A prior near-ignorance Gaussian Process model for nonparametric regression

Francesca Mangili

IDSIA - Istituto Dalle Molle di studi sull'Intelligenza Artificiale, USI - SUPSI, Lugano - Switzerland

Introduction

Consider the regression model

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

$\mathbf{v} = [v_1, \dots, v_n] :=$ white Gaussian noise;
$\mathbf{x} = [x_1, \dots, x_n] :=$ vector of covariates;
$\mathbf{y} = [y_1, \dots, y_n] :=$ vector of observations;
$f(x) :=$ unknown regression function.

Goals:

- make inferences about $f(x)$;
- model prior near ignorance about $f(x)$.

Gaussian Process (GP)

$$f(x) \sim \mathcal{GP}(\mu(x), k(x, x'))$$

$\mu(x) :=$ **mean function**.
Prior belief about shape of $f(x)$.
Usually set equal to 0.

$k(x, x') :=$ **covariance function**.
Example: squared exponential
$k_{\boldsymbol{\theta}}(x, x') = \sigma_k^2 \exp\left[-\frac{1}{2}\frac{(x-x')^2}{\ell^2}\right]$,
$\boldsymbol{\theta} = (\sigma_k, \ell) :=$ hyperparameters.

- A priori $f(\mathbf{x}) \sim \mathcal{N}(\mu(\mathbf{x}), \overbrace{K}^{[k_{\boldsymbol{\theta}}(x_i,x_j)]_{ij}})$.
- A posteriori

$$f(x)|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \sim \mathcal{GP}(\hat{\mu}(x), \hat{k}(x, x')),$$

with:

$$\hat{\mu}(x) = \mu(x) + \overbrace{\mathbf{k}_x^T}^{k_{\boldsymbol{\theta}}(x,\mathbf{x})} (\overbrace{K}^{K+\sigma_n^2 \mathbf{I}})^{-1}(\mathbf{y} - \mu(\mathbf{x})),$$

$$\hat{k}(x, x') = k_{\boldsymbol{\theta}}(x, x') - \mathbf{k}_x^T K_n^{-1} \mathbf{k}_x.$$

Imprecise GP (IGP)

Definition h-IGP:

Given a base kernel $k_{\boldsymbol{\theta}}(x, x')$, a function $h(x)$ and a constant $c > 0$ we define an h-IGP as the set

$$\mathcal{G}_h = \{GP\left(Mh(x), k_{\boldsymbol{\theta}} + k_h\right), \ M \geq 0\}$$

with $k_h = \frac{M+1}{c} h(x)h(x')$.

Given a prior in h-IGP the posterior mean is

$$E[f(x)] = \mathbf{k}_x^T K_n^{-1}(\mathbf{y} - \hat{y}(\mathbf{x})) + \hat{y}(x)$$

with

$$\hat{y}(x) = \frac{(M+1)h(\mathbf{x})^T K_n^{-1}\mathbf{y} + cM}{c + (M+1)h(\mathbf{x})^T K_n^{-1} h(\mathbf{x})} h(x).$$

Definition $\mathcal{H}$-IGP:

Given a set of functions $\mathcal{H}$ and a constant $c > 0$, we define an $\mathcal{H}$-IGP as the set

$$\mathcal{G}_{\mathcal{H}} = \{\mathcal{G}_h : h(x) \in \mathcal{H}\}.$$

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

Near-ignorance: If there exist both strictly positive and negative values of $h(x^*) \in \mathcal{H}$, then the $\mathcal{H}$-IGP is a model of prior ignorance about $E[f(x^*)]$.

Constant mean IGP (c-IGP)

Definition c-IGP:

We define the c-IGP as the $\mathcal{H}$-IGP with

$$\mathcal{H} = \{h(x) = \pm 1\}.$$

- Prior ignorance about $E[f(x^*)]$:

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

- A posteriori, if $\left|\frac{\mathbf{s}_k \mathbf{y}}{S_k}\right| \leq 1 + \frac{c}{S_k}$,

$$\left.\begin{array}{l}\overline{E}[f(x)]\\ \underline{E}[f(x)]\end{array}\right\} = \mathbf{k}_x^T K_n^{-1}\mathbf{y} + (1 - \mathbf{k}_x^T \mathbf{s}_k)\frac{\mathbf{s}_k^T}{S_k}\mathbf{y} \pm c\,\frac{|1 - \mathbf{k}_x^T \mathbf{s}_k|}{S_k}.$$

with $\mathbf{s}_k = K_n^{-1}\mathbb{1}_n$, $S_k = \mathbb{1}_n^T K_n^{-1}\mathbb{1}_n$.

- Parameter c determines the degree of imprecision of the model:

$$\overline{E}[f(x)] - \underline{E}[f(x)] = 2c\,\frac{|1 - \mathbf{k}_x^T \mathbf{s}_k|}{S_k}$$

Example:

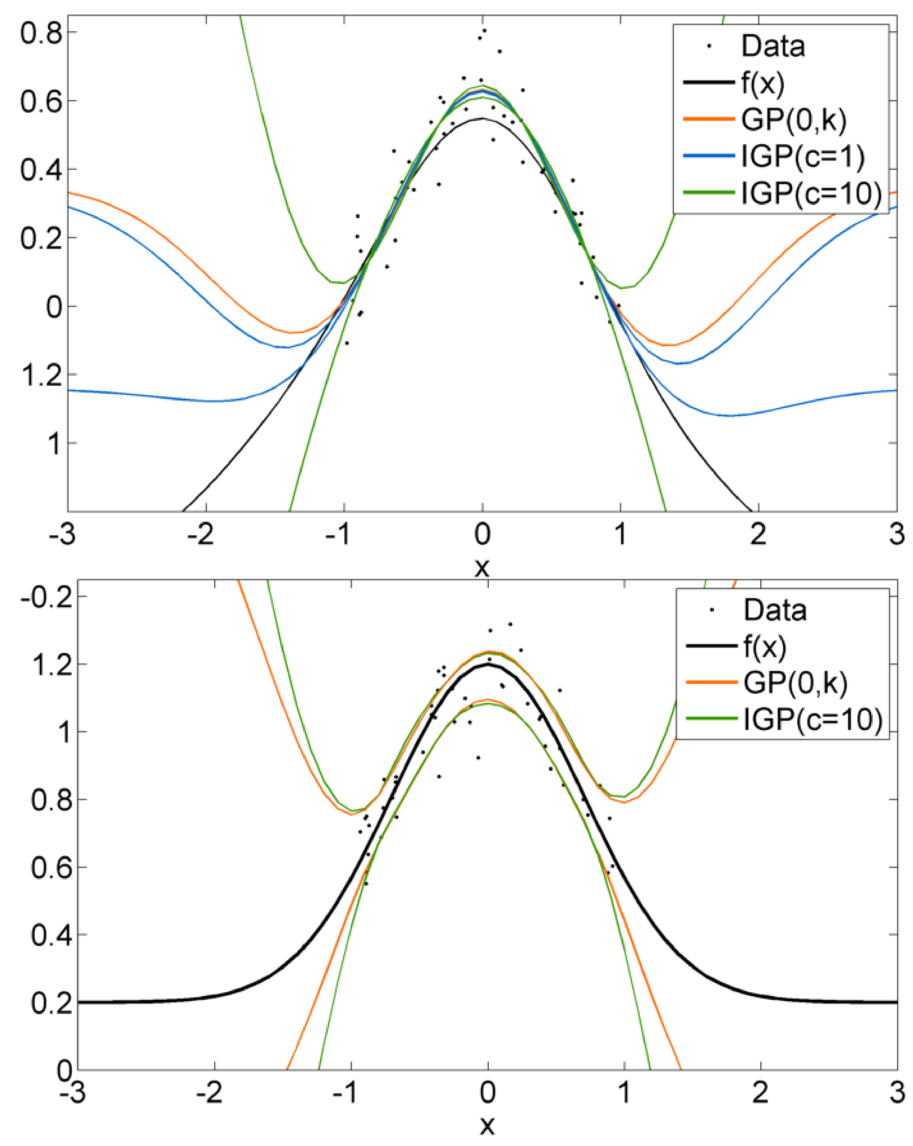

Figure 1: GP and c-IGP estimates of $E[f(x)]$ (upper) and its pointwise credible interval (bottom) given $n = 50$ observations.

Hypothesis testing

Goal: Compare $f_1(x)$ and $f_2(x)$ given two independent samples $(\mathbf{x}^{(1)}, \mathbf{y}^{(1)})$ and $(\mathbf{x}^{(2)}, \mathbf{y}^{(2)})$.

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

Hypothesis: $\Delta\mu(x) = E[f_1(x) - f_2(x)] \neq 0$ in a region of interest $\mathcal{X}_T$.

Procedure:

- Derive the credible region (CI) of $\Delta\mu(\mathbf{x}^*)$ ($\mathbf{x}^* :=$ vector of equispaced inputs $\in \mathcal{X}_T$) from the chi-squared random variable

$$\chi_s^2 = [\Delta\mu(\mathbf{x}^*)]^T (\hat{K}_\Delta^*)^{-1} [\Delta\mu(\mathbf{x}^*)]$$

Prior near-ignorance: $\underline{\chi}_s^2 = 0 \quad \overline{\chi}_s^2 \to +\infty$.

- If, a posteriori, $\mathbf{0} \notin$ CI then $f_1 \neq f_2$.
Indecision: If different priors entail different decisions, a robust decision cannot be made in $\mathcal{X}_T$.

Numerical example:

Case A: $x_i^{(1,2)} \sim U[-2, 2]$, $y_i^{(1,2)} = f(x_i) + v_i$
Case B: $x_i^{(1)} \sim U[-2, 2]$, $y_i^{(1)} = f(x_i) + v_i$,
$x_i^{(2)} \sim U[-2, 2]$, $y_i^{(2)} = g(x_i) + v_i$,
Case C: $x_i^{(1)} \sim U[-2, 0]$, $y_i^{(1)} = f(x_i) + v_i$,
$x_i^{(2)} \sim U[-2, 4]$, $y_i^{(2)} = g(x_i) + v_i$.

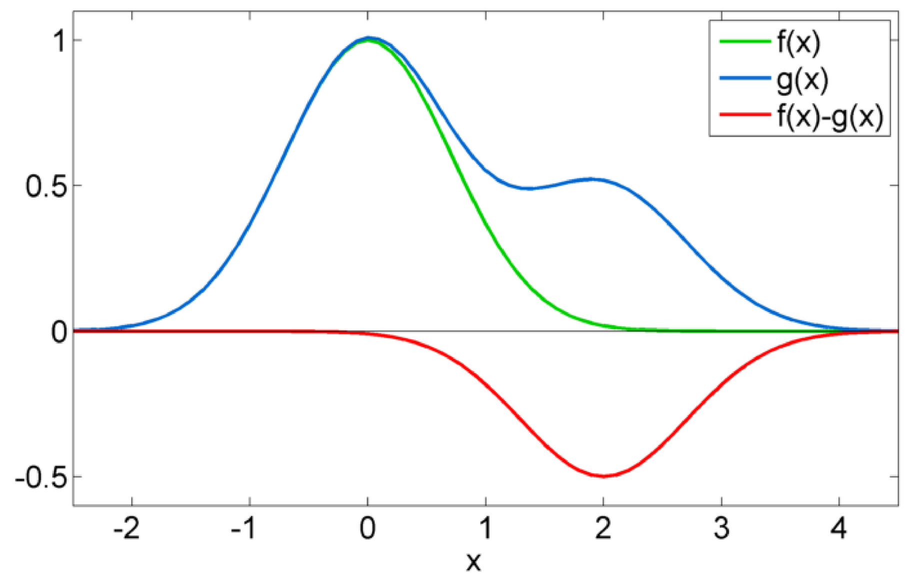

Figure 2: Functions f, g and $f - g$

Case	$\mathcal{X}_T$	GP n=50,	200	IGP n=50,	200
A	[-2 2]	0	0	0/**0**/2	0/**0**/2
A	[-2 4]	0	0	0/**2**/2	0/**2**/2
B	[-2 0]	0	0	0/**0**/2	0/**0**/2
B	[-2 2]	1	1	1/**1**/1	1/**1**/1
C	[-2 0]	0	0	0/**0**/2	0/**0**/2
C	[-2 2]	0	0	0/**2**/2	0/**2**/2

Table 1: Decisions for $c = 1/\mathbf{5}/10$. $0 \Rightarrow f_1 = f_2$, $1 \Rightarrow f_1 \neq f_2$, $2 \Rightarrow$ indecision.

Discussion: For $c = 5$ the IGP distinguishes whether there is no difference (rows 1,3,5) or the available data are not informative enough to make a decision (rows 2,6) .