Foundations of Machine Learning - Lab 1 Report

Johnny Joyce (jvjj1u19@soton.ac.uk)

October 12, 2019

1 Linear Algebra

In this section, we were given a matrix $\boldsymbol{U}$, whose columns are the eigenvectors of the symmetric 3×3 matrix $\boldsymbol{B}$. We notice that $\boldsymbol{U} \times \boldsymbol{U}^\intercal = \boldsymbol{I}$, where $\boldsymbol{I}$ is the 3×3 identity matrix. Therefore, by definition, we have that

$$\boldsymbol{U}^{-1} = \boldsymbol{U}^\intercal$$

Since $\boldsymbol{U}$ has this property, we call it an *orthogonal* matrix. The fact that U is orthogonal follows from the fact that any two distinct eigenvectors of a symmetric matrix are orthogonal to one another. The following proof taken from [1] demonstrates this.

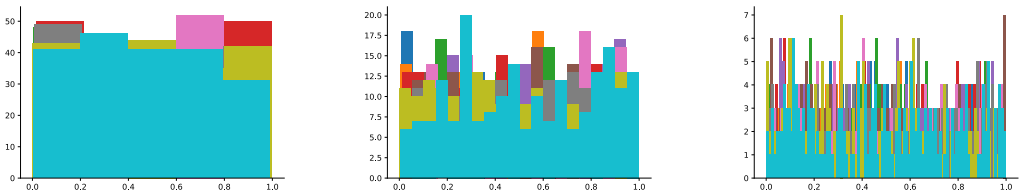
Lemma 1.1. *Let $\boldsymbol{B}$ be a symmetric 3×3 matrix with at least two distinct eigenvalues λ_1 and λ_2 and corresponding eigenvectors $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$. Then $\overrightarrow{v_1} \cdot \overrightarrow{v_2} = 0$.*

Proof. We have that

$$\lambda_1(\overrightarrow{v_1} \cdot \overrightarrow{v_2}) = (\lambda_1 \overrightarrow{v_1}) \cdot \overrightarrow{v_2} = (\boldsymbol{B}\overrightarrow{v_1}) \cdot \overrightarrow{v_2} = \overrightarrow{v_1} \cdot (\boldsymbol{B}^\intercal \overrightarrow{v_2}) = \overrightarrow{v_1} \cdot (\lambda_1 \overrightarrow{v_2}) = \lambda_2(\overrightarrow{v_1} \cdot \overrightarrow{v_2}).$$

So $(\lambda_1 - \lambda_2)(\overrightarrow{v_1} \cdot \overrightarrow{v_2}) = 0$. Therefore either $\lambda_1 - \lambda_2 = 0$ or $(\overrightarrow{v_1} \cdot \overrightarrow{v_2}) = 0$. But $\lambda_1 \neq \lambda_2$, so $(\overrightarrow{v_1} \cdot \overrightarrow{v_2}) = 0$, as required. □

2 Random Numbers and Univariate Distributions



(a) 5 bins (b) 20 bins (c) 100 bins

Figure 1: Histograms of 200 uniform random numbers over the interval $[0, 1]$ with (a) 5 bins (b) 20 bins (c) 100 bins

We were next tasked with creating a histogram of random numbers sampled from a uniform distribution. Figure 1 shows us that if we increase the number of bins whilst keeping the data itself constant, we obtain a representation that looks less similar to the original distribution.

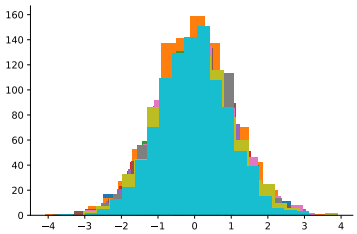
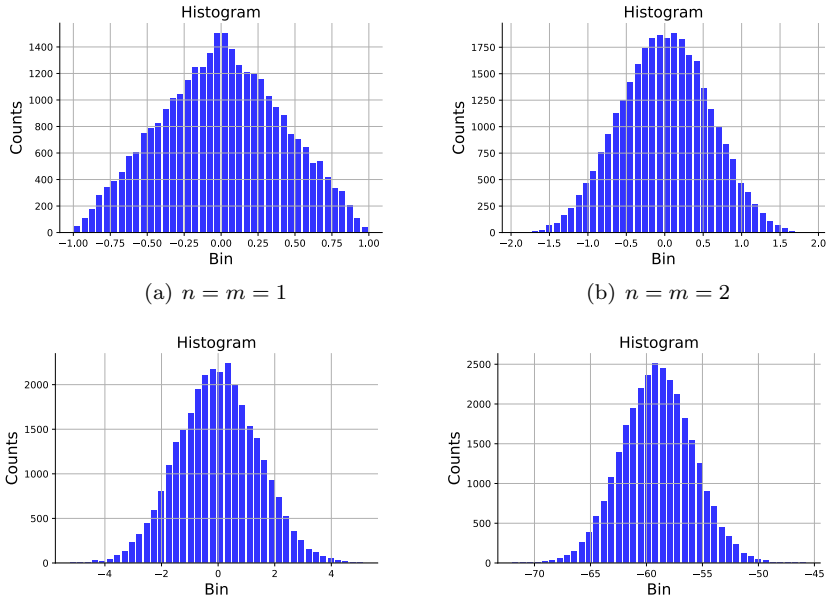


Figure 2: A histogram of 1000 random numbers sampled from a Gaussian distribution ($\mu = 0, \sigma = 1$) with 20 bins

Next, histograms were created where each value is the sum of n random variables minus the sum of m random variables, all of which are uniformly distributed. This resulted in a Gaussian distribution in the cases of Figure 3(b)(c)(d). We can see from Figure 3 that by increasing both n and m, we increase the standard deviation of the distribution. Furthermore, Figure 3(d) shows us that the mean of the distribution is proportional to $n - m$. This is to be expected, since if n is much greater than m, then we are adding far more numbers than we are subtracting, raising the mean of our distribution, and vice versa.

However, by observing Figure 3(a), we see that if we set $n = 1$ and $m = 1$, the resulting distribution is not Gaussian. We instead obtain a triangular distribution over the interval $[-1, 1]$ with a mean of 0. A brief proof of this result can be seen at [2].



(a) $n = m = 1$ (b) $n = m = 2$

(c) $n = m = 12$ (d) $n = 2, m = 120$

Figure 3: Histograms where each value is the sum of n random variables minus the sum of m random variables, all of which are uniformly distributed.

3 Uncertainty in Estimation

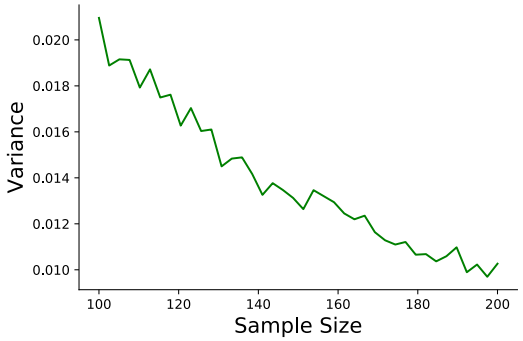


Figure 4: A line graph showing the variance for differing sample sizes

This section demonstrated the relationship between sample size and variance when given random samples. For each data point in Figure 4, random variables sampled from a Gaussian distribution ($\mu = 0, \sigma = 1$) were selected. The horizontal axis shows the size of the sample taken and the vertical axis shows the sample's respective variance. We can see that as the sample size increases, the variance decreases. This result is to be expected as it follows directly from the law of large numbers.

4 Bivariate Gaussian Distribution

This section involved plotting contour plots of Gaussian distributions over two variables. The resulting plots can be seen in Figure 5.

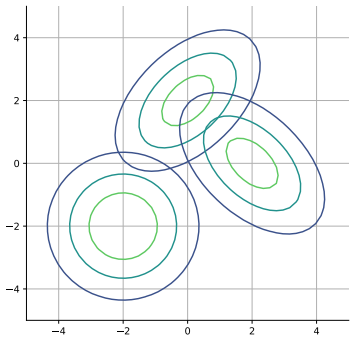


Figure 5: A contour plot of various bivariate Gaussian distributions

5 Sampling from a Multivariate Gaussian Distribution

A scatter plot was made of 10 000 samples from a bivariate Gaussian distribution ($\overrightarrow{\mu} = \overrightarrow{0}$, $\boldsymbol{C}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$). This is represented by the cyan scatters in Figure 6. We were also given the matrix $\boldsymbol{C} = \begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}$, which has Cholesky decomposition $\boldsymbol{A} = \begin{bmatrix}\sqrt{2} & 0\\ \frac{\sqrt{2}}{2} & \sqrt{\frac{3}{2}}\end{bmatrix}$. The magenta coordinates are the result of multiplying the cyan coordinates by $\boldsymbol{A}$; that is, each cyan coordinate $\mathbf{x}$ (in vector form) has a corresponding magenta coordinate $\boldsymbol{A}\mathbf{x}$.

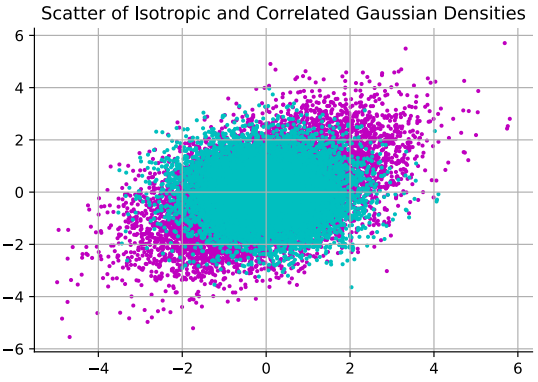


Figure 6: Two overlaid scatter plots. The cyan plot represents 10 000 samples from a bivariate Gaussian distribution ($\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1$). The magenta plot represents the samples from cyan plot multiplied by the Cholesky decomposition of the matrix $\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}$.

We can see from the shape of the magenta scatter plot that it has covariance matrix $\boldsymbol{C}$. To show this result, we apply the following theorem from lectures:

$$\text{If } \overrightarrow{x} \sim N(\overrightarrow{\mu},\boldsymbol{S}) \text{ and } \overrightarrow{y}=\boldsymbol{A}\overrightarrow{x} \text{ for some matrix } \boldsymbol{A}, \text{ then } \overrightarrow{y} \sim N(\boldsymbol{A}\overrightarrow{\mu},\ \boldsymbol{A}\ \boldsymbol{S}\ \boldsymbol{A}^\intercal)$$

To apply this, consider $\overrightarrow{x}$ to be any coordinate from our cyan distribution. Then $\overrightarrow{x} \sim N(\overrightarrow{0},\boldsymbol{I})$ (where $\boldsymbol{I}$ is the 2×2 identity matrix), so $\overrightarrow{y}$ would therefore be a coordinate from our magenta distribution by our theorem. Thus the covariance matrix of $\overrightarrow{y}$ is $\boldsymbol{A}\boldsymbol{I}\boldsymbol{A}^\intercal = \boldsymbol{A}\boldsymbol{A}^\intercal = \boldsymbol{C}$ since $\boldsymbol{A}$ is the Cholesky decomposition of $\boldsymbol{C}$. Furthermore, the mean will remain as $\overrightarrow{0}$ since our original mean was $\overrightarrow{0}$.

6 Distribution of Projections

This section uses the magenta scatter plot from Section 5. We created a graph that measures the variance of the set $Y_p \coloneqq \{x\cos(\theta) + y\sin(\theta) \mid (x, y) \in Y\}$, where $\theta \in \mathbb{R}$ is an independent variable and Y is the set of all coordinates from the magenta plot in Section 5. The resulting graph can be seen in Figure 7.

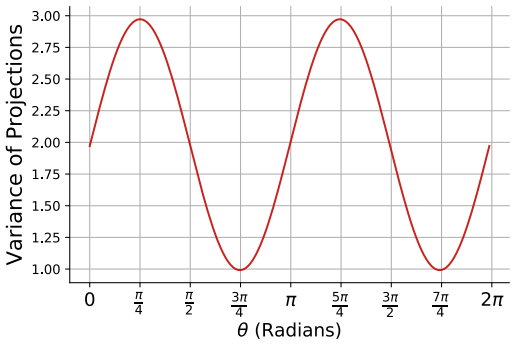


Figure 7: Graph of the variance of the set $\{x\cos(\theta) + y\sin(\theta) \mid (x, y) \in Y\}\}$, where Y is the set of points from the magenta plot in Figure 6.

Note that the projected variance in Figure 7 ranges from $\lambda_1 \coloneqq 1$ to $\lambda_2 \coloneqq 3$; these are equal to the eigenvalues of the covariance matrix $\boldsymbol{C}$ of Y. An interesting explanation for this result can be found at [3]. The peaks occur over the set $\{\frac{\pi}{4} + k\pi \mid k \in \mathbb{Z}\}$, which is the angle of a line from the origin over which the distribution has the highest variance (and vice versa for the troughs).

We know that bivariate Gaussian distributions have elliptic contours, making Y roughly elliptic. Therefore, we can form an ellipse by adding the projected variance to the mean of Y in each direction. Such an ellipse would have semi-major and semi-minor axes of lengths of 6 and 2 respectively (ignoring the rotation of the ellipse).

Ignoring the rotation of this ellipse, it would therefore satisfy the equation $\frac{x^2}{3^2} + \frac{y^2}{1^2} = 1$, which can be expressed parametrically as the set $\{(3\cos(t), 1\sin(t) \mid t \in [0, 2\pi]\}$. This shows us why our graph in Figure 7 appears sinusoidal. However it is worth noting that this graph is not perfectly sinusoidal since we are using a sample of random data rather than a contour curve.

References

[1] A. M. (https://math.stackexchange.com/users/742/arturo magidin), "Eigenvectors of real symmetric matrices are orthogonal." Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/82471 (version: 2017-11-05).

[2] L. Leemis, "Untitled page." `http://www.math.wm.edu/~leemis/chart/UDR/PDFs/StandarduniformStandardtriangular.pdf`.

[3] V. Spruyt, "A geometric interpretation of the covariance matrix." `https://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/`.