Imperial College London

COURSEWORK

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Mathematics for Machine Learning

Author:
Alexander Gaskell (CID: 01813313)

Date: October 18, 2019

1 Statistics and Probabilities

1.1

[8 marks]Compute the sample mean and the sample covariance matrix of the following dataset (use 1/N for the covariance matrix). Describe the computations you used to get to the answer.

$$D = [\begin{bmatrix}1\\2\\3\end{bmatrix}, \begin{bmatrix}-1\\0\\0\end{bmatrix}, \begin{bmatrix}-4\\4\\2\end{bmatrix}] = [\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3] \tag{1}$$

Sample mean is calculated as follows:

$$\overline{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n, \quad \boldsymbol{x} \in \mathbb{R}^3 \tag{2}$$

Hence to find the mean vector we sum horizontally for each row and divide by the number of samples (in this instance, 3). For example, to calculate $\overline{x_1}$, the calculation is $(1 + (-1) + (-4))/3 = -4/3$. Thus the sample mean is:

$$\overline{\boldsymbol{x}} = \begin{bmatrix}-4/3\\2\\5/3\end{bmatrix} \tag{3}$$

Sample covariance can be found as follows:

$$\mathrm{Cov}[D] = \mathrm{E}[(\boldsymbol{x} - \overline{\boldsymbol{x}})^2] = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \overline{\boldsymbol{x}})(\boldsymbol{x}_n - \overline{\boldsymbol{x}})^\top \tag{4}$$

This shows that the covariance is found by taking the mean of N outer products, where each outer product is composed of a single de-meaned sample multiplied by its transpose. So $\mathrm{Cov}[D]$ is calculated by the following:

$$\mathrm{Cov}[D] = \frac{1}{3}\left(\left(\boldsymbol{x}_1 - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_1^\top - \overline{\boldsymbol{x}}^\top\right) + \left(\boldsymbol{x}_2 - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_2^\top - \overline{\boldsymbol{x}}^\top\right) + \left(\boldsymbol{x}_3 - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_3^\top - \overline{\boldsymbol{x}}^\top\right)\right) \tag{5}$$

$$= \frac{1}{9}\begin{bmatrix}38 & -18 & 5\\-18 & 24 & 12\\5 & 12 & 14\end{bmatrix} = \begin{bmatrix}4.2 & -2.0 & 0.6\\-2.0 & 2.7 & 1.3\\0.6 & 1.3 & 1.6\end{bmatrix} \tag{6}$$

1.2

[9 marks]Generate two datasets $\{(x_1, x_2)_n\}$ of 100 data points each. The datasets have mean: $\mu = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and marginal variances $\sigma_1^2 = 2, \sigma_2^2 = 0.5$. Ensure that the shape of the datasets you generate is different. Visualize the two datasets and explain how you generated them so that their shapes are different.

Given the above specifications on mean vector and marginal covariances, a sample can be generated from drawing from the following Gaussian distribution:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \mathcal{N}(\begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 & \sigma^2_{x_1,x_2} \\ \sigma^2_{x_2,x_1} & 0.5 \end{bmatrix}) \tag{7}$$

Where $\sigma^2_{x_1,x_2}$ and $\sigma^2_{x_2,x_1}$ can be modified to change the shape of the distribution (provided the covariance matrix is positive semi-definite). The sample plotted in figure 1 is drawn from the following distribution:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \mathcal{N}(\begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix}) = G_1 \tag{8}$$

This can be contrasted with figure 2, drawn from the distribution:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \mathcal{N}(\begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{bmatrix}) = G_2 \tag{9}$$

Both of these samples are plotted below. For figure 1 below, the 100 samples drawn from G_1 appear as blue dots on the right-hand plot. They have been plotted on top of the contour plot of G_1, and alongside the mesh plot of G_1, both visualizing the probability density of G_1. G_2 has likewise been plotted below in figure 2.

Inspection of of the contour plots and sample plots in figures 1 and 2 show that setting the cross covariance terms to -0.5 has the effect of tilting the axis of the distribution (or "squish" the distribution along the axis $x_1 = 2x_2$). This is because the cross covariance terms dictate the correlation between x_1 and x_2, so setting this to -0.5 creates a negative relationship between x_1 and x_2 in G_2.

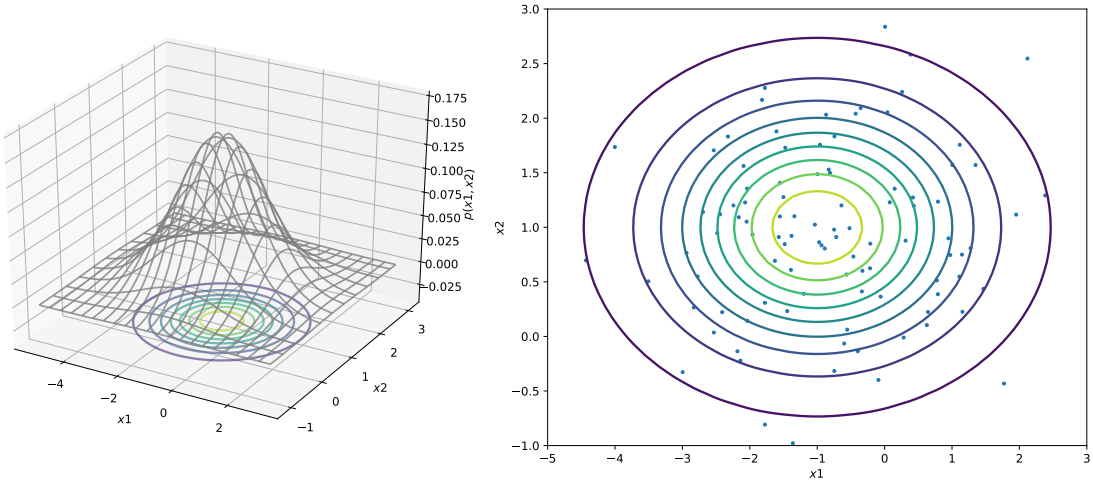

Figure 1: Pdf and contour plots for G_1 where $\sigma^2_{x_1,x_2} = \sigma^2_{x_2,x_1} = 0$

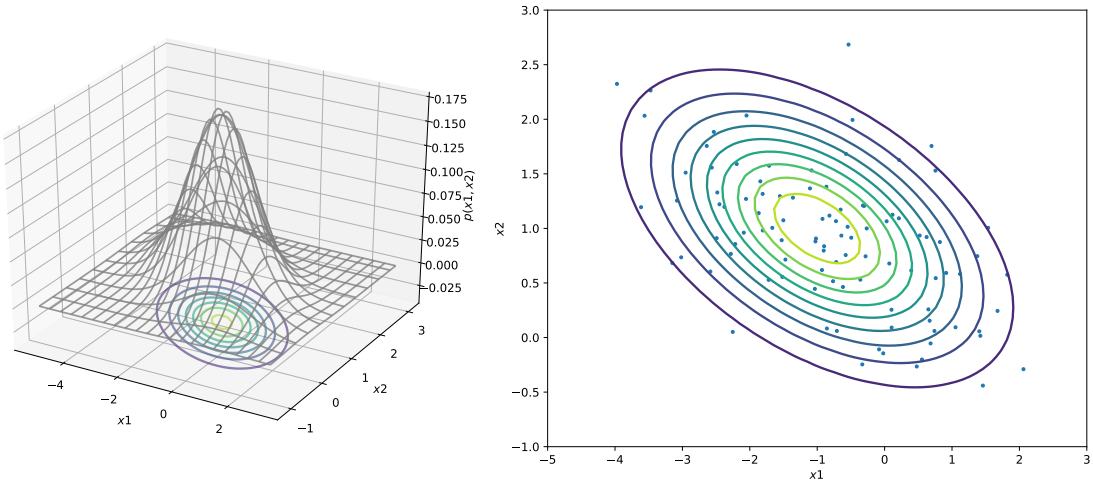

Figure 2: Pdf and contour plots for G_2 where $\sigma^2_{x_1,x_2} = \sigma^2_{x_2,x_1} = -0.5$

1.3

[27 marks] Nora and Noah spent the summer on writing a computer program that solves AI. However, they encounter the problem that their code seems to be failing randomly when compiling. Nora and Noah want to estimate the probability of successful compilation using a probabilistic model. They assume that when compiling the code N times (without any changes to the code) gives i.i.d. results. Furthermore, the probability of success can be described by a Bernoulli distribution with an unknown parameter μ. As good Bayesians, they place a conjugate Beta prior on this unknown parameter, where the parameters of this beta prior are $\alpha = 2, \beta = 2$. They have now run $N = 20$ experiments, and 6 of them successfully compiles, and 14 failed.

- **Compute the posterior distribution on μ (derive your result) and plot it.**

The prior follows a Beta$(2,2)$ distribution:

$$p(\mu|\alpha,\beta) = p(\mu|2,2) = \frac{\Gamma(2+2)}{\Gamma(2)\Gamma(2)}\mu^{2-1}(1-\mu)^{2-1} = 6\mu(1-\mu) \tag{10}$$

The likelihood can be modelled as a Binomial distribution given that it is multiple trials of a Bernoulli random variable. Hence the likelihood is:

$$p(x|N,\mu) = p(6|20,\mu) = \binom{20}{6}\mu^6(1-\mu)^{14} \tag{11}$$

By Bayes' theorem, we can compute the posterior as being proportional to the prior and the likelihood:

$$posterior = \frac{prior * likelihood}{evidence} \propto prior * likelihood \tag{12}$$

Hence the posterior can be drived as follows:

$$posterior \propto 6\mu(1-\mu) * \binom{20}{6}\mu^6(1-\mu)^{14} = 6\binom{20}{6}\mu^7(1-\mu)^{15} \tag{13}$$

Given that the final expression of equation 13 is (proportional to) a Beta distribution, we derive the posterior to be a $Beta(8,16)$ distribution. Thus:

$$posterior = p(\mu|x,N,\alpha,\beta) = Beta(\alpha+x,\beta+N-x) = Beta(8,16) \tag{14}$$

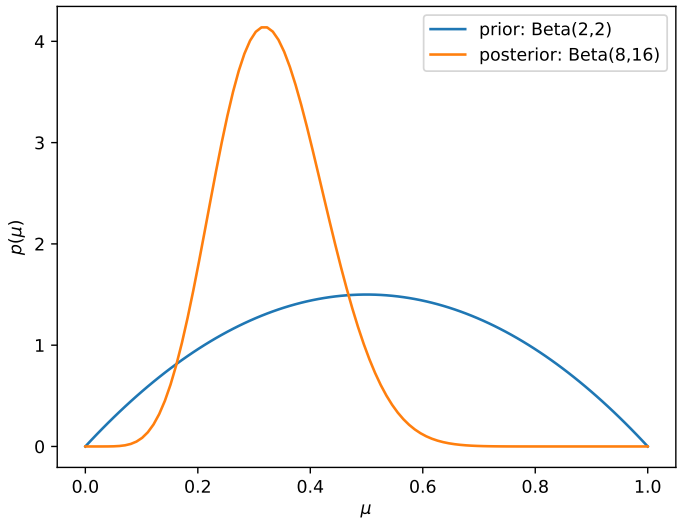

Figure 3: Pdf and contour plots for G_1 where $\sigma^2_{x_1,x_2} = \sigma^2_{x_2,x_1} = 0$

Figure 3 plots the prior and the posterior, with distributions Beta$(2,2)$ and Beta$(8,16)$ respectively.

- **What has changed from the prior to the posterior? Describe properties of the prior and the posterior.**

As shown in figure 3, the shape of the posterior (with $\alpha = 8, \beta = 16$) is different to the the shape of the prior (with $\alpha = 2, \beta = 2$). While the prior was symmetric and centered on the [0,1] interval with $mean = median = mode = 0.5$, these properties no longer hold in the posterior. By inspection, the posterior is a narrower distribution and taller distribution.

For a Beta distribution, it is straightforward to compare the expected value, variance and mode as these all have closed-form solutions. Expected value is:

$$\mathbb{E}[\mu] = \frac{\alpha}{\alpha + \beta} \tag{15}$$

Computing this, we find the expected value has fallen from 0.5 to 1/3. This is expected as the experiment showed a success rate of 30% so the posterior expected value would be in the interval $(0.3, 0.5)$. We can also compute the variance of a Beta distribution as follows:

$$\text{Var}[\mu] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{16}$$

We find the variance falls from 0.05 to $1/600 = 0.0017$. Again, this is expected as we have conducted an experiment on μ thereby reducing total uncertainty, which is refelcted in a lower variance. We can also compute the mode of the distribution by finding the value μ which maximizes $p(\mu|x, N, \alpha, \beta) = Beta(8, 16)$; i.e. by taking the derivative of the pdf and setting it equal to zero. Doing so yields the following expression for the mode:

$$\text{Mode}[\mu] = \frac{\alpha - 1}{\alpha + \beta - 2} \tag{17}$$

The mode for the prior is 0.5, and the mode for the posterior is 0.32. As expected, the mode has also fallen. In addition, we see that in the posterior the mode is now less than the mean, so the posterior is positively skewed.

To summarize, running the experiment has changed the distribution on μ by reducing the mean, variance and the mode, and has introduced a small positive skew.

2 Graphical Models

2.1

[16 marks] Given a factorized joint distribution, draw the corresponding directed graphical model (you can scan in a picture or use the tikz-bayesnet)

$$p(a,b,c,d,e,f) = p(a|b,c)p(c|b)p(d)p(e|d,a)p(f|c,d,e)p(b) \tag{18}$$

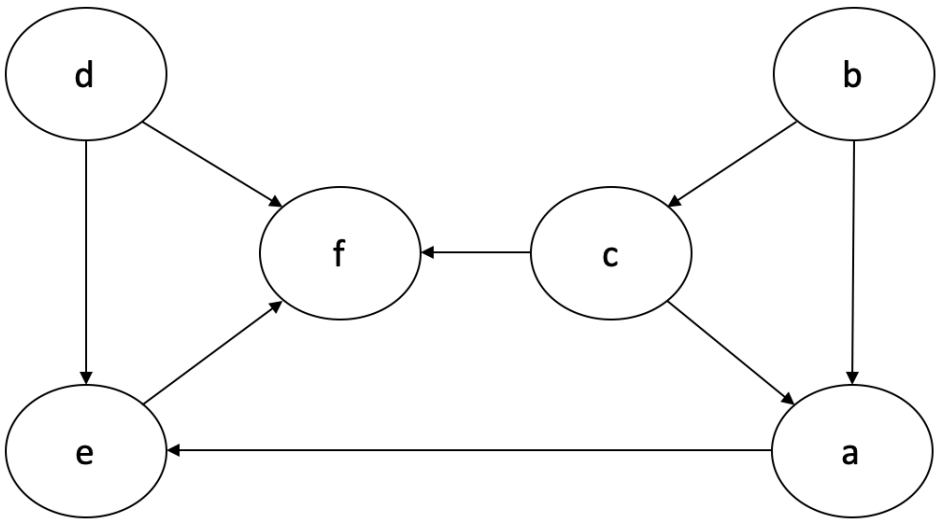

Figure 4: Graphical model for p(a, b, c, d, e, f) $p(a$

2.2

[40 marks] Determine whether the random variables in the graphical are conditionally independent.

a) Conditionally independent: e blocks the path as arrows meet head to head at e and neither e nor its descendants are observed

b) Not conditionally independent: $h - i - d$; arrows meet head to tail at j and j is unobserved

c) Conditionally independent: e blocks the path as arrows meet head to head at e (and e's descendants) are unobserved

d) Not conditionally independent: $j - d - e$; arrows meet head to tail at d and d is unobserved

e) Not conditionally independent: $b - d - e$; arrows meet head to tail at d and d isn't observed

f) Not conditionally independent: $j - h - i - c$; arrows meet head to head at h but k is a descendent of h so h is unblocked. arrows meet tail to tail at i and i unobserved so is also unblocked

g) Not conditionally independent: $a - b - d - k$; all paths are head to tail and no set is observed so none are blocked

h) Not conditionally independent: $a - b - d - k$; all paths are head to tail and e is not on the path so path is unblocked

i) Not conditionally independent: $h - i - d$; arrows tail to tail at i but i isn't observed

j) Not conditionally independent: $b - j - h$; arrows meet head to head at j but e is a descendant of j so j is unblocked

k) Not conditionally independent: $h - i - c$; arrows head to tail at i but i isn't observed

l) Not conditionally independent: $a - b - j - f$; arrows head to tail at b and b isn't observed; arrows are head to head at at j but k is a descendant of j so j is unblocked

m) Not conditionally independent: $i - h - j - b - a$; h is unblocked as arrows arrive head to head but j is a descendant of h; j is unblocked as arrows are head to head but j is observed; arrows are head to tail at b and b is unobserved to b is unobserved

n) Not conditionally independent: $h - j - d - e - g$; arrows are head to tail at j and d and neither are observed so both are unblocked; arrows are head to head at e and e is observed so unblocked

o) Conditionally independent: e blocks the path (arrows are head to head at e and e is unobserved with no descendants)

p) Conditionally independent: same as above

q) Conditionally independent: j and d blocks the paths (j blocks $a - h$ as arrows are head to head at j but j and descendants are unobserved; d blocks $a - i$ as arrows are head to head at d and d and its descendants are unobserved)

r) Conditionally independent: j and d block paths (j blocks the direct path from $b \rightarrow h$ as arrows are head to head at j and and nothing is observed; d blocks $d - d - i - h$ as arrows are head to head at d but nothing is observed

s) Conditionally independent: j and d block paths (same reason as question r) as g is not a descendant of j or d)

t) Conditionally independent: j and d block paths (j blocks $a-b-j-h-i$ as arrows are head to head at j and nothing is observed; d blocks the remaining paths as arrows are head to head at d and nothing is observed)