimensional ellipsoid centered at point μ.
- If it is assumed that the data within a group k follows a MVN distribution with mean μ_k and covariance $\mathbf{\Sigma}_k$, then the scatter of the data should be roughly elliptical.
- The mean fixes the location of the scatter and the covariance affects the shape of the ellipsoid.

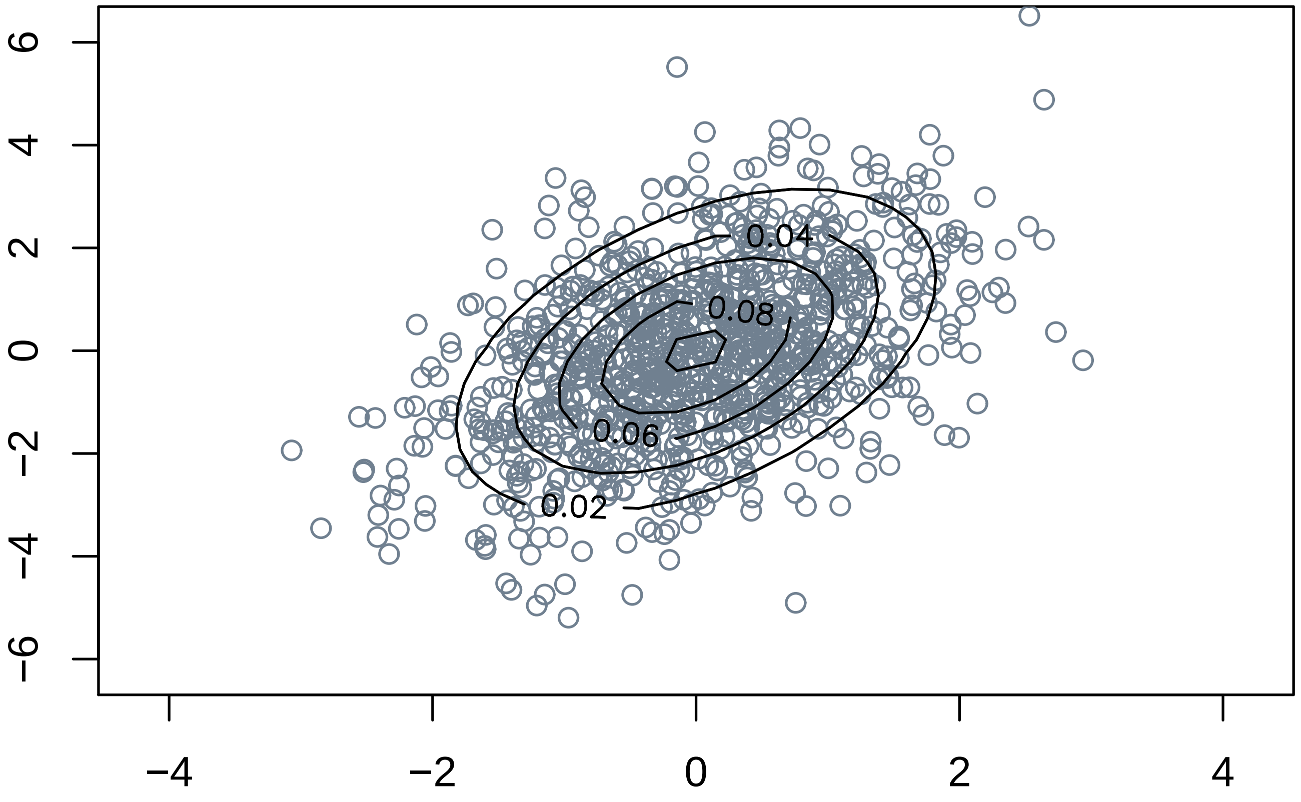
Normal Contours

- For example, the contour plot of a MVN $\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 3 \end{pmatrix}\right]$ is:



Normal Contours: Data

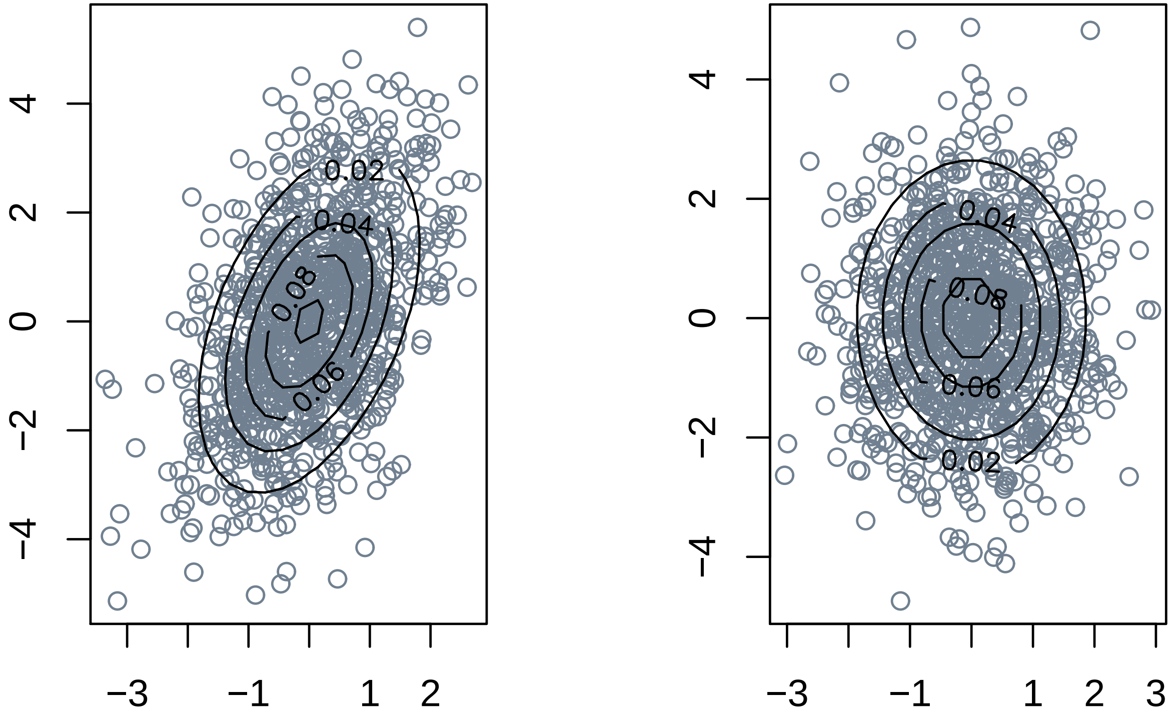
- Sampling from this distribution and overlaying the results on the contour plot gives:



Shape of Scatter

- If we assume that the data within each group follows a MVN distribution with mean μ_k and covariance $\mathbf{\Sigma}_k$, then we also assume that the scatter is roughly elliptical.

- The mean sets the location of this scatter and the covariance sets the shape of the ellipse.

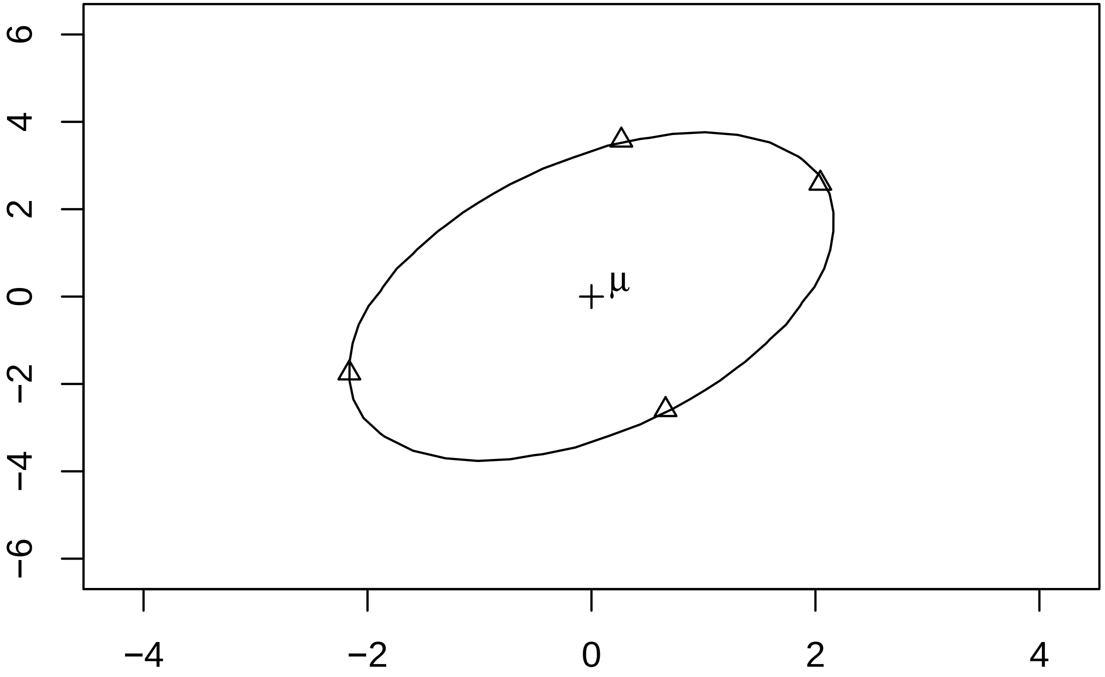


Mahalanobis Distance

- The Mahalanobis distance from a point $\mathbf{x}$ to a mean μ is D, where

$$D^2 = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu).$$

- Two points have the same Mahalanobis distance if they are on the same ellipsoid centered on μ (as defined earlier).



Which Is Closest?

- Suppose we wish to find the mean μ_k that a point $\mathbf{x}$ is closest to as measured by Mahalanobis distance.

- That is, we want to find the k that minimizes the expression:

$$(\mathbf{x} - \mu_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \mu_k)$$

- The point $\mathbf{x}$ is closer to μ_k than it is to μ_l (under Mahalanobis distance) when:

$$(\mathbf{x} - \mu_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \mu_k) < (\mathbf{x} - \mu_l)^T \mathbf{\Sigma}_l^{-1} (\mathbf{x} - \mu_l).$$

- Note that this is a quadratic expression for $\mathbf{x}$.

When Covariance is Equal

- If $\mathbf{\Sigma}_k = \mathbf{\Sigma}$ for all k, then the previous expression becomes:

$$(\mathbf{x} - \mu_k)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu_k) < (\mathbf{x} - \mu_l)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu_l).$$

- This can be simplified as:

$$-2\mathbf{x}^T \mathbf{\Sigma}^{-1} \mu_k + \mu_k^T \Sigma^{-1} \mu_k < -2\mathbf{x}^T \mathbf{\Sigma}^{-1} \mu_l + \mu_l^T \Sigma^{-1} \mu_l$$

$$\Leftrightarrow \quad -2\mu_k^T \mathbf{\Sigma}^{-1} \mathbf{x} + \mu_k^T \Sigma^{-1} \mu_k < -2\mu_l^T \mathbf{\Sigma}^{-1} \mathbf{x} + \mu_l^T \Sigma^{-1} \mu_l$$

- This is now a linear expression for $\mathbf{x}$
- *Note the names of 'linear' discriminant analysis and 'quadratic' discriminant analysis.*

Estimating Equal Covariance

- In LDA we need to pool the covariance matrices of individual classes.
- Remember that the sample covariance matrix Q for a set of n observations of dimension m is the matrix whose elements are

  $$q_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)$$

  for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, m$.
- Then the pooled covariance matrix is defined as:

  $$Q_p = \frac{1}{n-g} \sum_{l=1}^{g} (n_l - 1) Q_l$$

  Where g is the number of classes, Q_l is the estimated sample covariance matrix for class l, n_l is the number of data points in class l, whilst n is the total number of data points.

Estimating Equal Covariance

- This formula arises from summing the squares and cross products over data points in all classes:

$$W_{ij} = \sum_{l=1}^{g} \sum_{k=1}^{n_l} (x_{ki} - \overline{x}_{li})(x_{kj} - \overline{x}_{lj})$$

  for $i = 1, \ldots, m$ and $j = 1, \ldots, m$.

- Hence:

$$W = \sum_{l=1}^{g} (n_l - 1) Q_l$$

- Given n data points falling in g groups, we have $n - g$ degrees of freedom because we need to estimate the g group means.

- This results in the previous formula for the pooled covariance matrix:

$$Q_p = \frac{W}{n - g}$$

Modelling Assumptions

- Both LDA and QDA are *parametric* statistical methods.
- In order to classify a new observation $\mathbf{x}$ into one of the known K groups, we need to know $\mathbb{P}(\mathbf{x} \in k|\mathbf{x})$ for $k = 1, \ldots K$.
- That is to say, we need to know the posterior probability of belonging to each of the possible groups given the data.
- Then classify the new observation as belonging to the class which has largest posterior probability.
- Bayes' Theorem states that the posterior probability of observation $\mathbf{x}$ belonging to group k is:

$$\mathbb{P}(\mathbf{x} \in k|\mathbf{x}) = \frac{\pi_k f(\mathbf{x}_i|\mathbf{x} \in k)}{\sum_{l=1}^{K} \pi_l f(\mathbf{x}_i|\mathbf{x} \in l)}$$

Modelling Assumptions

- Discriminant analysis assumes that observations from group k follow a MVN distribution with mean μ_k and covariance $\mathbf{\Sigma}_k$.
- That is

$$f(\mathbf{x}|\mathbf{x} \in k) = f(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) = \frac{1}{(2\pi)^{\frac{m}{2}}|\mathbf{\Sigma}_k|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x} - \mu_k)\right]$$

- Discriminant analysis (as presented here) also assumes values for $\pi_k = \mathbb{P}(\mathbf{x} \in k)$, which is the proportion of population objects belonging to class k (this can be known or estimated).
- Note that $\sum_{k=1}^{K} \pi_k = 1$.
- Typically, $\pi_k = 1/K$ is used.
- π_k are sometimes referred to as prior probabilities.
- Using all this we can compute $\mathbb{P}(\mathbf{x} \in k|\mathbf{x})$ and assign data points to groups so as to maximise this probability.

Some Calculations

- The probability of $\mathbf{x}$ belonging to group k conditional on $\mathbf{x}$ being known satisfies:
$$\mathbb{P}(\mathbf{x} \in k|\mathbf{x}) \propto \pi_k f(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k).$$
- Hence,
$$\mathbb{P}(\mathbf{x} \in k|\mathbf{x}) > \mathbb{P}(\mathbf{x} \in l|\mathbf{x}) \Leftrightarrow \pi_k f(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) > \pi_l f(\mathbf{x}|\mu_l, \mathbf{\Sigma}_l)$$
- Taking logarithms and substituting in the probability density function for a MVN distribution we find after simplification:
$$\log \pi_k - \frac{1}{2} \log |\mathbf{\Sigma}_k| - \frac{1}{2}(\mathbf{x} - \mu_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x} - \mu_k)$$
$$> \log \pi_l - \frac{1}{2} \log |\mathbf{\Sigma}_l| - \frac{1}{2}(\mathbf{x} - \mu_l)^T \mathbf{\Sigma}_l^{-1}(\mathbf{x} - \mu_l)$$

Linear Discriminant Analysis

- If equal covariances are assumed then $\mathbb{P}(\mathbf{x} \in k|\mathbf{x}) > \mathbb{P}(\mathbf{x} \in l|\mathbf{x})$ if and only if:

$$\log \pi_k + \mathbf{x}^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k > \log \pi_l + \mathbf{x}^T \mathbf{\Sigma}^{-1} \mu_l - \frac{1}{2} \mu_l^T \Sigma^{-1} \mu_l$$

- Hence the name linear discriminant analysis.

- If $\pi_k = 1/K$ for all k, then this reduces further.

$$\left(\mathbf{x} - \frac{1}{2} (\mu_k + \mu_l) \right)^T \mathbf{\Sigma}^{-1} (\mu_k - \mu_l) > 0$$

Quadratic Discriminant Analysis

- No simplification arises in the unequal covariance case, hence $\mathbb{P}(\mathbf{x} \in k|\mathbf{x}) > \mathbb{P}(\mathbf{x} \in l|\mathbf{x})$ if and only if:

$$\log \pi_k - \frac{1}{2} \log |\mathbf{\Sigma}_k| - \frac{1}{2}(\mathbf{x} - \mu_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \mu_k)$$

$$> \log \pi_l - \frac{1}{2} \log |\mathbf{\Sigma}_l| - \frac{1}{2}(\mathbf{x} - \mu_l)^T \mathbf{\Sigma}_l^{-1} (\mathbf{x} - \mu_l)$$

- Hence the name quadratic discriminant analysis.

- If $\pi_k = 1/K$ for all k, then some simplification arises.

Summary

- In LDA the decision boundary between class k and class l is given by:

$$\log \frac{P(k|\mathbf{x})}{P(l|\mathbf{x})} = \log \frac{\pi_k}{\pi_l} + \log \frac{f(\mathbf{x}|k)}{f(\mathbf{x}|l)} = 0$$

- Unlike k-nearest neighbour, both LDA and QDA are model based classifiers where P(data|group) is assumed to follow a MVN distribution:
  - The model based assumption allows for the generation of the probability for class membership.
  - The MVN assumption means that groups are assumed to follow an elliptical shape.
- Whilst LDA assumes groups have the same covariance matrix, QDA permits different covariance structures between groups.