Multivariate Normal Distribution

Max Turgeon

STAT 7200–Multivariate Statistics

Building the multivariate density i

- Let $Z \sim N(0,1)$ be a standard (univariate) normal random variable. Recall that its density is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right).$$

- Now if we take $Z_1, \ldots, Z_p \sim N(0,1)$ independently distributed, their joint density is

Building the multivariate density ii

$$\begin{aligned}\phi(z_1, \ldots, z_p) &= \prod_{i=1}^p \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z_i^2\right) \\ &= \frac{1}{(\sqrt{2\pi})^p} \exp\left(-\frac{1}{2} \sum_{i=1}^p z_i^2\right) \\ &= \frac{1}{(\sqrt{2\pi})^p} \exp\left(-\frac{1}{2} \mathbf{z}^T \mathbf{z}\right),\end{aligned}$$

where $\mathbf{z} = (z_1, \ldots, z_p)$.

- More generally, let $\mu \in \mathbb{R}^p$ and let Σ be a $p \times p$ positive definite matrix.

Building the multivariate density iii

  - Let $\Sigma = LL^T$ be the Cholesky decomposition for Σ.
- Let $\mathbf{Z} = (Z_1, \ldots, Z_p)$ be a standard (multivariate) normal random vector, and define $\mathbf{Y} = L\mathbf{Z} + \mu$. We know from a previous lecture that
  - $E(\mathbf{Y}) = LE(\mathbf{Z}) + \mu = \mu$;
  - $\mathrm{Cov}(\mathbf{Y}) = L\mathrm{Cov}(\mathbf{Z})L^T = \Sigma$.
- To get the density, we need to compute the inverse transformation:

$$\mathbf{Z} = L^{-1}(\mathbf{Y} - \mu).$$

Building the multivariate density iv

- The Jacobian matrix J for this transformation is simply L^{-1}, and therefore

$$
\begin{aligned}
|\det(J)| &= |\det(L^{-1})| \\
&= \det(L)^{-1} \qquad \text{(positive diagonal elements)} \\
&= \sqrt{\det(\Sigma)}^{-1} \\
&= \det(\Sigma)^{-1/2}.
\end{aligned}
$$

Building the multivariate density v

- Plugging this into the formula for the density of a transformation, we get

$$
\begin{aligned}
f(y_1, \ldots, y_p) &= \frac{1}{\det(\Sigma)^{1/2}} \phi(L^{-1}(\mathbf{y} - \mu)) \\
&= \frac{1}{\det(\Sigma)^{1/2}} \left(\frac{1}{(\sqrt{2\pi})^p} \exp\left(-\frac{1}{2} (L^{-1}(\mathbf{y} - \mu))^T L^{-1}(\mathbf{y} - \mu) \right) \right) \\
&= \frac{1}{\det(\Sigma)^{1/2} (\sqrt{2\pi})^p} \exp\left(-\frac{1}{2} (\mathbf{y} - \mu)^T (LL^T)^{-1} (\mathbf{y} - \mu) \right) \\
&= \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2} (\mathbf{y} - \mu)^T \Sigma^{-1} (\mathbf{y} - \mu) \right).
\end{aligned}
$$

Example i

```r
set.seed(123)

n <- 1000; p <- 2
Z <- matrix(rnorm(n*p), ncol = p)

mu <- c(1, 2)
Sigma <- matrix(c(1, 0.5, 0.5, 1), ncol = 2)
L <- t(chol(Sigma))
```

Example ii

```r
Y <- L %*% t(Z) + mu
Y <- t(Y)

colMeans(Y)
```

```
## [1] 1.016128 2.044840
```

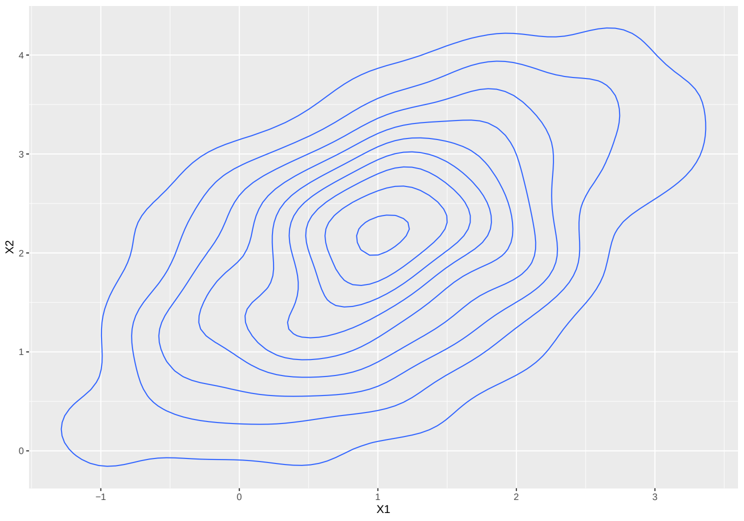
```r
cov(Y)
```

```
##           [,1]      [,2]
## [1,] 0.9834589 0.5667194
## [2,] 0.5667194 1.0854361
```

Example iii

```r
library(tidyverse)
Y %>%
  data.frame() %>%
  ggplot(aes(X1, X2)) +
  geom_density_2d()
```

Example iv



Example v

```r
library(mvtnorm)

Y <- rmvnorm(n, mean = mu, sigma = Sigma)

colMeans(Y)
```

```
## [1] 0.9812102 1.9829380
```

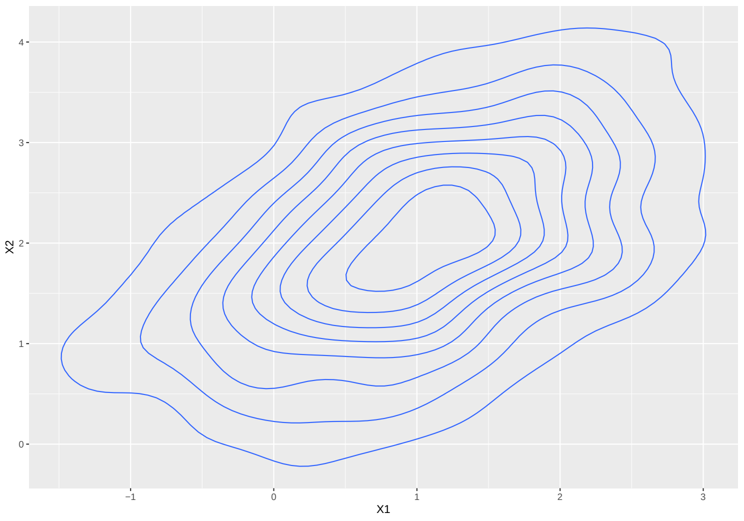
```r
cov(Y)
```

Example vi

```
##              [,1]      [,2]
## [1,] 0.9982835 0.4906990
## [2,] 0.4906990 0.9489171
```

```
Y %>%
  data.frame() %>%
  ggplot(aes(X1, X2)) +
  geom_density_2d()
```

Example vii



Characteristic function i

- Using a similar strategy, we can derive the characteristic function of the multivariate normal distribution.
- Recall that the characteristic function of the univariate standard normal distribution is given by

$$\varphi(t) = \exp\left(\frac{-t^2}{2}\right).$$

Characteristic function ii

- Therefore, if we have $Z_1, \ldots, Z_p \sim N(0,1)$ independent, the characteristic function of $\mathbf{Z} = (Z_1, \ldots, Z_p)$ is

$$\begin{aligned}
\varphi_{\mathbf{Z}}(\mathbf{t}) &= \prod_{i=1}^{p} \exp\left(\frac{-t_i^2}{2}\right) \\
&= \exp\left(\sum_{i=1}^{p} \frac{-t_i^2}{2}\right) \\
&= \exp\left(\frac{-\mathbf{t}^T\mathbf{t}}{2}\right).
\end{aligned}$$

Characteristic function iii

- For $\mu \in \mathbb{R}^p$ and $\Sigma = LL^T$ positive definite, define $\mathbf{Y} = L\mathbf{Z} + \mu$. We then have

$$\begin{aligned}
\varphi_{\mathbf{Y}}(\mathbf{t}) &= \exp\left(i\mathbf{t}^T\mu\right)\varphi_{\mathbf{Z}}(L^T\mathbf{t}) \\
&= \exp\left(i\mathbf{t}^T\mu\right)\exp\left(\frac{-(L^T\mathbf{t})^T(L^T\mathbf{t})}{2}\right) \\
&= \exp\left(i\mathbf{t}^T\mu - \frac{\mathbf{t}^T\Sigma\mathbf{t}}{2}\right).
\end{aligned}$$

Alternative characterization

A p-dimensional random vector $\mathbf{Y}$ is said to have a multivariate normal distribution if and only if every linear combination of $\mathbf{Y}$ has a *univariate* normal distribution. - **Note**: In particular, every component of $\mathbf{Y}$ is also normally distributed.

Proof i

This result follows from the Cramer-Wold theorem. Let $\mathbf{u} \in \mathbb{R}^p$. We have

$$\begin{aligned}
\varphi_{\mathbf{u}^T\mathbf{Y}}(t) &= \varphi_{\mathbf{Y}}(t\mathbf{u}) \\
&= \exp\left(it\mathbf{u}^T\mu - \frac{\mathbf{u}^T\Sigma\mathbf{u}t^2}{2}\right).
\end{aligned}$$

This is the characteristic function of a univariate normal variable with mean $\mathbf{u}^T\mu$ and variance $\mathbf{u}^T\Sigma\mathbf{u}$.

Proof ii

Conversely, assume $\mathbf{Y}$ has mean μ and Σ, and assume $\mathbf{u}^T\mathbf{Y}$ is normally distributed for all $\mathbf{u} \in \mathbb{R}^p$. In particular, we must have

$$\varphi_{\mathbf{u}^T\mathbf{Y}}(t) = \exp\left(it\mathbf{u}^T\mu - \frac{\mathbf{u}^T\Sigma\mathbf{u}t^2}{2}\right).$$

Now, let's look at the characteristic function of $\mathbf{Y}$:

Proof iii

$$
\begin{aligned}
\varphi_{\mathbf{Y}}(\mathbf{t}) &= E\left(\exp\left(i\mathbf{t}^T\mathbf{Y}\right)\right) \\
&= E\left(\exp\left(i(\mathbf{t}^T\mathbf{Y})\right)\right) \\
&= \varphi_{\mathbf{t}^T\mathbf{Y}}(1) \\
&= \exp\left(i\mathbf{t}^T\mu - \frac{\mathbf{t}^T\Sigma\mathbf{t}}{2}\right).
\end{aligned}
$$

This is the characteristic function we were looking for. $\square$

Counter-Example i

- Let $\mathbf{Y}$ be a mixture of two multivariate normal distributions $\mathbf{Y}_1, \mathbf{Y}_2$ with mixing probability p.
- Assume that

$$\mathbf{Y}_i \sim N_p(0, (1-\rho_i)I_p + \rho_i \mathbf{1}\mathbf{1}^{\mathbf{T}}),$$

where $\mathbf{1}$ is a p-dimensional vector of 1s.

  - In other words, the diagonal elements are 1, and the off-diagonal elements are ρ_i.

Counter-Example ii

- First, note that the characteristic function of a mixture distribution is a mixture of the characteristic functions:

$$\varphi_{\mathbf{Y}}(\mathbf{t}) = p\varphi_{\mathbf{Y}_1}(\mathbf{t}) + (1-p)\varphi_{\mathbf{Y}_2}(\mathbf{t}).$$

- Therefore, unless $p = 0, 1$ or $\rho_1 = \rho_2$, the random vector $\mathbf{Y}$ does **not** follow a normal distribution.
- But the components of a mixture are the mixture of each component.
    - Therefore, all components of $\mathbf{Y}$ are univariate standard normal variables.

Counter-Example iii

- In other words, **even if all the margins are normally distributed, the joint distribution may not follow a multivariate normal**.

Useful properties i

- If $\mathbf{Y} \sim N_p(\mu, \Sigma)$, A is a $q \times p$ matrix, and $b \in \mathbb{R}^q$, then

$$A\mathbf{Y} + b \sim N_q(A\mu + b, A\Sigma A^T).$$

- If $\mathbf{Y} \sim N_p(\mu, \Sigma)$ then all subsets of $\mathbf{Y}$ are normally distributed; that is, write
  - $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$;
  - $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$.
  - Then $\mathbf{Y}_1 \sim N_r(\mu_1, \Sigma_{11})$ and $\mathbf{Y}_2 \sim N_{p-r}(\mu_2, \Sigma_{22})$.

Useful properties ii

- Assume the same partition as above. Then the following are equivalent:
  - $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent;
  - $\Sigma_{12} = 0$;
  - $\mathrm{Cov}(\mathbf{Y}_1, \mathbf{Y}_2) = 0$.

Exercise (J&W 4.3)

Let $(Y_1, Y_2, Y_3) \sim N_3(\mu, \Sigma)$ with $\mu = (3, 1, 4)$ and

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Which of the following random variables are independent? Explain.

1. Y_1 and Y_2.
2. Y_2 and Y_3.
3. (Y_1, Y_2) and Y_3.
4. $0.5(Y_1 + Y_2)$ and Y_3.
5. Y_2 and $Y_2 - \frac{5}{2}Y_1 - Y_3$.

Conditional Normal Distributions i

- **Theorem**: Let $\mathbf{Y} \sim N_p(\mu, \Sigma)$, where
  - $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$;
  - $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$.
- Then the *conditional distribution* of $\mathbf{Y}_1$ given $\mathbf{Y}_2 = \mathbf{y}_2$ is multivariate normal $N_r(\mu_{1|2}, \Sigma_{1|2})$, where
  - $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \mu_2)$
  - $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Proof i

Let B be a matrix of the same dimension as Σ_{12}. We will look at the following linear transformation of $\mathbf{Y}$:

$$\begin{pmatrix} I & -B \\ 0 & I \end{pmatrix} \mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 - B\mathbf{Y}_2 \\ \mathbf{Y}_2 \end{pmatrix}.$$

Using the properties of the mean, we have

$$\begin{pmatrix} I & -B \\ 0 & I \end{pmatrix} \mu = \begin{pmatrix} \mu_1 - B\mu_2 \\ \mu_2 \end{pmatrix}.$$

Proof ii

Similarly, using the properties of the covariance, we have

$$\begin{pmatrix} I & -B \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ -B^T & I \end{pmatrix}$$
$$= \begin{pmatrix} \Sigma_{11} - B\Sigma_{21} - \Sigma_{12}B^T + B\Sigma_{22}B^T & \Sigma_{12} - B\Sigma_{22} \\ \Sigma_{21} - \Sigma_{22}B^T & \Sigma_{22} \end{pmatrix}.$$

Proof iii

Recall that subsets of a multivariate normal variable are again multivariate normal:

$$\mathbf{Y}_1 - B\mathbf{Y}_2 \sim N\left(\mu_1 - B\mu_2, \Sigma_{11} - B\Sigma_{21} - \Sigma_{12}B^T + B\Sigma_{22}B^T\right),$$
$$\mathbf{Y}_2 \sim N(\mu_2, \Sigma_{22}).$$

If we take $B = \Sigma_{12}\Sigma_{22}^{-1}$, the two off-diagonal blocks of the covariance matrix above become 0. This implies that $\mathbf{Y}_1 - B\mathbf{Y}_2$ is independent of $\mathbf{Y}_2$.

Proof iv

Given $B = \Sigma_{12}\Sigma_{22}^{-1}$, we can deduce that

$$\mathbf{Y}_1 - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Y}_2 \sim N\left(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{1|2}\right),$$

where

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Using the fact that $\mathbf{Y}_1 - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Y}_2$ and $\mathbf{Y}_2$ are independent, we can conclude that

$$\mathbf{Y}_1 - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Y}_2 = \mathbf{Y}_1 - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{y}_2 \mid \mathbf{Y}_2 = \mathbf{y}_2.$$

Proof v

Finally, by adding $\Sigma_{12}\Sigma_{22}^{-1}\mathbf{y}_2$ to the right-hand side, we get

$$\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2 \sim N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \mu_2), \Sigma_{1|2}\right).$$

. $\square$

Conditional Normal Distributions ii

- **Theorem**: Let $\mathbf{Y}_2 \sim N_{p-r}(\mu_2, \Sigma_{22})$ and assume that $\mathbf{Y}_1$ given $\mathbf{Y}_2 = \mathbf{y}_2$ is multivariate normal $N_r(A\mathbf{y}_2 + b, \Omega)$, where Ω does not depend on $\mathbf{y}_2$. Then $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim N_p(\mu, \Sigma)$, where
    - $\mu = \begin{pmatrix} A\mu_2 + b \\ \mu_2 \end{pmatrix}$;
    - $\Sigma = \begin{pmatrix} \Omega + A\Sigma_{22}A^T & A\Sigma_{22} \\ \Sigma_{22}A^T & \Sigma_{22} \end{pmatrix}$.
- Proof: **Exercise** (e.g. compute joint density).

Exercise

- Let $\mathbf{Y}_2 \sim N_1(0, 1)$ and assume

$$\mathbf{Y}_1 \mid \mathbf{Y}_2 = y_2 \sim N_2 \left(\begin{pmatrix} y_2 + 1 \\ 2y_2 \end{pmatrix}, I_2 \right).$$

  Find the joint distribution of $(\mathbf{Y}_1, \mathbf{Y}_2)$.

Another important result i

- Let $\mathbf{Y} \sim N_p(\mu, \Sigma)$, and let $\Sigma = LL^T$ be the Cholesky decomposition of Σ.
- We know that $\mathbf{Z} = L^{-1}(\mathbf{Y} - \mu)$ is normally distributed, with mean 0 and covariance matrix

$$\mathrm{Cov}(\mathbf{Z}) = L^{-1}\Sigma(L^{-1})^T = I_p.$$

- Therefore $(\mathbf{Y} - \mu)^T\Sigma^{-1}(\mathbf{Y} - \mu)$ is the sum of *squared* standard normal random variables.
  - In other words, $(\mathbf{Y} - \mu)^T\Sigma^{-1}(\mathbf{Y} - \mu) \sim \chi^2(p)$.
  - This can be seen as a generalization of the univariate result $\left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2(1)$.

Another important result ii

- From this, we get a result about the probability that a multivariate normal falls within an *ellipse*:

$$P\left((\mathbf{Y}-\mu)^T\Sigma^{-1}(\mathbf{Y}-\mu)\leq\chi^2(\alpha;p)\right)=1-\alpha.$$

  - We can use this to construct a confidence region around the sample mean.

Application i

- We can use the result above to construct a graphical test of multivariate normality.
  - **Note**: The chi-square distribution does not yield a good approximation for large p. A more accurate graphical test can be constructed using a beta distribution.
- *Procedure*: Given a random sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ of p-dimensional random vectors:
  - Compute $D_i^2 = (\mathbf{Y}_i - \bar{\mathbf{Y}})^T S^{-1} (\mathbf{Y}_i - \bar{\mathbf{Y}})$.
  - Compare the (observed) quantiles of the D_i^2s with the (theoretical) quantiles of a $\chi^2(p)$ distribution.

Application ii

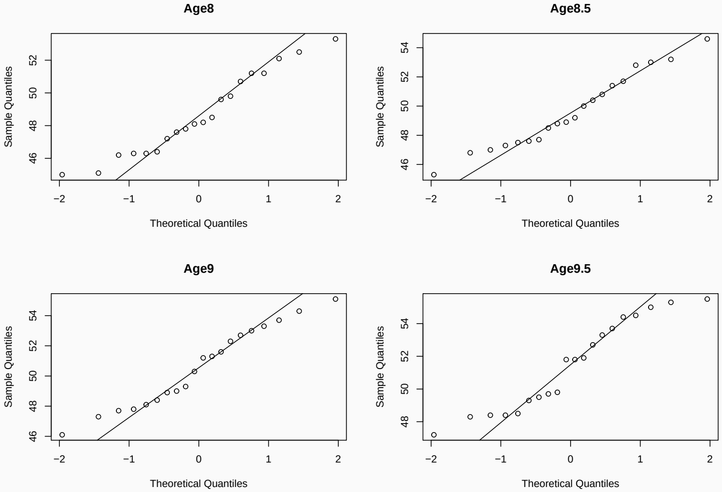
```r
# Ramus data, Timm (2002)
main_page <- "https://maxturgeon.ca/w20-stat7200/"
ramus <- read.csv(paste0(main_page, "Ramus.csv"))
head(ramus, n = 5)
```

```
##   Age8 Age8.5 Age9 Age9.5 ID
## 1 47.8   48.8 49.0   49.7  1
## 2 46.4   47.3 47.7   48.4  2
## 3 46.3   46.8 47.8   48.5  3
## 4 45.1   45.3 46.1   47.2  4
## 5 47.6   48.5 48.9   49.3  5
```

Application iii

```r
var_names <- c("Age8", "Age8.5",
               "Age9", "Age9.5")

par(mfrow = c(2, 2))
for (var in var_names) {
  qqnorm(ramus[, var], main = var)
  qqline(ramus[, var])
}
```

Application iv



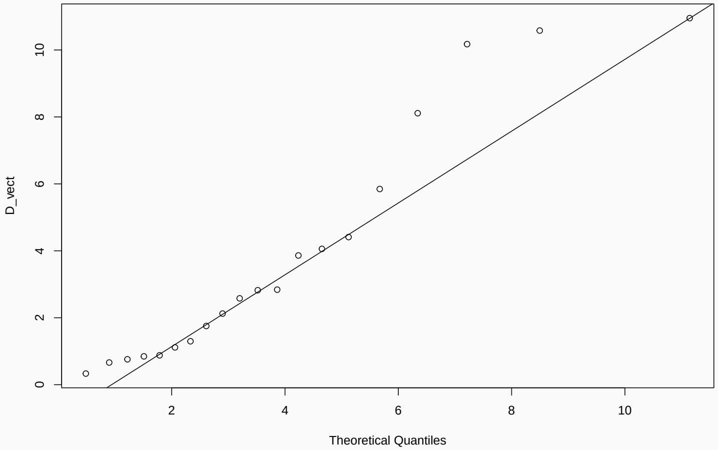
Application v

```
ramus <- ramus[,var_names]
sigma_hat <- cov(ramus)

ramus_cent <- scale(ramus, center = TRUE,
                    scale = FALSE)

D_vect <- apply(ramus_cent, 1, function(row) {
  t(row) %*% solve(sigma_hat) %*% row
})
```

Application vi

```r
qqplot(qchisq(ppoints(D_vect), df = 4),
       D_vect, xlab = "Theoretical Quantiles")
qqline(D_vect, distribution = function(p) {
  qchisq(p, df = 4)
  })
```

Application vii



Estimation

Sufficient Statistics i

- We saw in the previous lecture that the multivariate normal distribution is completely determined by its mean vector $\mu \in \mathbb{R}^p$ and its covariance matrix Σ.
- Therefore, given a sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim N_p(\mu, \Sigma)$ $(n > p)$, we only need to estimate (μ, Σ).
    - Obvious candidates: sample mean $\bar{\mathbf{Y}}$ and sample covariance S_n.

Sufficient Statistics ii

- Write down the *likelihood*:

$$
\begin{aligned}
L &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2} (\mathbf{Y}_i - \mu)^T \Sigma^{-1} (\mathbf{Y}_i - \mu) \right) \right) \\
&= \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{Y}_i - \mu)^T \Sigma^{-1} (\mathbf{Y}_i - \mu) \right)
\end{aligned}
$$

- If we take the (natural) logarithm of L and drop any term that does not depend on (μ, Σ), we get

$$
\ell = -\frac{n}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{Y}_i - \mu)^T \Sigma^{-1} (\mathbf{Y}_i - \mu).
$$

Sufficient Statistics iii

- If we can re-express the second summand in terms of $\bar{\mathbf{Y}}$ and S_n, by the Fisher-Neyman factorization theorem, we will then know that $(\bar{\mathbf{Y}}, S_n)$ is jointly **sufficient** for (μ, Σ).
- First, we have

Sufficient Statistics iv

$$
\begin{aligned}
\sum_{i=1}^{n}(\mathbf{Y}_i-\mu)(\mathbf{Y}_i-\mu)^T &= \sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}}+\bar{\mathbf{Y}}-\mu)(\mathbf{Y}_i-\bar{\mathbf{Y}}+\bar{\mathbf{Y}}-\mu)^T \\
&= \sum_{i=1}^{n}\Big((\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T + (\mathbf{Y}_i - \bar{\mathbf{Y}})(\bar{\mathbf{Y}} - \mu)^T \\
&\qquad + (\bar{\mathbf{Y}} - \mu)(\mathbf{Y}_i - \bar{\mathbf{Y}})^T + (\bar{\mathbf{Y}} - \mu)(\bar{\mathbf{Y}} - \mu)^T\Big) \\
&= \sum_{i=1}^{n}(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T + n(\bar{\mathbf{Y}} - \mu)(\bar{\mathbf{Y}} - \mu)^T \\
&= (n-1)S_n + n(\bar{\mathbf{Y}} - \mu)(\bar{\mathbf{Y}} - \mu)^T.
\end{aligned}
$$

Sufficient Statistics v

- Next, using the fact that $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$, we have

Sufficient Statistics vi

$$
\begin{aligned}
\sum_{i=1}^{n}(\mathbf{Y}_i - \mu)^T\Sigma^{-1}(\mathbf{Y}_i - \mu) &= \mathrm{tr}\left(\sum_{i=1}^{n}(\mathbf{Y}_i - \mu)^T\Sigma^{-1}(\mathbf{Y}_i - \mu)\right)\\
&= \mathrm{tr}\left(\sum_{i=1}^{n}\Sigma^{-1}(\mathbf{Y}_i - \mu)(\mathbf{Y}_i - \mu)^T\right)\\
&= \mathrm{tr}\left(\Sigma^{-1}\sum_{i=1}^{n}(\mathbf{Y}_i - \mu)(\mathbf{Y}_i - \mu)^T\right)\\
&= (n-1)\mathrm{tr}\left(\Sigma^{-1}S_n\right)\\
&\qquad + n\mathrm{tr}\left(\Sigma^{-1}(\bar{\mathbf{Y}} - \mu)(\bar{\mathbf{Y}} - \mu)^T\right)\\
&= (n-1)\mathrm{tr}\left(\Sigma^{-1}S_n\right)\\
&\qquad + n(\bar{\mathbf{Y}} - \mu)^T\Sigma^{-1}(\bar{\mathbf{Y}} - \mu).
\end{aligned}
$$

Maximum Likelihood Estimation i

- Going back to the log-likelihood, we get:

$$\ell = -\frac{n}{2}\log|\Sigma| - \frac{(n-1)}{2}\mathrm{tr}\left(\Sigma^{-1}S_n\right) - \frac{n}{2}(\bar{\mathbf{Y}}-\mu)^T\Sigma^{-1}(\bar{\mathbf{Y}}-\mu).$$

- First, fix Σ and maximise over μ. The only term that depends on μ is

$$-\frac{n}{2}(\bar{\mathbf{Y}}-\mu)^T\Sigma^{-1}(\bar{\mathbf{Y}}-\mu).$$

- We can maximise this term by minimising

$$(\bar{\mathbf{Y}}-\mu)^T\Sigma^{-1}(\bar{\mathbf{Y}}-\mu).$$

Maximum Likelihood Estimation ii

- But since Σ^{-1} is positive definite, we have

$$(\bar{\mathbf{Y}} - \mu)^T \Sigma^{-1} (\bar{\mathbf{Y}} - \mu) \geq 0,$$

  with equality if and only if $\mu = \bar{\mathbf{Y}}$.
- In other words, the log-likelihood is maximised at

$$\hat{\mu} = \bar{\mathbf{Y}}.$$

- Now, we can turn our attention to Σ. We want to maximise

$$\ell = -\frac{n}{2}\log|\Sigma| - \frac{(n-1)}{2}\mathrm{tr}\left(\Sigma^{-1} S_n\right) - \frac{n}{2}(\bar{\mathbf{Y}} - \mu)^T \Sigma^{-1} (\bar{\mathbf{Y}} - \mu).$$

Maximum Likelihood Estimation iii

- At $\mu = \bar{\mathbf{Y}}$, it reduces to

$$-\frac{n}{2}\log|\Sigma| - \frac{(n-1)}{2}\mathrm{tr}\left(\Sigma^{-1}S_n\right).$$

- Write $V = (n-1)S_n$. We then have

$$-\frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}V\right).$$

- Maximising this quantity is equivalent to minimising

$$\log|\Sigma| + \frac{1}{n}\mathrm{tr}\left(\Sigma^{-1}V\right),$$

and by adding the constant $\log|nV^{-1}|$, we get

$$\log|\Sigma|+\frac{1}{n}\mathrm{tr}\left(\Sigma^{-1}V\right)+\log|nV^{-1}| = \log|nV^{-1}\Sigma|+\mathrm{tr}\left(n^{-1}\Sigma^{-1}V\right)$$

Maximum Likelihood Estimation iv

- Set $T = nV^{-1}\Sigma$. Our maximum likelihood problem now reduces to minimising

$$\log|T| + \operatorname{tr}\left(T^{-1}\right).$$

- Let $\lambda_1, \ldots, \lambda_p$ be the (positive) eigenvalues of T. We now have

$$\begin{aligned} \log|T| + \operatorname{tr}\left(T^{-1}\right) &= \log\left(\prod_{i=1}^{p} \lambda_i\right) + \sum_{i=1}^{p} \lambda_i^{-1} \\ &= \sum_{i=1}^{p} \log \lambda_i + \lambda_i^{-1}. \end{aligned}$$

Maximum Likelihood Estimation v

- Each summand can be minimised individually, and the minimum occurs at $\lambda_i = 1$. In other words, the (overall) minimum is when $T = I_p$, which is equivalent to

$$\Sigma = \frac{n-1}{n} S_n = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T.$$

- *In other words*: $\left(\bar{\mathbf{Y}}, \frac{n-1}{n} S_n\right)$ are the **maximum likelihood estimators** for (μ, Σ).

. $\square$

Maximum Likelihood Estimators

- Since the multivariate normal density is "well-behaved", we can deduce the usual properties:
  - **Consistency**: $(\bar{\mathbf{Y}}, \hat{\Sigma})$ converges in probability to (μ, Σ).
  - **Efficiency**: Asymptotically, the covariance of $(\bar{\mathbf{Y}}, \hat{\Sigma})$ achieves the Cramér-Rao lower bound.
  - **Invariance**: For any transformation $(g(\mu), G(\Sigma))$ of (μ, Σ), its MLE is $(g(\bar{\mathbf{Y}}), G(\hat{\Sigma}))$.

Visualizing the likelihood i

```r
library(mvtnorm)
set.seed(123)

n <- 50; p <- 2

mu <- c(1, 2)
Sigma <- matrix(c(1, 0.5, 0.5, 1), ncol = p)

Y <- rmvnorm(n, mean = mu, sigma = Sigma)
```

Visualizing the likelihood ii

```r
loglik <- function(mu, sigma, data = Y) {
  # Compute quantities
  y_bar <- colMeans(Y)
  quad_form <- t(y_bar - mu) %*% solve(sigma) %*%
    (y_bar - mu)

  -0.5*n*log(det(sigma)) -
    0.5*(n - 1)*sum(diag(solve(sigma) %*% cov(Y))) -
    0.5*n*drop(quad_form)
}
```

Visualizing the likelihood iii

```
grid_xy <- expand.grid(seq(0, 2, length.out = 32),
                       seq(0, 4, length.out = 32))

head(grid_xy, n = 5)
```

```
##         Var1 Var2
## 1 0.00000000    0
## 2 0.06451613    0
## 3 0.12903226    0
## 4 0.19354839    0
## 5 0.25806452    0
```

Visualizing the likelihood iv

```r
contours <- purrr::map_df(seq_len(nrow(grid_xy)),
                          function(i) {
  # Where we will evaluate loglik
  mu_obs <- as.numeric(grid_xy[i,])
  # Evaluate at the pop covariance
  z <- loglik(mu_obs, sigma = Sigma)
  # Output data.frame
  data.frame(x = mu_obs[1],
             y = mu_obs[2],
             z = z)
})
```

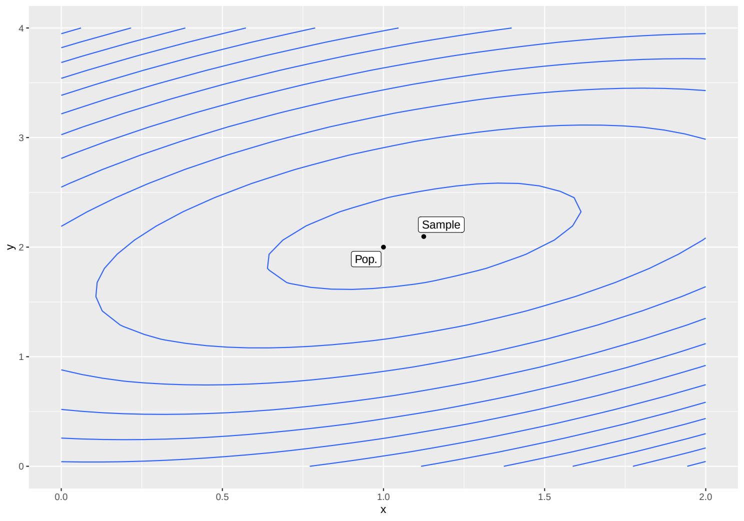
Visualizing the likelihood v

```r
library(tidyverse)
library(ggrepel)
# Create df with pop and sample means
data_means <- data.frame(x = c(mu[1], mean(Y[,1])),
                         y = c(mu[2], mean(Y[,2])),
                         label = c("Pop.", "Sample"))
```

Visualizing the likelihood vi

```r
ggplot(contours, aes(x, y)) +
  geom_contour(aes(z = z)) +
  geom_point(data = data_means) +
  geom_label_repel(data = data_means,
                   aes(label = label))
```

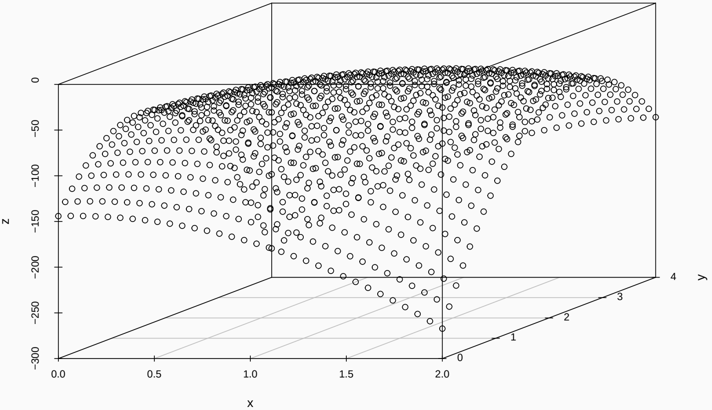
Visualizing the likelihood vii



Visualizing the likelihood viii

```r
library(scatterplot3d)
with(contours, scatterplot3d(x, y, z))
```

Visualizing the likelihood ix



Sampling Distributions

- Recall the univariate case:
  - $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$;
  - $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$;
  - $\bar{X}$ and s^2 are independent.
- In the multivariate case, we have similar results:
  - $\bar{\mathbf{Y}} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$;
  - $(n-1)S_n = n\hat{\Sigma}$ follows a *Wishart* distribution with $n-1$ degrees of freedom;
  - $\bar{\mathbf{Y}}$ and S_n are independent.
- **We will prove the last two properties later.**

Bayesian analysis i

- In *Frequentist* statistics, parameters are fixed quantities that we are trying to estimate and about which we want to make inference.
- In *Bayesian* statistics, parameters are given a distribution that models the uncertainty/knowledge we have about the underlying population quantity.
    - And as we collect data, our knowledge changes, and so does the distribution.

Bayesian analysis ii

- Some vocabulary:
  - **Prior distribution**: Distribution of the parameters *before* data collection/analysis. It represents our *current* knowledge.
  - **Posterior distribution**: Distribution of the parameters *after* data collection/analysis. It represents our *updated* knowledge.
- **Bayesian statistics** is based on the following updating rule:

$$\text{Posterior distribution} \propto \text{Prior distribution} \times \text{Likelihood}.$$

Bayesian analysis iii

- We will look at the posterior distribution of the multivariate normal mean μ, assuming Σ is known, when the prior is also normally distributed.
- Let's start with a single p-dimensional observation $\mathbf{Y} \sim N(\mu, \Sigma)$. The log-likelihood (keeping only terms depending on μ) is equal to

$$\log L(\mathbf{Y} \mid \mu) \propto -\frac{1}{2}(\mathbf{Y} - \mu)^T \Sigma^{-1} (\mathbf{Y} - \mu).$$

- Let $p(\mu) = N(\mu_0, \Sigma_0)$ be the prior distribution for μ. On the log scale, we have

$$\log p(\mu) \propto -\frac{1}{2}(\mu - \mu_0)^T \Sigma_0^{-1} (\mu - \mu_0).$$

Bayesian analysis iv

- Using the updating rule, we have

$$\log p(\mu \mid \mathbf{Y}) \propto -\frac{1}{2}(\mathbf{Y}-\mu)^T\Sigma^{-1}(\mathbf{Y}-\mu)-\frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0).$$

- If we expand both quadratic forms and only keep terms that depend on μ, we get

$$\begin{aligned}\log p(\mu \mid \mathbf{Y}) \propto -\frac{1}{2}\Big(&\mu^T\Omega^{-1}\mu - (\mathbf{Y}^T\Sigma^{-1} + \mu_0^T\Sigma_0^{-1})\mu \\ &-\mu^T(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0)\Big),\end{aligned}$$

where $\Omega^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$.

Bayesian analysis v

- Since Ω^{-1} is the sum of two positive definite matrices, it is itself positive definite.
- Using the Cholesky decomposition, we can write $\Omega^{-1} = U^T U$ with U triangular and invertible. We therefore have

$$
\begin{aligned}
\log p(\mu \mid \mathbf{Y}) \propto -\frac{1}{2} \Big(& \mu^T U^T U \mu - (\mathbf{Y}^T \Sigma^{-1} + \mu_0^T \Sigma_0^{-1}) U^{-1} U \mu \\
& -\mu^T (U^T)(U^T)^{-1} (\Sigma^{-1} \mathbf{Y} + \Sigma_0^{-1} \mu_0) \Big) \\
\propto -\frac{1}{2} \Big(& (U\mu)^T (U\mu) - (\mathbf{Y}^T \Sigma^{-1} + \mu_0^T \Sigma_0^{-1}) U^{-1} (U\mu) \\
& -(U\mu)^T (U^T)^{-1} (\Sigma^{-1} \mathbf{Y} + \Sigma_0^{-1} \mu_0) \Big).
\end{aligned}
$$

Bayesian analysis vi

- Set $\nu = (U^T)^{-1}(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0)$ and complete the square:

$$\begin{aligned}
\log p(\mu \mid \mathbf{Y}) &\propto -\frac{1}{2}\left((U\mu)^T(U\mu) - \nu^T(U\mu) - (U\mu)^T\nu\right)\\
&\propto -\frac{1}{2}\left((U\mu - \nu)^T(U\mu - \nu) - \nu^T\nu\right)\\
&\propto -\frac{1}{2}\left((\mu - U^{-1}\nu)^T U^T U(\mu - U^{-1}\nu) - \nu^T\nu\right)\\
&\propto -\frac{1}{2}\left((\mu - U^{-1}\nu)^T \Omega^{-1}(\mu - U^{-1}\nu) - \nu^T\nu\right).
\end{aligned}$$

Bayesian analysis vii

- Now, note that

$$\begin{aligned} U^{-1}\nu &= U^{-1}(U^T)^{-1}(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0) \\ &= (U^TU)^{-1}(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0) \\ &= \Omega(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0) \\ &= \left(\Sigma^{-1} + \Sigma_0^{-1}\right)^{-1}(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0). \end{aligned}$$

Bayesian analysis viii

- Moreover, we have

$$
\begin{aligned}
\nu^T\nu &= \left((U^T)^{-1}(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0)\right)^T \left((U^T)^{-1}(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0)\right) \\
&= \left(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0\right)^T (U)^{-1}(U^T)^{-1} \left(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0\right) \\
&= \left(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0\right)^T (U^TU)^{-1} \left(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0\right) \\
&= \left(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0\right)^T \Omega \left(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0\right).
\end{aligned}
$$

Bayesian analysis ix

- In other words, $\nu^T\nu$ does not depend on μ, and therefore we can drop it from our expression above. The conclusion is that the log-posterior distribution is proportional to

$$-\frac{1}{2}\left((\mu - \Omega(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0))^T\Omega^{-1}(\mu - \Omega(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0))\right).$$

- As a function of μ, this is the kernel of a multivariate normal density:

$$p(\mu \mid \mathbf{Y}) \sim N\left(\Omega(\Sigma^{-1}\mathbf{Y} + \Sigma_0^{-1}\mu_0), \Omega\right).$$

Bayesian analysis x

- Now, assume we have a random sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. We know that
$$\bar{\mathbf{Y}} \sim N(\mu, n^{-1}\Sigma).$$
- Therefore, the posterior distribution of μ given the random sample is
$$p(\mu \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_n) \sim N\left(\Omega(n\Sigma^{-1}\bar{\mathbf{Y}} + \Sigma_0^{-1}\mu_0), \Omega\right),$$
where $\Omega = \left(n\Sigma^{-1} + \Sigma_0^{-1}\right)^{-1}$.

A few comments

- The inverse covariance matrix $n\Sigma^{-1} + \Sigma_0^{-1}$ is also called the *precision* matrix.
    - We can see that the larger the sample size n, the less significant the prior precision Σ_0^{-1} becomes.
- The posterior mean is a (scaled) linear combination of the sample mean and prior mean.
    - Again, as the sample size increases, the less significant the prior mean becomes.