Gaussian Processes

Advanced Statistical Inference

Simone Rossi

Multivariate Gaussian: Marginalization and Conditioning

Let $\boldsymbol{z} = (z_1, z_2, z_3)^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 4 & 2 & 0 \\ 2 & 3 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

Partition the vector as $\boldsymbol{z} = (\boldsymbol{z}_a^\top, z_b)^\top$ where $\boldsymbol{z}_a = (z_1, z_2)^\top$ and $z_b = z_3$, so that

$$\boldsymbol{\mu}_a = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \mu_b = 0, \quad \boldsymbol{\Sigma}_{aa} = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}, \quad \boldsymbol{\Sigma}_{ab} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \Sigma_{bb} = 2.$$

1. **Marginalization.** State the general rule for the marginal $p(\boldsymbol{z}_a)$ of a partitioned Gaussian. Apply it to compute $p(\boldsymbol{z}_a) = p(z_1, z_2)$ explicitly, giving its mean vector and covariance matrix. Analyze all the dimensions of the matrices and vectors involved to verify that they are consistent.

2. **Conditioning.** State the general formula for the conditional $p(\boldsymbol{z}_a \mid z_b)$:

$$p(\boldsymbol{z}_a \mid z_b) = \mathcal{N}\left(\underbrace{\boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\Sigma_{bb}^{-1}(z_b - \mu_b)}_{\boldsymbol{\mu}_{a|b}}, \underbrace{\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\Sigma_{bb}^{-1}\boldsymbol{\Sigma}_{ab}^\top}_{\boldsymbol{\Sigma}_{a|b}} \right).$$

   For the observation $z_b = 2$, compute numerically:

   - the conditional mean $\boldsymbol{\mu}_{a|b}$;
   - the conditional covariance $\boldsymbol{\Sigma}_{a|b}$.

3. **Interpretation.** Compare $\boldsymbol{\Sigma}_{aa}$ and $\boldsymbol{\Sigma}_{a|b}$. Which is "larger"? What does this tell you about the effect of observing z_b on our uncertainty about $\boldsymbol{z}_a$? Which component (z_1 or z_2) is affected more, and why?

From Parametric to Non-Parametric Models

1. In Bayesian linear regression we place a prior over weights $\boldsymbol{w}$ and obtain a posterior $p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X})$. Explain in words what it means to instead place a prior directly over *functions* $f : \mathcal{X} \to \mathbb{R}$.

2. A Gaussian process is defined as: a collection of random variables such that every finite subset has a joint Gaussian distribution. Let $f \sim \mathcal{GP}(m, k)$ where $m(\boldsymbol{x}) = 0$ and $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\right)$.

   - Write the joint distribution $p(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2))$ for two inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$.
   - What is the marginal distribution of $f(\boldsymbol{x}_1)$?

Kernel Functions

1. The squared exponential (RBF) kernel is

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\ell^2}\right).$$

   - What does the lengthscale ℓ control? What happens as $\ell \to 0$ and $\ell \to \infty$?
   - What does the signal variance σ_f^2 control?

2. For inputs $x_1 = 0$, $x_2 = 1$, $x_3 = 3$ with $\sigma_f^2 = 1$ and $\ell = 1$, compute the 3×3 kernel matrix $\boldsymbol{K}$ where $K_{ij} = k(x_i, x_j)$.

3. A valid kernel must be symmetric and positive semi-definite. Verify that the RBF kernel is symmetric. Why is positive semi-definiteness important for the covariance matrix of a GP?

4. Consider the linear kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}'$. What kind of functions does a GP with this kernel represent?

GP Regression (Exact Inference)

Consider a noise-corrupted observation model:

$$y_n = f(\boldsymbol{x}_n) + \varepsilon_n, \qquad \varepsilon_n \sim \mathcal{N}(0, \sigma_n^2),$$

with prior $f \sim \mathcal{GP}(0, k)$.

1. Write the joint distribution of the observations $\boldsymbol{y} = (y_1, \ldots, y_N)^\top$ and the latent function values $\boldsymbol{f} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N))^\top$.

2. The joint distribution of training outputs $\boldsymbol{y}$ and test function values $\boldsymbol{f}_\star = f(\boldsymbol{X}_\star)$ is Gaussian:

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{f}_\star \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} \boldsymbol{K} + \sigma_n^2 \boldsymbol{I} & \boldsymbol{K}_\star^\top \\ \boldsymbol{K}_\star & \boldsymbol{K}_{\star\star} \end{pmatrix}\right),$$

where $\boldsymbol{K} = k(\boldsymbol{X}, \boldsymbol{X})$, $\boldsymbol{K}_\star = k(\boldsymbol{X}_\star, \boldsymbol{X})$, $\boldsymbol{K}_{\star\star} = k(\boldsymbol{X}_\star, \boldsymbol{X}_\star)$. Derive the posterior predictive mean and covariance:

$$\mu_\star = \boldsymbol{K}_\star(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1}\boldsymbol{y}, \qquad \Sigma_\star = \boldsymbol{K}_{\star\star} - \boldsymbol{K}_\star(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1}\boldsymbol{K}_\star^\top.$$

3. Suppose we have a single training point $x_1 = 0$ with $y_1 = 1$, and we want to predict at $x_\star = 1$. Use the RBF kernel with $\sigma_f^2 = 1$, $\ell = 1$, and $\sigma_n^2 = 0.1$.

   - Compute $K = k(x_1, x_1)$, $k_\star = k(x_\star, x_1)$, $k_{\star\star} = k(x_\star, x_\star)$.
   - Compute the posterior predictive mean $\mu_\star$ and variance $\sigma_\star^2$.

4. **Two training points.** Now suppose we have two training points $x_1 = 0$, $y_1 = 1$ and $x_2 = 2$, $y_2 = -1$, and we want to predict at $x_\star = 1$. Use the RBF kernel with $\sigma_f^2 = 1$, $\ell = 1$, $\sigma_n^2 = 0.25$.

   - Build the 2×2 matrix $\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}$ and compute its inverse.
   - Compute the posterior predictive mean $\mu_\star$ and variance $\sigma_\star^2$ at $x_\star = 1$.

5. **Effect of noise.** Repeat the single-training-point prediction from the previous section ($x_1 = 0$, $y_1 = 1$, $x_\star = 1$, RBF with $\sigma_f^2 = 1$, $\ell = 1$) for two noise levels: $\sigma_n^2 = 0.01$ and $\sigma_n^2 = 4$.

   - Compute $\mu_\star$ and $\sigma_\star^2$ for each case.
   - Describe qualitatively how increasing σ_n^2 affects the posterior mean and variance at the test point, and give an intuitive explanation.

Hyperparameter Optimisation

1. The marginal likelihood (model evidence) for GP regression is

$$\log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = -\tfrac{1}{2}\boldsymbol{y}^\top(\boldsymbol{K}_\theta + \sigma_n^2 \boldsymbol{I})^{-1}\boldsymbol{y} - \tfrac{1}{2}\log|\boldsymbol{K}_\theta + \sigma_n^2 \boldsymbol{I}| - \tfrac{N}{2}\log 2\pi.$$

   Identify the three terms and give an intuitive explanation of the role each plays in selecting the kernel hyperparameters $\boldsymbol{\theta} = (\ell, \sigma_f^2, \sigma_n^2)$.

2. Why is maximising the marginal likelihood preferable to maximising the training data likelihood $p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{f})$ for choosing hyperparameters?

3. What is the computational complexity of exact GP regression in terms of the number of training points N? What bottleneck causes this? Name one strategy to reduce this cost.