University of BRISTOL

Introduction to multivariate distributions

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.

UOB Open

What will we cover today?

- We will introduce the concept of a random vector.
- We will introduce the family of multivariate Gaussian distributions.
- We also considered parameter estimation for multivariate Gaussian distributions.

UOB Open

Multivariate distributions

We often need to think about distributions involving multiple features.

```
## # A tibble: 9 x 8
## # Groups:   species [3]
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
## 1 Adelie  Dream            37.3          16.8              192        3000 fema~
## 2 Adelie  Torge~           33.5          19                190        3600 fema~
## 3 Adelie  Biscoe           45.6          20.3              191        4600 male
## 4 Chinst~ Dream            49.6          18.2              193        3775 male
## 5 Chinst~ Dream            58            17.8              181        3700 fema~
## 6 Chinst~ Dream            52.7          19.8              197        3725 male
## 7 Gentoo  Biscoe           49.6          15                216        4750 male
## 8 Gentoo  Biscoe           43.6          13.9              217        4900 fema~
## 9 Gentoo  Biscoe           49.5          16.1              224        5650 male
## # ... with 1 more variable: year <int>
```

n examples

d features

To model the relationships between these features we must consider multivariate distributions.

Univariate random variables

Given a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, a **random variable** is a mapping $X : \Omega \to \mathbb{R}$ such that for every a, $b \in \mathbb{R}$, $\{\omega \in \Omega \ : \ X(\omega) \in [a, b]\} \in \mathcal{E}$ is an event.

UOB Open

Univariate random variables

Given a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, a **random variable** is a mapping $X : \Omega \to \mathbb{R}$ such that for every a, $b \in \mathbb{R}$, $\{\omega \in \Omega \ : \ X(\omega) \in [a, b]\} \in \mathcal{E}$ is an event.

Discrete random variables are specified by a **probability mass function** $p_X : \mathbb{R} \to [0, 1]$ with $\sum_{x \in \mathbb{R}} p_X(x) = 1$. For all a, $b \in \mathbb{R}$ we have

$$\mathbb{P}(X \in [a, b]) = \sum_{x \in [a,b]} p_X(x).$$

Univariate random variables

Given a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, a **random variable** is a mapping $X : \Omega \to \mathbb{R}$ such that for every a, $b \in \mathbb{R}$, $\{\omega \in \Omega \ : \ X(\omega) \in [a, b]\} \in \mathcal{E}$ is an event.

Discrete random variables are specified by a **probability mass function** $p_X : \mathbb{R} \to [0, 1]$ with $\sum_{x \in \mathbb{R}} p_X(x) = 1$. For all a, $b \in \mathbb{R}$ we have

$$\mathbb{P}(X \in [a, b]) = \sum_{x \in [a,b]} p_X(x).$$

Continuous random variables are specified by a **probability density function** $f_X : \mathbb{R} \to [0, \infty)$ with $\int_{-\infty}^{\infty} f_X(x) = 1$. For all $a, b \in \mathbb{R}$ we have

$$\mathbb{P}(X \in [a, b]) = \int_a^b f_X(x) dx.$$

Gaussian random variables

A classical example of continuous random variable X is a Gaussian with parameters (μ, σ)

The associated density $f_{\mu,\sigma} : \mathbb{R} \to [0, \infty)$ given by

$$f_{\mu,\sigma}(x) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad \text{for all} \quad x \in \mathbb{R}$$

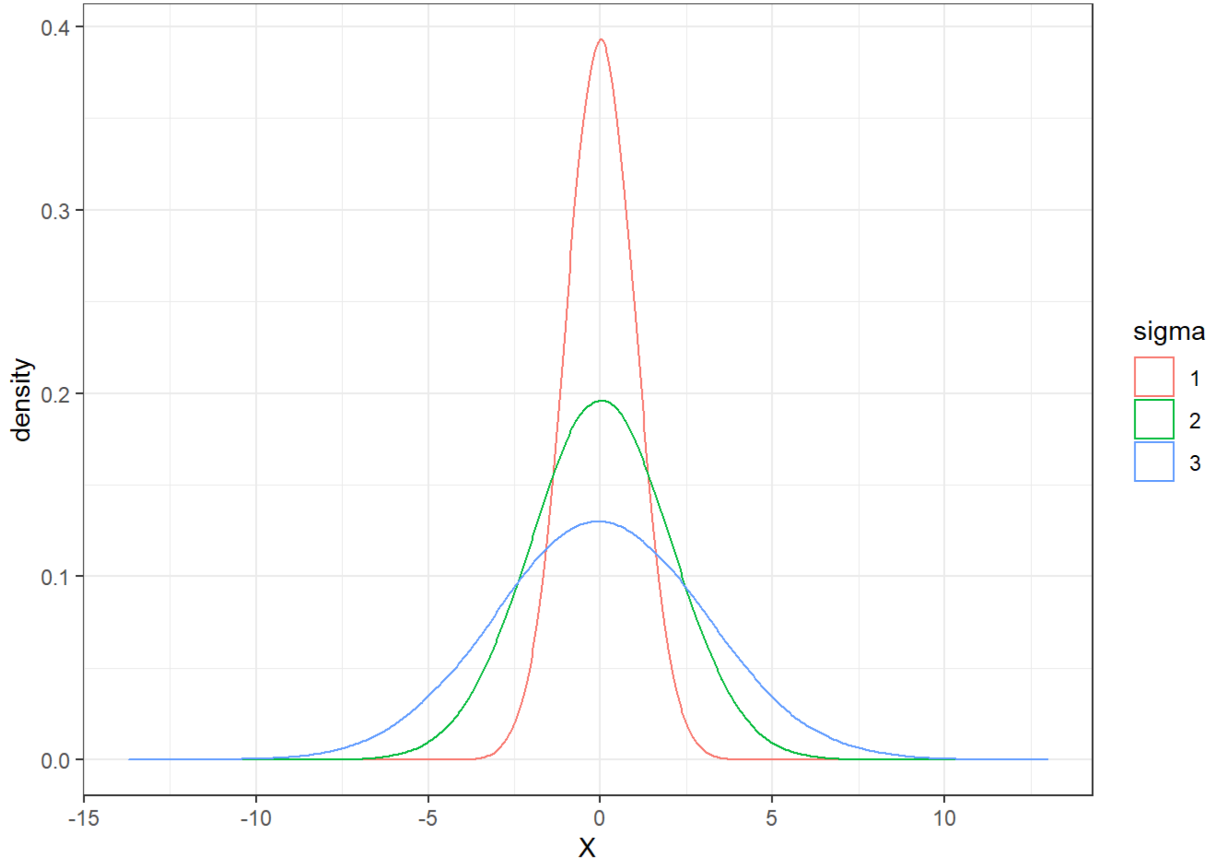
We have $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$

A Gaussian random variable is often referred to as a normal random variable.

We often write $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ to mean X is Gaussian with parameters (μ, σ)

UOB Open

Gaussian random variables



Given $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ we have $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$

UOB Open

Continuous random vectors

Given a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, a **random vector** is a mapping $X : \Omega \to \mathbb{R}^d$ such that for every $a_1, \ldots, a_d$, $b_1, \ldots, b_d \in \mathbb{R}$ with each $a_j \leq b_j$, $\{\omega \in \Omega \ : \ X(\omega) \in \prod_{j=1}^d [a_j, b_j]\} \in \mathcal{E}$ is an event.

UOB Open

Continuous random vectors

Given a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, a **random vector** is a mapping $X : \Omega \to \mathbb{R}^d$ such that for every $a_1, \ldots, a_d$, $b_1, \ldots, b_d \in \mathbb{R}$ with each $a_j \leq b_j$, $\{\omega \in \Omega \ : \ X(\omega) \in \prod_{j=1}^d [a_j, b_j]\} \in \mathcal{E}$ is an event.

A multivariate **probability density function** is a function $f_X : \mathbb{R}^d \to [0, \infty)$ with

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x_1, \ldots, x_d) = 1.$$

Continuous random vectors

Given a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, a **random vector** is a mapping $X : \Omega \to \mathbb{R}^d$ such that for every $a_1, \ldots, a_d$, $b_1, \ldots, b_d \in \mathbb{R}$ with each $a_j \leq b_j$, $\{\omega \in \Omega \ : \ X(\omega) \in \prod_{j=1}^{d}[a_j, b_j]\} \in \mathcal{E}$ is an event.

A multivariate **probability density function** is a function $f_X : \mathbb{R}^d \to [0, \infty)$ with

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x_1, \ldots, x_d) = 1.$$

A **continuous random vector** is a random vector $X : \Omega \to \mathbb{R}^d$ with a probability density function $f_X : \mathbb{R}^d \to [0, \infty)$ such that every $a_1, \ldots, a_d$, $b_1, \ldots, b_d \in \mathbb{R}$ with each $a_j \leq b_j$,

$$\mathbb{P}\left(X \in [a_1, b_1] \times \ldots \times [a_d, b_d]\right) = \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} f_X(x_1, \ldots, x_d)$$

Multivariate Gaussians

A classic example of a continuous random vector X is a multivariate the Gaussian.

Multivariate Gaussians

A classic example of a continuous random vector X is a multivariate the Gaussian.

Its parameters are a) A mean vector $\mu = \mathbb{E}[X] \in \mathbb{R}^d$

b) A covariance matrix $\Sigma = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right] \in \mathbb{R}^{d \times d}.$

Multivariate Gaussians

A classic example of a continuous random vector X is a multivariate the Gaussian.

Its parameters are a) A mean vector $\mu = \mathbb{E}[X] \in \mathbb{R}^d$

b) A covariance matrix $\Sigma = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right] \in \mathbb{R}^{d \times d}.$

The probability density function $f_{\mu,\Sigma} : \mathbb{R}^d \to (0, \infty)$ is given by

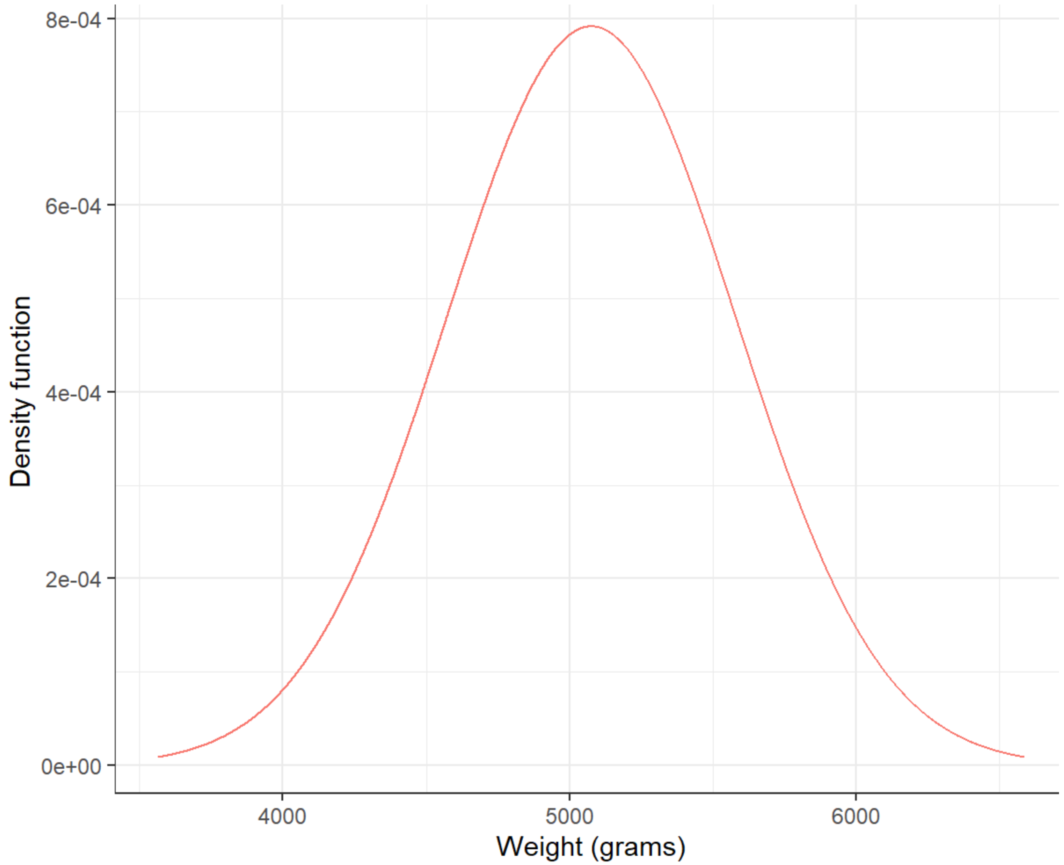
$$f_{\mu,\Sigma}(x) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)\right).$$

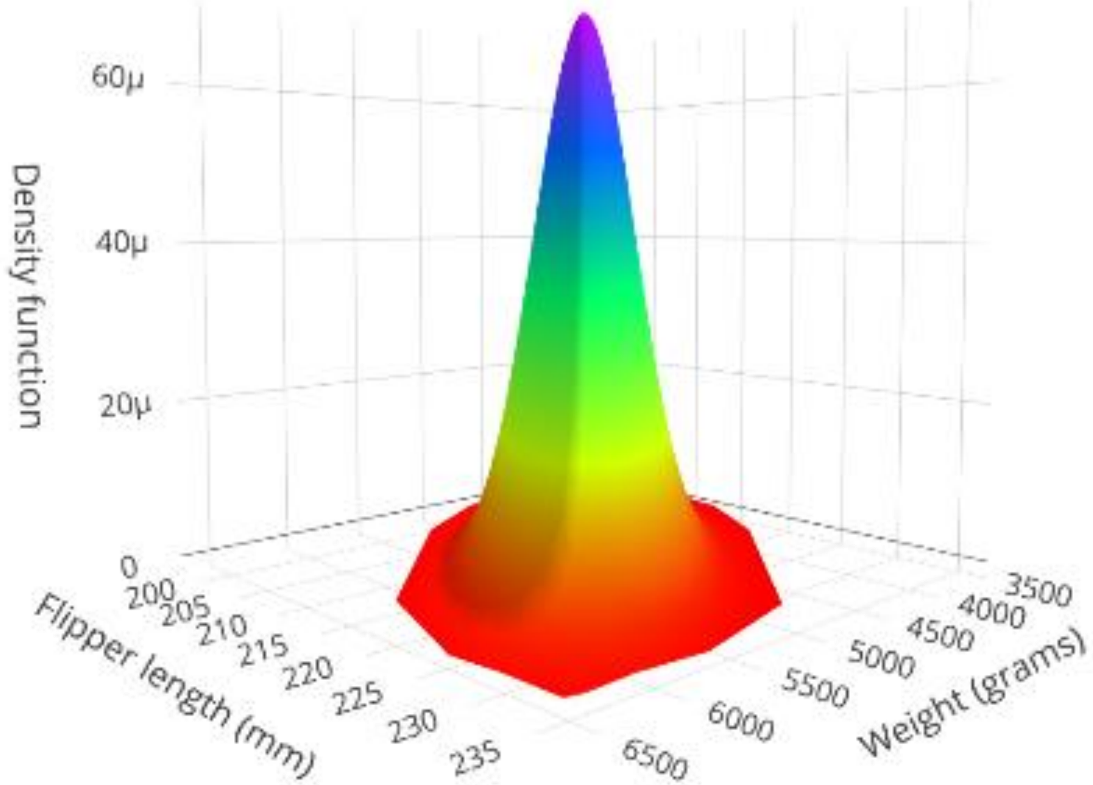
This generalizes the univariate Gaussian we discussed previously.

UOB Open

Multivariate Gaussians

A univariate Gaussian and a bivariate Gaussian





Parameter estimation for multivariate Gaussians

Suppose $X_1, \cdots, X_n \in \mathcal{N}(\mu, \Sigma)$ are i.i.d. samples from a multivariate Gaussian

with parameters $\mu = \mathbb{E}[X] \in \mathbb{R}^d$ and $\Sigma = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right] \in \mathbb{R}^{d \times d}$.

Parameter estimation for multivariate Gaussians

Suppose $X_1, \cdots, X_n \in \mathcal{N}(\mu, \Sigma)$ are i.i.d. samples from a multivariate Gaussian

with parameters $\mu = \mathbb{E}[X] \in \mathbb{R}^d$ and $\Sigma = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right] \in \mathbb{R}^{d \times d}$.

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \in \mathbb{R}^d$ is both the MVUE and the MLE for $\mu \in \mathbb{R}^d$

Parameter estimation for multivariate Gaussians

Suppose $X_1, \cdots, X_n \in \mathcal{N}(\mu, \Sigma)$ are i.i.d. samples from a multivariate Gaussian

with parameters $\mu = \mathbb{E}[X] \in \mathbb{R}^d$ and $\Sigma = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right] \in \mathbb{R}^{d \times d}$.

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \in \mathbb{R}^d$ is both the MVUE and the MLE for $\mu \in \mathbb{R}^d$

$\hat{\Sigma}_{\mathrm{U}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^\top \in \mathbb{R}^{d \times d}$ is the MVUE for $\Sigma \in \mathbb{R}^{d \times d}$

Parameter estimation for multivariate Gaussians

Suppose $X_1, \cdots, X_n \in \mathcal{N}(\mu, \Sigma)$ are i.i.d. samples from a multivariate Gaussian

with parameters $\mu = \mathbb{E}[X] \in \mathbb{R}^d$ and $\Sigma = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right] \in \mathbb{R}^{d \times d}$.

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \in \mathbb{R}^d$ is both the MVUE and the MLE for $\mu \in \mathbb{R}^d$

$\hat{\Sigma}_{\mathrm{U}} = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^\top \in \mathbb{R}^{d \times d}$ is the MVUE for $\Sigma \in \mathbb{R}^{d \times d}$

$\hat{\Sigma}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^\top \in \mathbb{R}^{d \times d}$ is the MLE for $\Sigma \in \mathbb{R}^{d \times d}$

UOB Open

Parameter estimation for multivariate Gaussians

Let's fit a multivariate model for our Gentoo penguins

```
penguins_gwf<-penguins%>%
  filter(species=="Gentoo")%>%
  select(body_mass_g,flipper_length_mm)
```

Parameter estimation for multivariate Gaussians

Let’s fit a multivariate model for our Gentoo penguins

```
penguins_gwf<-penguins%>%
  filter(species=="Gentoo")%>%
  select(body_mass_g,flipper_length_mm)
```

```
mu_gwf<-map_dbl(penguins_gwf,~mean(.x,na.rm=1)) # MLE estimate of the mean
mu_gwf
```

```
##      body_mass_g flipper_length_mm
##         5076.016           217.187
```

Parameter estimation for multivariate Gaussians

Let’s fit a multivariate model for our Gentoo penguins

```
penguins_gwf<-penguins%>%
  filter(species=="Gentoo")%>%
  select(body_mass_g,flipper_length_mm)
```

```
mu_gwf<-map_dbl(penguins_gwf,~mean(.x,na.rm=1)) # MLE estimate of the mean
mu_gwf
```

```
##       body_mass_g flipper_length_mm
##          5076.016           217.187
```

```
Sigma_gwf<-cov(penguins_gwf,use="complete.obs") # MVUE estimate of the covariance
Sigma_gwf
```

```
##                   body_mass_g flipper_length_mm
## body_mass_g        254133.180        2297.14448
## flipper_length_mm    2297.144          42.05491
```

Parameter estimation for multivariate Gaussians

```
mu_gwf<-map_dbl(penguins_gwf,~mean(.x,na.rm=1)) # MLE estimate of the mean
mu_gwf
```

```
##       body_mass_g flipper_length_mm
##          5076.016           217.187
```

```
Sigma_gwf<-cov(penguins_gwf,use="complete.obs") # MVUE estimate of the covariance
Sigma_gwf
```

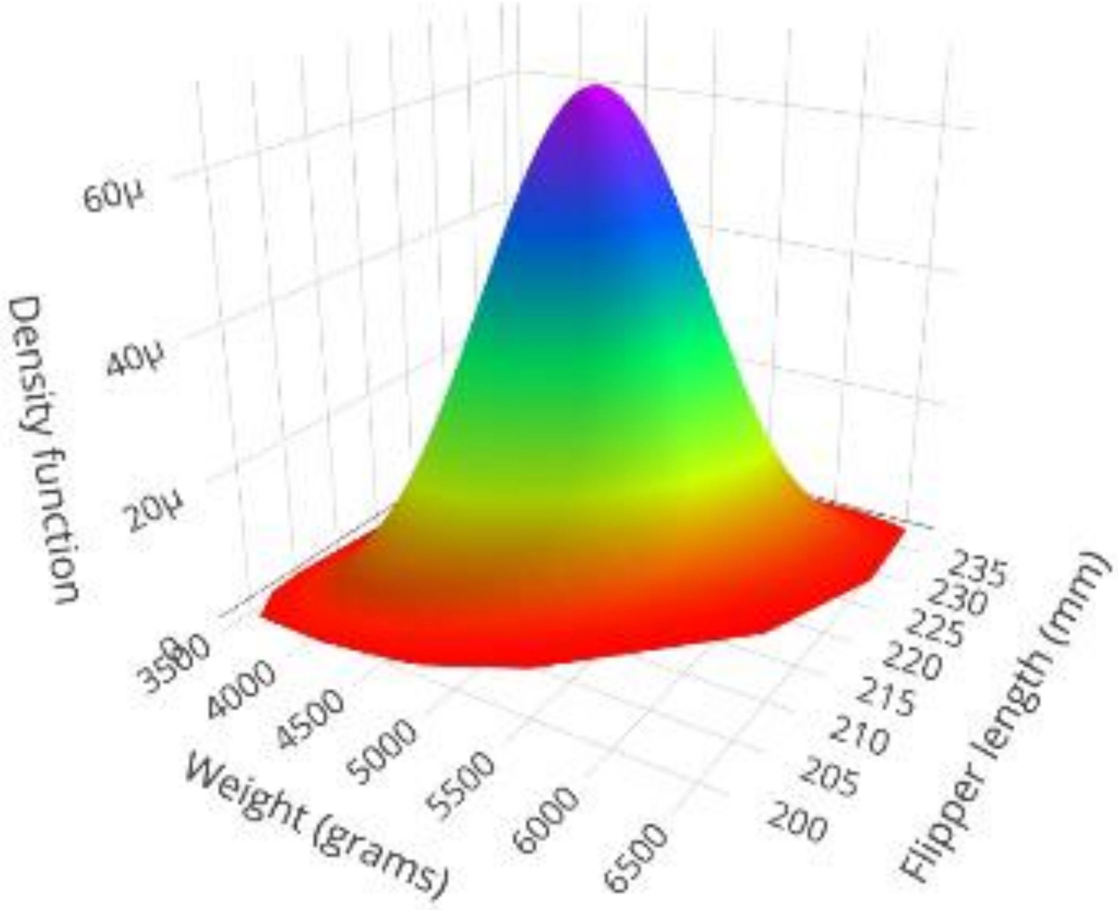
```
##                   body_mass_g flipper_length_mm
## body_mass_g        254133.180        2297.14448
## flipper_length_mm    2297.144          42.05491
```

```
Sigma_gwf_MLE<-cov(penguins_gwf,use="complete.obs")*(n-1)/n # MLE estimate of the covariance
Sigma_gwf_MLE
```

```
##                   body_mass_g flipper_length_mm
## body_mass_g        252083.719        2278.61912
## flipper_length_mm    2278.619          41.71576
```

Parameter estimation for multivariate Gaussians

Let’s fit a multivariate model for our Gentoo penguins



What have we covered?

- We introduced the concept of a random vector.
- We saw that continuous random vectors can be understood via probability density functions.
- We introduced the concept of a multivariate Gaussian distribution.
- We also considered parameter estimation for multivariate Gaussian distributions.

UOB Open

Thanks for listening!

Henry W J Reeve

henry.reeve@bristol.ac.uk

EMATM0061

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.