Machine Learning and Neural Networks (MATH3431) Epiphany term, 2024

Handout 11: Gaussian process regression

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim. To introduce the Gaussian process regression as a kernel method.

Reading list & references:

(1) Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: Springer.
- Ch. 6.4 Gaussian process

(2) Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning (Vol. 1, p. 159). Cambridge, MA: MIT press.
- Chapter 2, Regression (supplementary)

(3) Roustant, O., Ginsbourger, D., & Deville, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Journal of statistical software, 51, 1-55.
- Supplementary material related to the the implementation of GP in R computing environment.

1. Intro and motivation

Note 1. As motivation for the Gaussian process regression, we "Kernelize" the standard Bayesian normal linear regression in the machine learning framework.

Note 2. Consider the predictive rule $h(x) = \eta(x)$, and cast it in a linear form $\eta(x) = (\psi(x))^\top w$ where $\psi(x) = (\psi_1(x), ..., \psi_d(x))$ is a vector of basis functions mapping the input space $\mathcal{X}$ into a feature space $\mathcal{F}$. Assume there is available a set of observables $\{z_i = (x_i, y_i)\}_{i=1}^n$. We associate the learning problem with the Bayesian linear regression model

(1.1)
$$\begin{cases} y_i|\psi(x_i), w, \sigma^2 & \overset{\text{ind}}{\sim} \mathrm{N}\left(\eta(x_i), \sigma^2\right),\ i = 1, ..., n \\ \eta(\cdot) & = (\psi(\cdot))^\top w \\ w & \sim \mathrm{N}(\mu_0, V_0) \end{cases} \quad \text{equiv.} \quad \begin{cases} y|\eta, \sigma^2 \sim \mathrm{N}\left(\eta, I\sigma^2\right) & \text{(sampl. distr.)} \\ \eta = \Psi w & \text{(linear model restr.)} \\ w \sim \mathrm{N}(\mu_0, V_0) & \text{(prior)} \end{cases}$$

where $[\Psi]_{i,j} = \psi_j(x_i)$.

Note 3. The marginal likelihood is

(1.2)
$$f(y) = \mathrm{N}\left(y|\Psi^\top \mu_0,\ \Psi V_0 \Psi^\top + I\sigma^2\right)$$

where $\mathrm{N}(y|\mu, \Sigma)$ denotes the pdf of the Normal distribution with mean μ,and covariance matrix Σ.

Note 4. [1]The predictive distribution of a new outcome y_* at a new input x_* given the observables $\{z_i = (x_i, y_i)\}_{i=1}^n$ is

$$f\left(y_* | x_*, \{(x_i, y_i)\}\right) = \mathrm{N}\left(\mu_*\left(x_*\right), \sigma_*^2\left(x_*\right)\right)$$

with

$$\mu_*\left(x_*\right) = \psi\left(x_*\right)^\top \mu_0 + \frac{1}{\sigma^2} \overbrace{\psi\left(x_*\right)^\top V \Psi}^{K(x_*, X)=} \left(\overbrace{\Psi^\top V \Psi}^{K(X,X)=} + \sigma^2 \right)^{-1} \left(\Psi^\top \mu_0 - y \right) \tag{1.3}$$

$$\sigma_*^2\left(x_*\right) = \left(\underbrace{\psi\left(x_*\right)^\top V \psi\left(x_*\right)}_{=K(x_*, x_*)} + \sigma^2 \right) - \underbrace{\psi\left(x_*\right)^\top V \Psi}_{=K(x_*, X)} \left(\underbrace{\Psi^\top V \Psi}_{=K(x,X)} + \sigma^2 \right)^{-1} \underbrace{\left(\psi\left(x_*\right)^\top V \Psi \right)^\top}_{=K(X, x_*)} \tag{1.4}$$

according to Proposition 44.

Note 5. In the prior part of (1.1), let's assume $\mu_0 = 0$ (arguably) denoting complete ignorance whether $\eta(\cdot)$ is positive or negative. By applying Kernel trick in (1.3) and (1.4), the feature space always enters in the form inner products. In fact we can define a kernel $K\left(x, x'\right) = \left\langle L\psi\left(x\right), L\psi\left(x'\right) \right\rangle = \psi\left(x\right)^\top V \psi\left(x'\right)$ where L is such that $V = L^\top L$, in terms of Section 4 in Handout 7: Kernel methods. We can denote $K\left(x_*, X\right) = \psi\left(x_*\right)^\top V \Psi$, and $K\left(x_*, x_*\right) = \psi\left(x_*\right)^\top V \psi\left(x_*\right)$.

2. The Gaussian process regression model

Definition 6. Gaussian process (GP) is a collection of random variables $\{f(x);\, x \in \mathcal{X}\}$, indexed by label x, where any finite collection of those variables has a multivariate normal distribution. It is fully specified by its mean and covariance functions. It is denoted as

$$f(\cdot) \sim \mathrm{GP}(\mu(\cdot), C(\cdot, \cdot))$$

with mean

$$\mu(x) := \mathrm{E}(f(x)),\ x \in \mathcal{X}$$

and covariance function

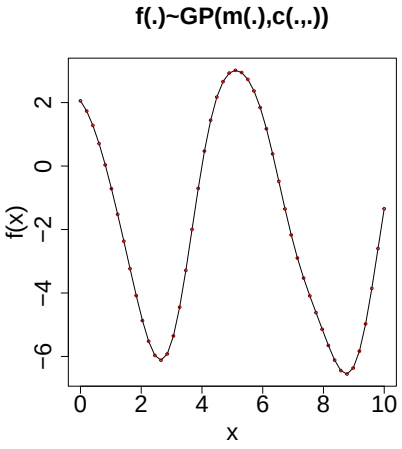
$$C(x, x') := \mathrm{Cov}(f(x), f(x')),\ x, x' \in \mathcal{X}$$

Note 7. Essentially, GP is a distribution defined over functions.

Example 8. One way to simulate a GP realization $f(\cdot) \sim \mathrm{GP}(\mu(\cdot), C(\cdot, \cdot))$, with $\mu(x) = 1 + 0.01x$, and $C(x, x') = \exp\left(-\frac{1}{2}\frac{|x - x'|^2}{1.5}\right)$ in the range $x \in [0, 10]$ is given in the following R code.

[1]No need to memorize the formulas in (1.2), (1.3), and (1.4). The material in Notes 2, 3, and 4 is given as a motivation for the Gaussian process regression.

```
rm(list=ls())
set.seed(99)
n <- 50 #descretize the problem
x <- seq(from = 0, to = 10, length = n)
mu_x <- matrix(1,n)
cov_x_x <- matrix(nrow = n, ncol = n)
for (i in 1:n) {
  mu_x[i] <- 1+0.01*x[i]
  for (j in 1:n) {
    cov_x_x[i,j]<- 10*exp(-0.5*(x[i]-x[j])^2/1.5)
  }
}
f <- rmvnorm(n = 1, mean = mu_x , sigma = cov_x_x)
plot(x,f,type="l",
  main="f(.)~GP(m(.),c(.,.))",
  ylab="f(x)",xlab="x",
  cex.axis=2, cex.lab=2, cex.main=2, cex.sub=2)
points(x, f, cex = .5, col = "dark red")
```

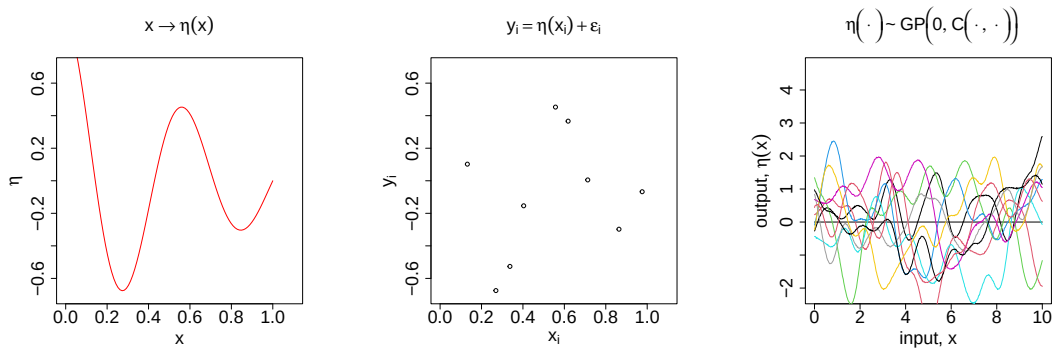


Note 9. Consider a function $\eta : \mathcal{X} \to \mathbb{R}$ with $\eta(x) = \langle \psi(x), w \rangle$ where $\psi(x)$ is a vector of known basis (feature) functions mapping from the input space $\mathcal{X}$ to the feature space $\mathcal{F}$, and $w \in \mathbb{R}^d$ is an unknown vector a priori following a normal distribution $w \sim \mathrm{N}(0, V)$, where the prior mean is set to zero denoting complete uncertainty about the sign of w's. Then the marginal $\eta(\cdot)$ follows a Normal distribution as a linear transformation of Normal variates with mean $\mathrm{E}(\eta(x)) = 0$ and covariance $\mathrm{Cov}(\eta(x), \eta(x')) = \psi(x)^\top V \psi(x')$ for any $x, x' \in \mathcal{X}$. Based on the Kernel trick and Definition 6, we can equivalently specify $\eta(\cdot) \sim \mathrm{GP}(0, C(\cdot,\cdot))$ for some corresponding kernel / covariance function $C(x, x') = \psi(x)^\top V \psi(x')$.

Note 10. We introduce the concept of Gaussian process regression in the machine learning framework below.

Note 11. Consider the predictive rule $h(x) = \eta(x)$, and assume that $\eta : \mathcal{X} \to \mathbb{R}$ with unknown formula (possibly up to a set of properties, we will discuss this later) and $\mathcal{X} \subseteq \mathbb{R}^d$.

Example 12. Figure 2.1a shows a function $\eta(\cdot)$. We pretend that we do not know $\eta(\cdot)$ but we wish to recover it. To recover $\eta(x)$, we collect training data set as in Figure 2.1b.



(A) (B) (C)

FIGURE 2.1. A toy example. (2.1a) shows the true $\eta(x)$. Fig 2.1b shows the training sample $\{z_i = (x_i, y_i)\}$ s.t. $y_i = \eta(x_i) + 0$. Fig 2.1c shows several realisations of the prior $\eta(\cdot) \sim \text{GP}(0, C(\cdot,\cdot))$. R code for the plots/running example is available from `https://github.com/georgios-stats/Machine_Learning_and_Neural_Networks_III_Epiphany_2024/tree/main/Lecture_handouts/code/10.Gaussian_process_regression/plots.R`

Note 13. For training purposes, assume there is available a set of observables $\{z_i = (x_i, y_i)\}_{i=1}^n$ whose sampling distribution is such that

$$y_i = \eta(x_i) + \epsilon_i,\ \epsilon_i \overset{\text{iid}}{\sim} \text{N}\left(0, \sigma^2\right),\ i = 1, ..., n$$

or equivalently

$$y_i | \eta(\cdot), \sigma^2 \overset{\text{iid}}{\sim} \text{N}\left(\eta(x_i), \sigma^2\right),\ i = 1, ..., n \tag{2.1}$$

for some unknown $\sigma^2 > 0$. Statistical model (2.1) can result by considering a quadratic loss $\ell(h, z = (x, y)) = \frac{1}{\sigma^2}(h(x) - y)^2$ and sampling distribution with pdf

$$\text{pr}(y | \{x_i, y_i\}) \propto \exp\left(-\sum_{i=1}^n \ell(h(x_i), (x_i, y_i))\right).$$

As $\eta(x)$ is assumed to be unknown, according to the Bayesian paradigm and by taking advantage of Note 9, we assign a GP prior on $\eta(\cdot)$

$$\eta(\cdot) \sim \text{GP}(\mu(\cdot|\beta), C(\cdot,\cdot|\phi)) \tag{2.2}$$

where μ is parametrized by unknown β (e.g. $\mu(x|\beta) = x^\top \beta$), and C is parametrized by unknown ϕ (e.g. $C(x, x'|\phi) = \exp\left(-\frac{1}{2\phi}\|x - x'\|_2^2\right)$; radial/Gaussian kernel). Summing up, the Bayesian model

$$\begin{cases} y_i | \eta(\cdot), \sigma^2 & \overset{\text{iid}}{\sim} \text{N}\left(\eta(x_i), \sigma^2\right),\ i = 1, ..., n \\ \eta(\cdot) | \beta, \phi & \sim \text{GP}(\mu(\cdot|\beta), C(\cdot,\cdot|\phi)) \end{cases}$$

up to some unknown tuning parameters σ^2, β , and $\phi > 0$.

Notation 14. From now on, to easy the notation, x, σ^2, β , and ϕ are suppressed from the conditioning; e.g. we use $C(\cdot,\cdot)$ instead of $\mathrm{C}(\cdot,\cdot|\phi)$.

Note 15. A realization of (2.2) (with ref to Fig 2.1c) can be simulated by setting a finite vector $x := (x_1, ..., x_m)$, computing μ such that $[\mu]_i = \mu(x_i)$ and C such that $[C]_{i,j} = C(x_i, x_j)$ and drawing am m dimensional vector η from the multivariate Normal distribution $\mathrm{N}(\mu, C)$. See Fig 2.1c, and the R code provided int he caption.

Note 16. Consider $\eta_* = \eta(X_*)$ where $X_* = (x_{*,1}, x_{*,2}, ..., x_{*,m})^\top$ is a vector of new inputs of any length $m > 0$. The joint distribution of $(\eta_*, y)^\top$ is

$$\begin{pmatrix}\eta_* \\ y\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mu(X_*) \\ \mu(X)\end{pmatrix}, \begin{pmatrix}C(X_*, X_*) & C(X_*, X) \\ C(X, X_*) & C(X, X) + I\sigma^2\end{pmatrix}\right) \tag{2.3}$$

where $C(X, X_*)$ is a Gram matrix over X and X_* such as $[C(X, X_*)]_{i,j} = C(x_i, x_{*,j})$.

Note 17. The conditional distribution of $\eta_* = \eta(X_*)$ given the training sample $\{z_i = (x_i, y_i)\}$, as results from 2.3 (Proposition 44),

$$\eta_*|y \sim \mathrm{N}(\mu_*(X_*), C_*(X_*, X_*)) \tag{2.4}$$

is a normal distribution, with mean

$$\mu_*(X_*) = \mathrm{E}(\eta_*|y) = \mu(X_*) + C(X_*, X)\left(C(X, X) + I\sigma^2\right)^{-1}(y - \mu(X)) \tag{2.5}$$

at X_* and with covariance function

$$C_*(X_*, X_*) = \mathrm{Cov}(\eta_*|y) = C(X_*, X_*) - C(X_*, X)\left(C(X, X) + I\sigma^2\right)^{-1}(C(X_*, X))^\top \tag{2.6}$$

Note 18. Comparing (2.6) with (1.4), the extra $+\sigma^2$ in the first term of the right-hand-side part is because in Note 17 we predict (and are interested in) the "underline pattern only" $\eta_* = \eta(x_*)$ while in Note we predict the "experimental outcome" $y_* = \eta_* + \epsilon$, $\epsilon \sim \mathrm{N}\left(0, \sigma^2\right)$4.

Note 19. Because X_* is of any finite length, and the derivations in Note 17, by definition of GP, the predictive distribution of $\eta(\cdot)$ given the data $\{z_i = (x_i, y_i)\}$ is the Gaussian process

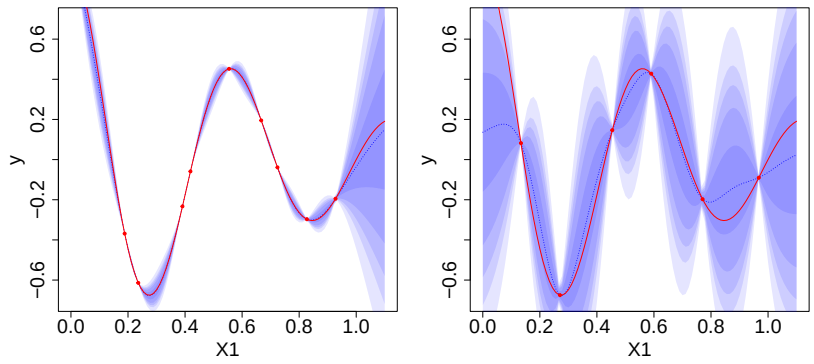
$$\eta(\cdot) \,|\, \{(x_i, y_i)\} \sim \mathrm{GP}(\mu_*(\cdot), C_*(\cdot,\cdot)) \tag{2.7}$$

with mean function and covariance function

$$\mu_*(x_*) = \mu(x_*) + C(x_*, X)\left(C(X, X) + I\sigma^2\right)^{-1}(y - \mu(X)) \tag{2.8}$$

$$C_*\left(x_*, x_*'\right) = C\left(x_*, x_*'\right) - C(x_*, X)\left(C(X, X) + I\sigma^2\right)^{-1} C\left(X, x_*'\right) \tag{2.9}$$

for any points $x_*, x_*' \in \mathcal{X}$. If I consider $X_* = (x_*, x_*')^\top$, (2.8) results as the first block of (2.5), and (2.9) results as the top off-diagonal block of (2.6).



(A) (B)

FIGURE 2.2. Predictive GP regressions given different numbers of training data points, for some values of σ^2, ϕ, and β. Fig 2.2a shows the predictive GP given $n = 10$ data-points. Fig 2.2b shows the predictive GP given $n = 6$ data-points.

Note 20. Note that the posterior expected rule at $x_* \in \mathcal{X}$ is

$$\begin{aligned} \mathrm{E}\left(h\left(x_*\right)|y\right) &= \mathrm{E}\left(\eta\left(x_*\right)|y\right) = \mu\left(x_*\right) + C\left(x_*, X\right)\left(C\left(X,X\right) + I\sigma^2\right)^{-1}\left(y - \mu\left(X\right)\right) \\ &= \sum_{i=1}^{n} \alpha_i C\left(x_i, x_*\right), \quad \text{by assuming that } \mu\left(\cdot\right)=0 \end{aligned} \tag{2.10}$$

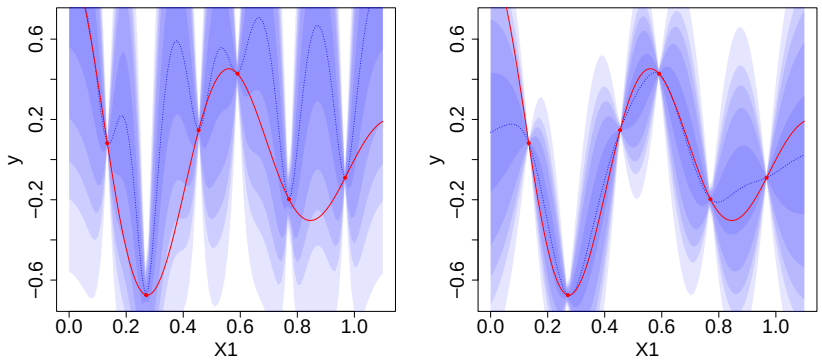
where $\alpha = y^\top \left(C\left(X,X\right) + I\sigma^2\right)^{-1} C\left(X, x_*\right)$. This is in accordance to the Representation theorem (Theorem 19 in Handout 7 Kernel methods) with reference to the Bayesian linear regression (Note 2).

Example 21. Given some values for σ^2, ϕ, and β (see Section σ^2, ϕ, and β), the predictive GP regression for $\eta\left(\cdot\right)$ is represented in Figure 2.2a. The red dots are training data points $\{z_i = (x_i, y_i)\}$. The red line is the predictive GP mean (2.8). The blue shades is the ares between the 95% quantiles of the Normal distribution $\mathrm{N}\left(\mu_*\left(x_*\right), C_*\left(x_*, x_*\right)\right)$ at each point x_* of the inputs. Note that $\mathrm{N}\left(\mu_*\left(x_*\right), C_*\left(x_*, x_*\right)\right)$ is just a snapshot of 2.7 at point x_*.

Note 22. (Related to Example 21) Intuitively, we can imagine that the conditional $\eta\left(\cdot\right)|y$ in Figure 2.2a results from the marginal/prior $\eta\left(\cdot\right)$ in Figure 2.1c by forcing all the lines in Figure 2.1c to pass through the data points in Figure 2.1c. So the more the data points, the smaller the uncertainty around the predictive mean; compare Fig 2.2a using 10 examples and Fig 3.1b using 5 examples.

3. Training (via Empirical Bayes)

Note 23. Recall that the mean and covariance functions in (2.7) depend on tunable parameters σ^2, ϕ, and β. When the number of training examples is small, the behavior of (2.7) is sensitive to these hyperparameters. In Figure 3.1, there are two instances of GP regression given 6 examples where the tunable parameters are different.



(A) Careless choice of ϕ Here $\sigma^2 = 0$, and $\beta = 0$.

(B) Wise choice of ϕ Here $\sigma^2 = 0$, and $\beta = 0$.

FIGURE 3.1. Sensitivity in the tunable parameters ϕ. Here $\sigma^2 = 0$, and $\beta = 0$.

Note 24. The marginal likelihood $f\left(y|\sigma^2, \phi, \beta\right)$ of y given the known parameters σ^2, ϕ, β results from (2.3) as

$$y|\sigma^2, \phi, \beta \sim \mathrm{N}\left(\mu\left(X|\beta\right), C\left(X, X|\phi\right) + I\sigma^2\right). \tag{3.1}$$

Note 25. Let $\mu_\beta = \mu\left(X|\beta\right)$ and $C_\phi = C\left(X, X|\phi\right)$. To learn the unknown hyper-parameters $\theta = \left(\sigma^2, \phi, \beta\right)$ according to classical training methods, we can specify the empirical risk function $\hat{R}\left(\sigma^2, \phi, \beta\right)$ by using the marginal likelihood in (3.1), as $\hat{R}\left(\sigma^2, \phi, \beta\right) = -2\log\left(f\left(y|\sigma^2, \phi, \beta\right)\right)$ [2]. Hence, we need to minimize

$$\begin{aligned}\left(\hat{\sigma}^2, \hat{\phi}, \hat{\beta}\right) &= \underset{\sigma^2,\phi,\beta}{\arg\min}\left(-2\log\left(\mathrm{N}\left(y \,|\, \mu_\beta, C_\phi + I\sigma^2\right)\right)\right) \\ &= \underset{\sigma^2,\phi,\beta}{\arg\min}\left(\underbrace{\log\left(\left|C_\phi + I\sigma^2\right|\right) + \left(y - \mu_\beta\right)^\top \left(C_\phi + I\sigma^2\right)^{-1}\left(y - \mu_\beta\right)}_{\hat{R}(\sigma^2,\phi,\beta)}\right).\end{aligned} \tag{3.2}$$

Note 26. (3.2) can be solved via GD (Algorithm 4 Handout 2 Gradient descent), or the Stochastic Gradient (Handout 3: Stochastic gradient descent) as the required gradient ca be easily computed as

$$\begin{aligned}\frac{\mathrm{d}R}{\mathrm{d}\beta_j} &= \left(C_\phi + I\sigma^2\right)^{-1}\left(y - \mu_\beta\right)\frac{\mathrm{d}\mu_\beta}{\mathrm{d}\beta_j} \\ \frac{\mathrm{d}R}{\mathrm{d}\phi_j} &= \mathrm{tr}\left(\left(C_\phi + I\sigma^2\right)^{-1}\left[\frac{\partial C_\phi}{\partial \phi_j}\right]\right) + \left(y - \mu\right)^\top\left(C_\phi + I\sigma^2\right)^{-1}\left[\frac{\partial C_\phi}{\partial \phi_j}\right]\left(C_\phi + I\sigma^2\right)^{-1}\left(y - \mu\right) \\ \frac{\mathrm{d}R}{\mathrm{d}\sigma^2} &= \mathrm{tr}\left(\left(C_\phi + I\sigma^2\right)^{-1}\right) + \left(y - \mu\right)^\top\left(C_\phi + I\sigma^2\right)^{-1}\left(C_\phi + I\sigma^2\right)^{-1}\left(y - \mu\right)\end{aligned}$$

No need t memorized

[2] Equivalently Empirical Bayes training procedures

4. Examples of Covariance Functions

Note 27. The covariance function $C : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ with $C(x, x')$ describes how much two random variables x, x' change together.

Note 28. Covariance function is a functional parameter of the GP prior (2.2). Different covarince functions represent different properties, hence they impose different prior info in the GP regression model; they are crusial parameters.

Note 29. Any positive definite Kernel as described in Section 4 in Handout 7 Kernel methods can be used as a covarince function. Consequently kernel construction approached and theories introduced can be use for the covariance functions as well.

Definition 30. Stationary covariance function is called a covariance function $C : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ whose image can be written as $C(x, x') = C(\|x - x'\|)$ namely, the dependence between any pair of input points x , x' is a function of their distance and only.

Note 31. In 1-D Gaussian process, one way to understand the characteristic length-scale of the process (if this exists) is in terms of the number of upcrossings of a level u. Consider stationary covariance function $C(x, x') = C(\|x - x'\|)$. The expected number of upcrossings $\mathrm{E}(N_u)$ of the level u on the unit interval by a zero-mean, stationary, is

$$\mathrm{E}(N_u) = \frac{1}{2\pi}\sqrt{\frac{-C''(0)}{C'(0)}}\exp\left(-\frac{u^2}{2C(0)}\right)$$

Note 32. Popular covariance functions are

Gaussian covariance function: given as

$$C(r) = \exp\left(-\frac{1}{2\phi^2}r^2\right)$$

- It is infinitely differentiable, which means that the GP is very smooth.
- The parameter ϕ is called lengthscale.
- The number of upcrossing at level u is $\mathrm{E}(N_u) = \left(2\phi^2\right)^{-1}$ meaning that smaller ϕ represents more upcrossings, hence represents smaller scale dependences

Exponential Covariance Function: given as

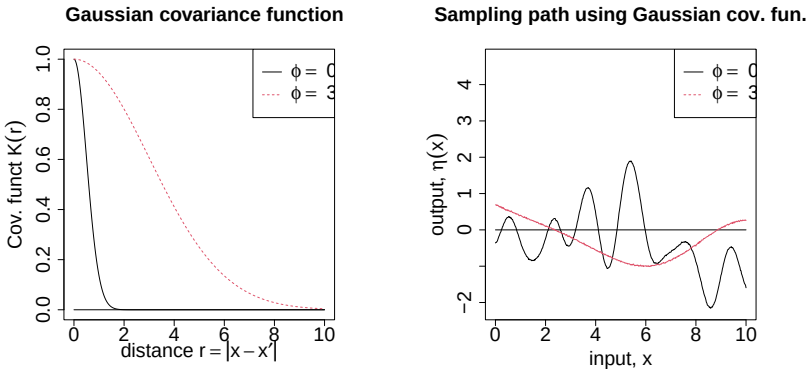
$$C(r) = \exp\left(-\frac{1}{\phi}|r|\right)$$

- It is not differentiable at $r = 0$

Matern Class of Covariance Functions: given as

$$C_\nu(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{\phi}\right)^\nu B_\nu\left(\frac{\sqrt{2\nu}r}{\phi}\right)$$

where $B_\nu(\cdot)$ is a modified Bessel functions (description of Bessel functions is out of the scope). Matern covariance function gives the Exponential one for $\nu = 1/2$, and the Gaussian one for $\nu \to \infty$.

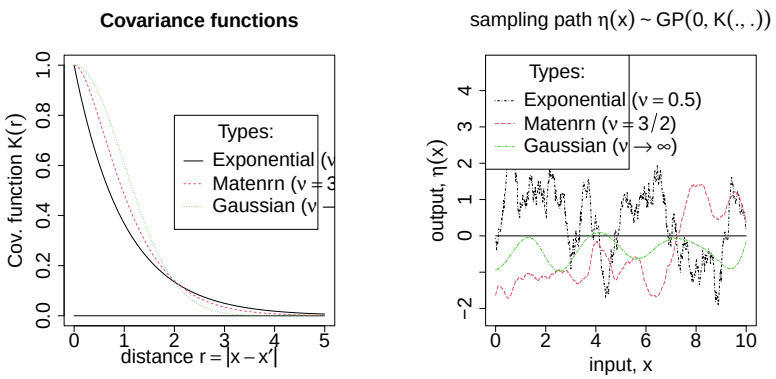


(A) (B)

FIGURE 4.1. Investigation of the effect of choosing a different length scale ϕ

Example 33. We investigate the effect of choosing a different length scale ϕ in Figure 4.1. We consider the Gaussian covariance function. Figure 4.1a shows that the smaller the length scale ϕ is the more it focuses on the small scale dependences. Figure 4.1a shows that the smaller the length scale ϕ is the more upcrossings the associated GP has. This agrees with Note 31. Bottom line, it is reasonable for the researcher to specify a GP prior with a covariance with a smaller length scale ϕ if there is some a priori knowledge that the line $\eta(\cdot)$ may have strong small scale variation (eg changes more frequently).

Example 34. We investigate the effect of choosing a different covariance functions: Gaussian, Exponential, Matern 3/2 covariance functions (Note 32) in Figure 4.2. In Fig 4.2a, we observe that (a) covariance function reduces with distance that is closer points are expected to have larger dependence; (b) Gaussian cov gives more weight to small scale dependences (closer points) than the exponential. In Fig 4.2b, we see that the GP using a Gaussian cov is smoother than the GP using an Exponential cov. as expected as the former is infinitely differentiable while the latter is not differentiable. In general, Matern ($\nu = 3/2$) 's behavior is between those of Gaussian and Exponential, as expected.



(A) (B)

FIGURE 4.2. Investigation of the effect of choosing a different covariance functions: Gaussian, Exponential, Matern 3/2 covariance functions. (4.2a) shows the covariance function $C(x, x')$ against the distance $|x - x'|$. (4.2a) shows the covariance function $C(x, x')$ against the distance $|x - x'|$. (4.2b) shows realizations of GP using different covariance functions.

Note 35. Anisotropic versions of these isotropic covariance functions can be created anisotropy by setting $r(x, x') = (x - x')^\top M (x - x')$ for some positive semi-definite matrix M. If M is diagonal this implements the use of different length-scales on different dimension of inputs. Off-diagonal elements of M implement cross-dimensional dependencies in the inputs.

5. Practical maters

Note 36. One may consider some low degree polynomial form $\mu(x|\beta) = \sum_{j=0}^{p} x^j \beta_j$, however in this case $\mu(\cdot|\beta)$ and $C(\cdot, \cdot|\phi)$ may compete as $C(\cdot, \cdot|\phi)$ can express such behaviors according to Note 9. For this reason, the usual specification of $\mu(\cdot|\beta)$ in (2.2) is $\mu(x|\beta) = 0$ implying a priori complete uncertainty about the sign of $\eta(x)$ at each x.

Note 37. (Geostistical model) Thinking as a statistician, one may decompose (2.2) as

$$\eta(\cdot) = \mu(\cdot|\beta) + \xi(\cdot|\phi)$$

where $\mu(\cdot|\beta)$ is modeled as a low degree polynomial (e.g., 2nd degree) representing large scale dependences (see polynomial regression in Term 1), and $\xi(\cdot|\phi) \sim \text{GP}(0, C(\cdot, \cdot|\phi))$ representing lower scale dependence by using an appropriate kernel. Seeing the big picture, (2.1) can be re-stated as $y_i = \mu(x_i|\beta) + \xi(x_i|\phi) + \epsilon_i$ where therm ϵ_i represents noise (no dependence or so short scale dependence that can be considered as noise in the model), $\xi(x_i|\phi)$ represents low-scale dependence (about nearby inputs points), and $\mu(x_i|\beta)$ represents large scale dependence (about very distant inputs points).

6. Practice, implementation, and code

Below is some practical examples on the implementation of Gaussian process regression in R programming environment by using the R packages: DiceKriging and DiceOptim.

- Roustant, O., Ginsbourger, D., & Deville, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Journal of statistical software, 51, 1-55.

The examples below were created for the undergraduate programme SURF 2016 at Purdue University (July 8, 2016) however they are suitable to the course for your practice.

Toy example
`https://github.com/georgios-stats/Intro_GPR_SURF_2016/blob/master/Numerical_example.ipynb`

The Piston Simulation function model in 2D
`https://github.com/georgios-stats/Intro_GPR_SURF_2016/blob/master/Practice_2D.ipynb`

Practice Catalytic Reaction 5D
`https://github.com/georgios-stats/Intro_GPR_SURF_2016/blob/master/Practice_CatalyticReaction_5D_solution.ipynb`

Practice Piston 7D
`https://github.com/georgios-stats/Intro_GPR_SURF_2016/blob/master/Practice_Piston_7D_solution.ipynb`

Practice Robot Arm 8D
`https://github.com/georgios-stats/Intro_GPR_SURF_2016/blob/master/Practice_Robot_Arm_8D_solution.ipynbb`

Appendix A. Multivariate Normal distribution[3] $x|\mu,\Sigma \sim \mathrm{N}_d(\mu,\Sigma)$

Definition 38. A d-dimensional random variable $x \in \mathbb{R}^d$ is said to have a multivariate Normal (Gaussian) distribution, if for every d-dimensional fixed vector $\alpha \in \mathbb{R}^d$, the random variable $\alpha^\top x$ has a univariate Normal (Gaussian) distribution.

Definition 39. We denote the d-dimensional Normal distribution with mean μ and covariance matrix $\Sigma \geq 0$ as $\mathrm{N}_d(\mu,\Sigma)$.

Notation 40. The d-dimensional standardized Normal distribution is $\mathrm{N}_d(0,I)$.

Proposition 41. *Let random variable $x \sim N_d(\mu,\Sigma)$, fixed vector $c \in \mathbb{R}^q$ and fixed matrix $A \in \mathbb{R}^q \times \mathbb{R}^d$. The random vector $y = c + Ax$ has distribution $y \sim N_q\left(c + A\mu, A\Sigma A^\top\right)$.*

Proposition 42. *Let a d-dimensional random vector $x \sim N_{(any)}(\mu,\Sigma)$.*

1. *Let $y = Ax$ and $z = Bx$, where $A \in \mathbb{R}^{q\times d}$ and $B \in \mathbb{R}^{k\times d}$: The vectors $y = Ax$ and $z = Bx$ are independent if and only if $A\Sigma B^\top = 0$.*
2. *Let $x = (x_1,...,x_d)^\top$: The $x_1,...,x_d$ are mutually independent if and only if the corresponding off diagonal parts of the Σ are zero.*

Proposition 43. *Any sub-vector of a vector with multivariate Normal distribution has a multivariate Normal distribution.*

Proposition 44. *[Marginalization & conditioning] Let $x \sim N_d(\mu,\Sigma)$. Consider partition such that*

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}; \qquad \Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{21}^\top \\ \Sigma_{21} & \Sigma_2 \end{bmatrix},$$

where $x_1 \in \mathbb{R}^{d_1}$, and $x_2 \in \mathbb{R}^{d_2}$ Then:

1. *For the marginal, it is $x_1 \sim N_{d_1}(\mu_1,\Sigma_1)$.*
2. *For the conditional, if $\Sigma_1 > 0$, it is*

$$x_2|x_1 \sim N_{d_2}(\mu_{2|1},\Sigma_{2|1})$$

where

$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_1^{-1}(x_1-\mu_1) \text{ and } \Sigma_{2|1} = \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top \tag{A.1}$$

Proposition 45. *The density function of the d-dimensional Normal distribution with mean μ and covariance matrix Σ, when Σ is symmetric positive definite matrix ($\Sigma > 0$), exists and it is equal to*

$$f(x) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right) \tag{A.2}$$

[3] More detailed material about the Multivariate Normal distribution can be found in the can be found in "Handout 2: Revision in mixture of probability distributions" of the module "Bayesian Statistics III/IV (MATH3341/4031)" Michaelmas term, 2021 available from `https://github.com/georgios-stats/Bayesian_Statistics_Michaelmas_2021/blob/main/Lecture_handouts/02_Revision_in_mixture_of_probability_distributions.pdf`. The material in this section is just a sub-set of the statements in the referenced handout.