STAT520P Diagnostic Problem Set

Geoff Pleiss

The purpose of this problem set is to ensure that you feel comfortable with multivariate Gaussian distributions and their manipulations. (They come up a lot in Bayesian optimization.)

If you have a strong background in Bayesian statistics, this problem set should be fairly straightforward. (Hopefully you will learn a new derivation or two!) If you are new to Gaussian distributions, this problem set should build fluency that you will need for the class. If these problems feels extremely difficult, then you will likely find this course to be technically overwhelming.

A quick note on notation. Variable names should use the following convention:

- deterministic scalars will be represented by lowercase/non-bold letters (e.g. a, θ, etc.);
- deterministic vectors will be represented by lowercase/bold letters (e.g. $\boldsymbol{a}$, $\boldsymbol{\theta}$, etc.);
- deterministic matrices will be represented by uppercase/bold letters (e.g. $\boldsymbol{A}$, $\boldsymbol{\Theta}$, etc.); and
- all random variables—scalar, vector, or matrix—will be represented by uppercase/non-bold letters (e.g. A, Θ, etc.).

(For the rest of the course, we will often use the same notation for deterministic and random variables. However, I am differentiating them in this problem set for clarity.)

$p(Y = \boldsymbol{a})$ refers to the density of the random variable Y evaluated at $\boldsymbol{a}$. $\mathcal{N}(a; \mu, \sigma^2)$ refers to the function that evaluates the μ-mean σ^2-variance Gaussian density on the scalar $a \in \mathbb{R}$; i.e.

$$\mathcal{N}\left(a; \mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\tfrac{1}{2\sigma^2}(a-\mu)^2\right). \tag{1}$$

With a slight abuse of notation, $Y \sim \mathcal{N}(\mu, \sigma^2)$ should be read as "the random variable Y is Gaussian distributed with mean μ and variance σ^2"—i.e. $p(Y = a) = \mathcal{N}(a; \mu, \sigma^2)$. Analogous notation will be used for multivariate Gaussian distributions (but you will first have to derive the density!).

In this problem set, you will deriving properties of Gaussian distributions from first principles. You should solve all of these problems using only the following rules:

1. the sum rule—$p(Y = \boldsymbol{a}) = \int p(Y = \boldsymbol{a}, Z = \boldsymbol{b}) \mathrm{d}\boldsymbol{b}$;
2. the product rule—$p(Y = \boldsymbol{a}, Z = \boldsymbol{b}) = p(Y = \boldsymbol{a} \mid Z = \boldsymbol{b}) p(Z = \boldsymbol{b}) = p(Z = \boldsymbol{b} \mid Y = \boldsymbol{a}) p(Y = \boldsymbol{a})$; with $p(Y = \boldsymbol{a}, Z = \boldsymbol{b}) = p(Y = \boldsymbol{a}) p(Z = \boldsymbol{b})$ if and only if Y and Z are independent;
3. the change of variables formula—if $\boldsymbol{g}(\cdot)$ is a differentiable and bijective function, then
$$p(Y = \boldsymbol{a}) = \det\left(\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{a})\right) \, p(\boldsymbol{g}(Y) = \boldsymbol{g}(\boldsymbol{a})),$$
where $\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{a})$ is the Jacobian matrix of $\boldsymbol{g}$ evaluated at $\boldsymbol{a}$;
4. linearity of expectation—$\mathbb{E}[\boldsymbol{A}Y + \boldsymbol{B}Z + \boldsymbol{c}] = \boldsymbol{A}\mathbb{E}[Y] + \boldsymbol{B}\mathbb{E}[Z] + \boldsymbol{c}$;
5. the *univariate* Gaussian density (Eq. 1); and
6. any linear algebraic identities that you want.

1) The Univariate Linear Gaussian Identity

Consider the univariate Gaussian random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$.

1. Since $\mathcal{N}(a; \mu, \sigma^2)$ is a density, we have that

   $$\int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(a-\mu)^2} \mathrm{d}a = 1. \tag{2}$$

   Prove that $\mathbb{E}[Y - \mu] = 0$ and $\mathbb{E}[(Y - \mu)^2] = \sigma^2$ by differentiating both sides of Eq. (2).

2. Let $Y, Y' \sim \mathcal{N}(0, 1)$ be two i.i.d. standard Gaussian random variables. Write out the joint density $p(Y = (b - a), Y' = a)$ and simplify.

3. Using your answer above, prove that $\int_{-\infty}^{\infty} p(Y = (b - a)) p(Y' = a) \mathrm{d}a = (4\pi)^{-1/2} \exp(-\frac{1}{2^2} b^2)$. (Hint: you should be able to prove this in 4 lines by completing the square and using Eq. (2).)

4. Based on the previous result, what can you say about the distribution of the random variable $Z = Y + Y'$?

The previous result is a special case of the *linear Gaussian identity*, which is arguably the most powerful property of Gaussian distributions. More generally, if Y and Y' are independent Gaussian random variables with $Y \sim \mathcal{N}(\mu, \sigma^2)$ and $Y' \sim \mathcal{N}(\mu', \sigma'^2)$, then for any $a, b, c \in R$, we have

$$(aY + bY' + c) \sim \mathcal{N}\left(a\mu + b\mu' + c, \;\; a^2\sigma^2 + b^2\sigma'^2\right). \tag{3}$$

You can prove this result with the same techniques as above, but it requires more bookkeeping.

2) Multivariate Gaussian Random Variables

Definition: Let Y be a d-dimensional vector-valued random variable. Y is *multivariate Gaussian* if and only all linear combination of its entries are univariate Gaussian; i.e. for all $\boldsymbol{c} \in \mathbb{R}^d$, we have that $\boldsymbol{c}^\top Y \sim \mathcal{N}(\mu, \sigma^2)$ for some $\mu, \sigma \in \mathbb{R}$.

1. Let $U = \begin{bmatrix} U_1 & \ldots & U_d \end{bmatrix}$ be a random d-dimensional vector, where $U_1, \ldots, U_d$ are all i.i.d. standard Gaussian random variables. ($U_1, \ldots, U_d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.) Prove that U meets the definition of a multivariate Gaussian random variable.

2. Consider the random vector $Y = \boldsymbol{L}U + \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ and $\boldsymbol{L}$ are deterministic. Prove that Y also meets the definition for a multivariate Gaussian random variable, and compute its mean and covariance.

3. Let Z be a multivariate Gaussian random variable where $\mathbb{E}[Z] = \boldsymbol{\mu}$ and $\mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^\top] = \boldsymbol{L}\boldsymbol{L}^\top$. Prove that, for any $a \in R$ and $c \in \mathbb{R}^d$, we have that $p((\boldsymbol{c}^\top Z) = a) \; = \; p((\boldsymbol{c}^\top(\boldsymbol{L}^\top U + \boldsymbol{\mu})) = a)$. (Hint: use the fact that the density of a univariate normal distribution is determined by its mean and variance.)

The last fact, taken together with the Cramér-Wold theorem, implies that $p(Z = \boldsymbol{a}) = p((\boldsymbol{L}U + \mu) = \boldsymbol{a})$ for all $\boldsymbol{a} \in \mathbb{R}^d$. In other words, *two multivariate Gaussian random variables are equal in distribution if they share the same mean and covariance.* We will exploit this fact to derive a density for Z.

4. Write the joint density $p(U = \boldsymbol{a})$ as a matrix.
5. Assume that $\boldsymbol{L}$ is a square matrix, and define $\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^\top$. Using the change-of-variables formula, prove that the density of $\boldsymbol{L}U + \boldsymbol{\mu}$ is

$$\mathcal{N}\left(\boldsymbol{a}; \boldsymbol{\mu}, \boldsymbol{K}\right) := \tfrac{1}{|2\pi \boldsymbol{K}|^{1/2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{a} - \boldsymbol{\mu})^\top \boldsymbol{K}^{-1}(\boldsymbol{a} - \boldsymbol{\mu})\right). \tag{4}$$

These last two results demonstrate that if Z is multivariate Gaussian with mean μ and covariance $\boldsymbol{K}$, then the density of Z is given by Eq. (4). Moreover, we have also demonstrated that $\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^\top$, and therefore the covariance must be positive semi-definite.

6. Consider the following multivariate Gaussian, written in block matrix form:

$$\begin{bmatrix} Y \\ Y' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{K}' \\ \boldsymbol{K}'^\top & \boldsymbol{K}'' \end{bmatrix}\right),$$

   where Y is d-dimensional and Y' is d'-dimensional. Prove that if $\boldsymbol{K}' = \boldsymbol{0}$ then Y and Y' are independent Gaussian random variables.

7. Prove the following generalization of the linear Gaussian identity: if $Y \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^\top)$ and $Y' \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{L}'\boldsymbol{L}'^\top)$ are independent multivariate Gaussian random variables, then

$$p\left(\boldsymbol{A}Y + \boldsymbol{B}Y' + \boldsymbol{c}\right) \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\mu}' + \boldsymbol{c},\ \boldsymbol{A}\boldsymbol{L}\boldsymbol{L}^\top\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{L}'\boldsymbol{L}'^\top\boldsymbol{B}^\top). \tag{5}$$

   (Hint: you can prove this in 3-5 lines using the previous results and some clever linear algebra.)

---

3) Marginal and Conditional Distributions

Using results from the previous problems, answers to these sub-problems should each be about 1-5 lines long!

Consider the following multivariate Gaussian, written in block matrix form:

$$\begin{bmatrix} Y \\ Y' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{K}' \\ \boldsymbol{K}'^\top & \boldsymbol{K}'' \end{bmatrix}\right),$$

where Y is d-dimensional and Y' is d'-dimensional.

1. Without performing any integration, prove that the marginal density of Y is equal to

$$p(Y = \boldsymbol{a}) = \int p\left(\begin{bmatrix} Y \\ Y' \end{bmatrix} = \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{a}' \end{bmatrix}\right) \mathrm{d}\boldsymbol{a}' = \mathcal{N}(\boldsymbol{a}; \boldsymbol{\mu}, \boldsymbol{K}). \tag{6}$$

2. Define the random variable Z such that

$$\begin{bmatrix} Y \\ Z \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{K}'^{\top}\boldsymbol{K}^{-1} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} Y \\ Y' \end{bmatrix}.$$

   Prove that Y and Z are independent, and derive the distribution of Z. (If the matrix on the right hand side seems arbitrary for you, then remind yourself about Gaussian elimination.)

3. Combine the previous two results to show that

$$p\left(Y' = \boldsymbol{a}' \mid Y = \boldsymbol{a}\right) = \mathcal{N}\left(\boldsymbol{a}';\ \boldsymbol{K}'^{\top}\boldsymbol{K}^{-1}\boldsymbol{a},\ \boldsymbol{K}'' - \boldsymbol{K}'^{\top}\boldsymbol{K}^{-1}\boldsymbol{K}\right). \tag{7}$$

   (Hint: use the product rule, and the fact that Z is determined by Y and Y'.)

---

4) (Optional Bonus!) The Cholesky Factorization

In this problem, we're going to derive an algorithm for drawing a sample from a d-dimensional multivariate Gaussian distribution $Y = \mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$. We will assume that we have access only to univariate standard Gaussian samples $\epsilon_1, \ldots, \epsilon_d \sim \mathcal{N}(0, 1)$.[1]

Using the product rule, we can factorize the density of Y as:

$$p(Y = \boldsymbol{a}) = p(Y_1 = a_1) \times p(Y_2 = a_2 \mid Y_1 = a_1) \times p(Y_3 = a_3 \mid Y_2 = a_2, Y_1 = a_1) \times \ldots \times p(Y_d = a_d \mid Y_{-d} = \boldsymbol{a}_{-d}),$$

where Y_{-d} and $\boldsymbol{a}_{-d}$ refer to vectors that contain all but the d^{th} entry. This decomposition defines a sequential sampling procedure:

1. First, we will sample Y_1.
2. Then, after "observing" Y_1, we will sample $Y_2 \mid Y_1$.
3. After observing Y_2, we will sample $Y_3 \mid Y_2, Y_1$.
4. We will continue this process until we have sampled $Y_d \mid Y_{d-1}, \ldots, Y_1$.

1. We can "transform" ϵ_1 into Y_1 using a simple affine transformation $Y_1 = \ell_{11}\epsilon_1$ for some scalar ℓ_{11}. Write ℓ_{11} terms of k_{11}, where k_{ij} is the ij^{th} entry of $\boldsymbol{K}$. Explain your answer.

2. We can "transform" ϵ_1, ϵ_2 into Y_2 through an affine transformation $Y_2 = \ell_{21}\epsilon_1 + \ell_{22}\epsilon_2$ for some scalars ℓ_{21} and ℓ_{22}. Write ℓ_{21} and ℓ_{22} in terms of ℓ_{11}, k_{21}, k_{22}. Explain your answer.

3. Define $\boldsymbol{L}_2 = \begin{bmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{bmatrix}$. Without explicitly performing matrix multiplication, prove that

$$\boldsymbol{L}_2 \boldsymbol{L}_2^{\top} = \begin{bmatrix} k_{11} & k_{21}^{\top} \\ k_{21} & k_{22} \end{bmatrix}.$$

   (Hint: consider the second moment of the random variable $\boldsymbol{L}_2 \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$.)

[1] You can generate these via `numpy.random.randn`, `torch.randn`, etc.

4. Define $\boldsymbol{\epsilon}_{1:2} = \begin{bmatrix}\epsilon_1 & \epsilon_2\end{bmatrix}^\top$. We can "transform" $\boldsymbol{\epsilon}_{1:2}, \epsilon_3$ into Y_3 through the affine transformation

$$Y_3 = (\boldsymbol{\ell}_{3,1:2}^\top)\boldsymbol{\epsilon}_{1:2} + \ell_{33},$$

for some vector $\boldsymbol{\ell}_{3,1:2} \in \mathbb{R}^2$ and scalar ℓ_{33}. Write $\boldsymbol{\ell}_{3,1:2}$ and ℓ_{33} in terms of $\boldsymbol{L}_2$, k_{33}, and $\boldsymbol{k}_{3,1:2} := \begin{bmatrix}k_{31} & k_{32}\end{bmatrix}^\top$. Explain your answer.

5. Without any formal proof, generalize the steps above into a recursive algorithm to compute a lower-triangular matrix $\boldsymbol{L}_d$ such that

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_d \end{bmatrix} = \boldsymbol{L}_d \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_d \end{bmatrix}.$$

and $\boldsymbol{L}_d \boldsymbol{L}_d^\top = \boldsymbol{K}$. $\boldsymbol{L}_d$ is known as the Cholesky factorization of $\boldsymbol{K}$.

From the previous results, we have proven the following (remarkable) facts about multivariate Gaussian random variables:

1. any multivariate Gaussian random variable is a rotation/shift of independent Gaussian random variables,

2. affine transformations and linear combinations of Gaussians are Gaussian (Eq. 5),

3. multivariate Gaussian random variables are closed under marginalization (Eq. 6), and

4. multivariate Gaussian conditionals are Gaussian (Eq. 7).

Moreover, we have also derived their density, mean, and variance from first principles, and we have also found a connection between the Cholesky factorization and sequential sampling!