


Homework #No. 1 (Probability, Bayesian Linear Regression)

For questions, please refer to Moodle.
Released on 25/09/2023

GENERAL INSTRUCTIONS

- Submission of solutions is not mandatory but solving the exercises are highly recommended. The master solution will be released after the exercise deadline.
- Part of the exercises are available on Moodle as a quiz. These problems are marked with .

Exercise 0: Warm-Up

Four points are chosen uniformly at random on the surface of a sphere. What is the probability that the center of the sphere lies inside the tetrahedron whose vertices are at the four points?

Solution:

Having placed 3 points A, B, C , the 4th D will enclose the center in the tetrahedron if and only if it lies in the spherical triangle $A'B'C'$, where P' is directly opposite to P (so that the center lies on PP'). The probability of this is the area of ABC divided by the area of the sphere. So taking the area of the sphere as 1, we want to find the expected area of ABC . But the 8 triangles $ABC, A'BC, AB'C, ABC', A'B'C, AB'C', A'BC', A'B'C'$ are all equally likely and between them partition the surface of the sphere. So the expected area of ABC , and hence the required probability, is just $\frac{1}{8}$.

For a more formal method, assume the sphere is centered at the origin, and that the first point P_0 is located at the north pole of the sphere, with the three remaining points then located at random locations on the sphere. We can assume that these remaining points are chosen in this format:

First a diameter $P_{i_1}, P_{i_2}, i \in \{1, 2, 3\}$ is fixed and then one of the two end-points $\{P_{i_1}, P_{i_2}\}$ is selected as a vertex of the tetrahedron. The eight possible tetrahedra

$$P_0, P_{j_1}, P_{j_2}, P_{j_3}$$

such that j is equal to 1 or 2 are equally likely. Further, we can assume that the result is an honest tetrahedron and that the origin does not lie on any face. (Recall that the plane through three noncollinear points P_1, P_2, P_3 consists of all affine combinations

$$\alpha_1 P_1 + \alpha_2 P_2 + \alpha_3 P_3$$

Such that

$$\alpha_1 + \alpha_2 + \alpha_3 = 1$$

With probability one, neither the fourth vertex nor the origin lies in the plane through any three vertices.)

In particular, the four vertex vectors

$$\vec{P}_0, \vec{P}_{11}, \vec{P}_{21}, \vec{P}_{31}$$

must be linearly dependent, so there exists a non-zero 4-tuple (w, x, y, z) for which

$$w\vec{P}_0 + x\vec{P}_{11} + y\vec{P}_{21} + z\vec{P}_{31} = \vec{0}$$

Then since

$$\vec{P}_{i1} = -\vec{P}_{i2}$$

The equation have the solution of

$$(w, x, y, z), (w, x, y, -z), (w, x, -y, z), (w, -x, y, z), \\ (w, x, -y, -z), (w, -x, -y, z), (w, -x, y, -z), (w, -x, -y, -z).$$

Each point in the tetrahedron with vertices

$$P_0, P_{1j_1}, P_{2j_2}, P_{3j_3}$$

can be uniquely represented as a convex combination

$$\beta_0 P_0 + \beta_1 \vec{P}_{1j_1} + \beta_2 \vec{P}_{2j_2} + \beta_3 \vec{P}_{3j_3}$$

where

$$\beta_i \geq 0, \beta_1 + \beta_2 + \beta_3 + \beta_4 = 0$$

Hence the origin is contained in the tetrahedron P

$$P_0 P_{1j_1} P_{2j_2} P_{3j_3}$$

if and only if the 4-tuple solving the associated vector equation

$$w \vec{P}_0 + x \vec{P}_{11} + y \vec{P}_{21} + z \vec{P}_{31} = \vec{0}$$

consists of four coordinates of the same sign. Since only one of the above eight solutions has this property, only one of the eight equally likely tetrahedra contains the origin, and hence the probability that the origin is contained in the randomly chosen tetrahedron is $\frac{1}{8}$.

Exercise 1: Bayes Rule

Imagine you have installed a sophisticated alarm system in your home. In case of a thief, it reacts with a 100% probability and sends you a notification. However, an alarm can also be triggered by an earthquake in 10% of the cases. Let's introduce the following random variables: Alarm (A) to represent whether the alarm at your home is triggered (1 for activated, 0 for not activated); Theft (T) to indicate whether theft is occurring (1 for theft, 0 for no theft); Earthquake (E) to represent whether an earthquake is happening (1 for earthquake, 0 for no earthquake). For your neighborhood, the probability of theft is $p(T = 1) = 2 \times 10^{-4}$ and of earthquake is $p(T = 1) = 10^{-2}$. Conditional probabilities of alarm $p(A = 1|T, E)$ can be summarised by

	E=0	E=1
T=0	0	0.1
T=1	1	1

So the full probabilistic model is $p(A, T, E) = P(A|T, E)P(T)P(E)$

- (a) You are busy at university and receive a notification from the alarm system. How would you react? Compute the probability of $p(T = 1|A = 1)$.

Solution: Let us compute this probability using Bayes theorem

$$p(T = 1|A = 1) = \frac{p(A = 1|T = 1)p(T = 1)}{\sum_T p(A = 1|T)p(T)} \\ P(A = 1|T) = \sum_E p(A = 1|T, E)p(E).$$

This results in

$$P(A = 1|T = 1) = 1 \sum_E p(E) = 1$$

$$\begin{aligned} P(A = 1|T = 0) &= p(A = 1|T = 0, E = 0)p(E = 0) + p(A = 1|T = 0, E = 1)p(E = 1) \\ &= 0 + 0.1 \times 10^{-2} = 10^{-3} \end{aligned}$$

Hence,

$$p(T = 1|A = 1) = \frac{1 \times 2 \times 10^{-4}}{1 \times 2 \times 10^{-4} + 10^{-3}(1 - 2 \times 10^{-4})} \approx \frac{2}{12} \approx 17\%.$$

- (b) How much would the $p(T = 1|A = 1)$ be changed if prior probability $p(T = 1)$ is 10 times higher in your neighborhood?

Solution: $p(T = 1|A = 1) \approx 67\%$

- (c) Your trusty radio (variable R) sometimes broadcasts information about earthquakes in the area (1 for radio reporting, 0 for no radio reporting). Let's assume $P(R = 1|E = 1) = 0.5$, while $P(R = 1|E = 0) = 0$. Assume the following joint probability factorization $p(A, T, E, R) = P(A|T, E)P(R|E)P(T)P(E)$. How would your posterior estimate of the probability of theft change if you heard a radio broadcast about an earthquake together with receiving a notification? Compute the probability of $p(T = 1|A = 1, R = 1)$.

Solution: Using marginalisation and conditioning, from the joint distribution $p(x, y, z)$, we can compute:

$$p(x|y) = \frac{p(x, y)}{p(y)} = \frac{\int p(x, y, z) dz}{\int \int p(x, y, z) dz dx}$$

For our problem formulation:

$$\begin{aligned} p(T = 1|A = 1, R = 1) &= \frac{p(T = 1, A = 1, R = 1)}{p(A = 1, R = 1)} \\ &= \frac{\sum_E P(A = 1|T = 1, E)P(R = 1|E)P(T = 1)P(E)}{\sum_{E, T} P(A = 1|T, E)P(R = 1|E)P(T = 1)P(E)} \approx \frac{1}{51} \approx 2\% \end{aligned}$$

Exercise 2: True/False Question

- (a) Let $X \sim \mathcal{N}(\mu, \Sigma)$ be a Gaussian random vector.

☒ Any affine transformation of X , such as $Y = M^T X + b$, is also Gaussian.

☒ True ☐ False

☒ All non-affine transformations of X are *not* Gaussian.

☐ True ☒ False

Solution: The first part is clear from the tutorial/lecture. For the second part, we bring a counterexample: we construct a non-affine transformation that keeps a Gaussian, Gaussian.

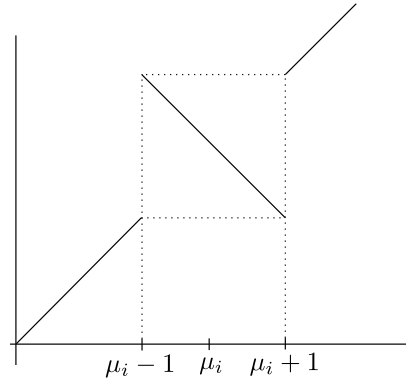
Let $X \in \mathbb{R}^d, X \sim \mathcal{N}(\mu, \Sigma)$. Define the transformation $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ as

$$\Phi \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \phi_i(x_i) \\ \vdots \\ x_d \end{pmatrix}, \quad (1)$$

which means that function Φ will only affect one coordinate and keeps all other coordinates unchanged. We define the function ϕ_i as

$$\phi_i(x_i) = \begin{cases} -(x_i - \mu_i) + \mu_i & |x_i - \mu_i| < 1 \\ x_i & \text{otherwise} \end{cases} \quad (2)$$

Intuitively, ϕ_i simply flips the all the “mass” in a neighborhood of μ_i in i th direction. Due to the symmetry of the Gaussian distribution around its mean, you could easily imagine that the transformation still preserves the Gaussian property. Also, this function cannot be an affine transformation since it is not continuous, see the figure below.



To prove this, one could use the change of variables formula from the tutorial. Let

$$Y = \Phi(X), \quad f_Y(y) = f_X(\Phi^{-1}(y)) \cdot |\det D\Phi^{-1}(y)|, \quad \forall y \in \mathbb{R}^d. \quad (3)$$

There are two cases for y :

1) if $|y_i - \mu_i| < 1$, then $x_i = \phi_i^{-1}(y_i) = -(y_i - \mu_i) + \mu_i$. Also note that in this case, the Jacobian matrix is simply a diagonal matrix with -1 in its i th position and 1 elsewhere:


$$D\Phi^{-1}(y) = \text{diag}(1, \dots, -1, \dots, 1),$$

hence, its determinant is -1 . We have

$$\begin{aligned} f_Y(y) &= f_X(\Phi^{-1}(y)) \cdot |\det D\Phi^{-1}(y)| \\ &= f_X(\Phi^{-1}(y)) = f_X \begin{pmatrix} y_1 \\ \vdots \\ -(y_i - \mu_i) + \mu_i \\ \vdots \\ y_d \end{pmatrix} = f_X \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_d \end{pmatrix} \quad f_X \text{ is symmetrical w.r.t. } \mu_i \\ &= f_X(y) \end{aligned}$$

2) if $|y_i - \mu_i| > 1$, $f_Y(y) = f_X(y)$ since Φ is the identity function in this case. Thus Y has an identical pdf as X , and is of course a Gaussian variable.

(b) Let X, Y, Z be independent standard normal random variables.

 The random variable $\frac{X+YZ}{\sqrt{1+Z^2}}$ is Gaussian.

■ True □ False


Solution: In this question, it is difficult to calculate distribution of $W := \frac{X+YZ}{\sqrt{1+Z^2}}$ since Z is in the denominator. The trick here is to first condition W on Z , and then integrate over distribution of Z . As a clarification, we inherit the notation definition from lecture slides, lowercase p means the probability density function. For $z \in \mathbb{R}$,

$$\begin{aligned} p(W|Z=z) &= p\left(\frac{X+YZ}{\sqrt{1+Z^2}} \mid Z=z\right) \\ &= p\left(\frac{X+Yz}{\sqrt{1+z^2}}\right) && X, Y \perp Z \\ &= \mathcal{N}(W; 0, 1) && \text{as } X + Yz \sim \mathcal{N}(0, 1 + z^2) \end{aligned}$$

which means that $P(W|Z=z)$ is always $\mathcal{N}(0, 1)$, and is *independent from the value of Z* .

Using total probability formula (or sum rule + product rule on slide 1, page 33), we have

$$\begin{aligned} p(W) &= \int p(W|Z=z) f_Z(z) dz && \text{law of total probability} \\ &= \mathcal{N}(W; 0, 1) \int f_Z(z) dz && p(W|Z) \text{ does not depend on } Z \\ &= \mathcal{N}(W; 0, 1) \end{aligned}$$

 Let $\alpha \sim \text{Unif}(0, 1)$ be independent of X and Y . Then $X \cos(\alpha) + Y \sin(\alpha) \sim \mathcal{N}(0, 1)$.


■ True □ False

Solution: Basically, we will use the same trick as used in the previous question. Let $Z = X \cos \alpha + Y \sin \alpha$, we have

$$\begin{aligned} p(Z|\alpha=a) &= p(X \cos \alpha + Y \sin \alpha | \alpha = a) \\ &= p(X \cos a + Y \sin a) && X, Y \perp \alpha \\ &= \mathcal{N}(Z; 0, \cos^2 a + \sin^2 a) \\ &= \mathcal{N}(Z; 0, 1). \end{aligned}$$

By integrating over α one obtains the result. Note that the distribution of α is not important at all, only its independence from X and Y is important.

(c) Let $X \sim \mathcal{N}(0, 1)$ and Z be ± 1 with equal probability and independent of X .


 The random variable $Y = ZX$ is standard Gaussian.

☒ True ☐ False

Solution:

$$\begin{aligned} p(Y) &= \sum_z p(ZX|Z=z) p(Z=z) && \text{law of total probability} \\ &= \frac{1}{2}p(X) + \frac{1}{2}p(-X) \end{aligned}$$

Since X is a standard Gaussian, the density at X is the same as the density at $-X$, hence $p(Y) = p(X)$.

 The vector $(X, Y) \in \mathbb{R}^2$ is a Gaussian random vector.

☐ True ☒ False

Solution: The answer is no. If it was a Gaussian random vector, any linear combination would be Gaussian. Now consider the linear combination $W = X + Y$. Notice that

$$\begin{aligned} \Pr\{W=0\} &= \Pr\{X=-Y\} \\ &= \Pr\{X=-Y|Z=1\} \Pr\{Z=1\} + \Pr\{X=-Y|Z=-1\} \Pr\{Z=-1\} \\ &= 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \\ &= \frac{1}{2}. \end{aligned}$$

Hence, W cannot be Gaussian, as it has a point mass of $\frac{1}{2}$ at 0.

Exercise 3: Bayesian Inference

A simple example of Bayesian inference is mean estimation. Assume we have a set of random variables (our data) X_1, \dots, X_N , each of which is distributed according to $\mathcal{N}(\mu, \sigma^2)$. Suppose that the variance σ^2 is a known constant available to us, and our goal is to estimate μ . To do things in a Bayesian way, we start off with a prior $p(\mu)$, which we assume for convenience to be Gaussian $\mathcal{N}(\mu_0, \sigma_0^2)$. Each of μ_0 and σ_0^2 can be endowed with a prior probability distribution again and this process can go on (this way of modelling things is called Hierarchical Bayesian Modeling). However, in our case, suppose that we have an idea about μ_0 and σ_0^2 and treat them as constants (this approach is called Empirical Bayes). Hence, the posterior distribution of μ can be written as

$$p(\mu | \{X_1, \dots, X_N\}, \sigma^2, \mu_0, \sigma_0^2).$$

(a) Write down the posterior density in terms of $\mu, X_1, \dots, X_N, \sigma^2, \mu_0, \sigma_0^2$.

Solution: The posterior, based on the Bayes rule, has the following form:

$$p(\mu | \{X_1, \dots, X_N\}, \sigma^2, \mu_0, \sigma_0^2) \propto \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_i - \mu)^2} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2}$$

- (b) Prove that this density is still a Gaussian. Compute the mean and variance of this distribution.

Solution: Let $\bar{X} = \frac{1}{N} \sum X_i$ be the sample mean. The strategy is to complete the squares (for μ). We have

$$\begin{aligned} & \sum_{i=1}^N \frac{1}{2\sigma^2} (X_i - \mu)^2 + \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \\ &= \left(\frac{N}{2\sigma^2} + \frac{1}{2\sigma_0^2} \right) \mu^2 - \left(\frac{1}{\sigma^2} \sum X_i + \frac{1}{\sigma_0^2} \mu_0 \right) \mu + C \\ &= \left(\frac{N}{2\sigma^2} + \frac{1}{2\sigma_0^2} \right) \mu^2 - \left(\frac{N}{\sigma^2} \bar{X} + \frac{1}{\sigma_0^2} \mu_0 \right) \mu + C \\ &= \frac{1}{2\tilde{\sigma}^2} (\mu - \tilde{\mu})^2 + C', \end{aligned}$$

where C, C' are constants not related to μ , and

$$\begin{aligned} \tilde{\sigma}^2 &= \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \right)^{-1} \\ \tilde{\mu} &= \left(\frac{N}{\sigma^2} \bar{X} + \frac{1}{\sigma_0^2} \mu_0 \right) \cdot \tilde{\sigma}^2 = \frac{N/\sigma^2}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} \bar{X} + \frac{1/\sigma_0^2}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} \mu_0. \end{aligned}$$

Hence,

$$p(\mu \mid \{X_1, \dots, X_N\}, \sigma^2, \mu_0, \sigma_0^2) \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2).$$

- (c) Compare the MAP and MLE for this problem. Can you explain what is the role of the prior? Specifically, can you observe the dependence on μ_0 and σ_0^2 and how they effect the MAP?

Solution:

The MLE for this problem is simply \bar{X} (check this), while the MAP estimate is $\tilde{\mu}$. Notice that this estimate depends on μ_0 and σ_0 and N in a clear way: the larger N gets, the less dependency on μ_0 and σ_0 will be (the prior gets forgotten). Moreover, one observes that for a higher value of σ_0 (a non-binding prior), the tendency to μ_0 also gets smaller. In the limit $\sigma_0 \rightarrow \infty$, one recovers the MLE solution.

Exercise 4: Multivariate Gaussian Distribution

A vector-valued random variable $x \in \mathbb{R}^d$ is said to have a multivariate normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbf{S}_{++}^d$ if its pdf is:

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

Prove the following facts, stated in the first lecture:

- (a) Every marginal of a Gaussian vector is Gaussian (slide [49]),

Solution: Consider $x = \begin{bmatrix} x_A \\ x_B \end{bmatrix}$, $\mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}$, and $\Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$, the joint distribution:

$$p(x) = p(x_A, x_B) = \frac{1}{Z} \exp \left(-\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right).$$

The following notations can ease the computation:

$$V = \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1}, \quad \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} = \begin{bmatrix} \Delta_A \\ \Delta_B \end{bmatrix}$$

$$\begin{aligned} p(x_A) &= \frac{1}{Z} \int_{x_B} \exp \left(-\frac{1}{2} \begin{bmatrix} \Delta_A \\ \Delta_B \end{bmatrix}^T \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} \begin{bmatrix} \Delta_A \\ \Delta_B \end{bmatrix} \right) dx_B = \\ &= \frac{1}{Z} \exp \left(-\frac{1}{2} \left[\Delta_A^T (V_{AA} - V_{AB} V_{BB}^{-1} V_{BA}) \Delta_A \right] \right) \cdot \\ &\quad \cdot \int_{x_B} \exp \left(-\frac{1}{2} \left[(\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A)^T V_{BB} (\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A) \right] \right) dx_B \end{aligned}$$

$$\begin{aligned} p(x_A) &= \frac{1}{Z_A} \exp \left(-\frac{1}{2} \left[\Delta_A^T (V_{AA} - V_{AB} V_{BB}^{-1} V_{BA}) \Delta_A \right] \right) = \\ &= \frac{1}{Z_A} \exp \left(-\frac{1}{2} \left[\Delta_A^T \Sigma_{AA}^{-1} \Delta_A \right] \right) \end{aligned}$$

Note: $\frac{1}{2} z^T A z + b^T z + c = \frac{1}{2} (z + A^{-1} b)^T A (z + A^{-1} b) + c - b^T A^{-1} b$

(b) Conditioning on a subset of variables of a joint Gaussian is Gaussian (slide [50]).

Solution:

$$\begin{aligned} p(x_B | x_A) &= \frac{p(x_A, x_B)}{p(x_A)} = \\ &= \frac{1}{Z'} \exp \left(-\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) = \\ &= \frac{1}{Z'} \exp \left(-\frac{1}{2} \left[\Delta_A^T (V_{AA} - V_{AB} V_{BB}^{-1} V_{BA}) \Delta_A \right] \right) \cdot \\ &\quad \cdot \exp \left(-\frac{1}{2} \left[(\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A)^T V_{BB} (\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A) \right] \right) = \\ &= \frac{1}{Z''} \exp \left(-\frac{1}{2} \left[(\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A)^T V_{BB} (\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A) \right] \right) \end{aligned}$$

$$x_B | x_A \sim \mathcal{N}(\underbrace{\mu_B - V_{BB}^{-1} V_{BA} (x_A - \mu_A)}_{=\mu_{B|A}}, \underbrace{\Sigma_{BB} - \Sigma_{BA} \Sigma_{AA}^{-1} \Sigma_{AB}}_{=\Sigma_{B|A} = V_{BB}^{-1}})$$

Recall the following fact about characteristic functions, which will help you to solve the next questions:
For a random vector $X \in \mathbb{R}^d$, define its characteristic function φ_X as

$$\varphi_X(t) = \mathbb{E}[\exp(it^\top X)], \quad \text{for all } t \in \mathbb{R}^d.$$

The characteristic function completely identifies a distribution. For a multivariate Normal distribution $\mathcal{N}(\mu, \Sigma)$, its characteristic function can be computed explicitly:

$$\varphi(t) = \exp(it^\top \mu - \frac{1}{2}t^\top \Sigma t).$$

- (c) [🔍] Let $X = (X_1, \dots, X_d)$ be a d -dimensional standard Gaussian random vector, that is, $X \sim \mathcal{N}_d(0, I)$. Define $Y = AX + \mu$, where A is a $d \times d$ matrix and $\mu \in \mathbb{R}^d$. What is the distribution of Y ?

Solution: Let us compute the characteristic function of Y . Define $s = A^\top t$. We have

$$\begin{aligned} \varphi_Y(t) &= \mathbb{E}[\exp(it^\top Y)] \\ &= \mathbb{E}[\exp(it^\top AX) \cdot \exp(it^\top \mu)] \\ &= \mathbb{E}[\exp(is^\top X)] \cdot \exp(it^\top \mu) \\ &= \varphi_X(s) \cdot \exp(it^\top \mu) \\ &= \exp(-\frac{1}{2}s^\top s + it^\top \mu) \\ &= \exp(it^\top \mu - \frac{1}{2}t^\top AA^\top t), \end{aligned}$$

which means that $Y \sim \mathcal{N}(\mu, AA^\top)$.

$$1. \mathcal{N}(\mu, A) \quad 2. \mathcal{N}(\mu, A^\top A) \quad 3. \mathcal{N}(\mu, A^2) \quad 4. \mathcal{N}(\mu, AA^\top)$$

- (d) [🔍] If B is an $r \times d$ matrix, what is the distribution of BY ?

$$1. \mathcal{N}(B\mu, BAA^\top B^\top) \quad 2. \mathcal{N}(B\mu, BAA^\top) \quad 3. \mathcal{N}(\mu, BAA^\top B^\top) \quad 4. \mathcal{N}(\mu, BAA^\top)$$

Solution: With the same argument as the previous question, one gets $BY \sim \mathcal{N}(B\mu, BAA^\top B^\top)$.

- (e) [🔍] Let $X = (X_1, X_2)$ be a bivariate Normal random variable with mean $\mu = (1, 1)$ and covariance matrix $\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$. What is the mean and the variance of the conditional distribution of $Y = X_1 + X_2$ given $Z = X_1 - X_2 = 0$?

Solution: The correct answer mean 2, variance $\frac{20}{3}$

First, take a look at the following facts:

Let A, B be events. The definition of conditional probability $\mathbb{P}(A | B)$ assumes that $\mathbb{P}(B) \neq 0$. So one essentially cannot condition on events of zero probability in the usual way. The following is a workaround to this issue.

Let X, Y be random variables with joint density f and joint CDF F . For $\varepsilon > 0$ and $x, y \in \mathbb{R}$, we compute

$$\begin{aligned}\mathbb{P}(X \leq x | Y \in [y, y + \varepsilon]) &= \frac{\mathbb{P}(X \leq x, Y \in [y, y + \varepsilon])}{\mathbb{P}(Y \in [y, y + \varepsilon])} \\ &= \frac{F(x, y + \varepsilon) - F(x, y)}{F_Y(y + \varepsilon) - F_Y(y)} \\ &= \frac{[F(x, y + \varepsilon) - F(x, y)] / \varepsilon}{[F_Y(y + \varepsilon) - F_Y(y)] / \varepsilon}.\end{aligned}$$

Now if $\varepsilon \rightarrow 0$, the right-hand side has the limit $\frac{\partial_y F(x, y)}{f_Y(y)}$, and the left-hand side can be regarded as $\mathbb{P}(X \leq x | Y = y)$. Taking derivative with respect to x gives the conditional density

$$f_{X|Y}(x | y) = \frac{f(x, y)}{f_Y(y)}.$$

One can use this density to compute probabilities like $\mathbb{P}(X \in A | Y = y) = \iint_A \frac{f(x, y)}{f_Y(y)} dx dy$.

We present two approaches for this exercise:

APPROACH 1. Note that $Z = 0$ implies $X_1 = X_2$. Furthermore, by the definition of Y , we have $X_1 = X_2 = Y/2$ given $Z = 0$. Hence, the marginal density of Y given $Z = 0$ is proportional to

$$f_{Y|Z}(y | 0) = \frac{f_{Y,Z}(y, 0)}{f_Z(0)} \propto f_{Y,Z}(y, 0) \propto f_X \left[\begin{pmatrix} y/2 \\ y/2 \end{pmatrix} \right].$$

The last equality is due to the fact that the linear map $(x_1, x_2) \mapsto (x_1 + x_2, x_1 - x_2)$ has constant determinant of -2 . Thus, by a change of variables formula, the density changes by a constant factor. We then have

$$\begin{aligned}f_X \left[\begin{pmatrix} y/2 \\ y/2 \end{pmatrix} \right] &\propto \exp \left(-\frac{1}{2} \begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix}^T \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix} \right) \\ &= \exp \left(-\frac{1}{2} \begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix}^T \frac{1}{5} \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix} \right) \\ &= \exp \left(-\frac{1}{2} \frac{(y - 2)^2}{\frac{20}{3}} \right).\end{aligned}$$

Clearly, the conditional distribution of Y given $Z = 0$ is hence Normal with mean 2 and variance $\frac{20}{3}$.

In this problem, we used the following trick which prevents a lot of computational headaches. If one is trying to derive the density of a random variable X at x , that is, $f_X(x)$, it is easier to neglect all *multiplicative* terms that do not include x . The reason is simply because $\int_{\mathbb{R}} f_X(x) dx = 1$.

Two important examples are single variable Normal random variables and multivariate Gaussian vectors. In the first case, following the trick above, we conclude that if a density function is of the form

$$f(x) \propto \exp(-ax^2 + bx)$$

for $a > 0$ and $b \in \mathbb{R}$, by completing the squares, we obtain

$$-ax^2 + bx = -a\left(x - \frac{b}{2a}\right)^2 + \frac{b^2}{4a},$$

and thus, by removing the terms that do not depend on x , we get

$$f(x) \propto \exp\left(-\frac{1}{2} \frac{\left(x - \frac{b}{2a}\right)^2}{1/2a}\right),$$

meaning that the distribution is a normal distribution with mean $\frac{b}{2a}$ and variance $1/2a$.

The situation for multivariate normal distribution is the same. One needs only to create a proper quadratic form in the exponent to get the familiar multivariate Gaussian density.

APPROACH 2. We define the random vector R as

$$R = \begin{pmatrix} Y \\ Z \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}}_{=A} X.$$

Notice that R is a linear transformation of a Gaussian vector, and by part (a), it is a Gaussian vector. Thus, we only need to compute its mean and covariance matrix. By linearity of expectation, the mean μ_R of R is

$$\mathbb{E}[R] = A \mathbb{E}[X] = A\mu = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

The covariance matrix Σ_R of R is also given by part (c):

$$\Sigma_R = A \Sigma A^\top = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 7 & 1 \\ 1 & 3 \end{pmatrix}$$

The conditional density of Y given $Z = 0$ is then given by

$$\begin{aligned} f_{Y|Z}(y | 0) &= \frac{f_{Y,Z}(y, 0)}{f_Z(0)} \propto f_{Y,Z}(y, 0) \\ &\propto \exp\left(-\frac{1}{2} \begin{pmatrix} y-2 \\ 0 \end{pmatrix}^\top \begin{pmatrix} 7 & 1 \\ 1 & 3 \end{pmatrix}^{-1} \begin{pmatrix} y-2 \\ 0 \end{pmatrix}\right) \\ &= \exp\left(-\frac{1}{2} \begin{pmatrix} y-2 \\ 0 \end{pmatrix}^\top \frac{1}{20} \begin{pmatrix} 3 & -1 \\ -1 & 7 \end{pmatrix} \begin{pmatrix} y-2 \\ 0 \end{pmatrix}\right) \\ &= \exp\left(-\frac{1}{2} \frac{(y-2)^2}{\frac{20}{3}}\right). \end{aligned}$$

Clearly, the conditional distribution of Y given $Z = 0$ is hence Normal with mean 2 and variance $\frac{20}{3}$.

Exercise 5: Online Bayesian Linear Regression

As you have seen in the lecture, BLR with Gaussian prior and Gaussian likelihood has a closed-form posterior, which is also Gaussian and its mean and covariance matrix can be written in terms of the observed data and the noise level σ_n . Concretely, if X is the data matrix and y is the vector of responses, the posterior has the form

$$p(w \mid X, y) = \mathcal{N}(w; (X^\top X + \sigma_n^2 I)^{-1} X^\top y, (\sigma_n^{-2} X^\top X + I)^{-1})$$

Instead of assuming that the whole data *is available offline* (as a data matrix X), we are now in a situation where the data *is coming one by one in an i.i.d. stream*. That is, at round $t \in \mathbb{N}$, a new datapoint (x_t, y_t) is observed, where $x_t \in \mathbb{R}^d$ is the feature vector and $y_t \in \mathbb{R}$ is the response.

Our goal is to keep track of the posterior of w at the end of each round. That is, letting $p_t(w) := p(w \mid \{(x_i, y_i)\}_{i=1}^t)$ to be the posterior of w after round t , we want to be able to compute p_t and the predictive distribution for a desired data point x^* .

To achieve this task, one idea is to construct the data matrix $X^{(t)} \in \mathbb{R}^{t \times d}$ and response vector $y^{(t)} \in \mathbb{R}^t$ and do the computations described in the lecture *every round*. This idea gives the correct posterior; however, it is computationally prohibitive: as larger and larger amounts of data is collected, the data matrix grows larger, and the computation complexity, as well as memory requirement grows with t .

- (a) Can you design an algorithm that *updates* the posterior (as opposed to recalculating it from scratch) in a smarter way? The requirement is that the memory should not grow as $O(t)$. As a hint, write down p_{t+1} recursively as a function of p_t and try to do the computations in a smart way.

Solution: As discussed in the lecture and reminded above,

$$p_t(w) = p(w \mid \{(x_i, y_i)\}_{i=1}^t) = \mathcal{N}(w; \bar{\mu}, \bar{\Sigma}),$$

with

$$\begin{aligned}\bar{\mu} &= (X^\top X + \sigma_n^2 I)^{-1} X^\top y \\ \bar{\Sigma} &= (\sigma_n^{-2} X^\top X + I)^{-1}\end{aligned}$$

The argument follows by just noting that

$$X^\top X = \sum_{i=1}^t x_i x_i^\top, \quad X^\top y = \sum_{i=1}^t y_i x_i.$$

This means that after observing the $(t+1)$ th data point, we have that

$$X_{\text{new}}^\top X_{\text{new}} = X^\top X + x_{t+1} x_{t+1}^\top,$$

and

$$X_{\text{new}}^\top y_{\text{new}} = X^\top y + y_{t+1} x_{t+1}.$$

Hence, by just keeping $X^\top X$ (which is a $d \times d$ matrix) and $X^\top y$ (which is a vector in \mathbb{R}^d), and updating as above, we do not need to keep the whole data in the memory.

- (b) If d is large, computing the inverse every round is very expensive. Can you use the recursive structure you found in the previous question to bring down the computational complexity of every round to $O(d^2)$?

Solution: First, remind Woodbury's identity for a symmetric invertible matrix A and a vector x (with compatible dimensions):

$$(A + xx^\top)^{-1} = A^{-1} - A^{-1}x(1 + x^\top A^{-1}x)^{-1}x^\top A^{-1} = A^{-1} - \frac{(A^{-1}x)(A^{-1}x)^\top}{1 + x^\top A^{-1}x}$$

If one knows the inverse of A , then the computation of the inverse of $(A + xx^\top)$ is of $O(d^2)$, which is much better than computing the inverse of $(A + xx^\top)$ from scratch.

As observed in the solution, one has to compute $(X^\top X + \sigma_n^2 I)^{-1}$ for finding $\bar{\mu}$ and $\bar{\Sigma}$. Using the identity above, we can write

$$(X_{\text{new}}^\top X_{\text{new}} + \sigma_n^2 I)^{-1} = (\underbrace{X^\top X + \sigma_n^2 I}_A + x_{t+1}x_{t+1}^\top)^{-1}.$$

where $A^{-1} = \sigma_n^2 \bar{\Sigma}$ has been computed in the last step.

The rest is basically plugging the matrices into the Woodbury's formula.