Introduction and learning objectives

Many widely used models, including diffusion models and variational autoencoders (VAEs), are built on properties of the multivariate Gaussian; for example, the formulas in Problem 2 give the distribution of the noisy data in a diffusion model. We cover VAEs in Week 11 and diffusion models in Week 12.

This tutorial works with the multivariate Gaussian from Lecture 2. We first find the distribution of some of its variables given the rest, then use the result for a Gaussian prior with a linear-Gaussian likelihood to derive the Bayesian linear regression posterior, which Lecture 2 stated without derivation. Both results are in this week's reading (PRML Sections 2.3.1 and 2.3.3).

By the end of the tutorial, you should be able to:

- Identify the mean and covariance of a Gaussian from the quadratic and linear terms in its exponent.
- Expand the exponent of a partitioned Gaussian in terms of the blocks of the precision matrix $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$.
- Derive the mean and covariance of the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ by matching these terms.
- Match Bayesian linear regression to the linear-Gaussian model and derive the posterior mean $\mathbf{m}_N$ and covariance $\mathbf{S}_N$ for a general Gaussian prior.

1 Multivariate Gaussian Distribution

Let $\mathbf{x} \in \mathbb{R}^D$. Recall that the multivariate Gaussian distribution has a probability density function of

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

where $\boldsymbol{\mu} \in \mathbb{R}^D$ is the mean vector and $\mathbf{\Sigma} \in \mathbb{R}^{D\times D}$ is the covariance matrix. We can partition these D variables into two groups $\mathbf{x}_a$ and $\mathbf{x}_b$ and talk about the conditional probability $p(\mathbf{x}_a|\mathbf{x}_b)$ and marginal probability $p(\mathbf{x}_a)$. The parameters of the Gaussian distribution are partitioned accordingly with

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{bmatrix}.$$

In many cases, we can greatly simplify some of the conditional and marginal expressions by introducing the precision matrix $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$ and its corresponding partitioning

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab} \\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb} \end{bmatrix}. \tag{1}$$

Note that the blocks of $\mathbf{\Lambda}$ are not simply the inverses of the corresponding blocks of $\mathbf{\Sigma}$ (for example: $\mathbf{\Lambda}_{aa} \neq \mathbf{\Sigma}_{aa}^{-1}$ in general).

Problem 1. Let us try to find the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$. Before that, note that the terms of $\mathcal{N}(\mathbf{x})$ that are outside of the exponent are independent of $\mathbf{x}$. These constant terms are there to normalize the distribution so that the integral of $\mathcal{N}(\mathbf{x})$ equates to 1. As such, we can identify the parameters of a Gaussian distribution solely based on the terms inside the exponent

$$\begin{aligned} z &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ &= -\frac{1}{2}\left(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right). \end{aligned}$$

Note that given a multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, the coefficients of the quadratic and linear terms in $\mathbf{x}$ inside the brackets are $\boldsymbol{\Sigma}^{-1}$ and $-2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ respectively. Using this observation, we can find the parameters of other multivariate Gaussian distributions by matching them with these coefficients.

Let us use this coefficient-matching trick to find the mean $\boldsymbol{\mu}_{a|b}$ and covariance $\boldsymbol{\Sigma}_{a|b}$ of the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$. By substituting $\mathbf{x}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ with their partitioned forms, show that

$$z = -\frac{1}{2}\left((\mathbf{x}_a-\boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) + 2(\mathbf{x}_a-\boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) + (\mathbf{x}_b-\boldsymbol{\mu}_b)^\top \boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)\right).$$

Using the above result, show that

$$\begin{aligned} \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b). \end{aligned}$$

Problem 2. Let us now look at another important property of the multivariate Gaussian distribution. Suppose the prior $p(\mathbf{x})$ and likelihood $p(\mathbf{y}|\mathbf{x})$ defined over variables $\mathbf{x}$ and $\mathbf{y}$ have probability densities

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}),$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

where the mean of $\mathbf{y}$ is affine in $\mathbf{x}$. Then the posterior $p(\mathbf{x}|\mathbf{y})$ and evidence $p(\mathbf{y})$ have probability densities

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^\top\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}),$$
$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top)$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1}$.

This result gives the posterior in Bayesian linear regression (Lecture 2). Take a Gaussian prior over the weights, $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$, and the likelihood $p(t|\mathbf{w}, \mathbf{x}) = \mathcal{N}(t|\mathbf{w}^\top\mathbf{x}, \beta^{-1})$; Lecture 2 used the special case $\mathbf{m}_0 = \mathbf{0}$, $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$. Stacking the targets $\mathbf{t} = [t_1, \ldots, t_N]^\top$, we can write the joint distribution as $p(\mathbf{t}|\mathbf{w}, \mathbf{X}) = \mathcal{N}(\mathbf{t}|\mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I})$. Show that the posterior $p(\mathbf{w}|\mathbf{t}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ where

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{X}^\top\mathbf{t}),$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X}.$$