Statistical Learning - Report 2:
Multivariate means and multiple testing

Julia Kiczka

April 11, 2025

Problem 1

1. Chi-square distribution and F distribution

Let $Z_1, Z_2, \ldots, Z_p$ be independent standard normal random variables, i.e., $Z_i \sim \mathcal{N}(0,1)$. Then, the chi-squared distributed random variable with p degrees of freedom is defined as:

$$X = \sum_{i=1}^{p} Z_i^2 \sim \chi^2_p.$$

Similarly, let $X_1 \sim \chi^2_{d_1}$ and $X_2 \sim \chi^2_{d_2}$ be independent chi-squared random variables with d_1 and d_2 degrees of freedom, respectively. The F-distributed random variable with d_1 and d_2 degrees of freedom is given by:

$$F = \frac{(X_1/d_1)}{(X_2/d_2)} \sim F_{d_1,d_2}.$$

2. What distribution will $F_{p,n-p}$ approximately follow for $p = 4$ and $n = 1000$?

We want to investigate the case where n is much larger than p. To do this, we look closely at the F distribution when d_2 goes to ∞.

$$F_{p,n-p} \sim \frac{\chi^2_{(p)}/p}{\chi^2_{(n-p)}/(n-p)}.$$

By the Law of Large Numbers (LLN), we have:

$$\frac{\chi^2_{(n-p)}}{n-p} \to E(\chi^2_{(1)}) = 1 \quad \text{as} \quad n \to \infty.$$

Thus, applying Slutsky's theorem, since $\chi^2_{(n-p)}/(n-p) \to 1$ in probability, we treat it as a constant, leading to:

$$\frac{\chi^2_{(p)}/p}{\chi^2_{(n-p)}/(n-p)} \to \frac{\chi^2_{(p)}}{p} \quad \text{as} \quad n \to \infty.$$

For $p = 4$ and $n = 1000$, $F_{p,n-p}$ will follow approximately $\frac{\chi^2_{(4)}}{4}$.

3. Prove that $n(\bar{X} - \mu)^T \Sigma^{-1} (\bar{X} - \mu) \sim \chi^2_p$

We are given that

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}_p(\mu, \Sigma),$$

which means each X_i follows a p-dimensional multivariate normal distribution with mean vector $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. We know that for every positive-definite matrix Σ there exist one unique root of this matrix, namely $\Sigma^{\frac{1}{2}}$, such that $\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}=\Sigma$ and $\Sigma^{\frac{1}{2}}$ is also positive-definite. The sample mean is defined as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Since the X_i are normally distributed, their linear combination $\bar{X}$ is also normally distributed:

$$\bar{X} \sim \mathcal{N}_p\left(\mu, \frac{1}{n}\Sigma\right).$$

Let's define the transformed vector:

$$Z = \Sigma^{-1/2}(\bar{X} - \mu).$$

Since

$$\bar{X} - \mu \sim \mathcal{N}_p\left(0, \frac{1}{n}\Sigma\right),$$

applying a linear transformation to a multivariate normal variable follows the rule:

$$AX \sim \mathcal{N}_p(A\mu, A\Sigma A^T).$$

Here, setting $A = \Sigma^{-1/2}$, we get:

$$Z = \Sigma^{-1/2}(\bar{X} - \mu) \sim \mathcal{N}_p\left(0, \Sigma^{-1/2} \cdot \frac{1}{n}\Sigma \cdot \Sigma^{-1/2}\right).$$

Since

$$\Sigma^{-1/2}\Sigma\Sigma^{-1/2} = \Sigma^{-1/2}(\Sigma^{1/2}\Sigma^{1/2})\Sigma^{-1/2} = (\Sigma^{-1/2}\Sigma^{1/2})\Sigma^{1/2}\Sigma^{-1/2} = II = I.$$

we obtain:

$$Z \sim \mathcal{N}_p\left(0, \frac{1}{n}I\right).$$

Consider the quadratic form, substituting Z:

$$n(\bar{X} - \mu)^T\Sigma^{-1}(\bar{X} - \mu) = n(\Sigma^{1/2}Z)^T\Sigma^{-1}(\Sigma^{1/2}Z)$$

Next, we simplify the expression using the property of transposition $(AB)^T = B^TA^T$:

$$n(\Sigma^{1/2}Z)^T\Sigma^{-1}(\Sigma^{1/2}Z) = nZ^T(\Sigma^{1/2})^T\Sigma^{-1}\Sigma^{1/2}Z$$

Since $\Sigma^{1/2}$ is symmetric, $(\Sigma^{1/2})^T = \Sigma^{1/2}$, and $\Sigma^{-1}\Sigma^{1/2} = \Sigma^{-1/2}$, we have:

$$nZ^T\Sigma^{1/2}\Sigma^{-1/2}Z = nZ^TZ$$

Thus, the quadratic form simplifies to:

$$n(\bar{X} - \mu)^T\Sigma^{-1}(\bar{X} - \mu) = nZ^TZ$$

Since $Z \sim \mathcal{N}_p(0, \frac{1}{n}I)$, we rewrite Z as:

$$Z = \frac{1}{\sqrt{n}}W, \quad \text{where} \quad W \sim \mathcal{N}_p(0, I).$$

Thus,

$$Z^TZ = \left(\frac{1}{\sqrt{n}}W\right)^T\left(\frac{1}{\sqrt{n}}W\right) = \frac{1}{n}W^TW.$$

Multiplying by n:

$$nZ^TZ = W^TW.$$

Since $W \sim \mathcal{N}_p(0, I)$, we conclude that

$$W^TW \sim \chi^2_p.$$

4. Let $X_1, \dots, X_n \sim \mathcal{N}_p(\mu, \Sigma)$. Assume we do not know either μ or Σ. We want to test the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Recall the Hotelling T^2 statistic:

$$T^2 := n(\bar{X} - \mu_0)^T S^{-1}(\bar{X} - \mu_0),$$

where S denotes the sample covariance matrix.

(a) Assume H_0 is true. What distribution does T^2 follow? Use this to write down a test, which rejects at level α.

(b) Explain what happens to T^2 if we gather more observations and n goes to infinity. What happens to the probability of rejecting H_0?

(c) Assume H_0 is false. Explain what happens to T^2 as n goes to infinity. What happens to the probability of rejecting H_0?

a)

When the null hypothesis is true, then

$$\frac{n-p}{(n-1)p} T^2$$

is distributed as $F_{p,n-p}$. What is more, large values of T^2 lead to rejection of the null hypothesis. It follows that we reject H_0 at level α if:

$$\frac{n-p}{(n-1)p} T^2 \geq q_{F_{p,n-p}}(1-\alpha),$$

where $q_{F_{p,n-p}(1-\alpha)}$ is the $1-\alpha$ quantile of $F_{p,n-p}$.

b)

$$T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p}.$$

We know that:

$$\frac{(n-1)p}{n-p} \to p \text{ as } n \to \infty$$

From previous points:

$$F_{p,n-p} \to \frac{\chi^2_p}{p}$$

Which leads to (Slutsky's theorem):

$$T^2 \to p \cdot \frac{\chi^2_p}{p} = \chi^2_p \text{ as } n \to \infty.$$

This asymptotic result implies that the probability of rejecting H_0 is:

$$P_{H_0}\left(T^2 \geq q_{\chi^2_p(1-\alpha)}\right) \to \alpha \quad \text{as } n \to \infty.$$

where $q_{\chi^2_p(1-\alpha)}$ is the $1-\alpha$ quantile of χ^2 distribution .

c)

Let us assume that H_1 is true, then the power of the above test, under $n \to \infty$ is:

$$P_{H_1}\left(T^2 \geq q_{\chi^2_p(1-\alpha)}\right).$$

Under H_1, T^2 has the noncentral χ^2_p distribution with noncentrality parameter given by:

$$\lambda = n(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0).$$

Since $(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0)$ is a fixed positive quantity, it follows that as $n \to \infty$, we obtain $\lambda \to \infty$. A key result in probability theory states that for large λ, the noncentral chi-squared distribution can be approximated as:

$$\chi^2_p(\lambda) \approx N(p + \lambda, 2(p + 2\lambda)).$$

Thus, the expected value of T^2 under H_1 is:

$$E[T^2 \mid H_1] = p + \lambda = p + n(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0).$$

Since the critical value $q_{\chi^2_p(1-\alpha)}$ remains fixed, as $n \to \infty$, the mean of T^2 increases indefinitely, shifting the entire distribution to the right. Consequently, the probability that T^2 exceeds $q_{\chi^2_p(1-\alpha)}$ approaches 1:

$$P(T^2 \geq q_{\chi^2_p(1-\alpha)} \mid H_1) \to 1.$$

Thus, the power of the test approaches 1 as $n \to \infty$.

Problem 2

Random vector

$$X = (1.7, 1.6, 3.3, 2.7, -0.04, 0.35, -0.5, 1.0, 0.7, 0.8)$$

comes from the 10-dimensional multivariate normal distribution $N(\mu, I)$.

1. Which hypotheses would be rejected by the Bonferroni multiple testing procedure?
2. Which hypotheses would be rejected by the BH multiple testing procedure?
3. Assume that only the first three coordinates of μ are different from zero. What is the False Discovery Proportion of the Bonferroni and the BH procedures?

Computing p-values

Each X_i follows $N(\mu_i, 1)$, so under H_0:

$$X_i \sim N(0, 1).$$

The p-value for testing H_{0i} is:

$$p_i = 2 \cdot (1 - \Phi(|X_i|)),$$

where Φ is the standard normal cumulative distribution function. Using a normal table:

$$\Phi(1.7) \approx 0.9554 \Rightarrow p_1 = 2(1 - 0.9554) = 0.0892$$

$$\Phi(1.6) \approx 0.9452 \Rightarrow p_2 = 0.1096$$

$$\Phi(3.3) \approx 0.9990 \Rightarrow p_3 = 0.001$$

$$\Phi(2.7) \approx 0.9965 \Rightarrow p_4 = 0.007$$

Remaining values give larger p-values.

1) Bonferroni Correction

Bonferroni correction controls FWER by setting the threshold:

$$\alpha_B = \frac{\alpha}{p}.$$

For $\alpha = 0.05$,

$$\alpha_B = \frac{0.05}{10} = 0.005.$$

We reject H_{0i} if $p_i < 0.005$.
Only $p_3 = 0.001$ is below this threshold. So, only H_{03} is rejected.

2) Benjamini-Hochberg (BH) Procedure

The BH procedure controls FDR. It sorts p-values in ascending order and finds the largest k such that:

$$p_{(k)} \leq \frac{k}{p}\alpha.$$

After comparing sorted p-values with thresholds, we conclude that only two hypotheses (H_{03} and H_{04}) are rejected.

3) False Discovery Proportion

We assume only the first three coordinates (μ_1, μ_2, μ_3) are nonzero.

- **Bonferroni:** Rejected only H_{03} → No false discoveries (FDP = 0).
- **BH:** Rejected H_{03} and H_{04}.
  - H_{03} is a true rejection.
  - H_{04} is a false rejection.

$$\text{FDP} = \frac{\text{false rejections}}{\text{total rejections}} = \frac{1}{2} = 0.5.$$

Simulation

Consider the sequence of independent random variables $X_1, \ldots, X_p$ such that $X_i \sim N(\mu_i, 1)$ and the problem of multiple testing the hypotheses $H_{0i} : \mu_i = 0$, for $i \in \{1, \ldots, p\}$. For $p = 5000$ and $\alpha = 0.05$, we use simulations (at least 1000 replicates) to estimate FWER, FDR, and power of the Bonferroni and Benjamini-Hochberg multiple testing procedures under the following setups:

- **Setup A**: $\mu_1 = \cdots = \mu_{10} = \sqrt{2 \log p}$, $\mu_{11} = \cdots = \mu_p = 0$
- **Setup B**: $\mu_1 = \cdots = \mu_{500} = \sqrt{2 \log p}$, $\mu_{501} = \cdots = \mu_p = 0$

Simulation and Results

The simulation was conducted by generating 5000 normally distributed test statistics under two different setups, applying the Bonferroni and Benjamini-Hochberg multiple testing procedures, and computing the Family-Wise Error Rate (FWER), False Discovery Rate (FDR), and statistical power. Each setup was repeated over 10 iterations, and the average values are reported in the table below.

	FWER Bonf	FDR Bonf	Power Bonf	FWER BH	FDR BH	Power BH
Setup A	0.047	0.011	0.386	0.279	0.051	0.546
Setup B	0.042	0.000	0.386	1.000	0.045	0.903

Table 1: Results of the multiple testing simulations averaged over 10 iterations.

Bonferroni correction is a conservative approach that controls the Family-Wise Error Rate (FWER), which is the probability of making one or more false discoveries across multiple tests. This method adjusts the significance threshold by dividing the alpha value (significance level) by the number of tests. As a result, it tends to be very stringent, significantly reducing the likelihood of false positives but at the cost of reduced statistical power. Consequently, it is most suitable when the primary goal is to minimize false positives and maintain a high degree of caution in declaring significant results.On the other hand, the Benjamini-Hochberg (BH) procedure controls the False Discovery Rate (FDR), which is the expected proportion of false positives among the rejected hypotheses. Compared to Bonferroni, the BH method is less strict, allowing for a greater number of rejections while still controlling the rate of false discoveries. This results in higher power, meaning that the test is more likely to detect true effects when they exist. The BH method is particularly useful when the objective is to identify more true signals without excessively penalizing for false positives.

Project: Printing Bank Notes

The Swiss banknote dataset consists of 100 measurements on genuine banknotes. The dataset includes six numerical features related to the physical dimensions of the banknotes. The objective of this study is to analyze these measurements and assess whether they follow a multivariate normal distribution.

Problem Formulation

Given a dataset $X = \{X_1, X_2, X_3, X_4, X_5, X_6\}$ where:

- X_1 - Length of the bill
- X_2 - Height of the bill (left)
- X_3 - Height of the bill (right)
- X_4 - Distance of the inner frame to the lower border
- X_5 - Distance of the inner frame to the upper border
- X_6 - Length of the diagonal of the central picture

1) Load the data, produce scatter plots and qq-plots of the data and discuss validity of the assumption that the data are from a multivariate normal distribution.

QQ plots display points aligning closely along a straight line, indicating that the data is likely normally distributed, with minimal outliers. Likewise, scatterplots reveal predominantly elliptical and bell-shaped patterns.

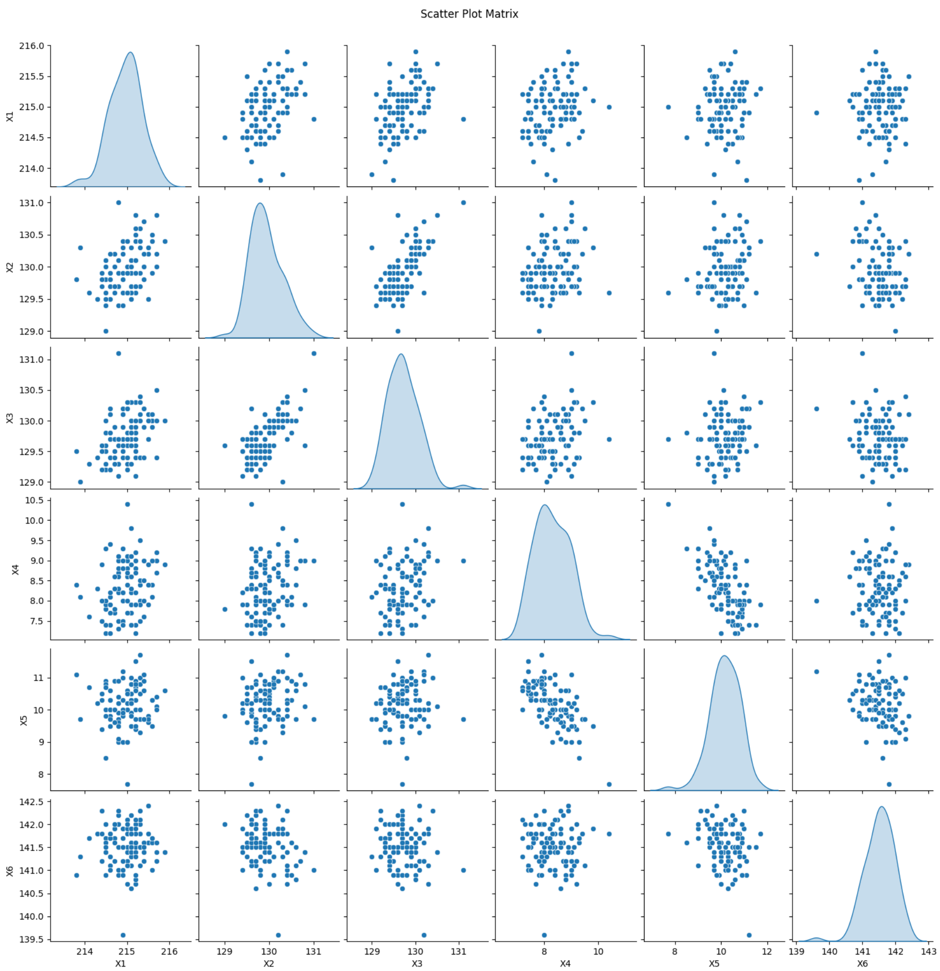

Figure 1: Scatterplot for the data.

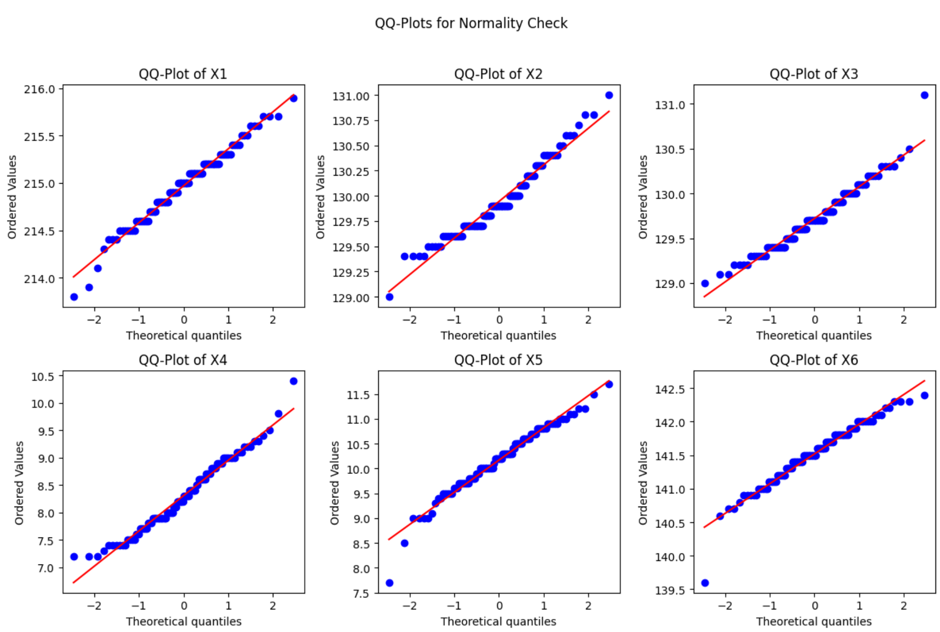

Figure 2: QQ-plots for the data.

2) Evaluate estimators of the vector of means and the covariance matrix.

The estimated mean vector is given by:

$$\mu = \begin{bmatrix} 214.969 \\ 129.943 \\ 129.720 \\ 8.305 \\ 10.168 \\ 141.517 \end{bmatrix}$$

The estimated covariance matrix is:

$$\Sigma = \begin{bmatrix} 0.150 & 0.058 & 0.057 & 0.057 & 0.014 & 0.005 \\ 0.058 & 0.133 & 0.086 & 0.057 & 0.049 & -0.043 \\ 0.057 & 0.086 & 0.126 & 0.058 & 0.031 & -0.024 \\ 0.057 & 0.057 & 0.058 & 0.413 & -0.263 & -0.000 \\ 0.014 & 0.049 & 0.031 & -0.263 & 0.421 & -0.075 \\ 0.005 & -0.043 & -0.024 & -0.000 & -0.075 & 0.200 \end{bmatrix}$$

3) Write an Python function that is verifying if a point lies inside of the six dimensional ellipsoid that serve as the 95% confidence region for the mean value of bank notes based on the Hotelling's T^2 statistics

```python
import numpy as np
import scipy.stats as stats

def is_inside_confidence_ellipsoid(point, mean, cov, n_samples, confidence=0.95):
    """
    Checks if a given point lies inside the 95% confidence ellipsoid
    based on Hotelling's T-squared statistic.

    Parameters:
    point (array-like): The 6D point to be checked.
    mean (array-like): The 6D mean vector of the distribution.
    cov (array-like): The 6x6 covariance matrix.
    n_samples (int): The number of samples used for estimation.
    confidence (float): Confidence level (default is 0.95).

    Returns:
    bool: True if the point is inside the confidence ellipsoid, False otherwise.
    """
    point = np.array(point)
    mean = np.array(mean)
    cov = np.array(cov)

    # Compute inverse of covariance matrix
    cov_inv = np.linalg.inv(cov)

    # Compute the squared Mahalanobis distance (Hotelling's T-squared statistic)
    diff = point - mean
    T2_stat = diff.T @ cov_inv @ diff *  n_samples

    # Compute critical value from F-distribution
    p = len(mean)  # Dimensionality (should be 6)
    F_critical = stats.f.ppf(confidence, p, n_samples - p) * (p * (n_samples - 1)) / (
        n_samples - p)

    return T2_stat <= F_critical
```

4) Check if the obtained mean values are within the Hotelling's confidence region that was obtained based on the original sample of bank notes.

LENGTH	LEFT	RIGHT	BOTTOM	TOP	DIAGONAL
214.97	130.00	129.67	8.30	10.16	141.52

Table 2: Measurements of the banknote

To determine whether the newly obtained mean vector $\boldsymbol{\mu}_{\text{new}}$ lies within the 95% confidence region constructed from the original sample, we use Hotelling's T^2 statistic:

$$T^2 = (\boldsymbol{\mu}_{\text{new}} - \boldsymbol{\mu}_{\text{orig}})^T \Sigma^{-1} (\boldsymbol{\mu}_{\text{new}} - \boldsymbol{\mu}_{\text{orig}}) \cdot n$$

where:

- $\boldsymbol{\mu}_{\text{new}}$ is the newly obtained mean vector.
- $\boldsymbol{\mu}_{\text{orig}}$ is the original mean vector estimated from the sample.
- Σ is the sample covariance matrix.
- n is the number of observations in the original sample.

The critical value is computed using the F-distribution:

$$F_{\text{crit}} = F_{p,n-p,\alpha} \cdot \frac{p(n-1)}{n-p}$$

where:

- $p = 6$ (dimensionality of the data).
- $\alpha = 0.95$ (for a 95% confidence level).

From our calculations:

$$T^2 = 13.91, \quad F_{\text{crit}} = 13.88$$

Since $T^2 > F_{\text{crit}}$, the new mean vector $\boldsymbol{\mu}_{\text{new}}$ lies **outside** the confidence ellipsoid, indicating a statistically significant deviation from the original sample mean.

5) Check if the new mean vector falls within the Bonferroni's confidence rectangular region for the mean value of the old bank note dimensions.

Bonferroni's confidence rectangular region is defined by constructing individual confidence intervals for each component μ_i of the mean vector. The intervals are given by

$$\mu_i \in \left[\bar{x}_i - t_{\alpha/(2p),n-1} \cdot \frac{s_i}{\sqrt{n}}, \quad \bar{x}_i + t_{\alpha/(2p),n-1} \cdot \frac{s_i}{\sqrt{n}} \right]$$

where $\bar{x}_i$ is the sample mean, s_i is the sample standard deviation, n is the sample size, and $t_{\alpha/(2p),n-1}$ is the Bonferroni-adjusted critical value from the t-distribution. Each mean component must fall within its respective interval for the new mean vector to be inside the confidence region. Since all components of the new mean satisfy this condition, we conclude that the new mean lies within the Bonferroni confidence rectangular region. The new mean vector lies outside Hotelling's confidence ellipsoid but inside Bonferroni's confidence rectangular region. This suggests that while the new mean shows a statistically significant deviation when considering the joint correlation structure of the variables (as in Hotelling's test), it does not exceed the bounds when each variable is considered independently (as in Bonferroni's method). Hotelling's test is more sensitive because it accounts for correlations between dimensions, while Bonferroni's method is more conservative due to its multiple comparisons adjustment. The result highlights that deviations in individual dimensions may not be large enough to be flagged independently but are collectively significant when considered together.

6) Plot the projection of both confidence regions to the one-dimensional spaces marked by the axes: X_i for $i = 1, \ldots, 6$. Mark the projection of the vector of means on the obtained confidence intervals. Comment what you observed.

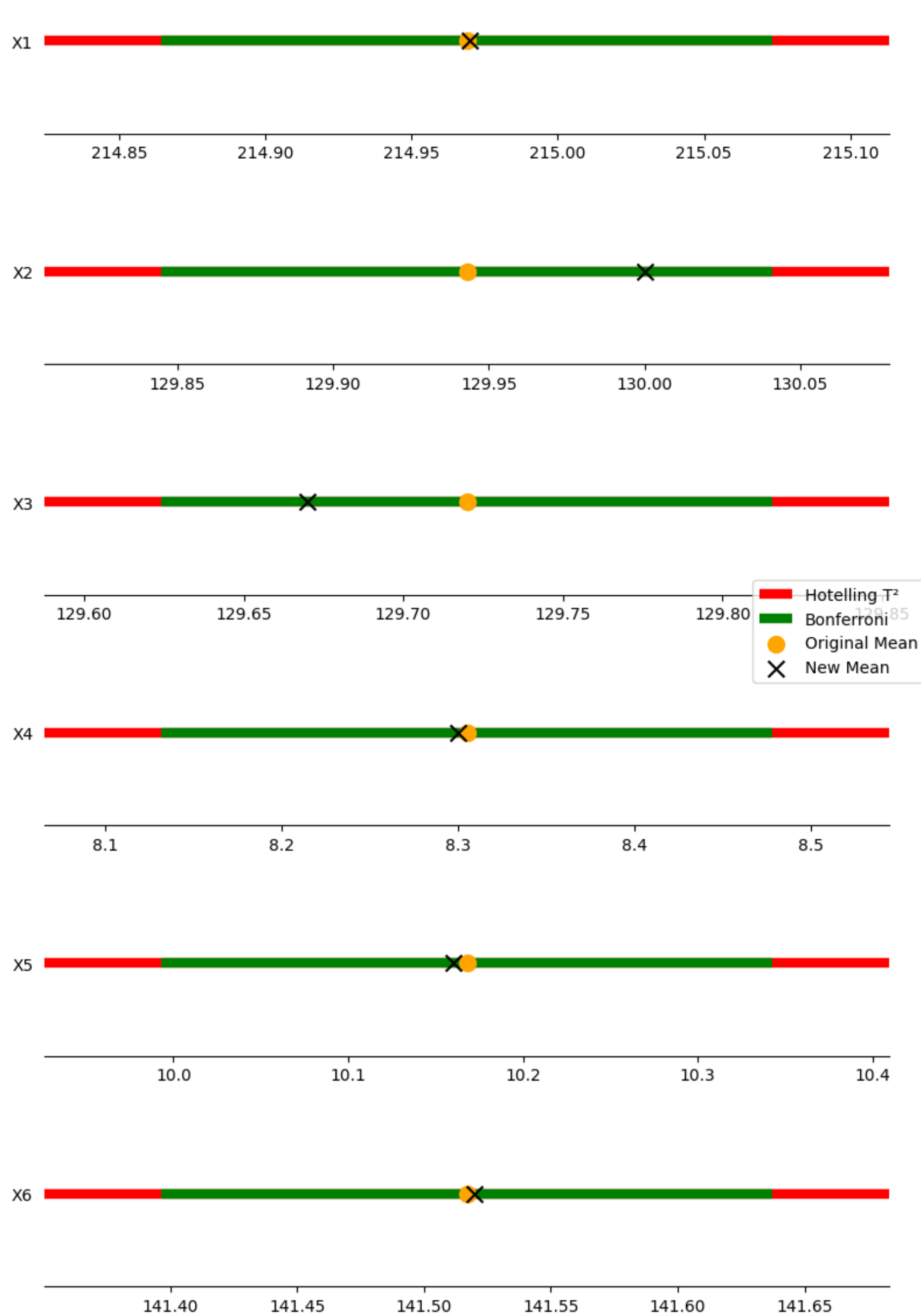

Figure 3: Projection of both confidence regions to the two-dimensional spaces marked by the pairs of axes.

The Bonferroni confidence intervals (green) are narrower than the Hotelling's T^2 region (red) since Bonferroni applies a stricter correction. We can observe that the new mean lies inside this one-dimensional projections of both types of confidence intervals.

7) Plot the projection of both confidence regions to the two-dimensional spaces marked by the pairs of axes: $X_i, X_j, i \neq j$. Mark the projection of the vector of means. Comment what you observed.

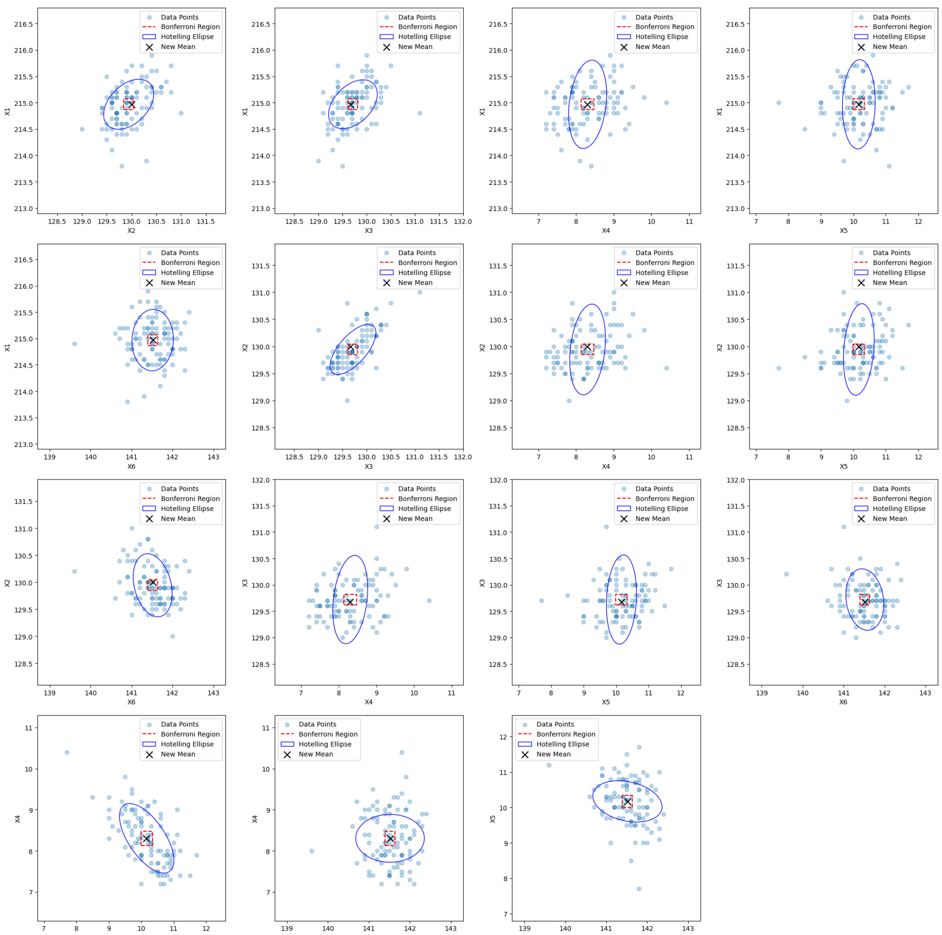

Figure 4: Projection of both confidence regions to the two-dimensional spaces marked by the pairs of axes.

As suggested during labs, quantiles needed to create the ellipses come from $F_\alpha(p, n-p)$ (not χ^2 distribution). The equation for an ellipse is defined as follows:

$$(x - \mu_{i,j})^T {\Sigma_{i,j}}^{-1} (x - \mu_{i,j}) = q_{F_\alpha(p,n-p)}$$

where $\Sigma_{i,j}$ and $\mu_{i,j}$ are sub-matrix and sub-vector corresponding to indices (i,j) of original covariance matrix and mean vector of the data. The Hotelling's confidence region (blue) forms ellipses in the two-dimensional spaces, accounting for correlation between the variables. The Bonferroni method (red), being more conservative, forms rectangular confidence regions due to the independent treatment of each dimension. We can observe that the new mean lies inside this two-dimensional projections of both types of confidence intervals.

8) Interpret geometrically the fact that the mean values of the bank note dimensions from the new production line fail to belong to the Hotelling's confidence region. Relate to the previously created graphs.

Hotelling's confidence region accounts for the joint distribution of all six dimensions, making it a stricter multivariate test. The new mean vector falls outside this region, indicating a statistically significant overall deviation from the original production. However, in 1D and 2D projections, the new means remain within confidence intervals and ellipses, as marginal deviations are not individually significant. This discrepancy arises because lower-dimensional projections ignore interdependencies, underestimating the total deviation. The result highlights that only a full multivariate analysis captures the accumulated differences across all dimensions.

9) Check if the vector of means are within Hotelling's confidence region and Bonferroni's confidence region. Comment your findings.

LENGTH	LEFT	RIGHT	BOTTOM	TOP	DIAGONAL
214.99	129.95	129.73	8.51	9.96	141.55

Table 3: Measurements of the banknote

This point lies inside Hotelling's confidence ellipsoid, since $T^2 = 12.649 < F_{\text{crit}} = 13.881$, but falls outside the Bonferroni confidence intervals:

$$\begin{aligned}\text{Bonferroni Lower Bounds:} \quad & [214.8646, 129.8450, 129.6243, 8.1319, 9.9933, 141.3967] \\ \text{Bonferroni Upper Bounds:} \quad & [215.0734, 130.0410, 129.8157, 8.4781, 10.3427, 141.6373]\end{aligned}$$

Hotelling's test suggests that the new mean is not significantly different in the multivariate space. However, the Bonferroni-adjusted confidence intervals detect deviations in individual dimensions (BOTTOM, TOP), indicating inconsistencies in these specific measurements. This implies that while the new banknotes generally conform to the original specifications, the (TOP, BOTTOM) dimensions may require further calibration to ensure full compliance.

10) Is this value acceptable based on the original sample of the bank notes, or the production line still needs some tuning? Explain your answer.

LENGTH	LEFT	RIGHT	BOTTOM	TOP	DIAGONAL
214.9473	129.9243	129.6709	8.3254	10.0389	141.4954

Table 4: Measurements of the banknote

This point lies inside Hotelling's confidence ellipsoid, since $T^2 = 8.640 < F_{\text{crit}} = 13.881$, and falls inside the Bonferroni confidence intervals:

$$\begin{aligned}\text{Bonferroni Lower Bounds:} \quad & [214.8646, 129.8450, 129.6243, 8.1319, 9.9933, 141.3967] \\ \text{Bonferroni Upper Bounds:} \quad & [215.0734, 130.0410, 129.8157, 8.4781, 10.3427, 141.6373]\end{aligned}$$

The latest results indicate that both tests now accept the new production line.