Chapter 11: Gaussian Process Regression– Theory[a]

We consider data points $\{(y_i^0, x_i^0)\}_{i=1}^n$ in $\mathbb{R} \times \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^p$, and assume the following regression model

$$Y_i^0 = f(x_i^0) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}_1(0, \sigma^2), \quad f \in \mathcal{F} \tag{11.1}$$

where $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}\}$ is a space of infinite dimension and $\sigma^2 > 0$.

In this chapter we consider a Bayesian approach for estimating f. As usual in Bayesian statistics, this requires to choose a prior distribution $\pi_f(\cdot|\sigma^2)$ for f and a prior distribution π_{σ^2} for σ^2.

The inference on (f, σ^2) is then based on the posterior distribution:

$$\pi(f, \sigma^2 | y_{1:n}^0) \propto p(y_{1:n}^0 | f, \sigma^2) \pi_f(f|\sigma^2) \pi_{\sigma^2}(\sigma^2). \tag{11.2}$$

Remark: Unlike in the previous chapters we use the notation $y_{1:n}^0$ (and not y^0) for the vector $(y_1^0, \ldots, y_n^0)$ because later we will study the behaviour of Bayesian posterior quantities as $n \to \infty$.

In Gaussian process regression (GPR) the prior $\pi_f(\cdot|\sigma^2)$ is a Gaussian process and is specified through the choice of a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (see Definition 4.2).

Using a GP is not the only way to specify a prior distribution $\pi_f(\cdot|\sigma^2)$ on $\mathcal{F}$. However, GP regression has the advantage to be (i) popular not only in statistics but also in machine learning and (ii) to be easy to use in practice, thanks to the various methods developed in the machine learning literature to efficiently approximate $\pi(f, \sigma^2 | y_{1:n}^0)$.

Remark: The fact that GP regression is easy to use in practice does not mean that it is always the 'best' method to use!

[a] The main references for this chapter are [12, Chapter 2] and [7].

Gaussian process

Definition 11.5 (Gaussian process) *A Gaussian process (GP) is a collection of random variables, any finite number of which has a joint Gaussian distribution.*

A Gaussian process $W = (W(x))_{x\in\mathcal{X}}$ indexed by $x \in \mathcal{X}$ is therefore completely characterised by its mean function $\mu : \mathcal{X} \to \mathbb{R}$ and kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and thus we write $W \sim \mathrm{GP}(\mu, k)$.

Hence, if $W \sim \mathrm{GP}(\mu, k)$ then, for all $(x'_1, \ldots, x'_m) \in \mathcal{X}^m$ and $m \in \mathbb{N}$,

$$(W(x'_1), \ldots W(x'_m)) \sim \mathcal{N}_m\Big(\big(\mu(x'_1), \ldots, \mu(x'_m)\big), \big(k(x'_i, x'_j)\big)_{i,j=1}^m\Big).$$

Figure 11.1 below shows four realizations of a $\mathrm{GP}(0, k)$ process (with k the Matérn kernel with parameter $(\alpha, \gamma) = (2, 0.1)$; see below).

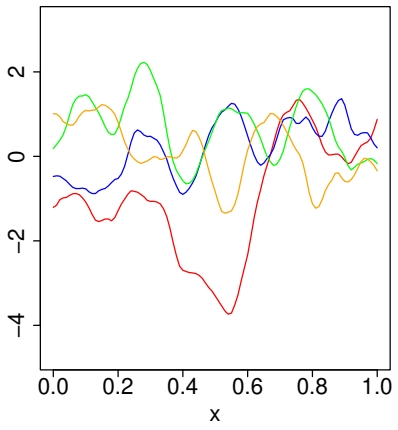

Figure 11.1: Four draws from a $\mathrm{GP}(0, k)$ process ($\mathcal{X} = [0, 1]$).

GP regression with known σ^2

We first consider the following non-parametric Bayesian regression model

$$Y_i^0 = f(x_i^0) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}_1(0, \sigma^2), \quad f \sim \mathrm{GP}(0, k), \quad \sigma^2 \sim \delta_\lambda \qquad (11.3)$$

with $\lambda > 0$ a hyperparameter.

Notation: δ_λ is the Dirac mass at λ, i.e $\sigma^2 \sim \delta_\lambda$ means that $\sigma^2 = \lambda$ with probability one.

Remark: Here we follow the (very) common practice to consider as a prior for f a GP with mean function $\mu \equiv 0$ but what follows can be easily generalized to an arbitrary mean function μ.

Letting $\boldsymbol{K}_n = \big(k(x_i^0, x_j^0)\big)_{i,j=1}^n$ be the Gram matrix and

$$k_n(x) = \big(k(x_1^0, x), \ldots, k(x_n^0, x)\big), \quad \forall x \in \mathcal{X}$$

we obtain the following expression for the posterior distribution of f given $y_{1:n}^0$ in the model (11.3).

Proposition 11.1 *Consider the Bayesian model* (11.3)*. Then,* $f | y_{1:n}^0 \sim \mathrm{GP}(f_n, k_n)$ *where* $f_n : \mathcal{X} \to \mathbb{R}$ *and* $k_n : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *are defined by*

$$f_n(x) = k_n(x)^\top \big(\boldsymbol{K}_n + \lambda \boldsymbol{I}_n\big)^{-1} y_{1:n}^0, \qquad x \in \mathcal{X}$$
$$k_n(x, x') = k(x, x') - k_n(x)^\top \big(\boldsymbol{K}_n + \lambda \boldsymbol{I}_n\big)^{-1} k_n(x'), \quad (x, x') \in \mathcal{X}^2.$$

Proof of Proposition 11.1

Let $m \in \mathbb{N}$, $(x'_1, \ldots, x'_m) \in \mathcal{X}^m$, $f'_{1:m} = \big(f(x'_1), \ldots, f(x'_m)\big)$ and

$$\boldsymbol{K}'_m = \big(k(x'_i, x'_j)\big)_{i,j=1}^m, \quad \boldsymbol{K}'_{mn} = \big(k_n(x'_1) \quad \ldots \quad k_n(x'_m)\big)^\top.$$

Then, under (11.3),

$$\begin{pmatrix} Y^0_{1:n} \\ f'_{1:m} \end{pmatrix} \sim \mathcal{N}_{n+m}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{K}_n + \lambda \boldsymbol{I}_n & (\boldsymbol{K}'_{mn})^\top \\ \boldsymbol{K}'_{mn} & \boldsymbol{K}'_m \end{pmatrix} \right) \tag{11.4}$$

so that, under (11.3),

$$f'_{1:m} | y^0_{1:n} \sim \mathcal{N}_m\Big(\boldsymbol{K}'_{mn}(\boldsymbol{K}_n + \lambda \boldsymbol{I}_n)^{-1} y^0_{1:n}, \boldsymbol{K}'_m - \boldsymbol{K}'_{mn}(\boldsymbol{K}_n + \lambda \boldsymbol{I}_n)^{-1}(\boldsymbol{K}'_{mn})^\top\Big).$$

Then, the result follows by noting that

$$\boldsymbol{K}'_{mn}(\boldsymbol{K}_n + \lambda \boldsymbol{I}_n)^{-1} y^0_{1:n} = \begin{pmatrix} k_n(x'_1)^\top (\boldsymbol{K}_n + \lambda \boldsymbol{I}_n)^{-1} y^0_{1:n} \\ \vdots \\ k_n(x'_m)^\top (\boldsymbol{K}_n + \lambda \boldsymbol{I}_n)^{-1} y^0_{1:n} \end{pmatrix} = \begin{pmatrix} f_n(x'_1) \\ \vdots \\ f_n(x'_m) \end{pmatrix}$$

while, using the shorthand $\boldsymbol{Q} = (\boldsymbol{K}_n + \lambda \boldsymbol{I}_n)^{-1}$, we have

$$\begin{aligned} &\boldsymbol{K}'_m - \boldsymbol{K}'_{mn} \boldsymbol{Q}\, \boldsymbol{K}^\top_{mn} \\ &= \begin{pmatrix} k(x'_1, x'_1) - k_n(x'_1)^\top \boldsymbol{Q}\, k_n(x'_1) & \ldots & k(x'_1, x'_m) - k_n(x'_1)^\top \boldsymbol{Q}\, k_n(x'_m) \\ & \vdots & \\ k(x'_m, x'_1) - k_n(x'_m)^\top \boldsymbol{Q}\, k_n(x'_1) & \ldots & k(x'_m, x'_m) - k_n(x'_m)^\top \boldsymbol{Q}\, k_n(x'_m) \end{pmatrix} \\ &= \big(k_n(x'_i, x'_j)\big)_{i,j=1}^m. \end{aligned}$$

□

Two comments and marginal likelihood

It is useful to remark that:

1. The mean function f_n can be re-written as follows:

$$f_n = \sum_{i=1}^{n} \alpha_{n,i} k(x_i^0, \cdot), \quad \alpha_n = \big(\boldsymbol{K}_n + \lambda \boldsymbol{I}_n\big)^{-1} y_{1:n}^0. \tag{11.5}$$

2. For all $x \in \mathcal{X}$, $f_n(x)$ is, under the model (11.3), the Bayes estimator of $f(x)$ associated to both the quadratic and the absolute error loss function.

From (11.4) we readily deduce the expression for the marginal distribution of $y_{1:n}^0$ in the model (11.3).

Proposition 11.2 *Consider the Bayesian model* (11.3)*. Then,*

$$Y_{1:n}^0 \sim \mathcal{N}_n\big(0, \boldsymbol{K}_n + \lambda \boldsymbol{I}_n\big).$$

Remark: The density of the marginal distribution of $Y_{1:n}^0$, called the marginal likelihood and denoted by $p(y_{1:n}^0|\lambda, k)$ in what follows, is often used to choose the hyperparameters of the model (see below).

GP regression with unknown σ^2

We now consider the following non-parametric Bayesian model:

$$\begin{aligned} Y_i^0 &= f(x_i^0) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}_1(0, \sigma^2), \\ f|\sigma^2 &\sim \mathrm{GP}\big(0, (\sigma^2/\lambda)k\big), \quad \frac{1}{\sigma^2} \sim \mathrm{Gamma}(a_0, b_0) \end{aligned} \tag{11.6}$$

where $\lambda, a_0, b_0 > 0$ are some hyperparameters.

In (11.6), the prior distribution $\pi_f(\cdot|\sigma^2)$ of f and the prior distribution π_σ^2 of σ^2 are chosen to ease the computations since, as shown below, the corresponding posterior distribution of (f, σ^2) given $y_{1:n}^0$ is tractable.

Remark: In (11.6) the prior density π_{σ^2} is given by

$$\pi_{\sigma^2}(z) = \frac{b_0^{a_0}}{\Gamma(a_0)} \Big(\frac{1}{z}\Big)^{a_0+1} e^{-b_0/z}, \quad z > 0.$$

Proposition 11.3 *Consider the Bayesian model* (11.6)*. Then,* $\pi(f, \sigma^2|y_{1:n}^0)$ *is such that*

$$f|(\sigma^2, y_{1:n}^0) \sim \mathrm{GP}\big(f_n, (\sigma^2/\lambda)k_n\big), \quad \frac{1}{\sigma^2}\big|y_{1:n}^0 \sim \mathrm{Gamma}(a_n, b_n)$$

where f_n *and* k_n *are as defined in Proposition 11.1 while*

$$a_n = a_0 + \frac{n}{2}, \quad b_n = b_0 + \frac{\lambda}{2}(y_{1:n}^0)^\top (\boldsymbol{K}_n + \lambda \boldsymbol{I}_n)^{-1} y_{1:n}^0.$$

Remark: The mean function f_n is therefore the same in the two Bayesian models (11.3) and (11.6).

Posterior distribution of f and marginal likelihood in (11.6)

From Proposition 11.3 we can deduce the expression for the posterior distribution of f given $y_{1:n}^0$.

Corollary 11.1 *Consider the Bayesian model* (11.6)*. Then, for all* $m \in \mathbb{N}$ *and* $(x'_1, \dots, x'_m) \in \mathcal{X}^m$*, we have*

$$\big(f(x'_1), \dots, f(x'_m)\big) \big| y_{1:n}^0 \\ \sim t_m\Big(2a_n, \big(f_n(x'_1), \dots, f_n(x'_m)\big), \frac{b_n}{\lambda a_n}\big(k_n(x'_i, x'_j)\big)_{i,j=1}^m\Big)$$

where $t_m(\nu, \mu, \Sigma)$ *denotes the* m*-dimensional* t*-distribution with* $\nu > 0$ *degrees of freedom, location parameter* μ *and scale matrix* Σ*.*

Remark: It is sometimes said that $f|y_{1:n}^0$ is a Student-t process.

Finally, the marginal likelihood of the observations in the model (11.6) with unknown noise variance σ^2 is given in the following corollary.

Corollary 11.2 *Consider the Bayesian model* (11.6)*. Then,*

$$Y_{1:n}^0 \sim t_n\big(2a_0, 0, (b_0/a_0)(\boldsymbol{K}_n/\lambda + \boldsymbol{I}_n)\big).$$

Proof of Proposition 11.3 and of Corollaries 11.1-11.2

To prove Proposition 11.3 note that $\pi(f, \sigma^2 | y^0_{1:n}) = \pi(f|\sigma^2, y^0_{1:n})\pi(\sigma^2|y^0_{1:n})$ where

- $\pi(f|\sigma^2, y^0_{1:n})$ is obtained from Proposition 11.1, by first replacing λ by σ^2 and then replacing k by $(\sigma^2/\lambda)k$.
- $\pi(\sigma^2|y^0_{1:n}) \propto \pi(y^0_{1:n}|\sigma^2)\pi_{\sigma^2}(\sigma^2)$, with $\pi(y^0_{1:n}|\sigma^2)$ obtained by replacing λ by σ^2 in Proposition 11.2.

Then, to conclude the proof of Proposition 11.3 suffices to note that

$$\pi(y^0_{1:n}|\sigma^2)\pi_{\sigma^2}(\sigma^2) \propto \Big(\frac{1}{\sigma^2}\Big)^{a_0+1+\frac{n}{2}} \exp\Big(-\frac{1}{\sigma^2}\Big(b_0 + \frac{\lambda}{2}(y^0_{1:n})^\top(\boldsymbol{K}_n + \lambda \boldsymbol{I}_n)^{-1} y^0_{1:n}\Big)\Big).$$

As preliminary computations to prove Corollaries 11.1 and 11.2 we show the following result: Let (Z, W) be such that $Z|W \sim \mathcal{N}_m(\mu, W\boldsymbol{\Sigma})$ and $(1/W) \sim \text{Gamma}(a, b)$ for some real numbers $a, b > 0$ and covariance matrix $\boldsymbol{\Sigma}$. Then,

$$Z \sim t_m(2a, \mu, b\boldsymbol{\Sigma}) \tag{11.7}$$

To show (11.7) let $p(z)$ be the pdf of Z and note that

$$p(z) \propto \int_0^\infty w^{-\frac{2a+m+2}{2}} \exp\Big(-\frac{2b + (z-\mu)^\top \boldsymbol{\Sigma}^{-1}(z-\mu)}{2w}\Big)\mathrm{d}w.$$

Let

$$x = c/w, \quad c = \frac{2b + (Z-\mu)^\top \boldsymbol{\Sigma}^{-1}(Z-\mu)}{2}$$

so that, using the change of variable formula,

$$\begin{aligned}
p(z) &= c^{-\frac{2a+m+2}{2}} \int_\infty^0 \Big(-\frac{c}{x^2}\Big) x^{\frac{2a+m+2}{2}} e^{-x}\mathrm{d}x \\
&= c^{-\frac{2a+m}{2}} \int_0^\infty x^{\frac{2a+m}{2}-1} e^{-x}\mathrm{d}x \\
&= c^{-\frac{2a+m}{2}} \Gamma\Big(\frac{2a+m}{2}\Big) \\
&\propto c^{-\frac{2a+m}{2}} \\
&= \Big(\frac{2b + (Z-\mu)^\top \boldsymbol{\Sigma}^{-1}(Z-\mu)}{2}\Big)^{-\frac{2a+m}{2}} \\
&= \Big(\frac{2ba + (Z-\mu)^\top (\boldsymbol{\Sigma}/a)^{-1}(Z-\mu)}{2a}\Big)^{-\frac{2a+m}{2}} \\
&\propto \Big(1 + \frac{(Z-\mu)^\top (b\boldsymbol{\Sigma}/a)^{-1}(Z-\mu)}{2a}\Big)^{-\frac{2a+m}{2}}
\end{aligned}$$

where the last term is the normalized density of the $t_m(2a, \mu, (b/a)\boldsymbol{\Sigma})$ distribution. This shows (11.7).

The result of Corollary 11.1 is then obtained by applying (11.7), with $a = a_n$, $b = b_n$, $\mu = (f_n(x'_1), \ldots, f_n(x'_n))$ and $\boldsymbol{\Sigma} = \lambda^{-1}(k_n(x'_i, x'_j))_{i,j=1}^m$.

Finally, Proposition 11.3, we have

$$Y^0_{1:n}|\sigma^2 \sim \mathcal{N}_m\Big(0, \sigma^2(\boldsymbol{K}_n/\lambda + \boldsymbol{I}_n)\Big), \quad (1/\sigma^2) \sim \Gamma(a_0, b_0)$$

and thus Corollary 11.2 is proved by applying (11.7) with $m = n$, $a = a_0$, $b = b_0$, $\mu = 0$ and $\boldsymbol{\Sigma} = \boldsymbol{K}_n/\lambda + \boldsymbol{I}_n$. □

Highest posterior density regions

One of the main motivation for using Bayesian methods is that they allow, through the computation of credible sets, to easily derive a measure of uncertainty about the "parameter" (here the function) of interest.

Definition 11.6 *A credible set at level* $(1-\alpha)$ *for* $f(x)$ *is a set* $C_\alpha(x) \subset \mathbb{R}$ *such that*

$$\pi\big(f(x) \in C_\alpha(x) | y^0_{1:n}\big) \geq 1-\alpha. \tag{11.8}$$

The highest posterior density (HPD) region $C^*_\alpha(x)$ *is the smallest set* $C_\alpha(x)$ *satisfying* (11.8).

- If σ^2 is known, by Proposition 11.1 the HPD region at level $1-\alpha$ is given by

$$C^*_\alpha(x) = \Big[f_n(x) - z_{1-\frac{\alpha}{2}}\sqrt{k_n(x,x)},\ f_n(x) + z_{1-\frac{\alpha}{2}}\sqrt{k_n(x,x)}\Big]$$

  where $z_{1-\frac{\alpha}{2}} = 1-\Phi(-\alpha/2)$ with Φ c.d.f. of the $\mathcal{N}_1(0,1)$ distribution.

- If σ^2 is unknown, by Corollary 11.1, the HPD region at level $1-\alpha$ for $f(x)$ given by

$$C^*_\alpha(x) = \left[f_n(x) - t_{2a_n,1-\frac{\alpha}{2}}\sqrt{\frac{b_n k_n(x,x)}{\lambda a_n}},\ f_n(x) + t_{2a_n,1-\frac{\alpha}{2}}\sqrt{\frac{b_n k_n(x,x)}{\lambda a_n}}\right]$$

  where $t_{2a_n,1-\frac{\alpha}{2}} = 1 - F_{2a_n}(-\alpha/2)$ with F_{2a_n} denote the c.d.f. of the $t_1(2a_n, 0, 1)$ distribution.

Uncertainty quantification: Three comments

1. Under the two Bayesian models (11.3) and (11.6), the posterior distribution of $\big(f(x'_1), \ldots, f(x'_m)\big)$ given $y^0_{1:n}$ is unimodal and symmetric around $\big(f_n(x'_1), \ldots, f_n(x'_m)\big)$.

   However, the tails of the posterior distribution for f are thinner under (11.3) than under (11.6), the reason being that knowing σ^2 increases the amount of available information and thus reduces the uncertainty about f. For this reason, the interval $C^*_\alpha(x)$ is, in general, larger when σ^2 is unknown that when σ^2 is known.

   $\Longrightarrow$ The conclusion of the models (11.3) and (11.6) regarding the uncertainty about f is different.

2. Recall that, for parametric models, HPD regions are asymptotically valid confidence intervals if the model is well-specified. More precisely, if C^*_α is the HPD region at level $(1-\alpha)$ for the parameter of the model then, as $n \to \infty$, the probability that C^*_α contains the true parameter value converges to $(1-\alpha)$.

   For nonparametric models with fixed hyperparameters, it is not the case that HPD regions are asymptotically valid confidence intervals. In particular, there exist examples where the coverage of HPD regions is disastrous (see [15] an references therein).

3. Since $f_n(x)$ is both the posterior mean of $f(x)$ and the centre of the HPD regions, the posterior distribution of f is often visualized by plotting the mean function f_n together with the lower and upper bounds indicating the HPD region at a chosen level $(1-\alpha)$ (typically at level 95%).

GP regression with rescaled kernels

Let $\tilde{k} = ak$ for some constant $a > 0$, so that $\tilde{k}$ is also a positive definite kernel. As illustrated in Figure 11.2 the parameter $a > 0$ regulates the amplitude of the functions sampled from $\mathrm{GP}(0, ak)$.

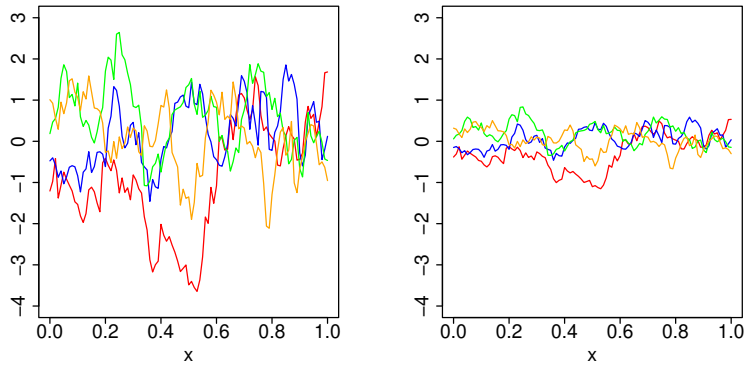

Figure 11.2: Four draws from a $\mathrm{GP}(0, ak)$ process ($\mathcal{X} = [0, 1]$) with $a = 1$ (left) and $a = 0.1$ right.

Consider the model (11.3) with known variance and let $f_{n,\lambda,a}$ and $k_{n,\lambda,a}$ be respectively the mean and the kernel of the distribution of $f|y^0_{1:n}$ under the kernel $\tilde{k} = ak$.

Proposition 11.4 *We have* $f_{n,\lambda,a} = f_{n,\lambda/a,1}$ *and* $k_{n,\lambda,a} = ak_{n,\lambda/a,1}$.

Proof: Let $r = \lambda/a$. Then, for every $x \in \mathcal{X}$ we have

$$f_{n,\lambda,a}(x) = ak_n(x)^\top \left(a\boldsymbol{K}_n + \lambda \boldsymbol{I}_n\right)^{-1} y^0_{1:n} = ak_n(x)^\top \left(a(\boldsymbol{K}_n + \boldsymbol{I}_n \lambda/a)\right)^{-1} y^0_{1:n} = k_n(x)^\top \left(\boldsymbol{K}_n + (\lambda/a)\boldsymbol{I}_n\right)^{-1} y^0_{1:n} = f_{n,r,1}(x)$$

and for every $x, x' \in \mathcal{X}$ we have

$$k_{n,\lambda,a}(x, x') = ak(x, x') - ak_n(x)^\top \left(a\boldsymbol{K}_n + \lambda \boldsymbol{I}_n\right)^{-1} ak_n(x') = a\Big(k(x, x') - k_n(x)^\top \left(\boldsymbol{K}_n + \boldsymbol{I}_n(\lambda/a)\right)^{-1}\Big) = ak_{n,r,1}(x, x').$$

The result follows. □

Remark: Proposition 11.4 shows that the mean function depends on λ and a only through the ratio $r = \lambda/a$ and that, for a given value of r, the posterior standard deviation and the length of the HPD regions for $f(x)$ are proportional to $a^{1/2}$.

Properties of the mean function f_n

We first recall that, by the Moore-Aronszajn theorem, there exists a unique RKHS $(H(k), < \cdot, \cdot >_k)$ for which k is the reproducing kernel, i.e. for which k is such that

$$f(x) =< f, k(x, \cdot) >_k, \quad \forall x \in \mathcal{X}, \ \forall f \in H(k). \quad \text{[Reproducing property]}$$

Let $\| \cdot \|_k$ be the norm on $H(k)$ induced by the inner product $< \cdot, \cdot >_k$ and recall that $f \in H(k)$ if and only if there exist a sequence $(a_i)_{i \geq 1}$ in $\mathbb{R}$ and a sequence $(x'_i)_{i \geq 1}$ in $\mathcal{X}$ such that[a]

$$\lim_{m \to \infty} \Big\| \sum_{i=1}^{m} a_i k(x'_i, \cdot) - f \Big\|_k = 0. \tag{11.9}$$

Remark: See [11] for a good reference on RKHS.

We then have the following two important properties for the mean function f_n:

- Using (11.5) and (11.9), we see that $f_n \in H(k)$ for all $n \geq 1$. Therefore, contrary to what happens for parametric models, for nonparametric models the impact of the prior distribution does not vanish as $n \to \infty$.

- Given a penalty $\lambda_n > 0$ the kernel ridge estimate of the function f in (11.1) is

$$\hat{f}_{\lambda_n} \in \underset{f \in H(k)}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i^0 - f(x_i^0))^2 + \lambda_n \| f \|_k \right\}.$$

  It can be shown that $\hat{f}_{\lambda_n} = f_n$ when $\lambda_n = \lambda / n$.

[a] It is interesting to mention that if k is bounded then (11.9) implies that $\sum_{i=1}^{m} a_i k(x'_i, \cdot)$ converges uniformly to f. Indeed, assume $|k(x, x')| \leq 1$ for all $x, x' \in \mathcal{X}$. Then, $|f(x)|^2 = | < f, k(x, \cdot) >_k |^2 \leq \|f\|_k^2 \, \|k(x, \cdot)\|_k^2 = \|f\|_k^2 \, k(x, x) \leq \|f\|_k^2$ where the first inequality uses Cauchy Schwartz. Hence, $\sup_{x \in \mathcal{X}} \left| \sum_{i=1}^{m} a_i k(x'_i, x) - f(x) \right| = \sup_{x \in \mathcal{X}} \left| \left(\sum_{i=1}^{m} a_i k(x'_i, \cdot) - f \right)(x) \right| \leq \left\| \sum_{i=1}^{m} a_i k(x'_i, \cdot) - f \right\|_k$ and the result follows.

Support of the posterior distribution for f

Let $\mathcal{X} = [0,1]$ and $k(x,x') = \min(x,x')$. Then it can be shown that

- If $f \sim \mathrm{GP}(0,k)$ then f is a Brownian motion and thus

$$\mathbb{P}\Big(f \text{ is nowhere differentiable}\Big) = 1.$$

- On the other hand, it can be shown that

$$H(k) = \Big\{f : [0,1] \to \mathbb{R} : \int_0^1 |f'(x)|^2 \mathrm{d}x < \infty\Big\}.$$

Therefore, in the Bayesian model (11.3),

$$f_n \in H(k), \quad \pi\big(H(k)|y^0_{1:n}\big) = 0, \quad \forall n \geq 1. \quad \text{(!!!)}$$

This example leads to the following key observations:

1. The elements in $H(k)$ are (typically) "smoother" than the draws from the prior (as in the above example).

2. The RKHS $H(k)$ is (typically) not in the support of the posterior distribution (in the sense that $\pi\big(H(k)|y^0_{1:n}\big) = 0$).

3. The support of the posterior distribution is (usually) much larger than $H(k)$, since for any continuous function f_0 and $\epsilon > 0$ the set $\{f : \sup_x |f(x) - f_0(x)| < \epsilon\}$ has (typically) a strictly positive mass under $\pi(f|y^0_{1:n})$.

Another view of GP regression

Let $f_{1:n} = \big(f(x_1^0), \ldots, f(x_n^0)\big)$ and consider the model (11.3) where σ^2 is assumed to be known.

Remark that

$$\pi(f, f_{1:n}|y_{1:n}^0) = \pi(f|f_{1:n}, y_{1:n}^0)\pi(f_{1:n}|y_{1:n}^0).$$

Since $y_{1:n}^0$ depends on f only through $f_{1:n}$, it follows that the conditional distribution of f given $(f_{1:n}, y_{1:n}^0)$ does not dependent on $y_{1:n}^0$ and is therefore determined by the prior only.

Consequently,

$$\pi(f, f_{1:n}|y_{1:n}^0) = \pi_f(f|f_{1:n})\pi(f_{1:n}|y_{1:n}^0) \tag{11.10}$$

so that the observations $y_{1:n}^0$ are only used to learn the n dimensional vector $\big(f(x_1^0), \ldots, f(x_n^0)\big)$ (which makes sense!).

Therefore, letting $\theta_i = f(x_i^0)$ for $i = 1, \ldots n$, we can see GP regression as a two steps procedure where

- We first estimate $(\theta_1, \ldots, \theta_n)$ in the Bayesian model

$$y_i^0 \sim \mathcal{N}_1(\theta_i, \lambda), \quad (\theta_1, \ldots, \theta_n) \sim \mathcal{N}_n(0, \boldsymbol{K}_n)$$

- We then use the GP prior to extrapolate, since by (11.10) we can sample from $\pi(f|y_{1:n}^0)$ by first sampling $f_{1:n}$ from $\pi(f_{1:n}|y_{1:n}^0)$, and then sampling f from $\pi_f(f|f_{1:n})$.

Remark: The same interpretation holds when σ^2 is unknown.

Choosing the kernel k

Recall that

- $f_n(x)$ is the Bayes estimator of $f(x)$ under the quadratic and the error loss functions (which are the two most popular loss functions),
- $f_n(x)$ is the centre of the HPD regions for $f(x)$,
- f_n is related to the kernel ridge regression estimate of f.

For the above reasons, the function f_n is usually the main object of interest in GP regression and thus, recalling that $f_n \in H(k)$, we often choose k based on the properties of $H(k)$.

The following proposition gives two general general properties on the link between k and $H(k)$

Proposition 11.5 *Let k be a continuous and bounded kernel on $\mathcal{X} \times \mathcal{X}$. Then, all functions in $H(k)$ are continuous. In addition, if k_1 and k_2 are two bounded kernels and $k = k_1 + k_2$ then*

$$H(k) = \Big\{ f_1 + f_2,\ f_1 \in H(k_1),\ f_2 \in H(k_2) \Big\}.$$

Proof: The result is a consequence of the fact that if k is bounded then the convergence w.r.t. the $\|\cdot\|_k$ norm implies the uniform convergence. □

In what follows we focus of the Gaussian and of Matérn kernels, but other kernels are popular in GP regression (see Section 4.2 of [12], and in particular Table 4.3, p.94).

The Gaussian kernel

We recall (see Chapter 4, page 73) that the Gaussian kernel k_γ (also called squared exponential kernel) is defined by

$$k_\gamma(x, x') = \exp\Big(-\frac{\|x - x'\|^2}{\gamma}\Big), \quad (x, x') \in \mathcal{X}.$$

The bandwidth parameter $\gamma > 0$ controls the wiggliness of the functions sampled from the Gaussian process. In particular, the smaller γ is the more wiggly are the functions sampled from the GP process, as illustrated in Figure 11.3.

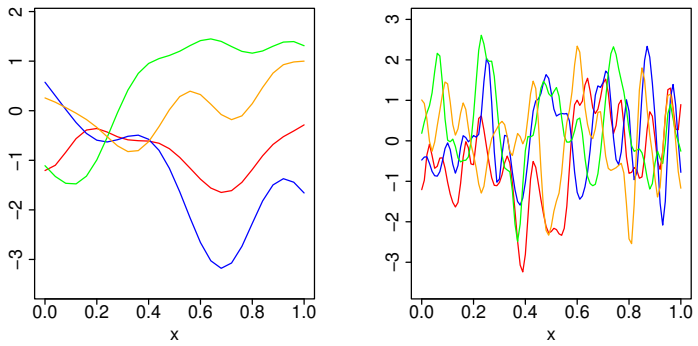

Figure 11.3: Four draws from a GP$(0, k_\gamma)$ process, with $\gamma^2 = 0.05$ (left plot) and $\gamma^2 = 0.001$ (right plot).

Remark: In Figure 11.3 the support of the GP is the same in the two plots but the smaller γ the more likely it is to sample a wiggly function f.

Some properties of the Gaussian kernel

- All functions in $H(k_\gamma)$ are continuous (this follows from Proposition 11.4).
- $\int_{\mathcal{X}} f^2(x)\mathrm{d}x < \infty$ for all $f \in H(k_\gamma)$.
- The functions in $H(k_\gamma)$ are infinitely many times differentiable.
- $H(k_\gamma)$ contains no polynomials, and hence no constant functions (except the zero function).
- If $\gamma_2 < \gamma_1$ then $H(k_{\gamma_1}) \subset H(k_{\gamma_2})$.

On the other hand, the support of the $\mathrm{GP}(0, k_\gamma)$ is the space of all continuous functions (and is therefore much bigger than the RKHS $H(k_\gamma)$).

The Matérn kernel

We recall (see Chapter 4, page 73) that the Matérn kernel $k_{\alpha,\gamma}$ is defined by

$$k_{\alpha,\gamma}(x,x') = \frac{2^{1-\alpha}}{\Gamma(\alpha)}\Big(\frac{\sqrt{2\alpha}\|x-x'\|}{\gamma}\Big)^{\alpha} K_{\alpha}\Big(\frac{\sqrt{2\alpha}\|x-x'\|}{\gamma}\Big),$$

with K_α the modified Bessel function of the 2nd kind of order α.

Some particular cases:

$$k_{1/2,\gamma}(x,x') = \exp\Big(-\frac{\|x-x'\|}{\gamma}\Big), \qquad \text{[exponential kernel]}$$

$$k_{3/2,\gamma}(x,x') = \Big(1+\frac{\sqrt{3}\|x-x'\|}{\gamma}\Big)\exp\Big(-\frac{\|x-x'\|}{\gamma}\Big)$$

$$k_{5/2,\gamma}(x,x') = \Big(1+\frac{\sqrt{3}\|x-x'\|}{\gamma}+\frac{5\|x-x'\|^2}{3\gamma^2}\Big)\exp\Big(-\frac{\|x-x'\|}{\gamma}\Big)$$

To understand the role of α on the set $H(k_{\alpha,\gamma})$ and on the support of the $\mathrm{GP}(0,k_{\alpha,\gamma})$ we need to introduce the notion of Sobolev spaces and that of a Hölder class of functions.

Remark: The role of the parameter γ is the same as for the Gaussian kernel.

Sobolev spaces and the space $H(k_{\alpha,\gamma})$

For $u \in \mathbb{N}_0^p$ and $f : \mathcal{X} \to \mathbb{R}$, let $D^u f$ be the function defined by

$$D^u f(x) = \frac{\partial^{|u|}}{\partial x_1^{u_1} \dots x_d^{u_d}} f(x), \quad x \in \mathcal{X}, \quad |u| = \sum_{i=1}^{p} u_i$$

and let $\|D^u f\|_{L_2(\mathcal{X})} = \left(\int_{\mathcal{X}} (D^u f(x))^2 \mathrm{d}x \right)^{1/2}$ be its $L_2(\mathcal{X})$ norm (with the convention $\|D^u f\|_{L_2(\mathcal{X})} = \infty$ if $D^u f$ does not exist).

For every integer $s \geq 1$ let

$$W_2^s(\mathcal{X}) = \left\{ f : \mathcal{X} \to \mathbb{R} : \ \|D^u f\|_{L_2(\mathcal{X})} < \infty, \, \forall u \in \mathbb{N}_0^p : \ |u| \leq s \right\}.$$

be the Sobolev space of order 2.

We then have the following result for the space $H(k_{\alpha,\gamma})$.

Theorem 11.1 ([7], Example 2.6) *Let $\mathcal{X} = [0,1]^p$ and α be such that $\alpha + p/2$ is an integer. Then, $H(k_{\alpha,\gamma}) = W_2^{\alpha+p/2}(\mathcal{X})$.*

Under the assumptions of the theorem it follows taht

- If $\alpha_2 < \alpha_1$ then $H(k_{\alpha_1,\gamma}) \subset H(k_{\alpha_2,\gamma})$.
- The larger the dimension p of the input variable x is the smoother are the functions in $H(k_{\alpha,\gamma})$.

Remark: Theorem 11.1 shows that the parameter $\alpha > 0$ of the Matérn kernel controls the smoothness of the functions in $H(k_{\alpha,\gamma})$, with the larger α the smoother the functions in $H(k_{\alpha,\gamma})$.

Remark: It can be shown that $k_{\alpha,\gamma} \to k_\gamma$ as $\alpha \to \infty$.

Hölder class of functions and support of $\mathrm{GP}(0, k_{\alpha,\gamma})$

For $\beta > 0$ we write $\beta = j + \eta$ with $j \in \mathbb{N}_0$ and $\eta \in (0, 1]$, and we let $\mathcal{C}^\beta(\mathcal{X})$ be such that $f \in \mathcal{C}^\beta(\mathcal{X})$ if and only if

1. The partial derivative of f order $(u_1, \ldots, u_p)$ exists for all $u \in \mathbb{N}_0^p$ such that $|u| \leq j$.

2. All the jth derivatives of f are Hölder continuous functions of order η.

Example: If $\beta \leq 1$ (so that $j = 0$) then $\mathcal{C}^\beta(\mathcal{X})$ is the set of Hölder continuous functions of order β.

We the have the following result for the support of $\mathrm{GP}(0, k_{\alpha,\gamma})$.

Theorem 11.2 ([16]) *Let $\mathcal{X} = [0, 1]^p$. Then, the $GP(0, k_{\alpha,\gamma})$ process takes values in $\mathcal{C}^{\underline{\alpha}}(\mathcal{X})$ for all $\underline{\alpha} < \alpha$.*

Let $u \in \mathbb{N}_0^p$ be such that $\alpha \leq |u| \leq \alpha + p/2$. Then, $\|D^u f\|_{L_2(\mathcal{X})} < \infty$ for all $f \in H(k_{\alpha,\gamma})$ while for every $\underline{\alpha} < \alpha$ the set $\mathcal{C}^{\underline{\alpha}}(\mathcal{X})$ contains function f such that $\|D^u f\|_{L_2(\mathcal{X})} = \infty$.

Therefore, the functions sampled from the $\mathrm{GP}(0, k_{\alpha,\gamma})$ process are less smooth than those in $H(k_{\alpha,\gamma})$ (and thus less smooth than the mean function f_n).

The above result also shows that the larger α the smoother the functions sampled from the $\mathrm{GP}(0, k_{\alpha,\gamma})$ process (see Figure 11.4 below for an illustration).

Impact of α on $\mathrm{GP}(0, k_{\alpha,\gamma})$ and some comments

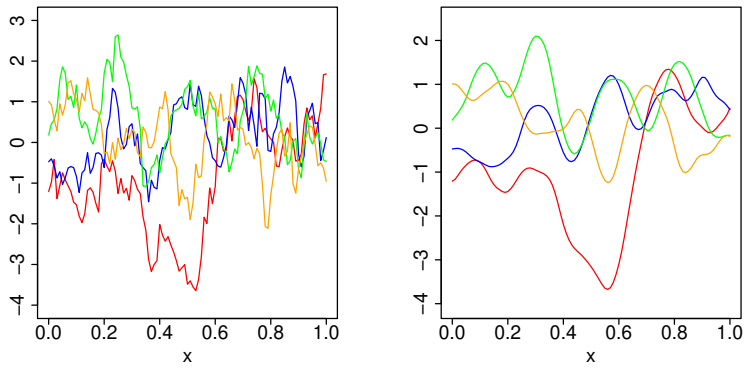

Figure 11.4: Four draws from a $\mathrm{GP}(0, k_{\alpha,0.1})$ process, with $\alpha = 1/2$ (left plot) and $\alpha = 7/2$ (right plot).

Some comments:

1. Both the Gaussian and the Matérn kernels can be written as $k(x, x') = h(\|x - x'\|/\gamma)$ for some function $h : [0, \infty) \to \mathbb{R}$. Kernels of this form are said to be translation invariant, in the sense that $k(x + z, x' + z) = k(x, x')$.

2. In practice, when using a translation invariant kernel, it may be worth choosing a different bandwidth for each component of x, that is to let k be such that

$$k(x, x') = h\left(\sqrt{\sum_{j=1}^{p} \frac{(x_j - x'_j)^2}{\gamma_j^2}}\right). \tag{11.11}$$

   Remark: This however increases the number of parameters to choose.

   Remark: If $p > 1$ and only one bandwidth parameter is used, if $k(x, x') = h(\|x - x'\|/\gamma)$ then it is important that the variables $\{x^0_{(j)}\}_{j=1}^p$ are on the same scale before choosing γ (see below for how to choose γ in practice).

Convergence result: Set up and preliminary remarks

We assume henceforth that there exists a function f_0 and a $\sigma_0^2 > 0$ such that

$$Y_i^0 = f_0(x_i^0) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}_1(0, \sigma_0^2). \tag{11.12}$$

As shown above, the support of the posterior distribution is much larger than the RKHS $H(k)$, the function space the mean function f_n belongs to. This suggests that the posterior distribution should be able to learn f_0 even when k is such that $f_0 \notin H(k)$.

However, the rate at which the posterior distribution learns f_0 is very sensitive to the choice of k, as shown in the following result.

Theorem 11.3 ([16]) *Assumes* (11.12), *where*

$$\mathcal{X} = [0,1]^p, \quad X_i^0 \overset{\text{iid}}{\sim} g(x)\mathrm{d}x \tag{11.13}$$

with the density g such that

$$\exists c > 0 : \; c \leq g(x) \leq c^{-1}, \quad \forall x \in \mathcal{X}, \tag{11.14}$$

and consider the Bayesian model (11.3) *with $\lambda = \sigma_0^2$ and $k = k_{\alpha,\gamma}$ for an $\alpha > p/2$. Assume that $f_0 \in \mathcal{C}^\beta(\mathcal{X}) \cap W_2^\beta(\mathcal{X})$ for some $\beta > p/2$. Then,*

$$\mathbb{E}\big[\|f_n - f_0\|_{L_2(\mathcal{X})}^2\big]^{1/2} = \mathcal{O}\big(n^{-\frac{\min(\alpha,\beta)}{2\alpha+p}}\big). \tag{11.15}$$

Reminder: $\|g\|_{L_2(\mathcal{X})} = \big(\int_{\mathcal{X}} g(x)^2 \mathrm{d}x\big)^{1/2}$.

Comments on Theorem 11.3

1. The conclusion of Theorem 11.3 holds for any choice of $\alpha > p/2$. However, since
$$\frac{\min(\alpha, \beta)}{2\alpha + p} \to 0$$
as $\alpha \to \infty$ and as $\alpha \to 0$, it follows that the convergence rate deteriorates as α increases/decreases.

2. Theorem 11.3 shows that even if $f_n \in H(k_{\alpha,\gamma})$ for all n and $f_0 \notin H(k_{\alpha,\gamma})$, the mean function f_n converges in some sense to f_0.

   In fact, it is easily checked that the best rate in (11.15) is obtained for $\alpha = \beta$, in which case $f_0 \notin H(k_{\alpha,\gamma})$ (recall that, by Theorem 11.1, $H(k_{\alpha,\gamma}) = W_2^{\alpha+p/2}(\mathcal{X})$).

3. When $\alpha = \beta$ Theorem 11.3 shows that
$$\mathbb{E}\big[\|f_n - f_0\|^2_{L_2(\mathcal{X})}\big]^{1/2} = \mathcal{O}(n^{-\frac{\beta}{2\beta+p}}). \tag{11.16}$$
   The rate in (11.16) turns out to be optimal.

   More precisely, the rate in (11.16) is minimax, in the sense that there exists a constant $c > 0$ such that
$$\liminf_{n\to\infty} \inf_{\tilde{f}_n} \sup_{f_0 \in \mathcal{C}^\beta(\mathcal{X})} n^{\frac{\beta}{2\beta+p}} \mathbb{E}\big[\|\tilde{f}_n - f_0\|^2_{L_2(\mathcal{X})}\big]^{1/2} \geq c$$
   were the $\inf_{\tilde{f}_n}$ is on all the estimators $\tilde{f}_n$ of f_0.

Model hyperparameters in a non-parametric setting

As shown/mentioned above, and contrary to what happens for parametric models, in the nonparametric setting

- The HPD regions can have a very poor asymptotic coverage, even when the model is well-specified.
- The convergence rate of f_n towards f_0 is sub-optimal (i.e. not minimax), unless β is known (which is never the case!).

The theory suggests that the two above problems can be solved using data driven hyperparameters.

In particular, using the data to choose λ and the parameters entering in the definition of the kernel k is particularly important.

Indeed, λ is a key hyperparameter since:

- The link between GPR and kernel ridge regression shows that λ has and impact on the "regularity" (i.e. the norm) of f_n (both when σ^2 is known and when σ^2 is unknown).
- In the model with known σ^2, the convergence result of Theorem 11.3 holds assuming $\lambda = \sigma_0^2$.

Letting ψ denote the vector that contains all the parameters of the kernel[a], ψ is a key hyperparameter since it can influence

- The function space the mean function f_n belongs to.
- The support of the posterior distribution.
- The convergence rate of f_n towards f_0.

[a] For instance $\psi = \gamma$ for the Gaussian kernel and $\psi = (\alpha, \gamma)$ for the Matérn kernel.

Choosing ψ: Hierarchical prior

In this approach, we treat the hyperparameter ψ as a "regular" parameter that we need to learn from the data. Following the standard Bayesian approach, we then choose a prior distribution π_ψ for this new parameter and base the inference on the posterior distribution (assuming that σ^2 is known)

$$\pi(f, \psi | y^0_{1:n}) = \pi(f|\psi, y^0_{1:n})\pi(\psi|y^0_{1:n}), \quad \pi(\psi|y^0_{1:n}) \propto p(y^0_{1:n}|\lambda, k)\pi_\psi(\psi).$$

In this case, the posterior distribution of f given $y^0_{1:n}$ is given by

$$\pi(f|y^0_{1:n}) = \int \pi(f|\psi, y^0_{1:n})\pi(\psi|y^0_{1:n})\mathrm{d}\psi$$

where $\pi(f|\psi, y^0_{1:n})$ is the posterior distribution associated with model (11.3) when ψ is taken as parameter for k.

More precisely, let $k = k_\gamma$ and consider the following Bayesian model that uses a hierarchical prior for the bandwidth parameter γ

$$\begin{aligned} &Y^0_i = f(x^0_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}_1(0, \sigma^2), \quad \pi_{\sigma^2} = \delta_\lambda \\ &f|\gamma \sim \mathrm{GP}(0, k_{\gamma^2}), \quad \gamma^{-1} \sim \mathrm{Gamma}(a'_0, b'_0) \end{aligned} \tag{11.17}$$

where $a'_0, b'_0 > 0$ are some hyperparameters.

We then have the following result.

Theorem 11.4 ([17]) *Let $\mathcal{X} = [0,1]^p$ and assume (11.12) with fixed design $(x^0_i)_{i\geq 1}$. Assume that $\lambda = \sigma_0^2$ and that $f_0 \in \mathcal{C}^\beta(\mathcal{X})$ for a $\beta > 0$. Then, the posterior distribution associated with (11.17) concentrates on f_0 at rate*

$$\mathcal{O}\Big(n^{-\frac{\beta}{2\beta+p}}(\log n)^{\frac{4\beta+p}{4\alpha+2p}}\Big).$$

Remark: Compared to the minimax rate there is an extra log-term, and thus the procedure is nearly adaptive.

Hierarchical prior: Some comments

1. The rational behind Theorem 11.4 is that, as mentioned above, increasing/decreasing γ makes the functions sampled from $\text{GP}(0, k_\gamma)$ smoother/rougher, making the $\text{GP}(0, k_\gamma)$ more suitable as a prior for smooth/rough functions.

2. It can be shown that the credible sets obtained by inflating (by a sufficiently large constant $M > 0$) the length of the HPD regions derived from the hierarchical model (11.17) contain (in some sense) f_0 with probability tending to one as $n \to \infty$ (see [14] for more details).

3. The posterior distribution associated to the model (11.17) is intractable, and difficult to approximate.

   For this latter reason, the empirical Bayes approach is usually preferred in practice.

Choosing (λ, ψ): Empirical Bayes

In empirical Bayes methods, the parameters of the prior distribution are chosen so that to maximize the marginal likelihood of the observations $y^0_{1:n}$.

Following this approach, we choose $(\lambda, \psi) = (\lambda_n, \psi_n)$ where, by Proposition 11.2 (and considering the case where σ^2 is known)

$$\begin{aligned}(\lambda_n, \psi_n) &\in \operatorname*{argmax}_{\lambda,\psi} \log p(y^0_{1:n}|\lambda, k) \\ &= \operatorname*{argmax}_{\lambda,\psi} \Big(-\frac{1}{2}\log|\boldsymbol{K}_n + \lambda \boldsymbol{I}_n| - \frac{1}{2}(y^0_{1:n})^\top \big(\boldsymbol{K}_n + \lambda \boldsymbol{I}_n\big)^{-1} y^0_{1:n} \Big).\end{aligned}$$

Remark: This optimization problem is not convex.

Remark: When the bandwidth parameter γ of a kernel is chosen using an empirical Bayes approach the same result as for the hierarchical prior holds for the "inflated" HPD regions (see [14]).

The empirical Bayes approach is often viewed as a cheap approximation of the hierarchical prior approach. Indeed, since in this latter approach the posterior distribution of the hyperparameters is

- proportional to $p(y^0_{1:n}|\lambda, k)\pi_\psi(\psi)$
- expected to concentrate on a given value

it is reasonable to expect that the two approaches will lead to a similar posterior distribution $\pi(f|y^0_{1:n})$ (at least for n large enough).

Remark: Empirical Bayes methods are not truly Bayesian, since the prior distribution also depends on the observations $y^0_{1:n}$.