# Machine Learning Lab 1

Student Name: ZHIKUN ZHU Student ID: 29356822

Q2. Histograms for both uniform distribution and normal distribution have been plot with 1000 numbers, and they distributed as their Probability Density Function (PDF). Then, the number of bins of function hist(·) raised from 40 to 200 with a step of 40. As a result, the histogram of normal distribution shows increasingly like its PDF as shown in Fig 1.

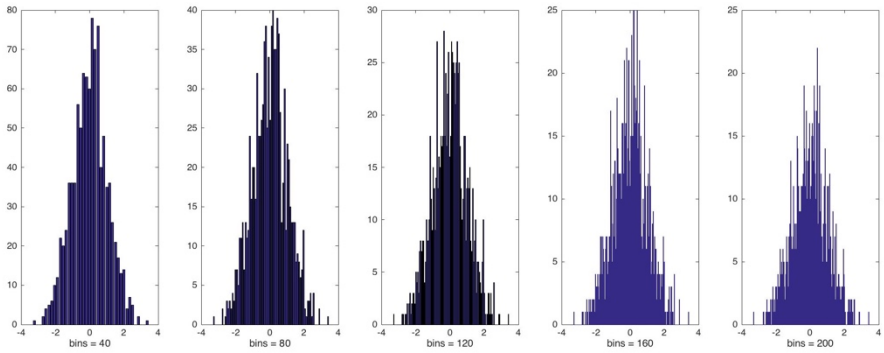


*Fig.1 Histogram change with number of bins*

Fig.1 illustrated that with the increase of bins, the height of each bin was dropped, which is obvious. Besides, the histograms increasingly approached to the theoretical PDF of normal distribution, which will be more distinct if the random number is much larger than 1000, 10,000 for example.

As the second programme was to be implemented, the outcomes(Fig.2(a)) showed that the distribution is highly like a normal distribution, which can be explained by central-limit theorem. However, it can only be used in a very large group of independent identically distributions. In this case, the PDF of the liner combination of 24 uniform distributions is still have a difference from normal distribution. Besides, when I test the particular case for the distribution: $\mathrm{sum}\big(rand(1,1)\big) - \mathrm{sum}\big(rand(1,1)\big)$ with MATLAB, I found its output is like a normal distribution (Fig.2(b)), which is not correct. It took me hours to find out that it is because the 'N' and 'bins' in the programme is too small. The histogram (Fig.2(c)) would be correct if they are ten times larger.

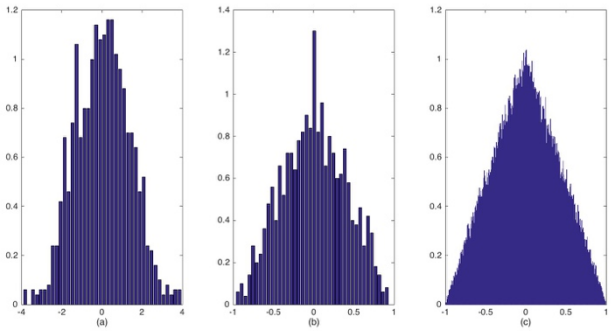


*Fig.2 Histogram outputs for the second programme*

Q3. The following Fig.3 shows the scatter plots for X and Y, respectively.

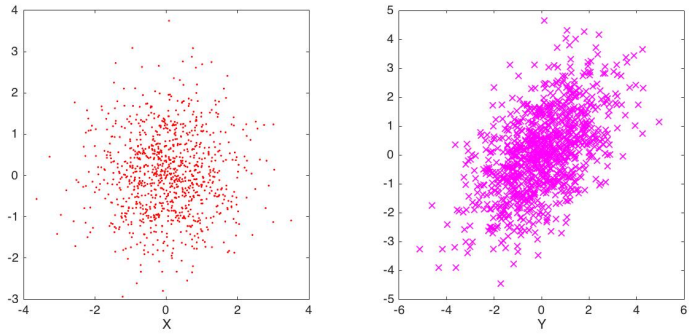


*Fig.3 The scatter plots for X and Y*

It is obvious from Fig.3 that there is no correlation between each column of X, because its dots distributed like a circle. Nevertheless, we can find the positive correlation of two columns of Y in the scatter plot itself.

Since X is a random matrix, so the script was tested several times to get the difference of empirical and theoretical variance. The outcome is that the difference is getting smaller as the number of data points increasing.

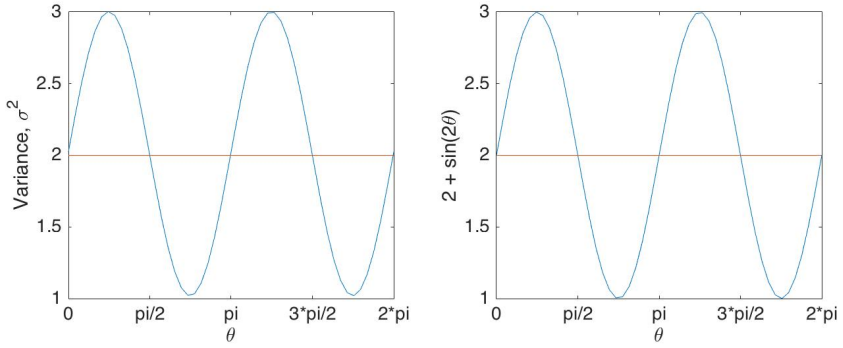


*Fig.4 Variance as a function of $\theta$ (left) and sinusoidal function: 2+ $sin(2\theta)$ (right)*

By calculating the eigenvectors, which can be found in Eq.1, eigenvector 1 is point to 3/4 π and eigenvector 2 is pointed to π /4 (According to rectangular coordinate system). So, when $\theta$=0, π /2, π , 3 π /2, variance of the projections of Y*u reaches 2. And when $\theta$= π/4, 5π/4, where u parallels with eigenvector 2, it reaches maximum value. Similarly, when $\theta$=3π/4, 2π, it reaches minimum with 1.

Mathematical expression of the variance of Y*u is as follow:

$$\begin{aligned} \mathrm{Var}(Yu) &= u'Cu = u'Q\Lambda Q'u \\ &= \left(sin(\theta)\ cos(\theta)\right)\begin{pmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} sin(\theta) \\ cos(\theta) \end{pmatrix} \qquad (Eq.1) \\ &= 2 + sin(2\theta) \end{aligned}$$

On the occasion when covariance matrix replaced by the one shown in the question, two columns of Y become negative correlated. And the eigenvector 2 is pointed to 5π/4, where the other remain the same.