Homework 1

Jacob Thielemier

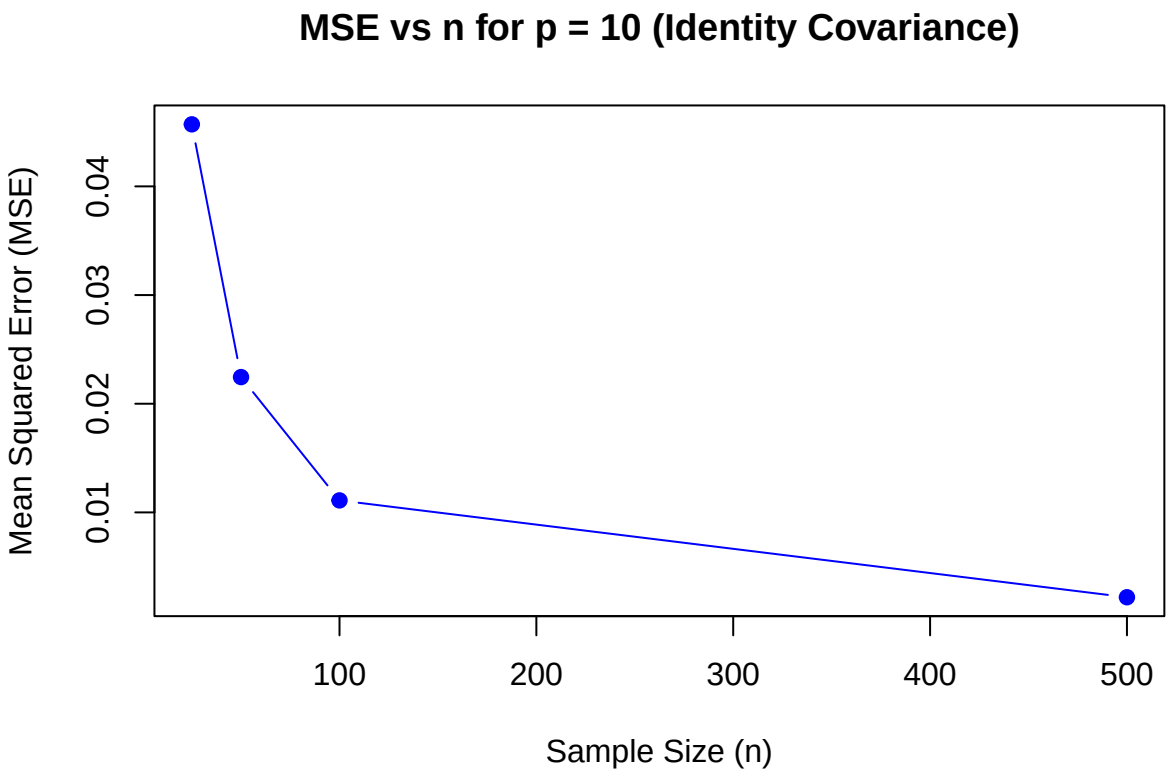
2025-02-15

Question 1

Part a.

Table 1: MSE for Different Sample Sizes (p = 10, Identity Covariance)

Sample.Size..n.	MSE
25	0.045708
50	0.022451
100	0.011105
500	0.002193

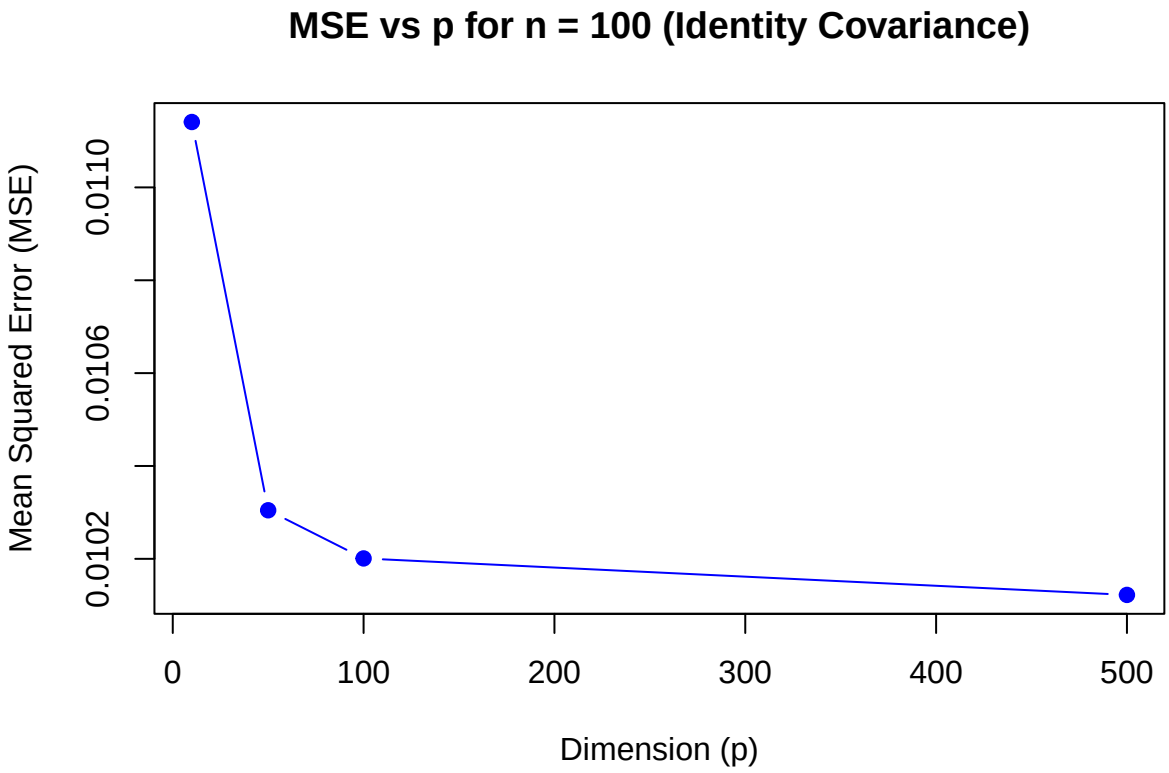


Part b.

Note: I had processing issues with so many iterations so I changed it to use 5000 iterations for p <= 100, and 500 for p > 100

Table 2: MSE for Different Dimensions (n = 100, Identity Covariance)

Dimension..p.	MSE
10	0.011141
50	0.010305
100	0.010201
500	0.010122



Part c.

Table 3: MSE for Different Sample Sizes (p = 10, AR(1) Covariance with rho = 1/sqrt(10))

Sample.Size..n.	MSE
25	0.046981
50	0.022952
100	0.011274

Sample.Size..n.	MSE
500	0.002227

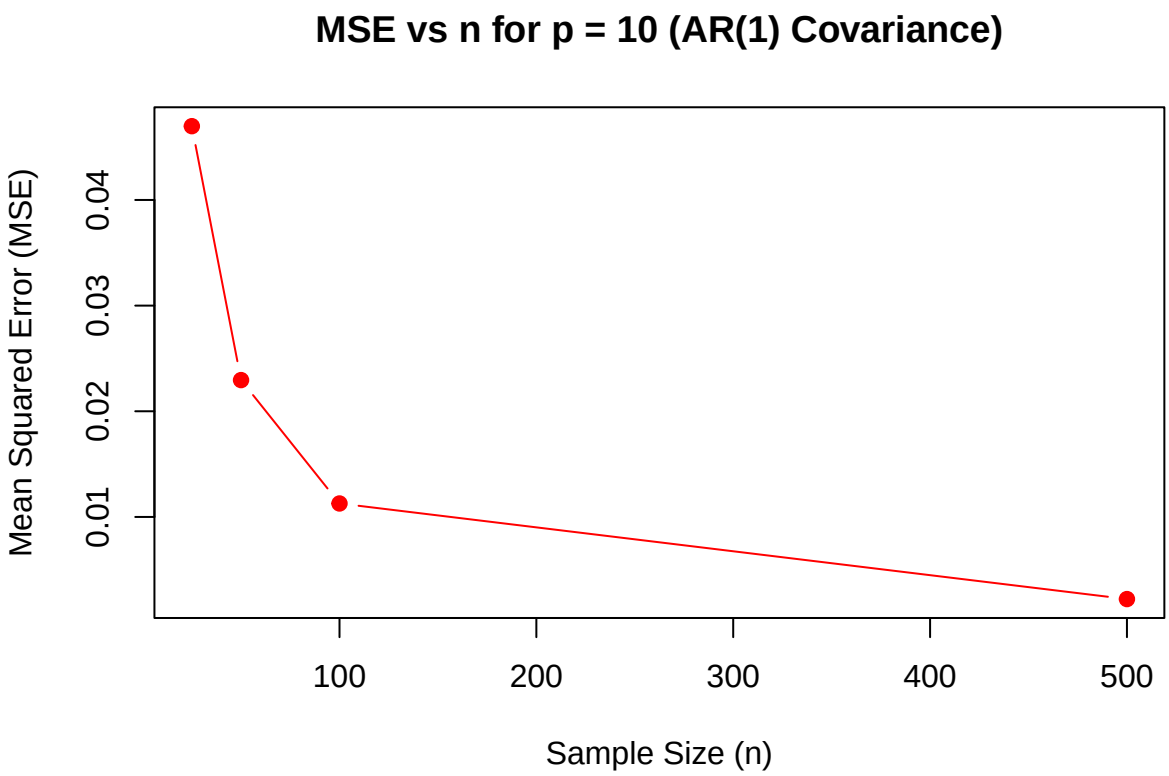
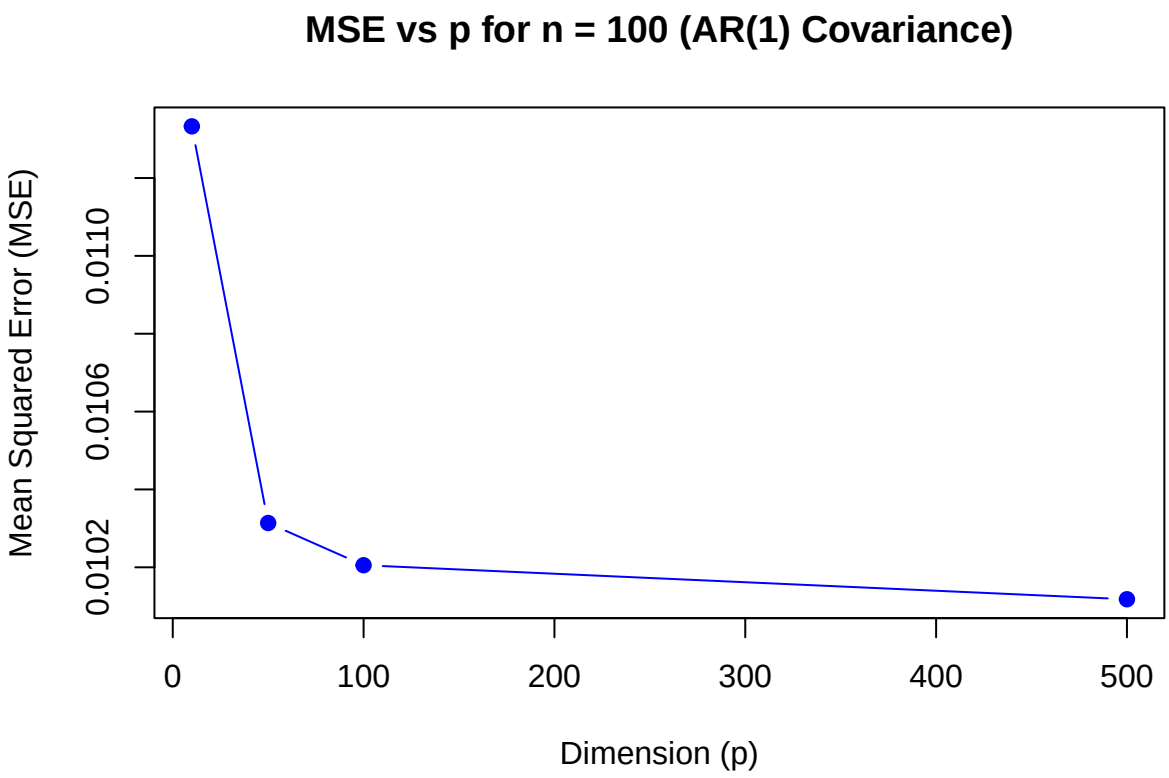


Table 4: MSE for Different Dimensions (n = 100, AR(1) Covariance with rho = 1/sqrt(p))

Dimension..p.	MSE
10	0.011333
50	0.010314
100	0.010205
500	0.010118



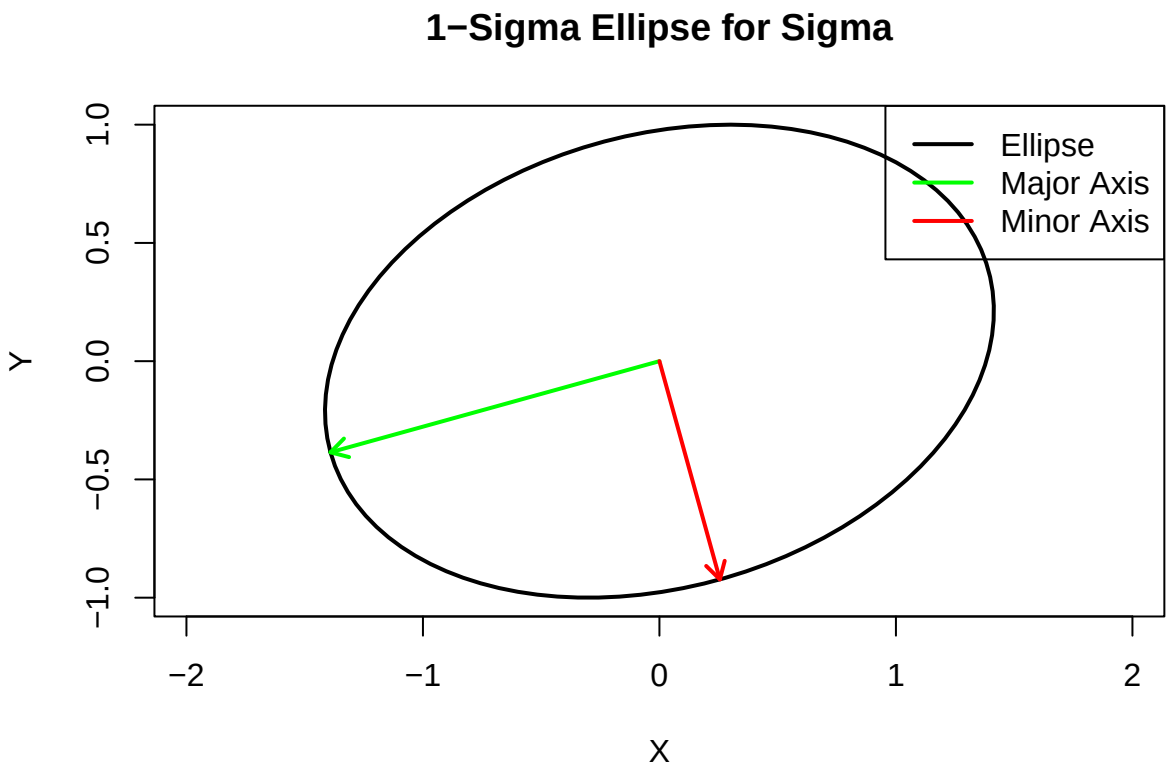
Question 2

Part a.

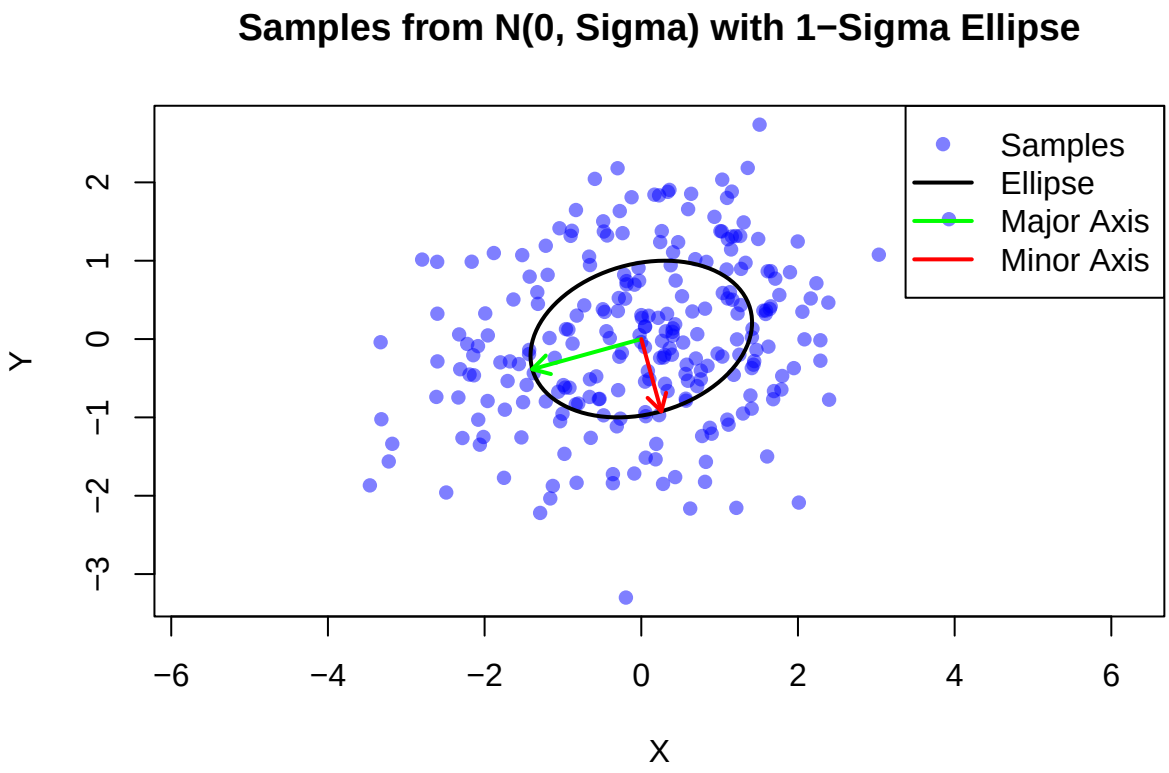
Table 5: Spectral Decomposition of Sigma

Component	Eigenvalue	Eigenvector_1	Eigenvector_2
Major Axis	2.083	-0.964	-0.267
Minor Axis	0.917	0.267	-0.964

Part b.


1−Sigma Ellipse for Sigma
Ellipse
Major Axis
Minor Axis
Y
X
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2

Part c.



Question 3

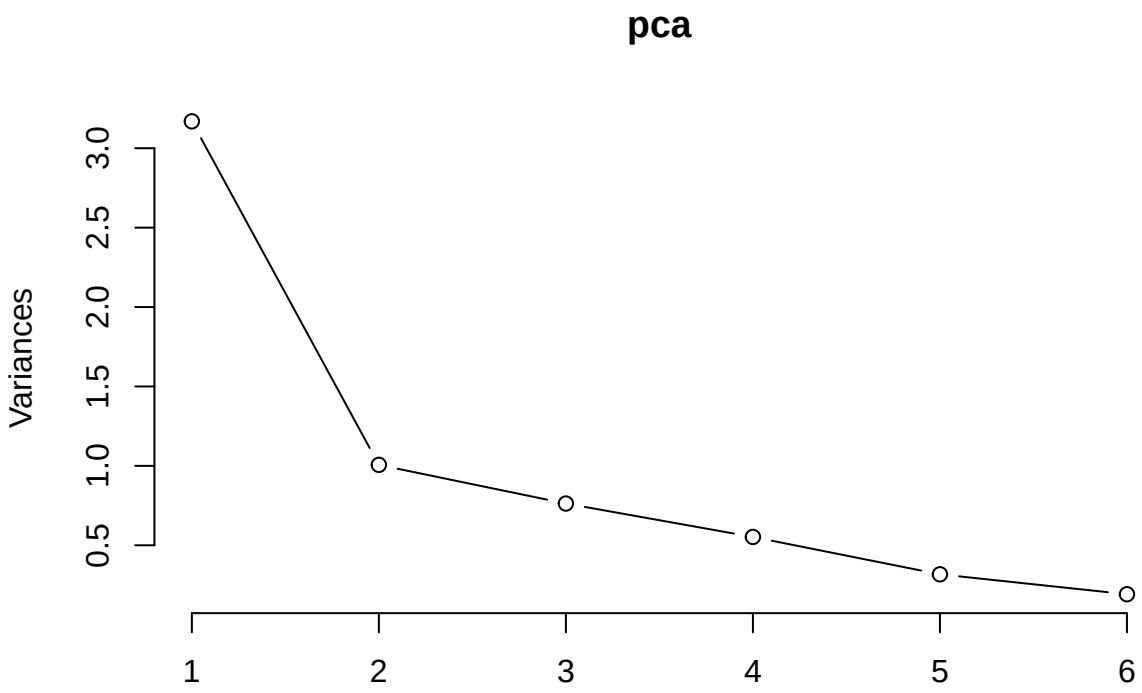
Part a.

Table 6: 95% Confidence Interval for Complaints Coefficient

Variable	Lower_Bound	Upper_Bound
Complaints	0.28	0.946

Part b.

Based off the elbow in the scree plot below I should keep two PCs.



Part c.

Table 7: 95% Confidence Interval for the Effect of PC1

	2.5 %	97.5 %
pc1	2.673	6.573

Part d.

Using the table below and comparing to Part .a I get an estimated coefficient for complaints of –0.216 and a 95% C.I. of (–0.39, –0.04). In Part d. I use the loading of complaints from the PCA and the 95% C.I. from PC1. I get a 95% C.I. for the effect of complaints of (–0.38, –0.06).

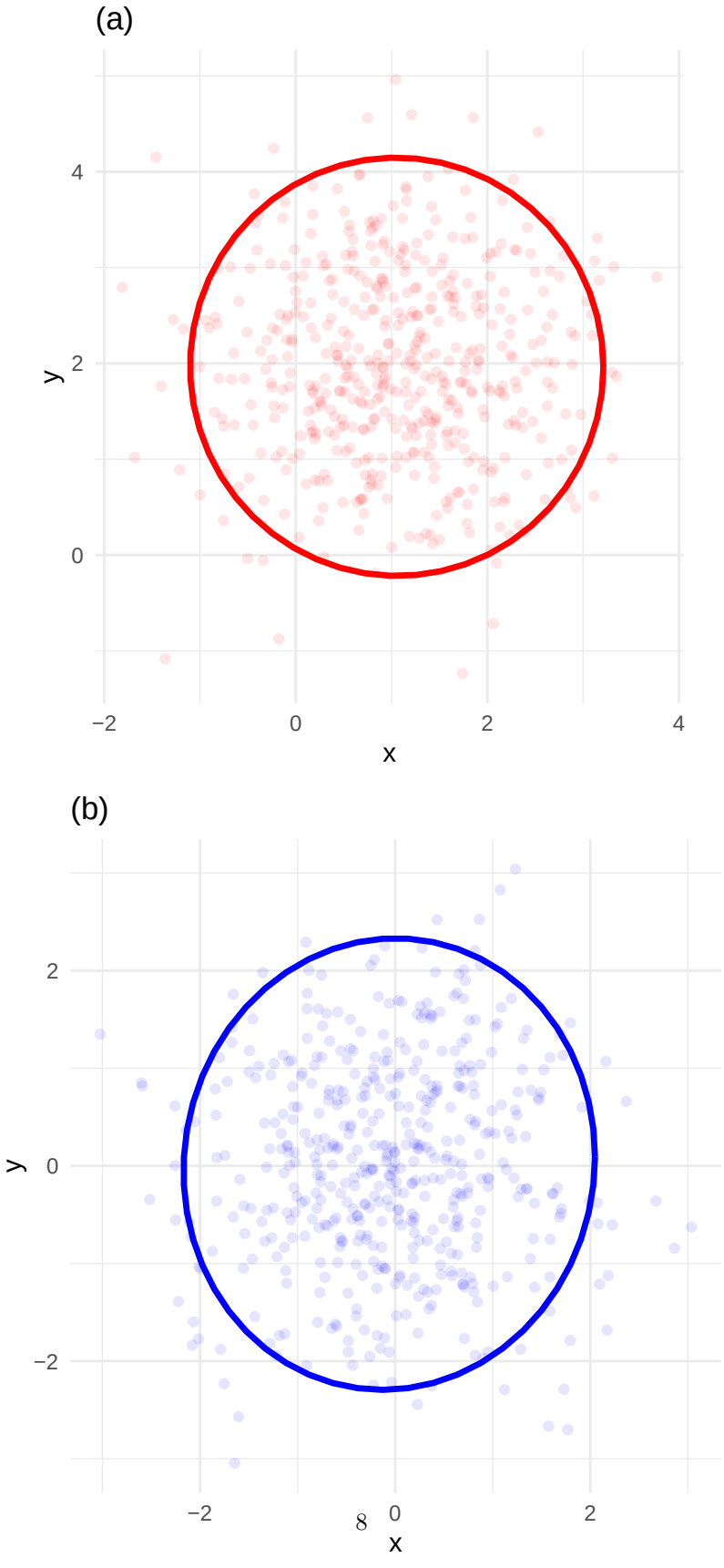
This small difference in the C.I.'s means using PC1 to represent the covariates retains much of the information of complaints and provides an effect estimate that aligns well with the full model.

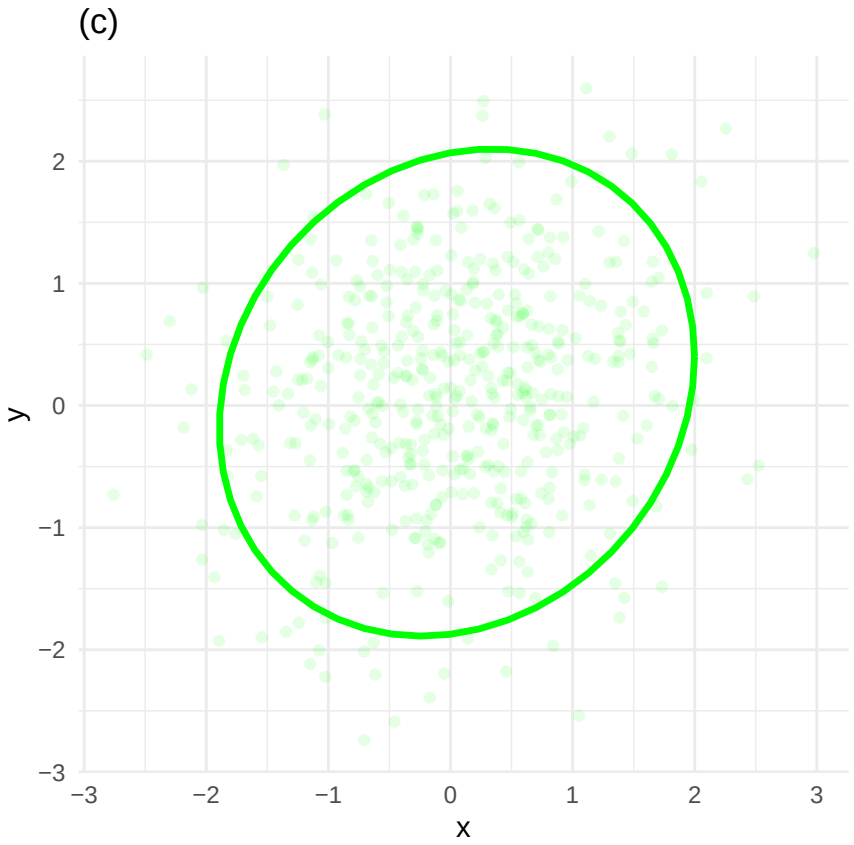
Table 8: 95% Confidence Interval for the Effect of 'Complaints' Based on PC1

	Lower Bound	Upper Bound
Complaints (From PC1 Model)	1.174	2.888

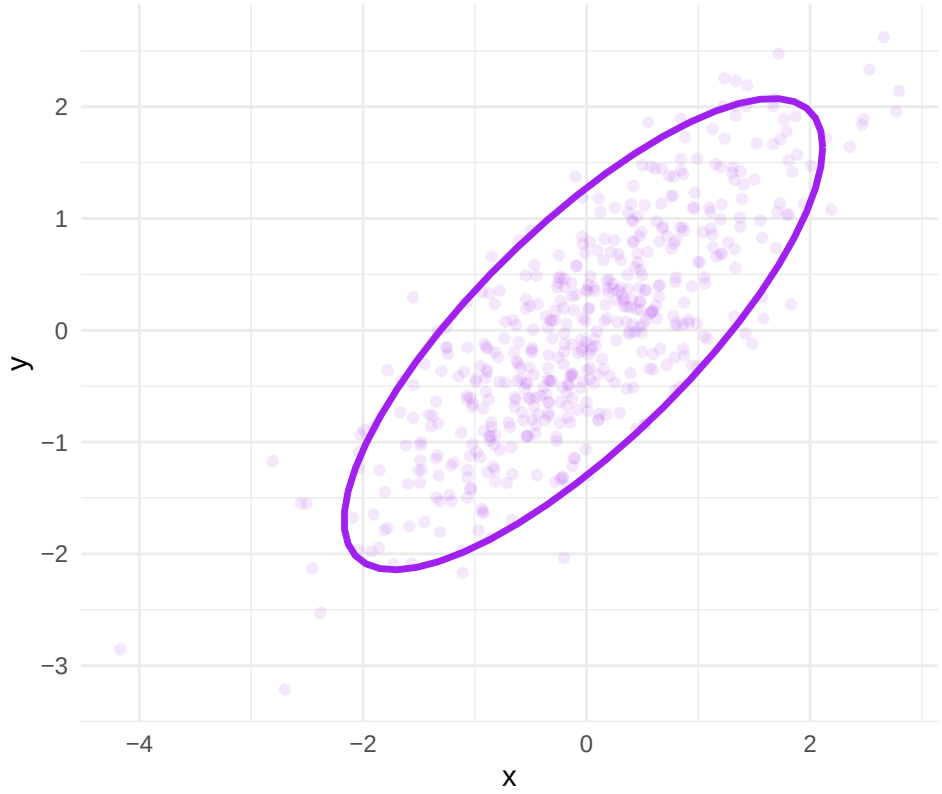
Question 4

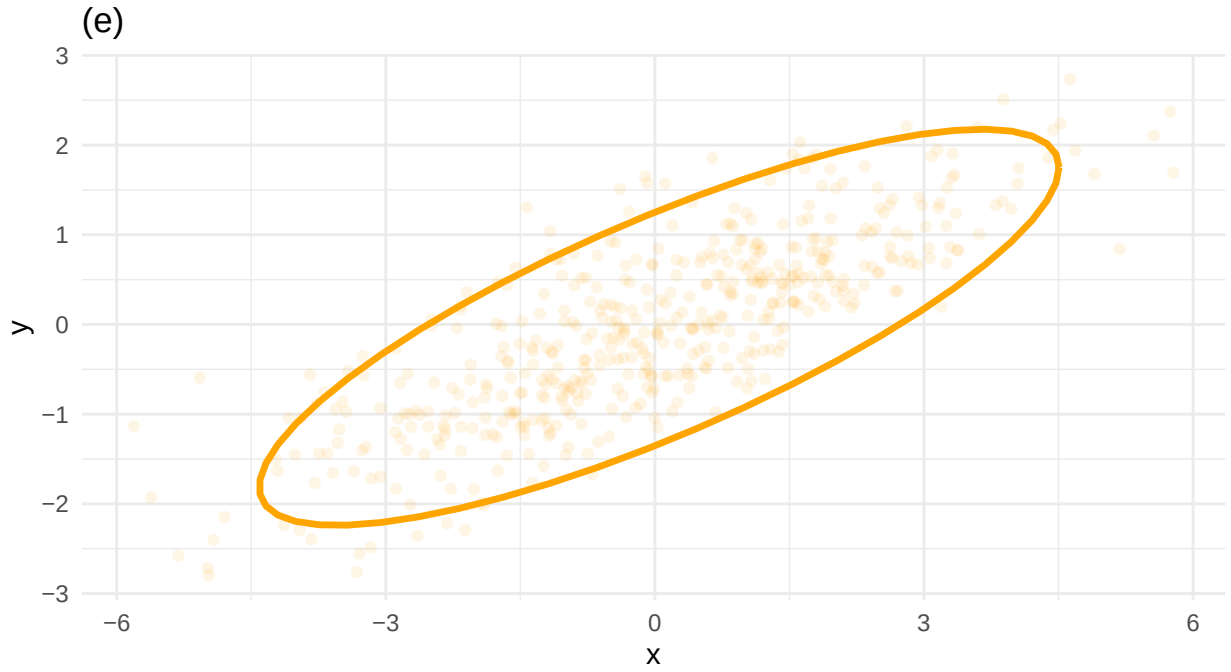
2.6




(c)
2
1
0
−1
−2
−3
y
−3
−2
−1
0
1
2
3
x

(d)





2.10

Need to show that the principal axes of the bivariate normal distribution lie along the 45° and 135° lines, with lengths $2\sqrt{c(1+\rho)}$ and $2\sqrt{c(1-\rho)}$

The covariance matrix for a bivariate normal distribution with correlation ρ is:

$$\Sigma = \begin{bmatrix} c & \rho c \\ \rho c & c \end{bmatrix}$$

To find the principal axes, compute the eigenvalues and eigenvectors of Σ.

The equation of Σ is:

$$\det \begin{bmatrix} c-\lambda & \rho c \\ \rho c & c-\lambda \end{bmatrix} = 0.$$

Expanding the determinant:

$$(c-\lambda)^2 - (\rho c)^2 = 0$$

$$(c-\lambda-\rho c)(c-\lambda+\rho c) = 0.$$

Solving for λ, obtain:

$$\lambda_1 = c(1+\rho), \quad \lambda_2 = c(1-\rho).$$

These eigenvalues represent the variances along the principal directions.

To find the eigenvectors:

$$\begin{bmatrix} c-\lambda & \rho c \\ \rho c & c-\lambda \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0.$$

For $\lambda_1 = c(1+\rho)$:

$$\begin{bmatrix} c-c(1+\rho) & \rho c \\ \rho c & c-c(1+\rho) \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0.$$

$$\begin{bmatrix} -c\rho & \rho c \\ \rho c & -c\rho \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0.$$

Setting $v_1 = v_2$, the eigenvector is:

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

For $\lambda_2 = c(1-\rho)$, the eigenvector is:

$$\mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

The eigenvectors indicate that the principal axes lie along the directions:

- 45° (eigenvector $\mathbf{v}_1$)
- 135° (eigenvector $\mathbf{v}_2$)

Define the transformation:

$$y_1 = \frac{z_1 + z_2}{\sqrt{2}}, \quad y_2 = \frac{z_1 - z_2}{\sqrt{2}}.$$

This transformation rotates the coordinate system to align with the principal axes.

The standard deviations along the principal axes are:

$$\sigma_1 = \sqrt{\lambda_1} = \sqrt{c(1+\rho)},$$

$$\sigma_2 = \sqrt{\lambda_2} = \sqrt{c(1-\rho)}.$$

The principal axes of the ellipse are:

$$2\sigma_1 = 2\sqrt{c(1+\rho)}, \quad 2\sigma_2 = 2\sqrt{c(1-\rho)}.$$

2.12

Need to show that if:

$$\Pr(X \geq 0, Y \geq 0) = \alpha$$

For the bivariate normal distribution:

$$N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right),$$

Then the correlation coefficient ρ is:

$$\rho = \cos(1 - 2\alpha)\pi.$$

Define new random variables:

$$X = U, \quad Y = \rho U + \sqrt{1 - \rho^2}V,$$

Where $U \sim N(0, 1)$ and $V \sim N(0, 1)$ are independent standard normal variables.

The probability I want is:

$$\Pr(X \geq 0, Y \geq 0) = \alpha.$$

The chance of both values being positive in a bivariate normal distribution is α.

Because the joint normal distribution is symmetric, the probability relates to an angle θ, with the first quadrant as a sector. A known formula gives:

$$\Pr(X \geq 0, Y \geq 0) = \frac{1}{4} + \frac{1}{2\pi}\tan^{-1}\rho.$$

Setting this equal to α:

$$\frac{1}{4} + \frac{1}{2\pi}\tan^{-1}\rho = \alpha.$$

2.25

(a) To determine the rank of Σ, compute its determinant:

$$\det(\Sigma) = (4 \cdot 1) - (2 \cdot 2) = 4 - 4 = 0.$$

Since the determinant is zero, the matrix is singular, meaning at least one eigenvalue is zero.

Find the eigenvalues λ by solving:

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

$$\begin{vmatrix}4 - \lambda & 2\\ 2 & 1 - \lambda\end{vmatrix} = (4 - \lambda)(1 - \lambda) - (2 \cdot 2) = 4 - 4\lambda + \lambda - \lambda^2 - 4 = -\lambda^2 + 3\lambda.$$

Setting this to zero:

$$-\lambda^2 + 3\lambda = 0$$

$$\lambda(\lambda - 3) = 0.$$

The eigenvalues are $\lambda_1 = 3$ and $\lambda_2 = 0$. Since there is exactly one nonzero eigenvalue, the rank of Σ is 1.

(b) Since Σ has rank 1, its columns are linearly dependent. Show the columns as multiples of a single vector. The covariance matrix shows that:

$$\Sigma = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \begin{bmatrix} 4 & 2 \end{bmatrix}.$$

Choosing:

$$a = \begin{bmatrix} 2 \\ 1 \end{bmatrix},$$

Define Y as a standard normal variable with variance 1:

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

Since $X = a'Y$, the density of Y is the standard normal density:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, \quad y \in \mathbb{R}.$$

Question 5

11.1

The characteristic equation is found by solving:

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

$$\begin{vmatrix} 1-\lambda & \rho \\ \rho & 1-\lambda \end{vmatrix} = (1-\lambda)(1-\lambda) - \rho^2.$$

$$(1-\lambda)^2 - \rho^2 = 0.$$

$$(1 - \lambda - \rho)(1 - \lambda + \rho) = 0.$$

The eigenvalues are:

$$\lambda_1 = 1 + \rho, \quad \lambda_2 = 1 - \rho.$$

For $\lambda_1 = 1 + \rho$, solve:

$$\begin{bmatrix} 1-(1+\rho) & \rho \\ \rho & 1-(1+\rho) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0.$$

$$\begin{bmatrix} -\rho & \rho \\ \rho & -\rho \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0.$$

Setting the first row equation:

$$-\rho x + \rho y = 0 \quad \Rightarrow \quad x = y.$$

Choosing $x = \frac{1}{\sqrt{2}}$, gives the eigenvector:

$$v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}.$$

For $\lambda_2 = 1 - \rho$, solve:

$$\begin{bmatrix} -(-\rho) & \rho \\ \rho & -(-\rho) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0.$$

$$\begin{bmatrix} \rho & \rho \\ \rho & \rho \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0.$$

Setting the first row equation:

$$\rho x + \rho y = 0 \quad \Rightarrow \quad x = -y.$$

Choosing $x = \frac{1}{\sqrt{2}}$, gives the eigenvector:

$$v_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

The characteristic vectors of A are:

$$\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \quad \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix},$$

Corresponding to eigenvalues $1 + \rho$ and $1 - \rho$, respectively.

11.5

(a) If all eigenvalues are equal $\lambda_1 = \lambda_2 = \cdots = \lambda_p = \lambda$, then the covariance matrix takes the form:

$$\Sigma = \lambda I_p,$$

Where I_p is the $p \times p$ identity matrix. This indicates that the distribution is **isotropic**, meaning it has the same spread in all directions.

The ellipsoid of constant density in this case is a **sphere** centered at the mean, since all principal axes have the same length.

(b) The covariance matrix has one distinct eigenvalue λ_1 and all others equal to λ (where $\lambda_1 > \lambda$). This means:

$$\Sigma = \lambda I_p + (\lambda_1 - \lambda) vv',$$

Where v is the eigenvector corresponding to λ_1.

The ellipsoid of constant density in this case is an **elongated ellipsoid**, stretched along the principal axis corresponding to λ_1. This means the spread of the distribution is greater in one direction compared to the others.

11.10

Problem gives:

- $U_1 = \beta^{(1)'} X$ is the first population principal component with variance:

$$\mathcal{V}(U_1) = \lambda_1.$$

- $V_1 = b^{(1)'} X$ is the first sample principal component with sample variance l_1, based on sample covariance matrix S.
- S^* is the covariance matrix of a second independent sample.

Show that:

$$b^{(1)'} S^* b^{(1)} \leq \lambda_1.$$

The first principal component direction $\beta^{(1)}$ is the eigenvector corresponding to the largest eigenvalue λ_1 of the population covariance matrix Σ, meaning:

$$\Sigma \beta^{(1)} = \lambda_1 \beta^{(1)}.$$

The first sample principal component direction $b^{(1)}$ is chosen to maximize:

$$b' S b,$$

Subject to $\|b\| = 1$. The sample estimate $b^{(1)}$ converges to $\beta^{(1)}$ as the sample size increases.

Since S^* is another independent estimate of the population covariance matrix Σ, its eigenvalues approximate those of Σ, but with sampling variability. But for any unit vector b:

$$b' S^* b \leq \lambda_1.$$

Since $b^{(1)}$ is an estimate of $\beta^{(1)}$, and S^* approximates Σ, the largest possible variance explained by any direction is bounded by λ_1, the maximum eigenvalue of Σ.

$$b^{(1)'} S^* b^{(1)} \leq \lambda_1.$$

Showe above is that the variance captured by the first sample principal component in an independent sample cannot exceed the maximum population variance λ_1.

11.17

Principal Components Analysis of Painted Turtle Measurements

The data consist of 24 observations on three variables (Length, Width, Height in mm). After computing the sample means and (unbiased) covariance matrix,

$$S = \frac{1}{n-1}\sum_{i=1}^{24}(x_i - \bar{x})(x_i - \bar{x})^T,$$

One obtains:

$$S \approx \begin{pmatrix} 140.32 & 79.50 & 38.26 \\ 79.50 & 50.04 & 22.82 \\ 38.26 & 22.82 & 11.67 \end{pmatrix}.$$

The principal components are found by solving:

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

This yields the eigenvalues (which are the principal-component variances)

$$\lambda_1 \approx 197.3, \quad \lambda_2 \approx 3.55, \quad \lambda_3 \approx 1.18,$$

With total variability:

$$\lambda_1 + \lambda_2 + \lambda_3 \approx 202.0.$$

98% of the total variation is captured by the first component.

A corresponding set of normalized eigenvectors (principal-component loadings) is:

$$v_1 \approx \begin{pmatrix} 0.844 \\ 0.490 \\ 0.218 \end{pmatrix}, \quad v_2 \approx \begin{pmatrix} -0.380 \\ 0.686 \\ -0.624 \end{pmatrix}, \quad v_3 \approx \begin{pmatrix} 0.377 \\ -0.536 \\ -0.755 \end{pmatrix}.$$

Since the loadings for the first component are all positive and of similar magnitude, PC1 is interpreted as an overall factor, while the remaining components (with much smaller variances) account for minor shape differences.

$$S \approx \begin{pmatrix} 140.32 & 79.50 & 38.26 \\ 79.50 & 50.04 & 22.82 \\ 38.26 & 22.82 & 11.67 \end{pmatrix}, \quad \lambda_1 \approx 197.3, \quad \lambda_2 \approx 3.55, \quad \lambda_3 \approx 1.18,$$

$$v_1 \approx (0.844,\ 0.490,\ 0.218), \quad v_2 \approx (-0.380,\ 0.686,\ -0.624), \quad v_3 \approx (0.377,\ -0.536,\ -0.755).$$