A prior near-ignorance Gaussian Process model for nonparametric regression

Francesca Mangili
IDSIA, USI-SUPSI, Lugano, Switzerland
{francesca}@idsia.ch

Abstract

A Gaussian Process (GP) defines a distribution over functions and thus it is a natural prior distribution for learning real-valued functions from a set of noisy data. GPs offer a great modeling flexibility and have found widespread application in many regression problems. A GP is fully defined by a mean function that represents our prior belief about the shape of the regression function and a covariance function, relating the function values at different covariates. In the absence of prior information, one typically assumes a GP with zero mean function. Therefore, a priori, it is assumed that the regression function is constantly equal to zero. The aim of this paper is to model a situation of prior near-ignorance about the GP mean function. For this we consider the set of all GPs with fixed covariance function and constrant mean function free to vary from $-\infty$ to $+\infty$. We apply the model with constant mean function to hypothesis testing; in particular we test the equality of two regression functions and show that the use of a prior near-ignorance model allows the test to automatically detect when a reliable decision cannot be made based on the available data. Finally, we propose a generalization of this model that allows considering other sets of prior mean functions.

Keywords: Gaussian Process, prior near-ignorance, nonparametric regression, hypothesis testing, Bayesian nonparametrics.

1. Introduction

Gaussian processes (GPs) extend multivariate Gaussian distributions to infinite dimensionality, thus defining a distribution over functions that can be used as prior distribution for inferences about an unknown function $f(x)$.

GPs have found widespread use in different application domains such as classification, regression etc. [9, 8, 6, 12, 11, 4]. The reason of such success can be attributed to the great modelling flexibility of GPs, which are often used in situations where little is known about $f(x)$. However, GPs are not completely free-form, since a GP is completely specified by its mean function and covariance function. The covariance function describes the relation between observations from the same process. A multitude of possible families exists for the covariance function, including squared exponential, polynomial, periodic, etc. (see [12]), among which the squared exponential family is by far the most popular. On the other side, the mean function represents our prior belief about the form of the regression function. In the absence of prior knowledge, which is typically the case, the mean function is assumed to be zero everywhere and, to comply with this assumption, data are transformed to have zero mean. However, this seems quite a poor representation of the condition of prior ignorance about $f(x)$. In this work we improve this representation by considering a set of GP priors with mean functions free to vary in the set of all constant functions. As the expectation of $f(x^*)$ at the covariate x^* w.r.t. the prior GPs can vary in $[-\infty, +\infty]$, this set of priors is a model of prior ignorance about $f(x)$. Prior ignorance and learning from data are usually conflicting properties [13, Sec. 7.4],[10, 14]. However, in [3, 2] it is shown that, for Gaussian distributions, if we let the variance to depend on the mean, prior near-ignorance and learning from data can be guaranteed at the same time. In this work, we apply this idea to GPs. In order for the GP model to be able to learn from data, we add to the covariance function a constant term increasing with the prior mean function.

We will use this set of priors to test the difference between two regression function given two samples of noisy observations. A nonparametric Bayesian test for the equality of regression functions based on GPs is described in [1]. In that work it is assumed that the covariates of the two samples cover the same range of values, and the comparison between the regression functions is limited to that range of values, assuming that, having no data outside of it, nothing can be stated about the difference or equality of the two functions. Using the Imprecise GP (IGP) it is possible to perform the equality test without worrying about the distribution of the covariates, as the imprecise approach is able to identify those instances where the decision is prior dependent and thus it automatically detects when a reliable decision cannot be made.

Finally, we introduce a IGP model that generalizes the previous one by

considering all GPs with mean function $Mh(x)$ where $M > 0$ and $h(x)$ belongs to a set of functions $\mathcal{H}$. We derive the conditions that $\mathcal{H}$ has to satisfy to make prior near-ignorance and leaning hold for the IGP model. From this model we can derive the IGP with constant mean as well as well as other models considering different/larger sets of prior mean functions.

2. Gaussian Process

Consider the regression model

$$y = f(x) + v, \tag{1}$$

where $x \in \mathcal{X} \subseteq \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ and $v \sim \mathcal{N}(0, \sigma_n^2)$ is a white noise, and assume that we observe the data (x_i, y_i) for $i = 1, \ldots, n$. Our goal is to employ these observations to make inferences about the unknown function $f(x)$. Following the Bayesian estimation approach, we place a prior distribution on $f(x)$, and employ the observations to compute its posterior distribution; finally we use this posterior to make inferences about $f(x)$. Since $f(x)$ is a function, the Gaussian process is a natural prior distribution for it [6, 12]. Formally,

Definition 1. *Let $\mu : \mathbb{R} \to \mathbb{R}$ and $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ be a positive definite symmetric function.*[1] *A function $f(x)$ with $x \in \mathbb{R}$ is said to be distributed according to a Gaussian process with mean function μ and covariance kernel k if for any finite set of covariates $x_1^*, \ldots, x_m^*$, the vector $[f(x_1^*), \ldots, f(x_m^*)]^T$ has a multivariate m-dimensional Gaussian distribution with mean $[\mu(x_1^*), \ldots, \mu(x_m^*)]^T$ and covariance matrix with (i, j)-th entry $k(x_i^*, x_j^*)$, $i, j = 1, \ldots, m$.*

In the following, $GP(\mu(x), k_{\boldsymbol{\theta}}(x, x'))$ will denote a GP with mean function $\mu(x)$ and covariance function $k_{\boldsymbol{\theta}}(x, x') : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$. The subscript $\boldsymbol{\theta}$ has been introduced to highlight that the covariance function usually depends on a vector of hyperparameters $\boldsymbol{\theta}$ [12]. If $f(x) \sim GP(\mu(x), k_{\boldsymbol{\theta}}(x, x'))$, then, for any fixed m points $\mathbf{x}^* = [x_1^*, \ldots, x_m^*]^T$, the vector $\mathbf{f}^* = [f(x_1^*), \ldots, f(x_m^*)]^T$ is Gaussian distributed:

$$p(\mathbf{f}^* | \mathbf{x}^*, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, K^{**}), \tag{2}$$

[1] A symmetric function $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is said to be positive definite if for any $\mathbf{x} = [x_1^*, \ldots, x_m^*]^T$ with $x_i^* \in \mathbb{R}$, the $m \times m$ matrix $[k(x_i^*, x_j^*)]_{ij}$ is positive definite.

with mean $\boldsymbol{\mu}^* = \mu(\mathbf{x}^*)$ and covariance matrix $K^{**} = [k_{\boldsymbol{\theta}}(x_i^*, x_j^*)]_{ij}$ for each $i, j = 1, \ldots, m$. Consider a set of n inputs $\mathbf{x} = [x_1, \ldots, x_n]^T$ and a vector of noisy output data $\mathbf{y} = [y_1, \ldots, y_n]^T$. Based on the training data (x_i, y_i) for $i = 1, \ldots, n$, and given a test input $\mathbf{x}^*$, we wish to find the posterior distribution of $\mathbf{f}^* = [f(x_1^*), \ldots, f(x_m^*)]^T$. From (1) and the properties of the Gaussian distribution, it follows that [12, Sec. 2.2]:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}^* \end{bmatrix}, \begin{bmatrix} K + \sigma_n^2 \mathbf{I} & K^{*T} \\ K^* & K^{**} \end{bmatrix} \right), \tag{3}$$

where $\boldsymbol{\mu} = \mu(\mathbf{x})$, $K = [k_{\boldsymbol{\theta}}(x_i, x_j)]_{ij}$, $i, j = 1 \ldots, n$ and $K^* = [k_{\boldsymbol{\theta}}(x_i^*, x_j)]_{ij}$, $i = 1 \ldots, m$, $j = 1 \ldots, n$. When σ_n^2 is not known, it can also be considered a hyperparameter. Hence, we introduce the extended vector $\boldsymbol{\theta}_n = [\boldsymbol{\theta}, \sigma_n^2]$ of all model hyperparameters, including the noise variance. The posterior distribution of $\mathbf{f}^*$ is then

$$p(\mathbf{f}^* | \mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_n) = \mathcal{N}(\mathbf{f}^*; \hat{\boldsymbol{\mu}}^*, \hat{K}^{**}), \tag{4}$$

with posterior mean and covariance given by:

$$\hat{\boldsymbol{\mu}}^* = \boldsymbol{\mu}^* + K^*(K + \sigma_n^2 \mathbf{I})^{-1}(\mathbf{y} - \boldsymbol{\mu}), \tag{5}$$

$$\hat{K}^{**} = K^{**} - K^*(K + \sigma_n^2 \mathbf{I})^{-1} K^{*T}. \tag{6}$$

Once we have computed $p(\mathbf{f}^* | \mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_n)$ we can make any inference about $\mathbf{f}^*$.

GP models use a kernel to define the covariance between any two function values: $Cov(f(x), f(x')) = k_{\boldsymbol{\theta}}(x, x')$. A popular choice is the squared exponential kernel:

$$k_{\boldsymbol{\theta}}(x, x') = \sigma_k^2 \exp\left[-\frac{1}{2} \frac{(x - x')^2}{\ell^2} \right], \tag{7}$$

with hyperparameters $\boldsymbol{\theta} = (\sigma_k, \ell) > 0$. This kernel assumes that the correlation between two function values decreases with the distance of their covariates. Observations whose covariates have a distance much larger than the lenghtscale ℓ are almost uncorrelated. A multitude of other possible families of covariance functions exists (polynomial, periodic, etc.), and more can be obtained by kernel composition, as positive definite kernels (i.e. those which define valid covariance functions) are closed under addition and multiplication. Once we have selected a kernel or a particular kernel composition, we

must determine the values of the hyperparameters $\boldsymbol{\theta}_n$. The proper Bayesian procedure is to choose a prior for $\boldsymbol{\theta}_n$ and then determine the posterior distribution of the quantities of interest. For instance, inferences on $\mathbf{f}^*$ can be carried out by marginalizing out $\boldsymbol{\theta}_n$:

$$p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}) = \int p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_n) p(\boldsymbol{\theta}_n|\mathbf{x}^*, \mathbf{x}, \mathbf{y})) d\boldsymbol{\theta}_n.$$

No closed form solution exists for $p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y})$ or for the posterior of the hyperparameters and, therefore, inferences must be computed numerically by Markov Chain Monte Carlo methods (MCMC). The convergence of MCMC methods can be quite slow when the dimension of $\boldsymbol{\theta}_n$ is high and, therefore, when we are not interested in the posterior distribution of $\boldsymbol{\theta}_n$, we can approximate the marginal of $\mathbf{f}^*$ by plugging the maximum a-posteriori (MAP) estimate for $\boldsymbol{\theta}_n$ into (4). In other words, we maximize w.r.t. $\boldsymbol{\theta}_n$ the joint marginal probability of $\mathbf{y}$ and $\boldsymbol{\theta}_n$, whose logarithm can be computed analytically [12, Ch.2]:

$$\begin{aligned} \log p(\mathbf{y}, \boldsymbol{\theta}_n|\mathbf{x}) = &- \frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}^*)^T (K_n)^{-1} (\mathbf{y} - \boldsymbol{\mu}^*) + \\ &- \frac{1}{2} \log |K_n| - \frac{n}{2} \log 2\pi + \log p(\boldsymbol{\theta}_n) \end{aligned} \tag{8}$$

where $K_n = K + \sigma_n^2 \boldsymbol{I}$.

3. Imprecise Gaussian Process with constant mean function

In this section we relax the assumption of zero-mean function and consider a set of GPs with constant mean function varying from $-\infty$ to $+\infty$.

Definition 2. *Given a covariance kernel $k_{\boldsymbol{\theta}}(x, x')$ and a constant $c > 0$, we define the constant mean Imprecise Gaussian Process (c-IGP) as the set of GPs:*

$$\mathcal{G}_c = \left\{ GP\left(Mh, k_{\boldsymbol{\theta}}(x, x') + \frac{1 + M}{c} \right) : \ h = \pm 1, \ M \geq 0 \right\}.$$

As discussed below, the constant parameter c determines the degree of posterior imprecision.
The c-IGP includes all GPs with constant mean function and covariance function made of two components: a first one, $k_{\boldsymbol{\theta}}(x, x')$, hereafter referred to as *base kernel*, which is chosen according to the specific application and is identical for all GPs in $\mathcal{G}_c$, and a constant component $(M+1)/c$ proportional

to $M+1$. The constant component allows the model to learn from data, as it forces the covariance to increase with the prior mean. In [2, pag. 22], it is shown for the one-parameter exponential family that if the product $n_0|y_0|$ of the number of pseudo-observations n_0 (which represent the strength of the prior and for a Gaussian prior is given by the inverse of its variance) and the absolute value of the pseudo-observation y_0 (which represent our prior opinion about the parameter value and for a Gaussian prior is given by its mean) is bounded, then learning from data is guaranteed. For the c-IGP model, this holds for each individual x^* because the prior about $f(x^*)$ is a Gaussian distribution with mean Mh (corresponding to y_0) and variance $k_{\boldsymbol{\theta}}(x^*, x^*) + \frac{M+1}{c}$ (corresponding to $1/n_0$), and thus:

$$n_0|y_0| = \frac{M|h|}{k_{\boldsymbol{\theta}}(x^*, x^*) + \frac{M+1}{c}} \leq c.$$

Notice however that this guarantees only learning from data with covariate equal to x^*.

Proposition 1. *The c-IGP is a model of prior ignorance about the expectation of $f(x^*)$ in the sense that for any covariate x^* it holds*

$$\inf_{M,h} E[f(x^*)] = -\infty, \quad \sup_{M,h} E[f(x^*)] = +\infty.$$

The proof of this and the following propositions and theorems can be found in the Appendix.

A posteriori we have the following result.

Theorem 1. *Let $\mathbf{x}$ be a vector of inputs and $\mathbf{y}$ a set of noisy observations of $f(\mathbf{x})$ with $f(x) \sim GP(Mh, k_{\boldsymbol{\theta}} + \frac{M+1}{c})$, and let $\mathbf{k}_x = [k_{\boldsymbol{\theta}}(x, x_1), \ldots, k_{\boldsymbol{\theta}}(x, x_n)]^T$. The posterior distribution of f is a GP with mean function*

$$\hat{\mu}(x) = \mathbf{k}_x^T K_n^{-1}(\mathbf{y} - \hat{y}\mathbb{1}_n) + \hat{y} \tag{9}$$

with $\hat{y} = \frac{(M+1)\mathbf{s}_k^T\mathbf{y} + cMh}{c + (M+1)S_k}$, and covariance function

$$\begin{aligned}\hat{k}(x, x') =& k_{\boldsymbol{\theta}}(x, x') - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_{x'} + \\ & \frac{(M+1)(1 - \mathbf{k}_x^T\mathbf{s}_k)(1 - \mathbf{k}_{x'}^T\mathbf{s}_k)}{c + (M+1)S_k},\end{aligned} \tag{10}$$

where $\mathbf{s}_k = K_n^{-1}\mathbb{1}_n$, $S_k = \mathbb{1}_n^T K_n^{-1}\mathbb{1}_n$, and $\mathbb{1}_n$ is a n-dimensional vector of ones.

The posterior mean function is the same that would have been obtained from the prior $GP(\hat{y}, k_\theta(x, x'))$. We can interpret $\hat{y}$ as an adjusted mean obtained by combining the prior mean Mh and a weighted average of the observations $\mathbf{y}$. This is due to the constant term in the covariance function which introduces a correlation between all function values (does not matter how distant their covariates are). As this constant term goes to infinity, that is, as $c \to 0$, the adjusted mean becomes $\hat{y} \to \frac{\mathbf{s}_k^T}{S_k}\mathbf{y}$ which is independent of h and M and the posterior distribution is a GP with mean and covariance functions

$$\begin{aligned}\lim_{c\to 0}\hat{\mu}(x) &= \mathbf{k}_x^T K_n^{-1}\mathbf{y} + (1 - \mathbf{k}_x^T s_k)\frac{s_k^T}{S_k}\mathbf{y},\\ \lim_{c\to 0}\hat{k}(x, x') &= k_{\boldsymbol{\theta}}(x, x') - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_{x'} + \\ &\qquad \frac{(1 - \mathbf{k}_x^T\mathbf{s}_k)(1 - \mathbf{k}_{x'}^T\mathbf{s}_k)}{S_k}.\end{aligned}$$

Notice that, for $c \to 0$ we have a precise model, as IGP posterior inferences are not influenced by the mean function of the prior and converge to a single GP. This prior can, thus, be interpreted as a partially uninformative prior (inferences still depend on the base kernel).

As discussed in Section 2, MAP estimates of the hyperparameters $\boldsymbol{\theta}_n$ are used in the model. However the different priors in the IGP set produce different estimates, whereas the IGP model here proposed requires the same set of hyperparameters for all priors. The issue is, then, which of the IGP priors should be used to estimate $\boldsymbol{\theta}_n$. Notice that posterior inferences obtained for any c always encompass those obtained for $c \to 0$ (see Theore 3 below). Hence, we use the MAP estimate of $\boldsymbol{\theta}_n$ given by this prior.

Theorem 2. *MAP estimates of the hyperparameters of the* $GP(Mh, k_{\boldsymbol{\theta}} + \frac{M+1}{c})$ *with* $c \to 0$ *are obtained by maximizing* $L(\mathbf{y}, \boldsymbol{\theta}_n|\mathbf{x}) + \log p(\boldsymbol{\theta}_n)$ *where*

$$L(\mathbf{y}, \boldsymbol{\theta}_n|\mathbf{x}) = \frac{1}{2}\left(\mathbf{y}^T K_n^{-1}\mathbf{y} - \frac{(\mathbf{y}^T\mathbf{s}_k)^2}{S_k} - \log S_k|K_n|\right) \tag{11}$$

is, up to an additive constant, the logarithm of the joint marginal likelihood of $\mathbf{y}, \boldsymbol{\theta}_n$

From (9) we can derive the upper and lower expectations of $f(x)$.

Theorem 3. *Under the c-IGP model, if* $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| \leq 1 + \frac{c}{S_k}$*, the upper and lower bounds,* $\overline{\mu}(x)$ *and* $\underline{\mu}(x)$*, of* $\hat{\mu}(x)$ *are*

$$\overline{\mu}(x) = \mathbf{k}_x^T K_n^{-1}\mathbf{y} + (1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k^T}{S_k}\mathbf{y} + c\ \frac{|1 - \mathbf{k}_x^T\mathbf{s}_k|}{S_k} \tag{12}$$

$$\underline{\mu}(x) = \mathbf{k}_x^T K_n^{-1}\mathbf{y} + (1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k^T}{S_k}\mathbf{y} - c\ \frac{|1 - \mathbf{k}_x^T\mathbf{s}_k|}{S_k}, \tag{13}$$

which, if $1 - \mathbf{k}_x^T\mathbf{s}_k \geq 0$*, are obtained for* $M \to \infty,\ h = 1$ *(upper) and* $M \to \infty, h = -1$ *(lower), while, if* $1 - \mathbf{k}_x^T\mathbf{s}_k < 0$*, are obtained for* $M \to \infty,\ h = -1$ *(upper) and* $M \to \infty, h = 1$ *(lower).*

If, instead, $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| > 1 + \frac{c}{S_k}$ *and* $(1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k\mathbf{y}}{S_k} > 0$*, the upper bound is found for* $M \to \infty$ *and* $h = 1$ *and the lower for* $M = 0$*; they are given by*

$$\overline{\mu}(x) = \mathbf{k}_x^T K_n^{-1}\mathbf{y} + (1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k^T}{S_k}\mathbf{y} + c\ \frac{1 - \mathbf{k}_x^T\mathbf{s}_k}{S_k}$$

$$\underline{\mu}(x) = \mathbf{k}_x^T K_n^{-1}\mathbf{y} + \ (1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k^T\mathbf{y}}{c + S_k}.$$

Finally, if $(1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k\mathbf{y}}{S_k} < 0$*, the upper bound is found for* $M = 0$ *and the lower for* $M \to \infty$ *and* $h = 1$*.*

From Theorem 3 we can see that the imprecision of the model verifies

$$\overline{\mu}(x) - \underline{\mu}(x) \geq 2c\ \frac{|1 - \mathbf{k}_x^T\mathbf{s}_k|}{S_k},$$

where the equality holds if the condition $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| \leq 1 + \frac{c}{S_k}$ is verified. In this case, we can see that the imprecision is symmetric with respect to the posterior mean of the prior with $c \to 0$. Parameter c determines the degree of imprecision of the model. A large value of c implies a large imprecision. For $c \to \infty$ we have a vacuous model that cannot learn from data.

When the condition $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| \leq 1 + \frac{c}{S_k}$ is verified, by a simple rewriting of equations (12)-(13) it can be seen that $\overline{\mu}(x)$ and $\underline{\mu}(x)$ are equivalent to the posterior mean given the prior $GP(\hat{y}, k_\theta(x, x'))$ with adjusted mean

$$\hat{y}_b = \frac{\mathbf{s}_k^T}{S_k}\mathbf{y} \pm \frac{c}{S_k}.$$

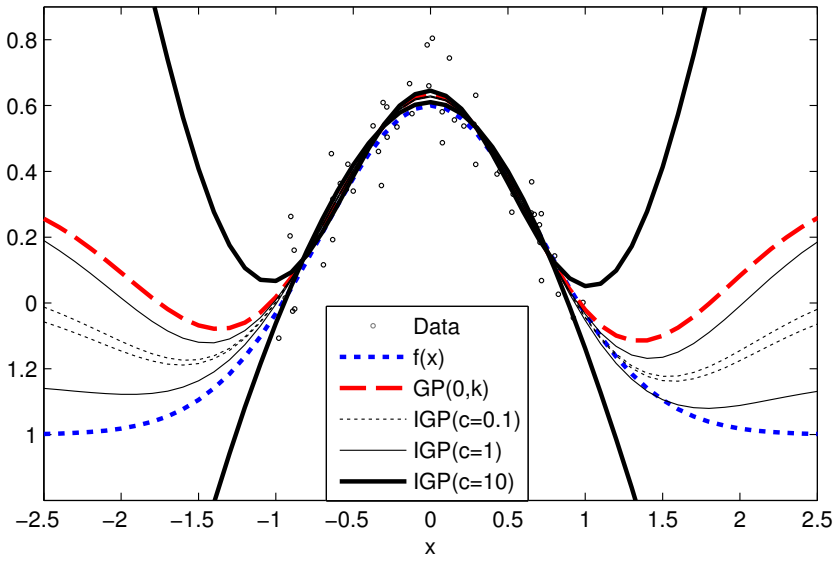

Figure 1: GP and c-IGP estimates of the function $f(x)$ given $n = 50$ observations.

Example 1. *A sample of $n = 50$ observations affected by Gaussian noise with $\sigma_n = 0.1$ is drawn from the function $f(x) = \exp(-x^2)$. The covariates $x_1, \ldots, x_n$ are uniformly distributed in $[-1, 1]$, i.e., $x \sim U[-1, 1]$. The function is modeled by the precise GP process $GP(0, k_\theta)$ and the c-IGP with the squared-exponential kernel in (7) as base kernel k_θ. Figure 1 shows the posterior expectation of the GP and the upper and lower expectations of the c-IGP for different values of c. Notice that in the region where there are observations ($x \in [-1, 1]$) the imprecision remains very small even when c is large, whereas it increases significantly outside this region.*

It is often useful to compute pointwise credible intervals $CI_f(x, \alpha) = [\underline{f}_{x,\alpha}, \overline{f}_{x,\alpha}]$ for the value of $f(x)$. Using a GP prior $f(x) \sim GP(\mu(x), k_\theta(x, x'))$, a posterior $(1 - \alpha)\%$ credible interval for the value of $f(x)$ is $CI_f(x, \alpha) = [\hat{\mu}(x) - z_{\alpha/2}\sqrt{\hat{k}(x, x)}, \hat{\mu}(x) + z_{\alpha/2}\sqrt{\hat{k}(x, x)}]$ with $z_{\alpha/2}$ the $1 - \alpha/2$ percentile of the standard normal distribution. Hence we have that the posterior probability $P\big(f(x) \in CI_f(x, \alpha)\big)$ is $1 - \alpha$. In the imprecise case, we define the credible interval by imposing that the upper posterior probabilities $\overline{P}\big(f(x) < \underline{f}_{x,\alpha}\big)$ and $\overline{P}\big(f(x) > \overline{f}_{x,\alpha}\big)$ are equal to $\alpha/2$. This implies that $P\big(f(x) \in CI_f(x, \alpha)\big) \geq 1 - \alpha$ for all GPs in $\mathcal{G}$.

Theorem 4. *Under the c-IGP model, the interval* $CI_\alpha = [\underline{f}_{x,\alpha} = \underline{\mu}(x) - z_{\alpha/2}\sigma_{f_x}, \overline{f}_{x,\alpha} = \overline{\mu}(x) + z_{\alpha/2}\sigma_{f_x}]$ *with*

$$\sigma^2_{f_x} = k_{\boldsymbol{\theta}}(x,x) - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_x + \frac{(1-\mathbf{k}_x^T\mathbf{s}_k)^2}{S_k},$$

verifies

$$\overline{P}\big(f(x) < \underline{f}_x\big) \leq \alpha/2, \quad \overline{P}\big(f(x) > \overline{f}_x\big) \leq \alpha/2$$

where the equality holds if $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| \leq 1 + \frac{c}{S_k}$.

Notice that with respect to the precise model $GP(0, k_{\boldsymbol{\theta}}(x,x))$ the value of $\sigma^2_{f_x}$ increases only for the term $\frac{(1-\mathbf{k}_x^T\mathbf{s}_k)^2}{S_k}$. Moreover, $\sigma^2_{f_x}$ does not depend on c and is the same given by the precise model with $c \to 0$. Then, the width of the pointwise CIs for $c > 0$ increases only for a term equal to the difference between the upper and lower expectation of $f(x)$. Figure 2 compares the credible intervals obtained using the prior $GP(0, k_{\boldsymbol{\theta}}(x,x))$ and the c-IGP model. As for the expectation, the width of the CIs remains small for $x \in [-1,1]$ and increases significantly outside this region.

Data analyst are often interested also in simultaneous credible regions (SCR) for the value of f at multiple covariate values. Given a vector of m covariates $\mathbf{x}^*$, a $(1-\alpha)\%$ SGR for $f(\mathbf{x}^*)$ includes all vectors $\mathbf{f}^*$ that verify

$$(\mathbf{f} - \hat{\boldsymbol{\mu}}^*)^T(\hat{K}^{**})^{-1}(\mathbf{f} - \hat{\boldsymbol{\mu}}^*) < \chi^{-1}(1-\alpha|m), \tag{14}$$

where $\chi^{-1}(1-\alpha|m)$ is the $(1-\alpha)$-quantile of a Chi-squared distribution with m degrees of freedoms. For the the condition 14 to be verified by all priors in the c-IGP, it has to be verified by the upper bound of $(\mathbf{f}-\hat{\boldsymbol{\mu}}^*)^T(\hat{K}^{**})^{-1}(\mathbf{f}-\hat{\boldsymbol{\mu}}^*)$, which can be found by solving numerically an optimization problem. An example of such optimization is given in the next section in the context of hypothesis testing.

4. Application: hypothesis test for the equality of two functions

An equality test is used to detect differences between two regression functions $f_1(x)$ and $f_2(x)$ given the two independent samples $D_1 = (\mathbf{x}^{(1)}, \mathbf{y}^{(1)})$ and $D_2 = (\mathbf{x}^{(2)}, \mathbf{y}^{(2)})$ of, respectively, n_1 and n_2 observations. Our aim is to extend the Bayesian test based on the GP presented in [1] using the c-IGP model. The approach in [1] assumes the same GP prior $GP(0, k_{\boldsymbol{\theta}})$ for the

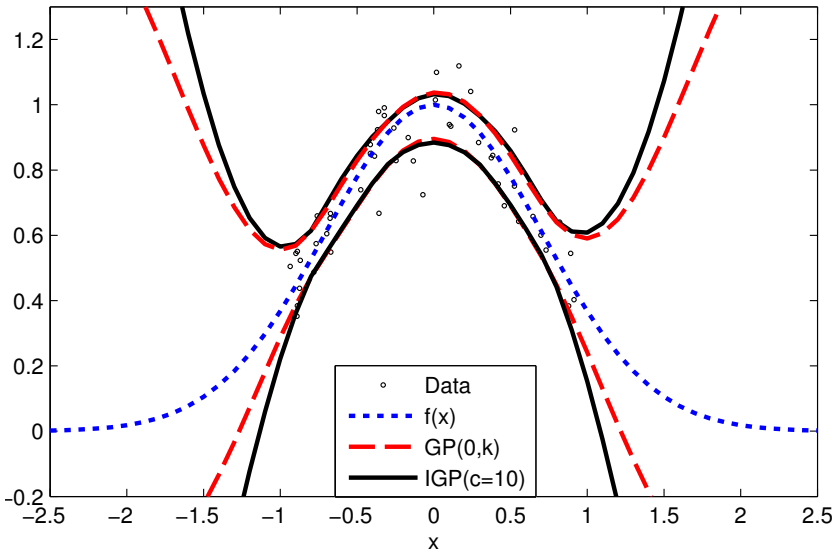

Figure 2: GP and c-IGP estimates of pointwise credible intervals for the value of $f(x)$ in Example 1.

two functions f_1 and f_2; the two posterior distributions share the same hyperparameters. Here, we assume the same c-IGP set of priors $\mathcal{G}$ for the two functions, that is,

$$f_i \sim GP\left(M_i h_i, k_{\boldsymbol{\theta}}(x, x') + \frac{M_i + 1}{c}\right),$$

with $i = 1, 2$, $h_i = \pm 1$ and $M_i \geq 0$. As a consequence, we are assuming that f_1 and f_2 are two GPs with the same base kernel $k_{\boldsymbol{\theta}}(x, x')$. Instead, their prior mean functions $M_1 h_1$ and $M_2 h_2$ can be different, as they are free to vary in the set of all constant functions. We assume that the two samples $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are affected by Gaussian noise with variance, respectively, σ_1^2 and σ_2^2. The hyperparameters $\boldsymbol{\theta}, \sigma_1, \sigma_2$ are obtained considering for both f_1 and f_2 the prior with $c \to 0$. Then, after combining the two datasets $\{D_1, D_2\}$, we maximize the joint marginal probability of $(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \boldsymbol{\theta}, \sigma_1, \sigma_2)$ with respect to $\boldsymbol{\theta}, \sigma_1, \sigma_2$. Assuming that f_1 and f_2 are independent Gaussian processes, we have that

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \boldsymbol{\theta}, \sigma_1, \sigma_2) = \\ p(\mathbf{y}^{(1)} | \mathbf{x}^{(1)}, \boldsymbol{\theta}, \sigma_1) p(\mathbf{y}^{(2)} | \mathbf{x}^{(2)}, \boldsymbol{\theta}, \sigma_2).$$

Then, the logarithm of the joint marginal of $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \boldsymbol{\theta}, \sigma_1^2, \sigma_2^2$ is

$$\sum_{i=1}^{2} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}, \sigma_i) + \log p(\boldsymbol{\theta}, \sigma_1, \sigma_2)$$

where $\log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}, \sigma_i)$, is given in (11), up to an additive constant.

In the precise approach, given the prior $GP(0, k_\theta)$, we compute from (4) the posterior marginal GPs $p(\mathbf{f}_1^*|\mathbf{x}^*, D_1)$ and $p(\mathbf{f}_2^*|\mathbf{x}^*, D_2)$ at the $m = n_1 + n_2$ test inputs $\mathbf{x}^* = \{x_i^* : i = 1, \ldots, m,\ x_i^* \in [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}]\}$. In this way, the equality of the two functions is tested at the covariates of the observations, that is, where we have the experimental evidence. Moreover, it is assumed that the observation covariates $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ cover the same region of the covariate space. This is done to avoid testing the equality in regions where there are no observations for one or both functions, as in these region we do not expect to be able to state any conclusion about equality or difference of the two functions. If applied in such regions, the precise test would always assign very large posterior probability to the hypothesis that there is no difference between the functions. Using a IGP model, we can test the equality assumption in any subset $\mathcal{X}_T$ of the covariate space $\mathcal{X}$ by taking the m test inputs $\mathbf{x}^*$ so to cover uniformly the region of interest $\mathcal{X}_T$. If all priors in the IGP set entail the same decision, we retain it, if instead they lead to different decisions we conclude that a robust decision cannot be made in $\mathcal{X}_T$. This way, we can automatically identify a situation where data do not allow to state any conclusion.

Let us denote the mean and covariance functions of the posterior distributions of $\mathbf{f}_1^*$ and $\mathbf{f}_2^*$ as $\hat{\mu}^{(i)}(x)$ and $\hat{k}^{(i)}(x, x')$, $i = 1, 2$. Since the difference of two Gaussian variables is Gaussian, it follows that the posterior of the GP $\Delta f(x) = f^1(x) - f^2(x)$ is also a GP with mean and covariance functions $\Delta\hat{\mu}(x) = \hat{\mu}^{(1)}(x) - \hat{\mu}^{(2)}(x)$ and $\hat{k}_\Delta(x, x') = \hat{k}^{(1)}(x, x') + \hat{k}^{(2)}(x, x')$. Let $\Delta\mathbf{f}^*$, $\Delta\hat{\boldsymbol{\mu}}^*$ and $\hat{K}_\Delta^*$ be the difference, the mean and the covariance functions evaluated at the test covariates $\mathbf{x}^*$, then, we say that the two functions are equal with posterior probability $1 - \alpha$ if the credible region for $\Delta\mathbf{f}^*$ includes the zero vector or, in other words, if:

$$(\Delta\hat{\boldsymbol{\mu}}^*)^T(\hat{K}_\Delta^*)^{-1}\Delta\hat{\boldsymbol{\mu}}^* \leq \chi^{-1}(1 - \alpha|\nu), \tag{15}$$

where ν is the number of positive eigenvalues of $\hat{K}_\Delta^*$. In practice, as the number m of test inputs is likely to be considerably larger than the dimensionality of the covariance function, the matrix $\hat{K}_\Delta^*$ is not full rank. Thus,

we decompose it as PDP^T, where D is the diagonal matrix of the eigenvalues $\lambda_1, \ldots, \lambda_m$ (sorted in descending order), and retain only the sub-matrices $P_\nu D_\nu P_\nu^T$ corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_\nu$ which verify the condition $\lambda_{\nu+1} / \sum_{i=1}^m \lambda_i < \epsilon$, where ϵ is a small, positive constant. In the example below, we use $\epsilon = 0.0001$.

In the c-IGP model, the inference about $\chi_s^2(M_1, M_2, h_1, h_2) = (\Delta\hat{\boldsymbol{\mu}}^*)^T (\hat{K}_\Delta^*)^{-1} \Delta\hat{\boldsymbol{\mu}}^*$ depends on the choice of the prior, that is on the value of $M_1,\ M_2$, and of $h_1,\ h_2$.

Proposition 2. *The c-IGP model is a prior ignorance model for inferences about χ_s^2, i.e.,*

$$\underline{\chi}_s^2 = 0 \quad \overline{\chi}_s^2 \to = +\infty.$$

A posteriori let $\hat{\mu}_0^{(i)}(x) = k_x^{(i)T} K_n^{(i)-1} \mathbf{y}^{(i)}$ be the posterior mean functions obtained from a GP with zero mean and covariance function $k_\theta(x, x')$ when $\mathbf{x} = \mathbf{x}^{(i)}$, $\mathbf{y} = \mathbf{y}^{(i)}$, and let $\hat{k}_0^{(i)}(x, x')$ be the covariance function obtained from (10) when $c \to \infty$. For a given value of M_1 and M_2 the posterior expectation of the GP $\Delta f(x)$, that is $\Delta\hat{\mu}(x)$, can be derived from (9) and is

$$\Delta\hat{\mu}_{M_1,M_2}(x) = \hat{\mu}_0^{(1)}(x) - \hat{\mu}_0^{(2)}(x) + \mu_c^{(1)}(x) + \mu_c^{(2)}(x)$$

where $\mu_c^{(i)}(x) = (1 - \mathbf{k}_x^{(i)T} K_n^{(i)-1}) \frac{\mathbf{s}_k^{(i)T} \mathbf{y}^{(i)} + t_i c}{c(1-|t_i|) + S_k^{(i)}}$, $t_i = \frac{Mh}{M+1}$ and $\mathbf{k}_x^{(i)}$ and $K_n^{(i)}$ are obtained by evaluating the covariance functions at the training covariates $\mathbf{x}^{(i)}$, $i = 1, 2$. The lower/upper bounds for $\chi_s^2(M_1, M_2, h_1, h_2)$ are obtained by minimizing/maximizing w.r.t. $t_i \in [-1, 1]$ the statistic:

$$\chi_s^2 = \Delta\hat{\boldsymbol{\mu}}^{*T} (\hat{K}_{M_1,M_2}^\Delta)^{-1} \Delta\hat{\boldsymbol{\mu}}^*, \tag{16}$$

where $\hat{K}_{M_1,M_2}^\Delta = [\hat{k}_0^{(1)}(x_i, x_j) + \hat{k}_0^{(2)}(x_i, x_j) + \hat{k}_c^{(1)}(x_i, x_j) + \hat{k}_c^{(2)}(x_i, x_j)]_{i,j}$, and $\hat{k}_c^{(i)}(x, x') = \frac{(M_i+1)[1-\mathbf{k}_x^{(i)T}\mathbf{s}_k^{(i)}][1-\mathbf{k}_{x'}^{(i)T}\mathbf{s}_k^{(i)}]^T}{c+(M_i+1)S_k^{(i)}}$, $i = 1, 2$.

4.1. Numerical example

Let us consider two samples D_1 and D_2 that we wish to compare on the subset $\mathcal{X}_T = [a, b]$ of the covariate space. Assuming an observation noise $v \sim \mathcal{N}(0, \sigma_n = 0.2)$, we sample D_1 and D_2 from:

Case A: $x_i^{(1,2)} \sim U[-2, 2]$, $y_i^{(1,2)} = f(x_i) + v_i$,

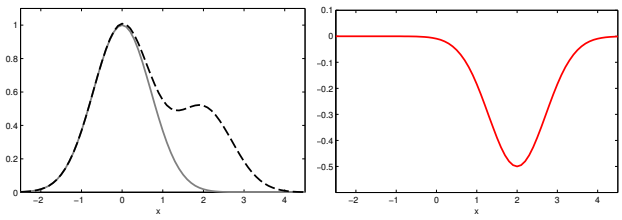

Figure 3: Left: Functions f (continuous line) and g (dashed line). Right: difference $f - g$.

$$\begin{aligned}
\text{Case B: } & x_i^{(1)} \sim U[-2,2], \quad y_i^{(1)} = f(x_i) + v_i, \\
& x_i^{(2)} \sim U[-2,2] \quad y_i^{(2)} = g(x_i) + v_i, \\
\text{Case C: } & x_i^{(1)} \sim U[-2,0], \quad y_i^{(1)} = f(x_i) + v_i, \\
& x_i^{(2)} \sim U[-2,4], \quad y_i^{(2)} = g(x_i) + v_i, \\
\text{Case D: } & x_i^{(1)} \sim U[-2,2], \quad y_i^{(1)} = f(x_i) + v_i, \\
& x_i^{(2)} \sim U[-2,4], \quad y_i^{(2)} = g(x_i) + v_i,
\end{aligned}$$

where $f(x) = \exp(-x^2)$ and $g(x) = f(x) + 0.5f(x-2)$ (see Figure 3). For each scenarion the two datasets D_1 and D_2 have been simulated only once. More extensive simulations are left to future work. We have tested the difference between the two samples for different test subsets $\mathcal{X}_T \in [-2, b]$. The difference $f(x) - g(x)$ is about zero for $x < 0$, is large ($> \sigma_n$) in the interval $[1, 3]$ and is small ($< \sigma_n$) in $[0, 1]$. Therefore, we expect to easily detect a difference between the two samples when $b > 1$, whereas for $b < 1$ the decision is more difficult and for $b < 0$ we can assume that the two functions are equal. Table 1 shows the decisions for the precise and the imprecise tests at different values of c and b. One can notice that for $c = 10$ we are most often undecided (save when the decision is simple, e.g., in cases B and D when $b > 1$ and thus all tests recognize the difference) as the imprecision is very large in this case.

On the other side, for $c = 1$ the test makes almost always the same decision as the precise test, as the imprecision is very small in this case. When $c = 5$ we have a better balance between robustness and power: the IGP test makes the same decision as the precise one when there is enough information to make a robust decision, whereas it is undecided when the decision is difficult due to the lack of information. For instance, in case A with $b = 2$ the precise test always issues a no difference decision. The same happens in case C, although the two situations are very different, because in

		GP		IGP	
Case	b	n=50	n=200	n=50	n=200
A	2	0	0	0/0/2	0/0/2
A	4	0	0	0/2/2	0/2/2
B	0	0	0	0/0/2	0/0/2
B	1	0	1	0/2/2	1/1/1
B	2	1	1	1/1/1	1/1/1
B	4	1	1	1/1/1	1/1/1
C	0	0	0	0/0/2	0/0/2
C	1	0	0	0/0/2	0/0/2
C	2	0	0	0/2/2	0/2/2
C	4	0	0	0/2/2	0/2/2
D	0	0	0	0/0/0	0/0/2
D	1	0	1	2/2/2	1/1/1
D	2	1	1	1/1/1	1/1/1
D	4	1	1	1/1/1	1/1/1

Table 1: Decisions of the precise test for $c = 1/5/10$, where 0 indicates that the two functions are equal with posterior probability $1 - \alpha$, 1 indicates that the two functions are different (i.e., the posterior probability that they are equal is less than α), 2 indicates indecision (i.e., the decision depends on the prior).

the first case $f_1 = f_2$ and we can observe both functions on the entire set $\mathcal{X}_T$, whereas in the second case $f_1 \neq f_2$ but we cannot see it as we observe f_1 only in the range $[-2, 0]$ where the two function are almost identical. On the other side, the imprecise test detects the difference of the two situations, and in case A it correctly issues a no difference decision, whereas in case C it is undecided, thus acknowledging that there is not enough information to make a decision. Something similar can be observed also in case D: when $b = 0$ both the precise and imprecise tests issue a no difference decision as in this range the two function can be actually considered identical; when, instead, $b = 1$, the functions are different, but, since the difference is small, it cannot be clearly detected with only $n = 50$ data. However, the imprecise test recognizes that the decision is somehow difficult and is undecided, whereas the precise test can only decide that there is no difference. For $n = 200$, the information is enough to make both tests detect a difference.

5. A generalization of the IGP model

In this Section we generalize the IGP with constant mean by considering an IGP with mean function proportional to an arbitrary function $h(x)$.

Definition 3. *Given a function $h(x)$ and a constant $c > 0$, we define an Imprecise Gaussian Process with base mean function $h(x)$ (hIGP) the set of GPs:*

$$\mathcal{G}_{h(x)} = \left\{ GP\left(Mh(x), k_{\boldsymbol{\theta}}(x, x') + \frac{M+1}{c} h(x)h(x') \right),\ M \geq 0 \right\} \quad .$$

A posteriori we have the following result.

Theorem 5. *Let $f(x) \sim GP(Mh(x), k_{\boldsymbol{\theta}} + \frac{M+1}{c} h(x), h(x'))$ and $\mathbf{h} = h(\mathbf{x})$. The posterior distribution of f is a GP with mean function*

$$\hat{\mu}(x) = \mathbf{k}_x^T K_n^{-1} (\mathbf{y} - \hat{y}(\mathbf{x})) + \hat{y}(x) \tag{17}$$

with $\hat{y}(x) = \frac{(M+1)\mathbf{h}^T K_n^{-1}\mathbf{y} + cM}{c + (M+1)\mathbf{h}^T K_n^{-1}\mathbf{h}} h(x)$, and covariance function

$$\begin{aligned} \hat{k}(x, x') =& k_{\boldsymbol{\theta}}(x, x') - \mathbf{k}_x^T K_n^{-1} \mathbf{k}_{x'} + \\ & \frac{(M+1)(h(x) - \mathbf{k}_x^T K_n^{-1}\mathbf{h})^T (h(x') - \mathbf{k}_{x'}^T K_n^{-1}\mathbf{h})}{c + (M+1)\mathbf{h}^T K_n^{-1}\mathbf{h}}. \end{aligned} \tag{18}$$

We can further generalize the IGP model in Definition 3 by letting $h(x)$ free to vary in a set of functions $\mathcal{H}$.

Definition 4. *Given a set of functions $\mathcal{H}$ and a constant $c > 0$, we define an Imprecise Gaussian Process with set of base mean functions $\mathcal{H}$ ($\mathcal{H}$-IGP) the set of GPs:*

$$\mathcal{G}_{\mathcal{H}} = \left\{\mathcal{G}_{h(x)} : h(x) \in \mathcal{H}\right\}. \tag{19}$$

From Theorem 5 we can see that not all $\mathcal{H}$-IGP verify learning. In fact, if $\mathcal{H}$ include functions that are zero at all training covariates $\mathbf{x}$ so that $\mathbf{h}^T K_n^{-1} \mathbf{h} = 0$, posterior inferences are vacuous.

Proposition 3. *Any set $\mathcal{H}$-IGP such that $\mathbf{h}$ is a nonzero vector for all $h(x) \in \mathcal{H}$ can learn from the observations $\mathbf{x}, \mathbf{y}$.*

Moreover, not all $\mathcal{H}$-IGP verify prior-ignorance about $E[f(x^*)]$ for all x^*. If, for instance, $h(x^*) = 0$ all $h(x) \in \mathcal{H}$, then a priori $E[f(x_i^*)] = Mh(x^*) = 0$ for all M.

Proposition 4. *If it exist $h^+(x^*) \in \mathcal{H} : h^+(x^*) > 0$ and $h^-(x^*) \in \mathcal{H} : h^-(x^*) < 0$, then the $\mathcal{H}$-IGP is a model of prior ignorance about the expectation of $f(x^*)$ in the sense thatit verifies*

$$\inf_{M,h(x)} E[f(x_i^*)] = -\infty, \quad \sup_{M,h(x)} E[f(x_i^*)] = +\infty.$$

By properly selecting the set $\mathcal{H}$ one can obtain IGP models that verify both prior near-ignorance and learning.

Example 2. *The c-IGP model presented in Section 3 is an $\mathcal{H}$-IGP model with set of base mean functions $\mathcal{H}_c = \{h(x) = -1,\ h(x) = 1\}$. This set verifies the conditions of both Proposition 4 and 3 an thus verifies both prior near-ignorance and learning.*

Example 3. *Let us consider the set $\mathcal{H} = \{h(x) = -x,\ h(x) = x\}$. It verifies the conditions of Proposition 4 for all covariates except $x = 0$ and verifies the condition of Proposition 3 provided that $\mathbf{x}$ is a nonzero vector. The corresponding $\mathcal{H}$-IGP includes GPs with mean function varying in the set of all linear functions with intercept in 0. It is a model of prior ignorance about $f(x)$ for all $x \neq 0$ and can learn from data with covariate $x \neq 0$.*

6. Conclusions

In this paper we have presented a model of prior near ignorance about the value of a regression function based on the Gaussian process. We have shown that this IGP model can be used to make inferences about the regression functions which are more robust with respect to the choice of the prior. In fact, for those subsets of $\mathcal{X}$ where there are many observations inferences almost coincide with the precise model, whereas in those subset with no observations the imprecision of the prediction is very high, thus reflecting the actual lack of knowledge. As a consequence of this, decisions based on this model are more reliable. For instance, we have applied the IGP to test the difference between regression functions, and shown that the IGP model allows us to acknowledge when the available data are not informative enough to make a robust decision. Although in this paper we have only consider univariate functions, the c-IGP model can be straightforward extended to the multivariate case where x is a vector of covariates.

A generalization of the IGP with constant mean has also been proposed, based on which it will be possible to develop other prior near ignorance models that consider different sets of prior mean functions. The study of these models and their properties will be the object of future work. Moreover, as a strong prior information is introduced in the model also by the base kernel, further research should focus on the development of models allowing for a weaker specification of the kernel function.

There are many techniques other than GPs available for nonparametric regression, e.g., splines, relevance vector machines, kernel smoothers, etc., that have not been considered in this work. Their relative strengths and weaknesses w.r.t. GPs are discussed in [12, Sec. 7]. As they are all precise methods, we can expect them to suffer from the same weaknesses of the precise GPs. The probabilistic formulation of GPs and the simple closed form expression of their posterior inferences, have made them a good starting point to develop an imprecise approach to nonparametric regression that, in the future, could be extended to other regression techniques, taking advantage also from the connections they have with GPs [12, Sec. 6].

Appendix

6.1. Proof of Proposition 1

This can be seen by considering that a priori $E[f(x_i^*)] = Mh(x_i^*)$ so that for $h(x_i^*) = \pm 1$ and $M \to \infty$ we have $E[f(x_i^*)] \to \pm\infty$.

6.2. Proof of Theorem 1

Miller in [7] proves the following

Lemma 1. *If A and $A+B$ are invertible, and B has rank 1, then*

$$(A+B)^{-1} = A^{-1} - \frac{1}{1+g}A^{-1}BA^{-1},$$

where $g = trace(BA^{-1})$ with $g \neq -1$.

From this, it follows that

$$(K_n + \frac{M+1}{c}\mathbb{1}_{nn})^{-1} = K_n^{-1} - \frac{M+1}{c+(M+1)S_k}\mathbf{s}_k\mathbf{s}_k^T, \tag{20}$$

where $\mathbb{1}_{nn}$ is a $n \times n$ dimensional matrix of ones. Then,

$$\begin{aligned}
\hat{\mu}(x) =& Mh + \left(\mathbf{k}_x + \frac{M+1}{c}\mathbb{1}_n\right)^T \\
& \left(K_n^{-1} - \frac{M+1}{c+(M+1)S_k}\mathbf{s}_k\mathbf{s}_k^T\right)(\mathbf{y} - Mh\mathbb{1}_n) \\
=& Mh + \left[\mathbf{k}_x^T K_n^{-1}\left(1 - \frac{(M+1)\mathbb{1}_n\mathbf{s}_k^T}{c+(M+1)S_k}\right) + \right. \\
& \left. \frac{(M+1)\mathbf{s}_k^T}{c+(M+1)S_k}\right](\mathbf{y} - Mh\mathbb{1}_n) \\
=& \mathbf{k}_x^T K_n^{-1}\left(\mathbf{y} - Mh(\mathbf{x}) - \frac{(M+1)\mathbb{1}_n\mathbf{s}_k^T}{c+(M+1)S_k}(\mathbf{y} - Mh(\mathbf{x}))\right) \\
& + Mh + \frac{(M+1)\mathbf{s}_k^T}{c+(M+1)S_k}(\mathbf{y} - Mh(\mathbf{x})) \\
=& \mathbf{k}_x^T K_n^{-1}\left(\mathbf{y} - \frac{(M+1)\mathbf{s}_k^T\mathbf{y} + cMh}{c+(M+1)S_k}\mathbb{1}_n\right) + \\
& \frac{(M+1)\mathbf{s}_k^T\mathbf{y} + cMh}{c+(M+1)S_k}.
\end{aligned}$$

Similarly for the covariance function we obtain:

$$\begin{aligned}
\hat{k}(x,x') =& k_\theta(x,x') + \frac{M+1}{c} - \left(\mathbf{k}_x + \frac{M+1}{c}\mathbb{1}_n\right)^T \\
& \left(K_n^{-1} - \frac{M+1}{c+(M+1)S_k}\mathbf{s}_k\mathbf{s}_k^T\right)\left(\mathbf{k}_{x'} + \frac{M+1}{c}\mathbb{1}_n\right) \\
=& k_\theta(x,x') - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_x + \\
& \frac{M+1}{c+(M+1)S_k}\left(\mathbf{k}_x^T\mathbf{s}_k\mathbf{s}_k^T\mathbf{k}_{x'} - \mathbf{k}_x^T\mathbf{s}_k - \mathbf{s}_k^T\mathbf{k}_{x'} + 1\right).
\end{aligned}$$

6.3. Proof of Theorem 2

From (8), the logarithm of the marginal probability of $\mathbf{y}, \theta_{\mathbf{n}}$ given the prior $GP\left(Mh, k_{\boldsymbol{\theta}}(x,x') + \frac{M+1}{c}\right)$ is

$$\begin{aligned}&\log p(\mathbf{y}, \theta_{\mathbf{n}}) = \\ &-\tfrac{1}{2}(\mathbf{y} - Mh\mathbb{1}_n)^T \left(K_n + \frac{M+1}{c}\mathbb{1}_{nn}\right)^{-1} (\mathbf{y} - Mh\mathbb{1}_n) + \\ &-\tfrac{1}{2}\log|K_n + \tfrac{M+1}{c}\mathbb{1}_{nn}| - \tfrac{n}{2}\log 2\pi + \log p(\theta_{\mathbf{n}}).\end{aligned} \tag{21}$$

From (25) we obtained that the first term on the r.h.s. of (21) is equal to

$$\frac{cM^2h^2S_k - 2cMh\mathbf{y}^T\mathbf{s}_k - (M+1)(\mathbf{y}^T\mathbf{s}_k)^2}{2c + 2(M+1)S_k} - \tfrac{1}{2}\mathbf{y}^T K_n^{-1}\mathbf{y}. \tag{22}$$

Based on the matrix determinant Lemma [5] which states:

Lemma 2. *Given a $n \times n$ invertible matrix A and two n dimensional vectors $\mathbf{u}, \mathbf{v}$*

$$|A + \mathbf{u}\mathbf{v}^T| = |A|(1 + \mathbf{v}^T A^{-1}\mathbf{u}),$$

we obtain

$$|K_n + \frac{M+1}{c}\mathbb{1}_{nn}| = |K_n|\frac{c + (M+1)S_k}{c}. \tag{23}$$

Then, from (22), (21) and (23), it follows

$$\begin{aligned}\log p(\mathbf{y}, \theta_{\mathbf{n}}) \quad &\xrightarrow{c\to 0} -\tfrac{1}{2}\left[\mathbf{y}^T K_n^{-1}\mathbf{y} - \tfrac{(\mathbf{y}^T\mathbf{s}_k)^2}{S_k}\right] + \\ &-\tfrac{1}{2}\log|K_n|\tfrac{M}{c}S_k - \tfrac{n}{2}\log 2\pi + \log p(\theta_{\mathbf{n}}).\end{aligned} \tag{24}$$

Finally, by considering that the terms $-\frac{1}{2}\log\frac{M}{c}$ and $-\frac{n}{2}\log 2\pi$ are constant with θ, we have that

$$\begin{aligned}\mathrm{argmax}_{\boldsymbol{\theta}}[\log p(\mathbf{y}, \boldsymbol{\theta}_n)] = \mathrm{argmax}_{\boldsymbol{\theta}} \Bigg[-\frac{1}{2}\Bigg(&\mathbf{y}^T K_n^{-1}\mathbf{y} - \frac{(\mathbf{y}^T\mathbf{s}_k)^2}{S_k} + \\ &- \log S_k|K_n|\Bigg) + \log p(\theta_{\mathbf{n}})\Bigg].\end{aligned}$$

6.4. Proof of Theorem 3

The derivative of (9) with respect to M is

$$\frac{\partial\hat{\mu}(x)}{\partial M} = (1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\pm c \pm S_k + \mathbf{s}_k^T\mathbf{y}}{(c + (M+1)S_k)^2}$$

If $\left|\frac{\mathbf{s}_k^T\mathbf{y}}{S_k}\right| \leq \frac{c}{S_k}+1$, the second term of the derivative above is positive for $h=1$ and negative for $h=-1$. Notice also that, for $M=0$, the values of $\hat{\mu}(x)$ in the two cases $h=\pm 1$ coincide. Then, if the first term $1-\mathbf{k}_x^T\mathbf{s}_k$ is positive, we have that $\hat{\mu}(x)$ increases with M for $h=1$ and decreases for $h=-1$ so that the upper is found for $h=1$ and $M\to\infty$ and the lower for $h=-1$ and $M\to\infty$. Vice versa, if the first term is negative, the upper is found for $h=-1$ and $M\to\infty$ and the lower for $h=1$ and $M\to\infty$.

If, instead, $\left|\frac{\mathbf{s}_k^T\mathbf{y}}{S_k}\right| > \frac{c}{S_k}+1$, the second term of the derivative above is always positive (if $\frac{\mathbf{s}_k^T\mathbf{y}}{S_k}>0$) or negative (otherwise); then, $\hat{\mu}(x)$ increases if $1-\mathbf{k}_x^T\mathbf{s}_k$ and $\frac{\mathbf{s}_k^T\mathbf{y}}{S_k}$ have the same sign, and decreases otherwise. In the first case, the upper is found for $h=1$ and $M\to\infty$ and the lower for $M=0$; viceversa, in the second case, the upper is found for $M=0$ and the lower for $h=-1$ and $M\to\infty$.

The value of the upper and lower can be derived from (9).

6.5. Proof of Theorem 4

For each GP in $\mathcal{G}$ the lower bound of a $(1-\alpha)\%$ credible interval is

$$\hat{\mu}(x) - z_{\alpha/2}\sqrt{\hat{k}(x,x)} \leq \underline{\mu}(x) - z_{\alpha/2}\sqrt{\hat{k}(x,x)}.$$

Moreover, from (10), it can be seen that $\hat{k}(x,x)$ increases with M, so that it maximum is found at $M\to\infty$ and is $\sigma^2_{f_x} = k_{\boldsymbol{\theta}}(x,x') - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_{x'} + \frac{(1-\mathbf{k}_x^T\mathbf{s}_k)^2}{S_k}$ so that

$$\hat{\mu}(x) - z_{\alpha/2}\sqrt{\hat{k}(x,x)} \leq \underline{\mu}(x) - z_{\alpha/2}\sigma_{f_x}.$$

where the equality holds when the lower expectation $\underline{\mu}(x)$ is found for $M\to\infty$, that is when $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| \leq 1+\frac{c}{S_k}$ (Theorem 3).

6.6. Proof of Proposition 2

It can be verified that the lower bound is found by choosing $M_1=M_2=0$. The upper is found, for instance, for $M_1=0$, $M_2\to\infty$, as we have

$$\begin{aligned}
\chi^2_s(0,M_2,h_1,h_2) &= M_2^2\mathbb{1}_m^T(2*K^{**}+\frac{M_2+2}{c})^{-1}\mathbb{1}_m \\
&= \frac{M_2^2}{2}\mathbb{1}_m^T\left[(K^{**})^{-1} - \frac{M_2+2}{2c+(M_2+2)S_{k^*}}\mathbf{s}_{k^*}\mathbf{s}_{k^*}^T\right]\mathbb{1}_m \\
&= M_2^2\frac{1}{2}\left[S_{k^*} - \frac{M_2+2}{2c+(M_2+2)S_{k^*}}S_{k^*}^2\right] \\
&= M_2^2\frac{cS_{k^*}}{2c+(M_2+2)S_{k^*}} \xrightarrow{M_2\to\infty} cM_2\to\infty,
\end{aligned}$$

where $S_{k^*} = \mathbb{1}_m^T(K^{**})^{-1}\mathbb{1}_m$, $\mathbf{s}_{k^*} = (K^{**})^{-1}\mathbb{1}_m$ and where we have used Lemma 1.

6.7. Proof of Theorem 5

Let us define $M' = \frac{M+1}{c}$, $S_h = \mathbf{h}^T K_n^{-1}\mathbf{h}$, $\mathbf{s}_h = K_n^{-1}\mathbf{h}$ and $D = 1 + M'S_h$. From Lemma 1, it follows that

$$(K_n + M'\mathbf{h}\mathbf{h}^T)^{-1} = K_n^{-1} - \frac{M'}{D}\mathbf{s}_h\mathbf{s}_h^T. \tag{25}$$

Then, $\hat{\mu}(x) =$

$$\begin{aligned}
=&Mh(x) + \left(\mathbf{k}_x + M'h(x)\mathbf{h}^T\right)^T \left(K_n^{-1} - \frac{M'}{D}\mathbf{s}_h\mathbf{s}_h^T\right)(\mathbf{y} - M\mathbf{h}) \\
=&Mh(x) + \mathbf{k}_x^T K_n^{-1}\mathbf{y} - \mathbf{k}_x^T\frac{M'}{D}\mathbf{s}_h\mathbf{s}_h^T\mathbf{y} + \frac{M'}{D}h(x)\mathbf{s}_h^T\mathbf{y} + \\
&- \frac{M}{D}\mathbf{k}_x^T\mathbf{s}_h - \frac{MM'}{D}h(x)S_h \\
=&\mathbf{k}_x^T K_n^{-1}\left(\mathbf{y} - \hat{y}(\mathbf{x})\right) + \hat{y}(x)
\end{aligned}$$

Similarly for the covariance function we obtain:

$$\begin{aligned}
\hat{k}(x,x') =&k_\theta(x,x') + M'h(x)h(x') - \left(\mathbf{k}_x + M'h(x)\mathbf{h}\right)^T \\
&\left(K_n^{-1} - \frac{M'}{D}\mathbf{s}_h\mathbf{s}_h^T\right)\left(\mathbf{k}_{x'} + M'h(x')\mathbf{h}\right) \\
=&k_\theta(x,x') - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_x + \frac{M'}{D}\left(\mathbf{k}_x^T\mathbf{s}_h\mathbf{s}_h^T\mathbf{k}_{x'} + \right. \\
&\left. -h(x)\mathbf{k}_x^T\mathbf{s}_h - h(x')\mathbf{s}_h^T\mathbf{k}_{x'} + h(x)h(x')\right).
\end{aligned}$$

6.8. Proof of Proposition 3

From (17) it follows that

$$\lim_{M\to+\infty}\hat{\mu}(x) = \mathbf{k}_x^T K_n^{-1}\mathbf{y} + (h(x) - \mathbf{k}_x^T K_n^{-1}\mathbf{h})\frac{\mathbf{h}^T K_n^{-1}\mathbf{y} + c}{\mathbf{h}^T K_n^{-1}\mathbf{h}},$$

which, if $\mathbf{h}$ is a nonzero vector, is bounded.

6.9. Proof of Proposition 4

This can be seen by considering that a priori $E[f(x_i^*)] = Mh(x_i^*)$ so that for $h(x) = h^+(x)$ and $M \to \infty$ we have $E[f(x_i^*)] \to +\infty$ and for $h(x) = h^-(x)$ and $M \to \infty$ we have $E[f(x_i^*)] \to -\infty$.

Acknowledgements

The author would like to thank Alessio Benavoli for his valuable comments and suggestions.

[1] A. Benavoli and F. Mangili. Gaussian Processes for Bayesian hypothesis tests. In *Proc 18th AIstat Conference*. Society for Artificial Intelligence and Statistics, 2015.

[2] A. Benavoli and M. Zaffalon. A model of prior ignorance for inferences in the one-parameter exponential family. *J of Stat Planning and Inference*, 142(7):1960 – 1979, 2012.

[3] A. Benavoli and M. Zaffalon. Prior near ignorance for inferences in the k-parameter exponential family. *Statistics*, 2014. in-press.

[4] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

[5] D. A. Harville. *Matrix algebra from a statistician's perspective*. Springer, 1997.

[6] D. J. MacKay. Introduction to Gaussian processes. *In Bishop, C. M., editor, Neural Networks and Machine Learning*, pages 133–166, 1998.

[7] K. S. Miller. On the inverse of the sum of matrices. *Mathematics Magazine*, 54(2):67–72, 1981.

[8] R. M. Neal. Regression and classification using gaussian process priors. In *Bernardo, et al. eds., Bayesian Statistics 6: Proc of the 6th Valencia international meeting*, volume 6, page 475, 1998.

[9] A. O'Hagan and J. F. C. Kingman. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):1–42, 1978.

[10] L. R. Pericchi and P. Walley. Robust Bayesian credible intervals and prior ignorance. *International Statistical Review*, 58:1–23, 1991.

[11] C. E. Rasmussen. The Gaussian Processes Web Site. http://www.gaussianprocess.org/, February 2011.

[12] C. E. Rasmussen and C. Williams. *Gaussian processes for machine learning*. The MIT Press, Cambridge, MA, USA, 2006.

[13] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

[14] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J of the Royal Statistical Society. Series B (Methodological)*, 58(1):3–57, 1996.