Multivariate Analyses—A Normal Approach

Hien Duy Nguyen
La Trobe University
(Contact–Email: h.nguyen5@latrobe.edu.au; Website: hiendn.github.io)

ICF Seminars, 2020

Outline

- Multivariate data
- The normal distribution
- Testing basic properties
- Regression and dimensionality reduction
- Classification and clustering

When are data multivariate?

- Suppose that

$$\boldsymbol{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_d \end{bmatrix} \in \mathbb{R}^d,$$

  is a random d-dimensional real vector.

- If $d = 1$, then we say that $\boldsymbol{Z}$ is **univariate**; otherwise, if $d > 1$, then we say that $\boldsymbol{Z}$ is **multivariate**.

- If we observe n instances of $\boldsymbol{Z}$, i.e.,

$$\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n,$$

  then we have a **multivariate data set** of size n.

Example of multivariate data

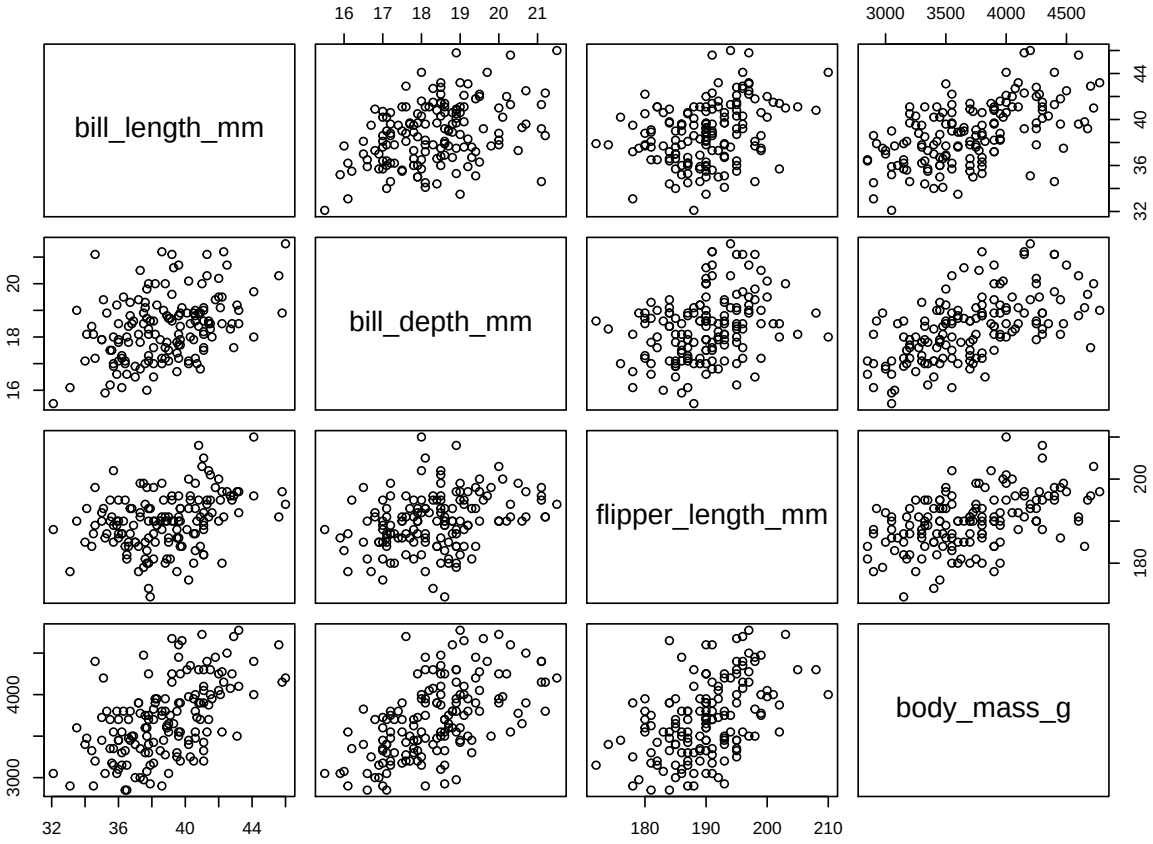

Figure: $n = 151$ sets of Adelie penguin measurements from the `penguins` data set.

Example of multivariate data

Figure: An independent sample of Adelie penguins.

Example of multivariate data

- Let $\boldsymbol{Z}_i \in \mathbb{R}^4$ be a vector of measurements from penguin i, where $i \in \{1, \ldots, n\}$ $(n = 151)$.
- We can write:

$$\boldsymbol{Z}_i = \begin{bmatrix} Z_{i1} \\ Z_{i2} \\ Z_{i3} \\ Z_{i4} \end{bmatrix} = \begin{bmatrix} \text{bill length}_i \\ \text{bill depth}_i \\ \text{flipper length}_i \\ \text{body mass}_i \end{bmatrix}.$$

- For penguin $i = 1$, we observe the **realization**

$$\boldsymbol{z}_1 = \begin{bmatrix} 39.1 & 18.7 & 181 & 3750 \end{bmatrix}^\top.$$

Summarizing multivariate data

- Let
$$\mathrm{E}(\boldsymbol{Z}) = \begin{bmatrix} \mathrm{E}(Z_1) & \mathrm{E}(Z_2) & \cdots & \mathrm{E}(Z_d) \end{bmatrix}^\top$$
be the **expectation vector** of $\boldsymbol{Z}$, which is a vector concatenation of the univariate expectations.
- Let
$$\mathrm{cov}(\boldsymbol{Z}) = \begin{bmatrix} \mathrm{cov}(Z_1, Z_1) & \mathrm{cov}(Z_1, Z_2) & \cdots & \mathrm{cov}(Z_1, Z_d) \\ \mathrm{cov}(Z_2, Z_1) & \mathrm{cov}(Z_2, Z_2) & \cdots & \mathrm{cov}(Z_2, Z_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(Z_d, Z_1) & \mathrm{cov}(Z_d, Z_2) & \cdots & \mathrm{cov}(X_d, X_d) \end{bmatrix},$$
be the **covariance matrix** of $\boldsymbol{Z}$, where
$$\mathrm{cov}(Z_j, Z_k) = \mathrm{E}([Z_j - \mathrm{E}(Z_j)]\,[Z_k - \mathrm{E}(Z_k)]),$$
is the **covariance** between variables j and k.

Summarizing multivariate data

- When $j = k$, we have

$$\begin{aligned}\operatorname{cov}(Z_j, Z_j) &= \mathrm{E}\left([Z_j - \mathrm{E}(Z_j)]^2\right) \\ &= \operatorname{var}(Z_j),\end{aligned}$$

  which is the **variance** for variable j.

- We can compute the **(Pearson) correlation** between variable j and k, by the definition:

$$\operatorname{cor}(Z_j, Z_k) = \frac{\operatorname{cov}(Z_j, Z_k)}{\sqrt{\operatorname{var}(Z_j)\operatorname{var}(Z_k)}}.$$

Estimating the summaries

- From data $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n$, we can estimate the expectation $\mathrm{E}(\boldsymbol{Z})$ by the **sample mean vector**:

$$\bar{\boldsymbol{Z}} = \frac{\sum_{i=1}^{n} \boldsymbol{Z}_i}{n} = \begin{bmatrix} \frac{\sum_{i=1}^{n} Z_{i1}}{n} \\ \vdots \\ \frac{\sum_{i=1}^{n} Z_{id}}{n} \end{bmatrix},$$

  where $\mathrm{E}(Z_j)$ is estimated by $\bar{Z}_j = n^{-1} \sum_{i=1}^{n} Z_{ij}$, for each variable j.

- Similarly, we can estimate $\mathrm{cov}(\boldsymbol{Z})$ by the **sample covariance matrix** $\boldsymbol{S}$, where the jth row and kth column element of $\boldsymbol{S}$ is the **sample covariance**:

$$S_{jk} = \frac{\sum_{i=1}^{n} \left(Z_{ij} - \bar{Z}_j\right)\left(Z_{ik} - \bar{Z}_k\right)}{n-1}.$$

Estimating the summaries

In R, we can use the `rowMeans()` and `cov()` functions to compute the sample mean vector and covariance matrix, respectively. In the case of the Adelie penguins, we have:

```
colMeans(data)

## [1]   38.79139   18.34636  189.95364 3700.66225
```

```
cov(data)

##             [,1]       [,2]        [,3]        [,4]
## [1,]   7.093725   1.268602    5.674265    670.3557
## [2,]   1.268602   1.480237    2.447497    321.4358
## [3,]   5.674265   2.447497   42.764503   1404.0309
## [4,] 670.355740 321.435762 1404.030905 210282.8918
```

Estimating the summaries

We can also use the `cor()` function to calculate the sample correlation matrix $\boldsymbol{R}$, where the jth row and kth column contains the sample correlation:

$$R_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}S_{kk}}}.$$

```r
cor(data)

##           [,1]      [,2]      [,3]      [,4]
## [1,] 1.0000000 0.3914917 0.3257847 0.5488658
## [2,] 0.3914917 1.0000000 0.3076202 0.5761382
## [3,] 0.3257847 0.3076202 1.0000000 0.4682017
## [4,] 0.5488658 0.5761382 0.4682017 1.0000000
```

The normal family of distributions

- For univariate random variable $Z \in \mathbb{R}$, we write

$$Z \sim \mathrm{N}\left(\mu, \sigma^2\right)$$

to denote that it is **normally distributed** with **mean and variance parameters** $\mu = \mathrm{E}(Z) \in \mathbb{R}$ and $\sigma^2 = \mathrm{var}(Z) \in (0, \infty)$, respectively.

- We say that $Z \sim \mathrm{N}\left(\mu, \sigma^2\right)$, if the probability of Z being in the interval (a, b) is given by the expression

$$\Pr(a \leq Z \leq b) = \int_a^b \phi\left(z; \mu, \sigma^2\right) \mathrm{d}z,$$

where

$$\phi\left(z; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(z-\mu)^2}{\sigma^2}\right]$$

is the **normal density function**, with parameters μ and σ^2.

Normal probabilities

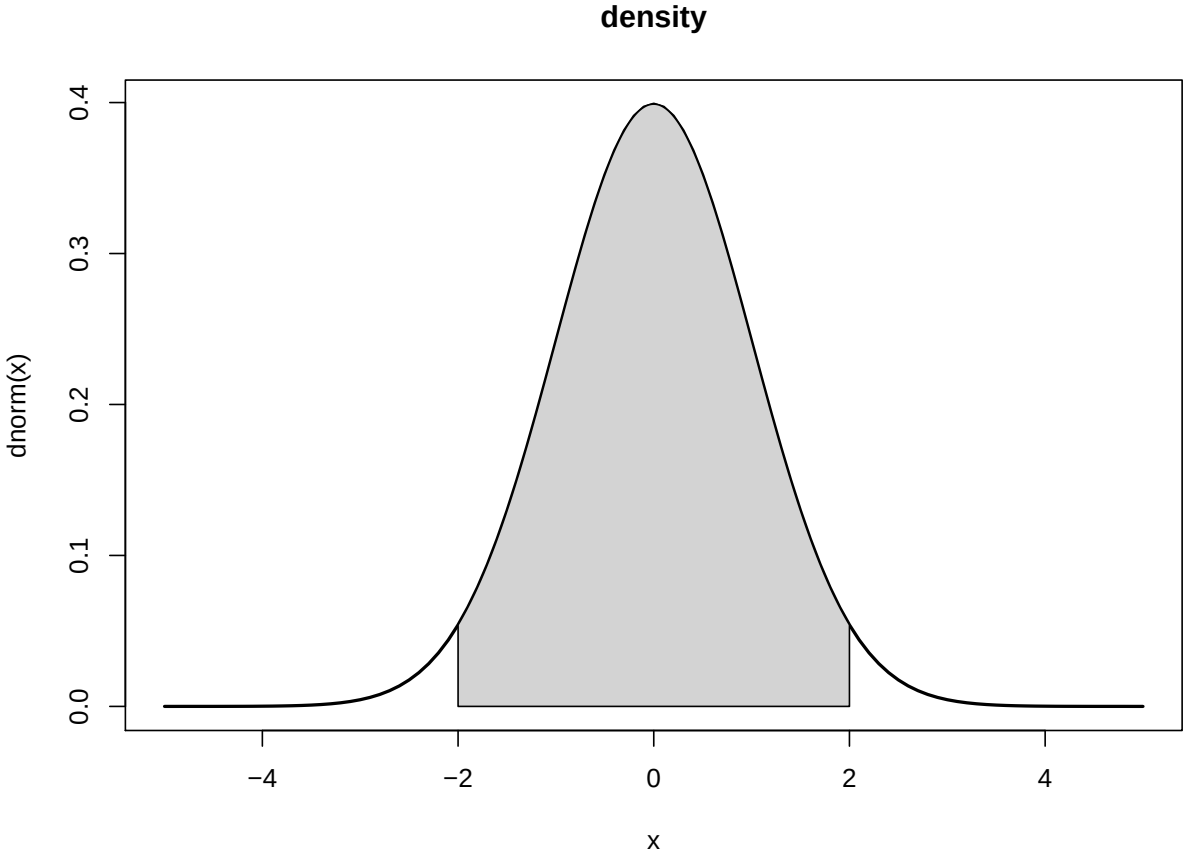

Figure: Probability that $Z \sim \mathrm{N}\left(\mu = 0, \sigma^2 = 1\right)$ is in the interval $(-2, 2)$.

Multivariate normal distributions

- For a multivariate random variable $\boldsymbol{Z} \in \mathbb{R}^d$, we write

$$\boldsymbol{Z} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

  to denote that it is **(multivariate) normally distributed** with **mean vector and covariance matrix parameters** $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{Z}) \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{Z})$, respectively.
- We say that $Z \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if the probability of Z being in the set $\mathbb{A}$ is given by the expression

$$\Pr(Z \in \mathbb{A}) = \int_{\mathbb{A}} \phi_d(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \mathrm{d}\boldsymbol{z},$$

  where

$$\phi_d(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right]$$

  is the **multivariate normal density function**, with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Normal probabilities

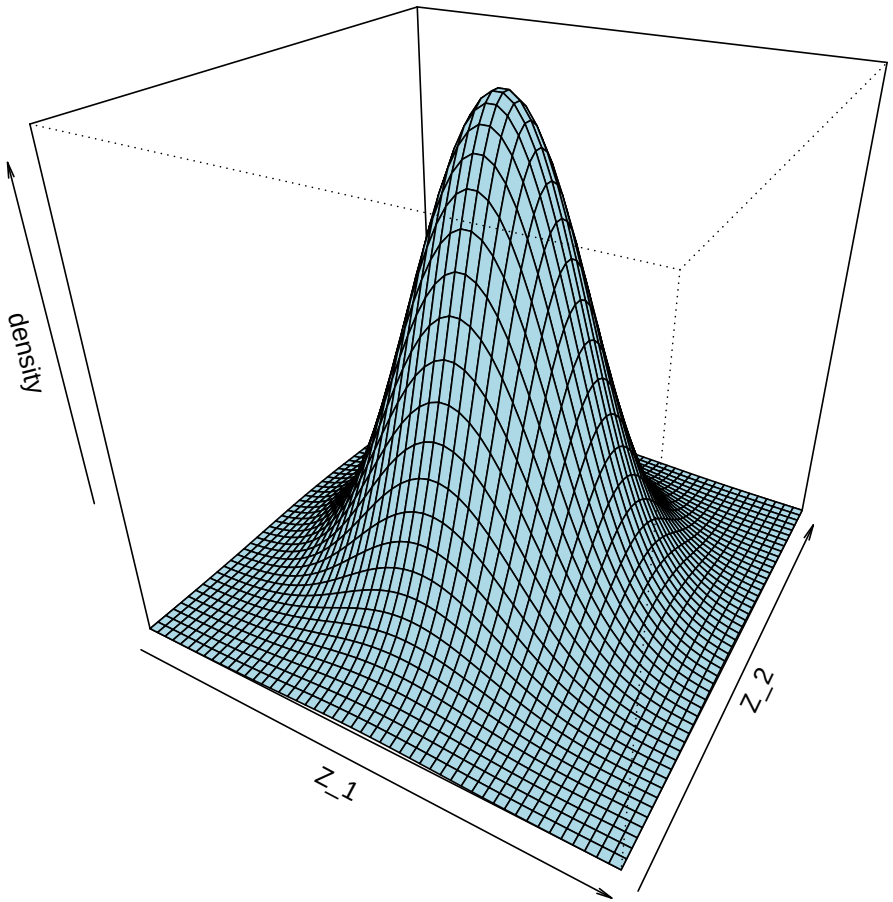

Figure: The multivariate normal density function $\phi_2(\boldsymbol{z}; \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I}_2)$, where $\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Fitted multivariate normal

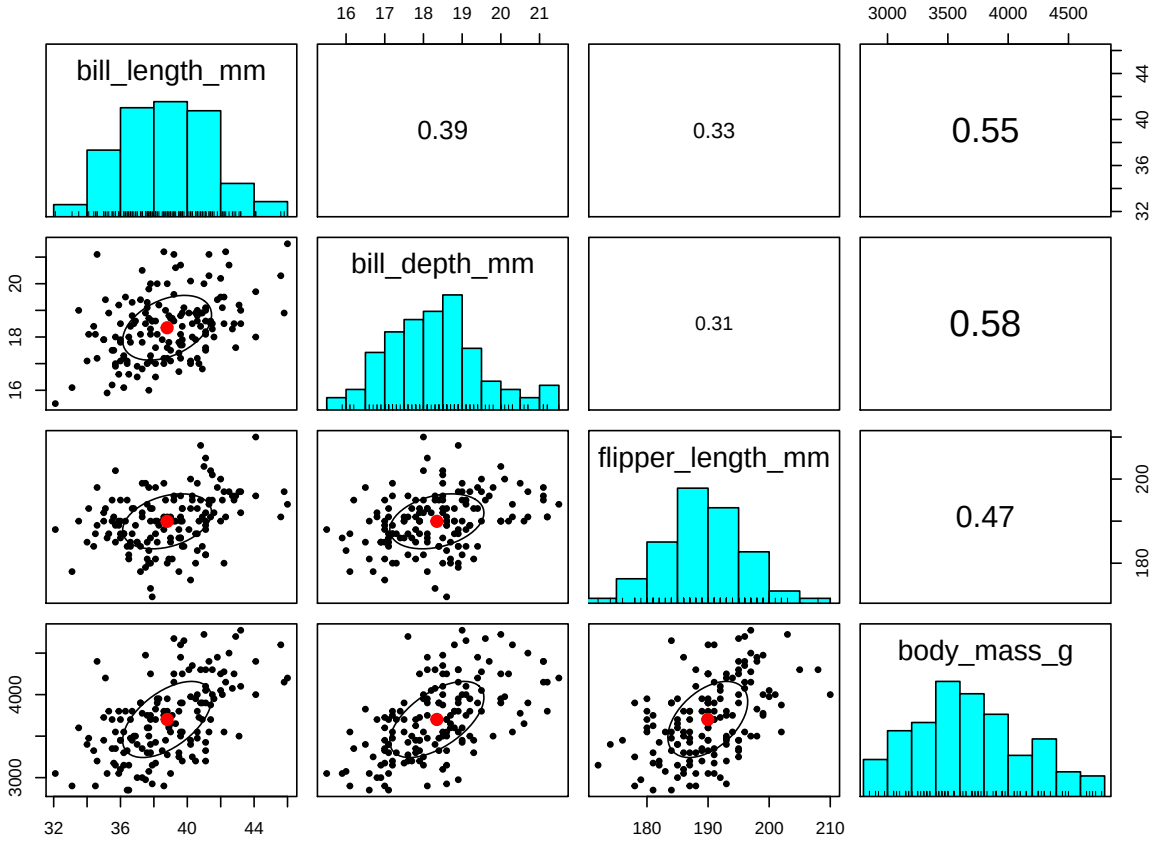

Figure: Mean and covariance/correlation estimates for the Adelie penguin data.

Closure under scalings and shifts

- If $\boldsymbol{A}$ and $\boldsymbol{b}$ are compatible matrices, and

$$\boldsymbol{Z} \sim \mathrm{N}_d\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$

then

$$\boldsymbol{A}\boldsymbol{Z} + \boldsymbol{b} \sim \mathrm{N}_d\left(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\right).$$

- This generalizes the univariate normal property that

$$aZ + b \sim \mathrm{N}\left(a\mu + b, a^2\sigma^2\right),$$

if

$$Z \sim \mathrm{N}\left(\mu, \sigma^2\right).$$

Closure under marginalization

- If $\boldsymbol{X} \in \mathbb{R}^{d_{\boldsymbol{X}}}$ and $\boldsymbol{Y} \in \mathbb{R}^{d_{\boldsymbol{Y}}}$, where $d_{\boldsymbol{X}} + d_{\boldsymbol{Y}} = d$ and

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim N_d \left(\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\mu}_{\boldsymbol{Y}} \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{XX}} & \boldsymbol{\Sigma}_{\boldsymbol{XY}} \\ \boldsymbol{\Sigma}_{\boldsymbol{XY}} & \boldsymbol{\Sigma}_{\boldsymbol{YY}} \end{bmatrix} \right),$$

  then $\boldsymbol{Y} \sim \mathrm{N}_{d_{\boldsymbol{Y}}} (\boldsymbol{\mu}_{\boldsymbol{Y}}, \boldsymbol{\Sigma}_{\boldsymbol{YY}})$.

- In particular, if

$$\boldsymbol{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_d \end{bmatrix} \sim \mathrm{N}_d \left(\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix} \right),$$

  then $Z_j \sim \mathrm{N}\left(\mu_j, \sigma^2 = \sigma_{jj}\right)$.

Closure under convolution

- If $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are independently multivariate normally distributed, with laws

$$\boldsymbol{Z}_i \sim \mathrm{N}_d\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right),$$

then

$$\sum_{i=1}^{n} \boldsymbol{Z}_i \sim \mathrm{N}_d\left(\sum_{i=1}^{n} \boldsymbol{\mu}_i, \sum_{i=1}^{n} \boldsymbol{\Sigma}_i\right).$$

- This generalizes the univariate normal property that

$$\sum_{i=1}^{n} Z_i \sim \mathrm{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right),$$

if $Z_1, \ldots, Z_n$ are independent normal random variables, where

$$Z_j \sim \mathrm{N}\left(\mu_j, \sigma_j^2\right),$$

for each $j \in \{1, \ldots, n\}$.

Closure under conditioning

- If $\boldsymbol{X} \in \mathbb{R}^{d_X}$ and $\boldsymbol{Y} \in \mathbb{R}^{d_Y}$, where $d_{\boldsymbol{X}} + d_{\boldsymbol{Y}} = d$ and

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim N_d \left(\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\mu}_{\boldsymbol{Y}} \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{XX}} & \boldsymbol{\Sigma}_{\boldsymbol{XY}} \\ \boldsymbol{\Sigma}_{\boldsymbol{XY}} & \boldsymbol{\Sigma}_{\boldsymbol{YY}} \end{bmatrix} \right),$$

  then

$$\boldsymbol{Y} | \boldsymbol{X} = \boldsymbol{x} \sim \mathrm{N}_{d_{\boldsymbol{Y}}} \left(\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}} \right),$$

  where

$$\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{X}} = \boldsymbol{\mu}_{\boldsymbol{Y}} + \boldsymbol{\Sigma}_{\boldsymbol{XY}} \boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-1} (\boldsymbol{x}_{\boldsymbol{X}} - \boldsymbol{\mu}_{\boldsymbol{X}}),$$

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}} = \boldsymbol{\Sigma}_{\boldsymbol{YY}} - \boldsymbol{\Sigma}_{\boldsymbol{XY}} \boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{XY}}.$$

- Consider the case where $\boldsymbol{\Sigma}_{\boldsymbol{XY}} = \boldsymbol{0}$, which implies independence between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Hotelling's two sample T^2 test

- Let $\boldsymbol{X}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_{\boldsymbol{X}}}$ and $\boldsymbol{Y}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_{\boldsymbol{Y}}}$ be **IID (independent and identically distributed)** random samples, such that

$$\boldsymbol{X} \sim \mathrm{N}_d\left(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}\right) \text{ and } \boldsymbol{Y} \sim \mathrm{N}_d\left(\boldsymbol{\mu}_{\boldsymbol{Y}}, \boldsymbol{\Sigma}\right).$$

- Then, we can test the hypotheses

$$\mathrm{H}_0 : \boldsymbol{\mu}_{\boldsymbol{X}} = \boldsymbol{\mu}_{\boldsymbol{Y}} \text{ versus } \mathrm{H}_1 : \boldsymbol{\mu}_{\boldsymbol{X}} \neq \boldsymbol{\mu}_{\boldsymbol{Y}}$$

via the T^2 **test statistic**

$$T^2 = \frac{n_{\boldsymbol{X}} n_{\boldsymbol{Y}}}{n_{\boldsymbol{X}} + n_{\boldsymbol{Y}}} \left(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}\right)^\top \boldsymbol{S}_{\text{pooled}}^{-1} \left(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}\right),$$

$$\boldsymbol{S}_{\text{pooled}} = \frac{\left(n_{\boldsymbol{X}} - 1\right) \boldsymbol{S}_{\boldsymbol{X}} + \left(n_{\boldsymbol{Y}} - 1\right) \boldsymbol{S}_{\boldsymbol{Y}}}{n_{\boldsymbol{X}} + n_{\boldsymbol{Y}} - 2},$$

where under the null hypothesis:

$$T^2 \sim T^2\left(d, n_{\boldsymbol{X}} + n_{\boldsymbol{Y}} - 2\right).$$

More penguins

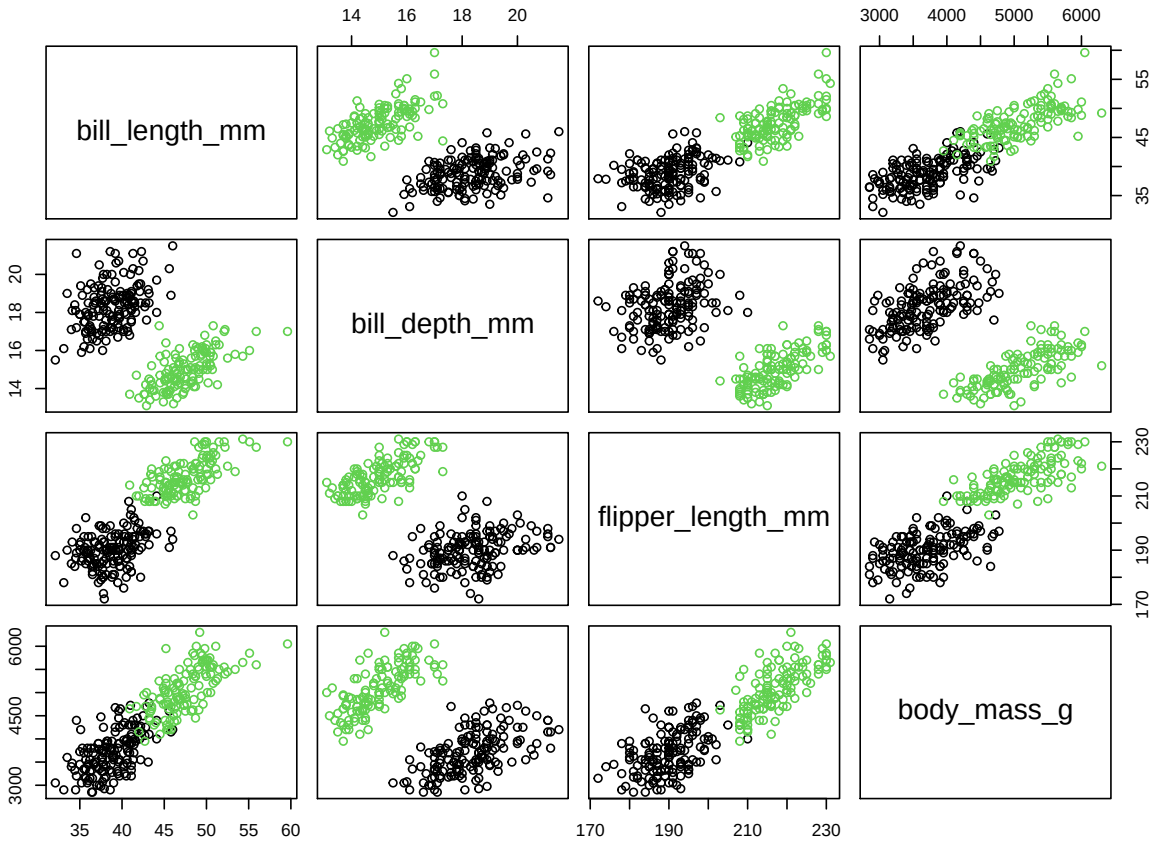

Figure: $n_{\boldsymbol{X}} = 151$ sets of Adelie penguin (black) measurements and $n_{\boldsymbol{Y}} = 123$ sets of Gentoo penguin (green) measurements from the `penguins` data set.

More penguins

Figure: An independent sample of Gentoo penguins.

A T^2 test for difference in means

In R, we can use the `hotel.test()` function from the `Hotelling` package in order to test for the difference in the mean measurements between the Adelie and Gentoo penguins:

```
colMeans(Adelie)

## [1]   38.79139   18.34636  189.95364 3700.66225

colMeans(Gentoo)

## [1]   47.50488   14.98211  217.18699 5076.01626

print(hotelling.test(Adelie,Gentoo))

## Test stat:  1134
## Numerator df:  4
## Denominator df:  269
## P-value:  0
```

A test for multivariate normality

- Suppose that $\boldsymbol{Z}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are IID random variables, and that we wish to test the null hypothesis

$$\mathrm{H}_0 : \boldsymbol{Z} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ for some } \boldsymbol{\mu} \text{ and } \boldsymbol{\Sigma}.$$

- We can consider the so-called **energy statistic for normality** (Szekely and Rizzo, 2005):

$$\mathscr{E} = n\left(\frac{2}{n}\sum_{i=1}^{n}\left\|\tilde{\boldsymbol{Z}}_i - \boldsymbol{W}\right\| - 2\frac{\Gamma\left[(d+1)/2\right]}{\Gamma\left[d/2\right]} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\tilde{\boldsymbol{Z}}_i - \tilde{\boldsymbol{Z}}_j\right\|\right),$$

where $\boldsymbol{W} \sim \mathrm{N}_d(\boldsymbol{0}, \mathbf{I}_d)$, $\|\cdot\|$ is the Euclidean norm, Γ is the Gamma function and for each $i \in \{1, \ldots, n\}$

$$\tilde{\boldsymbol{Z}}_i = \boldsymbol{S}^{-1/2}\left(\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}\right).$$

- The test statistic is known to be **consistent against all non-normal alternatives**.

Testing the normality of penguin data

We can use the `mvnorm.test()` function from the `energy` package to conduction a bootstrap tests for normality.

```
mvnorm.etest(Adelie, R = 999)

##
##  Energy test of multivariate normality: estimated parameters
##
## data:  x, sample size 151, dimension 4, replicates 999
## E-statistic = 1.119, p-value = 0.1231

mvnorm.etest(Gentoo, R = 999)

##
##  Energy test of multivariate normality: estimated parameters
##
## data:  x, sample size 123, dimension 4, replicates 999
## E-statistic = 1.2557, p-value = 0.01201
```

A test for equality of distributions

- Let $\boldsymbol{X}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_{\boldsymbol{X}}}$ and $\boldsymbol{Y}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_{\boldsymbol{Y}}}$ be IID random samples, from unknown distributions, and that we wish to test the null hypothesis:

$$\mathrm{H}_0: \text{ The distributions of } \boldsymbol{X} \text{ and } \boldsymbol{Y} \text{ are equal.}$$

- We consider the two-sample **energy statistic** of Szekely and Rizzo (2004): $\mathscr{E}$, where

$$\frac{n_{\boldsymbol{X}} + n_{\boldsymbol{Y}}}{n_{\boldsymbol{X}} n_{\boldsymbol{Y}}} \mathscr{E} = \frac{2}{n_{\boldsymbol{X}} n_{\boldsymbol{Y}}} \sum_{i=1}^{n_{\boldsymbol{X}}} \sum_{j=1}^{n_{\boldsymbol{Y}}} \|\boldsymbol{X}_i - \boldsymbol{Y}_j\| - \frac{1}{n_{\boldsymbol{X}}^2} \sum_{i=1}^{n_{\boldsymbol{X}}} \sum_{j=1}^{n_{\boldsymbol{X}}} \|\boldsymbol{X}_i - \boldsymbol{X}_j\| - \frac{1}{n_{\boldsymbol{Y}}^2} \sum_{i=1}^{n_{\boldsymbol{Y}}} \sum_{j=1}^{n_{\boldsymbol{Y}}} \|\boldsymbol{Y}_i - \boldsymbol{Y}_j\|.$$

- The test statistic is known to be **consistent against all alternatives** that satisfy $\|\boldsymbol{X}\| < \infty$ and $\|\boldsymbol{Y}\| < \infty$.

Testing the equality of distribution in the penguin data

We can use the `eqdist.etest()` function from the `energy` package to conduction a bootstrap test for whether the Adelie and Gentoo penguin samples arise from the same distribution.

```
eqdist.etest(rbind(Adelie,Gentoo),c(nrow(Adelie),nrow(Gentoo)),R=999)

##
##  Multivariate 2-sample E-test of equal distributions
##
## data:  sample sizes 151 123, replicates 999
## E-statistic = 113363, p-value = 0.001
```

Even more penguins

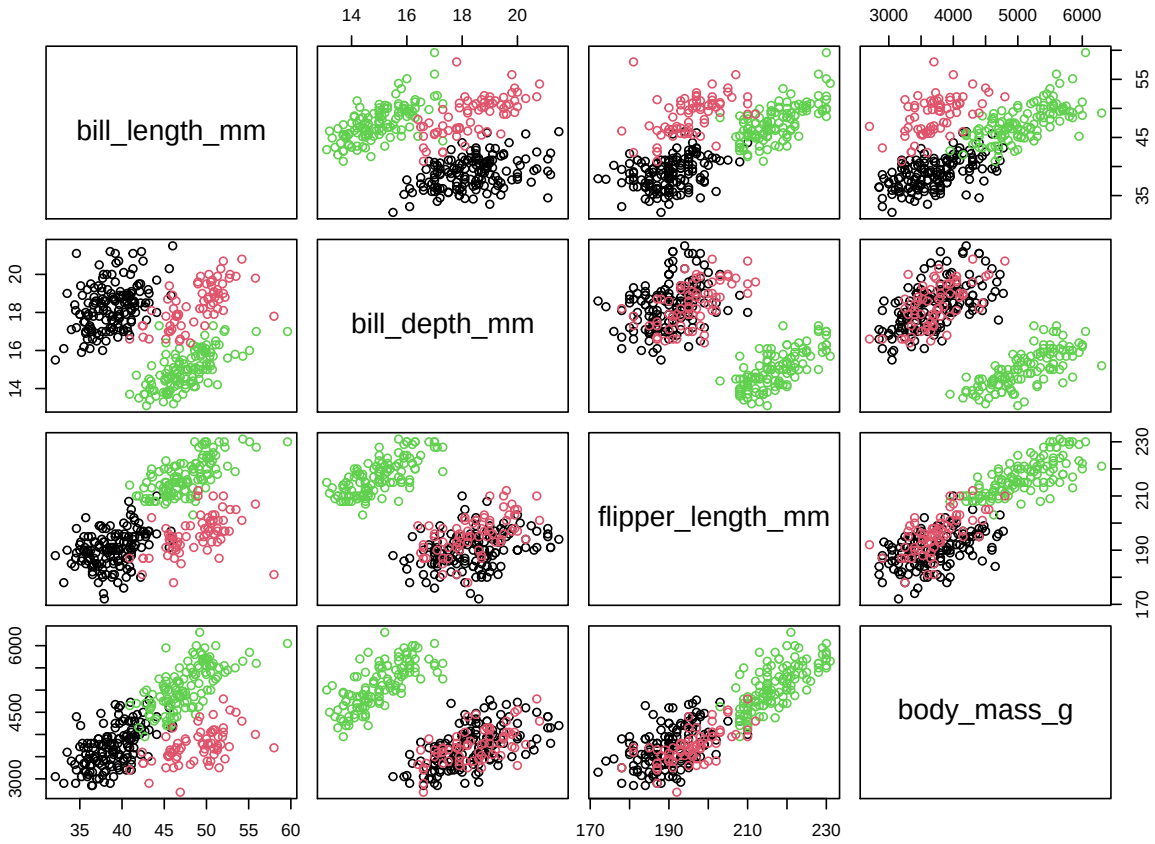

Figure: Measurements from 151 Adelie penguins (black), 123 Gentoo penguin (green), and 68 Chinstrap penguins (red), from the `penguins` data set.

Even more penguins

Figure: An independent sample of Chinstrap penguins.

Additional information

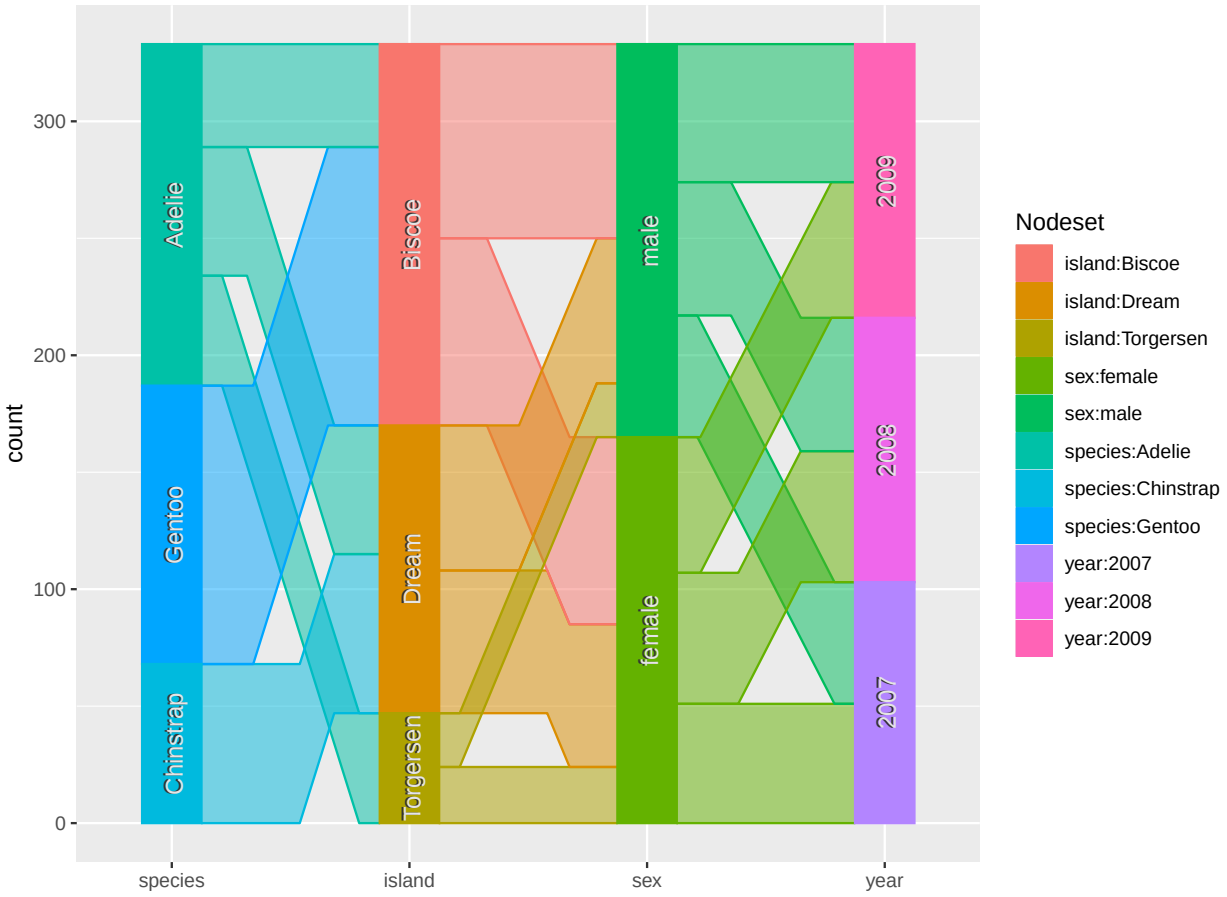

Figure: Covariate information for the penguins.

Multivariate regression

- Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \in \mathbb{R}^{d_{\boldsymbol{Y}}}$ be independent **response vectors** and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^{d_{\boldsymbol{X}}}$ be non-random **covariates (explanatory) vectors**.
- Suppose that for each $i \in \{1, \ldots, n\}$,

$$\boldsymbol{Y}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{x}_i + \boldsymbol{E}_i$$

$$\begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{id_{\boldsymbol{Y}}} \end{bmatrix} = \begin{bmatrix} \alpha_1 + B_{11}x_{i1} + \cdots + B_{1d_{\boldsymbol{X}}}x_{id_{\boldsymbol{X}}} \\ \vdots \\ \alpha_{d_{\boldsymbol{Y}}} + B_{d_{\boldsymbol{Y}}1}x_{i1} + \cdots + B_{d_{\boldsymbol{Y}}d_{\boldsymbol{X}}}x_{id_{\boldsymbol{X}}} \end{bmatrix} + \begin{bmatrix} E_{i1} \\ \vdots \\ E_{id_{\boldsymbol{Y}}} \end{bmatrix}$$

  where $\boldsymbol{E}_i \sim \mathrm{N}_{d_{\boldsymbol{Y}}}(\mathbf{0}, \boldsymbol{\Sigma})$ are a **random noise**, $\boldsymbol{\alpha}$ is a **vector of intercepts** and $\mathbf{B}$ is a **matrix of regression coefficients**.
- By closure under shifting, we have

$$\boldsymbol{Y}_i \sim \mathrm{N}_{d_{\boldsymbol{Y}}}(\boldsymbol{\alpha} + \mathbf{B}\boldsymbol{x}_i, \boldsymbol{\Sigma}).$$

Multivariate regression

- We wish to estimate $\mathbf{B}$ from the data, which can be done using the **ordinary least squares estimator**:

$$\begin{bmatrix} \hat{\boldsymbol{\alpha}}^\top \\ \hat{\mathbf{B}} \end{bmatrix} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y},$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & \boldsymbol{x}_1^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n^\top \end{bmatrix} \text{ and } \mathbf{Y} = \begin{bmatrix} \boldsymbol{Y}_1^\top \\ \vdots \\ \boldsymbol{Y}_n^\top \end{bmatrix}.$$

- Using the least squares estimator, we wish to test hypotheses of the form:

H_0 : Variable j has no effect on any dimension of the response.

Least squares estimation

We can use the `lm()` function in the `R` to compute the least squares estimates for the multivariate regression model

```
LM <- lm(cbind(bill_length_mm,bill_depth_mm,
               flipper_length_mm,body_mass_g)~.,
        data=data)
res <- coef(LM); colnames(res) <- NULL; res

##                         [,1]        [,2]       [,3]        [,4]
## (Intercept)      37.01488826 17.75987075 181.938201 3371.126516
## speciesChinstrap 10.34685159  0.18699767   6.116480   38.004609
## speciesGentoo     8.54973222 -3.40234890  28.571283 1374.656277
## islandDream      -0.52081120 -0.16993070   1.545461  -13.763279
## islandTorgersen   0.09814743  0.07825233   3.199515    6.470768
## sexmale           3.69752879  1.50619199   6.864346  667.682268
## year2008         -0.27243868 -0.21816805   4.028082    5.093966
## year2009          0.61269871 -0.14646949   4.887380    7.170751
```

Multivariate analysis of variance (MANOVA)

To conduct inference regarding the effects of the covariates on the responses, we use the `Anova()` function from the `car` package of Fox and Weisberg (2011).

```
Anova(LM)

##
## Type II MANOVA Tests: Pillai test statistic
##         Df test stat approx F num Df den Df    Pr(>F)
## species  2   1.50814  247.597      8    646 < 2.2e-16 ***
## island   2   0.04072    1.678      8    646    0.1004
## sex      1   0.64319  145.111      4    322 < 2.2e-16 ***
## year     2   0.20572    9.258      8    646 4.014e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Probabilistic principal component analysis (PPCA)

- In Tipping and Bishop (1999), the authors consider that we observe IID random variables $\boldsymbol{Y}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \in \mathbb{R}^d$, where

$$\boldsymbol{Y} | \boldsymbol{X} = \boldsymbol{x} \sim \mathrm{N}_d \left(\boldsymbol{\alpha} + \mathbf{W}\boldsymbol{x}, \sigma^2 \mathbf{I}_d \right),$$

  $\mathbf{W}$ is a **matrix of weights** and

$$\mathbf{X} \sim \mathrm{N}_p \left(\mathbf{0}, \mathbf{I}_p \right),$$

  is an unobserved **latent representation vector** for some lower dimension $p < d$.

- By the closure under conditioning, we have

$$\boldsymbol{Y} \sim \mathrm{N}_d \left(\boldsymbol{\alpha}, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I} \right).$$

Dimensionality reduction

- Using the data $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, we can compute the maximum likelihood estimates for the unknown parameters $\boldsymbol{\alpha}$, $\mathbf{W}$, and σ^2 explicitly, or via the incremental expectation–maximization algorithm.
- Upon obtaining the maximum likelihood estimates, we can summarize the information in $\boldsymbol{Y}_i$ by the estimated conditional expectation of $\boldsymbol{X}_i$, for each $i \in \{1, \ldots, n\}$:

$$\hat{\boldsymbol{x}}_i = \hat{\mathrm{E}}(\boldsymbol{X}_i | \boldsymbol{Y}_i = \boldsymbol{y}_i) = \left(\hat{\mathbf{W}}^\top \hat{\mathbf{W}} + \hat{\sigma}^2 \mathbf{I}_p\right) \hat{\mathbf{W}}^\top (\boldsymbol{Y}_i - \hat{\boldsymbol{\alpha}}).$$

- It can be shown that $\hat{\boldsymbol{x}}_i$ approaches the first p **principal component projections** as $\sigma^2 \to 0$.

Conducting PPCA

We can use the `pca()` function from the `pcaMethods` package to conduct PPCA on the Adelie penguin data. Here, we summarize the $d = 4$ measurements into $p = 2$ latent dimensions.

```
PPCA <- pca(scale(Adelie),method='ppca'); summary(PPCA)

## ppca calculated PCA
## Importance of component(s):
##                    PC1    PC2
## R2              0.5815 0.1779
## Cumulative R2 0.5815 0.7595
```

Latent representation and weights

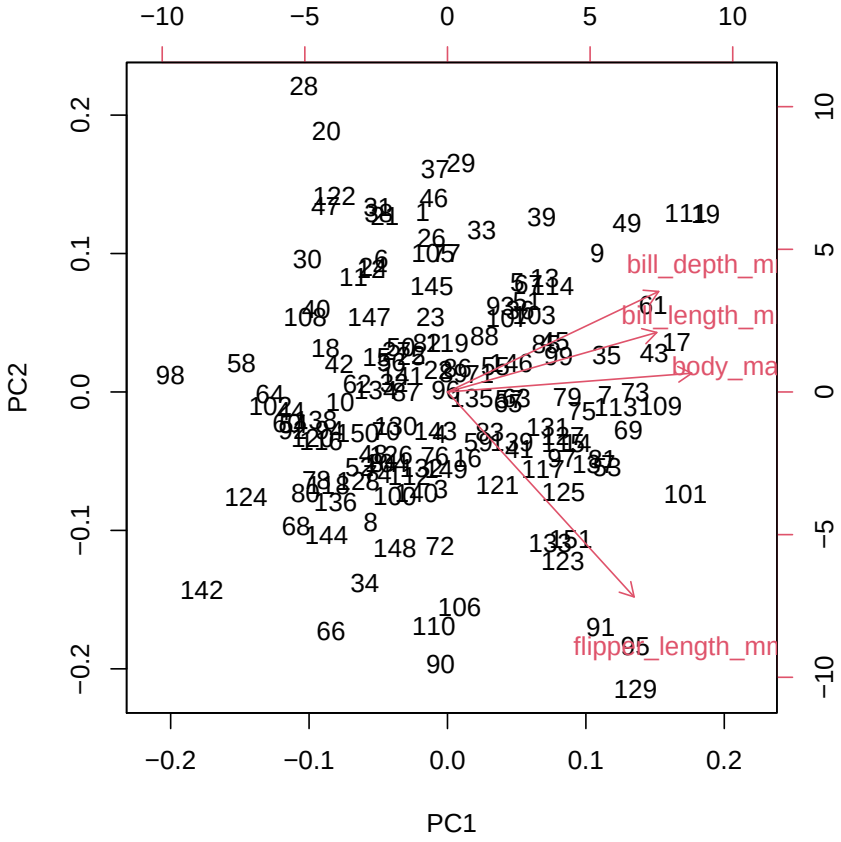

Figure: Latent variable projections are displayed .

Even more penguins

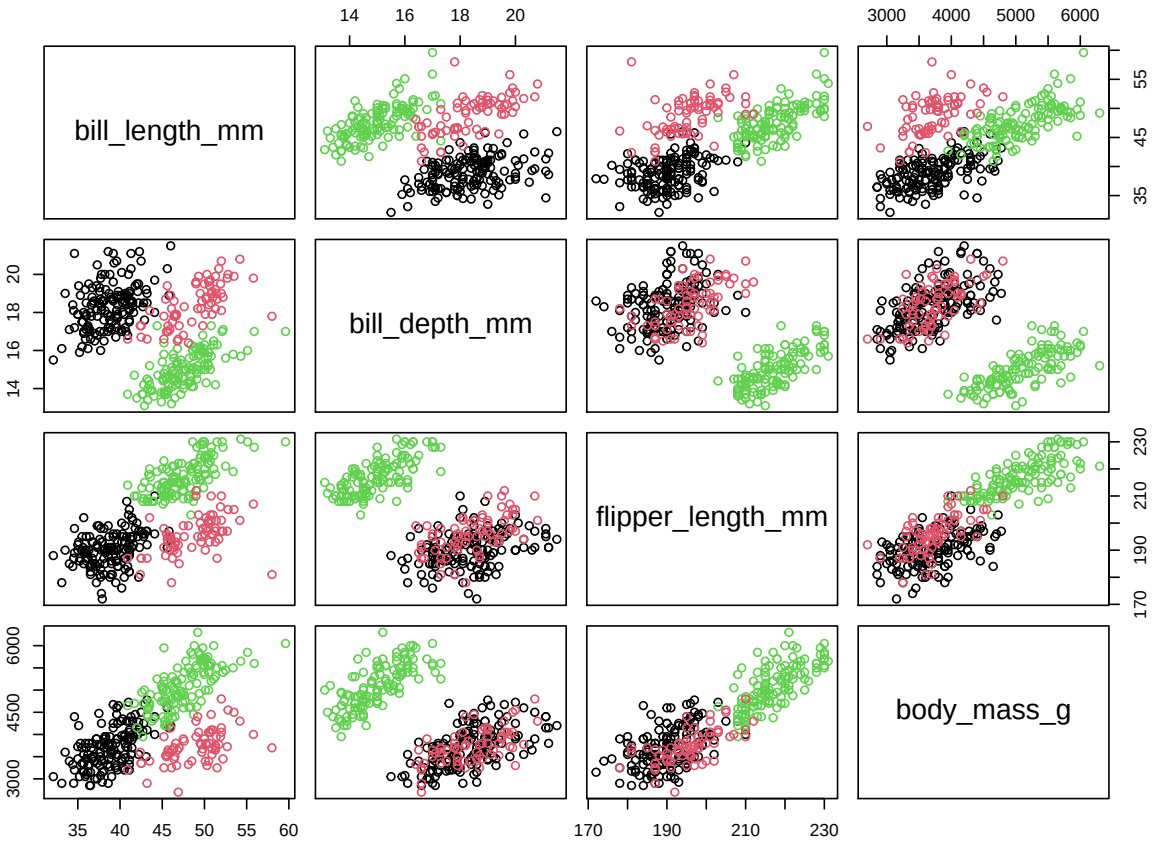

Figure: Measurements from 151 Adelie penguins (black), 123 Gentoo penguin (green), and 68 Chinstrap penguins (red), from the `penguins` data set.

Classification

- Let $\boldsymbol{Y}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \in \mathbb{R}^d$ be independent response random variables, each paired with a **label** $L, L_1, \ldots, L_n \in \{1, \ldots, K\}$ that are IID, where K is the total **number of categories**.
- For any $l \in \{1, \ldots, K\}$,

$$\Pr(L = l) = \pi_l \geq 0, \ \sum_{l=1}^{K} \pi_l = 1.$$

- Conditional on its label, suppose that the response is normally distributed, in the sense that

$$\boldsymbol{Y} | L = l \sim \mathrm{N}_d(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l),$$

where $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$ are parameters that determine the distribution of observations in category l.

Classification

- Suppose that we observe the response $\boldsymbol{Y}$ but do not observe its label L.
- For each $l \in \{1, \ldots, K\}$, we can compute the **a posteriori probability** that $L = l$, given our observation $\boldsymbol{Y}$ using Bayes rule:

$$\Pr(L = l | \boldsymbol{Y}) = \frac{\pi_l \phi_d(\boldsymbol{Y}; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}{\sum_{k=1}^{K} \pi_k \phi_d(\boldsymbol{Y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}.$$

- By the **maximum a posteriori principal**, we estimate L by the category that has the highest probability of occurring:

$$\hat{L} = \underset{l \in \{1, \ldots, K\}}{\arg\max} \Pr(L = l | \boldsymbol{Y}).$$

Quadratic discriminant analysis

- **Quadratic discriminant analysis (QDA)** uses the maximum a posteriori rule, but with the parameters estimated by their sample values:

$$\hat{\pi}_l = \frac{1}{n} \sum_{i=1}^{n} [\![L_i = l]\!] ,$$

$$\hat{\boldsymbol{\mu}}_l = \frac{1}{\sum_{i=1}^{n} [\![L_i = l]\!]} \sum_{i=1}^{n} [\![L_i = l]\!] \, \boldsymbol{Y}_i ,$$

$$\hat{\boldsymbol{\Sigma}}_l = \frac{1}{\sum_{i=1}^{n} [\![L_i = l]\!] - 1} \sum_{i=1}^{n} [\![L_i = l]\!] \, (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_l)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_l)^\top ,$$

  for each $l \in \{1, \ldots, K\}$.

- If we consider instead that $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K$, then we obtain the **linear discriminant analysis (LDA)** method.

Conducting QDA

We can use the `McclustDA()` function from the `mclust` package (Scrucca et al., 2016) to fit the QDA classifier to the `penguins` data.

```
MC <- MclustDA(data[,c(3:6)],data$species,G=1,modelNames = 'XXX')
table(predict(MC)$class,data$species)

##
##              Adelie Chinstrap Gentoo
##   Adelie        144         2      0
##   Chinstrap       2        66      0
##   Gentoo          0         0    119
```

Estimated model

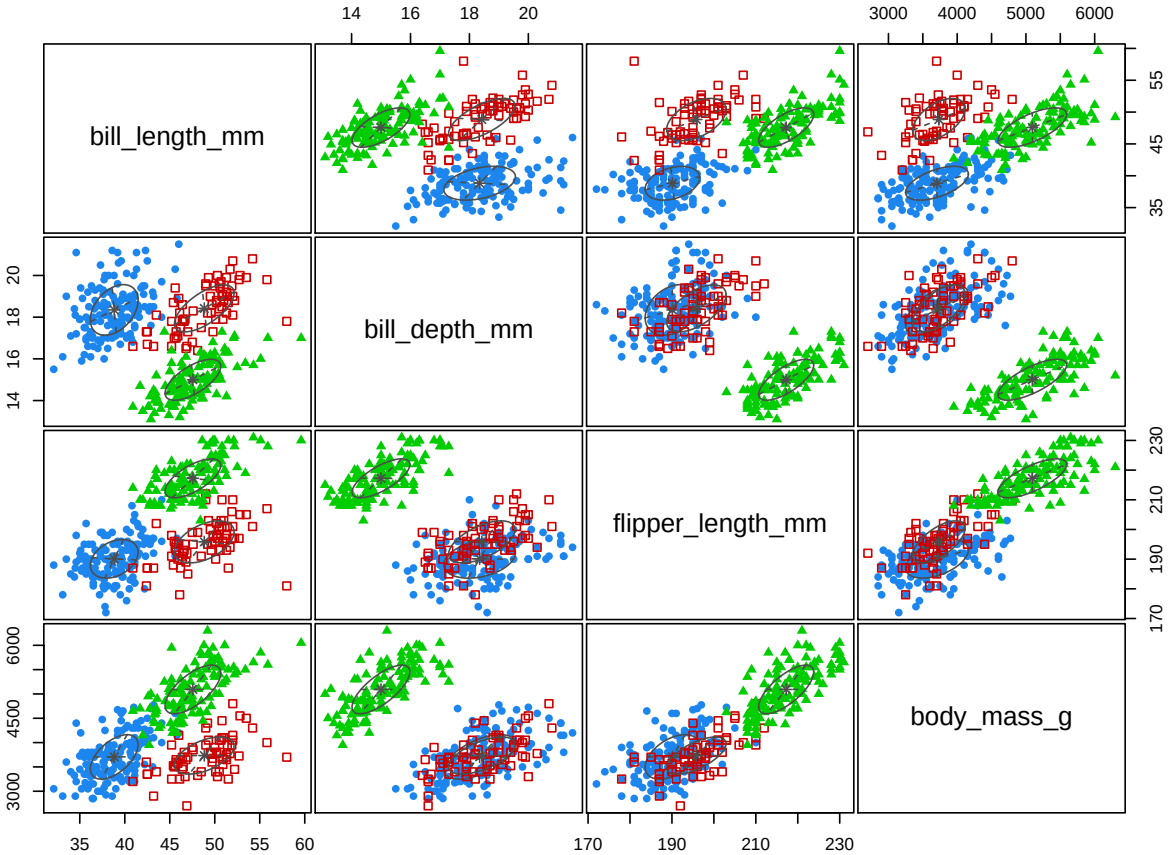

Figure: Fitted normal distributions for each of the penguin species. Blue, green and red indicate Adelie, Gentoo, and Chinstrap penguins, respectively.

Classification error

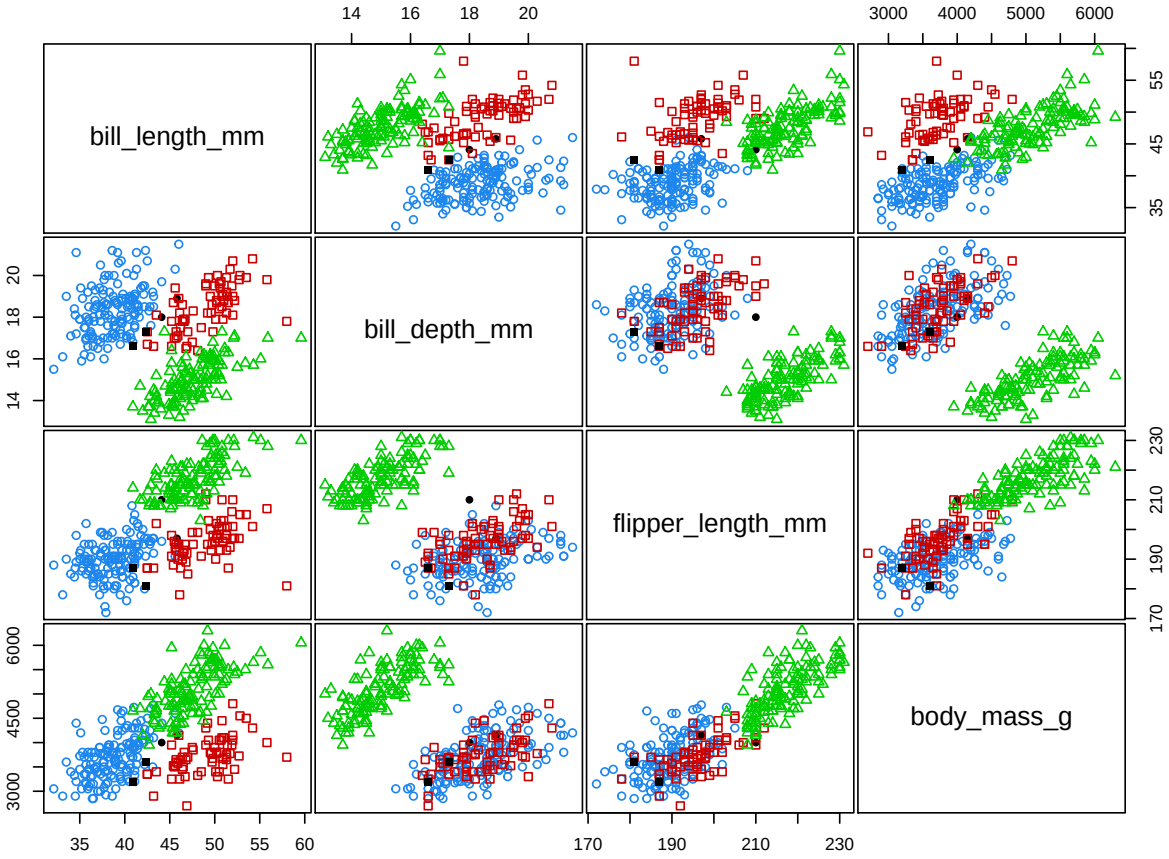

Figure: Incorrect decisions made from the maximum a posteriori rule. Blue, green and red indicate Adelie, Gentoo, and Chinstrap penguins, respectively.

Penguins without labels

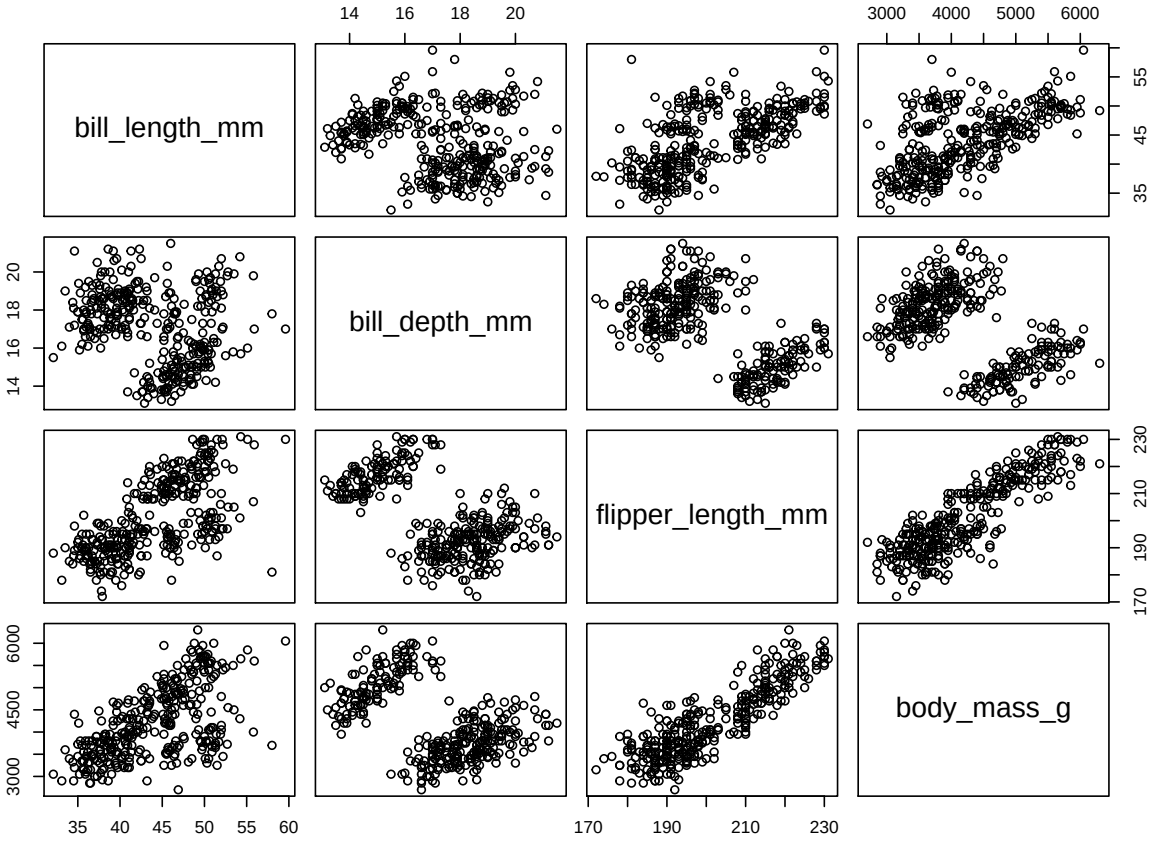

Figure: $n = 333$ measurements from an unknown number of species of penguins.

Cluster analysis

- Let $\boldsymbol{Y}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \in \mathbb{R}^d$ be independent response random variables, each paired with an unobserved **label** $L, L_1, \ldots, L_n \in \{1, \ldots, K\}$ that are IID, where K is the total **number of clusters (component/subpopulation)**.
- For any $l \in \{1, \ldots, K\}$,

$$\Pr(L = l) = \pi_l \geq 0, \ \sum_{l=1}^{K} \pi_l = 1.$$

- Conditional on its label, suppose that the response is normally distributed, in the sense that

$$\boldsymbol{Y} | L = l \sim \mathrm{N}_d(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l),$$

where $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$ are parameters that determine the distribution of observations in clusters l.

Cluster analysis

- Using the **law of total probability**, we have the fact that $\boldsymbol{Y}$ has distribution, characterized by the K **component finite mixture probability density**:

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \sum_{l=1}^{K} \pi_l \phi(\boldsymbol{y}; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l),$$

  where $\boldsymbol{\theta}$ contains the parameter elements π_l, $\boldsymbol{\mu}_l$, $\boldsymbol{\Sigma}_l$, for all $l \in \{1, \ldots, K\}$.

- We cluster the observations $\boldsymbol{Y}_i$ using the **maximum a posteriori rule**; that is, we say that observation i is allocated to cluster $\hat{L}_i \in \{1, \ldots, K\}$, if

$$\hat{L}_i = \underset{l \in \{1,\ldots,K\}}{\arg\max}\ \pi_l \phi_d(\boldsymbol{Y}; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) / f_{\boldsymbol{\theta}}(\boldsymbol{y}),$$

Estimating the finite mixture model

- Since $L_1, \ldots, L_n$ are unobserved, we can't estimate the elements of $\boldsymbol{\theta}$ in the same way as the QDA model.
- Instead, we must estimate $\boldsymbol{\theta}$ using only the information available from $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$.
- We can compute the maximum likelihood estimate for $\boldsymbol{\theta}$ using the **expectation–maximization (EM) algorithm**, as suggested in McLachlan and Krishnan (2008).
- The EM algorithm begins by proposing some initial estimate $\boldsymbol{\theta}^{(0)}$ for the elements of $\boldsymbol{\theta}$.

The EM algorithm

- For $r \geq 1$, at the rth E-step, we compute, for each l:

$$\begin{aligned}\tau_l^{(r)}(\boldsymbol{Y}_i) &= \mathrm{E}_{\boldsymbol{\theta}^{(r-1)}}\left[L_i | \boldsymbol{Y}_i\right] \\ &= \pi_l^{(r-1)} \phi_d\left(\boldsymbol{Y}_i; \boldsymbol{\mu}_l^{(r-1)}, \boldsymbol{\Sigma}_l^{(r-1)}\right) / f_{\boldsymbol{\theta}^{(r-1)}}(\boldsymbol{y}).\end{aligned}$$

- At the rth M-step, we compute the parameter updates $\boldsymbol{\theta}^{(r)}$, consisting of:

$$\pi_l^{(r)} = n^{-1} \sum_{i=1}^{n} \tau_l^{(r)}(\boldsymbol{Y}_i),$$

$$\boldsymbol{\mu}_l^{(r)} = \frac{1}{\sum_{i=1}^{n} \tau_l^{(r)}(\boldsymbol{Y}_i)} \sum_{i=1}^{n} \tau_l^{(r)}(\boldsymbol{Y}_i) \boldsymbol{Y}_i,$$

$$\boldsymbol{\Sigma}_l^{(r)} = \frac{1}{\sum_{i=1}^{n} \tau_l^{(r)}(\boldsymbol{Y}_i)} \sum_{i=1}^{n} \tau_l^{(r)}(\boldsymbol{Y}_i) \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_l^{(r)}\right) \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_l^{(r)}\right)^\top.$$

The EM algorithm

- We repeat the E- and M-steps until $r = R$, for some sufficiently large R at which stage we call $\boldsymbol{\theta}^{(R)}$ the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$.
- It is provable that $\boldsymbol{\theta}^{(r)}$ approaches a stationary point of the log-likelihood function

$$\mathscr{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i),$$

  as $r \to \infty$.
- We then estimate the **maximum a posteriori clustering rule** and allocate observation i into cluster

$$\hat{L}_i = \underset{l \in \{1,\ldots,K\}}{\arg\max}\ \hat{\pi}_l \phi_d\left(\boldsymbol{Y}; \hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}_l\right) / f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{y}).$$

Choosing K

- Thus far, we have assumed that is fixed; however, unless we have some prior knowledge regarding subpopulation structures, we must estimate K.
- Under regularity conditions, K can be consistently estimated by

$$\hat{K} = \underset{K \in \{1,2,\ldots\}}{\arg\min}\ \mathrm{BIC}_K,$$

where

$$\mathrm{BIC}_K = -2\mathscr{L}\left(\hat{\boldsymbol{\theta}}_K\right) + \dim_K \log(n)$$

is the **Bayesian information criterion**, where $\hat{\boldsymbol{\theta}}_K$ is the maximum likelihood estimator for the model with K clusters, and

$$\dim_K = Kd + K\frac{d(d+1)}{2} + K - 1.$$

Fitting the mixture model

We can use the `Mclust()` function from the `mclust` package (Scrucca et al., 2016) to fit the finite mixture model and select the optimal number of clusters K.

```
MC <- Mclust(data[,c(3:6)],G=1:10,modelNames = 'VVV')
table(predict(MC)$class,data$species)

##
##      Adelie Chinstrap Gentoo
##   1     144         3      0
##   2       2        65      0
##   3       0         0    119
```

BIC results

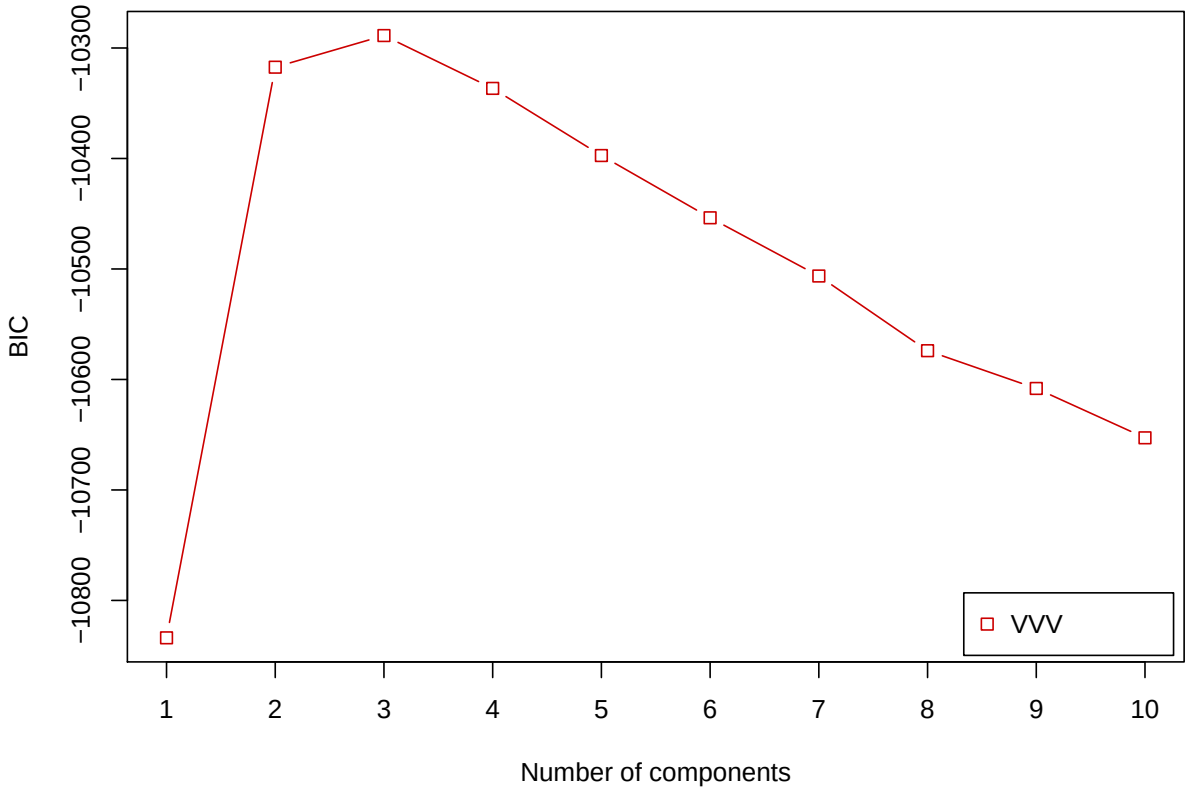

Figure: Negative BIC results for selecting the number of clusters K.

Clustering outcomes

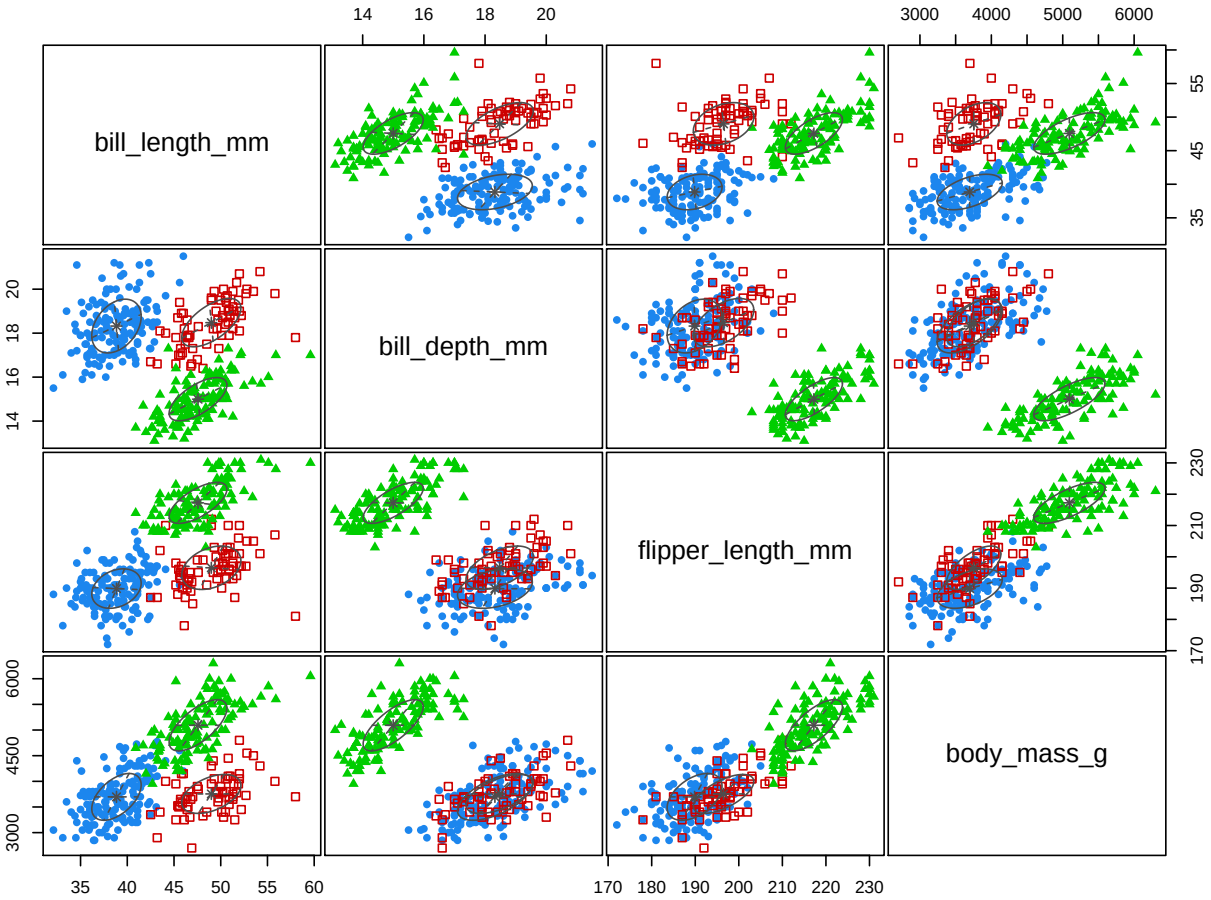

Figure: Cluster allocations using the maximum a posteriori rule.

References I

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage, Los Angeles.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm And Extensions*. Wiley, New York, 2 edition.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, 8:289–317.

Szekely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5:1249–1272.

Szekely, G. J. and Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, 93:58–80.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principle component analysis. *Journal of the Royal Statistical Society B*, 61:611–622.

Thank you!

Email: **h.nguyen5@latrobe.edu.au**

Website: **hiendn.github.io**