Gaussian Process Regression under Inequality Constraints

Andrés F. López-Lopera[1], François Bachoc[2], Nicolas Durrande[1,3], and Olivier Roustant[1]

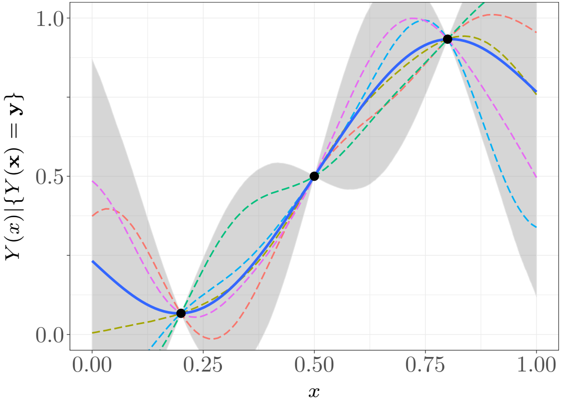
[1]École des Mines de Saint-Étienne (EMSE), France.
[2]Institut de Mathématiques de Toulouse (IMT), France.
[3]PROWLER.io, Cambridge, UK.

This work is funded by the chair of applied mathematics OQUAIDO.

2019 MASCOT-NUM Annual Conference

March 18, 2019

GPs are one of the most famous Bayesian approaches for placing prior distributions over the space of functions (Rasmussen and Williams, 2005).



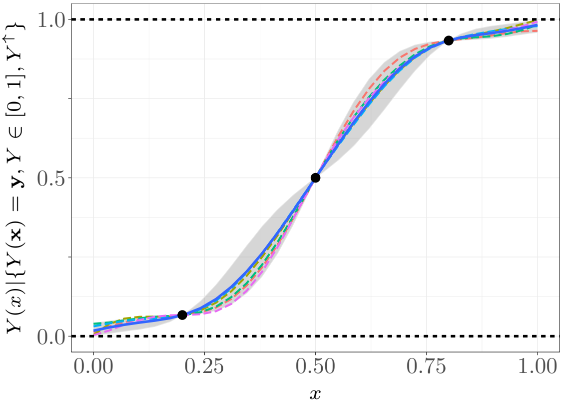
Mean function
Prediction intervals
⋯ Realizations

- Interpolation conditions

$$\mathcal{D} = (x_i, y_i)_{i=1}^n = (\mathbf{x}, \mathbf{y})$$

$$Y|\{Y(\mathbf{x}) = \mathbf{y}\} \sim \mathcal{GP}(\underbrace{m(x)}_{\text{cond. mean}}, \underbrace{c_\theta(x, x')}_{\text{cond. covariance}})$$

GPs are one of the most famous Bayesian approaches for placing prior distributions over the space of functions (Rasmussen and Williams, 2005).



Mean function

Prediction intervals

$\cdots$ Realizations

• Interpolation conditions

$$\mathcal{D} = (x_i, y_i)_{i=1}^n = (\mathbf{x}, \mathbf{y})$$

- - Boundedness condition

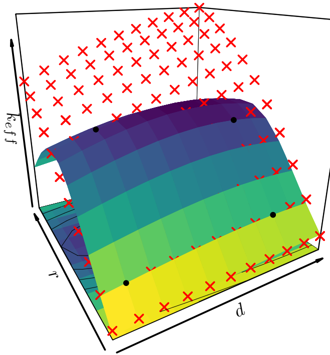
$$Y \in [0, 1]$$

$\uparrow$ Monotonicity condition

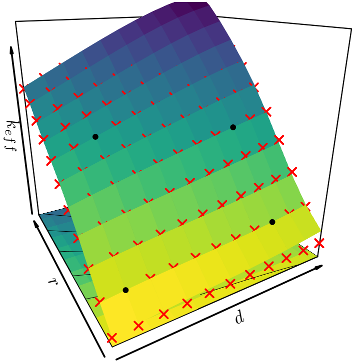
$$Y(x) \leq Y(x'), \quad \forall\, x \leq x'.$$

$$Y|\{Y(\mathbf{x}) = \mathbf{y}, Y \in [0,1], Y^{\uparrow}\} \sim \; ?$$

2D models for interpolating the IRSN's dataset using $n = 4$.



(a) Unconstrained GP



(b) GP with positivity & monotonicity

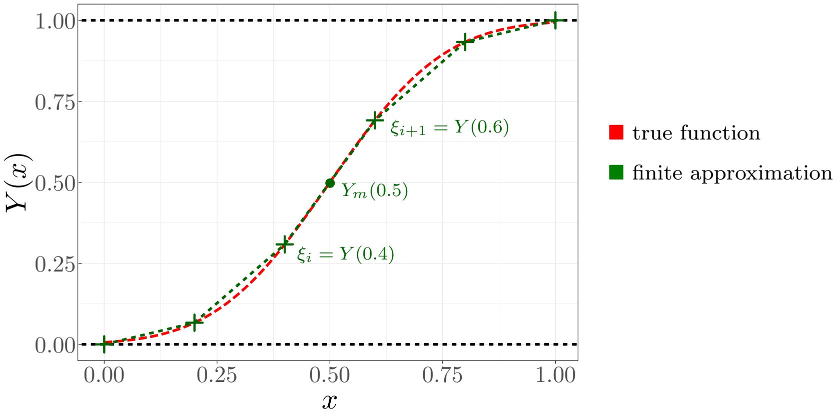
k_{eff}: effective neutron multiplication factor.

● interpolation points ✕ test data

Table of Contents

1D finite-dimensional Gaussian approximation

Finite-dimensional representation Y_m: also bounded & monotonic.



✦ Imposing constraints on the knots is enough (Maatouk and Bay, 2017).

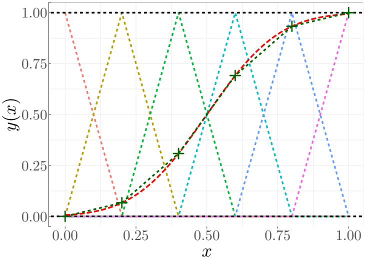
1D finite-dimensional Gaussian approximation

- Let the finite-dimensional Gaussian approximation be defined as

$$Y_m(x) = \sum_{j=1}^{m} \xi_j \phi_j(x), \text{ s.t. } \begin{cases} Y_m(x_i) = y_i & \text{(interpolation conditions)}, \\ \boldsymbol{l} \leq \boldsymbol{\Lambda\xi} \leq \boldsymbol{u} & \text{(linear inequality conditions)}, \end{cases} \quad (1)$$

where $x_i \in [0, 1]$, $y_i \in \mathbb{R}$ for $i = 1, \cdots, n$; and

- $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_m]^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_\theta)$ with covariance matrix $\boldsymbol{\Gamma}_\theta$,
- $(\boldsymbol{\Lambda}, \boldsymbol{l}, \boldsymbol{u})$ defines the inequality conditions, and
- $\phi_j : [0, 1] \mapsto \mathbb{R}$ are hat functions as in (Maatouk and Bay, 2017):



- Since the Gaussianity is preserved for linear operations, we have that

$$\boldsymbol{\Lambda}\boldsymbol{\xi} | \{\boldsymbol{\Phi}\boldsymbol{\xi} = \mathbf{y}\} \sim \mathcal{N}\left(\boldsymbol{\Lambda}\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}^\top\right), \quad \text{(conditional distribution)} \tag{2}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix of the distribution $\boldsymbol{\xi} | \{\boldsymbol{\Phi}\boldsymbol{\xi} = \mathbf{y}\} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Rasmussen and Williams, 2005):

$$\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^\top, \quad \boldsymbol{\mu} = \boldsymbol{\Gamma}\boldsymbol{\Phi}^\top\mathbf{K}^{-1}\mathbf{y}, \quad \boldsymbol{\Sigma} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\boldsymbol{\Phi}^\top\mathbf{K}^{-1}\boldsymbol{\Phi}\boldsymbol{\Gamma}. \tag{3}$$

- Since the Gaussianity is preserved for linear operations, we have that

$$\boldsymbol{\Lambda}\boldsymbol{\xi} | \{\boldsymbol{\Phi}\boldsymbol{\xi} = \mathbf{y}\} \sim \mathcal{N}\left(\boldsymbol{\Lambda}\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}^\top\right), \quad \text{(conditional distribution)} \tag{2}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix of the distribution $\boldsymbol{\xi} | \{\boldsymbol{\Phi}\boldsymbol{\xi} = \mathbf{y}\} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Rasmussen and Williams, 2005):

$$\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^\top, \quad \boldsymbol{\mu} = \boldsymbol{\Gamma}\boldsymbol{\Phi}^\top\mathbf{K}^{-1}\mathbf{y}, \quad \boldsymbol{\Sigma} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\boldsymbol{\Phi}^\top\mathbf{K}^{-1}\boldsymbol{\Phi}\boldsymbol{\Gamma}. \tag{3}$$

- Then, the posterior distribution follows (López-Lopera et al., 2018)

$$\boldsymbol{\Lambda}\boldsymbol{\xi} | \{\boldsymbol{\Phi}\boldsymbol{\xi} = \mathbf{y}, \boldsymbol{l} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\} \sim \mathcal{TN}(\boldsymbol{\Lambda}\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}^\top, \boldsymbol{l}, \boldsymbol{u}), \tag{4}$$

✦ The posterior distribution in (4) can be approximated via MC/MCMC (e.g. via Hamiltonian MC–HMC as in Pakman and Paninski, 2014).

- In practice, the posterior mode (MAP solution) can be used as starting point for MCMC algorithms aiming a rapid convergence.
- Let $\boldsymbol{\mu}^*$ be the mode that maximises the pdf $\boldsymbol{\xi}|\{\boldsymbol{\Phi}\boldsymbol{\xi} = \mathbf{y}, \boldsymbol{l} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\}$, i.e.

$$\boldsymbol{\mu}^* = \underset{\boldsymbol{\xi} \text{ s.t. } \boldsymbol{l} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}}{arg\,max} \{-[\boldsymbol{\xi} - \boldsymbol{\mu}]^\top \boldsymbol{\Sigma}^{-1} [\boldsymbol{\xi} - \boldsymbol{\mu}]\}. \quad (5)$$

- Then, the MAP estimate of Y_m is given by

$$Y_m^{\mathrm{MAP}}(x) = \sum_{j=1}^{m} \mu_j^* \phi_j(x). \quad (6)$$

- In practice, the posterior mode (MAP solution) can be used as starting point for MCMC algorithms aiming a rapid convergence.
- Let $\boldsymbol{\mu}^*$ be the mode that maximises the pdf $\boldsymbol{\xi} | \{\boldsymbol{\Phi}\boldsymbol{\xi} = \mathbf{y}, \boldsymbol{l} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\}$, i.e.

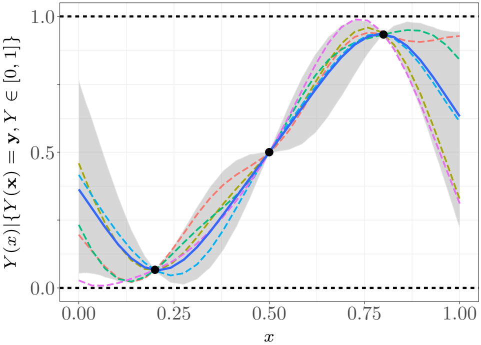
$$\boldsymbol{\mu}^* = \underset{\boldsymbol{\xi} \text{ s.t. } \boldsymbol{l} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}}{arg\,max} \{-[\boldsymbol{\xi} - \boldsymbol{\mu}]^\top \boldsymbol{\Sigma}^{-1} [\boldsymbol{\xi} - \boldsymbol{\mu}]\}. \tag{5}$$

- Then, the MAP estimate of Y_m is given by

$$Y_m^{\mathrm{MAP}}(x) = \sum_{j=1}^{m} \mu_j^* \phi_j(x). \tag{6}$$

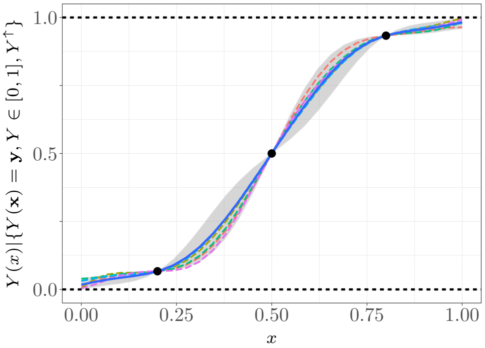
- The benefits of computing $\boldsymbol{\mu}^*$ are:
  - $\boldsymbol{\mu}^*$ can be easily used as a point estimate since its easy calculation,
  - it converges to the spline solution as $m \to \infty$ (Bay et al., 2016).

1D example with boundedness constraints via HMC



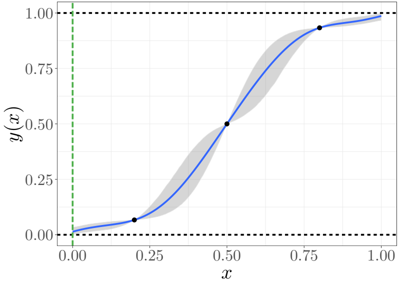
$$\underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}}_{\boldsymbol{l}} \leq \underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}}_{\boldsymbol{\Lambda}} \underbrace{\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \vdots \\ \xi_{m-1} \\ \xi_m \end{bmatrix}}_{\boldsymbol{\xi}} \leq \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}}_{\boldsymbol{u}}$$

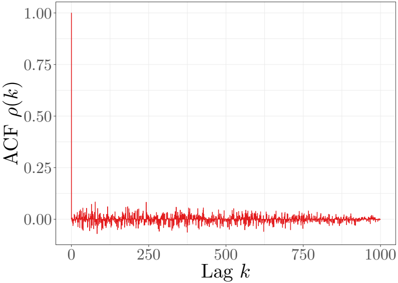
1D example with boundedness & monotonicity constraints via HMC



$$\underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}}_{\boldsymbol{l}} \leq \underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}}_{\boldsymbol{\Lambda}} \underbrace{\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \vdots \\ \xi_{m-1} \\ \xi_m \end{bmatrix}}_{\boldsymbol{\xi}} \leq \underbrace{\begin{bmatrix} 1 \\ \infty \\ \infty \\ \vdots \\ \infty \\ 1 \end{bmatrix}}_{\boldsymbol{u}}$$

GP under boundedness & monotonicity constraints ($m = 100,\ N = 10^3$)





MC/MCMC Method	CPU Time $[s]$	Effective sampling rate $[s^{-1}]$
(Maatouk and Bay, 2016)	-	-
(Botev, 2017)	73.7	0.012
Gibbs – (Taylor and Benjamini, 2017)	29.0	0.006
HMC – (Pakman and Paninski, 2014)	**12.6**	**0.061**

1D finite-dimensional Gaussian approximation with noisy data

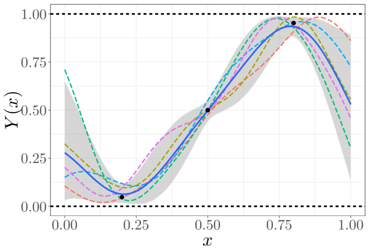
- Noise effects can also be considered (López-Lopera et al., 2019):

$$Y_m(x_i) + \varepsilon_i = y_i \tag{7}$$

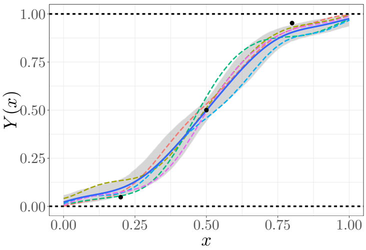
where $x_i \in [0, 1]$, $y_i \in \mathbb{R}$ and $\varepsilon_i \sim \mathcal{N}(0, \tau^2)$ with noise variance τ^2.

$$\mathbf{K} = \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^\top \quad \rightarrow \quad \mathbf{K} = \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^\top + \tau^2\mathbf{I}.$$

boundedness



boundedness & monotonicity



✦ Adding ε_i leads to more flexible GP models and less restrictive sample spaces improving the performance of MC/MCMC algorithms (López-Lopera et al., 2019).

Table of Contents

- The finite representation Y_m can be scaled to $d \geq 2$ by **tensorisation**:

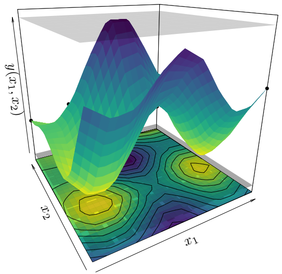
$$Y_m(\mathbf{x}) = \sum_{j_1,\cdots,j_d=1}^{m_1,\cdots,m_d} \left[\prod_{k=1,\cdots,d} \phi_{j_k}^k(x_k) \right] \xi_{j_1,\cdots,j_d}, \text{ s.t. } \begin{cases} Y_m(\mathbf{x}_i) = y_i, \\ \boldsymbol{\xi} \in \mathcal{C}, \end{cases} \tag{8}$$

where $\mathbf{x}_i \in [0,1]^d$, $y_i \in \mathbb{R}$ for $i = 1, \cdots, n$; and

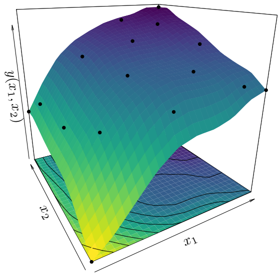
- $\xi_{j_1,\cdots,j_d} := [\xi_{1,\cdots,1}, \ \cdots, \ \xi_{m_1,\cdots,m_d}]^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_\theta)$ with covariance $\boldsymbol{\Gamma}_\theta$,
- $\mathcal{C}$ is a convex set of $\mathbb{R}^{m_1 \times \cdots \times m_d}$ composed by a set of linear inequalities,
- $\phi_{j_i}^i : [0,1] \mapsto \mathbb{R}$ are hat basis functions.

Extension to high dimensions: 2D examples

2D Examples under boundedness constr.



2D Examples under monotonicity constr.



● interpolation points ■ predictive mean ■ bounds

MC/MCMC Method	2D boundedness example CPU Time [s]	2D boundedness example Eff. sampling rate [s^{-1}]	2D monotonicity example CPU Time [s]	2D monotonicity example Eff. sampling rate [s^{-1}]
(Maatouk and Bay, 2016)	-	-	-	-
(Botev, 2017)	0.9	1.009	1488.3	0.001
(Taylor and Benjamini, 2017)	9.7	0.088	-	-
(Pakman and Paninski, 2014)	**0.6**	**1.493**	**8.6**	**0.093**

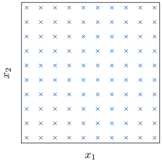
$$m_1 = m_2 = 10, \quad N = 10^3$$

MINES Saint-Étienne

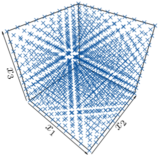
INSTITUT DE MATHÉMATIQUES DE TOULOUSE

Curse of dimensionality

- The cost of Y_m increases as d (or $m = m_1 \times \cdots \times m_d$) increases.



$m = m_1 \times m_2 = 10^2$

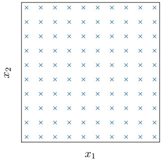


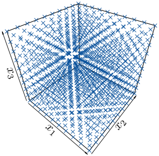
$m = m_1 \times m_2 \times m_3 = 10^3$

Extension to high dimensions

Curse of dimensionality

- The cost of Y_m increases as d (or $m = m_1 \times \cdots \times m_d$) increases.

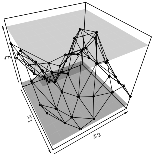


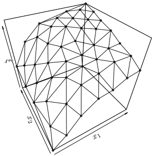


$m = m_1 \times m_2 = 10^2$ $\qquad$ $m = m_1 \times m_2 \times m_3 = 10^3$

However, this downside can be partially mitigated in different ways...

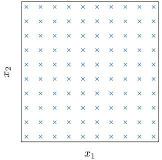
- Using a smarter construction of rectangular grids of the knots.
- Assuming further assumptions for complexity simplification.
  - e.g. inactive dimensions, additive conditions.
- Using other types of designs for the knots: e.g. Delaunay tessellation.

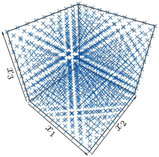




Curse of dimensionality

- The cost of Y_m increases as d (or $m = m_1 \times \cdots \times m_d$) increases.

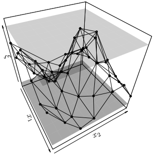


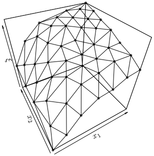


$m = m_1 \times m_2 = 10^2$ $\qquad$ $m = m_1 \times m_2 \times m_3 = 10^3$

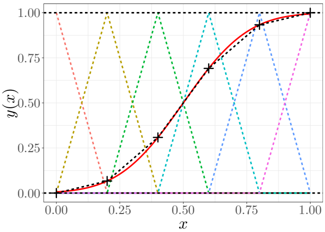
However, this downside can be partially mitigated in different ways...

- Using a smarter construction of rectangular grids of the knots. ✓
- Assuming further assumptions for complexity simplification.
  - e.g. inactive dimensions, additive conditions. ✓
- Using other types of designs for the knots: e.g. Delaunay tessellation.

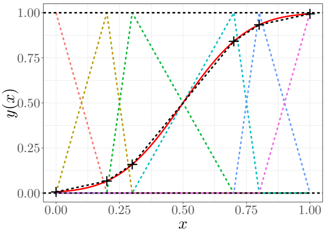




- Using a "smarter" construction of rectangular grids of the knots:



Standard hat basis functions.

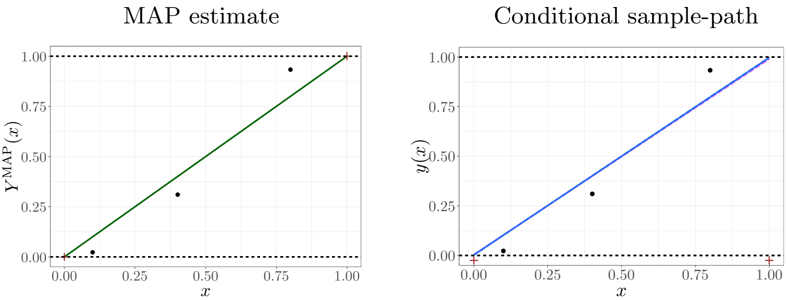


Asymmetric hat basis functions.

✦ This allows refining the grid in places requiring more quality of resolution.

- The "optimal" addition of knots can be achieved by maximising the MAP total-variation (MAP-TV) criterion,

$$t_*^{\mathrm{opt}} = \operatorname{argmax}_{t_*} \int_0^1 [Y_{m_i+1}^{\mathrm{MAP}}(x) - Y_{m_i}^{\mathrm{MAP}}(x)]^2 dx. \tag{9}$$

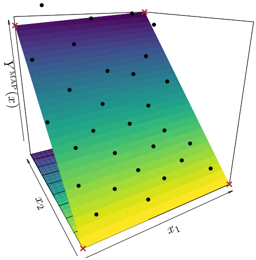


● training points + knots ■ MAP estimate
■ predictive mean ■ 90% confidence intervals

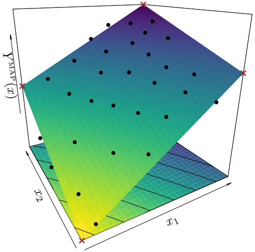
- For the multidimensional case ($d \geq 2$), the MAP-TV criterion is given by,

$$t_*^{\text{opt}} = \operatorname{argmax}_{t_*} \int_{\mathbf{x}\in[0,1]^d} \left[Y_{m_{i+1}}^{\text{MAP}}(\mathbf{x}) - Y_{m_i}^{\text{MAP}}(\mathbf{x})\right]^2 d\mathbf{x}. \tag{10}$$

Toy 1 – MAP estimate
$y(x_1, x_2) = x_2^2$



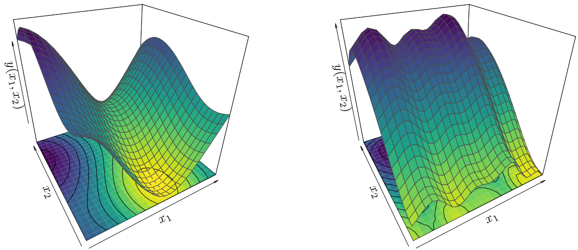
Toy 2 – MAP estimate
$y(x_1, x_2) = \arctan(5x_1) + \arctan(x_2)$



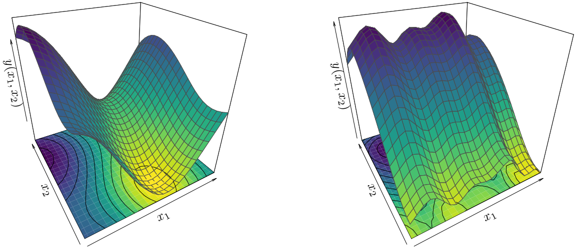
• training points + knots ■ MAP estimate

✦ This construction has been tested up to 5D applications!

Variation II: additive GPs under inequality constraints



2D examples of additive GPs $Y(x_1, x_2) = Y_1(x_1) + Y_2(x_2)$.



2D examples of additive GPs $Y(x_1, x_2) = Y_1(x_1) + Y_2(x_2)$.

- Let Y be the first-order additive process on $\mathbb{R}$ given by

$$Y(x_1, \cdots, x_d) = \sum_{\kappa=1}^{d} Y_\kappa(x_\kappa), \tag{11}$$

where the Y_κ's are centred GPs on $\mathbb{R}$ with covariances k_κ's.

- Assuming that $Y_1, \cdots, Y_d$ are independent, then Y is also a centred GP with covariance function given by (Durrande et al., 2012)

$$k(x_1, \cdots, x_d; x_1', \cdots, x_d') = \sum_{\kappa=1}^{d} k_\kappa(x_\kappa, x_\kappa').$$

- Assume that Y exhibits certain inequalities along each Y_κ, then Y_m is

$$Y_m(x_1, \cdots, x_d) = \sum_{\kappa=1}^{d} Y_\kappa^{m_\kappa}(x_\kappa), \ \text{s.t.} \ \begin{cases} Y_m(x_1^i, \cdots, x_d^i) + \varepsilon_i = y_i, \\ \boldsymbol{\xi}^\kappa \in \mathcal{C}_\kappa, \end{cases} \tag{12}$$

where $(x_1^i, \cdots, x_d^i) \in [0,1]^d$, $y_i \in \mathbb{R}$ and $\varepsilon_i \sim \mathcal{N}\left(0, \tau^2\right)$ for $i = 1, \cdots, n$.

- Assume that Y exhibits certain inequalities along each Y_κ, then Y_m is

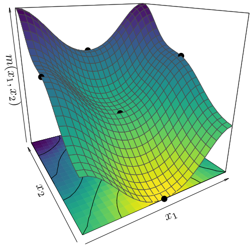
$$Y_m(x_1, \cdots, x_d) = \sum_{\kappa=1}^{d} Y_\kappa^{m_\kappa}(x_\kappa), \text{ s.t. } \begin{cases} Y_m(x_1^i, \cdots, x_d^i) + \varepsilon_i = y_i, \\ \boldsymbol{\xi}^\kappa \in \mathcal{C}_\kappa, \end{cases} \tag{12}$$

where $(x_1^i, \cdots, x_d^i) \in [0,1]^d$, $y_i \in \mathbb{R}$ and $\varepsilon_i \sim \mathcal{N}\left(0, \tau^2\right)$ for $i = 1, \cdots, n$.

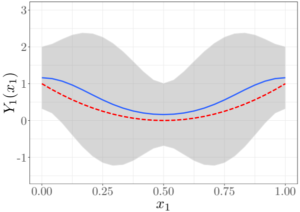
2D toy: $Y(x_1, x_2) = \underbrace{4(x_1 - 0.5)^2}_{Y_1(x_1)} + \underbrace{2x_2}_{Y_2(x_2)}$

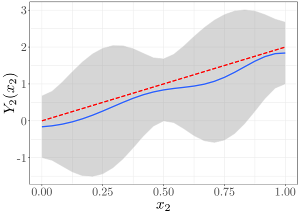
Predictive mean without constraints

$$m(x_1, x_2) = m_1(x_1) + m_2(x_2)$$



$(\sigma_1^2, \ell_1) = (\sigma_2^2, \ell_2) = (1, 0.2)$





- Assume that Y exhibits certain inequalities along each Y_κ, then Y_m is

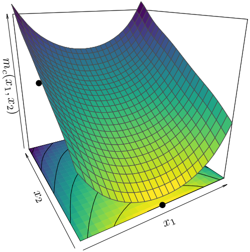
$$Y_m(x_1, \cdots, x_d) = \sum_{\kappa=1}^{d} Y_\kappa^{m_\kappa}(x_\kappa), \text{ s.t. } \begin{cases} Y_m(x_1^i, \cdots, x_d^i) + \varepsilon_i = y_i, \\ \boldsymbol{\xi}^\kappa \in \mathcal{C}_\kappa, \end{cases} \tag{12}$$

where $(x_1^i, \cdots, x_d^i) \in [0,1]^d$, $y_i \in \mathbb{R}$ and $\varepsilon_i \sim \mathcal{N}\left(0, \tau^2\right)$ for $i = 1, \cdots, n$.

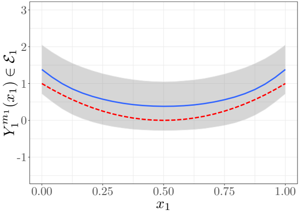
2D toy: $Y(x_1, x_2) = \underbrace{4(x_1 - 0.5)^2}_{Y_1(x_1)} + \underbrace{2x_2}_{Y_2(x_2)}$

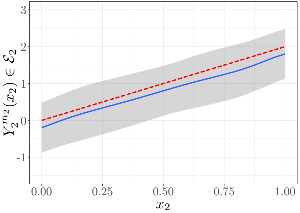
Predictive mean with constraints

$m_c(x_1, x_2) = m_{c,1}(x_1) + m_{c,2}(x_2)$



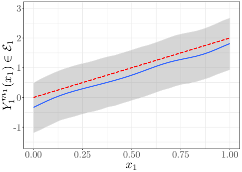
$(\sigma_1^2, \ell_1) = (\sigma_2^2, \ell_2) = (1, 0.2)$



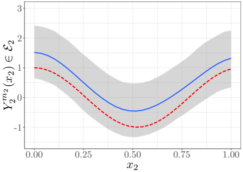


5D Toy example:

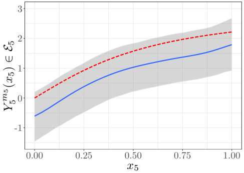
$$Y(x_1, \cdots, x_5) = \underbrace{2x_1}_{Y_1(x_1)} + \underbrace{\cos(6x_2)}_{Y_2(x_2)} + \underbrace{2x_3^2}_{Y_3(x_3)} + \underbrace{4(x_4 - 0.5)^2}_{Y_4(x_4)} + \underbrace{2\arctan(2x_5)}_{Y_5(x_5)}$$



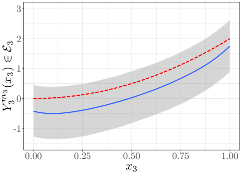
(a) Monotonicity



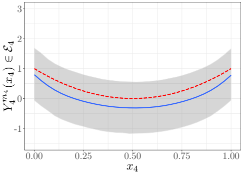
(b) No constraints



(e) Monotonicity



(c) Convexity



(d) Convexity

- We used 25 points from a maximin LHS.
- For κ=1,$\cdots$,5, we fixed $(\sigma_\kappa^2, \ell_\kappa) = (1, 0.2)$.

✦ Higher dimensions can be considered since HMC is performed in 1D!

MINES Saint-Étienne

INSTITUT DE MATHÉMATIQUES DE TOULOUSE

Table of Contents

Table of Contents

- The GP covariance parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ can be estimated via maximum likelihood (Stein, 1999).
  - Consider the observation vector

$$\mathbf{Y}_n = \left[Y(x_1), \cdots, Y(x_n) \right]^\top.$$

  - Let $\mathcal{L}_n(\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\mathbf{Y}_n)$ be the unconstrained log-likelihood

$$\mathcal{L}_n(\boldsymbol{\theta}) = -\frac{1}{2}\log(\det(\mathbf{R}_{\boldsymbol{\theta}})) - \frac{1}{2}\mathbf{Y}_n^\top \mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{Y}_n - \frac{n}{2}\log 2\pi, \tag{13}$$

  with $\mathbf{R}_{\boldsymbol{\theta}} = (k_{\boldsymbol{\theta}}(x_i, x_j))_{1 \leq i,j \leq n}$.

- Then, the MLE is given by

$$\widehat{\boldsymbol{\theta}}_n \in \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{argmax}\ \mathcal{L}_n(\boldsymbol{\theta}). \tag{14}$$

✗ Notice that the MLE does not take into account the inequality conditions.

- We studied a constrained likelihood in (López-Lopera et al., 2018).
  - Consider a convex set with linear inequality constraints $\mathcal{E}$.
  - Let $\mathcal{L}_{n,c}(\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\mathbf{Y}_n | Y \in \mathcal{E})$ be the constrained log-likelihood:

$$\begin{aligned} \mathcal{L}_{n,c}(\boldsymbol{\theta}) &= \log \frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_n) p_{\boldsymbol{\theta}}(Y \in \mathcal{E} | \mathbf{Y}_n)}{p_{\boldsymbol{\theta}}(Y \in \mathcal{E})} \\ &= \mathcal{L}_n(\boldsymbol{\theta}) + \log p_{\boldsymbol{\theta}}(Y \in \mathcal{E} | \mathbf{Y}_n) - \log p_{\boldsymbol{\theta}}(Y \in \mathcal{E}). \end{aligned} \tag{15}$$

- Then, the cMLE is given by

$$\widehat{\boldsymbol{\theta}}_{n,c} \in \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{argmax}\ \mathcal{L}_{n,c}(\boldsymbol{\theta}). \tag{16}$$

✓ Note that the cMLE accounts for the inequality constraints.

- The terms $p_{\boldsymbol{\theta}}(Y \in \mathcal{E} | \mathbf{Y}_n)$ and $p_{\boldsymbol{\theta}}(Y \in \mathcal{E})$ have to be approximated via MC.

Table of Contents

Consistency of the MLE & cMLE

- Let $\mathcal{E}_\kappa$ be one of the following convex set of functions (mild conditions)

$$\mathcal{E}_\kappa = \begin{cases} f \ : \ \mathbb{X} \to \mathbb{R}, f \text{ is } C^0 \text{ and } \forall \mathbf{x} \in \mathbb{X}, \ \ell \leq f(\mathbf{x}) \leq u & \text{if } \kappa = 0, \\ f \ : \ \mathbb{X} \to \mathbb{R}, f \text{ is } C^1 \text{ and } \forall \mathbf{x} \in \mathbb{X}, \ \forall i = 1, \cdots, d, \ \frac{\partial}{\partial x_i} f(\mathbf{x}) \geq 0 & \text{if } \kappa = 1, \\ f \ : \ \mathbb{X} \to \mathbb{R}, f \text{ is } C^2 \text{ and } \forall \mathbf{x} \in \mathbb{X}, \ \frac{\partial^2}{\partial \mathbf{x}^2} f(\mathbf{x}) \text{ is a non-negative definite matrix} & \text{if } \kappa = 2. \end{cases}$$

corresponding to boundedness, monotonicity, and convexity constraints.

- We consider $\boldsymbol{\theta}_0$ as the true unknown covariance parameters.

- Let $\mathcal{E}_\kappa$ be one of the following convex set of functions (mild conditions)

$$\mathcal{E}_\kappa = \begin{cases} f \;:\; \mathbb{X} \to \mathbb{R},\, f \text{ is } C^0 \text{ and } \forall \mathbf{x} \in \mathbb{X},\; \ell \leq f(\mathbf{x}) \leq u & \text{if } \kappa = 0, \\ f \;:\; \mathbb{X} \to \mathbb{R},\, f \text{ is } C^1 \text{ and } \forall \mathbf{x} \in \mathbb{X},\; \forall i = 1, \cdots, d,\; \frac{\partial}{\partial x_i} f(\mathbf{x}) \geq 0 & \text{if } \kappa = 1, \\ f \;:\; \mathbb{X} \to \mathbb{R},\, f \text{ is } C^2 \text{ and } \forall \mathbf{x} \in \mathbb{X},\; \frac{\partial^2}{\partial \mathbf{x}^2} f(\mathbf{x}) \text{ is a non-negative definite matrix} & \text{if } \kappa = 2. \end{cases}$$

corresponding to boundedness, monotonicity, and convexity constraints.

- We consider $\boldsymbol{\theta}_0$ as the true unknown covariance parameters.

Proposition (Asymptotic consistency of cMLE, López-Lopera et al. (2018))

Assume mild conditions (López-Lopera et al., 2018). Assume $\forall \varepsilon > 0$ and $\forall M < \infty$, (Consistency of the unconditional ML)

$$P\left(\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \varepsilon} (\mathcal{L}_n(\boldsymbol{\theta}) - \mathcal{L}_n(\boldsymbol{\theta}_0)) \geq -M \right) \xrightarrow[n \to +\infty]{} 0.$$

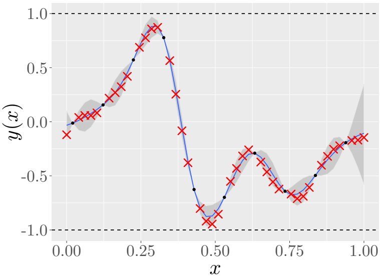
Then, (Consistency of the conditional cML)

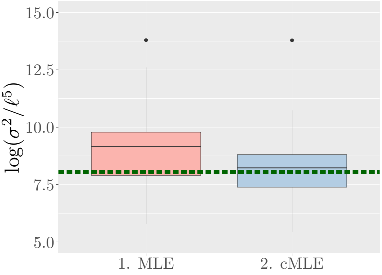
$$P\left(\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \varepsilon} (\mathcal{L}_{n,c}(\boldsymbol{\theta}) - \mathcal{L}_{n,c}(\boldsymbol{\theta}_0)) \geq -M \;\middle|\; Y \in \mathcal{E}_\kappa \right) \xrightarrow[n \to +\infty]{} 0.$$

Consequently, both MLE and cMLE are consistent estimators

$$\widehat{\boldsymbol{\theta}}_n \in \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{argmax}\; \mathcal{L}_n(\boldsymbol{\theta}) \xrightarrow[n \to +\infty]{P} \boldsymbol{\theta}_0, \qquad \widehat{\boldsymbol{\theta}}_{n,c} \in \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{argmax}\; \mathcal{L}_{n,c}(\boldsymbol{\theta}) \xrightarrow[n \to +\infty]{P | Y \in \mathcal{E}_\kappa} \boldsymbol{\theta}_0.$$

Assessment of the MLE and cMLE for **10^2 realisations** drawn from a constrained GP $Y \in [-1, 1]$ with Matérn 5/2 kernel and true parameters $\boldsymbol{\theta}_0 = (\sigma^2 = 1, \ell = 0.2)$.





(b) Consistently estimable ratio $\log(\sigma^2/\ell^5)$

- For instance, we focus on estimating a single variance parameter σ^2, i.e.

$$k_{\sigma^2}(x, x') = \sigma^2 k_1(x, x'),$$

with known correlation function k_1 and true variance parameter σ_0^2.

- For instance, we focus on estimating a single variance parameter σ^2, i.e.

$$k_{\sigma^2}(x, x') = \sigma^2 k_1(x, x'),$$

with known correlation function k_1 and true variance parameter σ_0^2.

Theorem (Asymptotic normality of MLE, Bachoc et al. (2018))

Assume mild conditions (Bachoc et al., 2018). Then, the MLE $\widehat{\sigma}_n^2$ of σ_0^2 conditioned on $\{Y \in \mathcal{E}_\kappa\}$ is asymptotically Gaussian distributed. More precisely, for $\kappa = 0, 1, 2$,

$$\sqrt{n}\left(\widehat{\sigma}_n^2 - \sigma_0^2\right) \xrightarrow[n \to +\infty]{\mathcal{L}|Y \in \mathcal{E}_\kappa} \mathcal{N}(0, 2\sigma_0^4). \quad (17)$$

- For instance, we focus on estimating a single variance parameter σ^2, i.e.

$$k_{\sigma^2}(x, x') = \sigma^2 k_1(x, x'),$$

with known correlation function k_1 and true variance parameter σ_0^2.

Theorem (Asymptotic normality of MLE, Bachoc et al. (2018))

Assume mild conditions (Bachoc et al., 2018). Then, the MLE $\widehat{\sigma}_n^2$ of σ_0^2 conditioned on $\{Y \in \mathcal{E}_\kappa\}$ is asymptotically Gaussian distributed. More precisely, for $\kappa = 0, 1, 2$,

$$\sqrt{n}\left(\widehat{\sigma}_n^2 - \sigma_0^2\right) \xrightarrow[n \to +\infty]{\mathcal{L}|Y \in \mathcal{E}_\kappa} \mathcal{N}(0, 2\sigma_0^4). \tag{17}$$

Theorem (Asymptotic normality of cMLE, Bachoc et al. (2018))

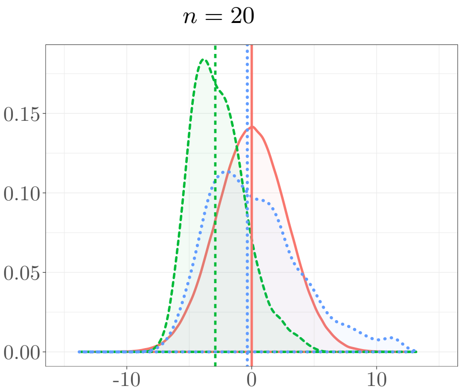
The cMLE $\widehat{\sigma}_{n,c}^2$ of σ_0^2 conditioned on $\{Y \in \mathcal{E}_\kappa\}$ is also asymptotically Gaussian distributed. More precisely, for $\kappa = 0, 1, 2$,

$$\sqrt{n}\left(\widehat{\sigma}_{n,c}^2 - \sigma_0^2\right) \xrightarrow[n \to +\infty]{\mathcal{L}|Y \in \mathcal{E}_\kappa} \mathcal{N}(0, 2\sigma_0^4). \tag{18}$$

✦ These results can be extended for the Matérn models (Bachoc et al., 2018).

Asymptotic normality of the MLE and cMLE: numerical illustration.

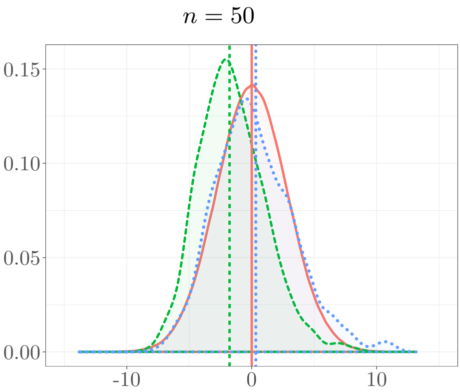
Assessment of the MLE and cMLE normality for **10^3 trajectories** drawn from a constrained. GP $Y \in [0, 1]$ with Matérn 5/2 kernel and parameters $\boldsymbol{\theta}_0 = (\sigma_0^2 = 0.2, \ell = 0.2)$.



■ limit dist. $\mathcal{N}(0, 2\sigma_0^4)$ ■ **MLE**: $\sqrt{n}\left(\widehat{\sigma}_n^2 - \sigma_0^2\right)$ ■ **cMLE**: $\sqrt{n}\left(\widehat{\sigma}_{n,c}^2 - \sigma_0^2\right)$

✦ For small values of n, **cMLE** seems to be more accurate.

Assessment of the MLE and cMLE normality for 10^3 **trajectories** drawn from a constrained. GP $Y \in [0,1]$ with Matérn 5/2 kernel and parameters $\boldsymbol{\theta}_0 = (\sigma_0^2 = 0.2, \ell = 0.2)$.

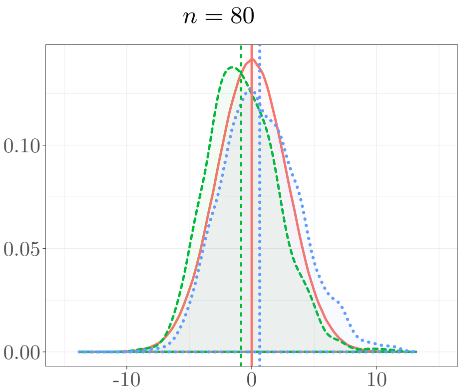


■ limit dist. $\mathcal{N}(0, 2\sigma_0^4)$ ■ **MLE**: $\sqrt{n}\left(\widehat{\sigma}_n^2 - \sigma_0^2\right)$ ■ **cMLE**: $\sqrt{n}\left(\widehat{\sigma}_{n,c}^2 - \sigma_0^2\right)$

✦ **cMLE** converges faster to the limit Gaussian distribution.

Asymptotic normality of the MLE and cMLE: numerical illustration.

Assessment of the MLE and cMLE normality for **10^3 trajectories** drawn from a constrained. GP $Y \in [0, 1]$ with Matérn 5/2 kernel and parameters $\boldsymbol{\theta}_0 = (\sigma_0^2 = 0.2, \ell = 0.2)$.



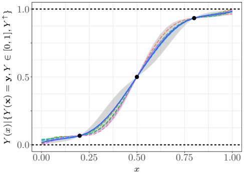
■ limit dist. $\mathcal{N}(0, 2\sigma_0^4)$ ■ **MLE**: $\sqrt{n}\left(\widehat{\sigma}_n^2 - \sigma_0^2\right)$ ■ **cMLE**: $\sqrt{n}\left(\widehat{\sigma}_{n,c}^2 - \sigma_0^2\right)$

✦ For large values of n , **MLE** and **cMLE** provide similar performances.

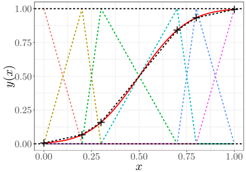
Table of Contents

Conclusions

- We further investigated (Maatouk and Bay, 2017) to account for sets of linear inequality constraints.



- We explored the use of MCMC (i.e. HMC) for the posterior approximation.

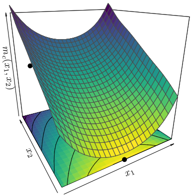


- We proposed variations of the model in high dimensions:
  - Noisy data,
  - “Optimal” rectangular designs,
  - Additive functions.

- We studied a constrained likelihood analysing their consistency and asymptotic normality.

$$\mathcal{L}_{n,c}(\boldsymbol{\theta}) = \mathcal{L}_n(\boldsymbol{\theta}) + \log p_{\boldsymbol{\theta}}(Y \in \mathcal{E}|\mathbf{Y}_n) - \log p_{\boldsymbol{\theta}}(Y \in \mathcal{E}).$$

  - **cMLE** is suggested for small number of obs n.
  - **MLE** can also be used when n is large.



Papers, technical reports & R packages

- ✦ A.F. López-Lopera, F. Bachoc, N. Durrande, and O. Roustant (2018). *Finite-dimensional Gaussian approximation with linear inequality constraints.* SIAM/ASA Journal on Uncertainty Quantification, 6(3): 1224–1255.
- – A.F. López-Lopera, F. Bachoc, N. Durrande, and O. Roustant (+2019). *Approximating Gaussian process emulators with linear inequality constraints and noisy observations via MC and MCMC.* (submitted).
- – F. Bachoc, A. Lagnoux, and A.F. López-Lopera (+2018). *Maximum likelihood estimation for Gaussian processes under inequality constraints* (submitted).
- ✦ A.F. López-Lopera. *LineqGPR: Gaussian process regression models with linear inequality constraints*, 2018. CRAN.

Working papers

- A.F. López-Lopera, F. Bachoc, N. Durrande, and O. Roustant (2019+). *Additive Gaussian processes under linear inequality constraints.*
- — (2019+). *Optimal grid construction for finite-dimensional Gaussian approximations under inequality constraints.*
- — (2019+). *Finite element methods for Gaussian processes under inequality constraints.*

Table of Contents

Bachoc, F., Lagnoux, A., and López-Lopera, A. F. (2018). Maximum likelihood estimation for Gaussian processes under inequality constraints. *ArXiv e-prints*.

Bay, X., Grammont, L., and Maatouk, H. (2016). Generalization of the Kimeldorf-Wahba correspondence for constrained interpolation. *Electronic journal of statistics*, 10(1):1580–1595.

Botev, Z. I. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):125–148.

Durrande, N., Ginsbourger, D., and Roustant, O. (2012). Additive covariance kernels for high-dimensional Gaussian process modeling. *Annales de la Faculté de Sciences de Toulouse*, 21(3):481–499.

López-Lopera, A. F., Bachoc, F., Durrande, N., and Roustant, O. (2018). Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1224–1255.

López-Lopera, A. F., Bachoc, F., Durrande, N., and Roustant, O. (2019). Approximating Gaussian Process Emulators with Linear Inequality Constraints and Noisy Observations via MC and MCMC. *arXiv e-prints*.

Maatouk, H. and Bay, X. (2016). *A New Rejection Sampling Method for Truncated Multivariate Gaussian Random Variables Restricted to Convex Sets*, pages 521–530. Springer International Publishing, Cham.

Maatouk, H. and Bay, X. (2017). Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, 49(5):557–582.

Pakman, A. and Paninski, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer New York.

Taylor, J. and Benjamini, Y. (2017). RestrictedMVN: multivariate normal restricted by affine constraints. `https://cran.r-project.org/web/packages/restrictedMVN/index.html`. [Online; 02-Feb-2017].