Gaussian processes under inequality constraints: Model selection and extension to high dimensions

Andrés F. López-Lopera and Mathis Deronzier

July 8, 2025

Workshop ANR JCJC GAP

Joint work with...

F. Bachoc
IMT, Toulouse

N. Durrande
Monumo, UK

O. Roustant
IMT, Toulouse

A. Lagnoux
IMT, Toulouse

M. Deronzier
IMT, Toulouse

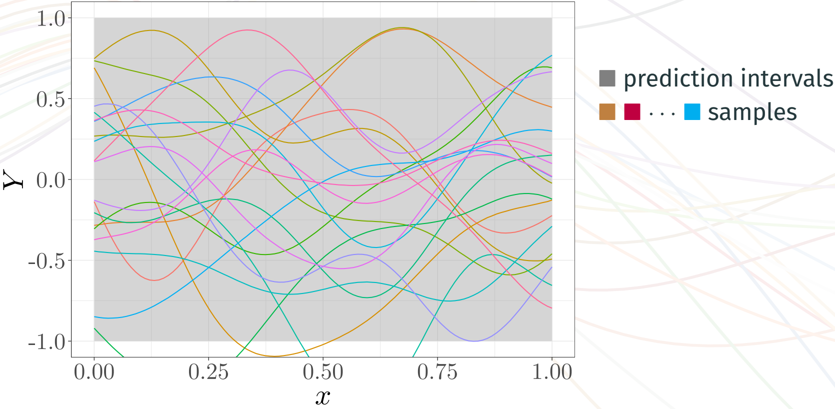
Ti John
Aalto University

Table of contents

Constrained Gaussian processes

Gaussian processes (GPs)

GPs form a flexible **prior over functions** [Rasmussen and Williams, 2005]:



- Let $\{Y(x); x \in \mathcal{D}\}$ be a stochastic process defined on a compact input space $\mathcal{D} \subseteq \mathbb{R}$ (e.g., $\mathcal{D} = [0, 1]$).
- Y is GP-distributed if, for all $x_1, \ldots, x_n \in \mathcal{D}$,

$$\mathbf{Y}_n := \left[Y(x_1), \ \cdots, \ Y(x_n)\right]^\top \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}).$$

with mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$.

- By convention, we denote the GP Y as

$$Y \sim \mathcal{GP}(\mu, k),$$

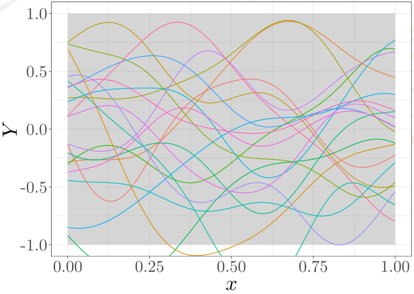
with mean function $\mu(x) = \mathbb{E}(Y(x))$ and covariance function (or kernel) $k(x, x') = \text{cov}(Y(x), Y(x')) = \mathbb{E}([Y(x) - \mu(x)][Y(x') - \mu(x')])$, for $x, x' \in \mathcal{D}$.

Gaussian processes (GPs)

· In practice, centered GPs priors Y are considered (i.e., $\mu(\cdot) = 0$). Then, Y is fully defined by its kernel function k.

Squared Exponential kernel: $$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell}\right),$$

where $\sigma^2 > 0$ and $\ell > 0$ are the variance and length-scale parameters.



prediction intervals

... samples

[link]

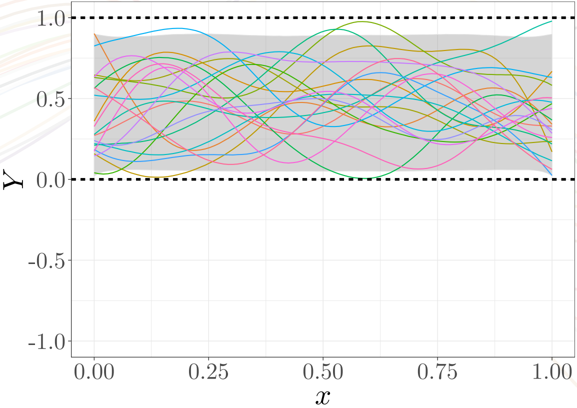
Gaussian processes (GPs)

- Let $(\mathbf{X}, \boldsymbol{y}) = (x_i, y_i)_{1 \leq i \leq n}$ a training dataset.
- Then, the conditional distribution $Y(x^*)|(\mathbf{Y}_n + \varepsilon = \boldsymbol{y})$, with $\varepsilon \sim \mathcal{N}\left(\mathbf{0}, \tau^2 \boldsymbol{I}_n\right)$, is Gaussian with mean and variance parameters:

$$\mu(x^*) = k(x^*, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \tau^2 \boldsymbol{I}_n)^{-1} \boldsymbol{y},$$
$$v(x^*) = k(x^*, x^*) - k(x^*, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \tau^2 \boldsymbol{I}_n)^{-1} k(\mathbf{X}, x^*),$$

where $k(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{n \times n}$ and $k(x^*, \mathbf{X}) = (k(\mathbf{X}, x^*))^\top \in \mathbb{R}^n$.

Our interest: GP-based priors satisfying some inequality constraints...



prediction intervals

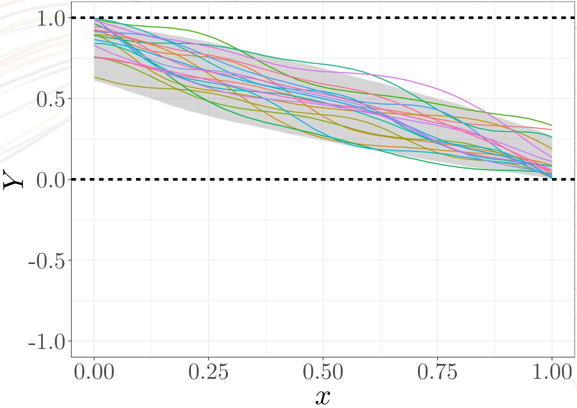
... samples

- - Boundedness condition

$Y \in [0, 1]$

Constrained GPs

Our interest: GP-based priors satisfying some inequality constraints...



■ prediction intervals

■ ■ ··· ■ samples

- - Boundedness condition

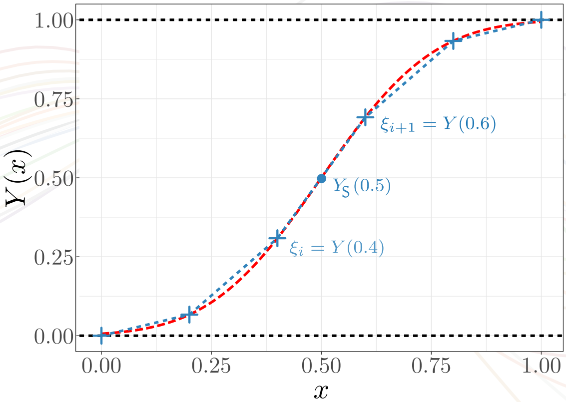
$$Y \in [0, 1]$$

↓ Monotonicity condition

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

CERAMATHS | Université Polytechnique HAUTS-DE-FRANCE

Finite-dimensional approximation of GPs



■ smooth function Y

■ piecewise approximation Y_S

Note that:

- If $\xi_j \in [0, 1]$ for $j = 1, \ldots, m$,

$$Y_S(0.5) \in [0, 1].$$

- Or if $\xi_j < \xi_{j+1}$ for $j = 1, \ldots, m-1$,

$$\xi_j < Y_S(0.5) < \xi_{j+1}.$$

Pro: imposing constraints over knots is enough [Maatouk and Bay, 2017]:

$$Y_S \in \mathcal{E} \Leftrightarrow \boldsymbol{\xi} \in \mathcal{C}.$$

Finite-dimensional approximation of GPs

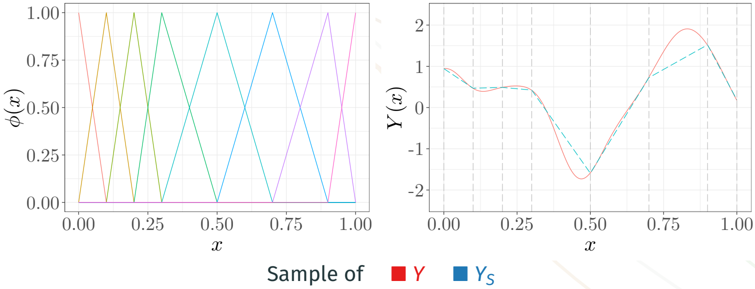
- Let Y_S be the finite-dimensional GP with an ordered set of knots:

$$S = \{t_0, \ldots, t_m\}, \quad \text{with} \quad 0 = t_0 < \cdots < t_m = 1,$$

such that

$$Y_S(x) = \sum_{j=1}^{m} Y(t_j)\phi_j(x), \tag{1}$$

where $x \in [0, 1]$, $Y \sim \mathcal{GP}(0, k_\theta)$, and $\phi_j : [0, 1] \mapsto \mathbb{R}$ are (asymmetric) hat basis functions.



CERAMATHS
Université Polytechnique HAUTS-DE-FRANCE

Finite-dimensional approximation of GPs

- Then, for regression tasks under inequality constraints, we have

$$Y_S(x) = \sum_{j=1}^{m} \xi_j \phi_j(x), \text{ s.t. } \begin{cases} Y_S(x_i) + \varepsilon_i = y_i & \text{(regression conditions)}, \\ \boldsymbol{l} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u} & \text{(linear inequality conditions)}, \end{cases} \quad (2)$$

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

- $\xi_j := Y(t_j)$ for $j = 1, \ldots, m$, i.e., $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_m]^\top \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\theta)$ with covariance matrix $\boldsymbol{\Sigma}_\theta = (k_\theta(t_j, t_{j'}))_{1 \leq j, j' \leq m}$
- $\varepsilon_i \sim \mathcal{N}(0, \tau^2)$, for $i = 1, \ldots, n$, with noise variance τ^2
- $(\boldsymbol{\Lambda}, \boldsymbol{l}, \boldsymbol{u})$ define the ineq. constraints. For instance, for the case of monotonicity, we have

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

· Since $Y \in \mathcal{E} \Leftrightarrow \boldsymbol{\xi} \in \mathcal{C}$, then *uncertainty quantification* relies on simulating the **truncated vector** $\boldsymbol{\xi}$ [López-Lopera et al., 2018]:

$$\boldsymbol{\Lambda}\boldsymbol{\xi} | \{\boldsymbol{\Phi}\boldsymbol{\xi} + \boldsymbol{\varepsilon} = \boldsymbol{y}, \boldsymbol{l} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\} \sim \mathcal{TN}(\boldsymbol{\Lambda}\boldsymbol{\mu}_c, \boldsymbol{\Lambda}\boldsymbol{\Sigma}_c\boldsymbol{\Lambda}^\top, \boldsymbol{l}, \boldsymbol{u}), \tag{3}$$

with conditional parameters $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ given by

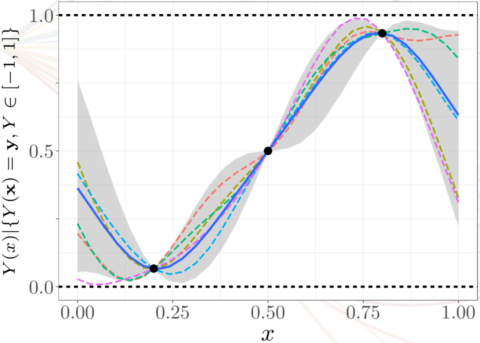
$$\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top + \tau^2\boldsymbol{I}, \quad \boldsymbol{\mu}_c = \boldsymbol{\Sigma}\boldsymbol{\Phi}^\top\mathbf{K}^{-1}\boldsymbol{y}, \quad \boldsymbol{\Sigma}_c = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\Phi}^\top\mathbf{K}^{-1}\boldsymbol{\Phi}\boldsymbol{\Sigma}. \tag{4}$$

$\ast$ Eq. (3) is computed via *Monte Carlo* (MC) or *Markov Chain MC* (MCMC):

- e.g., *Hamiltonian Monte Carlo* (HMC) [Pakman and Paninski, 2014]

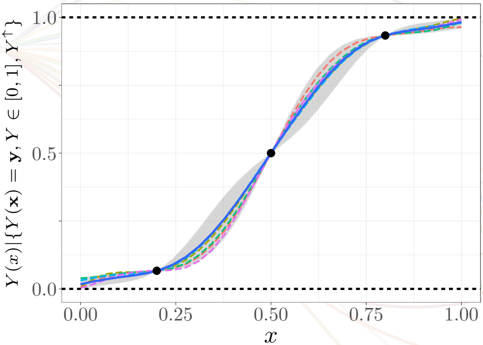
- A. López-Lopera, N. Durrande, F. Bachoc and O. Roustant, Finite-dimensional Gaussian approximation with linear inequality constraints, SIAM/ASA J. on Uncertainty Quantification, 2018.

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



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

The maximum a posteriori (mode) function in 1D

· Let $\widehat{\boldsymbol{\xi}}$ be the mode that maximises the pdf of $\boldsymbol{\xi} | \{\boldsymbol{\Phi}\boldsymbol{\xi} + \varepsilon = \boldsymbol{y}, \boldsymbol{l} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\}$:

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

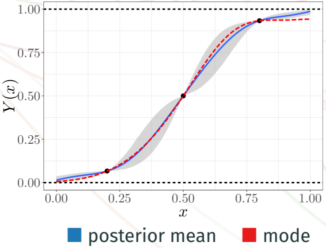
with $\widehat{\boldsymbol{\xi}} = [\widehat{\xi}_1, \ldots, \widehat{\xi}_m]^\top$.

· The *MAP estimate* of Y_S is given by

$$\widehat{Y}_S(x) = \sum_{j=1}^{m} \widehat{\xi}_j \phi_j(x). \tag{6}$$

Pro:

- $\widehat{Y}_S$ can be used as a point estimate
- Easy and fast calculations
- Starting point for MCMC
- Convergence to the spline solution as $m \to \infty$ [Bay et al., 2016]



CERAMATHS
Université Polytechnique HAUTS-DE-FRANCE

Correspondence between Bayes' estimation and optimal interpolation

· The usual Bayesian estimator of y is the mean of the posterior distribution of the $Y \sim \mathcal{GP}(0, k)$ given data $(\mathbf{X}, \boldsymbol{y}) = (x_i, y_i)_{1 \leq i \leq n}$:

$$\mu(x) = \mathbb{E}\left(Y(x) | \mathbf{Y}_n = \boldsymbol{y}\right)$$

· The estimator μ is the unique solution of the optimization problem [Kimeldorf and Wahba, 1970]:

$$\min_{h \in \mathcal{H} \cap I} ||h||^2_{\mathcal{H}},$$

where $\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS) associated to the kernel k, and I is the set of interpolant functions:

$$I := \{f : \mathcal{D} \to \mathbb{R} : f(x_i) = y_i, i = 1, \ldots, n\}.$$

· Bay et al. [2016] have shown that the mode that maximises the pdf of $Y(x) | \{\mathbf{Y}_n = \boldsymbol{y}, Y \in \mathcal{E}\}$ is the unique solution of the constrained opt. problem:

$$\min_{h \in \mathcal{H} \cap \mathcal{E} \cap I} ||h||^2_{\mathcal{H}}.$$

CERAMATHS | Université Polytechnique HAUTS-DE-FRANCE

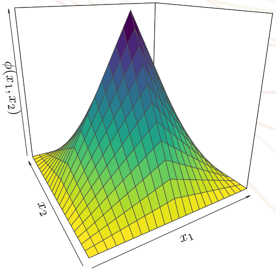
Constrained Gaussian processes: Extension to d dimensions

- The extension to d dimensions is obtained by **tensorization**:

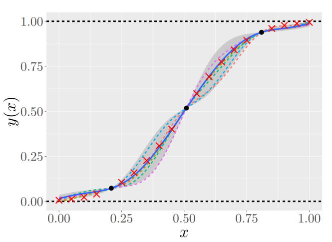
$$Y_S(\mathbf{x}) = \sum_{j_1,\ldots,j_d=1}^{m_1,\ldots,m_d} \left[\prod_{p=1,\ldots,d} \phi_{j_p}^{(p)}(x_p) \right] \xi_{j_1,\cdots,j_d}, \quad \text{s.t.} \begin{cases} Y_m(\mathbf{x}_i) + \varepsilon_i = y_i, \\ \boldsymbol{\xi} \in \mathcal{C}, \end{cases} \tag{7}$$

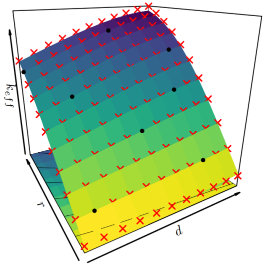
where $\mathbf{x}_i \in [0,1]^d$, $y_i \in \mathbb{R}$, $\varepsilon_i \sim \mathcal{N}(0, \tau^2)$, for $i = 1, \ldots, n$; and

- $\boldsymbol{\xi} = [\xi_{1,\ldots,1}, \ldots, \xi_{m_1,\ldots,m_d}]^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_\theta)$,
- $\mathcal{C}$ is a convex set of linear inequality constraints, and
- $\phi_{j_i}^{(i)} : [0,1] \mapsto \mathbb{R}$ are hat basis functions.



Risk assessment: nuclear safety and coastal flooding



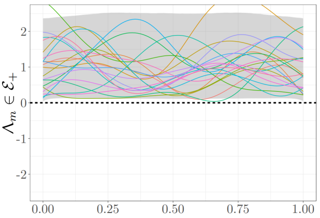


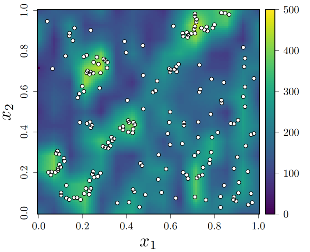
A. F. López-Lopera, N. Durrande, F. Bachoc and O. Roustant (2018). Finite-dimensional Gaussian approximation with linear inequality constraints. SIAM/ASA Journal on Uncertainty Quantification, 6(3).

A. F. López-Lopera, F. Bachoc, N. Durrande, J. Rohmer, D. Idier, and O. Roustant (2019). Approximating Gaussian process emulators with linear inequality constraints and noisy observations via MC and MCMC. In International Conference in Monte Carlo & Quasi-Monte Carlo Methods, Springer Proceedings in Mathematics & Statistics.

Constrained Gaussian processes: Application in spatial modeling

Geostatistics: Spatial location of redwood trees





· We considered Cox processes with a (non-negative) GP-distributed stochastic intensity function:

$$\Lambda_m(x) = \sum_{j=1}^{m} \phi_j(x)\xi_j \quad \text{s.t.} \quad \Lambda_m \in \mathcal{E}_+$$

A. F. López-Lopera, S. John and N. Durrande (2019). Gaussian process modulated Cox processes under linear inequality constraints. International Conference on Artificial Intelligence and Statistics (AISTATS).

Maximum likelihood estimation under constraints

- Consider $\{k_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, with $\boldsymbol{\Theta} \subset \mathbb{R}^p$, a parametric family of covariance functions where $\boldsymbol{\theta}$ defines the covariance parameters
- The maximum likelihood estimator, with log-likelihood function $\mathcal{L}_n(\boldsymbol{\theta}) := \log p_{\boldsymbol{\theta}}(\mathbf{Y}_n)$, is given by

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} \in \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ \mathcal{L}_n(\boldsymbol{\theta}).$$

Maximum likelihood estimation under constraints

- Consider $\{k_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, with $\boldsymbol{\Theta} \subset \mathbb{R}^p$, a parametric family of covariance functions where $\boldsymbol{\theta}$ defines the covariance parameters

- The maximum likelihood estimator, with log-likelihood function $\mathcal{L}_n(\boldsymbol{\theta}) := \log p_{\boldsymbol{\theta}}(\mathbf{Y}_n)$, is given by

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} \in \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ \mathcal{L}_n(\boldsymbol{\theta}).$$

- To account for inequality constraints, we can consider the conditional log-likelihood function

$$\begin{aligned} \mathcal{L}_{n,\mathcal{C}}(\boldsymbol{\theta}) &= \log p_{\boldsymbol{\theta}}(\mathbf{Y}_n | \boldsymbol{\xi} \in \mathcal{C}) \\ &= \log p_{\boldsymbol{\theta}}(\mathbf{Y}_n) + \log P_{\boldsymbol{\theta}}(\boldsymbol{\xi} \in \mathcal{C} | \boldsymbol{\Phi}\boldsymbol{\xi} = \mathbf{Y}_n) - \log P_{\boldsymbol{\theta}}(\boldsymbol{\xi} \in \mathcal{C}) \end{aligned}$$

Then, the constrained estimator is given by

$$\widehat{\boldsymbol{\theta}}_{\text{cMLE}} \in \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ \mathcal{L}_{n,\mathcal{C}}(\boldsymbol{\theta}),$$

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

F. Bachoc, A. Lagnoux and **A. F. López-Lopera** (2019). Maximum likelihood estimation for Gaussian processes under inequality constraints. Electronic Journal of Statistics, 13(2).

Asymptotic consistency of the MLE & cMLE

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

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

- Denote: $\boldsymbol{\theta}_0$ (true covariance parameters), $\widehat{\boldsymbol{\theta}}_n$ (MLE), $\widehat{\boldsymbol{\theta}}_{n,\mathcal{C}}$ (cMLE).

Proposition (Consistency of the MLE and cMLE)

Assume $\forall \varepsilon > 0$ *and* $\forall M < \infty$,

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

Then,

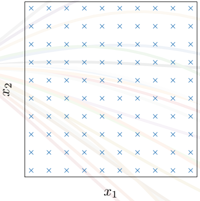
$$P\left(\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \geq \varepsilon} (\mathcal{L}_{n,\mathcal{C}}(\boldsymbol{\theta}) - \mathcal{L}_{n,\mathcal{C}}(\boldsymbol{\theta}_0)) \geq -M \mid Y \in \mathcal{E}_\kappa\right) \xrightarrow[n\to\infty]{} 0.$$

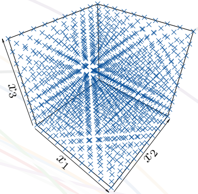
Consequently, both the MLE and cMLE are consistent estimators:

$$\widehat{\boldsymbol{\theta}}_n \in \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_n(\boldsymbol{\theta}) \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_0, \quad \widehat{\boldsymbol{\theta}}_{n,\mathcal{C}} \in \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{n,\mathcal{C}}(\boldsymbol{\theta}) \xrightarrow[n\to\infty]{P|Y\in\mathcal{E}_\kappa} \boldsymbol{\theta}_0.$$

Constrained Gaussian processes

- **Con:** the cost of Y_S increases as d increases.





- This drawback can be mitigated by considering:
  - a *“smarter” construction of rectangular grids* of knots thanks to the asymmetric construction of the hat basis functions
  - and/or *further assumptions for complexity simplification*
    → e.g., *inactive variables, additive structures*

F. Bachoc, **A. F. López-Lopera**, and O. Roustant (2022). Sequential construction and dimension reduction of Gaussian processes under inequality constraints. SIAM Journal on Mathematics of Data Science, 4(2).

A. F. López-Lopera, F. Bachoc, and O. Roustant (2022). High-dimensional additive Gaussian processes under monotonicity constraints. In Advances in Neural Information Processing Systems (NeurIPS), volume 35.

The MaxMod algorithm

The MaxMod algorithm in 1D

- Let $\widehat{Y}_S$ be the MAP function with an ordered set of knots:

$$S = \{t_0, \ldots, t_m\}, \quad \text{with} \quad 0 = t_0 < \cdots < t_m = 1.$$

- Here, we aim at adding a new knot t in S (where?)
- To do so, we aim at *maximising the total modification of the MAP*:

$$I_S(t) = \int_{[0,1]} \left(\widehat{Y}_{S \cup t}(x) - \widehat{Y}_S(x) \right)^2 dx. \tag{8}$$

- The integral in (8) has a closed-form expression.

- Let $\widehat{Y}_S$ be the MAP function with an ordered set of knots:

$$S = \{t_0, \dots, t_m\}, \quad \text{with} \quad 0 = t_0 < \cdots < t_m = 1.$$

- Here, we aim at adding a new knot t in S (where?)
- To do so, we aim at *maximising the total modification of the MAP*:

$$I_S(t) = \int_{[0,1]} \left(\widehat{Y}_{S \cup t}(x) - \widehat{Y}_S(x)\right)^2 dx. \tag{8}$$

- The integral in (8) has a closed-form expression.

Algorithm MaxMod (maximum modification of the MAP) in 1D

Input parameters: the initial subdivision $S^{(0)} \in \mathcal{S}$.

Sequential procedure: for $\kappa \in \mathbb{N}$, do:

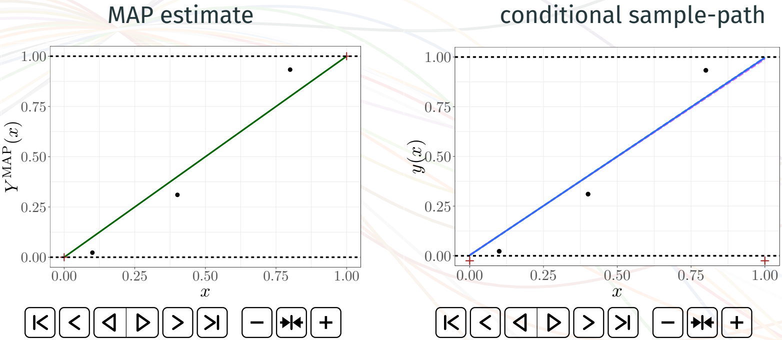
1: Set $t^\star_{\kappa+1} \in [0, 1]$ such that

$$I_{S^{(\kappa)}}(t^\star_{\kappa+1}) \geq \sup_{t \in [0,1]} I_{S^{(\kappa)}}(t)$$

2: $S^{(\kappa+1)} = S^{(\kappa)} \cup t^\star_{\kappa+1}$.

The MaxMod algorithm in 1D

1D example under boundedness and monotonicity constraints



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

CERAMATHS Université Polytechnique HAUTS-DE-FRANCE

The MaxMod algorithm in higher dimensions

- Let $\widehat{Y}_{\mathcal{J},S}$ be the MAP function with $|\mathcal{J}|$ active variables and ordered sets of knots $S_{\mathcal{J}}$ for $\mathcal{J} \subseteq \{1, \ldots, D\}$.
- Then, the criterion to maximise is given by

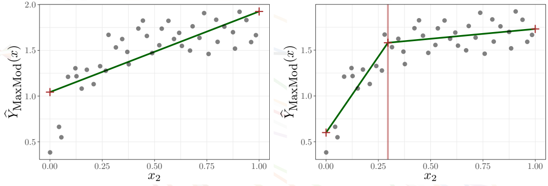
$$I_{\mathcal{J},S}(i,t) = \begin{cases} \frac{1}{N_{S,\mathcal{J},i}} \int_{[0,1]^d} \left(\widehat{Y}_{\mathcal{J},S \cup_i t}(\mathbf{x}) - \widehat{Y}_{\mathcal{J},S}(\mathbf{x})\right)^2 d\mathbf{x} & \text{if } i \in \mathcal{J}, \\ \frac{1}{N_{S,\mathcal{J},i}} \int_{[0,1]^{d+1}} \left(\widehat{Y}_{\mathcal{J}\cup\{i\},S+i}(\mathbf{x}) - \widehat{Y}_{\mathcal{J},S}(\mathbf{x})\right)^2 d\mathbf{x} & \text{if } i \notin \mathcal{J}, \end{cases} \quad (9)$$

where $N_{S,\mathcal{J},i}$ is the increase of the number of basis functions.

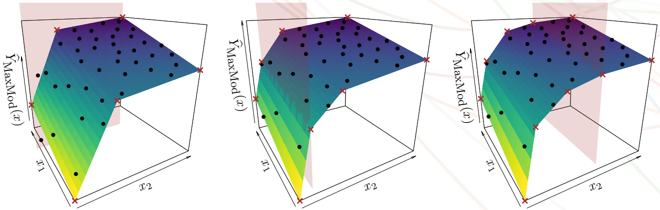
- F. Bachoc, A. López-Lopera, and O. Roustant. Sequential construction and dimension reduction of GPs under inequality constraints. SIAM J. on Maths. of Data Science, 2022.

2D example under monotonicity constraints

Evolution of the MaxMod algorithm using $f(x) = \frac{1}{2}x_1 + \arctan(10x_2)$



(a) iteration 0 (b) iteration 1

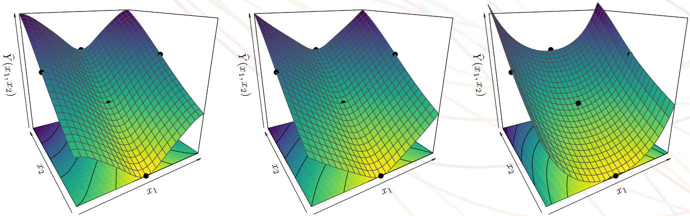


(c) iteration 2 (d) iteration 3 (e) iteration 4

• training points + knots ■ MAP estimate

The MaxMod algorithm in higher dimensions

- The constrained GP is tractable depending on $|\mathcal{J}|$ (nb of active variable).
- According to numerical tests, our framework is limited to $|\mathcal{J}| \leq 5$.
- Therefore, further assumptions are required to scale the model:
  - e.g., additive structures



Additive GP predictions using (left) the unconstrained GP mean, (center) the cGP mode and (right) the cGP mean via HMC. The constrained model accounts for both componentwise convexity and monotonicity conditions along x_1 and x_2, respectively.

CERAMATHS | Université Polytechnique HAUTS-DE-FRANCE

Extension to additive functions

- In high dimension, many statistical regression models are based on additive structures of the form:

$$y(\boldsymbol{x}) = y_1(x_1) + \cdots + y_d(x_d). \tag{10}$$

- Then GP priors can be placed over $y_1, \ldots, y_d$ [Durrande et al., 2012]

$$Y_i \sim \mathcal{GP}(0, k_i),$$

for $i = 1, \ldots, d$. Taking $Y_1, \ldots, Y_d$ as independent GPs, the process

$$Y(\boldsymbol{x}) = Y_1(x_1) + \cdots + Y_d(x_d)$$

is also a GP and its kernel is given by

$$k(\boldsymbol{x}, \boldsymbol{x}') = k_1(x_1, x_1') + \cdots + k_d(x_d, x_d'). \tag{11}$$

Finite-dimensional approximation of additive GPs

- For the constrained case, we can approximate Y_i by a finite-dimensional GP:

$$Y_{i,S_i}(x_i) = \sum_{j=1}^{m_i} \xi_{i,j}\phi_{i,j}(x_i),$$

with one-dimensional subdivision S_i, and m_i knots.

- We let $S = (S_1, \ldots, S_d)$. The finite-dimensional GP is written,

$$Y_S(\boldsymbol{x}) = \sum_{i=1}^{d} Y_{i,S_i}(x_i) = \sum_{i=1}^{d}\sum_{j=1}^{m_i} \xi_{i,j}\phi_{i,j}(x_i), \tag{12}$$

where $\xi_{i,j} = Y_i(t_{(j)}^{(S_i)})$ and $\phi_{i,j} : [0, 1] \mapsto \mathbb{R}$ are asymmetric hat basis functions.

- One can note that the total number of knots is given by $m = m_1 + \cdots + m_d$.

Finite-dimensional approximation of additive GPs

- Observe from (12) that, since $\xi_{i,j}$, for $i = 1, \dots, d$ and $j = 1, \dots, m_i$, are Gaussian distributed, then Y_{i,S_i} is a GP with kernel given by

$$\widetilde{k}_i(x_i, x_i') = \sum_{j=1}^{m_i} \sum_{\kappa=1}^{m_i} \phi_{i,j}(x_i)\phi_{i,\kappa}(x_i') k_i(t_{(j)}^{(S_i)}, t_{(\kappa)}^{(S_i)}). \tag{13}$$

Moreover, Y_S is a GP with kernel $\widetilde{k}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{d} \widetilde{k}_i(x_i, x_i')$.

- We let $\boldsymbol{\Sigma}_i = k_i(S_i, S_i)$ be the $m_i \times m_i$ covariance matrix of ξ_i.

Finite-dimensional approximation of additive GPs

- We consider the componentwise constraints $Y_{i,S_i} \in \mathcal{E}_i$, $i = 1, \ldots, d$ such that

$$Y_{i,S_i} \in \mathcal{E}_i \;\Leftrightarrow\; \boldsymbol{\xi}_i \in \mathcal{C}_i \tag{14}$$

where $\boldsymbol{\xi}_i = [\xi_{i,1}, \cdots, \xi_{i,m_i}]^\top$ and $\mathcal{C}_i = \{\boldsymbol{c} \in \mathbb{R}^{m_i} : \boldsymbol{l}_i \leq \boldsymbol{\Lambda}_i \boldsymbol{c} \leq \boldsymbol{u}_i\}$.

- Examples of constraints are monotonicity and componentwise convexity.
- Given the observations and the constraints, the MAP estimate is given by

$$\widehat{Y}_S(\boldsymbol{x}) = \sum_{i=1}^{d} \sum_{j=1}^{m_i} \widehat{\xi}_{i,j} \phi_{i,j}(x_i). \tag{15}$$

Finite-dimensional approximation of additive GPs

· As in (5), the vector $\widehat{\boldsymbol{\xi}} = [\widehat{\boldsymbol{\xi}}_1^\top, \ldots, \widehat{\boldsymbol{\xi}}_d^\top]^\top$ with $\widehat{\boldsymbol{\xi}}_i = [\widehat{\xi}_{i,1}, \ldots, \widehat{\xi}_{i,m_i}]^\top$ is given by

$$\widehat{\boldsymbol{\xi}} = \underset{\substack{\boldsymbol{\xi}=(\boldsymbol{\xi}_1^\top,\ldots,\boldsymbol{\xi}_d^\top)^\top \\ \boldsymbol{l}_i \leq \boldsymbol{\Lambda}_i \boldsymbol{\xi}_i \leq \boldsymbol{u}_i, i=1,\ldots,d}}{\operatorname{argmin}} (\boldsymbol{\xi} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{\xi} - \boldsymbol{\mu}_c), \tag{16}$$

where $\boldsymbol{\mu}_c = [\boldsymbol{\mu}_{c,1}^\top, \ldots, \boldsymbol{\mu}_{c,d}^\top]^\top$ is the $m \times 1$ vector with block i given by

$$\boldsymbol{\mu}_{c,i} = \boldsymbol{\Sigma}_i \boldsymbol{\Phi}_i^\top \left[\left(\sum_{p=1}^{d} \boldsymbol{\Phi}_p \boldsymbol{\Sigma}_p \boldsymbol{\Phi}_p^\top \right) + \tau^2 \boldsymbol{I}_n \right]^{-1} \boldsymbol{y}_n, \tag{17}$$

and $(\boldsymbol{\Sigma}_{c,i,j})_{i,j}$ is the $m \times m$ matrix with block (i,j) given by

$$\boldsymbol{\Sigma}_{c,i,j} = \mathbf{1}_{i=j} \boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_i \boldsymbol{\Phi}_i^\top \left[\left(\sum_{p=1}^{d} \boldsymbol{\Phi}_p \boldsymbol{\Sigma}_p \boldsymbol{\Phi}_p^\top \right) + \tau^2 \boldsymbol{I}_n \right]^{-1} \boldsymbol{\Phi}_j \boldsymbol{\Sigma}_j. \tag{18}$$

Remarks:

- $\boldsymbol{\Sigma}_{c,i,j}$ involves contributions of the cross-covariances.
- The inversion is computed efficiently for $m << n$ (matrix inv. lemma).

Additive MaxMod algorithm

- Consider an additive cGP model that uses only a subset $\mathcal{J} \subseteq \{1, \dots, d\}$ of active variables.
- Its mode function $\widehat{Y}_S$, from $\mathbb{R}^{|\mathcal{J}|}$ to $\mathbb{R}$, by, for $\boldsymbol{x} = (x_i; i \in \mathcal{J})$,

$$\widehat{Y}_S(\boldsymbol{x}) = \sum_{i \in \mathcal{J}} \sum_{j=1}^{m_i} \widehat{\xi}_{i,j} \phi_{i,j}(x_i). \quad (19)$$

- We measure this benefit by the squared-norm modification of the cGP mode

$$I_{S,i^\star} = \int_{[0,1]^{|\mathcal{J}|+1}} \left(\widehat{Y}_{S,i^\star}(\boldsymbol{x}) - \widehat{Y}_S(\boldsymbol{x}) \right)^2 d\boldsymbol{x} \ \text{ for } i^\star \notin \mathcal{J}, \quad (20)$$

$$I_{S,i^\star,t} = \int_{[0,1]^{|\mathcal{J}|}} \left(\widehat{Y}_{S,i^\star,t}(\boldsymbol{x}) - \widehat{Y}_S(\boldsymbol{x}) \right)^2 d\boldsymbol{x} \ \text{ for } i^\star \in \mathcal{J}. \quad (21)$$

- Both (20) and (21) have analytic expression assuming $x_i \sim \text{Uniform}(0, 1)$ for $i = 1, \dots, d$ (see López-Lopera et al. [2022]), where the computational cost is linear with respect to $m = \sum_{i \in \mathcal{J}} m_i$.

CERAMATHS | Université Polytechnique HAUTS-DE-FRANCE

Extension to additive functions

- For a new variable $i^\star \notin \mathcal{J}$, the new mode function is

$$\widehat{Y}_{S,i^\star}(\boldsymbol{x}) = \sum_{i\in\mathcal{J}} \sum_{j=1}^{m_i} \widetilde{\xi}_{i,j}\phi_{i,j}(x_i) + \sum_{j=1}^{2} \widetilde{\xi}_{i^\star,j}\phi_{i^\star,j}(x_{i^\star})$$

- We let $\phi_{i^\star,1}(u) = 1 - u$ and $\phi_{i^\star,2}(u) = u$ for $u \in [0,1]$.

Proposition (Computation of $I_{S,i^\star}$)

We have

$$I_{S,i^\star} = \sum_{i\in\mathcal{J}} \sum_{\substack{j,j'=1 \\ |j-j'|\le 1}}^{m_i} \eta_{i,j}\eta_{i,j'} E_{j,j'}^{(S_i)} - \sum_{i\in\mathcal{J}} \left(\sum_{j=1}^{m_i} \eta_{i,j} E_j^{(S_i)} \right)^2 + \frac{\eta_{i^\star}^2}{12} + \left(\sum_{i\in\mathcal{J}} \sum_{j=1}^{m_i} \eta_{i,j} E_j^{(S_i)} - \frac{\zeta_{i^\star}}{2} \right)^2,$$

where $\eta_{i,j} = \widehat{\xi}_{i,j} - \widetilde{\xi}_{i,j}$, $\eta_{i^\star} = \widetilde{\xi}_{i^\star,2} - \widetilde{\xi}_{i^\star,1}$, $\zeta_{i^\star} = \widetilde{\xi}_{i^\star,1} + \widetilde{\xi}_{i^\star,2}$, $E_j^{(S_i)} := \int_0^1 \phi_{i,j}(t)dt$ *and* $E_{j,j'}^{(S_i)} := \int_0^1 \phi_{i,j}(t)\phi_{i,j'}(t)dt$ *with explicit expressions in Lemma 1 [López-Lopera et al., 2022, Appendix A.3]. The matrices* $(E_{j,j'}^{(S_i)})_{1\le j,j'\le m_i}$ *are 1-band and the computational cost is linear w.r.t.* $m = \sum_{i\in\mathcal{J}} m_i$.

Additive MaxMod algorithm

- For a new t added to $S_{i^\star}$ with $i^\star \in \mathcal{J}$, the new mode function is

$$\widehat{Y}_{S,i^\star,t}(\mathbf{x}) = \sum_{i\in\mathcal{J}} \sum_{j=1}^{\widetilde{m}_i} \widetilde{\xi}_{i,j}\widetilde{\phi}_{i,j}(x_i),$$

where $\widetilde{m}_i = m_i$ for $i \neq i^\star$, $\widetilde{m}_{i^\star} = m_{i^\star} + 1$, $\widetilde{\phi}_{i,j} = \phi_{i,j}$ for $i \neq i^\star$, and $\widetilde{\phi}_{i^\star,j}$ is obtained from $S_{i^\star} \cup \{t\}$ as in Proposition 2.

Proposition (Computation of $I_{S,i^\star,t}$)

For $i \in \mathcal{J}\backslash\{i^\star\}$, let $\widetilde{S}_i = S_i$. Let $\widetilde{S}_{i^\star} = S_{i^\star} \cup \{t\}$. Recall that the knots in $S_{i^\star}$ are written $0 = t_{(1)}^{(S_{i^\star})} < \cdots < t_{(m_{i^\star})}^{(S_{i^\star})} = 1$. Let $\nu \in \{1, \dots, m_{i^\star} - 1\}$ be such that $t_{(\nu)}^{(S_{i^\star})} < t < t_{(\nu+1)}^{(S_{i^\star})}$. Then, with a linear cost w.r.t. $\widetilde{m} = \sum_{i\in\mathcal{J}} \widetilde{m}_i$, we have

$$I_{S,i^\star,t} = \sum_{i\in\mathcal{J}} \sum_{\substack{j,j'=1 \\ |j-j'|\le 1}}^{\widetilde{m}_i} \bar{\eta}_{i,j}\bar{\eta}_{i,j'} E_{j,j'}^{(\widetilde{S}_i)} - \sum_{i\in\mathcal{J}} \left(\sum_{j=1}^{\widetilde{m}_i} \bar{\eta}_{i,j} E_j^{(\widetilde{S}_i)} \right)^2 + \left(\sum_{i\in\mathcal{J}} \sum_{j=1}^{\widetilde{m}_i} \bar{\eta}_{i,j} E_j^{(\widetilde{S}_i)} \right)^2,$$

where $\bar{\eta}_{i,j} = \bar{\xi}_{i,j} - \widetilde{\xi}_{i,j}$, $\bar{\xi}_{i,j} = \widehat{\xi}_{i,j}$ for $i \neq i^\star$, $\bar{\xi}_{i^\star,j} = \widehat{\xi}_{i^\star,j}$ for $j \le \nu$, $\bar{\xi}_{i^\star,j} = \widehat{\xi}_{i^\star,j-1}$ for $j \geq \nu + 2$, and

$$\bar{\xi}_{i^\star,\nu+1} = \widehat{\xi}_{i^\star,\nu} \frac{t_{(\nu+1)}^{(S_{i^\star})} - t}{t_{(\nu+1)}^{(S_{i^\star})} - t_{(\nu)}^{(S_{i^\star})}} + \widehat{\xi}_{i^\star,\nu+1} \frac{t - t_{(\nu)}^{(S_{i^\star})}}{t_{(\nu+1)}^{(S_{i^\star})} - t_{(\nu)}^{(S_{i^\star})}}.$$

Numerical experiments

Numerical experiments: Monotonicity in hundreds of dimensions

- We consider the target function:

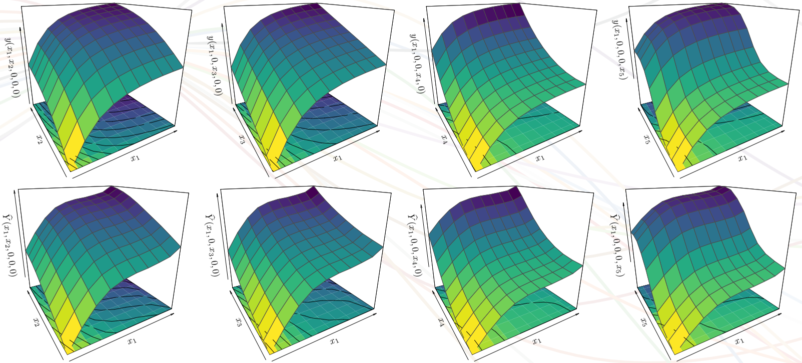
$$y(\boldsymbol{x}) = \sum_{i=1}^{d} \arctan\left(5\left[1 - \frac{i}{d+1}\right]x_i\right). \qquad (22)$$

with $\boldsymbol{x} \in [0,1]^d$. y exhibits decreasing growth rates as the index i increases.

Results (mean $\pm$ one standard deviation over 10 replicates) with $n = 2d$. For the computation of the cGP mean, 10^3 (†50) HMC samples are used.

d	m	CPU Time [s]		Q^2 [%]		
		cGP mode	cGP mean	GP mean	cGP mode	cGP mean
10	50	0.1 ± 0.1	0.1 ± 0.1	82.3 ± 6.2	83.8 ± 4.2	$\mathbf{88.1 \pm 1.7}$
100	500	0.4 ± 0.1	5.2 ± 0.5	89.8 ± 1.6	90.7 ± 1.4	$\mathbf{91.5 \pm 1.3}$
250	1250	4.2 ± 0.7	132.3 ± 26.3	91.7 ± 0.8	92.9 ± 0.6	$\mathbf{93.4 \pm 0.6}$
500	2500	37.0 ± 11.4	$^\dagger 156.9 \pm 40.5$	92.5 ± 0.6	93.8 ± 0.5	$^\dagger \mathbf{94.3 \pm 0.5}$
1000	5000	262.4 ± 35.8	$^\dagger 10454.3 \pm 3399.3$	92.6 ± 0.3	94.6 ± 0.2	$^\dagger \mathbf{95.1 \pm 0.2}$

Numerical experiments: Monotonicity in hundreds of dimensions



2D projections of the true profiles (top) and the constrained GP predictions (bottom)

Numerical experiments: Dimension reduction illustration

- We test the capability of MaxMod to account for dimension reduction considering the function in (22).
- In addition to $(x_1, \dots, x_d)$, we include $D - d$ virtual variables, indexed as $(x_{d+1}, \dots, x_D)$, which will compose the subset of inactive dimensions.
  - $\widehat{Y}_{\text{MaxMod}}$: the mode of the additive cGP and MaxMod.
  - $\widetilde{Y}_{\text{MaxMod}}$: the mode of the non-additive cGP and MaxMod.

Q^2 Performance of the MaxMod algorithm with $n = 10D$.

D	d	active dimensions	knots per dimension	$Q^2(\widetilde{Y}_{\text{MaxMod}})$ [%]	$Q^2(\widehat{Y}_{\text{MaxMod}})$ [%]
10	2	(1, 2)	(4, 3)	99.5	**99.8**
	3	(1, 2, 3)	(5, 5, 3)	97.8	**99.8**
	5	(1, 2, 3, 4, 5)	(4, 4, 4, 3, 2)	91.4	**99.8**
20	2	(1, 2)	(5, 3)	99.7	**99.8**
	3	(1, 2, 3)	(4, 4, 3)	99.0	**99.9**
	5	(1, 2, 3, 4, 5)	(5, 4, 3, 3, 2)	96.0	**99.7**

· The database contains a flood study conducted by the French multinational electric utility company EDF in the Vienne river [Petit et al., 2016].

· It is composed of $N = 2 \times 10^4$ simulations.

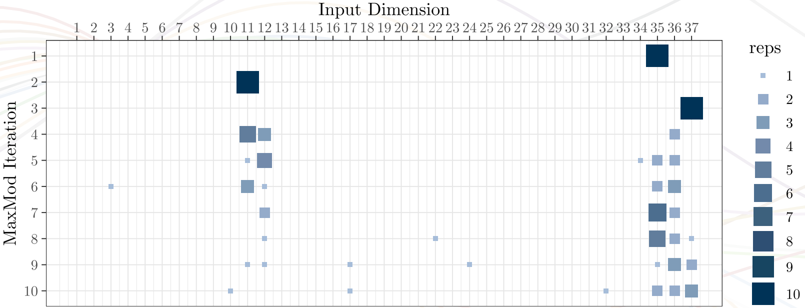
- 1 output: water level H
- 37 inputs depending on: a value of flow upstream, data on the geometry of the bed, and Strickler friction coefficients

· It is possible to identify that H is decreasing along the first 24 input dimensions and increasing along dimension 37.

· Petit et al. [2016] have shown that the additive assumption is realistic here, and that inputs 11, 35 and 37 explain most of the variance.

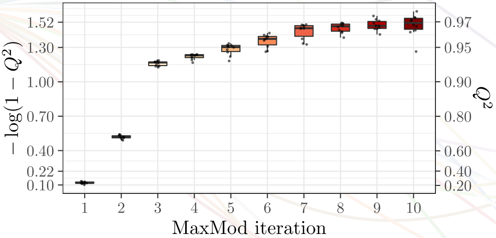
· We consider (approximated) LHD of size $n = 2d$ for training the cGP.

Numerical experiments: Flood study of the Vienne river



The choice made by MaxMod per iteration. Results are computed over 10 replicates. For the first panel, a bigger and darker square implies a more repeated choice.

Numerical experiments: Flood study of the Vienne river



Q^2 boxplots. Results are computed over 10 replicates. For the first panel, a bigger and darker square implies a more repeated choice.

Conclusions

- We combine the additive and constrained frameworks to propose an additive constrained GP prior and MaxMod algorithm.
- The corresponding mode predictor can be computed and posterior realizations can be sampled, both in a scalable way to high dimension.
  - We demonstrate the performance and scalability of the framework with examples with $d \leq 1000$ and in a real-world application with $d = 37$.
  - MaxMod identifies the most important input variables, with data size as low as $n = 2d$ in dimension d.
- Open-source R codes are available:

https://github.com/anfelopera/lineqGPR

Perspectives

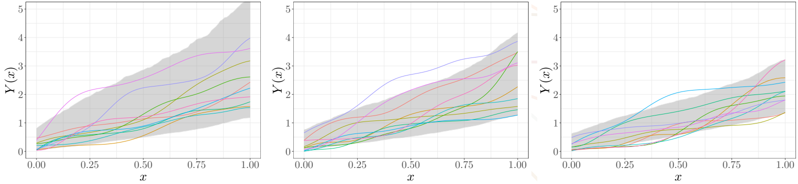
⋆ The extension to block-additivity is being studied by Mathis Deronzier (PhD student at the IMT), for instance for disjoint blocks:

$$Y(x_1, x_2, x_3) = Y(x_1, x_2) + Y(x_3).$$

We seek to study:

- Variable selection (structure of the blocks).
- Further real-world applications.

⋆ The extension to Student-t processes is studied in collaboration with Ari Pakman (Ben-Gurion University).



Samples of (from left to right) zero-mean t-processes with shape parameters $\nu = 4, 10$ and a zero-mean GP. Both processes are reinforced by positivity and monotonicity constraints, and consider SE kernels with $(\sigma^2, \ell) = (1, 0.2)$.

References

X. Bay, L. Grammont, and H. Maatouk. Generalization of the Kimeldorf-Wahba correspondence for constrained interpolation. *Electronic Journal of Statistics*, 2016.

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

George S. Kimeldorf and Grace Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2): 495 – 502, 1970.

A. F. López-Lopera, F. Bachoc, N. Durrande, and O. Roustant. Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 2018.

A. F. López-Lopera, F. Bachoc, and O. Roustant. High-dimensional additive Gaussian processes under monotonicity constraints. In *NeurIPS*. 2022.

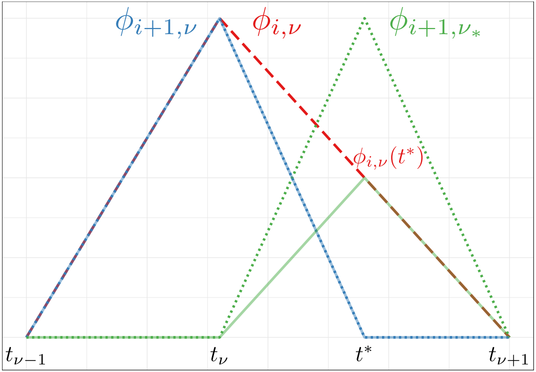
H. Maatouk and X. Bay. Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, 2017.

A. Pakman and L. Paninski. Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics*, 2014.

S. Petit, F. Zaoui, A.-L Popelin, C. Goeury, and N. Goutal. Couplage entre indices à base de dérivées et mode adjoint pour l'analyse de sensibilité globale. Application sur le code Mascaret. HAL e-prints, September 2016.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, 2005.

Projection of the hat function into the new basis space



Projection of the hat function into the new basis space.