A prior near-ignorance Gaussian Process model for nonparametric regression

Francesca Mangili

francesca@idsia.ch

Istituto "Dalle Molle" di Studi sull'Intelligenza Artificiale
Lugano (Switzerland)

http://www.ipg.idsia.ch/

ISIPTA 2015, Pescara

Introduction

Consider the regression model

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

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

Goals:

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

The 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
$k(x, x') = \sigma_k^2 \exp\left[-\frac{1}{2}\frac{(x-x')^2}{\ell^2}\right]$,
(σ_k, ℓ):= hyperparameters.

The Gaussian Process (GP)

Definition: A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

$$\begin{bmatrix} f(x_1) \\ f(x_2) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \end{bmatrix}, \quad \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) \\ k(x_2,x_1) & k(x_2,x_2) \end{bmatrix} \right)$$

Short notation:

$$f(\mathbf{x}) \quad \sim \mathcal{N}(\quad \mu(\mathbf{x}) \quad , \qquad k(\mathbf{x},\mathbf{x}) \qquad)$$

The Gaussian Process (GP)

Posterior

Observations: $(\mathbf{x}, \mathbf{y})$

Generic covariate: x

$$\begin{bmatrix} \mathbf{y} \\ f(x) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(\mathbf{x}) \\ \mu(x) \end{bmatrix} \begin{bmatrix} \overbrace{k(\mathbf{x},\mathbf{x}) + \sigma_\nu^2 I}^{K_n} & \overbrace{k(\mathbf{x},x)}^{\mathbf{k}_x} \\ k(x,\mathbf{x}) & k(x,x) \end{bmatrix} \right)$$

Then

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

with

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

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

The Imprecise Gaussian Process (IGP)

Definition: Given a base kernel $k(x, x')$, a function $h(x)$ and a constant $c > 0$ we define an Imprecise Gaussian Process with base mean function $h(x)$ (h-IGP) the set

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

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

If $h(x) \neq 0$

- a priori $\overline{E}[|f(x)|] = +\infty$
- the component k_h increases with the mean and thus

$$\frac{|\text{Prior mean of } f(x)|}{\text{Variance of } f(x)} = \frac{M|h(x)|}{k_{\boldsymbol{\theta}}(x, x) + \frac{M+1}{c} h(x)^2} \leq \frac{c}{h(x)} \quad \text{(bounded).}$$

The $\mathcal{H}$-IGP

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

Definition: We define an Imprecise Gaussian Process with set of base mean functions $\mathcal{H}$ ($\mathcal{H}$-IGP) as the set of GPs:

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

Near-ignorance: If there exist both strictly positive and strictly negative values of $h(x)$ for different $h \in \mathcal{H}$, then

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

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

The constant mean IGP (c-IGP)

Definition:
We define the costant mean IGP as the $\mathcal{H}$-IGP with

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

It verifies

- prior near-ignorance about $E[f(x)]$;
- learning.

The c-IGP

Posterior inferences:

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

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

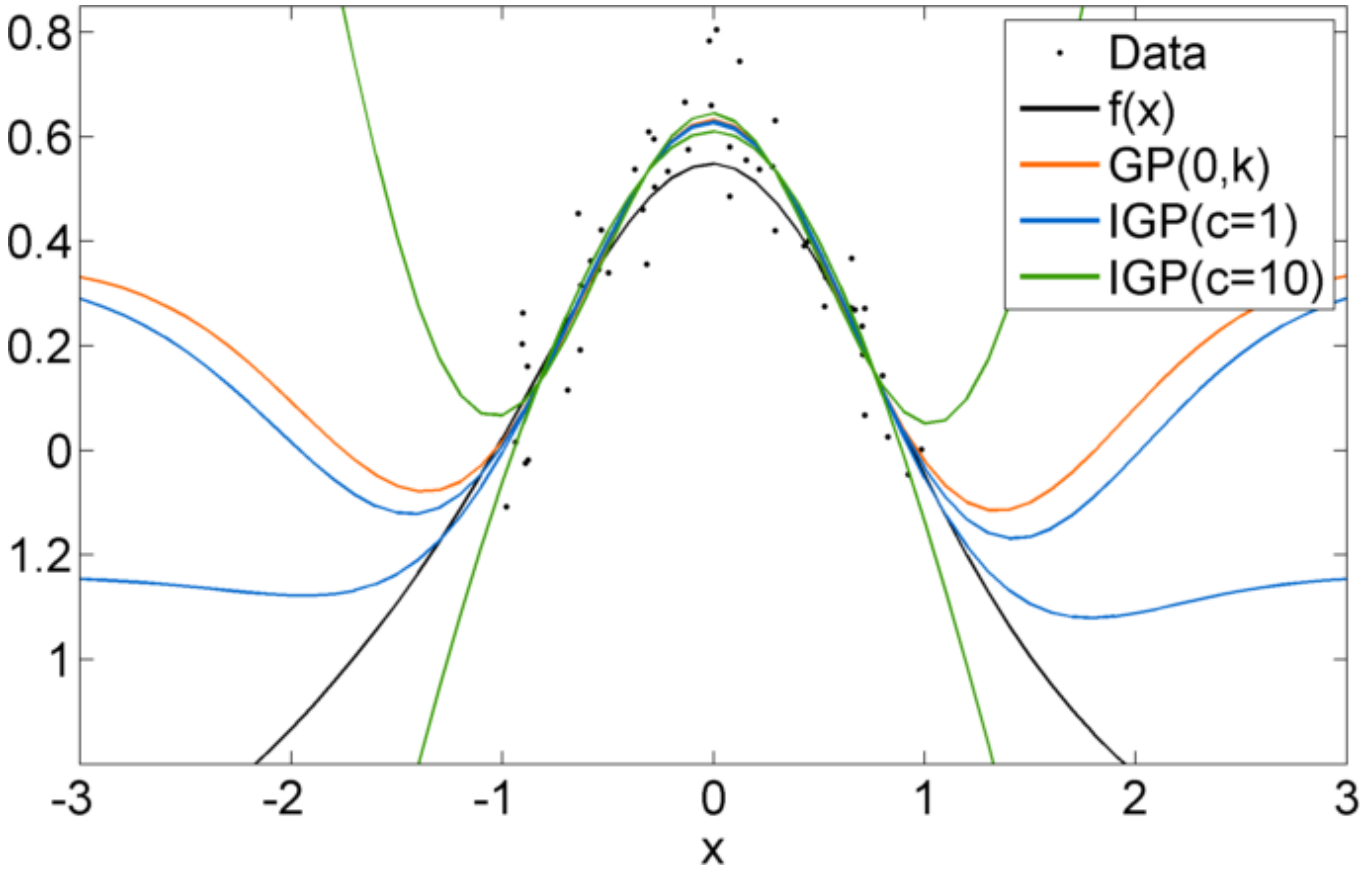
with $\mathbf{s}_k = K_n^{-1} \mathbb{1}_n$, $S_k = \mathbb{1}_n^T K_n^{-1} \mathbb{1}_n$.

- 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

Estimates of $E[f(x)]$ given $n = 50$ observations $(\mathbf{x}, \mathbf{y})$.



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

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

Procedure

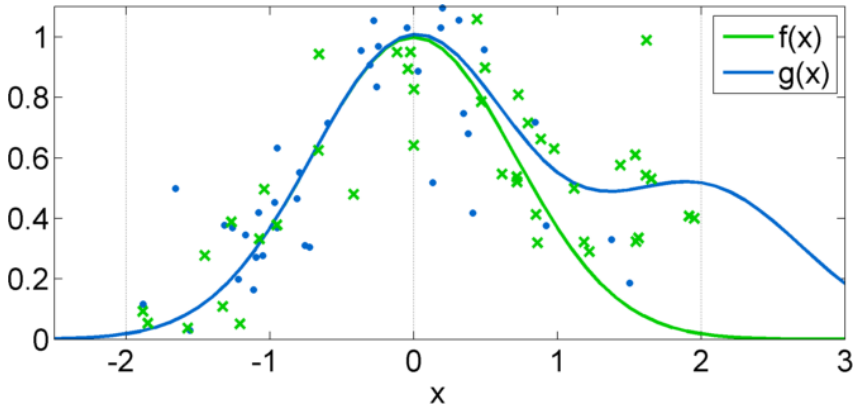
- Consider a vector $\mathbf{x}^*$ of equispaced inputs $\in \mathcal{X}_T$;
- Derive the credible region (CI) of $\Delta\mu(\mathbf{x}^*)$ from the chi-squared random variable

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

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

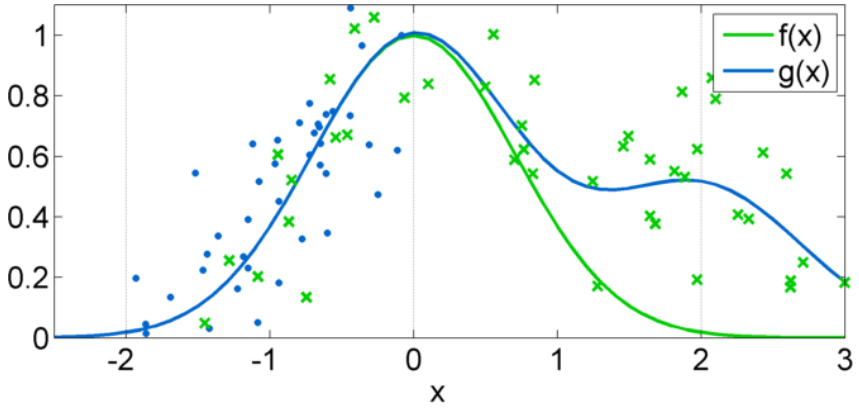
- If, a posteriori, $\mathbf{0} \notin$ CI conclude that $f_1 \neq f_2$.
  Indecision: If different priors entail different decisions, a robust decision cannot be made in $\mathcal{X}_T$.

Numerical example



$\mathcal{X}_T$	**GP**	**c-IGP**
Case A: $x_i^{(1)} \sim U[-2,2]$, $x_i^{(2)} \sim U[-2,2]$		
$[-2,\ 0]$	0	**0**
$[-2,\ 2]$	1	**1**

Numerical example



$\mathcal{X}_T$	GP	c-IGP
Case A: $x_i^{(1)} \sim U[-2,2]$, $x_i^{(2)} \sim U[-2,2]$		
$[-2,\ 0]$	0	**0**
$[-2,\ 2]$	1	**1**
Case B: $x_i^{(1)} \sim U[-2,0]$, $x_i^{(2)} \sim U[-2,4]$		
$[-2,\ 0]$	0	**0**
$[-2,\ 2]$	0	**2**

Conclusions

- We have presented a general framework for modeling prior near ignorance about $f(x)$ based on the Gaussian process (IGP).
- We have derived an IGP model with prior constant mean free to vary between $-\infty$ and $+\infty$:
  - with many observations the IGP and GP inferences almost coincide;
  - where there are no observations the imprecision of the IGP is very high, reflecting the actual lack of knowledge.
  - Applied to hypothesis testing, the IGP acknowledges when the available data are not informative enough to make a robust decision.
- Future research should focus on
  - the study of other prior near ignorance models based on different sets $\mathcal{H}$ of base mean functions;
  - the development of models allowing for a weaker specification of the kernel function.