Bayesian Optimization for High-Dimensional Problems

Tanguy APPRIOU [(1), (2)], David GAUDRIE [(1)], Didier RULLIERE [(2)]

(1) STELLANTIS

(2) École des Mines de Saint-Etienne, LIMOS

Journées scientifiques CIROQUO
24 mai 2023

OUTLINE

1) **Introduction**
   - Kriging surrogate models and Bayesian optimization
   - Challenges in high dimension

2) **High-dimensional surrogate via a combination of Kriging sub-models**

3) **Numerical results**

4) **Perspectives and current work**

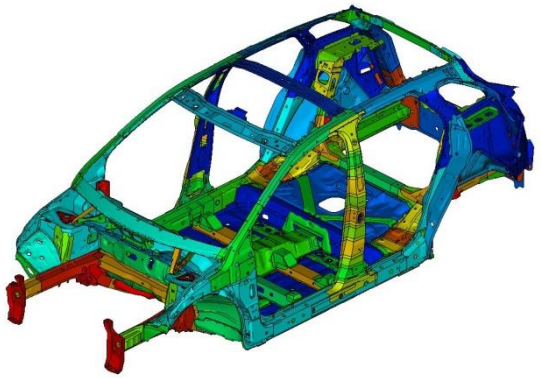
- We are interested in the optimization of a black-box function :

$$y : \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d \rightarrow y(\boldsymbol{x}) \in \mathbb{R}.$$

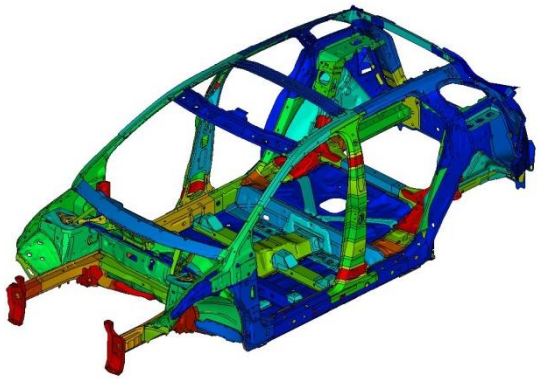
→ We want to find the best design :

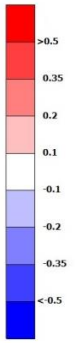
$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathcal{X}} y(\boldsymbol{x}) .$$

Initial model

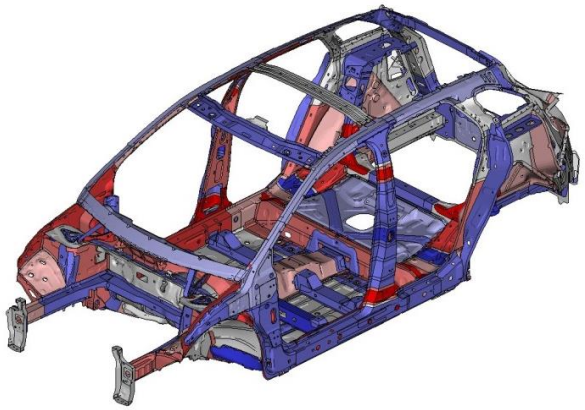


Optimised model





Thickness difference



Example : optimization of the Peugeot 3008 to minimize the vehicle weight while satisfying the norms for chock resistance.

KRIGING SURROGATE MODELS

- We are in the context where the black-box function y is expensive to evaluate :

→ Evaluating the function for a single design can take hours.

↳ We can only afford of few observations.

↳ We cannot use the usual optimization methods which require a large number of these evaluations.

Build a surrogate model

$$y(x) \approx \hat{y}(x)$$

Expensive true objective function

Cheap analytical approximation

- We dispose of n observations $\boldsymbol{Y} = \left(y(\boldsymbol{x_1}), \dots, y(\boldsymbol{x_n})\right)^T$ at the sample locations $\boldsymbol{X} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n})^T$.

→ The ordinary Kriging method approximates y as the realization of a Gaussian Process :

$$Y(.) \sim GP\left(\mu, k_{\sigma,\boldsymbol{\theta}}(.,.)\right).$$

- $k_{\sigma,\boldsymbol{\theta}}(.,.)$ is the covariance function (kernel) with σ^2 the variance of the GP and $\boldsymbol{\theta} \in \mathbb{R}^d$ the covariance length-scales.

- We obtain the Kriging predictors for the mean and predictive variance by conditioning the GP Y over $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{Y})$:

$$\hat{y}(\boldsymbol{x}) = E(Y(\boldsymbol{x})|\mathcal{D}) = \mu + k(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})^{-1}(\boldsymbol{Y} - \boldsymbol{1}\mu),$$

$$\hat{s}^2(\boldsymbol{x}) = Var(Y(\boldsymbol{x})|\mathcal{D}) = k(\boldsymbol{x}, \boldsymbol{x}) - k(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})^{-1}k(\boldsymbol{X}, \boldsymbol{x}).$$

COVARIANCE FUNCTION

The choice of the covariance function is very important to obtain a good prediction.

Popular choices of 1D stationary covariance are :

- Exponential : $k_{\sigma,\theta}(x,x') = \sigma^2 \exp\left(-\frac{|x-x'|}{\theta}\right)$,
- Gaussian : $k_{\sigma,\theta}(x,x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\theta^2}\right)$,
- Matérn 5/2 : $k_{\sigma,\theta}(x,x') = \sigma^2 \left(1+\sqrt{5}\frac{|x-x'|}{\theta}+\frac{5(x-x')^2}{3\theta^2}\right)\exp\left(-\sqrt{5}\frac{|x-x'|}{\theta}\right)$,

Typically, the hyperparameters are optimized to maximize the log-likelihood of the model :

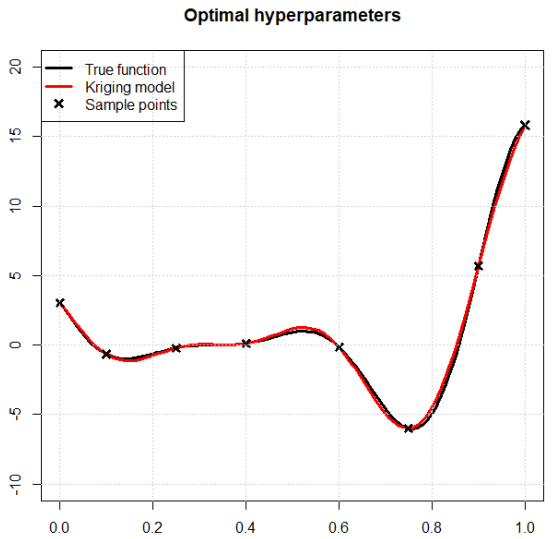
$$\mathcal{L}(\sigma,\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{Y}^T\boldsymbol{K}_{\sigma,\boldsymbol{\theta}}^{-1}\boldsymbol{Y} - \frac{1}{2}\log\left|\boldsymbol{K}_{\sigma,\boldsymbol{\theta}}\right| - \frac{n}{2}\log(2\pi)\,.$$

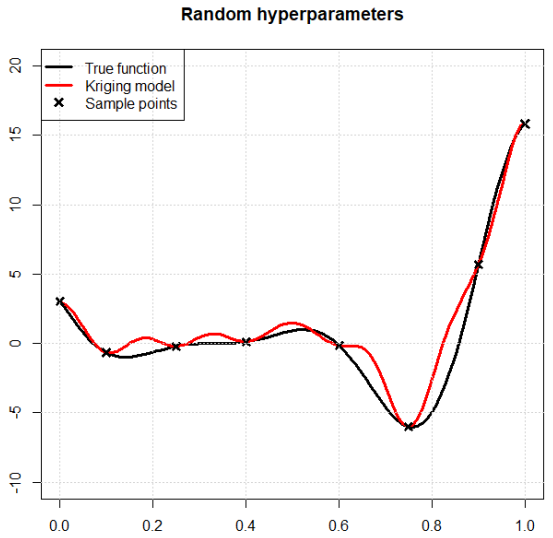
Denoting $\boldsymbol{R}$ the correlation matrix such that $\boldsymbol{K}_{\sigma,\boldsymbol{\theta}} = \sigma^2\boldsymbol{R}_{\boldsymbol{\theta}}$, the MLE estimator for σ^2 is :

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{Y}.$$

And we obtain the length-scales by solving the minimization problem :

$$\theta_{MLE} = \arg\min_{\boldsymbol{\theta}} \; -\frac{1}{2}\log(\hat{\sigma}^2_{MLE}) - \frac{1}{2}\log(|\boldsymbol{R}_{\boldsymbol{\theta}}|)\,.$$





EFFICIENT GLOBAL OPTIMIZATION

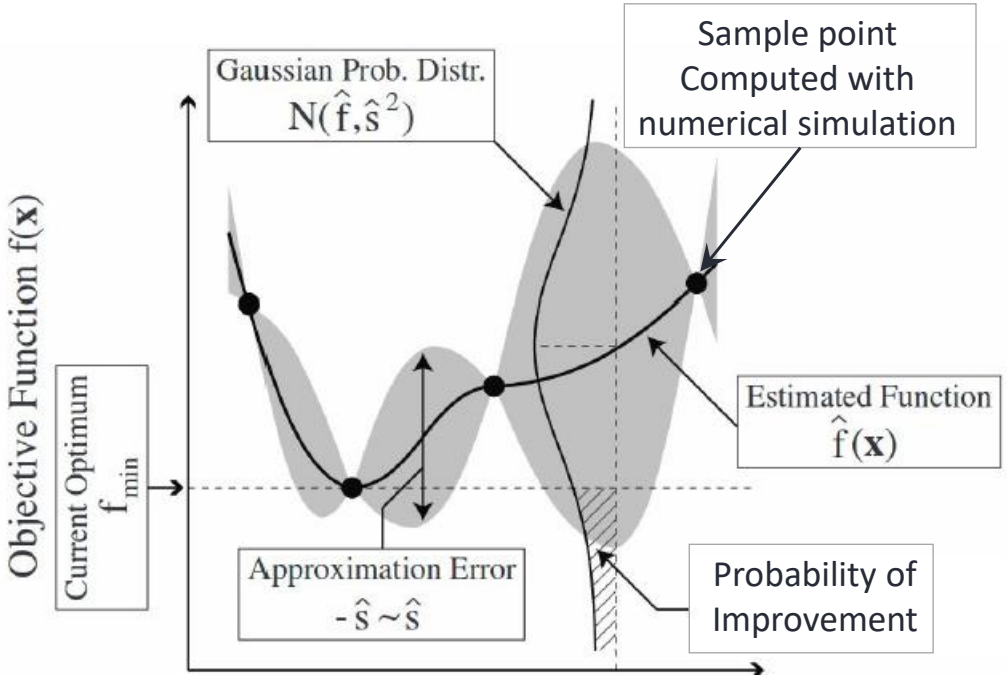
For Kriging-based optimization, we build the sampling plan sequentially by adding new training points to refine the model based on an acquisition criterion. This is called **efficient global optimization** (EGO) (see Jones et al., 1998).

A popular method to decide where to add new samples is to use a criterion called **Expected Improvement** (EI).

- The expected improvement is computed with both the Kriging estimate value and the model error value:

$$E[I(\boldsymbol{x})] = E\left(\left(y_{min} - Y(\boldsymbol{x})\right)^{+}\right).$$

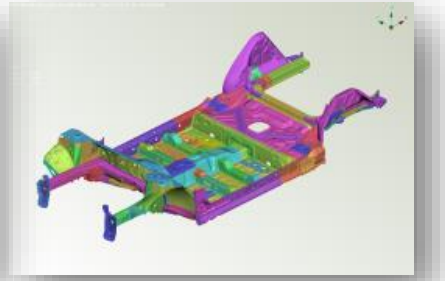
- EI balances local search around the optimum and global search where the Kriging model is not very accurate.

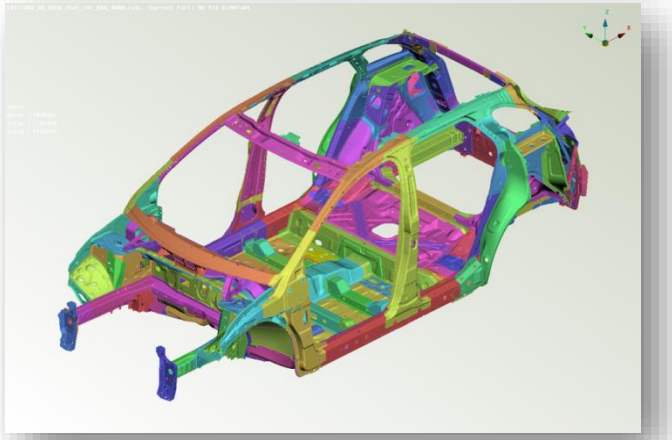


WHAT IS HIGH DIMENSION ?

53 sur périmètre superstructure

77 sur périmètre base





A total of 130 parameters for this example !

The dimension of the problem is the dimension of the design space.

→ That is, **the number of design variables in the problem**.

Typically, for a number of design variables superior to ≈ 20, the ordinary Kriging method begins to show its limits.

ISSUES OF KRIGING IN HIGH DIMENSION

- One issue is the **optimization of the hyperparameters**.

There is one length-scale hyperparameter per dimension, and all these hyperparameters need to be optimized.
→ The optimization of the hyperparameters is difficult :

- d-dimensional problem (with $d > 20$ up to $\approx 100 - 150$).
- **The optimization can be costly** due to the cost of the cost for the evaluation of the objective (log-likelihood) and its gradient is in $O(n^3)$.
- The shape of **the objective is flat is some areas**.
- When the training data is sparse (which is often the case for high dimensional problems since we cannot afford to compute too many observations), **the likelihood criterion over-fit the data which lead to a bad estimation of the hyperparameters.**

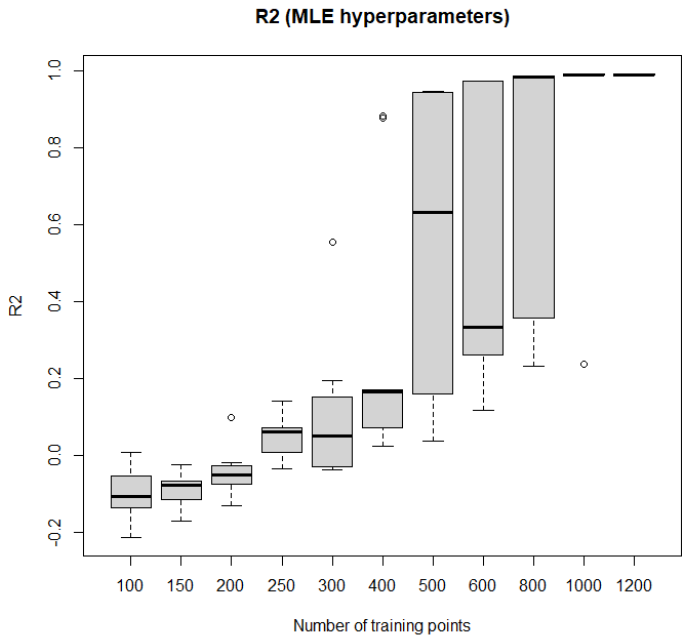
ISSUES OF KRIGING IN HIGH DIMENSION

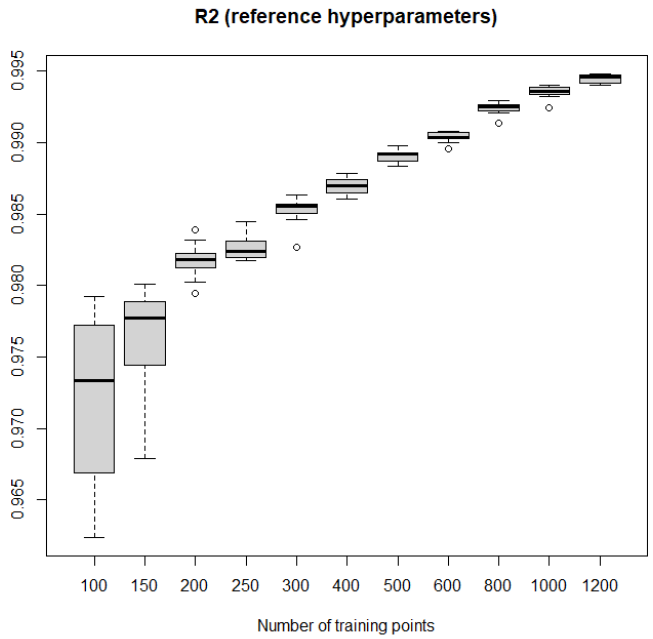
- An illustration of this difficulty: approximating the 50D sphere function:

$$f_{sphère}(x_1, \dots, x_d) = \sum_{i=1}^{d} (x_i - 0{,}5)^2 , \qquad 0 \leq x_i \leq 1.$$

We build a Kriging model using a varying number of training points and compare to a Kriging model with reference hyperparameters :

- 100 iterations for the hyperparameter optimization using the DiceKriging package in R.
- The reference hyperparameters are obtained by doing the optimization with 5000 points.
- The boxplots give the results for 10 different runs.





ISSUES OF KRIGING IN HIGH DIMENSION

- Several methods have been proposed to solve this issue :

- Reduction of the problem's dimension by embedding the design space into a lower-dimension space (see for example Constantine et al., 2015, Bouhlel et al., 2016).
- Additive Kriging where the function is assumed to be a sum of one-dimensional components (see for example Durrande et al., 2012).
- Penalized version of the likelihood to improve the robustness of the hyperparameter optimization (see for example RobustGaSP in Gu et al., 2018).

→ In the following, we present a method to **bypass the hyperparameter optimization** by combining Kriging sub-models with fixed length-scales.

This method is both:
- **Fast** since it avoids the expensive hyperparameter optimization,
- **Easily generalized** since it does not assume a particular form of the underlying function.

OUTLINE

1) Introduction
   - Kriging surrogate models and Bayesian optimization
   - Challenges in high dimension

2) **High-dimensional surrogate via a combination of Kriging sub-models**
   - Choice of the sub-models
   - Combination of the sub-models
   - Variance of the combination

3) Numerical results

4) Perspectives and current work

COMBINATION OF SUB-MODELS WITH FIXED LENGTH-SCALE

The motivation of the method is to avoid the costly and difficult optimization of the Kriging hyperparameters for high-dimensional problems.

→ We propose a model which is a combination of Kriging models with fixed length-scale (see preprint Appriou et al., 2022) :

$$M_{tot}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i(\boldsymbol{x}) M_i(\boldsymbol{x}), \qquad \text{with } M_i(x) = k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}_i) K_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{Y}_i - \mu_i) \text{ Kriging model with fixed length−scale vector } \boldsymbol{\theta}_i.$$

- The weights of the combination can be obtained in **closed-form** and does not require a numerical optimization.
- This method does not rely on reducing the dimension in order to **preserve the correlations between each design variables** and to ensure that there is **no loss of information due to a design space of reduced dimension**.
- This method is flexible since each sub-model can be constructed with **different subsets of points, different design variables, different covariance functions** …
- The complexity of the combination is $O(pn^3)$ (one inversion of the $n \times n$ covariance matrix for each of the p sub-models). For a reasonable number of sub-models, this is less than the cost of ordinary Kriging in $O(\alpha_{iter} n^3)$ where α_{iter} is the number of matrix inversion for the hyperparameter optimization.

CHOICE OF THE SUB-MODEL LENGTH-SCALES

An appropriate method to select the length-scales of each sub-model is essential for this method to work.

- **We want to have variety in the sub-models**, so that the combined model can select well-suited behaviors through the weights in the combination.

→ To have variety among the sub-models, we need **variety among the length-scales** as they are the main source of difference between the sub-models.

- We want to avoid too small or too large values of the length-scales:

- For too small values:

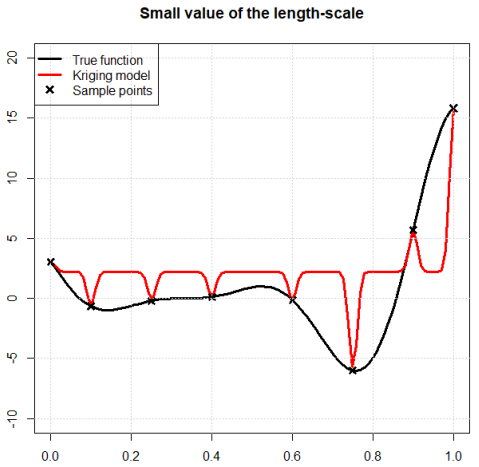
$$k_\theta(x_i, x_j) \longrightarrow 0 \text{ for all } i \neq j, \text{ and } \boldsymbol{K}_\theta \longrightarrow \sigma^2 \boldsymbol{I}_n.$$

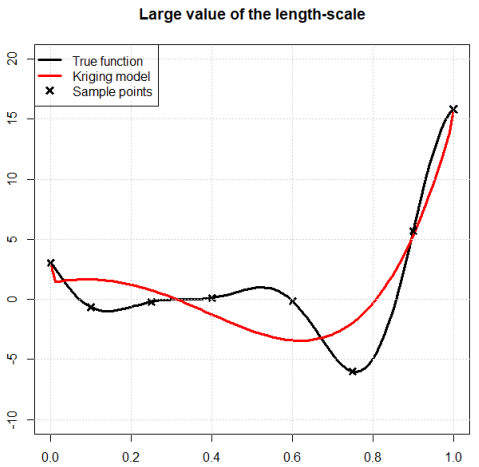
In this case, the Kriging model will return to its mean outside the observations.

- For too large values:

$$k_\theta(x_i, x_j) \longrightarrow 1, \text{ and } \boldsymbol{K}_\theta \longrightarrow \sigma^2 \mathbf{1}_{n \times n}.$$

In this case, the covariance matrix is ill-defined and its inversion will pose numerical issues.





First, we will study an example for which analytical expressions can be obtained.

- Assume that design points are distributed as a random vector $\boldsymbol{X} = \left(X^{(1)}, \dots, X^{(d)}\right)$ with i.i.d components with common variance σ_X^2 and kurtosis κ_X.

- We note D^2 the random square distance between two independent points $\boldsymbol{X}$ and $\boldsymbol{X}'$ of the design. For a large enough dimension:

$$D^2 = \sum_{k=1}^{d} (X_k - X_k')^2 \sim \mathcal{N}\left(2d\sigma_X^2, 2d\sigma_X^4(\kappa_X + 1)\right).$$

- For a Gaussian correlation:

$$R_\theta = e^{-\frac{1D^2}{2\theta^2}} \sim \text{Lognormal}\left(\frac{-\sigma_X^2}{\theta^2}d, \frac{\sigma_X^4}{2\theta^4}(\kappa_X + 1)d\right).$$

- We can finally obtain the entropy of the correlation:

$$H(R_\theta) = \mathbb{E}\left(-\log f_{R_\theta}(R_\theta)\right) = -\frac{\sigma_X^2}{\theta^2}d + \frac{1}{2}\ln\left(\frac{\sigma_X^4}{2\theta^4}d(\kappa_X + 1)2\pi\right) + \frac{1}{2}.$$

CHOICE OF THE SUB-MODEL LENGTH-SCALES

$$H(R_\theta) = \mathbb{E}\big(-\log f_{R_\theta}(R_\theta)\big) = -\frac{\sigma_X^2}{\theta^2}d + \frac{1}{2}\ln\left(\frac{\sigma_X^4}{2\theta^4}d(\kappa_X+1)2\pi\right) + \frac{1}{2}.$$

How to use the knowledge about this entropy ?

- When sampling the length-scales, we want to favor θ corresponding to high entropy values, which result in a high variability in the correlation.

- In the two degenerated cases of small and large length-scales: $R_{\theta_{small}} \longrightarrow \delta_0$ and $R_{\theta_{large}} \longrightarrow \delta_1$, which gives:

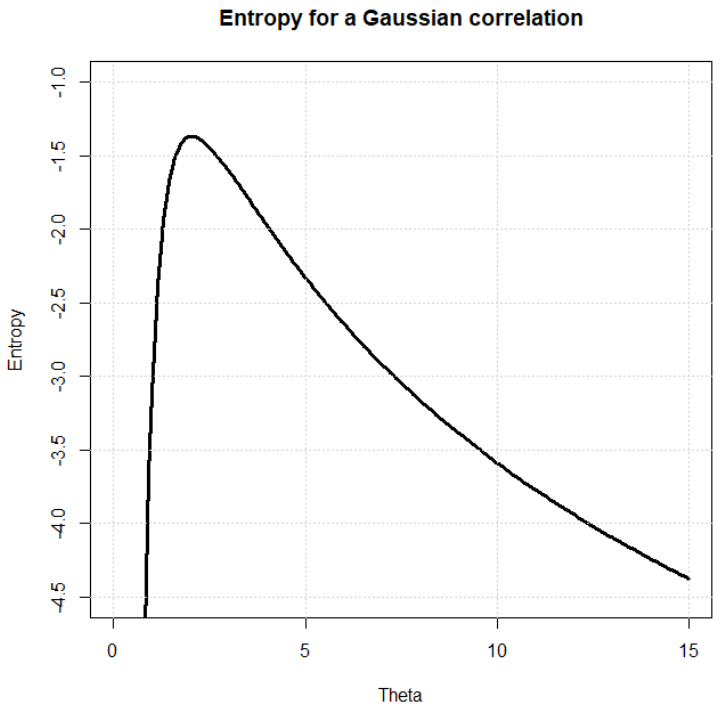
$$H\big(R_{\theta_{small}}\big) \longrightarrow -\infty \text{ and } H\left(R_{\theta_{large}}\right) \longrightarrow -\infty.$$

- We can also obtain the maximal entropy length-scale:

$$\arg\max_\theta H(R_\theta) = \sigma_X\sqrt{d}$$

- Finally, we will sample the length-scales using a positive transformation of the entropy:

$$f(\theta) \propto \exp\big(H(R_\theta)\big).$$



Entropy of a Gaussian correlation in 50D for a uniform design ($\sigma_X^2 = 1/12$ and $\kappa_X = 9/5$).

WEIGHTS OF THE COMBINATION

Now, we present the method used to obtain the weights in the combination: $M_{tot}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i M_i(\boldsymbol{x})$.

- One method (see for example Viana et al., 2009) relies on minimizing the LOOCV error of the combination:

$$e_{LOOCV}(M_{tot}) = \frac{1}{n}\sum_{k=1}^{n}\left(\sum_{i=1}^{p} w_i M_{i-k}(x_k) - y(x_k)\right)^2 = \boldsymbol{w}^T \boldsymbol{C}\boldsymbol{w}.$$

→ The components of the matrix $\boldsymbol{C}$ are : $c_{ij} = \frac{1}{N} {e_{CV}}_i^T e_{CV\,j}$, with $e_i^{(k)} = \left[K_i^{-1}Y\right]_k / \left[K_i^{-1}\right]_{k,k}$, $k = 1, \dots, n$.

The weights are then obtained by :

$$\boldsymbol{w}_{LOOCV} = \arg\min_{w} \boldsymbol{w}^T \boldsymbol{C}\boldsymbol{w}, \qquad \text{subject to } \mathbf{1}^T\boldsymbol{w} = 1 \qquad \Longrightarrow \boldsymbol{w}_{LOOCV} = \frac{\mathbf{1}^T\boldsymbol{C}^{-1}}{\mathbf{1}^T\boldsymbol{C}^{-1}\mathbf{1}} .$$

VARIANCE OF THE COMBINATION

- One of the main advantage of the Kriging method is that it naturally provides a measure of the model error. For a Kriging model $Y(.) \sim GP\left(\mu, k_{\sigma,\boldsymbol{\theta}}(.,.)\right)$:

$$\mathbb{E}\left(\left(M(x) - Y(x)\right)^2\right) = Var(Y(\boldsymbol{x})|Y(X)) = k(\boldsymbol{x},\boldsymbol{x}) - k(\boldsymbol{x},\boldsymbol{X})\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})^{-1}k(\boldsymbol{X},\boldsymbol{x})$$

→ This prediction error is essential to assess the model uncertainty when performing Bayesian optimization for example.

- For our combination of Kriging sub-models: $M_{tot}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i M_i(\boldsymbol{x})$.

We can obtain the error prediction for every individual sub-model, but the covariance structure between the sub-models is unknown.

→ We cannot directly access the prediction error of the combination.

VARIANCE OF THE COMBINATION

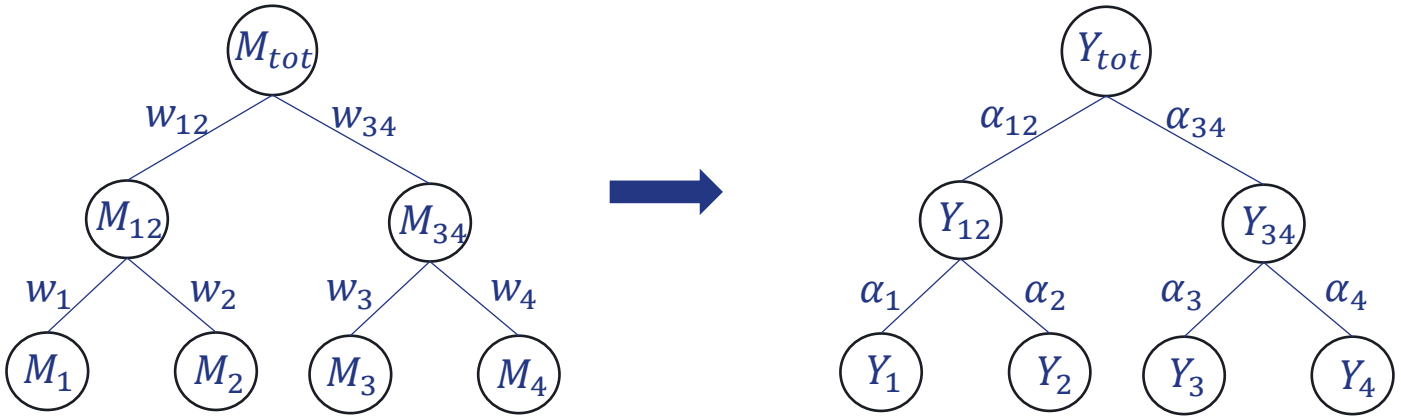
- To obtain the variance of the combination, we add the hypothesis that the underlying Gaussian Process Y is a combination (with different weights) of independent Gaussian Processes:

$$Y = \sigma_{tot}^2 \sum_{i=1}^{p} \alpha_p Y_p \,, \qquad \text{with } Y_p \sim GP\left(\mu_p, r_{\theta_p}(.,.)\right), \qquad \sum_{i=1}^{p} \alpha_p = 1, \qquad \text{and } \sigma_{tot}^2 \text{ the variance of the GP.}$$

Thus, the covariance of this GP is:

$$k_{tot}(.,.) = \sigma_{tot}^2 \sum_{i=1}^{p} \alpha_i^2 r_{\theta_i}(.,.)\,.$$

- To simplify the upcoming expressions, we will also assume that the sub-models (and the associated GPs) are combined following a binary tree structure:



VARIANCE OF THE COMBINATION

- The weights α in the combination of GPs are chosen to minimize the expected mean-square error of the combined model under the corresponding hypothesis:

$$\alpha^* = \arg\min_{\alpha} \mathbb{E}\left[\int_{\mathcal{X}} (wM_1(\boldsymbol{x}) + (1-w)M_2(\boldsymbol{x}) \; - \alpha Y_1(\boldsymbol{x}) + (1-\alpha)Y_2(\boldsymbol{x}))^2 d\boldsymbol{x}\right].$$

By approximation the global MSE using the LOOCV error, we obtain:

$$\alpha^* = \arg\min_{\alpha} \mathbb{E}_{Y=\alpha Y_1+(1-\alpha)Y_2}\, e_{LOOCV}(M_{tot}) = \frac{a_1(w)}{a_1(w)+a_2(w)}, \qquad \text{with:}$$

$$a_1(w) = w^2 \mathbb{E}_{Y=Y_2}\left(e_{LOOCV}(M_1)\right) + (1-w^2)\mathbb{E}_{Y=Y_2}\left(e_{LOOCV}(M_2)\right),$$

$$a_2(w) = (1-w)^2 \mathbb{E}_{Y=Y_1}\left(e_{LOOCV}(M_2)\right) + (1-(1-w)^2)\mathbb{E}_{Y=Y_1}\left(e_{LOOCV}(M_1)\right).$$

VARIANCE OF THE COMBINATION

- Once we obtain the weights α, the model uncertainty can be obtained as:

$$\mathbb{E}\left(\left(M_{comb}(\boldsymbol{x}) - Y(\boldsymbol{x})\right)^2\right) = \mathbb{E}\left(M_{comb}(\boldsymbol{x})^2 + Y(\boldsymbol{x})^2 - 2M_{comb}(\boldsymbol{x})Y(\boldsymbol{x})\right)$$
$$= Var\left(Y(\boldsymbol{x})\right) + Var\left(M_{comb}(\boldsymbol{x})\right) - 2cov\left(M_{comb}(\boldsymbol{x}), Y(\boldsymbol{x})\right)$$
$$= Var\left(Y(\boldsymbol{x})\right) + \boldsymbol{w}^T \boldsymbol{K_M}(\boldsymbol{x})\boldsymbol{w} - 2\boldsymbol{w}^T \boldsymbol{k_M}(\boldsymbol{x}),$$

With:

$$\left(K_M(\boldsymbol{x})\right)_{i,j} = Cov\left(M_i(\boldsymbol{x}), M_j(\boldsymbol{x})\right) = k_i(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{K}_i(\boldsymbol{X}, \boldsymbol{X})^{-1}\, Cov\left(Y(\boldsymbol{X}), Y(\boldsymbol{X})\right) \boldsymbol{K}_j(\boldsymbol{X}, \boldsymbol{X})^{-1} k_j(\boldsymbol{X}, \boldsymbol{x}),$$

$$\left(k_M(x)\right)_i = Cov\left(M_i(\boldsymbol{x}), Y(\boldsymbol{x})\right) = k_i(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{K}_i(\boldsymbol{X}, \boldsymbol{X})^{-1} Cov\left(Y(\boldsymbol{X}), Y(\boldsymbol{x})\right).$$

And:

$$Cov\left(Y(.), Y(.)\right) = k_{tot}(.,.) = \sigma_{tot}^2 \sum_{i=1}^{p} \alpha_i^2 r_{\theta_i}(.,.)\,.$$

VARIANCE OF THE COMBINATION

- Finally, the last step is to calibrate the amplitude of the variance using the amplitude hyperparameter σ_{tot}^2.

Generally, this can be done by observing that the normalized LOO errors should be normally distributed:

$$\frac{e_{LOO}}{\sqrt{Var_{LOO}}} \sim \mathcal{N}(0, \sigma_{tot}^2).$$

→ Thus, one way to obtain the amplitude is:

$$\sigma_{tot}^2 = Var\left(\frac{e_{LOO}}{\sqrt{Var_{LOO}}}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{{e_{LOO}}_i^2}{{Var_{LOO}}_i}.$$

However, this definition tends to give too large amplitudes due to the presence of many outliers in the LOO error.

To have an expression for the amplitude more robust to outliers and which overall give prediction interval that are better calibrated, we fit the empirical inter-quartile distance of the LOO error to that of a Gaussian distribution:

$$IQ\left(\frac{e_{LOO}}{\sigma_{tot}\sqrt{Var_{LOO}}}\right) = IQ_{norm} \quad \Leftrightarrow \sigma_{tot} = \frac{IQ\left(\frac{e_{LOO}}{\sqrt{Var_{LOO}}}\right)}{IQ_{norm}} = \frac{q_{0,75}\left(\frac{e_{LOO}}{\sqrt{Var_{LOO}}}\right) - q_{0,25}\left(\frac{e_{LOO}}{\sqrt{Var_{LOO}}}\right)}{IQ_{norm}}.$$

OUTLINE

1) Introduction
   - Kriging surrogate models and Bayesian optimization
   - Challenges in high dimension

2) High-dimensional surrogate via a combination of Kriging sub-models

3) **Numerical results**
   - Analytical test function
   - Real-world applications

4) Perspectives and current work

NUMERICAL RESULTS – ANALYTICAL TEST FUNCTION

We tested the method for the approximation of a GP trajectory in 50D (with isotropic length-scale $\theta_{true} = 2$ or $\theta_{true} = 3$) :

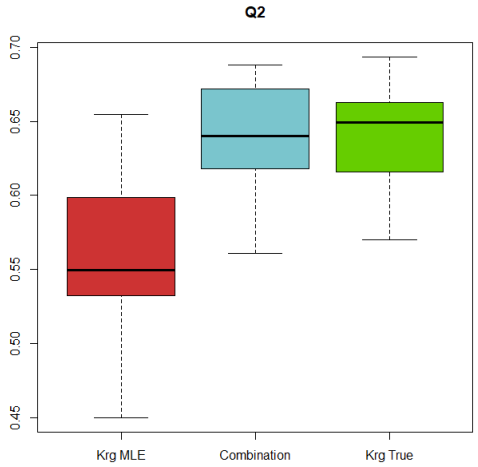
1. Sample a GP trajectory (known length-scale) in **dimension 50**.
2. Select **500 training points** on the trajectory and **5000 test points** to evaluate the precision.
3. Build 32 non-isotropic sub-models with different random length-scales each (6 levels in the tree structure).
4. Build an ordinary Kriging model with hyperparameters estimated by MLE to compare the performances (300 maximum iterations).
5. Build an ordinary Kriging model with the true length-scales (same as the trajectory). This model is the ideal model whose precision we want to approach.
6. Repeat the experiment 10 times.

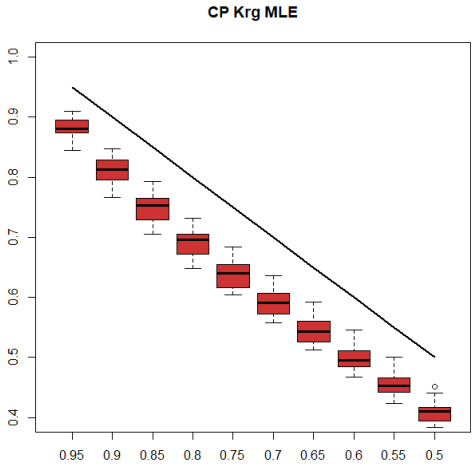
To measure the precisions for the 3 models, we compute the Q^2: $Q^2 = 1 - \frac{\sum_{i=1}^{n_{test}} \left(y_{test}(x_i) - \hat{y}(x_i)\right)^2}{\sum_{i=1}^{n_{test}} \left(y_{test}(x_i) - \frac{1}{n_{test}} \sum_{k=1}^{n_{test}} y_{test}(x_k)\right)^2}$

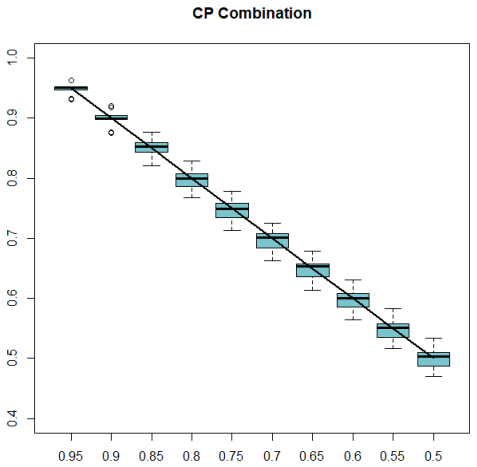
We also access the quality of the error prediction by computing the coverage probabilities for different levels.

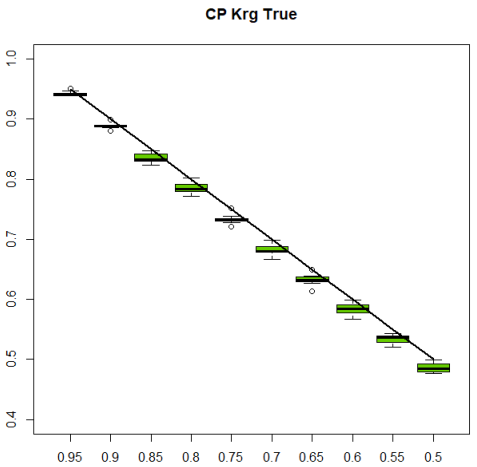
NUMERICAL RESULTS – ANALYTICAL TEST FUNCTION

$\theta_{true} = 3$





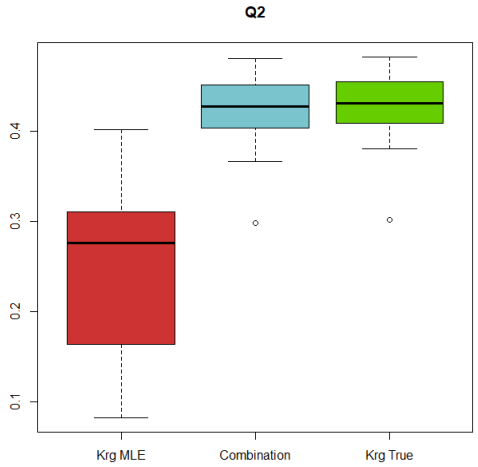


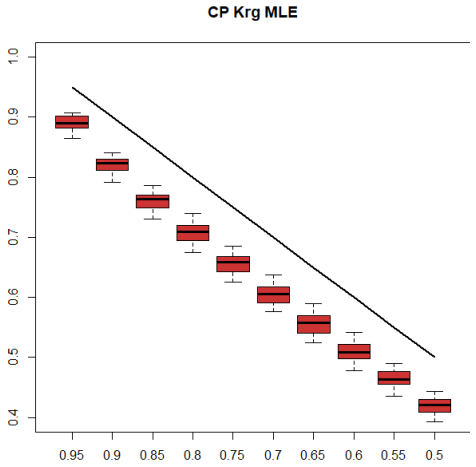


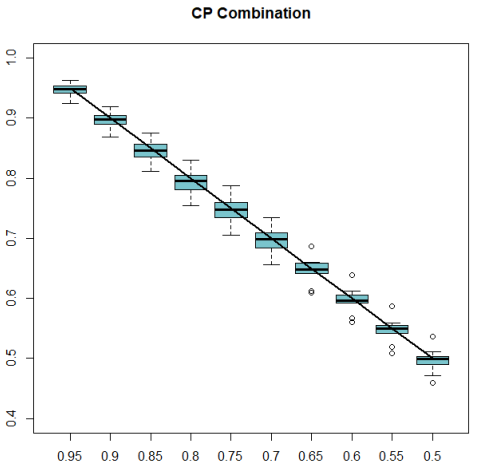
Average computational time:

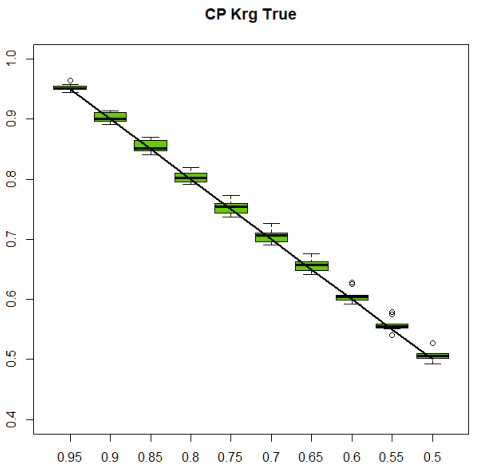
- Krg MLE: 2,9 mins
- Combination : 0,33 mins

$\theta_{true} = 2$









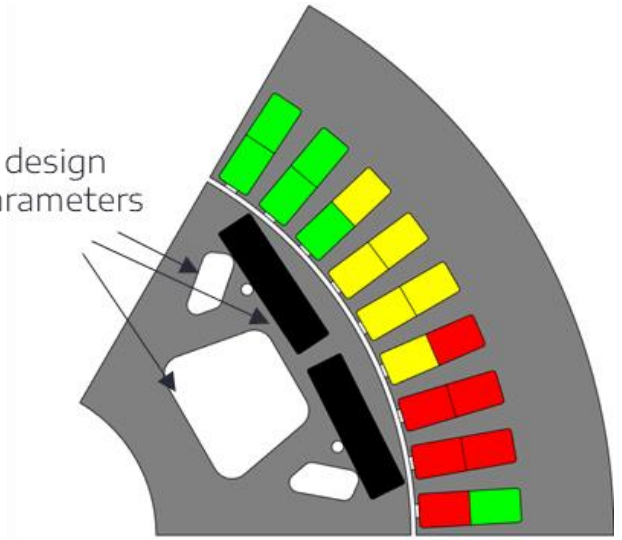
Average computational time:

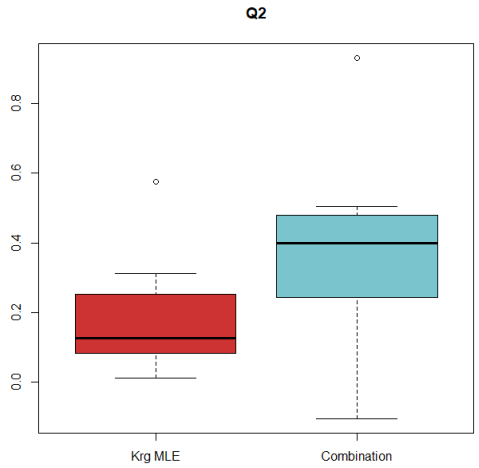
- Krg MLE: 3,4 mins
- Combination : 0,33 mins

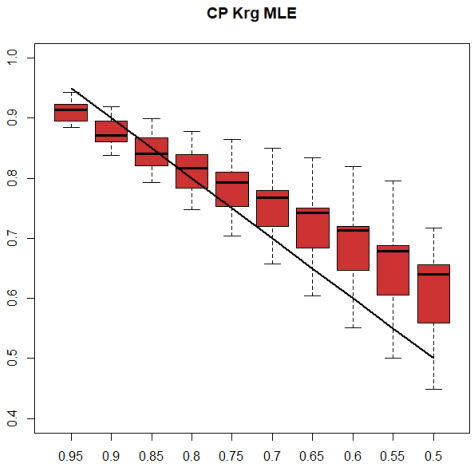
NUMERICAL RESULTS – REAL-WORLD APPLICATIONS

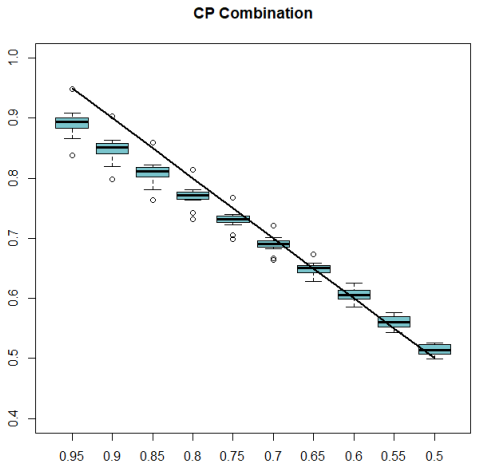
- Study of an electrical machine:

- 37 design variables,
- 500 training points,
- 4500 test points,
- 2 objectives and 10 constraints to surrogate,
- Average results over 10 runs.







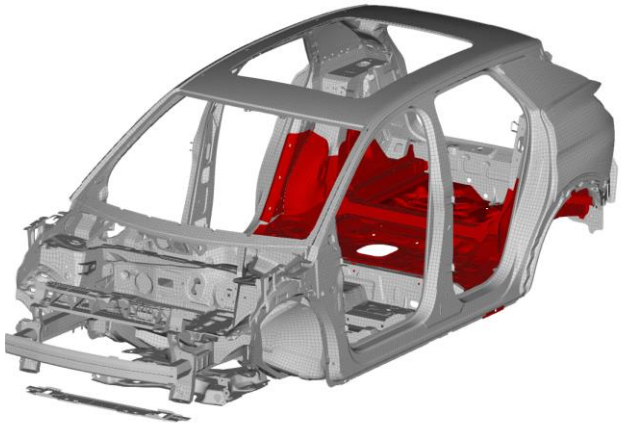


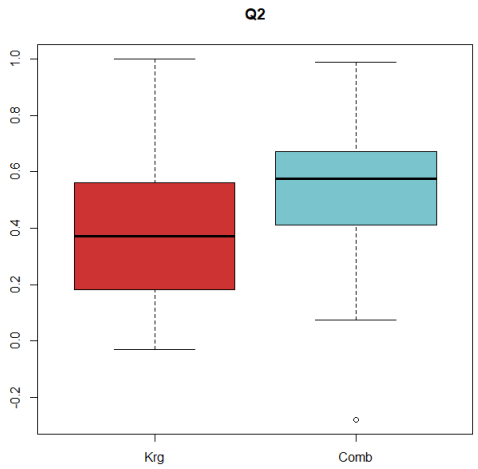
Average computational time:

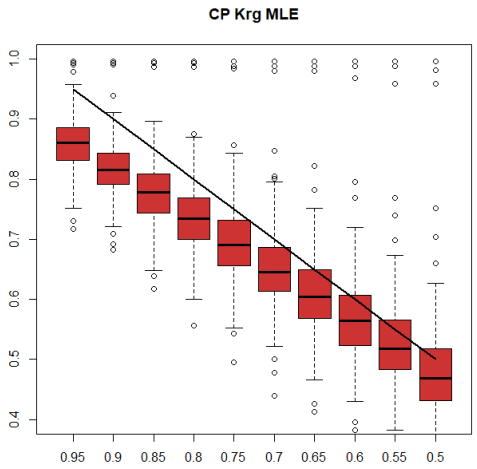
- Krg MLE: 17,1 mins
- Combination : 3,0 mins

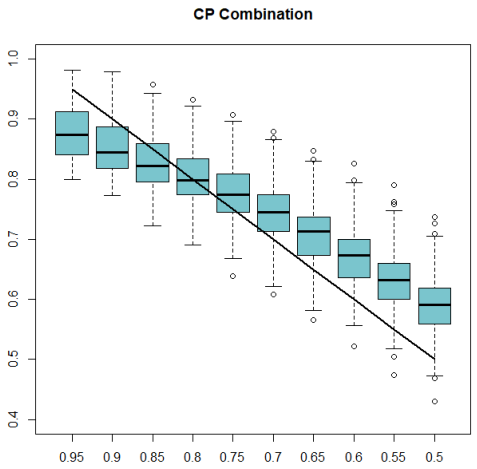
- Study of the Peugeot 3008 (vibratory comfort and rear crash safety) :

  - 48 design variables,
  - 300 training points,
  - 327 test points,
  - 2 objectives and 413 constraints (a surrogate model is built only for 190 constraints).









Computational time:

- Krg MLE: 220 mins
- Combination : 15,8 mins

OUTLINE

1) Introduction
   - Kriging surrogate models and Bayesian optimization
   - Challenges in high dimension

2) High-dimensional surrogate via a combination of Kriging sub-models

3) Numerical results

4) Perspectives and current work

PERSPECTIVES

- We developed **a model with better accuracy than the ordinary Kriging** in high dimension, especially when the length-scales are poorly estimated using MLE, and which is both easier and faster to construct.

- We also gave a method to obtain the prediction error for the combined model which gives prediction interval that are overall well-calibrated and suitable for Bayesian optimization.

Future work :

- Apply the combined model for Bayesian optimization and see the potential gains in both construction time and number of iterations required to find the optimum.

- There are still challenges in the acquisition criterion for Bayesian optimization:
- The acquisition function is very flat with only a few peaks which can be hard to find, especially so in high dimension.
- In high dimension, the volume near the borders of the design space becomes dominant. This can result in adding most of the new points near the borders.

- We can also diversify the sub-models using subsets of points or subsets of design variables for example.

- …

Thank you for your attention !

Contact :

Tanguy APPRIOU
tanguy.appriou@stellantis.com

REFERENCES

- Appriou, T., Rullière, D. and Gaudrie, D., 2022. Combination of High-Dimensional Kriging Sub-models.
- Bouhlel, M.A., Bartoli, N., Otsmane, A. and Morlier, J., 2016. Improving kriging surrogates of high-dimensional design models by Partial Least Squares dimension reduction. *Structural and Multidisciplinary Optimization*, *53*(5), pp.935-952.
- Cao, Y. and Fleet, D.J., 2014. Generalized product of experts for automatic and principled fusion of Gaussian process predictions. *arXiv preprint arXiv:1410.7827*.
- Constantine, P.G., Dow, E. and Wang, Q., 2014. Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM Journal on Scientific Computing*, *36*(4), pp.A1500-A1524.
- Deisenroth, M. and Ng, J.W., 2015, June. Distributed gaussian processes. In *International Conference on Machine Learning* (pp. 1481-1490). PMLR.
- Durrande, N., Ginsbourger, D. and Roustant, O., 2012. Additive covariance kernels for high-dimensional Gaussian process modeling. In *Annales de la Faculté des sciences de Toulouse: Mathématiques* (Vol. 21, No. 3, pp. 481-499).
- Gu, M., Palomo, J. and Berger, J.O., 2018. RobustGaSP: Robust Gaussian stochastic process emulation in R. *arXiv preprint arXiv:1801.01874*.
- Gu, M., Wang, X. and Berger, J.O., 2018. Robust Gaussian stochastic process emulation. *The Annals of Statistics*, *46*(6A), pp.3038-3066.

REFERENCES

- Jones, D.R., Schonlau, M. and Welch, W.J., 1998. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, *13*(4), pp.455-492.
- Liu, H., Ong, Y.S., Shen, X. and Cai, J., 2020. When Gaussian process meets big data: A review of scalable GPs. *IEEE transactions on neural networks and learning systems*, *31*(11), pp.4405-4423.
- Matheron, G., 1963. Principles of geostatistics. *Economic geology*, *58*(8), pp.1246-1266.
- Quinonero-Candela, J. and Rasmussen, C.E., 2005. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, *6*, pp.1939-1959.
- Rasmussen, C.E. and Williams, C.K., 2006. *Gaussian processes for machine learning*. Cambridge, MA: MIT press.
- Roustant, O., Ginsbourger, D. and Deville, Y., 2012. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of statistical software*, *51*, pp.1-55.
- Rullière, D., Durrande, N., Bachoc, F. and Chevalier, C., 2018. Nested Kriging predictions for datasets with a large number of observations. *Statistics and Computing*, *28*(4), pp.849-867.
- Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H.P., 1989. Design and analysis of computer experiments. *Statistical science*, *4*(4), pp.409-423.
- Viana, F.A., Haftka, R.T. and Steffen, V., 2009. Multiple surrogates: how cross-validation errors can help us to obtain the best predictor. *Structural and Multidisciplinary Optimization*, *39*(4), pp.439-457.

ANNEX - ISSUES OF KRIGING IN HIGH DIMENSION

- One issue in high dimension is **the increased size of the correlation matrix**. The higher the dimension, the more sample points are needed to obtain a Kriging model with a good accuracy. But Kriging models scale poorly for large numbers of observations.

→ $K(\boldsymbol{X}, \boldsymbol{X})^{-1}$ is necessary for the Kriging prediction (inverse of the covariance matrix of size $n \times n$ with n the number of sample points).

→ **Complexity in $\boldsymbol{O}(\boldsymbol{n^3})$** prohibitive for large number of sample points.

This issue was tackled in the literature using different techniques of :

- Sparse approximation of the covariance matrix (see for example Hensman et al., 2013, Quinonero-Candela et al., 2005).
- Combination of models with subsets of points (see for example Deisenroth et al., 2015, Cao et al., 2014, Rullière et al., 2018).

ANNEX – PREVIOUS METHOD FOR THE CHOICE OF THE LENGTH-SCALES

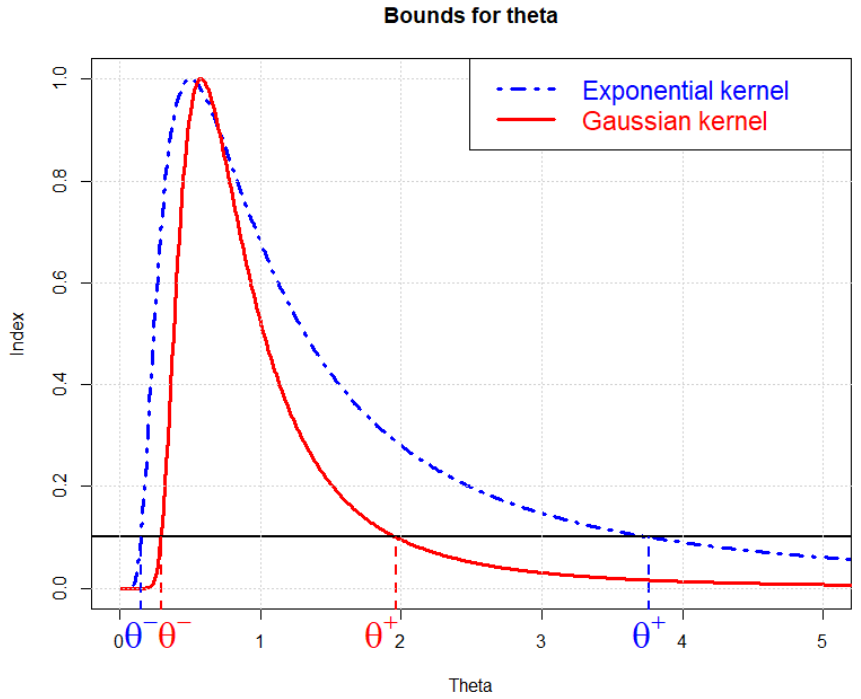
In previous presentations:

- Bounds for the length-scales were chosen based on the covariance function variations with respect to variations of the length-scales.
- The length-scale variation influence index was defined as :

$$I^{(\ell)}\left(\left\|\frac{\boldsymbol{r}}{\boldsymbol{\theta}}\right\|, \theta^{(\ell)}\right) = \left| \frac{\frac{\partial}{\partial \theta^{(\ell)}} k\left(\left\|\frac{\boldsymbol{r}}{\boldsymbol{\theta}}\right\|\right)}{\max\limits_{\theta^{(\ell)}, \boldsymbol{\theta} \in \mathcal{C}} \frac{\partial}{\partial \theta^{(\ell)}} k\left(\left\|\frac{\boldsymbol{r}}{\boldsymbol{\theta}}\right\|\right)} \right|$$

- The length-scales were chosen uniformly in the bounded interval:

$$\theta^{(\ell)} \sim \mathcal{U}\left[\theta_{min}^{(\ell)}, \theta_{max}^{(\ell)}\right], \qquad \ell = 1, \ldots, d.$$



An issue with this method is that it produces too many large or small samples of the length-scale.

→ In the following, we introduce an entropy-based sampling scheme to improve over this method.

ANNEX – EMPIRICAL ENTROPY

In practice, for any correlation function R_θ and any design plan $\boldsymbol{X}$.

1. For a given length-scale θ. We sample N values of the correlation for the design plan $\boldsymbol{X}$: $r_\theta^{(1)}, \dots r_\theta^{(N)}$.

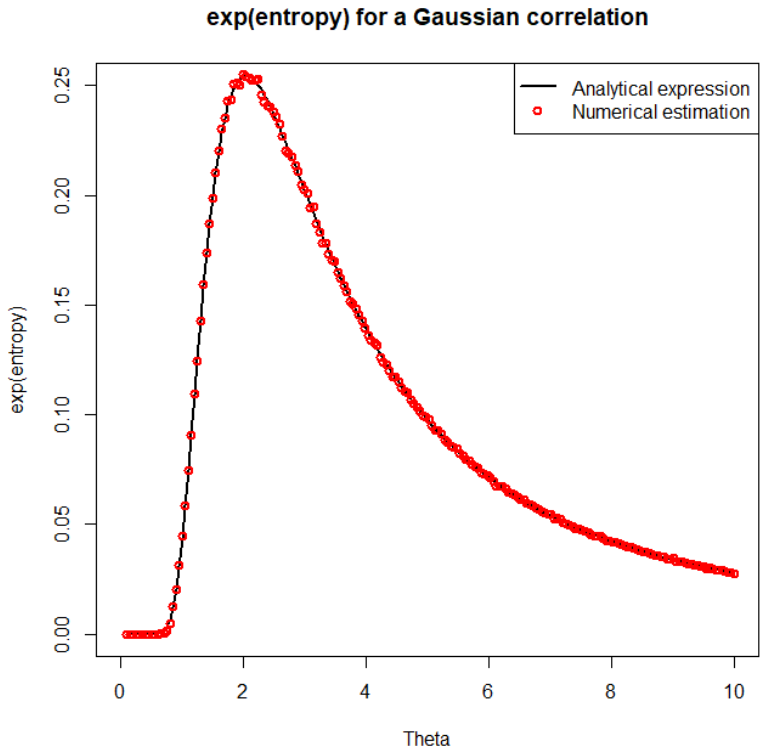
2. We make a kernel estimation $\hat{f}_{R_\theta}$ of the density of R_θ based on these samples.

3. We compute the empirical entropy:

$$\widehat{H}(R_\theta) = -\frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_{R_\theta}\left(r_\theta^{(i)}\right).$$

4. We define a grid of possible values for the length-scales $\theta_{grid}^{(\ell)},\ \ell = 1, \dots, q$, and we sample with probability:

$$P\left(\theta_{grid}^{(\ell)}\right) \propto \exp\left(H(R_\theta)\right).$$



→ We sample d length-scale values (one for each dimension) for each of the sub-models.