Statistical learning - Report 1: Vector and Matrix Algebra, Multivariate Normal Distribution

Julia Kiczka

March 20, 2025

Problem 1

Considered the matrix:

$$A = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$$

1. Is A symmetric?

A matrix is symmetric if $A^T = A$. Since:

$$A^T = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix} = A,$$

A is symmetric.

2. Spectral Decomposition of A

The characteristic equation is obtained from:

$$\det(A - \lambda I) = \begin{vmatrix} 3-\lambda & -1 \\ -1 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 - (-1)^2 = 9 - 6\lambda + \lambda^2 - 1 = \lambda^2 - 6\lambda + 8.$$

Solving $\lambda^2 - 6\lambda + 8 = 0$,

$$(\lambda - 4)(\lambda - 2) = 0.$$

Thus, the eigenvalues are $\lambda_1 = 4$ and $\lambda_2 = 2$.

For $\lambda_1 = 4$, solving $(A - 4I)x = 0$:

$$\begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.$$

This gives $x_1 = -x_2$, so an eigenvector is $e_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

For $\lambda_2 = 2$, solving $(A - 2I)x = 0$:

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.$$

This gives $x_1 = x_2$, so an eigenvector is $e_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

3. Spectral Decomposition Representation

In the formula $A = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T$, we can observe that $e_1 e_1^T$ and $e_2 e_2^T$ are marices.

$$e_1 e_1^T = \frac{1}{2}\begin{bmatrix} 1 \\ -1 \end{bmatrix}\begin{bmatrix} 1 & -1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$
$$e_2 e_2^T = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Thus,

$$A = 4 \cdot \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + 2 \cdot \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

4. Computing $\sqrt{A}$

Finding $\sqrt{A}$ Using Spectral Decomposition For a symmetric matrix A, the spectral theorem states that it can be decomposed as:

$$A = P\Lambda P^T$$

where P is an orthogonal matrix whose columns are the eigenvectors of A, and Λ is a diagonal matrix containing the eigenvalues λ_i.To define $\sqrt{A}$, we take the square root of each eigenvalue in Λ:

$$\sqrt{A} = P\sqrt{\Lambda}P^T, \quad \text{where} \quad \sqrt{\Lambda} = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots)$$

$$\sqrt{\Lambda} = \text{diag}(\sqrt{4}, \sqrt{2}) = \begin{bmatrix} 2 & 0 \\ 0 & \sqrt{2} \end{bmatrix}$$

The matrix P with the normalized eigenvectors as columns is:

$$P = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Verification: Multiplying $\sqrt{A}$ by itself:

$$\sqrt{A} \cdot \sqrt{A} = (P\sqrt{\Lambda}P^T)(P\sqrt{\Lambda}P^T) = P\sqrt{\Lambda}(P^T P)\sqrt{\Lambda}P^T = P(\Lambda)P^T = A.$$

Problem 2

Consider the spectral decomposition of a positive definite matrix:

$$A = P\Lambda P^T$$

where P is an orthonormal matrix containing eigenvectors e_i as columns, and Λ is a diagonal matrix containing the corresponding positive eigenvalues.

1. $P^T = P^{-1}$

A square matrix P is **orthonormal** if its columns (or rows) are **orthonormal vectors**. This means:

- Each column has unit length: $\|p_i\| = 1$.
- Each pair of columns is mutually perpendicular (orthogonal): $p_i^T p_j = 0$ for $i \neq j$.

Above conditions can written as:

$$P^T P = I.$$

By definition, the inverse of P satisfies:

$$P^{-1} P = I.$$

From the uniqueness of the inverse, we see that:

$$P^{-1} = P^T.$$

2. Determinant of Λ is the Product of the Diagonal Elements

For any square matrix $A = [a_{ij}]$ of size $n \times n$, the determinant is defined recursively as:

$$\det(A) = \sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det(A_{1j}),$$

where A_{1j} is the $(n-1) \times (n-1)$ submatrix obtained by removing the first row and jth column of A. This definition is known as Laplace expansion along the first row. Since Λ is a diagonal matrix:

$$\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

we have all off-diagonal elements equal to zero, i.e., $\Lambda_{ij} = 0$ for all $i \neq j$. The determinant can be computed using Laplace expansion along the first row:

$$\det(\Lambda) = \sum_{j=1}^{n} (-1)^{1+j} \lambda_{1j} \det(\Lambda_{1j}).$$

Since all off-diagonal elements are zero, only the term where $j = 1$ contributes, and we obtain:

$$\det(\Lambda) = \lambda_1 \det(\Lambda_{11}),$$

where Λ_{11} is the $(n-1) \times (n-1)$ diagonal matrix:

$$\Lambda_{11} = \operatorname{diag}(\lambda_2, \lambda_3, \ldots, \lambda_n).$$

Applying the same argument recursively, we get:

$$\det(\Lambda_{11}) = \lambda_2 \det(\Lambda_{22}),$$

where Λ_{22} is the $(n-2) \times (n-2)$ diagonal matrix:

$$\Lambda_{22} = \operatorname{diag}(\lambda_3, \ldots, \lambda_n).$$

Continuing this process until we reach the 1×1 case, we conclude:

$$\det(\Lambda) = \lambda_1 \lambda_2 \cdots \lambda_n.$$

Thus, for any diagonal matrix, the determinant is simply the product of its diagonal elements.

3. Determinant of A is the same as that of Λ

Taking the determinant on both sides of the spectral decomposition:

$$\det(A) = \det(P \Lambda P^T).$$

Using the property $\det(AB) = \det(A)\det(B)$, we get:

$$\det(A) = \det(P)\det(\Lambda)\det(P^T).$$

Since P is an orthogonal matrix, we know that $P^T P = I$, which implies $\det(P^T)\det(P) = \det(I)$, so $\det(P^T) = \frac{1}{\det(P)}$:

$$\det(A) = \det(P)\det(\Lambda)\det(P^T) = \det(P)\det(\Lambda)\frac{1}{\det(P)} = \det(\Lambda).$$

Thus, $\det(A) = \det(\Lambda)$.

4. Finding Λ^{-1}

Given that Λ is a diagonal matrix:

$$\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

the inverse of a diagonal matrix is straightforward to compute. Since the off-diagonal elements are all zero, the inverse of Λ, denoted as Λ^{-1}, is simply another diagonal matrix where each diagonal entry is the reciprocal of the corresponding entry in Λ:

$$\Lambda^{-1} = \text{diag}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_n}\right).$$

This property follows directly from the definition of matrix inversion. For any diagonal matrix $D = \text{diag}(d_1, d_2, \ldots, d_n)$, the inverse matrix D^{-1} satisfies:

$$D \cdot D^{-1} = I,$$

where I is the identity matrix. Since the only non-zero elements in D and D^{-1} are along the diagonal, each diagonal element of D^{-1} must be the reciprocal of the corresponding diagonal element of D in order to satisfy this equation. Therefore, for Λ, the inverse matrix is simply obtained by replacing each diagonal element λ_i with $\frac{1}{\lambda_i}$.

5. Verifying $A^{-1} = P\Lambda^{-1}P^T$

Given the spectral decomposition:

$$A = P\Lambda P^T,$$

we claim that the inverse is given by:

$$A^{-1} = P\Lambda^{-1}P^T.$$

To verify this, we compute:

$$AA^{-1} = (P\Lambda P^T)(P\Lambda^{-1}P^T).$$

Using the associativity of matrix multiplication:

$$AA^{-1} = P\Lambda(P^TP)\Lambda^{-1}P^T.$$

Since P is orthogonal, we have $P^TP = I$, so:

$$AA^{-1} = P\Lambda I\Lambda^{-1}P^T = P(\Lambda\Lambda^{-1})P^T.$$

Since $\Lambda\Lambda^{-1} = I$, it follows that:

$$AA^{-1} = PIP^T = I.$$

Thus, we confirm that:

$$A^{-1} = P\Lambda^{-1}P^T.$$

6. Checking on the Example from Problem 1

We use the matrix:

$$A = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}.$$

From Problem 1, we found:

$$P = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}.$$

Thus, its inverse is:

$$\Lambda^{-1} = \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

Now, computing $A^{-1} = P\Lambda^{-1}P^T$:

$$P\Lambda^{-1} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

Multiplying by P^T:

$$A^{-1} = \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

Computing the product:

$$A^{-1} = \frac{1}{2} \begin{bmatrix} \frac{1}{4} + \frac{1}{2} & -\frac{1}{4} + \frac{1}{2} \\ -\frac{1}{4} + \frac{1}{2} & \frac{1}{4} + \frac{1}{2} \end{bmatrix}.$$

$$A^{-1} = \frac{1}{2} \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}.$$

$$A^{-1} = \begin{bmatrix} \frac{3}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{3}{8} \end{bmatrix}.$$

Direct computation of A^{-1} using Gaussian elimination also gives:

$$A^{-1} = \frac{1}{8} \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}.$$

Since the computed expression using spectral decomposition matches this result, we confirm that:

$$A^{-1} = P\Lambda^{-1}P^T.$$

Problem 3

Parameters μ and Σ

The given data states that the weight W and length L of newborns follow a bivariate normal distribution:

$$(W, L) \sim N(\mu, \Sigma).$$

The mean vector μ is:

$$\mu = \begin{bmatrix} \mathbb{E}[W] \\ \mathbb{E}[L] \end{bmatrix} = \begin{bmatrix} 3343 \\ 49.8 \end{bmatrix}.$$

The covariance matrix Σ is given by:

$$\Sigma = \begin{bmatrix} \mathrm{Var}(W) & \mathrm{Cov}(W, L) \\ \mathrm{Cov}(W, L) & \mathrm{Var}(L) \end{bmatrix}.$$

We use the standard deviations:

$$\sigma_W = 528, \quad \sigma_L = 2.5, \quad \rho_{WL} = 0.75.$$

The covariance is:

$$\mathrm{Cov}(W, L) = \rho_{WL} \cdot \sigma_W \cdot \sigma_L = 0.75 \times 528 \times 2.5 = 990.$$

Thus, the covariance matrix is:

$$\Sigma = \begin{bmatrix} 528^2 & 990 \\ 990 & 2.5^2 \end{bmatrix} = \begin{bmatrix} 278784 & 990 \\ 990 & 6.25 \end{bmatrix}.$$

Joint Density Function

The probability density function of a bivariate normal distribution is:

$$f(W, L) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right),$$

where $x = \begin{bmatrix} W \\ L \end{bmatrix}$.

Eigenvalues and Eigenvectors of Σ

The eigenvalues λ of Σ satisfy:

$$\det(\Sigma - \lambda I) = 0.$$

Expanding:

$$\begin{vmatrix} 278784 - \lambda & 990 \\ 990 & 6.25 - \lambda \end{vmatrix} = (278784 - \lambda)(6.25 - \lambda) - (990)^2 = 0.$$

Expanding above expression results in quadratic equation:

$$\lambda^2 - 278790\lambda + 762300 = 0$$

Roots of the quadratic equation are the eigenvalues:

$$\lambda_1 \approx 2.73, \quad \lambda_2 \approx 278787.52$$

For each eigenvalue $\lambda_i, i \in \{1, 2\}$, we solve:

$$(\Sigma - \lambda_i I)v = 0.$$

which results in :

$$e_1 \approx \begin{bmatrix} 0.0035 \\ -0.9999 \end{bmatrix}, \quad e_2 \approx \begin{bmatrix} -0.9999 \\ -0.0035 \end{bmatrix}$$

Ellipses and the Covariance Matrix

Ellipses can be used to represent contours of constant probability density for a bivariate normal distribution. Given a random vector X with mean vector μ and covariance matrix Σ, the contours of constant probability density satisfy the equation:

$$(x - \boldsymbol{\mu})^T \Sigma^{-1}(x - \boldsymbol{\mu}) = c,$$

For a constant c, the equation defines an ellipse centered at $\boldsymbol{\mu}$. In Figure 1, the ellipses represent the contours of constant probability density for a bivariate normal distribution, where each ellipse corresponds to different χ^2 quantiles, specifically for the 0.75 and 0.9 percentiles, i.e., $\chi^2_{2,0.75}$ and $\chi^2_{2,0.9}$. The eigenvectors indicate the principal directions of variability in the data. The lengths of the eigenvectors are scaled by the square roots of the corresponding eigenvalues. The longer eigenvector corresponds to the direction with the most variability, while a shorter eigenvector represents the direction with less variability. The shape of the ellipse provides further insight: an elongated ellipse suggests a strong correlation between the variables, while a more circular shape indicates weak or no correlation.

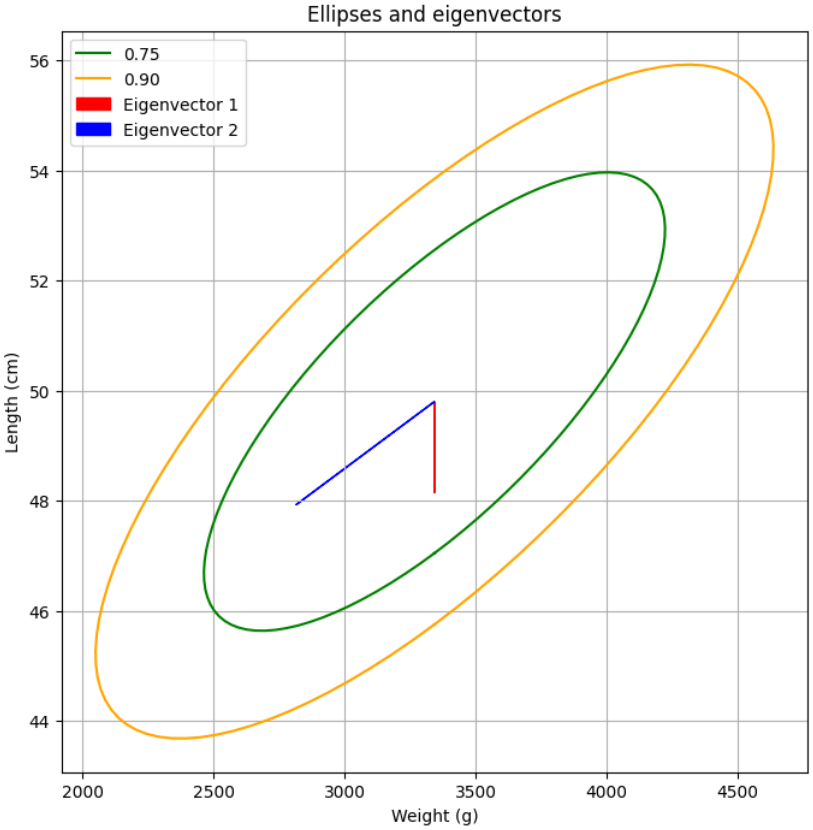

Figure 1: Elipses corresponding to the constant density contours of the joint distributions and eigenvectors scaled by the square roots of the corresponding eigenvalues.

Number of Parameters

A **bivariate normal distribution** (a normal distribution in two dimensions) is fully defined by **five parameters**:

1. **Mean Vector** (μ) – Contains the means of both variables (μ_1, μ_2), contributing **2 parameters**.
2. **Covariance Matrix** (Σ) – A 2×2 symmetric matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

   Since the covariance matrix is symmetric, it introduces **3 parameters**: two variances (σ_1^2, σ_2^2) and one covariance term (σ_{12}).

Thus, the total number of parameters in a bivariate normal distribution is:

$$2 \text{ (mean)} + 3 \text{ (covariance matrix)} = 5.$$

For a **p-dimensional normal distribution** in $\mathbb{R}^p$:

1. **Mean Vector** (μ), a $p \times 1$ vector, contributes p **parameters**.
2. **Covariance Matrix** (Σ), a $p \times p$ symmetric matrix, contributes:
   - p variance terms (diagonal elements).
   - $\frac{p(p-1)}{2}$ unique covariance terms (off-diagonal elements).

Thus, the total number of covariance parameters is:

$$p + \frac{p(p-1)}{2} = \frac{p(p+1)}{2}.$$

Therefore, the total number of parameters for a **p-dimensional normal distribution** is:

$$p + \frac{p(p+1)}{2} = \frac{p^2+3p}{2}.$$

Distribution of L

The variable L follows a **univariate normal distribution** because the marginal distribution of a bivariate normal distribution is also normal.

Marginal Distribution of L

Since we have the bivariate normal distribution:

$$(W, L) \sim N(\mu, \Sigma),$$

where

$$\mu = \begin{bmatrix} 3343 \\ 49.8 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 278784 & 990 \\ 990 & 6.25 \end{bmatrix},$$

the **marginal distribution** of L is given by:

$$L \sim N(\mu_L, \sigma_L^2),$$

where: $\mu_L = 49.8$ (mean of L) $\sigma_L^2 = \mathrm{Var}(L) = 6.25$ (variance of L, from the covariance matrix). Therefore, the distribution of L is:

$$L \sim N(49.8, 6.25),$$

or equivalently,

$$L \sim N(49.8, 2.5^2).$$

This means that L follows a **normal distribution** with mean 49.8 and standard deviation 2.5.

Estimating Newborn Length with the 3-σ Rule

Suppose the hospital records of a newborn child were lost. We aim to make an educated guess about the child's length based on statistical data.

From medical studies, the length of a full-term newborn follows an approximately normal distribution with:

- **Mean length**: $\mu = 49.8$ cm
- **Standard deviation**: $\sigma = 2.5$ cm

Best Guess (Expected Value)

Since the length follows a normal distribution, the best estimate (expected value) for the newborn's length is simply:

$$\hat{L} = \mu = 49.8 \text{ cm}$$

Accuracy Bounds (3-σ Rule)

The **3-σ rule** states that for a normal distribution, **99.7%** of values lie within $\mu \pm 3\sigma$. Applying this:

$$49.8 \pm (3 \times 2.5)$$

$$49.8 \pm 7.5$$

Thus, the estimated range for the newborn's length is:

$$\mathbf{42.3}\text{ cm} \leq \mathbf{L} \leq \mathbf{57.3}\text{ cm}$$

This means that with **99.7% confidence**, the newborn's length falls within the range **42.3 cm to 57.3 cm**.

Problem 4

Conditional Distribution of L

We are given that the joint distribution of length L and weight W is bivariate normal. The conditional distribution of L given $W = w$ is:

$$L|W = w \sim N\left(\mu_L + \rho\frac{\sigma_L}{\sigma_W}(w - \mu_W), \sigma_L^2(1 - \rho^2)\right)$$

Given the following values:

$$\mu_L = 49.8, \quad \sigma_L = 2.5, \quad \mu_W = 3343, \quad \sigma_W = 528, \quad \rho = 0.75, \quad w = 4025,$$

the mean of the conditional distribution is:

$$\mu_{L|W} = \mu_L + \rho\frac{\sigma_L}{\sigma_W}(w - \mu_W)$$

Substituting the values:

$$\mu_{L|W} = 49.8 + 0.75 \cdot \frac{2.5}{528} \cdot (4025 - 3343)$$

$$\mu_{L|W} = 49.8 + 0.75 \cdot \frac{2.5}{528} \cdot 682 = 49.8 + 0.75 \cdot 3.23 = 49.8 + 2.4225 \approx 52.22$$

The variance of the conditional distribution is:

$$\sigma_{L|W}^2 = \sigma_L^2(1 - \rho^2)$$

Substituting the values:

$$\sigma_{L|W}^2 = 2.5^2 \cdot (1 - 0.75^2) = 6.25 \cdot (1 - 0.5625) = 6.25 \cdot 0.4375 = 2.734375$$

Thus, the standard deviation is:

$$\sigma_{L|W} = \sqrt{2.734375} \approx 1.65$$

Therefore, the conditional distribution of L given $W = 4025$ is:

$$L|W = 4025 \sim N(52.22, 1.65^2)$$

Best Guess (Expected Value)

After knowing the weight, the best guess for the length is the mean of the above distribution:

$$L_{\text{best guess}} = 52.22\,\text{cm}$$

Accuracy Bounds Using the 3-σ Rule

Lower Bound:

$$52.22 - 3 \cdot 1.65 = 52.22 - 4.95 = 47.27\,\text{cm}$$

Upper Bound:

$$52.22 + 3 \cdot 1.65 = 52.22 + 4.95 = 57.17\,\text{cm}$$

Thus, based on the 3-σ rule, we can say that the best guess for the newborn's length is 52.22 cm.The accuracy bounds for this estimate, with 99.7% confidence, are between 47.27 cm and 57.17 cm.

Conclusions

By observing that the weight $W = 4025\,\text{g}$ is higher than the typical mean weight ($\mu_W = 3343\,\text{g}$), we can infer that the estimated length L is also expected to be above its average value. This is consistent with our finding that the conditional mean of $L|W$ which is equal to 52.22 is higher than the unconditional mean of L which is equal to 49.8. Moreover, the conditional variance in $L|W$ is smaller, indicating that the uncertainty in our estimate of length decreases once we know the weight. This reduction in variance suggests that knowing the weight provides us with a more precise prediction of the newborn's length.

Problem 5

Let X_1, X_2, X_3 be independent $N(\mu, \Sigma)$ random vectors of dimension p. Find the distribution of each of the following vectors and their joint distribution.

$$V_1 = \frac{1}{4}X_1 - \frac{1}{2}X_2 + \frac{1}{4}X_3$$

$$V_2 = \frac{1}{4}X_1 - \frac{1}{2}X_2 - \frac{1}{4}X_3$$

Distribution of V_1

Since X_1, X_2, X_3 are independent, the linear combination V_1 will also follow a multivariate normal distribution. The mean and covariance of V_1 are computed as follows:

Mean of V_1:

$$\mathbb{E}[V_1] = \frac{1}{4}\mathbb{E}[X_1] - \frac{1}{2}\mathbb{E}[X_2] + \frac{1}{4}\mathbb{E}[X_3] = \frac{1}{4}\mu - \frac{1}{2}\mu + \frac{1}{4}\mu = 0$$

Covariance of V_1:

Since X_1, X_2, X_3 are independent, the covariance matrix of V_1 is the sum of the covariances of the individual terms:

$$\text{Cov}(V_1) = \frac{1}{4^2}\Sigma + \left(\frac{-1}{2}\right)^2 \Sigma + \frac{1}{4^2}\Sigma = \frac{1}{16}\Sigma + \frac{1}{4}\Sigma + \frac{1}{16}\Sigma = \frac{2}{16}\Sigma + \frac{4}{16}\Sigma = \frac{6}{16}\Sigma = \frac{3}{8}\Sigma$$

Thus, the distribution of V_1 is:

$$V_1 \sim N\left(0, \frac{3}{8}\Sigma\right)$$

Distribution of V_2

Similarly, for V_2:

Mean of V_2:

$$\mathbb{E}[V_2] = \frac{1}{4}\mathbb{E}[X_1] - \frac{1}{2}\mathbb{E}[X_2] - \frac{1}{4}\mathbb{E}[X_3] = \frac{1}{4}\mu - \frac{1}{2}\mu - \frac{1}{4}\mu = -\frac{\mu}{2}$$

Covariance of V_2:

$$\text{Cov}(V_2) = \frac{1}{4^2}\Sigma + \left(\frac{-1}{2}\right)^2\Sigma + \left(\frac{-1}{4}\right)^2\Sigma = \frac{1}{16}\Sigma + \frac{1}{4}\Sigma + \frac{1}{16}\Sigma = \frac{2}{16}\Sigma + \frac{4}{16}\Sigma = \frac{6}{16}\Sigma = \frac{3}{8}\Sigma$$

Thus, the distribution of V_2 is:

$$V_2 \sim N\left(-\frac{\mu}{2}, \frac{3}{8}\Sigma\right)$$

Joint distribution

The mean of V_1 is 0 and for V_2 is $-\frac{\mu}{2}$, so the joint mean vector is:

$$\mathbb{E}[\mathbf{V}] = \begin{pmatrix} \mathbb{E}[V_1] \\ \mathbb{E}[V_2] \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{\mu}{2} \end{pmatrix}$$

The joint covariance matrix has the following block structure:

$$\text{Cov}(\mathbf{V}) = \begin{pmatrix} \text{Cov}(V_1, V_1) & \text{Cov}(V_1, V_2) \\ \text{Cov}(V_2, V_1) & \text{Cov}(V_2, V_2) \end{pmatrix}$$

We need to compute the following terms:

1. $\text{Cov}(V_1, V_1)$ — the variance of V_1
2. $\text{Cov}(V_2, V_2)$ — the variance of V_2
3. $\text{Cov}(V_1, V_2)$ — the covariance between V_1 and V_2
4. $\text{Cov}(V_2, V_1)$ — the covariance between V_2 and V_1

1. $\text{Cov}(V_1, V_1)$:

$$\text{Cov}(V_1, V_1) = \text{Cov}\left(\frac{1}{4}X_1 - \frac{1}{2}X_2 + \frac{1}{4}X_3, \frac{1}{4}X_1 - \frac{1}{2}X_2 + \frac{1}{4}X_3\right)$$

Since X_1, X_2, X_3 are independent:

$$\text{Cov}(X_1, X_2) = \text{Cov}(X_1, X_3) = \text{Cov}(X_2, X_3) = 0$$

Expanding the covariance:

$$\text{Cov}(V_1, V_1) = \frac{1}{16}\text{Cov}(X_1, X_1) + \frac{1}{4}\text{Cov}(X_2, X_2) + \frac{1}{16}\text{Cov}(X_3, X_3)$$

Thus, the covariance reduces to:

$$\text{Cov}(V_1, V_1) = \frac{1}{16}\Sigma + \frac{1}{4}\Sigma + \frac{1}{16}\Sigma = \frac{2}{16}\Sigma + \frac{4}{16}\Sigma = \frac{6}{16}\Sigma = \frac{3}{8}\Sigma$$

2. $\text{Cov}(V_2, V_2)$:
Similarly, for V_2, we compute:

$$\text{Cov}(V_2, V_2) = \text{Cov}\left(\frac{1}{4}X_1 - \frac{1}{2}X_2 - \frac{1}{4}X_3, \frac{1}{4}X_1 - \frac{1}{2}X_2 - \frac{1}{4}X_3\right)$$

Expanding this, since X_1, X_2, X_3 are independent:

$$\text{Cov}(V_2, V_2) = \frac{1}{16}\text{Cov}(X_1, X_1) + \frac{1}{4}\text{Cov}(X_2, X_2) + \frac{1}{16}\text{Cov}(X_3, X_3)$$

$$\text{Cov}(V_2, V_2) = \frac{1}{16}\Sigma + \frac{1}{4}\Sigma + \frac{1}{16}\Sigma = \frac{2}{16}\Sigma + \frac{4}{16}\Sigma = \frac{6}{16}\Sigma = \frac{3}{8}\Sigma$$

3. $\text{Cov}(V_1, V_2)$:

$$\mathrm{Cov}(V_1, V_2) = \mathrm{Cov}\left(\frac{1}{4}X_1 - \frac{1}{2}X_2 + \frac{1}{4}X_3, \frac{1}{4}X_1 - \frac{1}{2}X_2 - \frac{1}{4}X_3\right)$$

Expanding this, since X_1, X_2, X_3 are independent, all cross-covariances between different X_i's are zero and we get:

$$\mathrm{Cov}(V_1, V_2) = \frac{1}{16}\Sigma + \frac{1}{4}\Sigma + \frac{1}{16}\Sigma = \frac{2}{16}\Sigma - \frac{4}{16}\Sigma = \frac{1}{4}\Sigma$$

4. $\mathrm{Cov}(V_2, V_1)$:
Since covariance is symmetric, we know that:

$$\mathrm{Cov}(V_2, V_1) = \mathrm{Cov}(V_1, V_2) = \frac{1}{4}\Sigma$$

Final Joint Covariance Matrix:
Now, we can combine all the computed covariance terms to get the full joint covariance matrix:

$$\mathrm{Cov}(\mathbf{V}) = \begin{pmatrix} \frac{3}{8}\Sigma & \frac{1}{4}\Sigma \\ \frac{1}{4}\Sigma & \frac{3}{8}\Sigma \end{pmatrix}$$

Project 1 - Part one

1) Using the date estimate the mean and the covariance for the length and the weight of children

Mean of each column:

Column	Mean
Weight	3233.545
Length	49.238

Covariance matrix:

$$\Sigma = \begin{bmatrix} \mathrm{Var(Weight)} & \mathrm{Cov(Weight,\ Length)} \\ \mathrm{Cov(Weight,\ Length)} & \mathrm{Var(Length)} \end{bmatrix} = \begin{bmatrix} 220276.658 & 915.296 \\ 915.296 & 4.443 \end{bmatrix}$$

where:

- Var(Weight) = 220276.658 is the variance of the weight.
- Var(Length) = 4.443 is the variance of the length.
- Cov(Weight, Length) = 915.296 is the covariance between weight and length.

2) Verify graphically the normal distribution of the data. Use scatterplots and qq-plots for the marginal distributions.

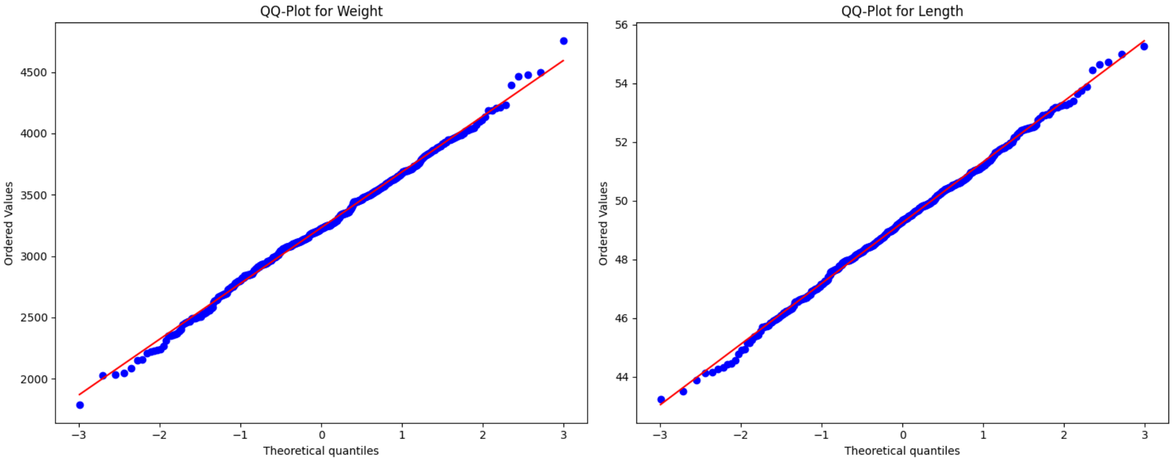

Figure 2: QQ-Plots for weight and height

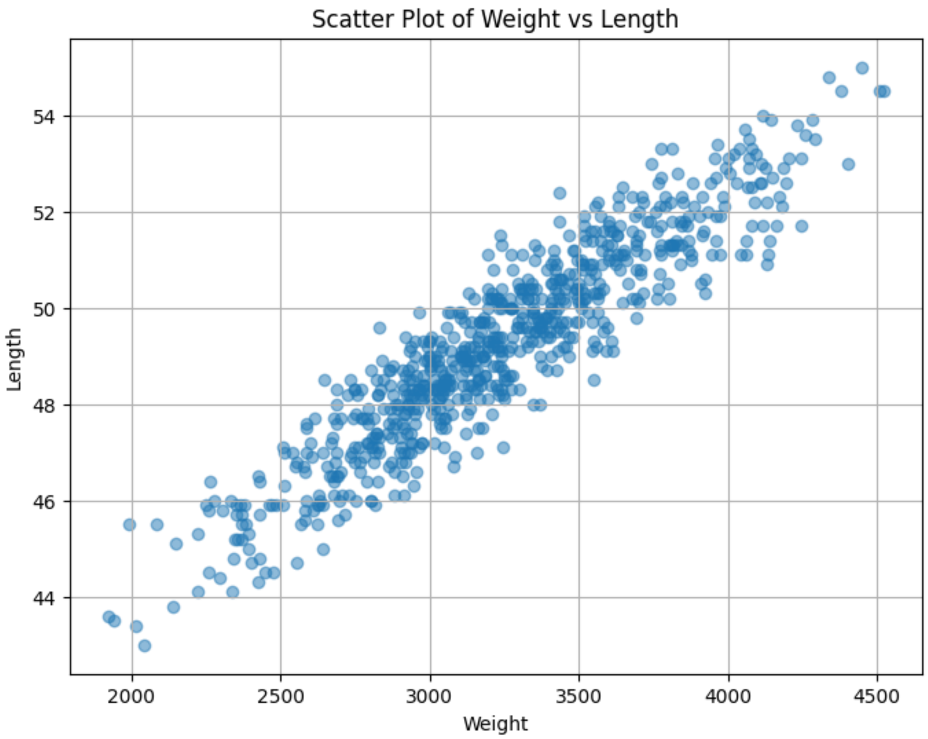

Figure 3: Scatterplot of the data (weight vs length)

QQ-plots and scatterplots enable us to verify that the data follows a multivariate normal distribution, for which the marginal distributions are known to be normal.

3) Find the ellipsoids that would serve classification regions for scores as described above.

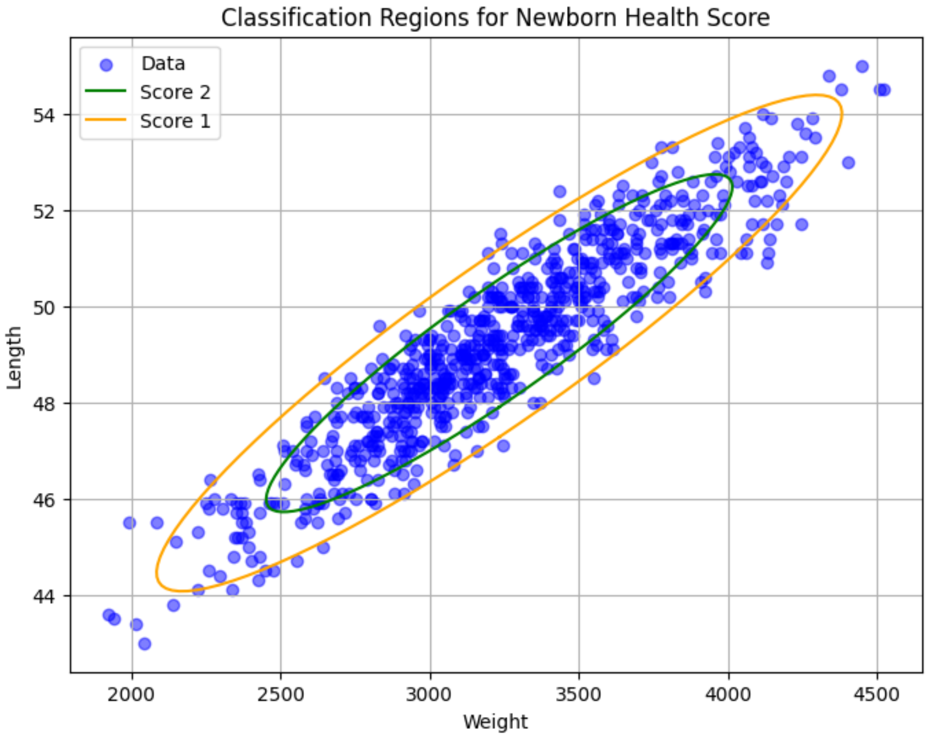

Figure 4: Ellipsoids as classification regions for scores

The ellipses represent classification regions for children's scores based on their weight and length. The classification is determined using confidence regions of the bivariate normal distribution.The equation defining these ellipses is:

$$(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c,$$

where:

- $\mathbf{x}$ is the vector of weight and length,
- $\boldsymbol{\mu}$ is the mean vector,
- Σ is the covariance matrix,
- c is a threshold based on **chi-squared quantiles** corresponding to the desired probability levels.

From the problem description, classification is based on the following confidence levels:

- The **95% quantile** ($c \approx \chi^2_{2,0.95}$) defines the **outer ellipsoid**, beyond which a score of **0** is assigned.
- The **75% quantile** ($c \approx \chi^2_{2,0.75}$) defines the **inner ellipsoid**, within which a score of **2** is assigned.
- For measurements falling **between** these ellipsoids, a score of **1** is assigned.

Thus, the ellipsoids serve as **decision boundaries** for assigning scores based on weight and length measurements.

4) How many children would score zero, one, and two, respectively? Illustrate this classification on the graphs.

Score	Count
0	38
1	157
2	541

Table 1: Classification scores and respective count

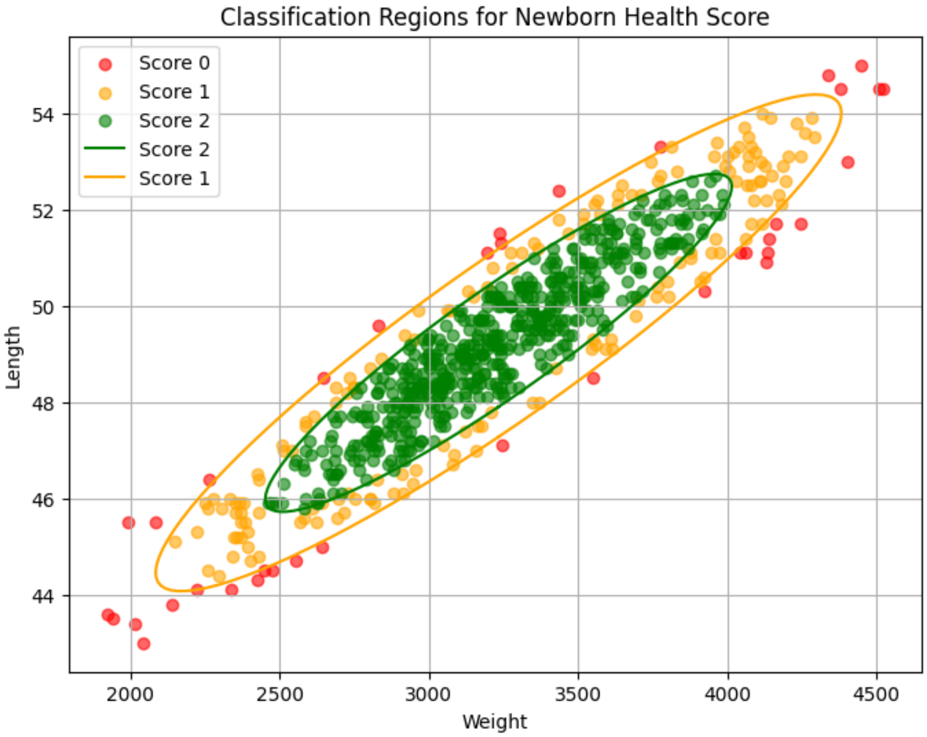

Figure 5: Ellipsoids as classification regions for scores

5) Find the Spectral Decomposition of the Estimated Covariance Matrix

The spectral decomposition of a matrix Σ involves expressing it as:

$$\Sigma = V\Lambda V^\top$$

where V is the matrix of eigenvectors, Λ is the diagonal matrix of eigenvalues, $V^\top$ is the transpose of V. Given the eigenvalue matrix Λ and eigenvector matrix V:

$$\Lambda = \begin{bmatrix} 0.63671 & 0 \\ 0 & 220280.463 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.004 & -0.999 \\ -0.999 & -0.004 \end{bmatrix}$$

We can reconstruct the covariance matrix Σ as:

$$\Sigma = V\Lambda V^\top = \begin{bmatrix} 0.004 & -0.999 \\ -0.999 & -0.004 \end{bmatrix} \begin{bmatrix} 0.637 & 0 \\ 0 & 220280.463 \end{bmatrix} \begin{bmatrix} 0.004 & -0.999 \\ -0.999 & -0.004 \end{bmatrix}^\top$$

We observe that the covariance matrix exhibits one dominant eigenvalue and one relatively small eigenvalue, leading to a highly elongated ellipse in the corresponding plot. The major axis of this ellipse is aligned with the eigenvector associated with the largest eigenvalue, signifying the principal direction of variance in the data.

6. Plot the data transformed according to $P^T X$, where P is the matrix made of the eigenvectors standing as the columns. Interpret the transformed data.

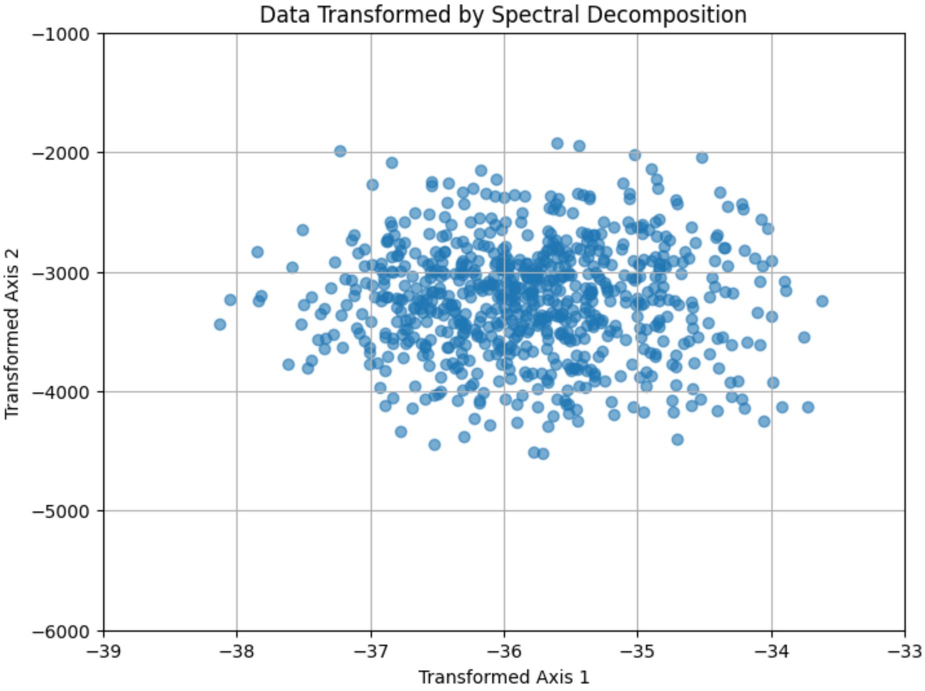

Figure 6: Data transformed according to $P^T X$, where P is the matrix made of the eigenvectors as columns.

If $X \sim \mathcal{N}(\mu, \Sigma)$ and P is an orthogonal matrix containing eigenvectors of Σ, then the transformed variable $Y = P^T X$ follows $Y \sim \mathcal{N}(P^T \mu, P^T \Sigma P)$. Since $P^T \Sigma P = \Lambda$ is a diagonal matrix with eigenvalues of Σ, the components of Y are uncorrelated, and independent if Σ is full rank. The plots present that data is expressed in a new coordinate system where the axes (basis vectors) are the eigenvectors of the covariance matrix. These eigenvectors are orthogonal, meaning they form a new set of uncorrelated features.

Project 1 - Part two

1) Using the data estimate the mean and the covariance for all four variables

Mean of each column:

Column	Mean
Father height	177.42
Mother height	166.92
Weight	3233.55
Length	49.24

Covariance matrix:

The covariance matrix Σ for the variables $\text{FatherHeight}, \text{MotherHeight}, \text{Weight}, \text{Length}$ is given by:

$$\Sigma = \begin{bmatrix} \text{Var(FatherHeight)} & \text{Cov(FatherHeight, MotherHeight)} & \text{Cov(FatherHeight, Weight)} & \text{Cov(FatherHeight, Length)} \\ \text{Cov(MotherHeight, FatherHeight)} & \text{Var(MotherHeight)} & \text{Cov(MotherHeight, Weight)} & \text{Cov(MotherHeight, Length)} \\ \text{Cov(Weight, FatherHeight)} & \text{Cov(Weight, MotherHeight)} & \text{Var(Weight)} & \text{Cov(Weight, Length)} \\ \text{Cov(Length, FatherHeight)} & \text{Cov(Length, MotherHeight)} & \text{Cov(Length, Weight)} & \text{Var(Length)} \end{bmatrix}$$

Substituting the given values:

$$\Sigma = \begin{bmatrix} 12.6121 & 0.6310 & 931.8590 & 3.2895 \\ 0.6310 & 9.7721 & 827.2878 & 2.8521 \\ 931.8590 & 827.2878 & 220276.6577 & 915.2955 \\ 3.2895 & 2.8521 & 915.2955 & 4.4433 \end{bmatrix}$$

2) Verify graphically the normal distribution of the data. Use scatterplots and qq-plots for the marginal distributions

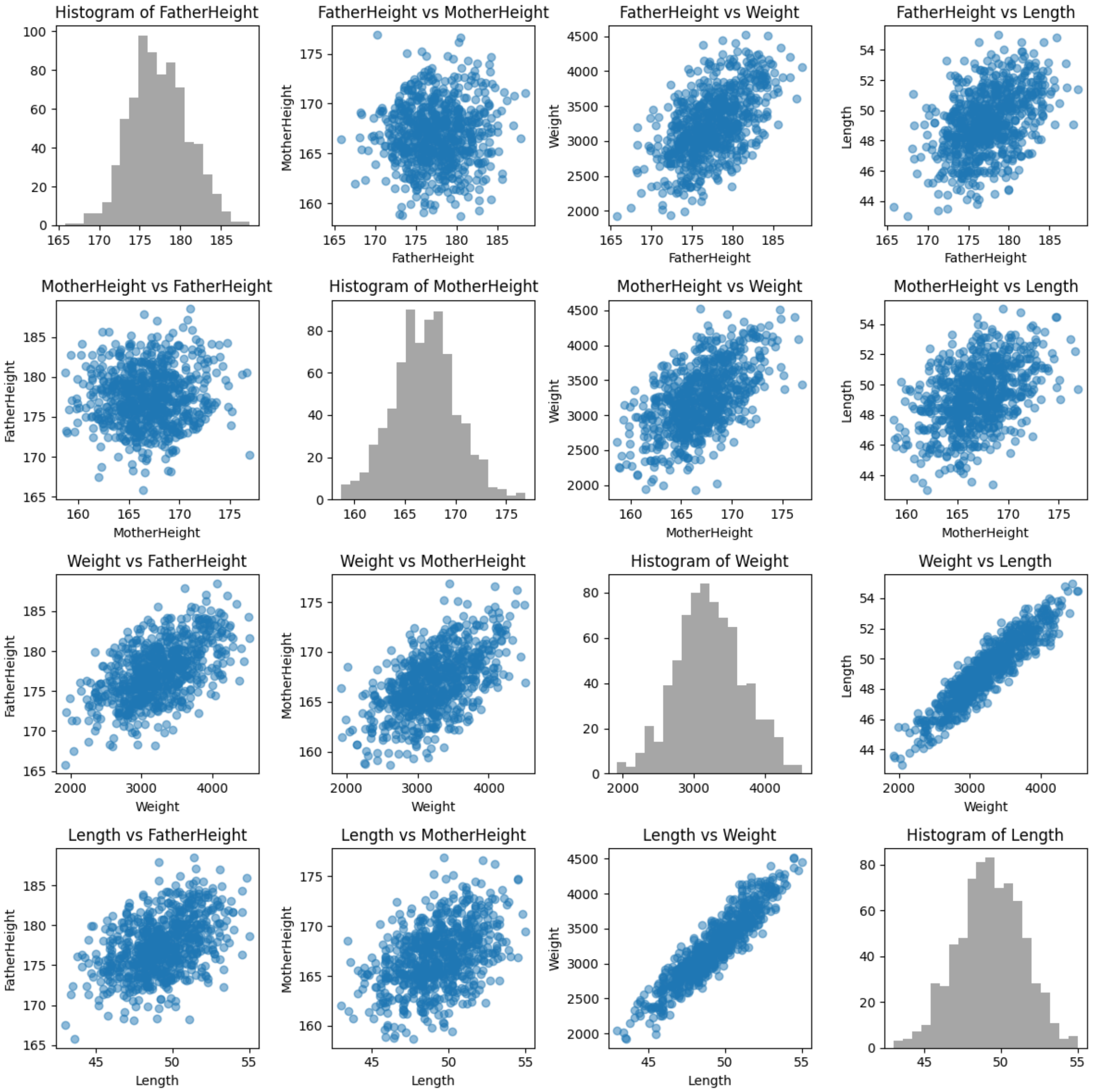

Figure 7: QQ-plots for the marginal distributions

QQ-plots and scatterplots enable us to verify that the data follows a multivariate normal distribution, for which the marginal distributions are known to be normal.

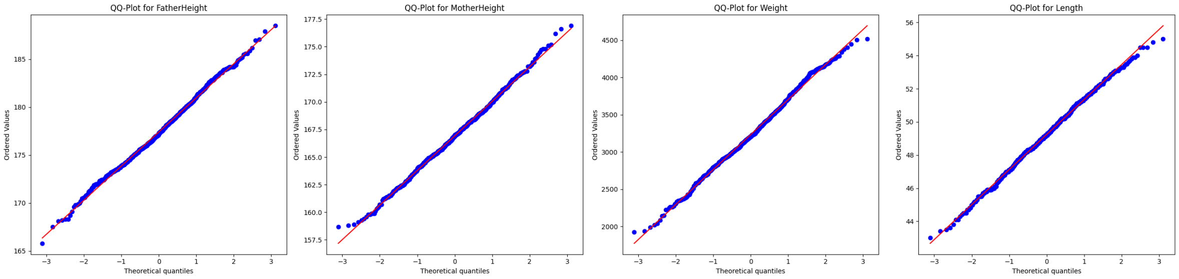

Figure 8: QQ-plots for the marginal distributions

3) Identify the conditional distribution of the weight and length of a child given the heights of parents. Find an estimate of the covariance matrix of the conditional distribution and compare it with the original unconditional covariance.

Conditional Distribution of Weight and Length Given Parent Heights:

Let X represent the heights of the father and mother, and Y represent the weight and length of the child. We aim to find the conditional distribution of Y given X. The original (unconditional) covariance matrix Σ_{YY} represents the covariance between weight and length without conditioning on the parents' heights. The conditional covariance matrix $\Sigma_{Y|X}$ is computed using the following formula:

$$\Sigma_{Y|X} = \Sigma_{YY} - \Sigma_{XY}\Sigma_{XX}^{-1}\Sigma_{XY}^{T}$$

Where: Σ_{XX} is the covariance matrix between the parents' heights, Σ_{XY} is the covariance between the parents' heights and the child's weight and length, Σ_{YY} is the covariance matrix for the child's weight and length.

Numerical Results:

The original (unconditional) covariance matrix is:

$$\Sigma_{YY} = \begin{bmatrix} 220276.658 & 915.295 \\ 915.295 & 4.443 \end{bmatrix}$$

The conditional covariance matrix given the parents' heights is:
Here is the matrix with the values rounded to three decimal places:

$$\Sigma_{Y|X} = \begin{bmatrix} 88857.995 & 456.846 \\ 456.846 & 2.844 \end{bmatrix}$$

The result shows that the conditional covariance is smaller than the unconditional covariance, which indicates that knowing the parents' heights reduces the variability of the child's weight and length.

4) How the ellipsoids based on the conditional distribution will look like?

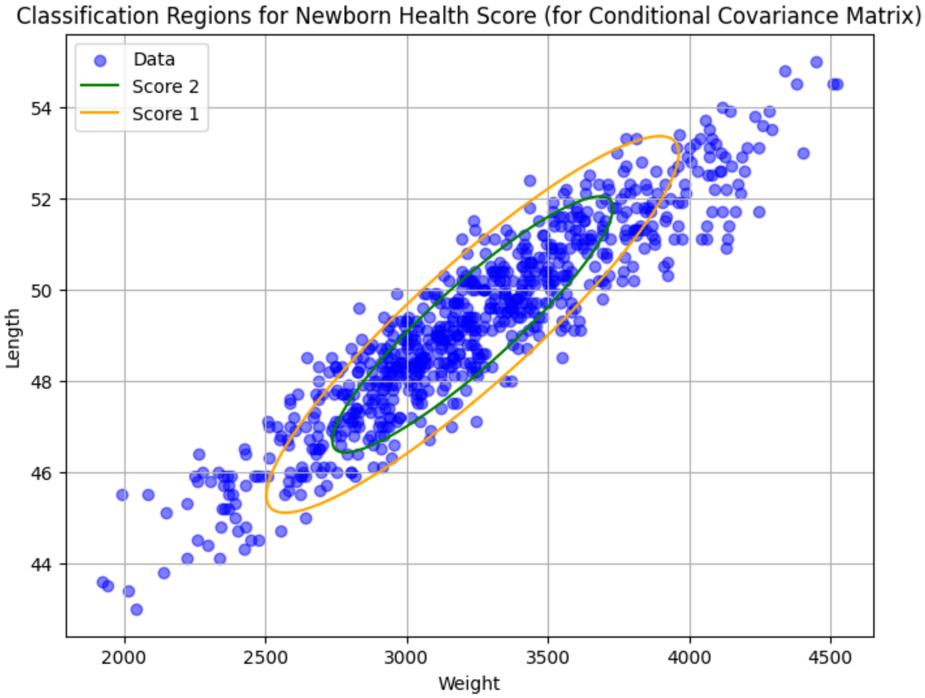

Figure 9: Ellipsoids based on conditional covariance matrix and original mean as classification regions for scores

Ellipses in Figure 9 are defined by the equations:

$$(x-\mu)^T\Sigma_{Y|X}^{-1}(x-\mu)=\chi_p^2$$

Where:

- x is a vector of points on the ellipsoid,
- μ is the original mean vector,
- $\Sigma_{Y|X}^{-1}$ is the inverse of the conditional covariance matrix,
- χ_p^2 is the quantile from the chi-squared distribution corresponding to the desired confidence level (0.75 and 0.9 for 75% or 90% confidence).

5)How many children would score zero, one, and two, respectively? Illustrate this classification on the graph and compare with the one obtained without considering the heights of parents.

Score	Count
0.0	40
1.0	137
2.0	559

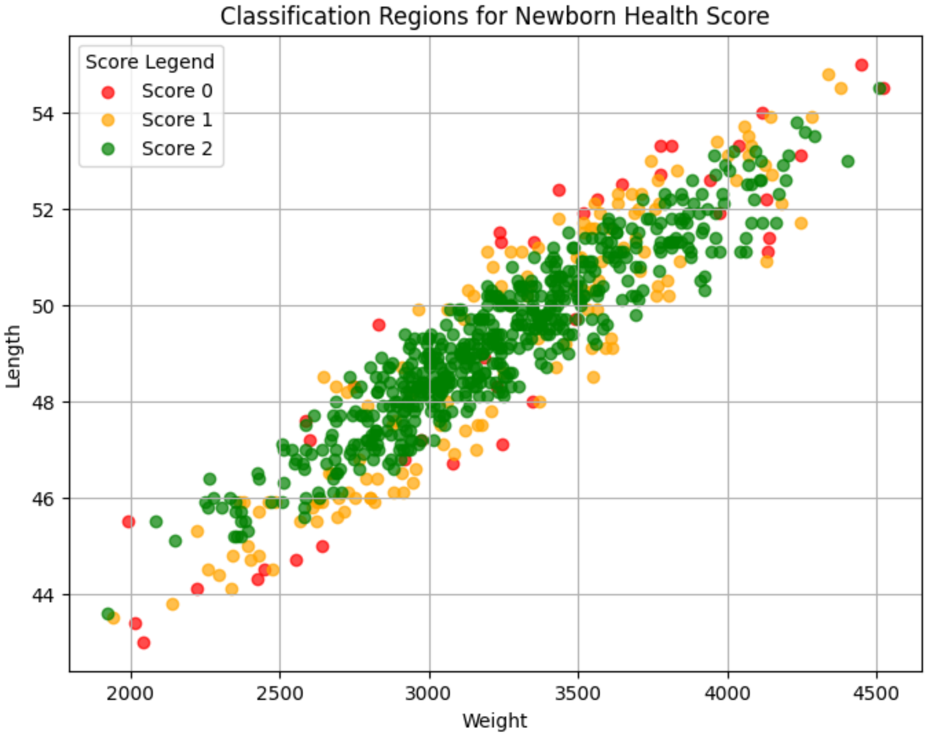

Figure 10: Ellipsoids based on conditional covariance matrix and conditional mean for every sample as classification regions.

Every data point in Figure 10 was classified with accordance to its own ellipse, which is defined with the equation:

$$(y - \mu_{Y|X})^T \Sigma_{Y|X}^{-1} (y - \mu_{Y|X}) = \chi_p^2$$

Where:

- $\mu_{Y|X}$ is the conditional mean vector (center of the ellipsoid), and is given by:

  $$\mu_{Y|X} = \mu_Y + \Sigma_{XY} \Sigma_{XX}^{-1} (X - \mu_X)$$

  where:

  - X is a data point (heights of the father and mother),
  - μ_Y is the mean of Y,
  - μ_X is the mean of X,
  - Σ_{XY} is the covariance between X and Y,
  - Σ_{XX} is the covariance of X.

- $\Sigma_{Y|X}^{-1}$ is the inverse of the conditional covariance matrix,
- χ_p^2 is the quantile from the chi-squared distribution corresponding to the desired confidence level (e.g., 0.75 or 0.9 for 75% or 90% confidence).

Classification is based on ellipses created for conditional covariance matrix and conditional mean computed for every sample. If we were to hypothesize about the reason for the change in classification while using conditional distribution, we can say that when we consider the weight and height of the parents, more children get good scores, so they are classified as healthy. This makes sense, because genetics related to height have an impact on the child's length. For example, a very short child shouldn't receive a bad score if their parents' heights are significantly below average. However, this observation may not fully explain the underlying factors influencing the classification, and further analysis would be needed to confirm any causal relationships.

6)Suppose that the father of a child is 185[cm] tall and mother is 178[cm] tall. Plot the classification ellipsoids for their child.

To draw ellipses based on conditional distribution, we will use conditional covariance matrix computed above, and we will compute conditional mean based on the formula:

$$\mu_{Y|X} = \mu_Y + \Sigma_{XY}\Sigma_{XX}^{-1}(X - \mu_X)$$

which becomes:

$$\mu_{Y|X} = [4651.537, 54.168]$$

when we assume X=[185, 178].

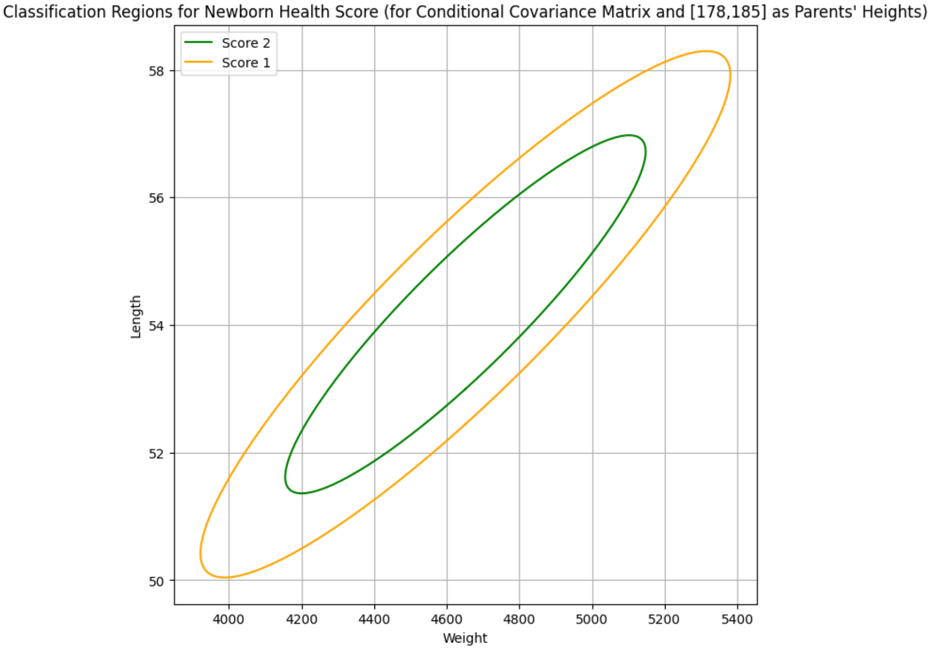

Figure 11: Ellipsoids based on conditional covariance matrix as classification regions for scores when we assume X=[185, 178].

We can observe that mother's and father's heights in this case are higher than the average, so the centre of the ellipse (mean) for the child's length and weight is also shifted towards higher values.

7) Find spectral decomposition of the estimated covariance matrix for the complete set of the data.

$$\Sigma = P\Lambda P^T$$

where

$$P = \begin{bmatrix} 0.122 & 0.564 & 0.816 & 0.004 \\ 0.149 & 0.803 & -0.577 & 0.004 \\ -0.005 & -0.005 & -0.001 & 1.000 \\ 0.981 & -0.193 & -0.014 & 0.004 \end{bmatrix}$$

and

$$\Lambda = \begin{bmatrix} 0.479 & 0 & 0 & 0 \\ 0 & 4.788 & 0 & 0 \\ 0 & 0 & 10.711 & 0 \\ 0 & 0 & 0 & 220287.510 \end{bmatrix}$$

8) Transform the data according to $P^T X$. Plot scatter plots of the transformed data.

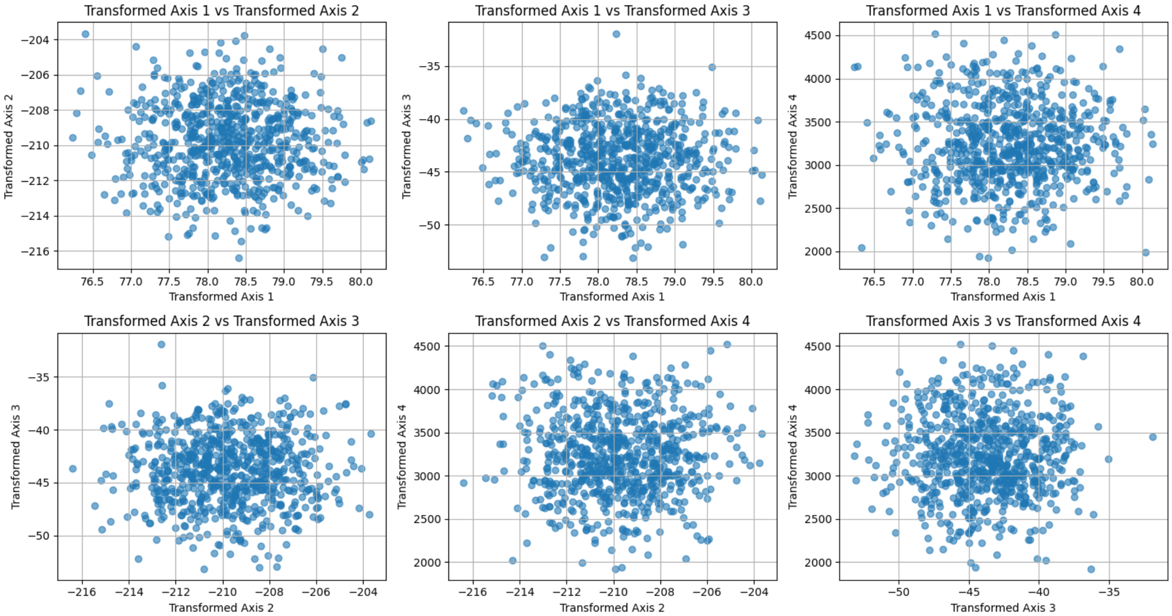

Figure 12: Data transformed according to $P^T X$

If $X \sim \mathcal{N}(\mu, \Sigma)$ and P is an orthogonal matrix containing eigenvectors of Σ, then the transformed variable $Y = P^T X$ follows $Y \sim \mathcal{N}(P^T \mu, P^T \Sigma P)$. Since $P^T \Sigma P = \Lambda$ is a diagonal matrix with eigenvalues of Σ, the components of Y are uncorrelated, and independent if Σ is full rank. The plots present that data is expressed in a new coordinate system where the axes (basis vectors) are the eigenvectors of the covariance matrix. These eigenvectors are orthogonal, meaning they form a new set of uncorrelated features.

A Python Code for Ellipse Plotting and scores counting - Project 1

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats

np.random.seed(42)
mean = df.mean()
cov = df.cov()

def mahalanobis_distance(x, mean, cov_inv):
    diff = x - mean
    return np.sqrt(diff @ cov_inv @ diff.T)

cov_inv = np.linalg.inv(cov)
chi2_75, chi2_95 = stats.chi2.ppf([0.75, 0.95], df=2)

scores = []
for i in range(len(df)):
    d2 = mahalanobis_distance(df.iloc[i].values, mean, cov_inv)
    if d2**2 <= chi2_75:
        scores.append(2)
    elif d2**2 <= chi2_95:
        scores.append(1)
    else:
        scores.append(0)

df["Score"] = scores
score_counts = df["Score"].value_counts().sort_index()
print(score_counts)
colors = {0: "red", 1: "orange", 2: "green"}
fig, ax = plt.subplots(figsize=(8, 6))

for score, color in colors.items():
    subset = df[df["Score"] == score]
    ax.scatter(subset["Weight"], subset["Length"], label=f"Score {score}", color=color, alpha=0.6)

def plot_ellipses(mean, cov, percentiles, ax):
    theta = np.linspace(0, 2 * np.pi, 100)
    eigvals, eigvecs = np.linalg.eigh(cov)
    ax_lengths = np.sqrt(eigvals)

    for p, label, color in zip(percentiles, ["Score 2", "Score 1"], ["green", "orange"]):
        d = np.sqrt(stats.chi2.ppf(p, df=2))
        ellipse = np.array([ax_lengths[0] * np.cos(theta), ax_lengths[1] * np.sin(theta)])
        rotated_ellipse = eigvecs @ ellipse
        ax.plot(mean[0] + d * rotated_ellipse[0], mean[1] + d * rotated_ellipse[1], label=label, colo

plot_ellipses(mean, cov, [0.75, 0.95], ax)
ax.set_title("Classification Regions for Newborn Health Score")
ax.set_xlabel("Weight")
ax.set_ylabel("Length")
ax.legend()
ax.grid(True)
plt.show()
```

A Python Code for Ellipse Plotting and scores counting - Project 2

```
import numpy as np
import pandas as pd
import scipy.stats as stats
from scipy.spatial.distance import cdist
import matplotlib.pyplot as plt

np.random.seed(42)

cov_matrix = df[[’Weight’, ’Length’, ’FatherHeight’, ’MotherHeight’]].cov()

mu_X = df[[’FatherHeight’, ’MotherHeight’]].mean().values
mu_Y = df[[’Weight’, ’Length’]].mean().values

Sigma_XX = cov_matrix[[’FatherHeight’, ’MotherHeight’]].loc[[’FatherHeight’, ’MotherHeight’]].values
Sigma_YY = cov_matrix[[’Weight’, ’Length’]].loc[[’Weight’, ’Length’]].values
Sigma_XY = cov_matrix[[’FatherHeight’, ’MotherHeight’]].loc[[’Weight’, ’Length’]].values
```

```
Sigma_XX_inv = np.linalg.inv(Sigma_XX)
Sigma_Y_given_X = Sigma_YY - np.dot(Sigma_XY, np.dot(Sigma_XX_inv, Sigma_XY.T))

def compute_conditional_mean(row):
    X = row[[’FatherHeight’, ’MotherHeight’]].values
    conditional_mean = mu_Y + np.dot(Sigma_XY, np.dot(Sigma_XX_inv, (X - mu_X)))
    return conditional_mean

def mahalanobis_distance(df, cov):
    inv_cov = np.linalg.inv(cov)
    distances = []
    for _, row in df.iterrows():
        mean = compute_conditional_mean(row)
        x = row[[’Weight’, ’Length’]].values
        diff = x - mean
        d = np.sqrt(diff @ inv_cov @ diff.T)
        distances.append(d)
    return np.array(distances)

def assign_scores(df, cov):
    distances = mahalanobis_distance(df, cov)

    threshold_75 = np.sqrt(stats.chi2.ppf(0.75, df=2))
    threshold_95 = np.sqrt(stats.chi2.ppf(0.95, df=2))

    scores = np.zeros(len(df))
    scores[distances <= threshold_75] = 2
    scores[(distances > threshold_75) & (distances <= threshold_95)] = 1
    scores[distances > threshold_95] = 0

    df[’Score’] = scores
    return df

df_with_scores = assign_scores(df, Sigma_Y_given_X)

print(df_with_scores[[’Weight’, ’Length’, ’Score’]])

score_counts = df_with_scores[’Score’].value_counts()
print("\nScore counts:")
print(score_counts)

colors = {0: "red", 1: "orange", 2: "green"}

fig, ax = plt.subplots(figsize=(8, 6))
for score, color in colors.items():
    subset = df_with_scores[df_with_scores["Score"] == score]
    ax.scatter(subset["Weight"], subset["Length"], label=f"Score {score}", color=color, alpha=0.7)

ax.set_title("Classification Regions for Newborn Health Score")
ax.set_xlabel("Weight")
ax.set_ylabel("Length")
ax.legend(title="Score Legend")
ax.grid(True)
plt.show()
```