Gaussian Processes – part 1

Gaussian pdf w/ mean μ & cov Σ.

$$f_X(x) = \left(\frac{1}{(2\pi)^k |\Sigma|}\right)^{1/2} \exp\left(-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right)$$

Q. Why do we care about Gaussians?

- Closed under linear operations:

$$X \sim \mathcal{N}(\mu, \Sigma), \quad Y = Ax + b$$

$$\Rightarrow \quad Y \sim \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$$

w/

$$\tilde{\mu} = A\mu + b \qquad ①$$

$$\tilde{\Sigma} = A\Sigma A^T \qquad ②$$

where ① follows by linearity of $\mathbb{E}$ & ② follows by computing the variance:

$$\text{cov}(y) = \mathbb{E}\left[(y-\tilde{\mu})(y-\tilde{\mu})^T\right] = \mathbb{E}\left[(Ax+b-(A\mu+b))(Ax+b-A\tilde{\mu}+b)^T\right]$$

$$= \mathbb{E}\left[A(x-\mu)(x-\mu)^T A^T\right] = A\,\text{cov}(x)\,A^T$$

- Central limit theorem
- Gaussian is "Maximum entropy" distribution w/ given mean/var.
- X, Y Jointly Gaussian & uncorrelated $\Rightarrow X \perp Y$
- Gauss-Markov theorem: Gaussian has really nice interpretation w/ least squares & plays nice w/ squared loss function in general

Before jumping into GPs, lets look at an important property of Gaussians: If two RVs X, Y are jointly Gaussian then the conditional distribution of $X \mid Y$ is Gaussian

<u>Thrm</u> Let X be a Gaussian RV w/ mean μ & cov Σ. Decompose $X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}$ then the distribution of $X_a \mid X_b$ is Gaussian. (well determ the mean and cov in the proof.)

<u>pf</u>

Partition as follows:

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

$$\Sigma^{-1} =: \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

Note: Σ is symmetric $\Rightarrow$ Σ_{aa}, Σ_{bb} symm & $\Sigma_{ab} = \Sigma_{ba}^T$

& Λ^{-1} symmetric

Tangent: If A symm then A^{-1} is symm. Follows from

$AA^{-1} = I \Rightarrow (AA^{-1})^T = I \Rightarrow (A^{-1})^T A^T = I \underset{\text{symm of } A}{=} (A^{-1})^T A = I$

$\Rightarrow (A^{-1})^T$ is inverse of A, but the inverse of A is uniquely A^{-1}.

The proof will follow by template matching. We show that the pdf of $x_a | x_b$ is of the form $c \exp\left(-(x_a-\mu_a)\tilde{\Sigma}(x-\mu_a)\right)$ which ⟹ it's Gaussian.

Using our partition,

$$-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu) = -\frac{1}{2}(x_a-\mu_a)^T\Lambda_{aa}(x_a-\mu_a)$$
$$+\frac{1}{2}(x_a-\mu_a)^T\Lambda_{ab}(x_b-\mu_b)$$
$$+\frac{1}{2}(x_b-\mu_b)^T\Lambda_{ba}(x_a-\mu_a) \qquad (3)$$
$$+\frac{1}{2}(x_b-\mu_b)^T\Lambda_{bb}(x_b-\mu_b)$$

$$(y_1 \;\; y_2)\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = (y_1 \;\; y_2)\begin{pmatrix} Ay_1 + By_2 \\ Cy_1 + Dy_2 \end{pmatrix}$$
$$= y_1 A y_1 + y_1 B y_2 + y_2 C y_1 + y_2 D y_2$$

Note, in general:

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu + \text{const} \qquad (4)$$

where we use Σ^{-1} is symmetric. Let's template match to the collect terms in (3) that are quadratic in x_a:

$$-\frac{1}{2}x_a^T\Lambda_{aa}x_a$$

Hence $\Sigma_{a|b} = \Lambda_{aa}$.

Now, collect terms (in ③) that are linear in x_a:

$$\frac{1}{2} x_a \Lambda_{aa} \mu_a + \frac{1}{2} \mu_a \Lambda_{aa} x_a + \frac{1}{2} x_a \Lambda_{ab} (x_b - \mu_b) + \frac{1}{2} (x_b - \mu_b) \Lambda_{ba} x_a$$

Use sym of Λ

$$= x_a \underbrace{\left(\Lambda_{aa} \mu_a + \Lambda_{ab} (x_b - \mu_b) \right)}_{= \Sigma_{a|b} \mu_{a|b} \text{ by } ④}$$

Know $\Sigma_{a|b} = \Lambda_{aa}$

$\Rightarrow$

$$\mu_{a|b} = \Sigma_{a|b} \left(\Lambda_{aa} \mu_a + \Lambda_{ab} (x_b - \mu_b) \right)$$

$$= \mu_a + \Lambda_{aa} (x_b - \mu_b)$$

Can also compute $\Sigma_{a|b}$ and $\mu_{a|b}$ in terms of Σ partition.
This is done using Schur complement. See Bishop book p. 87.

Get.

$$\mu_{a|b} = \mu_a + \Sigma_{ab} \Sigma_{bb}^{-1} (x_b - \mu_b)$$

(See end of these notes)

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba}$$

Tangent

Q. Is there a nice intuitive interpretation for how our estimate of where x_a should land (like, where's the mean) and how our confidence in how close it should land near the mean changes knowing x_b & how it changes w/ different cov matrices? e.g., what if we use a 2-d example w/ $\Sigma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$?

Ex:

Let $x \sim N(\mu, \Sigma)$

$$\Sigma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Then

$$\mu_{x_1 | x_2} = \mu_1 + \frac{a}{d}(x_2 - \mu_2)$$

$$\sigma_{x_1 | x_2} = a - \frac{bc}{d}$$

- Visualize this using cov

Cases to consider

- $\Sigma = I$
- Σ = diagonal — What does this mean about indep of x_1, x_2?
- How do we make the scatter plot show like ? (set $b<0$)
- What do a and d mean (var x_1, x_2) $b = \text{cov}(x_1, x_2)$. If we fix a, d, what's the valid range for b?

  ans: By Cauchy-Schwarz $b \in (-\sqrt{ad}, \sqrt{ad})$
- Note that knowing x_2 only changes $\mu_{x_1|x_2}$. The variance of the distribution follows from Σ alone.
- Take $|b|$ close to $\sqrt{ad}$ what happens to variance of conditional dist?

such complex distributions is that of probabilistic graphical models, which will form the subject of Chapter 8.

2.3.1 Conditional Gaussian distributions

An important property of the multivariate Gaussian distribution is that if two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian. Similarly, the marginal distribution of either set is also Gaussian.

Consider first the case of conditional distributions. Suppose $\mathbf{x}$ is a D-dimensional vector with Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and that we partition $\mathbf{x}$ into two disjoint subsets $\mathbf{x}_a$ and $\mathbf{x}_b$. Without loss of generality, we can take $\mathbf{x}_a$ to form the first M components of $\mathbf{x}$, with $\mathbf{x}_b$ comprising the remaining $D - M$ components, so that

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}. \tag{2.65}$$

We also define corresponding partitions of the mean vector $\boldsymbol{\mu}$ given by

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{2.66}$$

and of the covariance matrix $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}. \tag{2.67}$$

Note that the symmetry $\boldsymbol{\Sigma}^{\mathrm{T}} = \boldsymbol{\Sigma}$ of the covariance matrix implies that $\boldsymbol{\Sigma}_{aa}$ and $\boldsymbol{\Sigma}_{bb}$ are symmetric, while $\boldsymbol{\Sigma}_{ba} = \boldsymbol{\Sigma}_{ab}^{\mathrm{T}}$.

In many situations, it will be convenient to work with the inverse of the covariance matrix

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \tag{2.68}$$

which is known as the *precision matrix*. In fact, we shall see that some properties of Gaussian distributions are most naturally expressed in terms of the covariance, whereas others take a simpler form when viewed in terms of the precision. We therefore also introduce the partitioned form of the precision matrix

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \tag{2.69}$$

Exercise 2.22

corresponding to the partitioning (2.65) of the vector $\mathbf{x}$. Because the inverse of a symmetric matrix is also symmetric, we see that $\boldsymbol{\Lambda}_{aa}$ and $\boldsymbol{\Lambda}_{bb}$ are symmetric, while $\boldsymbol{\Lambda}_{ab}^{\mathrm{T}} = \boldsymbol{\Lambda}_{ba}$. It should be stressed at this point that, for instance, $\boldsymbol{\Lambda}_{aa}$ is not simply given by the inverse of $\boldsymbol{\Sigma}_{aa}$. In fact, we shall shortly examine the relation between the inverse of a partitioned matrix and the inverses of its partitions.

Let us begin by finding an expression for the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$. From the product rule of probability, we see that this conditional distribution can be evaluated from the joint distribution $p(\mathbf{x}) = p(\mathbf{x}_a, \mathbf{x}_b)$ simply by fixing $\mathbf{x}_b$ to the observed value and normalizing the resulting expression to obtain a valid probability distribution over $\mathbf{x}_a$. Instead of performing this normalization explicitly, we can obtain the solution more efficiently by considering the quadratic form in the exponent of the Gaussian distribution given by (2.44) and then reinstating the normalization coefficient at the end of the calculation. If we make use of the partitioning (2.65), (2.66), and (2.69), we obtain

$$\begin{aligned}
-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \\
-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\
-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b).
\end{aligned} \tag{2.70}$$

We see that as a function of $\mathbf{x}_a$, this is again a quadratic form, and hence the corresponding conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ will be Gaussian. Because this distribution is completely characterized by its mean and its covariance, our goal will be to identify expressions for the mean and covariance of $p(\mathbf{x}_a|\mathbf{x}_b)$ by inspection of (2.70).

This is an example of a rather common operation associated with Gaussian distributions, sometimes called 'completing the square', in which we are given a quadratic form defining the exponent terms in a Gaussian distribution, and we need to determine the corresponding mean and covariance. Such problems can be solved straightforwardly by noting that the exponent in a general Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const} \tag{2.71}$$

where 'const' denotes terms which are independent of $\mathbf{x}$, and we have made use of the symmetry of $\boldsymbol{\Sigma}$. Thus if we take our general quadratic form and express it in the form given by the right-hand side of (2.71), then we can immediately equate the matrix of coefficients entering the second order term in $\mathbf{x}$ to the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ and the coefficient of the linear term in $\mathbf{x}$ to $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, from which we can obtain $\boldsymbol{\mu}$.

Now let us apply this procedure to the conditional Gaussian distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ for which the quadratic form in the exponent is given by (2.70). We will denote the mean and covariance of this distribution by $\boldsymbol{\mu}_{a|b}$ and $\boldsymbol{\Sigma}_{a|b}$, respectively. Consider the functional dependence of (2.70) on $\mathbf{x}_a$ in which $\mathbf{x}_b$ is regarded as a constant. If we pick out all terms that are second order in $\mathbf{x}_a$, we have

$$-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a \tag{2.72}$$

from which we can immediately conclude that the covariance (inverse precision) of $p(\mathbf{x}_a|\mathbf{x}_b)$ is given by

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}. \tag{2.73}$$

Now consider all of the terms in (2.70) that are linear in $\mathbf{x}_a$

$$\mathbf{x}_a^{\mathrm{T}} \{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\} \tag{2.74}$$

where we have used $\boldsymbol{\Lambda}_{ba}^{\mathrm{T}} = \boldsymbol{\Lambda}_{ab}$. From our discussion of the general form (2.71), the coefficient of $\mathbf{x}_a$ in this expression must equal $\boldsymbol{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b}$ and hence

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \end{aligned} \tag{2.75}$$

where we have made use of (2.73).

The results (2.73) and (2.75) are expressed in terms of the partitioned precision matrix of the original joint distribution $p(\mathbf{x}_a, \mathbf{x}_b)$. We can also express these results in terms of the corresponding partitioned covariance matrix. To do this, we make use of the following identity for the inverse of a partitioned matrix

Exercise 2.24

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix} \tag{2.76}$$

where we have defined

$$\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}. \tag{2.77}$$

The quantity $\mathbf{M}^{-1}$ is known as the *Schur complement* of the matrix on the left-hand side of (2.76) with respect to the submatrix $\mathbf{D}$. Using the definition

$$\begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \tag{2.78}$$

and making use of (2.76), we have

$$\boldsymbol{\Lambda}_{aa} = (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1} \tag{2.79}$$

$$\boldsymbol{\Lambda}_{ab} = -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}. \tag{2.80}$$

From these we obtain the following expressions for the mean and covariance of the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \tag{2.81}$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}. \tag{2.82}$$

Comparing (2.73) and (2.82), we see that the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ takes a simpler form when expressed in terms of the partitioned precision matrix than when it is expressed in terms of the partitioned covariance matrix. Note that the mean of the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$, given by (2.81), is a linear function of $\mathbf{x}_b$ and that the covariance, given by (2.82), is independent of $\mathbf{x}_a$. This represents an example of a *linear-Gaussian* model.

Section 8.1.4