Statistical Modeling: First Assignment

Matteo Suardi

2025-03-24

Exercise

Consider 800 realizations of a bivariate Gaussian random variable (X_1, X_2) with the following parameter values:

- $\mu_1 = 0$
- $\mu_2 = -3$
- $\sigma_1^2 = 5$
- $\sigma_2^2 = 100$
- $\rho_{12} = -0.99$

1) Provide some measures of association of the generated data, and some bivariate plots. Comment.

Given the correlation, the theoretical covariance is computed as:

$$\sigma_{12} = \rho_{12} \cdot \sigma_1 \cdot \sigma_2$$

Let's first generate the data:

```
library(mvtnorm)
require(mvtnorm)

set.seed(123456)
means <- c(0, -3)
sigma1 <- sqrt(5)
sigma2 <- 10
rho <- -0.99
cov <- rho*(sigma1*sigma2)
sigma <- matrix(c(sigma1^2, cov, cov, sigma2^2), ncol=2)
n <- 800

xx <- rmvnorm(n, means, sigma = sigma)
colnames(xx) <- c("X1", "X2")
```

The dataset consists of 800 rows (observations) and 2 columns, representing the variables X_1 and X_2, respectively.

Descriptive statistics We briefly analyze the generated data by calculating the main descriptive statistics. But first, we recall that the components of a bivariate Gaussian-distributed random variable each follow a univariate Gaussian distribution. Specifically, in this case:

- $X_1 \sim N(0, 5)$ i.e., $\mu = 0$ (mean) and $\sigma^2 = 5$ (variance),
- $X_2 \sim N(-3, 100)$ i.e., $\mu = -3$ (mean) and $\sigma^2 = 100$ (variance)

We summarize the empirical data:

```
summary(xx)
```

```
##        X1                  X2
##  Min.   :-9.21181   Min.   :-33.498
##  1st Qu.:-1.48301   1st Qu.: -8.856
##  Median :-0.09168   Median : -2.369
##  Mean   :-0.10352   Mean   : -2.545
##  3rd Qu.: 1.33140   3rd Qu.:  3.621
##  Max.   : 6.95646   Max.   : 36.681
```

Empirical means and medians are close to their theoretical counterparts. Variances should also approximate theoretical values.

Empirical covariance and correlation The empirical covariance matrix describes how the two variables vary jointly. We expect empirical covariance values to approximate the theoretical covariance (-22.13707):

```
cov(xx)
```

```
##           X1        X2
## X1   4.74941 -21.00553
## X2 -21.00553  95.03658
```

The off-diagonal terms reflect the strong negative covariance between X_1 and X_2, close to the theoretical covariance (cov):

```
cov
```

```
## [1] -22.13707
```

This similarity indicates that the randomly generated data accurately reflect the theoretical distribution we initially imposed, confirming the validity of our random generation procedure. The slight difference observed is expected and explained by natural sampling variablity. Indeed, since our simulation is based on a finite sample of 800 observations, it's normal for the empirical covariance not to match exactly with the theoretical one. If we were to increase the sample size further, the empirical covariance would approach the theoretical one even more closely.

Next, consider the empirical correlation matrix, to quantify the linear association:

```
cor(xx)
```

```
##            X1         X2
## X1  1.0000000 -0.9887088
## X2 -0.9887088  1.0000000
```

The empirical correlation coefficient (approx. -0.99) matches the theoretical one and confirms a strong negative linear relationship, as expected theoretically.

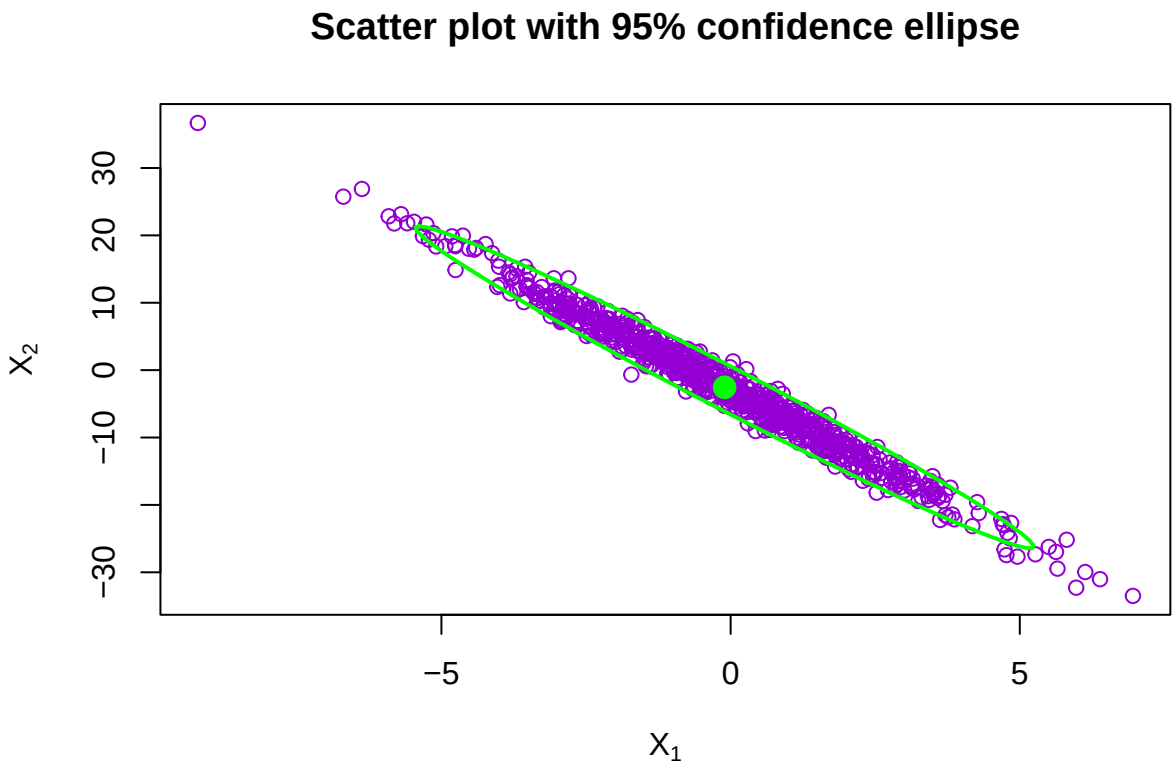
Scatter plot with confidence ellipse To visually illustrate this strong negative correlation, we plot the data with a 95% confidence ellipse:

```
require(car)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
plot(xx[,1], xx[,2],
     col="darkviolet",
     pch=1,
     main="Scatter plot with 95% confidence ellipse",
     xlab=expression(X[1]),
     ylab=expression(X[2]))

dataEllipse(xx[,1], xx[,2],
            levels = 0.95,
            col = "green",
            add = TRUE,
            plot.points = FALSE)
```



The elliptical shape visually demonstrates the strength of the negative correlation. In particular, the orientation (direction) of the ellipse reflects the sign of correlation: the negative slope indicates negative correlation. The elongation (ratio between length and width) of the ellipse reveals the magnitude of the correlation. The

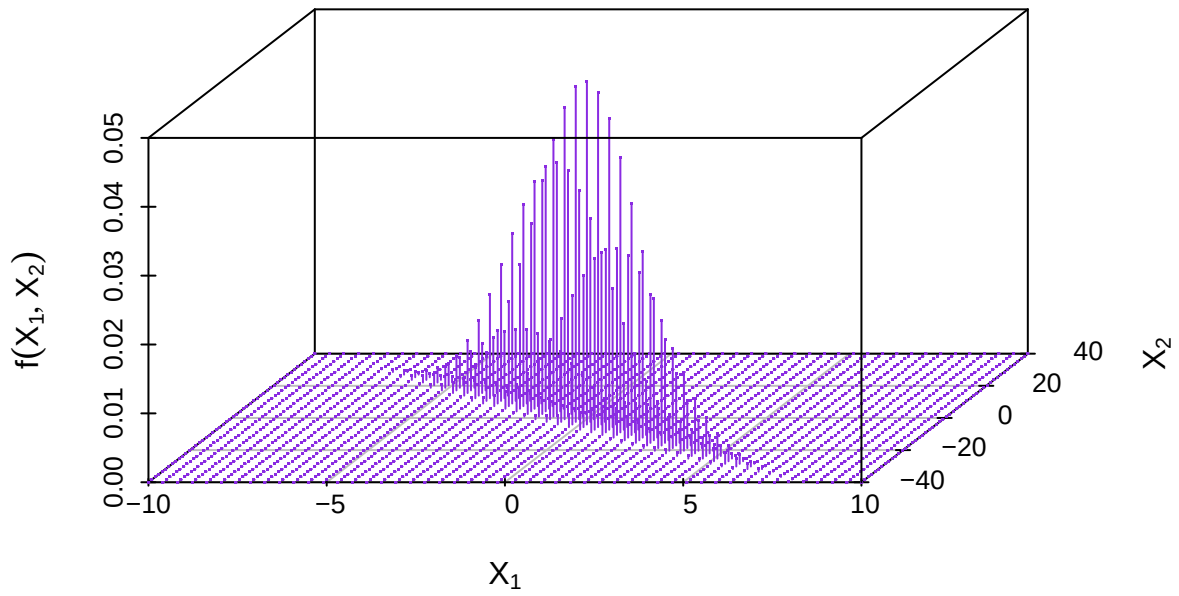
closer this ratio is to zero (very elongated), the stronger the correlation. Here, the ellipse is markedly elongated, highlighting the very strong negative correlation. The ellipse is quite narrow, indicating minimal variability perpendicular to the principal axis. This further supports the strong linear dependence between X_1 and X_2.

3D density representation The following 3D plot effectively illustrates the joint density function:

```
# ## 3D representation
## ----echo=TRUE, include=TRUE-----------------------------------------------

require(scatterplot3d)
```

```
## Loading required package: scatterplot3d
```

```
x1 <- seq(-10, 10, length = 50)
x2 <- seq(-40, 40, length = 50)

grid <- expand.grid(X1=x1, X2=x2)
density <- dmvnorm(grid, mean = means, sigma = sigma)

scatterplot3d(x = grid$X1,
              y = grid$X2,
              z = density,
              color="blueviolet",
              type="h",
              pch = ".",
              cex.symbols = 0.5,
              angle = 60,
              scale.y = 0.7,
              zlab = expression(f(X[1], X[2])),
              xlab = expression(X[1]),
              ylab = expression(X[2]),
              main = "3D Bivariate Normal Distribution: (X1,X2)~N(0,-3,5,100,22)")
```

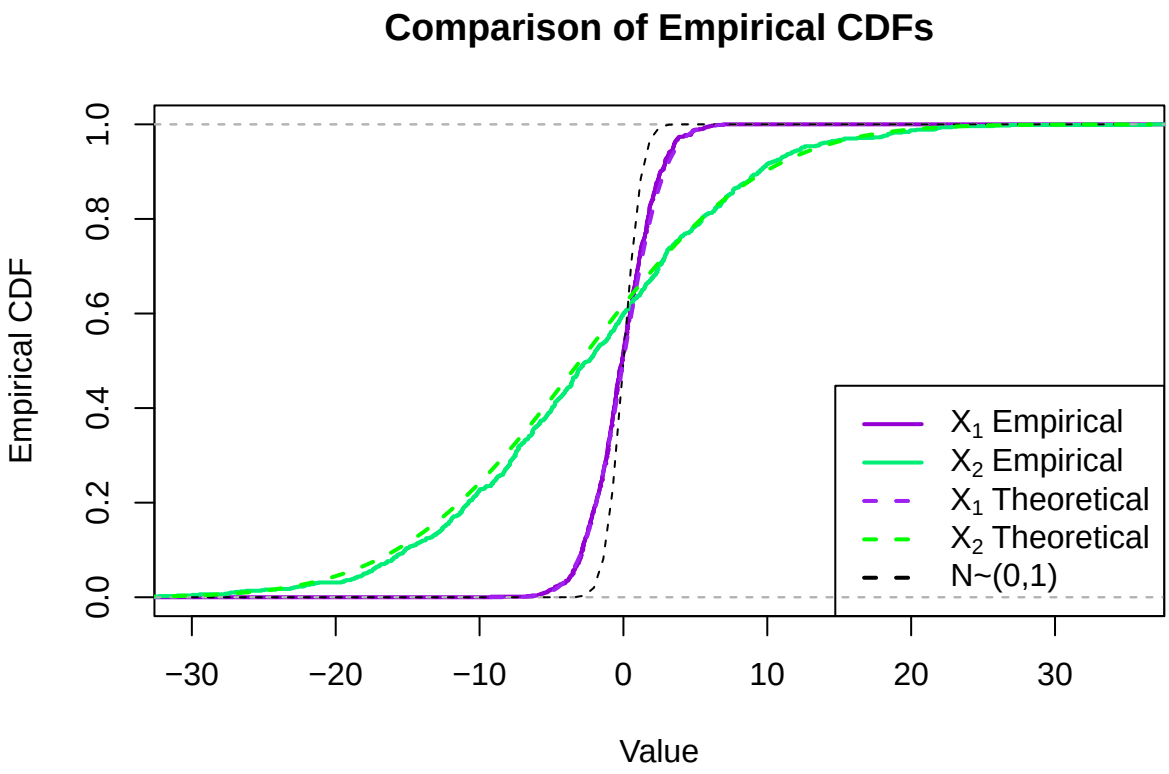


The highest density region (peak) clearly appears near the mean coordinates $(0, -3)$, precisely as expected based on our theoretical means. The elongated shape along the diagonal visually highlights the strong negative covariance $\sigma_{12} \approx -22.14$, confirming that high values of X_1 tend to associate with low values of X_2 and vice versa. The density decreases smoothly and symmetrically away from the peak, following a typical bell-shaped Gaussian distribution in two dimensions. The narrowness and elongation of the distribution surface clearly indicate a very strong negative correlation, resulting in limited dispersion perpendicular to the main axis of elongation.

2) Plot the two empirical cumulative distribution functions in a single graph and comment on each one and on the differences.

```
plot(ecdf(xx[,1]),
     do.points = FALSE,
     col = "darkviolet",
     lwd = 2,
     main = "Comparison of Empirical CDFs",
     xlab = "Value",
     ylab = "Empirical CDF",
     xlim = c(-30,35))

lines(ecdf(xx[,2]),
      do.points = FALSE,
      col = "springgreen2",
      lwd = 2)
```

```r
curve(pnorm(x), lty = 2, add = TRUE)


curve(pnorm(x, mean = 0, sd = sqrt(5)),
      from = min(xx[,1]),
      to = max(xx[,1]),
      col = "purple",
      lty = 2,
      lwd = 2,
      add = TRUE)

curve(pnorm(x, mean = -3, sd = 10),
      from = min(xx[,2]),
      to = max(xx[,2]),
      col = "green",
      lty = 2,
      lwd = 2,
      add = TRUE)


legend("bottomright",
       legend = c(expression(X[1]~"Empirical"), expression(X[2]~"Empirical"),
                  expression(X[1]~"Theoretical"), expression(X[2]~"Theoretical"),
                  "N~(0,1)"),
       col = c("darkviolet", "springgreen2", "purple", "green", "black"),
       lty = c(1,1,2,2, 2),
       lwd = 2)
```



The graph above displays the empirical cumulative distribution functions (ECDFs) of both variables X_1 and X_2.

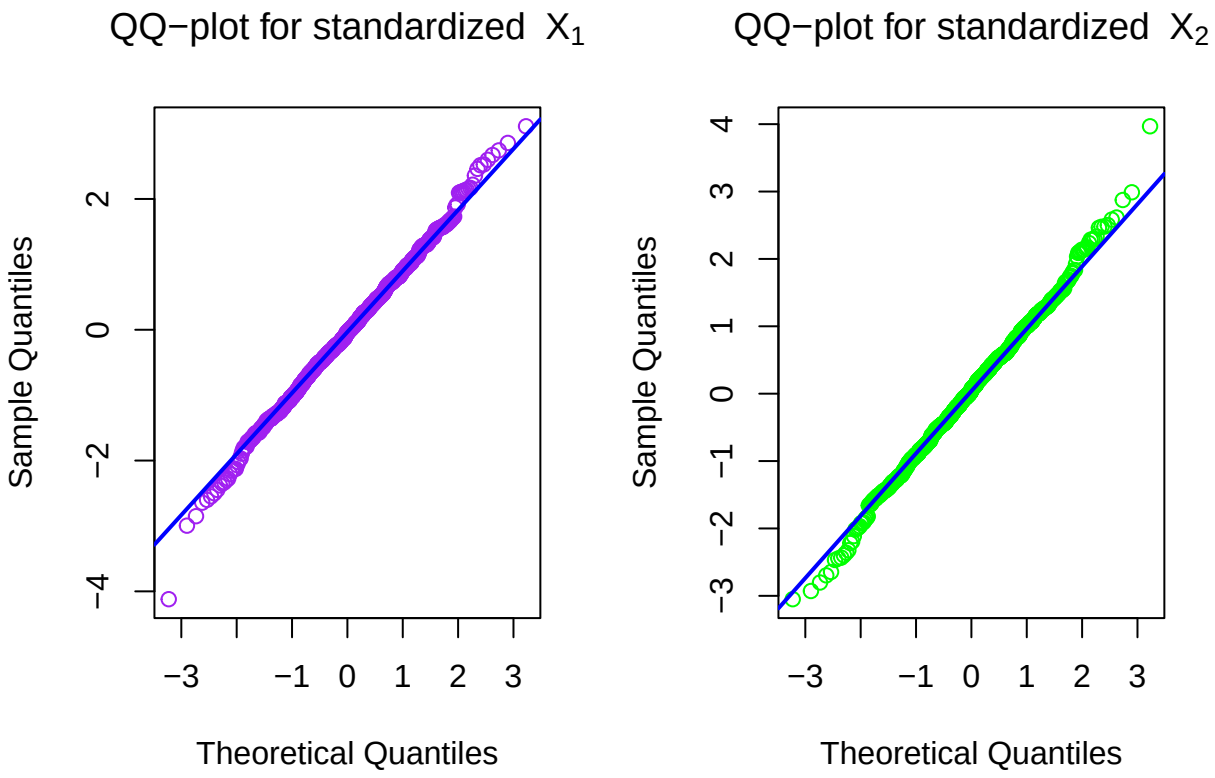
- **ECDF for** X_1 (purple line): it closely matches the theoretical Gaussian distribution $N \sim (0, 5)$, as evidenced by the dashed purple line. Its distribution is relatively concentrated around the mean ($\mu_1 = 0$) with smaller variance ($\sigma_1^2 = 5$), meaning the data points of X_1 are more closely grouped around their theoretical center, and tails are comparatively lighter.
- **ECFD for** X_2 (green line): this ECDF reflects a broader distribution, aligning well with the theoretical Gaussian distribution $N(-3, 100)$ (dashed green line). The greater variance ($\sigma_2^2 = 100$) manifests as wider dispersion, with heavier tails indicating that values of X_2 spread more widely around its theoretical mean ($\mu_2 = -3$). This characteristic results in a flatter slope of the ECDF curve, reflecting greater data variability.

Comparing both ECDFs highlights clearly the significant difference in their respective variances. The ECDF of X_2 indicates much greater variability and range that that of X_1, consistent with their theoretical definitions. Moreover, the close alignment between empirical and theoretical curves for both variables further confirms that our sample faithfully represents the specified theoretical distributions.

3) Provide the quantile-quantile plots on a single graphical window and interprete them.

We use standardized variables (subtracting theoretical means and dividing by theoretical standard deviations) to directly compare with the standard normal distribution $N(0, 1)$:

```r
par(mfrow = c(1,2))


qqnorm((xx[,1]-0)/sqrt(5), col = "purple",
       main = expression("QQ-plot for standardized "~X[1]),
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles")
qqline((xx[,1]-0)/sqrt(5), col = "blue", lwd = 2)


qqnorm((xx[,2]+3)/10, col = "green",
       main = expression("QQ-plot for standardized "~X[2]),
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles")
qqline((xx[,2]+3)/10, col = "blue", lwd = 2)
```



Standardizing variables allows direct comparison with the standard normal distribution, simplifying interpretation. The closer the empirical quantiles are to the diagonal reference line, the closer the distribution is to the theoretical Gaussian distribution. These QQ-plots provide visual evidence for assessing how closely each variable's distribution aligns with the theoretical Gaussian distributions:

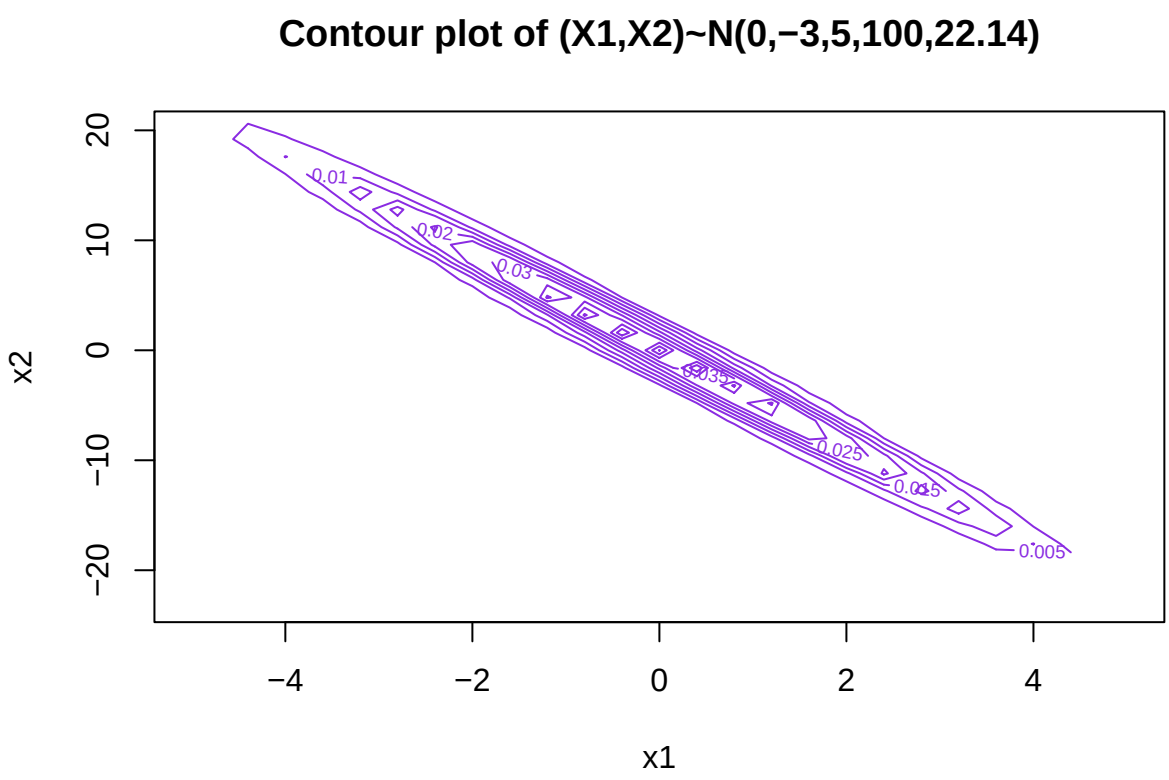
- **QQ-plot for standardized** X_1: the data points closely follow the reference line, confirming the empirical distribution of X_1 matches very well with its theoretical Gaussian $N(0, 5)$. Minor deviations observed at the extreme tails are normal in finite samples and do not suggest significant deviations from normality.

- **QQ-plot for standardized X_2**: the standardized X_2 points similarly align well with the theoretical reference line. Slight deviations visible at the tails reflect natural sample variability, particularly given the higher variance of X_2. Such small discrepancies are expected and do not indicate meaningful departures from the theoretical Gaussian assumption.

Overall, both QQ-plots strongly support the assumption of notmality for both variables, validating the robustness and correctness of the data-generation procedure.

4) Depict the contour plots obtained from the theoretical distribution and comment on their shapes.

```
require(mvtnorm)

x1 <- seq(-10, 10, length = 51)
x2 <- seq(-40, 40, length = 51)

dens <- matrix(dmvnorm(expand.grid(x1, x2),
                       sigma=sigma),
               ncol = length(x1))

contour(x1, x2, dens,
        main = "Contour plot of (X1,X2)~N(0,-3,5,100,22.14)",
        col = "blueviolet",
        xlim = c(-5,5),
        ylim = c(-23,20),
        xlab = "x1",
        ylab = "x2")
```



Contour plots represent the joint density function by lines of equal density (iso-density curves). The contour lines appear elliptical and notably elongated, clearly visualizing the strong negative correlation between the two variables. This is also enforced by the direction of the major axis of the ellipses, which runs diagonally. Specifically, as the values of X_1 increase, the values of X_2 tend to decrease and vice versa. The substantial elongation of the ellipses highlights a significant difference in variance along the principal axes, suggesting that one eigenvalue might be considerably larger than the other. The narrow width of the contours perpendicular to the main diagonal direction demonstrates very limited variability, reinforcing the strong linear dependency between the variables. The center of the contour ellipses visually corresponds closely to the theoretical mean vector $(0, -3)$, confirming the correct parameter specification and data generation. Inner contours represent regions of higher joint probability density, while outer contours indicate areas of lower probability density. The contours clearly convey that observations are densely packed around the central line, diminishing sharply as we move away from this line. Overall, the contour plot confirms both theoretically and visually all prior statistical interpretations, clearly demonstrating the strong negative correlation and variance-covariance structure of the bivariate Gaussian distribution.

5) Extract the eigenvalues and eigenvectors of the variance-covariance matrix and comment on them. Show the spectral decomposition.

The eigen-decomposition of the covariance matrix provides formal mathematical confirmation of the visual and statistical insights gained.

```
EV <- eigen(sigma)
```

```r
EV$values
```

```
## [1] 104.90515242   0.09484758
```

```r
EV$vectors
```

```
##            [,1]       [,2]
## [1,] -0.2163337 -0.9763195
## [2,]  0.9763195 -0.2163337
```

These values, obtained from the covariance matrix, allow us to gain further insight into the structure and orientation of the bivariate distribution.

The **eigenvalues** quantify the variability along the principal direction (principal axes). The larger eigenvalue (104.90515242) indicates the direction along which the data show the greatest variability. The smaller eigenvalue (0.09484758) indicates the perpendicular direction with smaller variability.

The **eigenvectors** represent the direction (axes) along which the covariance matrix is diagonalized. They are orthogonal, forming a rotation to diagonalize covariance and clearly indicate how data variability is structured and oriented.

Spectral decomposition The covariance matrix Σ can be represented using its eigen (spectral) decomposition as:

$$\Sigma = Q\Lambda Q^T$$

where Q is the matrix whose columns are the eigenvectors of Σ and A is the diagonal matrix whose diagonal elements are eigenvalues.

```r
Q <- EV$vectors
lambda <- diag(EV$values)

Q %*% lambda %*% t(Q)
```

```
##           [,1]      [,2]
## [1,]   5.00000 -22.13707
## [2,] -22.13707 100.00000
```

Comparing the reconstructed matrix with the original covariance matrix (sigma), we see they match exactly. This confirms the correctness of our spectral decomposition.

For the statistical interpertation, **principal directions** (given by eigenvectors) indicate how data are primarily oriented. Given our strong negative covariance and correlation, we expect one eigenvector to point diagonally in the direction of maximum spread (negative slope), and the other to be perpendicular. **Variability** is then explained by eigenvectors: due to the large difference between the eigenvalues, we observe one clear principal direction that captures most of the variability (larger eigenvalue), confirming a strong linear dependence between X_1 and X_2. The second direction (small eigenvalue) captures much less variance, reinforcing the idea of strong collinearity. Overall, the eigen decomposition reinforces our previous observations and provides a robust, mathematical characterization of the distribution's variance-covariance structure.

6) List the properties of the bivariate normal distribution.

- **Marginal distributions**: the marginal distributions of a bivariate normal are themselves univariate Gaussian distributions. Specifically:

$$X \sim N(\mu_X, \sigma_X^2)$$

and

$$Y \sim N(\mu_Y, \sigma_Y^2)$$

This means that taking only one of the two variables still results in a normal distribution.

- **Conditional distributions**: all conditional distributions derived from a bivariate normal distribution are also Gaussian distributions. In particular, the conditional distributions are given by:
- For the distribution of X given $Y = y$:

$$X|Y = y \sim N(\mu_X + \rho_{XY}\frac{\sigma_X}{\sigma_Y}(y - \sigma_Y), \sigma_X^2(1 - \rho_{XY}^2))$$

- Similarly, for the distribution of Y given $X = x$:

$$Y|X = x \sim N(\mu_X + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \sigma_X), \sigma_Y^2(1 - \rho_{XY}^2))$$

Thus, the conditional means are linearly related to the conditioning variable, and the conditional variances are constant.

- **Correlation and independence**: in general, two random variables being uncorrelated does not necessarily imply independence. However, the bivariate Gaussian distribution is a notable exception. If two jointly Gaussian random variables are uncorrelated ($\rho_{XY} = 0$), then they are also independent. Formally:

$$\rho_{XY} = 0 \iff f(x, y) = f(x)f(y)$$

- **Geometric properties: contour lines**: the contours (level curves) of constant density for a bivariate Gaussian distribution are always elliptical. The shape and orientation of these ellipses depend on the correlation coefficient ρ_{XY}. If $\rho_{XY} = 0$, the ellipses have axes parallel to the coordinate axes, resulting in symmetry with no diagonal orientation. When $\rho_{XY} \neq 0$, the ellipses is rotated. The direction of elongation (major axis) corresponds to the sign and magnitude of ρ_{XY}. A contour plot visualizes these elliptical shapes clearly, illustrating how variables vary jointly.
- **Linear combinations**: any linear combination of jointly Gaussian random variables is itself normally distributed. For example, if we define

$$Z = aX + bY$$

then:

$$Z \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY})$$

This property implies the closure of normal distributions under linear transformations.