

# Gaussian Process Regression

Uri Shaham

October 11, 2022

## 1 Preliminary: Multivariate Gaussians

**Definition 1.1** (multivariate normal distribution). *A random vector  $X \in \mathbb{R}^d$  is said to have a multivariate normal distribution with mean vector  $\mu \in \mathbb{R}^d$  and covariance matrix  $\Sigma$  if  $\Sigma$  is positive definite and  $x$  has density*

$$p(X; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X - \mu)^T \Sigma^{-1}(X - \mu)\right).$$

We can partition the  $d$  variables to two sets,  $A$  and  $B$ . In this case we can write  $X = \begin{bmatrix} X_A \\ X_B \end{bmatrix}$ ,  $\mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}$ ,  $\Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$ .

**Proposition 1.2.** *The conditional density  $p(X_A|X_B) = \frac{p(X_A, X_B; \mu, \Sigma)}{\int_{X_B} p(X_A, X_B; \mu, \Sigma) dX_B}$  are also multivariate normal:*

$$p(X_A|X_B) \sim \mathcal{N}(\mu_A + \Sigma_{AB} \Sigma_{BB}^{-1}(X_B - \mu_B), \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA}).$$

*Proof.*

$$\begin{aligned} p(X_A|X_B) &= \frac{p(X_A, X_B; \mu, \Sigma)}{\int_{X_B} p(X_A, X_B; \mu, \Sigma) dX_B} \\ &= \frac{1}{\int_{X_B} p(X_A, X_B; \mu, \Sigma) dX_B} \exp\left(-\frac{1}{2}(X - \mu)^T \Sigma^{-1}(X - \mu)\right) \\ &= \frac{1}{Z} \exp\left(-\frac{1}{2} \begin{bmatrix} X_A - \mu_A \\ X_B - \mu_B \end{bmatrix}^T \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} \begin{bmatrix} X_A - \mu_A \\ X_B - \mu_B \end{bmatrix}\right), \end{aligned} \tag{1}$$

where  $Z$  does not depend on  $X_A$ , and  $\Sigma^{-1} = V = \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix}$ . Observe that

$$\begin{aligned} \begin{bmatrix} X_A - \mu_A \\ X_B - \mu_B \end{bmatrix}^T \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} \begin{bmatrix} X_A - \mu_A \\ X_B - \mu_B \end{bmatrix} &= (X_A - \mu_A)^T V_{AA} (X_A - \mu_A) \\ &\quad + (X_A - \mu_A)^T V_{AB} (X_B - \mu_B) \\ &\quad + (X_B - \mu_B)^T V_{BA} (X_A - \mu_A) \\ &\quad + (X_B - \mu_B)^T V_{BB} (X_B - \mu_B). \end{aligned}$$

Retaining only terms depending on  $X_A$ , and using the fact that  $V_{AB} = V_{BA}^T$  we can thus have

$$p(X_A|X_B) \propto \exp\left(-\frac{1}{2} [X_A^T V_{AA} X_A - 2X_A^T V_{AA} \mu_A + 2X_A^T V_{AB} X_B]\right),$$

where the term inside the exponential is  $X_A^T V_{AA} X_A - 2X_A^T (V_{AA} \mu_A - V_{AB} X_B)$ . Completing the squares, we can write this as

$$(X_A - (\mu_A - V_{AA}^{-1} V_{AB} X_B))^T V_{AA} (X_A - (\mu_A - V_{AA}^{-1} V_{AB} X_B)) + c,$$

where  $c$  is a constant not depending on  $X_A$ . From this, we deduce that  $p(X_A|X_B)$  is normal with mean  $\mu = \mu_A - V_{AA}^{-1} V_{AB} X_B$  and covariance  $V_{AA}^{-1}$ . Finally, we recall the form of the inverse of a block matrix to have  $V_{AA} = (\Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA})^{-1}$ , and  $V_{BA} = -(\Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA})^{-1} - \Sigma_{AB} \Sigma_{BB}^{-1}$ .  $\square$

## 2 Gaussian Processes

Gaussian processes are extension of multivariate Gaussians from vectors to functions.

**Definition 2.1** (Gaussian process). *A GP with mean function  $m(\cdot)$  and covariance function  $k(\cdot, \cdot)$  is a stochastic process  $\{Z_t : t \in \mathcal{T}\}$  such that for every finite collection  $t_1, \dots, t_n$  of indices, the vector  $(Z_{t_1}, \dots, Z_{t_n})^T$  has a multivariate normal distribution with mean vector  $\mu = (m(Z_{t_1}), \dots, m(Z_{t_n}))^T$  and covariance matrix  $K$  such that  $K_{ij} = k(Z_{t_i}, Z_{t_j})$ . We*

Since the covariance function has to be positive definite, it makes sense that  $k$  will be a kernel function. A typical choice for  $k$  is the RBF function

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right).$$

When we say that a function  $f$  is a sample drawn from a GP prior, we can think of  $f$  as a sample from a infinite dimensional multivariate normal vector, where each entry corresponds to an index  $t \in \mathcal{T}$ . That is,  $f = \{f(x_t) : t \in \mathcal{T}\}$ .

## 3 Gaussian Process Regression

GPR is a popular tool to quantify prediction uncertainty. Let  $\{(x_i, y_i)\}$ ,  $i = 1, \dots, n$  be a training set, drawn from some data distribution  $\mathcal{D}$ . where  $x_i \in \mathbb{R}^d$ , and  $y_i \in \mathbb{R}$ . We model the data by  $y_i = f(x_i) + \epsilon_i$ , where  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is some function, drawn from a GP prior with zero mean and covariance function  $k$ , and the  $\epsilon_i$ 's are iid samples from zero mean normal distribution with variance  $\sigma^2$ . Let  $\{(x_j^*)\}$ ,  $j = 1, \dots, m$  be a test set, also drawn from the  $x$ -marginal distribution induced from  $\mathcal{D}$ . In vector form we can write

$$\vec{y} = \vec{f} + \vec{\epsilon},$$

and

$$\vec{y}^* = \vec{f}^* + \vec{\epsilon}^*,$$

where  $\vec{y} = (y_1, \dots, y_n)^T$ ,  $\vec{f} = (f(x_1), \dots, f(x_n))^T$ , and so on. Similarly, we write  $X = (x_1^T, \dots, x_n^T)^T$ , i.e., an  $n \times d$  matrix.

### 3.1 Prediction

Since  $f$  is a sample from a Gaussian process prior with zero mean vector and covariance function  $k$ , it follows that given  $X, X^*$

$$\begin{bmatrix} \vec{f} \\ \vec{f}^* \end{bmatrix} \sim \mathcal{N} \left( \vec{0}, \begin{bmatrix} k(X, X) & k(X, X^*) \\ k(X^*, X) & k(X^*, X^*) \end{bmatrix} \right)$$

where  $k(X, X^*)_{ij} = k(x_i, x_j^*)$ , and so on. Since both  $\vec{f}$  and  $\vec{\epsilon}$  are Gaussians, it follows that so is  $\vec{y}$ , i.e., given  $X, X^*$

$$\begin{bmatrix} \vec{Y} \\ \vec{Y}^* \end{bmatrix} \sim \mathcal{N} \left( \vec{0}, \begin{bmatrix} k(X, X) + \sigma^2 I & k(X, X^*) \\ k(X^*, X) & k(X^*, X^*) + \sigma^2 I \end{bmatrix} \right)$$

Finally, we are interested in the predictive distribution

$$p(\vec{Y}^* | X, X^*, \vec{y}).$$

recalling proposition 1.2, this distribution is multivariate Gaussian

$$p(\vec{Y}^* | X, X^*, \vec{y}) = \mathcal{N}(\mu^*, \Sigma^*),$$

with

$$\mu^* = k(X, X^*)(k(X, X) + \sigma^2 I)^{-1} \vec{y},$$

and

$$\Sigma^* = k(X^*, X^*) + \sigma^2 I - k(X^*, X)((k(X, X) + \sigma^2 I)^{-1} k(X, X^*)).$$

And that's :)

In particular, this gives us a measure of uncertainty in the prediction of  $y_j^*$ , which is  $\Sigma_{jj}^*$ , as

$$Y_j^* \sim \mathcal{N}(\mu_j^*, \Sigma_{jj}^*).$$