Family name: *Given names:* *Student ID:*

1	
2	
3	
4	
5	
6	
T	

STA 437/1005 — Mid-term Test — 2010-10-18

For all questions, show enough of your work to indicate how you obtained your answer. No books or notes are allowed. For all questions that have a numerical answer, give your answer as an actual number (eg, 5/4 or 1.25).

The questions are worth equal amounts; they may not be equally difficult.

Here are some of the formulas relating to the material we covered. These formulas might or might not be relevant to some of the questions on the mid-term test.

Sample covariance:

$$\mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})'$$

Covariance of a random vector:

$$\mathrm{Cov}(\mathbf{X}) = E[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))']$$

Covariance of transformed random vector:

$$\mathrm{Cov}(\mathbf{CX}) = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{C}'$$

Spectral decomposition:

$$\mathbf{A} = \lambda_1\mathbf{e}_1\mathbf{e}_1' + \cdots + \lambda_k\mathbf{e}_k\mathbf{e}_k'$$

Probability density function for multivariate normal:

$$f(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp(-(\mathbf{x}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)/2)$$

Conditional mean and covariance for multivariate normal:

$$\text{Mean of } \mathbf{X}_1 \text{ given } \mathbf{X}_2 = \mathbf{x}_2 = \mu_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \mu_2)$$

$$\text{Covariance of } \mathbf{X}_1 \text{ given } \mathbf{X}_2 = \mathbf{x}_2 = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

1. Suppose we have five observations of three variables, as follows:

$$\begin{array}{rrr} 57 & 72 & -11 \\ 45 & 67 & -9 \\ 46 & 69 & -10 \\ 53 & 71 & -11 \\ 49 & 71 & -9 \end{array}$$

   (a) Find the sample mean vector for this data.

   (b) Find the sample covariance matrix for this data. (Use the definition in which the divisor is the number of observations minus one.)

   (c) Find the sample correlation matrix for this data.

2. For three subjects, we take measurements of systolic blood pressure, denoted by X_1, X_2, and X_3, and of diastolic blood pressure, denoted by Y_1, Y_2, and Y_3, both in units of mmHg.

   Suppose that the blood pressure measurements are independent from one subject to another, and that the measurements on all subjects have the same bivariate distribution for $[X_i, Y_i]'$. The mean vector for $[X_i, Y_i]'$ is $[130, 85]'$. The standard deviation of X_i is 10 and the standard deviation of Y_i is 5. The correlation between X_i and Y_i is 0.5.

   (a) What is the conditional distribution of Y_1 given that $X_1 = 140$?

   (b) Suppose that from the systolic and diastolic blood pressure measurements we compute the average systolic blood pressure for the three subjects, denoted by S, and the average difference of systolic minus diastolic blood pressure for the three subjects, denoted by D. What is the mean vector and covariance matrix of $[S, D]'$?

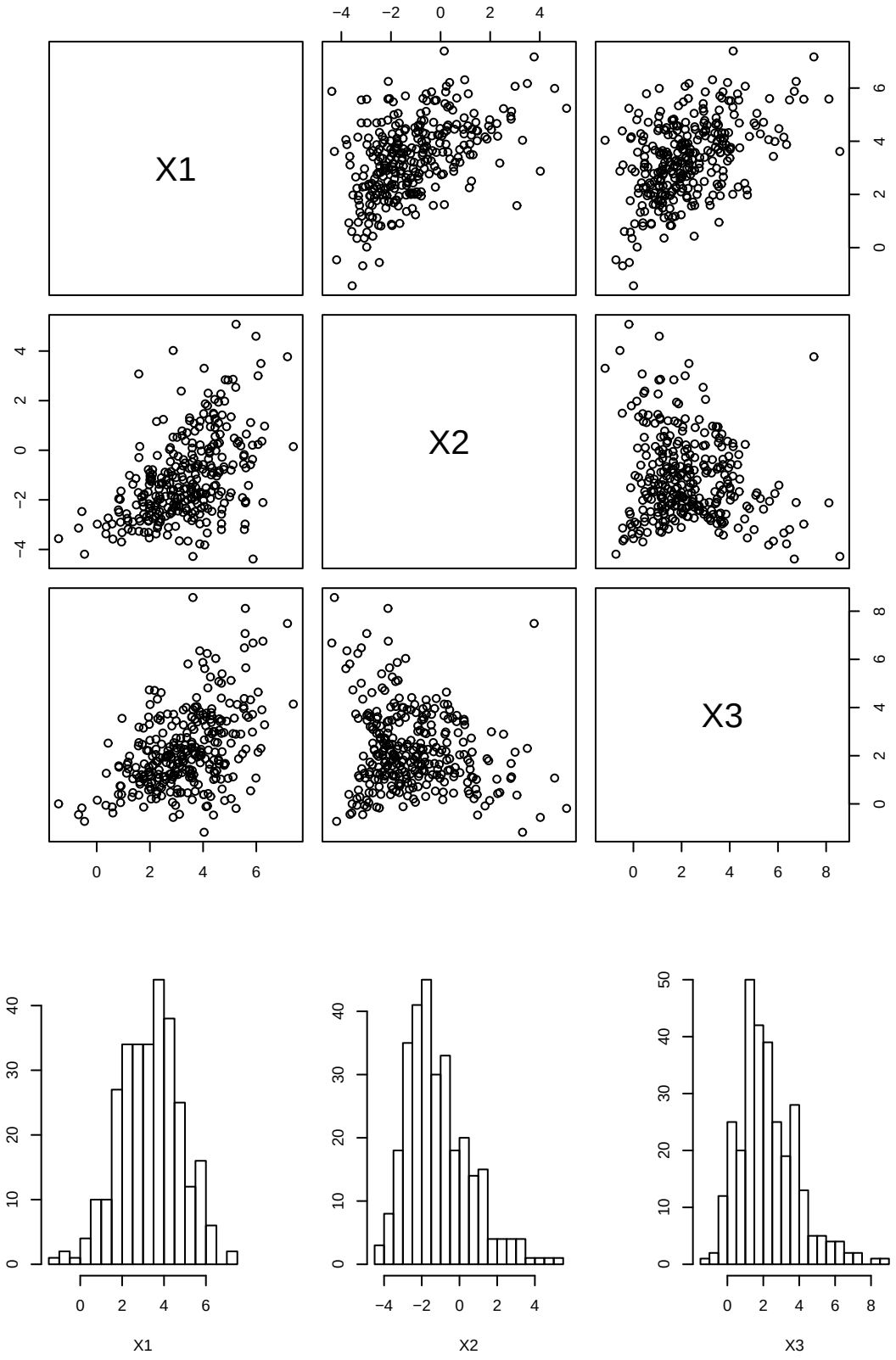
3. Prove the statements below, showing all details of the proofs.

   (a) Prove that if A is a square matrix, and e is an eigenvector of A with eigenvalue λ, then e is also an eigenvector of cA, where c is a scalar. Also, find the eigenvalue of e as an eigenvector of cA.

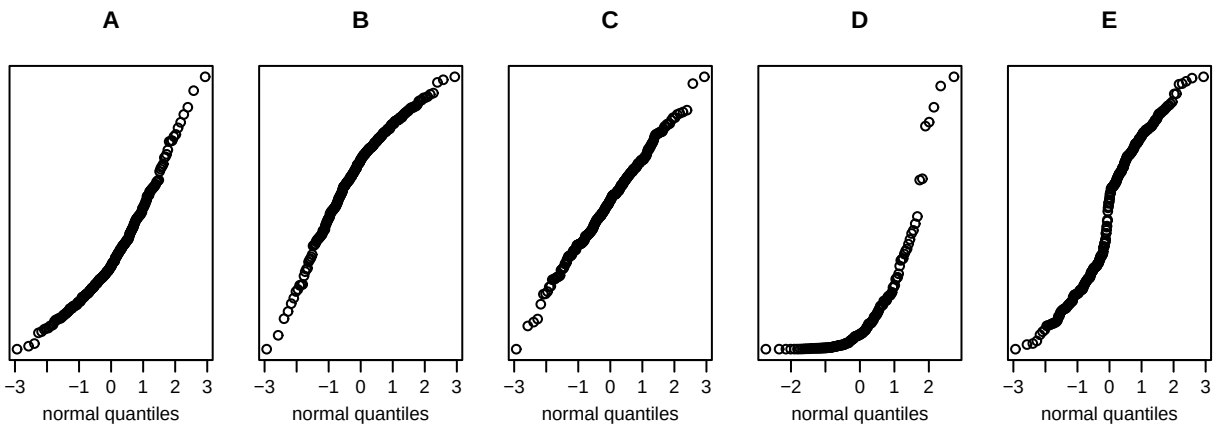
   (b) Prove that if A is a square matrix, and e is an eigenvector (of length one) of A with eigenvalue λ, then e is also an eigenvector of $A + ee'$. Also, find the eigenvalue of e as an eigenvector of $A + ee'$.

   (c) Prove that if A and B are symmetric, positive definite matrices with the same dimensions, then $A + B$ is also a symmetric, positive definite matrix.

4. Here are all pairwise scatterplots for 300 observations on three variables (X1, X2, and X3), along with histograms for each of the three variables:



Here are five quantile-quantile plots, with quantiles of the standard normal distribution on the horizontal axis and sample quantiles on the vertical axis. The scales on the vertical axes have been omitted.



(a) Which of the quantile-quantile plots above is for variable X1?
Write the letter here (no explanation required):

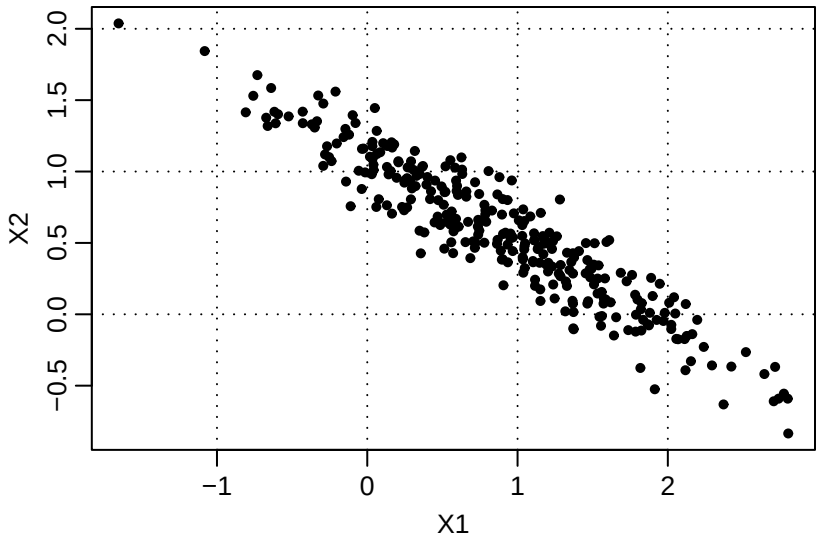
(b) Which of the quantile-quantile plots above is for variable X2?
Write the letter here (no explanation required):

(c) Do any of the observations in this data set appear to be outliers (which may be errors, or otherwise not be from the same distribution as the rest of the data)? If so, explain why you think so, and indicate which ones (eg, by saying how you have marked them on the plots on the previous page).

(d) Discuss whether or not there is good reason to believe that any of X1, X2, and X3 is not normally distributed, considering each variable separately, and ignoring any outliers that you identified in part (c) of this question.

(e) Discuss whether or not there is good reason to believe that $[X1\ X2\ X3]'$ does not have a multivariate normal distribution, ignoring any outliers that you identified in part (c) of this question.

5. Here is a scatterplot of 300 observations on two variables:



The sample mean vector for this data is $[0.88,\ 0.57]'$.

(a) Which of the following is the sample covariance matrix for this data?

$$\begin{bmatrix} 0.58 & 0.45 \\ 0.45 & 0.24 \end{bmatrix} \quad \begin{bmatrix} 4.89 & -1.37 \\ -1.37 & 2.20 \end{bmatrix} \quad \begin{bmatrix} 0.63 & -0.37 \\ -0.37 & 0.25 \end{bmatrix} \quad \begin{bmatrix} 0.68 & -0.32 \\ -0.32 & 0.71 \end{bmatrix}$$

(a) (b) (c) (d)

Write the letter for the correct answer here (no explanation required):

(b) Which of the following is a vector pointing in the direction of the first principal component?

$$\begin{bmatrix} 0.55 \\ 0.84 \end{bmatrix} \quad \begin{bmatrix} 0.85 \\ -0.52 \end{bmatrix} \quad \begin{bmatrix} 0.52 \\ -0.85 \end{bmatrix} \quad \begin{bmatrix} -0.55 \\ -0.84 \end{bmatrix} \quad \begin{bmatrix} -0.80 \\ -0.60 \end{bmatrix}$$

(a) (b) (c) (d) (e)

Write the letter for the correct answer here (no explanation required):

(c) What is the projection of the data point $[-0.12,\ 1.07]'$ on the first principal component?

(d) Write down a vector that points in the direction of the second principal component.

6. For each of the following questions, answer "yes" or "no" **and** give an explanation of your answer. No marks will be given for an answer without an explanation.

   (a) Suppose we have data on the height in metres and the weight in kilograms of 100 people, for which the sample correlation between height and weight is 0.59. Will the sample correlation between height and weight change if we re-express heights in feet and weights in pounds?

   (b) Suppose we have data on the lengths of arms, legs, and noses of 100 people. We intend to find the direction of the first principal component for this data, using the sample covariance matrix. Does it matter for this purpose whether we express all these lengths in inches, or instead express all these lengths in centimetres?

   (c) Suppose we have data on the shoulder height, weight, and milk production of 100 three-year-old Holstein cows. We intend to find the direction of the first principal component from the sample covariance matrix of this data. Does it matter for this purpose whether we measure height in centimetres, weight in kilograms, and milk production in litres per day, or instead measure height in inches, weight in pounds, and milk production in gallons per day?