Probability and Statistics Fall 2020
HW6 Matlab assignment

1. Bivariate normal distribution

If X and Y are two continuous random variables and they jointly form a normal distribution f(x, y). f(x, y) is called a bivariate normal distribution. In fact, more than two random variables can jointly form a multivariate normal distribution. Here, we look at a bivariate normal distribution as an example.

The probability density function of a bivariate normal distribution can be defined as:

$$\mathrm{f(x,y)} = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left[-\frac{z}{2(1-\rho^2)}\right]$$, where *z* is:

$$\mathrm{z} = \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y},$$

ρ is the correlation coefficient of X and Y.

1.(a) Now, please plot four bivariate normal distributions with the following parameters. Show each distribution in a 2D plot in a x-y plane with color representing f(x, y). **You are asked to turn in four plots here.** [Hint: You may consider using imagesc and colormap(jet) functions.]

Range of simulation:

X = from 0 to 100
Y = from 1000 to 2000
(For X and Y, an increment step of 1 is good enough. You are free to go finer increment steps.)

Parameters:

Distribution 1: $[\mu_x, \sigma_x, \mu_y, \sigma_y, \rho] = [50, 20, 1500, 200, 0]$
Distribution 2: $[\mu_x, \sigma_x, \mu_y, \sigma_y, \rho] = [50, 20, 1500, 200, 0.3]$
Distribution 3: $[\mu_x, \sigma_x, \mu_y, \sigma_y, \rho] = [50, 20, 1500, 200, 0.8]$
Distribution 4: $[\mu_x, \sigma_x, \mu_y, \sigma_y, \rho] = [50, 20, 1500, 200, -0.8]$

1.(b) From 1.(a), please explain the difference in four distributions and the effect of changing ρ.

Reference:

http://mathworld.wolfram.com/BivariateNormalDistribution.html

2. Decision boundaries

In statistics (and machine learning), people often want to separate samples from two populations (i.e., this is called a classification problem). In this problem, you are asked to simulate two bivariate normal distributions and find out the boundary that these two distribution would be equal. These boundaries are called decision boundaries. Points on the decision boundaries have equal values for both probability distributions (e.g., meaning that they are equally probable to belong to either one of the two bivariate normal distributions). As the name ‘decision boundary’ implies, points on either side of the decision boundaries belong to the closer bivariate normal distribution. Now, please simulate the following two cases and find out decision boundaries between two distributions.

Range of simulation:
X = from 0 to 100
Y = from 1000 to 2000
(For X and Y, an increment step of 1 is good enough. You are free to go finer increment steps.)

Case 1:
Distribution 1: $[\mu_x, \sigma_x, \mu_y, \sigma_y, \rho] = [25, 30, 1250, 300, 0]$
Distribution 2: $[\mu_x, \sigma_x, \mu_y, \sigma_y, \rho] = [75, 30, 1750, 300, 0]$

Case 2:
Distribution 1: $[\mu_x, \sigma_x, \mu_y, \sigma_y, \rho] = [25, 20, 1250, 200, 0]$
Distribution 2: $[\mu_x, \sigma_x, \mu_y, \sigma_y, \rho] = [75, 30, 1750, 300, 0]$

2.(a) **For each case, you are asked to turn in three 2D plots, including two distribution plots and one plot for the decision boundary.** [Hint: When finding the decision boundary, it is more convenient for you to check whether the difference of two distributions at a certain (x, y) is lower than 0.005 * max(f(x, y)).]

2.(b) **In addition, you are asked to compare the two decision boundaries and discuss how/why they are (same or different).** You are encouraged to experiment with different parameters to see how decision boundaries change accordingly.