Formation CIROQUO

Tanguy APPRIOU [1], [2]

(1) STELLANTIS

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

18 novembre 2024

R Package “combinedkriging”

Available on Github : https://github.com/TAppriou/combinedkriging

Notebook also on Github :

Or with the link, then make a copy :
https://colab.research.google.com/drive/1jMLOrt3PTepAgABOJGTz0jiu5-fueJXb?usp=sharing

- Package to build a combination of Kriging models with fixed length-scales.

→ Use for high-dimensional Bayesian optimization.

COMBINATION OF KRIGING MODELS

→ We propose a model which is a combination of Kriging models with random length-scales

$$M_{tot}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i(\boldsymbol{x}) M_i(\boldsymbol{x}),$$

with $M_i(x) = \mu_i + k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) K_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{Y} - \mu_i)$ Kriging model with fixed length−scale vector $\boldsymbol{\theta}_i$.

1) Choice of the weights

- Weights based on the sub-models variance: $\hat{s}_i^{\,2}(\boldsymbol{x}) = k_{\boldsymbol{\theta}_i}(\boldsymbol{x},\boldsymbol{x}) - k_{\boldsymbol{\theta}_i}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{K}_{\boldsymbol{\theta}_i}(\boldsymbol{X},\boldsymbol{X})^{-1}k_{\boldsymbol{\theta}_i}(\boldsymbol{X},\boldsymbol{x})$

  - Product of Experts (PoE): correspond to the weights of the best linear combination for independent sub-models.

$$w_{PoE_i}(\boldsymbol{x}) = \frac{\hat{s}_i^{-2}(\boldsymbol{x})}{\sum_{j=1}^{p}\hat{s}_j^{-2}(\boldsymbol{x})}$$

→ Do not depend on the observations.

→ Gives more weight to models with higher length-scales.

  - Generalized PoE (gPoE): corrects PoE by weighting the contribution of each sub-models.

$w_{gPoE_i}(\boldsymbol{x}) = \dfrac{\beta_i\hat{s}_i^{-2}(\boldsymbol{x})}{\sum_{j=1}^{p}\beta_j\hat{s}_j^{-2}(\boldsymbol{x})}$, the inner weights $\boldsymbol{\beta}$ can be obtained by LOOCV for example :

$$\beta^* = \arg\min_{\boldsymbol{\beta}} e_{LOOCV}\left(\sum_{i=1}^{p} w_{gPoE_i}(\boldsymbol{\beta})M_i\right), \qquad \text{subject to} \sum_{i=1}^{p} w_{gPoE_i}(\boldsymbol{\beta}) = 1\,.$$

→ No analytical expression.

CHOICE OF THE WEIGHTS

- Mixture of experts (MoE): weights based on the sub-models likelihoods.

$$w_{MoE_i} = \frac{\mathcal{L}(\boldsymbol{\theta}_i)}{\sum_{j=1}^{p} \mathcal{L}(\boldsymbol{\theta}_j)}, \qquad \text{where } \mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{Y}^T\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{Y} - \frac{1}{2}\log|\boldsymbol{K}_{\boldsymbol{\theta}}| - \frac{n}{2}\log(2\pi)$$

→ The likelihood is not always a good measure of model accuracy when few observations are available.

→ The likelihood of sub-models often differ by several order of magnitudes, thus this method often select a few sub-models instead of doing a combination.

CHOICE OF THE WEIGHTS

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

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

  - Normal LOOCV:

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

  → Weights can be negative or more than 1.

  - Diagonal LOOCV: to enforce $w_i \in [0,1]$, we keep only the diagonal terms of $\boldsymbol{C}$:

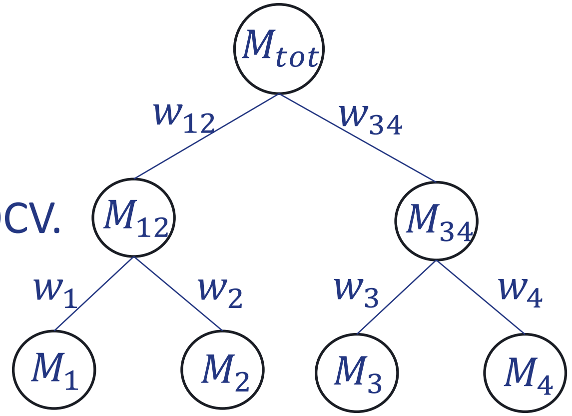
$$\boldsymbol{w}_{LOOCV_{diag}} = \frac{C_{diag}^{-1}\boldsymbol{1}}{\boldsymbol{1}^T C_{diag}^{-1}\boldsymbol{1}} \quad \Longrightarrow w_{LOOCV_{diag_i}} = \frac{{e_{LOOCV}}_i^{-1}}{\sum_{l=1}^{P} {e_{LOOCV}}_l^{-1}}$$

  → Tends to give weights to all models (close to average of all models)

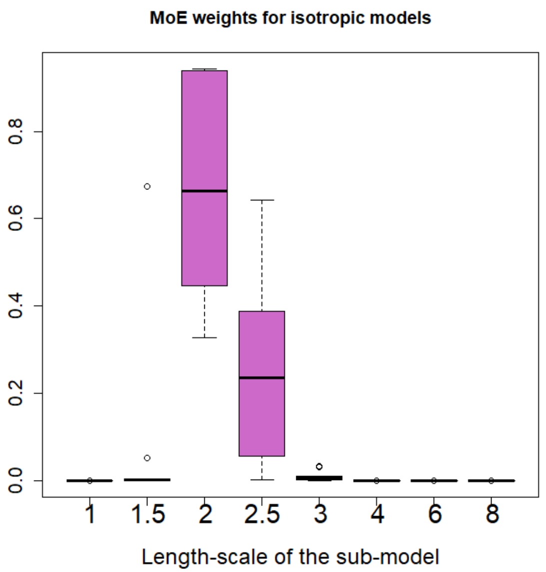
  - Binary LOOCV: to enforce $w_i \in [0,1]$, we combine the sub-models two by two using LOOCV.

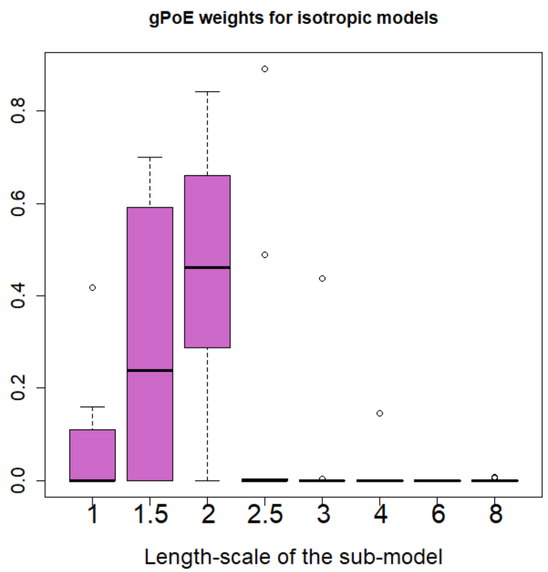
  → Number of sub-models limited to a power of 2.

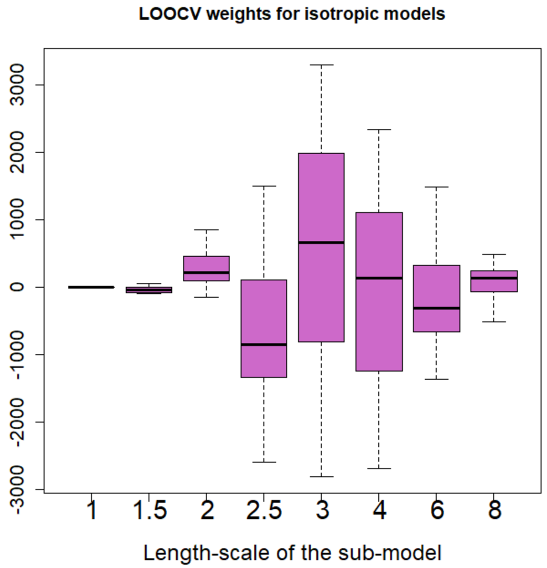
  → Can implement sparsity.

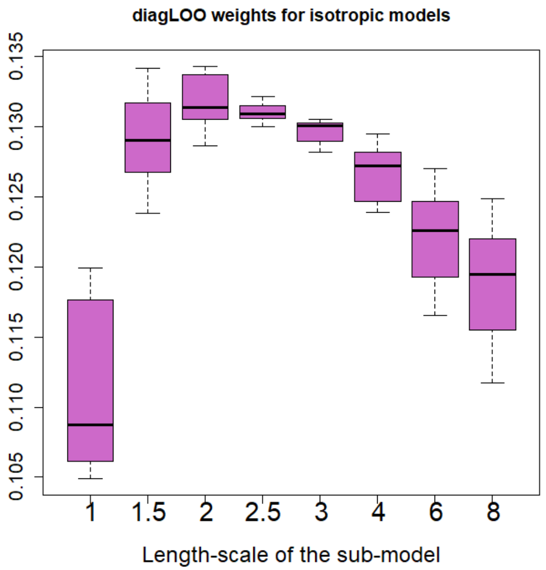


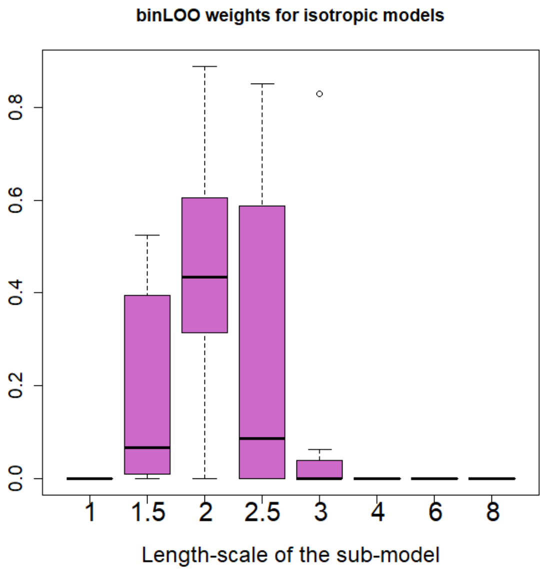
CHOICE OF THE WEIGHTS











→ We propose a model which is a combination of Kriging models with random length-scales

$$M_{tot}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i(\boldsymbol{x}) M_i(\boldsymbol{x}),$$

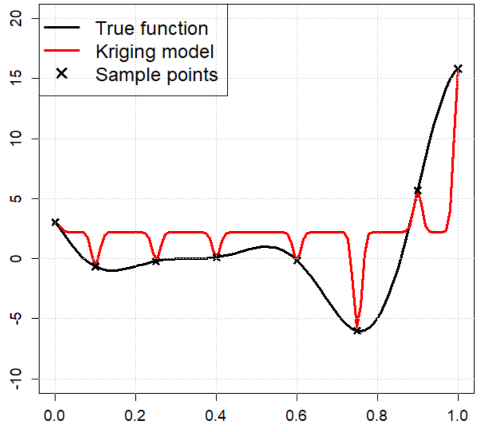
with $M_i(x) = \mu_i + k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) K_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{Y} - \mu_i)$ Kriging model with fixed length−scale vector $\boldsymbol{\theta}_i$.

2) Choice of the sub-model length-scales

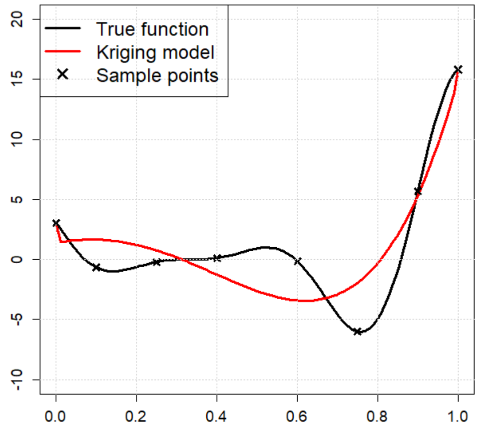
- We want to sample the length-scales in **a range of appropriate values** to avoid degenerate cases.

- For too small values: $k_\theta(x_i, x_j) \longrightarrow 0$ for all $i \neq j$, and $\boldsymbol{K}_\theta \longrightarrow \sigma^2 \boldsymbol{I}_n$.
- For too large values: $k_\theta(x_i, x_j) \longrightarrow 1$, and $\boldsymbol{K}_\theta \longrightarrow \boldsymbol{1}_{n \times n}$.

Small θ



Large θ (+nugget)



→ We need to define an interval $\left[\theta_{min}^{(\ell)}, \theta_{max}^{(\ell)}\right]$ for sampling the length-scales.

SAMPLING THE LENGTH-SCALES

- Assume the random vector of design points is $\boldsymbol{X} = \left(X^{(1)}, \dots, X^{(d)}\right)$ with i.i.d components.

We denote $\sigma^2 = Var\left(X^{(i)}\right)$, and $\kappa = E\left[\left(\frac{X^{(i)}-\mu}{\sigma}\right)^4\right]$.

We can obtain the distribution of the distance between two random points $\boldsymbol{X}$ and $\boldsymbol{X}'$:

$$D^2 = \sum_{i=1}^{d}\left(X^{(i)} - X'^{(i)}\right)^2 \sim \mathcal{N}\left(2d\sigma^2, 2d\sigma^4(\kappa+1)\right)$$

- Typical distances in the DoE can be obtained by taking the root of a 95% Gaussian confidence interval :

$$[r_{min}, r_{max}] = \left[\sigma\sqrt{2d - 1{,}96\sqrt{2(\kappa+1)d}}\,, \sigma\sqrt{2d + 1{,}96\sqrt{2(\kappa+1)d}}\right].$$

- If the standard deviation differs on every dimension ℓ :

$$\left[r_{min}^{(\ell)}, r_{max}^{(\ell)}\right] = \sigma^{(\ell)}[r_{min}, r_{max}] = \left[\sigma^{(\ell)}\sqrt{2d - 1{,}96\sqrt{2(\kappa+1)d}}\,,\ \sigma^{(\ell)}\sqrt{2d + 1{,}96\sqrt{2(\kappa+1)d}}\right]$$

SAMPLING THE LENGTH-SCALES

- For a fixed distance r, the impact of small variations of the length-scale on the correlation can be measure by the index :

$$I\left(\frac{r}{\theta}\right) = \left| \frac{\frac{\partial}{\partial \theta} k\left(\frac{r}{\theta}\right)}{\max_{\theta} \frac{\partial}{\partial \theta} k\left(\frac{r}{\theta}\right)} \right|$$

- For a given distance, a length-scale is considered influential if the index is more than a threshold for the typical distances :

$$\Theta_{adm}(r) = \left\{\theta : \ I^{(\ell)}\left(\frac{r}{\theta}\right) \geq \delta\right\}.$$

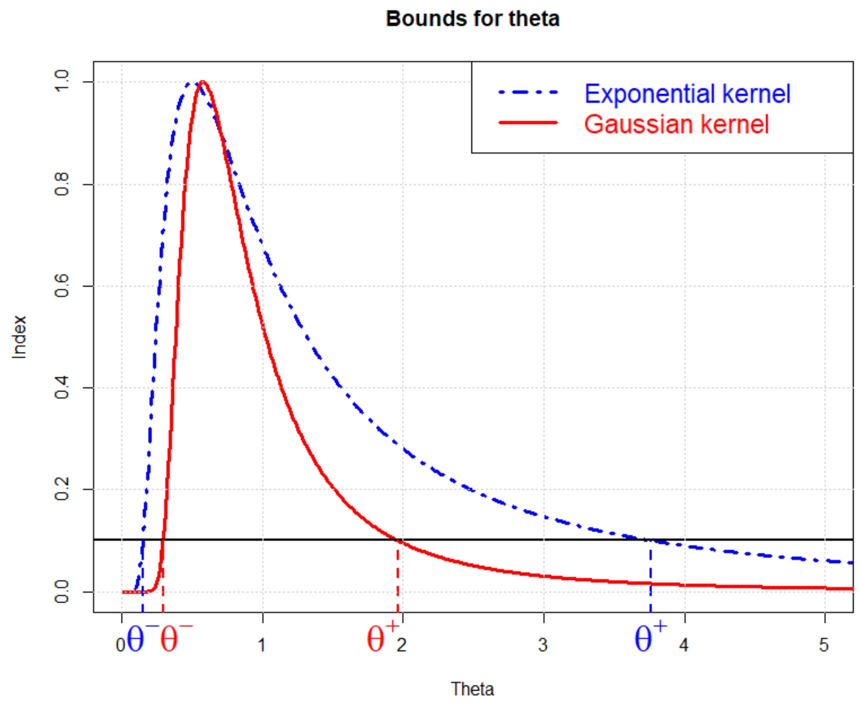
- Note that multiplying by a scale factor α changes the set of admissible length-scales by the same factor :

$$\Theta_{adm}(\alpha r) = \alpha \Theta_{adm}(r).$$

→ We only have to solve for $r = 1$ in θ :

$$I^{(\ell)}\left(\frac{1}{\theta}\right) = \delta.$$

- We denote $\theta^-(k)$ and $\theta^+(k)$ the smallest and largest roots of $I^{(\ell)}\left(\frac{1}{\theta}\right) = \delta$.



SAMPLING THE LENGTH-SCALES

- Putting both factors together :

For typical values of the inter-point distance $r \in \left[r_{min}^{(\ell)}, r_{max}^{(\ell)}\right]$, the length-scales bounds are chosen as :

$$\theta_{min}^{(\ell)} = \inf \bigcup_{r \in \left[r_{min}^{(\ell)}, r_{max}^{(\ell)}\right]} \Theta_{adm}(r) = r_{min}^{(\ell)} \theta^{-}(k), \qquad \theta_{max}^{(\ell)} = \sup \bigcup_{r \in \left[r_{min}^{(\ell)}, r_{max}^{(\ell)}\right]} \Theta_{adm}(r) = r_{max}^{(\ell)} \theta^{+}(k).$$

Finally, we get :

Influence of covariance family

$$\theta_{min}^{(\ell)} = \underbrace{\sigma^{(\ell)} r_{min}} \overbrace{\theta^{-}(k)} \quad \text{and} \quad \theta_{max}^{(\ell)} = \underbrace{\sigma^{(\ell)} r_{max}} \overbrace{\theta^{+}(k)}.$$

Influence of the design

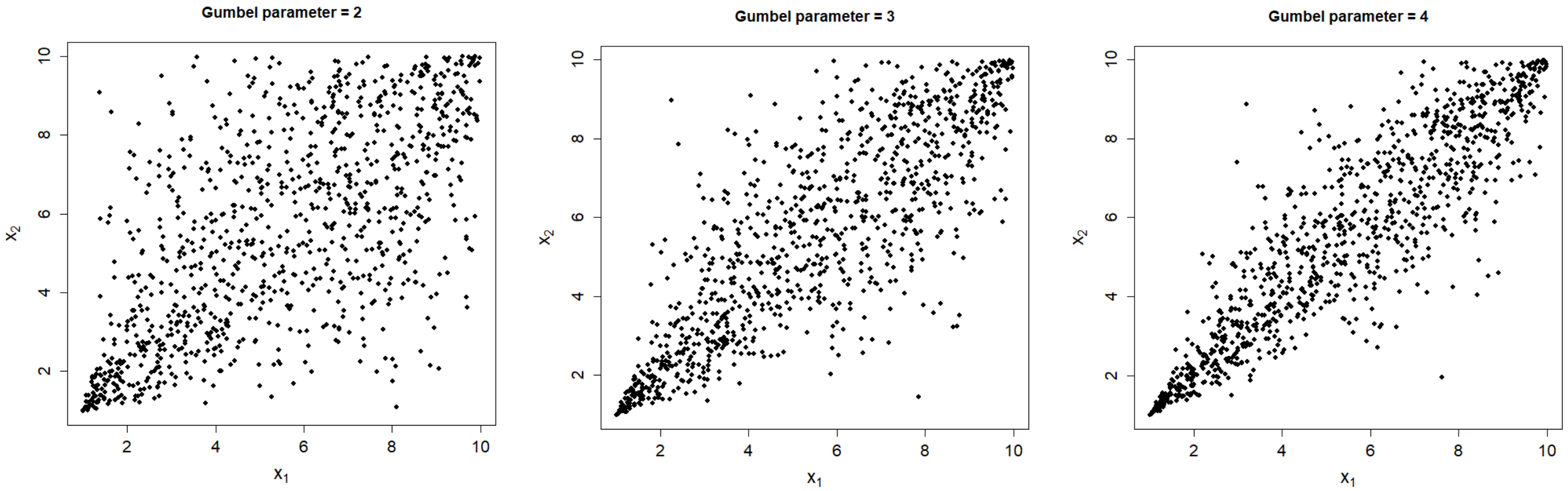
		Design influence			Kernel influence		Resulting bounds	
d	Kernel k	$\sigma^{(\ell)}$	r_{min}	r_{max}	$\theta^{-}(k)$	$\theta^{+}(k)$	$\theta_{min}^{(\ell)}$	$\theta_{max}^{(\ell)}$
10	Exponential	$\frac{1}{\sqrt{12}}$	2.31	5.89	0.15	3.76	0.10	6.39
	Matérn 3/2				0.21	2.74	0.14	4.66
	Matérn 5/2				0.23	2.44	0.15	4.15
	Gaussian				0.29	1.96	0.19	3.33
50	Exponential	$\frac{1}{\sqrt{12}}$	8.20	11.5	0.15	3.76	0.36	12.5
	Matérn 3/2				0.21	2.74	0.50	9.10
	Matérn 5/2				0.23	2.44	0.54	8.10
	Gaussian				0.29	1.96	0.69	6.51
$d \to \infty$	Exponential	$\frac{1}{\sqrt{12}}$	$\sqrt{2d}$	$\sqrt{2d}$	0.15	3.76	$0.061\sqrt{d}$	$1.54\sqrt{d}$
	Matérn 3/2				0.21	2.74	$0.086\sqrt{d}$	$1.12\sqrt{d}$
	Matérn 5/2				0.23	2.44	$0.094\sqrt{d}$	$1.00\sqrt{d}$
	Gaussian				0.29	1.96	$0.12\sqrt{d}$	$0.80\sqrt{d}$

Example of values for a uniform design plan ($\kappa = 9/5$), a standard deviation $\sigma^{(\ell)} = 1/\sqrt{12}$, and a threshold $\delta = 1/10$.

SAMPLING THE LENGTH-SCALES

- Uniform sampling: $\theta^{(\ell)} \sim \mathcal{U}\left[\theta_{min}^{(\ell)}, \theta_{max}^{(\ell)}\right],\ \ell = 1, \dots, d.$

- Sampling with copulas:



- Sampling based on the entropy of the correlations

SAMPLING THE LENGTH-SCALES

- Recall that D^2 the random square distance between two independent points $\boldsymbol{X}$ and $\boldsymbol{X}'$ of the design is :

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

- For a Gaussian correlation (for other kernels, we can use numerical approximations):

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

- We can finally obtain the entropy of the correlation:

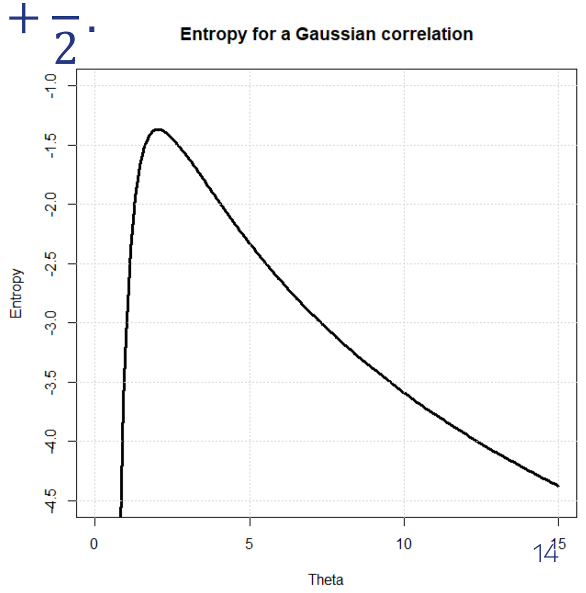
$$H(R_\theta) = \mathbb{E}\left(-\log f_{R_\theta}(R_\theta)\right) = -\frac{\sigma_X^2}{\theta^2}d + \frac{1}{2}\ln\left(\frac{\sigma_X^4}{2\theta^4}d(\kappa_X + 1)2\pi\right) + \frac{1}{2}.$$

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

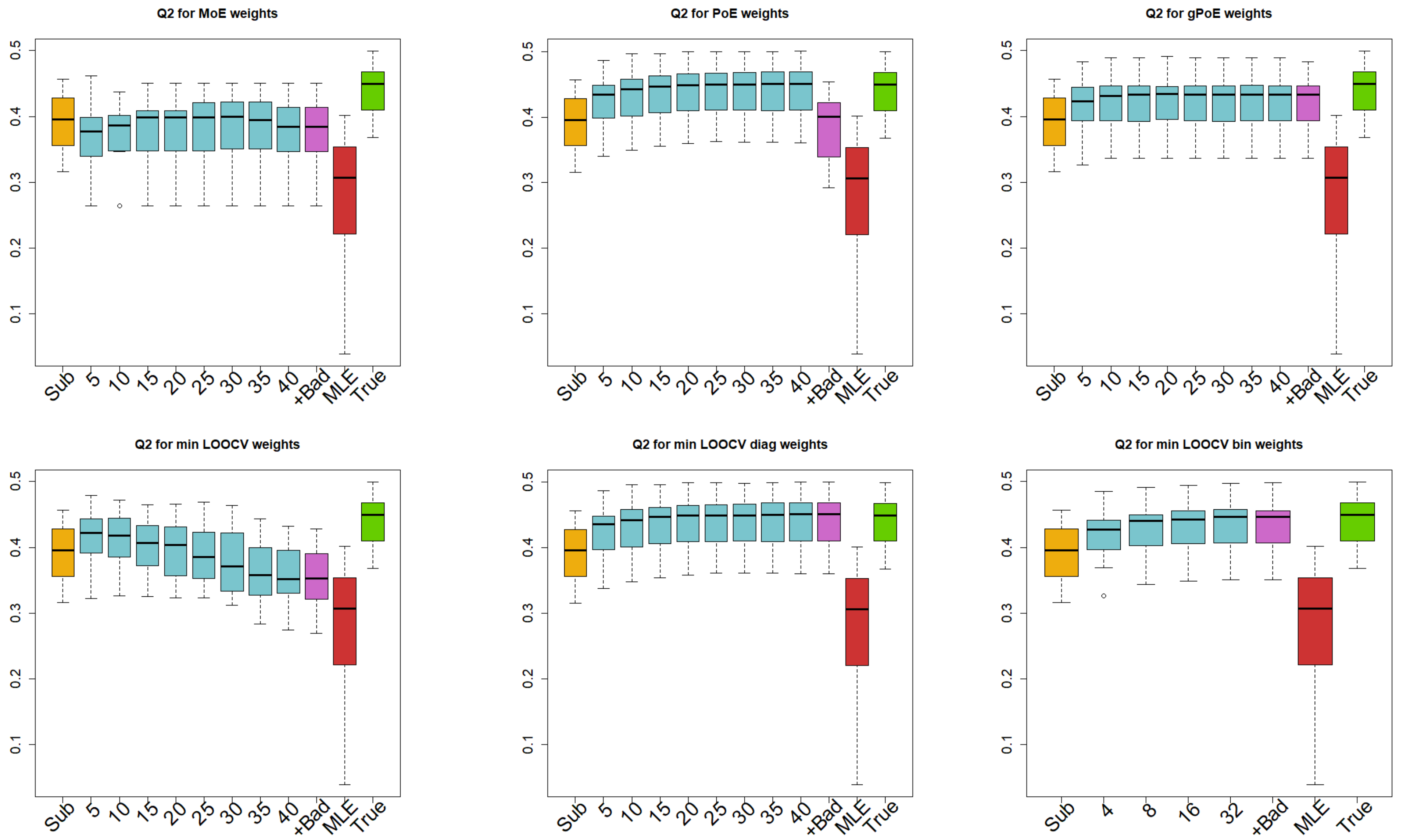
→ We will sample the length-scales using a positive transformation of the entropy:

$$f(\theta) \propto \exp\left(H(R_\theta)\right).$$

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



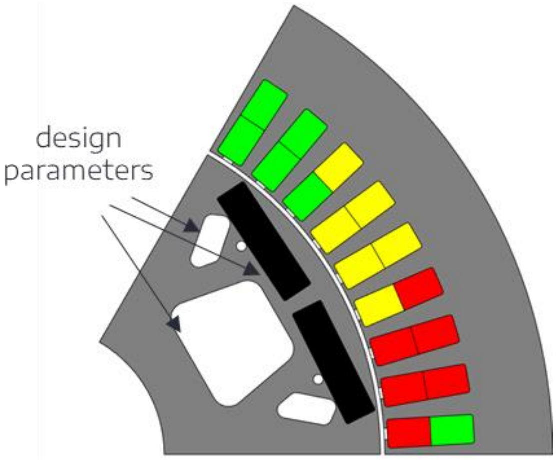
CHOICE OF THE WEIGHTS



- Résultats sur 2 jeux de données tests

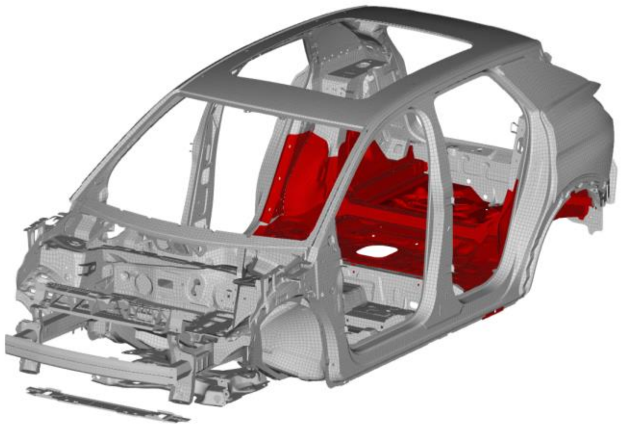
- Étude d'une machine électrique:

- 37 variables de design,
- 500 points d'apprentissage,
- 4500 points test,
- 2 objectifs et 10 contraintes à modéliser,
- Résultats moyennés sur 10 runs.

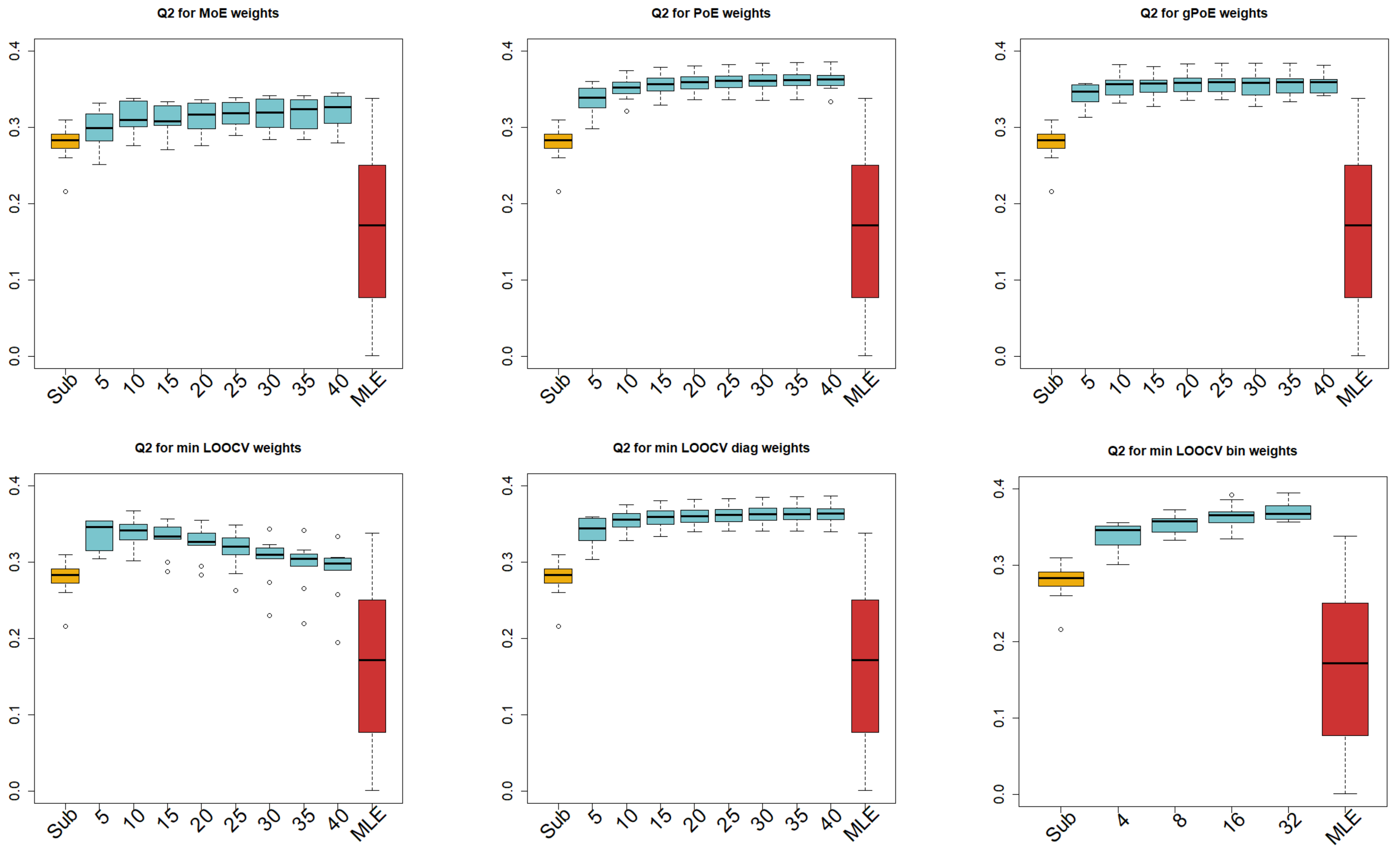


- Étude de la Peugeot 3008 (confort vibratoire et sécurité en crash arrière) :

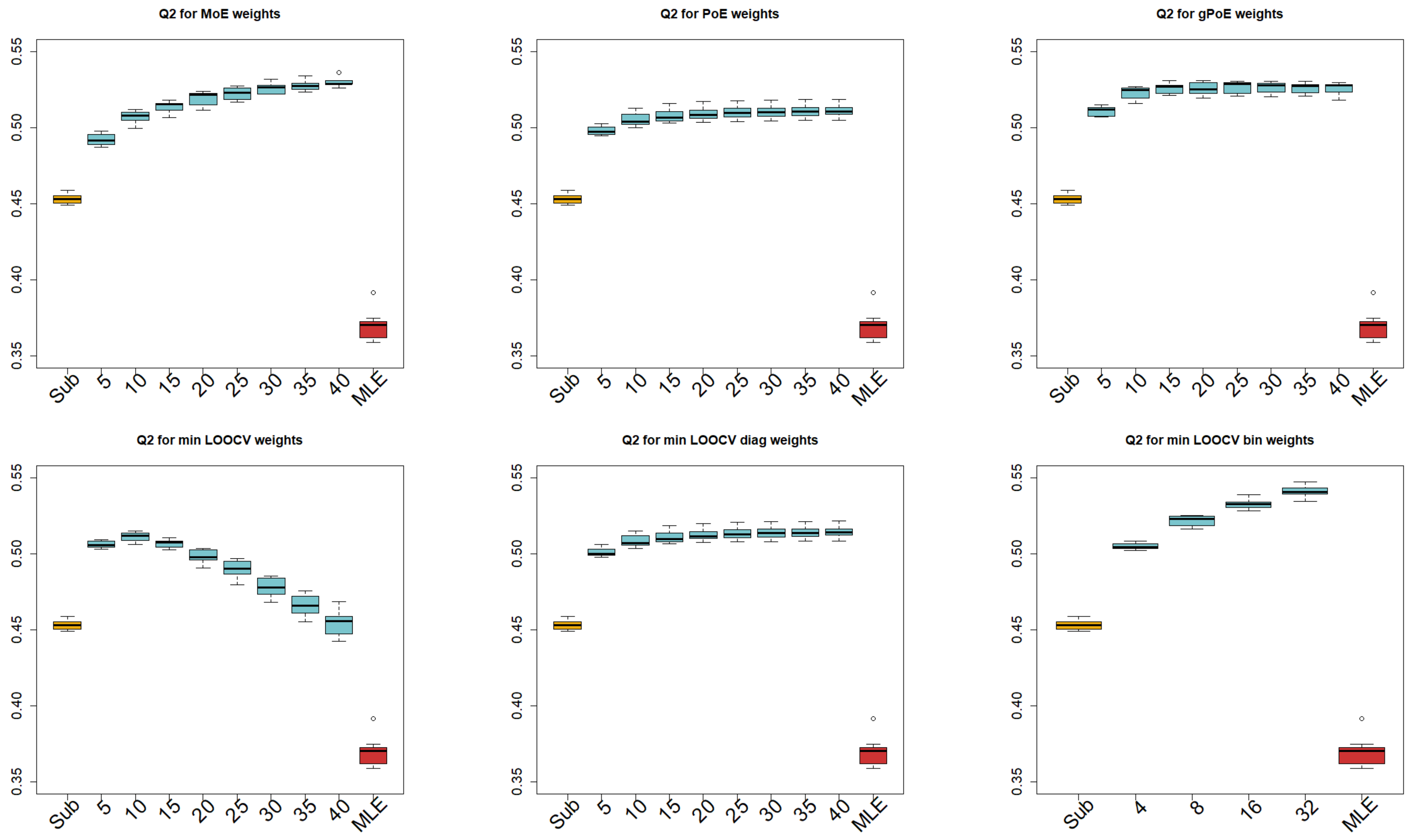
  - 48 variables de design,
  - 300 points d'apprentissage,
  - 327 points test,
  - 2 objectifs et 413 contraintes (un modèle est construit pour seulement 190 des contraintes).



CHOICE OF THE WEIGHTS



CHOICE OF THE WEIGHTS



COMBINATION OF KRIGING MODELS

→ We propose a model which is a combination of Kriging models with random length-scales

$$M_{tot}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i(\boldsymbol{x}) M_i(\boldsymbol{x}),$$

with $M_i(x) = \mu_i + k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) K_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{Y} - \mu_i)$ Kriging model with fixed length−scale vector $\boldsymbol{\theta}_i$.

3) Variance of the combination

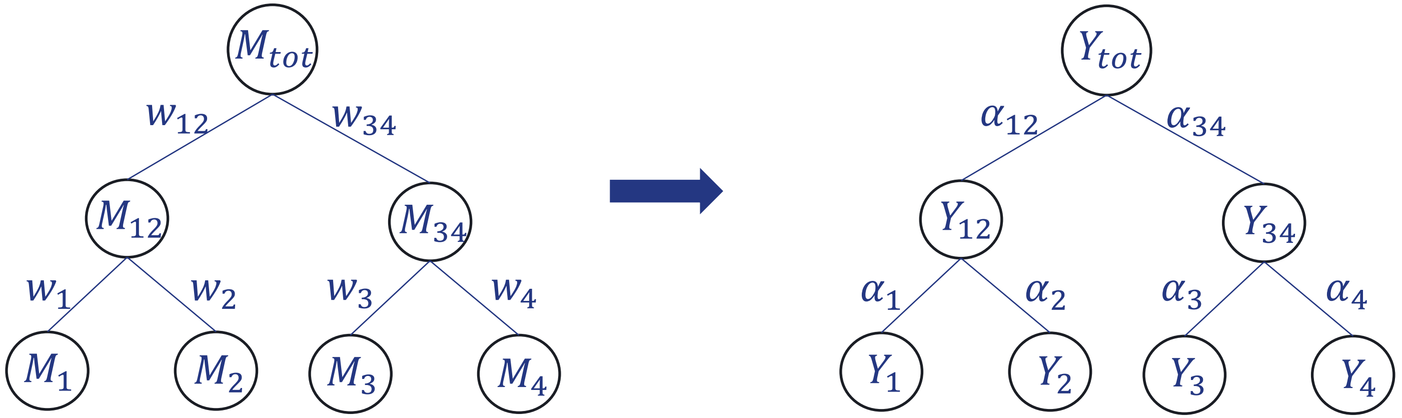
- To obtain the variance of the combination, we add the hypothesis that **the underlying Gaussian Process Y_{tot} is a combination (with different weights) of independent Gaussian Processes**:

$$Y_{tot} = \sigma_{tot}^2 \sum_{i=1}^{p} \alpha_i Y_i \,, \qquad \text{with } Y_i \sim \mathcal{GP}\left(\mu_i, r_{\boldsymbol{\theta}_i}(.,.)\right), \qquad \sum_{i=1}^{p} \alpha_i = 1, \qquad \text{and } \sigma_{tot}^2 \text{ the variance of the GP.}$$

Thus, the covariance of this GP is:

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

- 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** with respect to $Y_{tot} = \alpha Y_1 + (1-\alpha)Y_2$:

$$\alpha^* = \arg\min_{\alpha} \mathbf{E}\left[\mathbf{E}\left[\left(wM_1(\boldsymbol{x}) + (1-w)M_2(\boldsymbol{x}) \; - \alpha Y_1(\boldsymbol{x}) + (1-\alpha)Y_2(\boldsymbol{x})\right)^2 \mid Y_1, Y_2\right]\right].$$

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

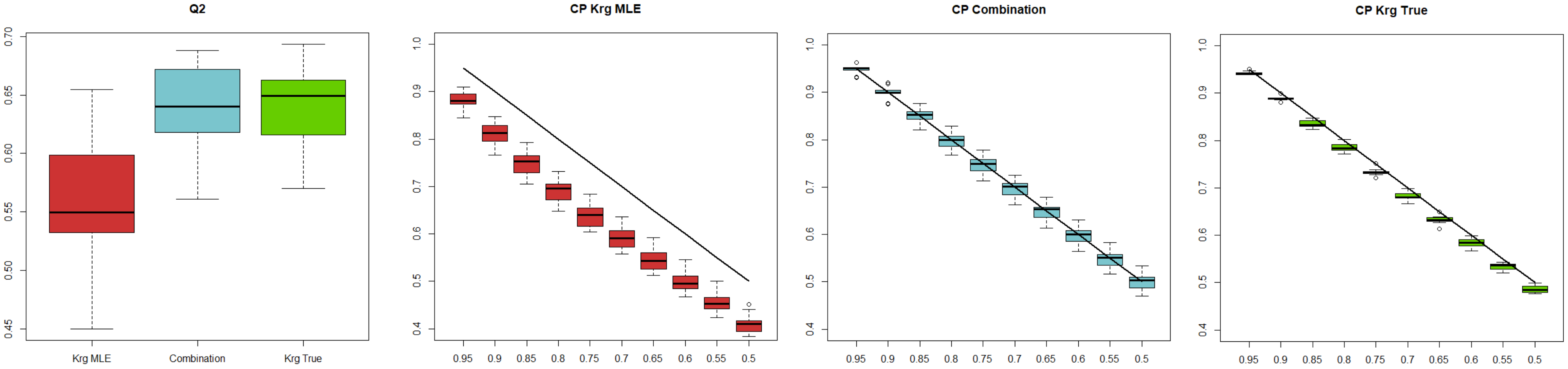
$$\alpha^* = \frac{a_1(w)}{a_1(w) + a_2(w)}, \qquad \text{with:} \begin{cases} a_1(w) = w^2\mathbf{E}(e_{LOOCV}(M_1)|Y_2) + (1-w^2)\mathbf{E}(e_{LOOCV}(M_2)|Y_2), \\ a_2(w) = (1-w)^2\mathbf{E}(e_{LOOCV}(M_2)|Y_1) + (1-(1-w)^2)\mathbf{E}(e_{LOOCV}(M_1)|Y_1). \end{cases}$$

Finally, the variance of the combination is obtained as:

$$\hat{s}^2(\boldsymbol{x}) = \mathbf{Var}(Y_{tot}(\boldsymbol{x})|\mathcal{D}) = k_{tot}(\boldsymbol{x},\boldsymbol{x}) - k_{tot}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{K}_{tot}(\boldsymbol{X},\boldsymbol{X})^{-1}k_{tot}(\boldsymbol{X},\boldsymbol{x}).$$

RÉSULTATS NUMÉRIQUES – DONNÉES SIMULÉES

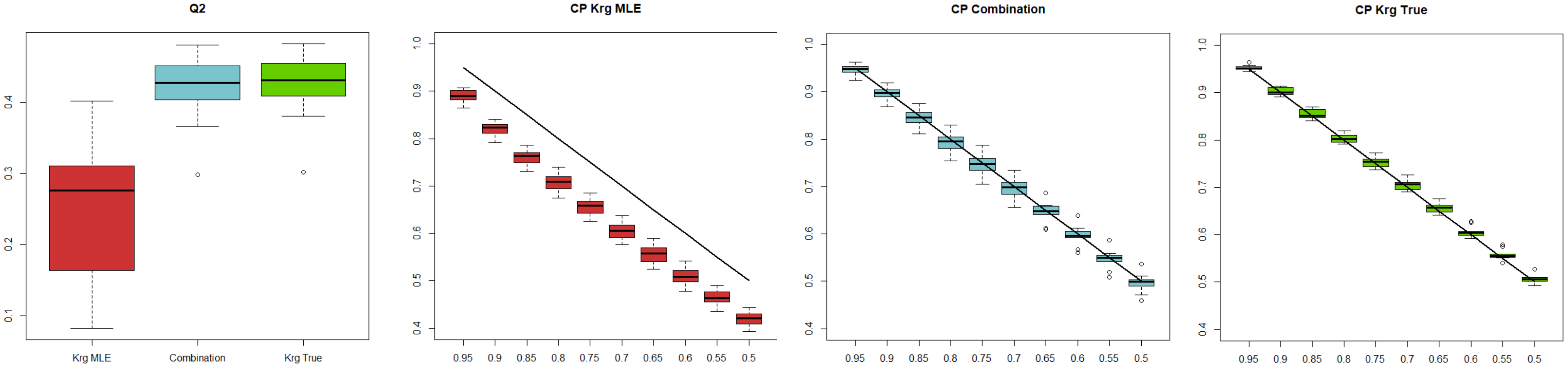
$$\theta_{true} = 3$$



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

TRUST REGIONS

- Trust regions are a class of methods to reduce the search space around interesting regions (see for example Eriksson et al., 2019, Diouane et al., 2023).

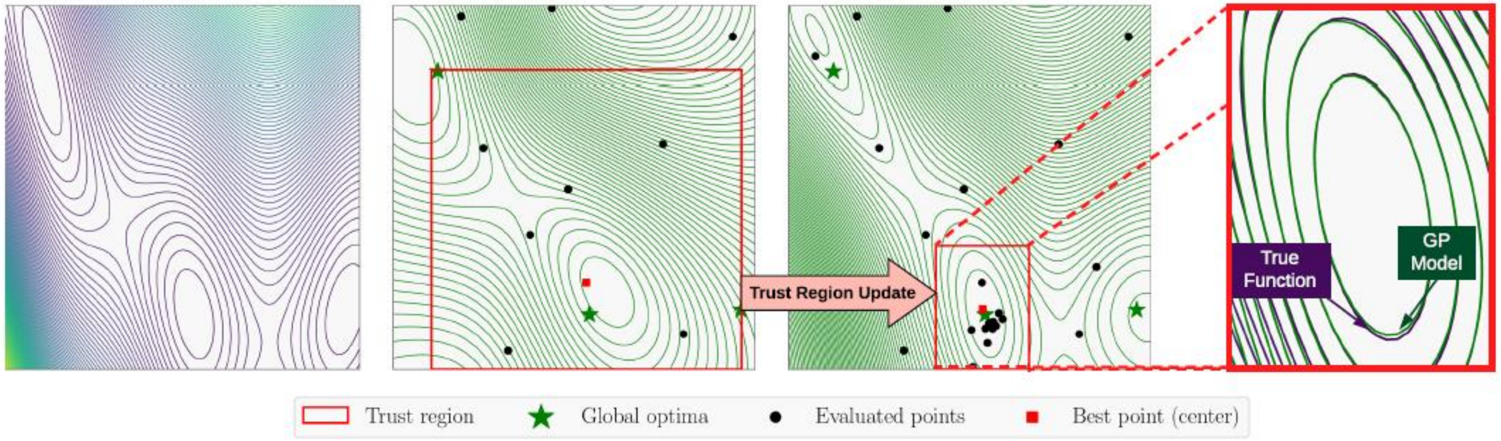


Figure from Eriksson et al., 2019:
Illustration of TuRBO for the 2D Branin function.

- A trust region is built around the current best value observed.
- The acquisition function is optimized inside the trust region to select new points.
- After a given number of consecutive failures (no improvement) → The size of the trust region is reduced.
- OR, after a given number of consecutive success (the best value was improved) → The size of the trust region is increased.

- In the benchmark, we use the **TREGO** implementation of trust regions (see Diouane et al., 2023).
- In TREGO, we alternate between global iterations of EGO, and local iterations inside a trust region.

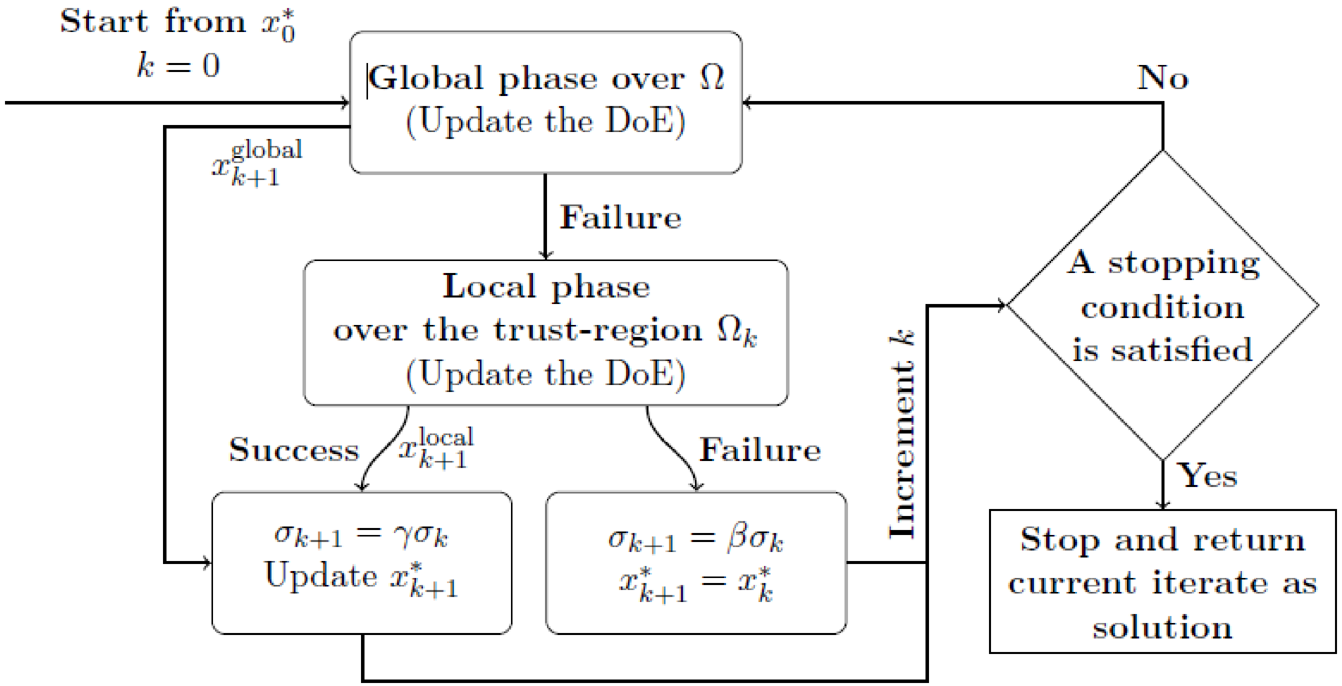
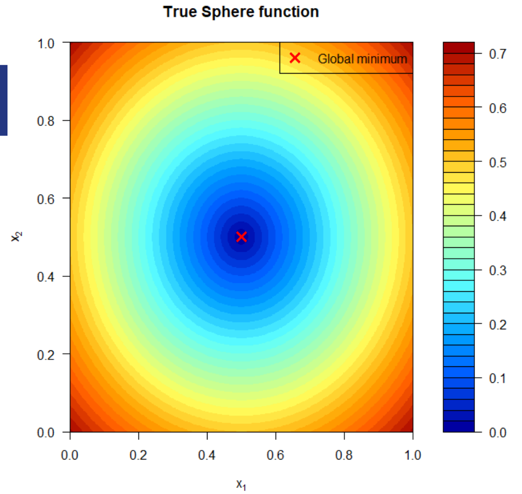


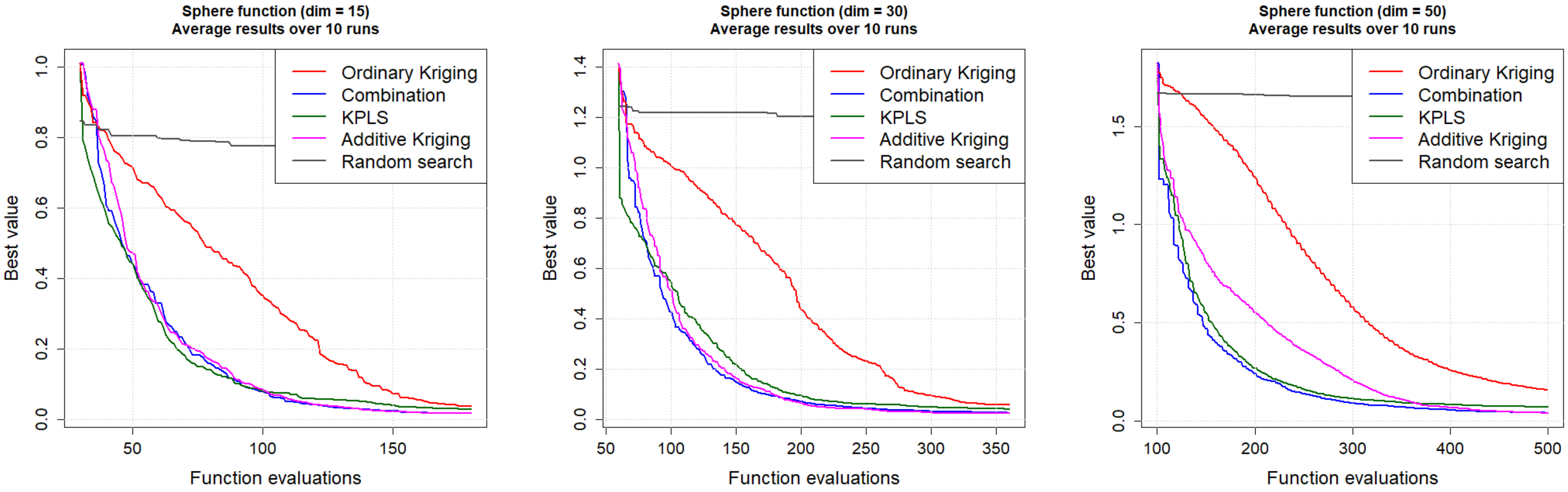
Figure from Diouane et al., 2023 :
Overview of the TREGO framework.



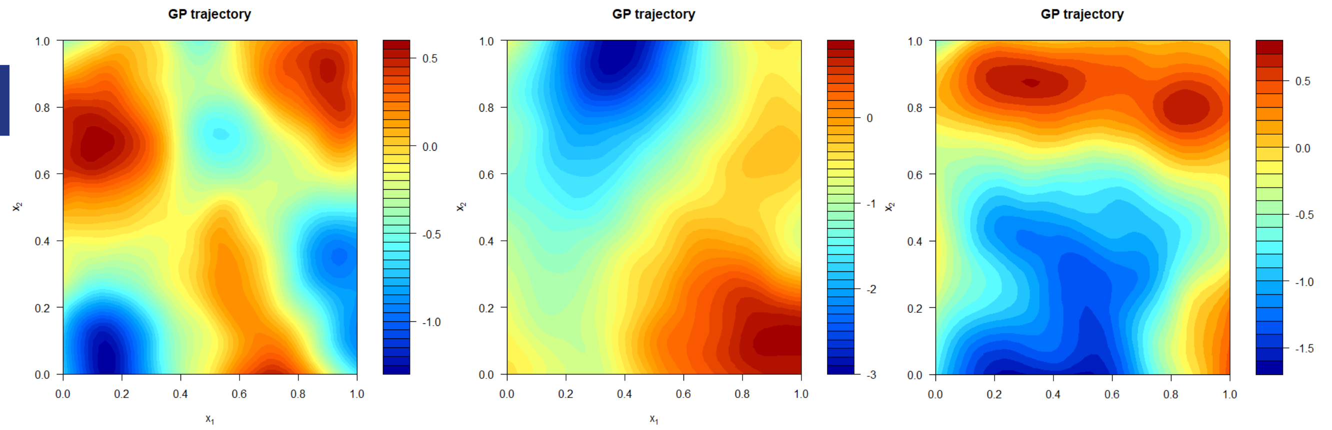
- Fonction sphère : $$f_{sphère}(x_1, \dots, x_d) = \sqrt{\sum_{i=1}^{d} (x_i - 0{,}5)^2}\,, \qquad 0 \leq x_i \leq 1.$$

→ Déceptivement difficile à modélisé avec des GPs quand on dispose de peu d'observations.

→ Facile à optimiser (fonction convexe).



RÉSULTATS NUMÉRIQUES

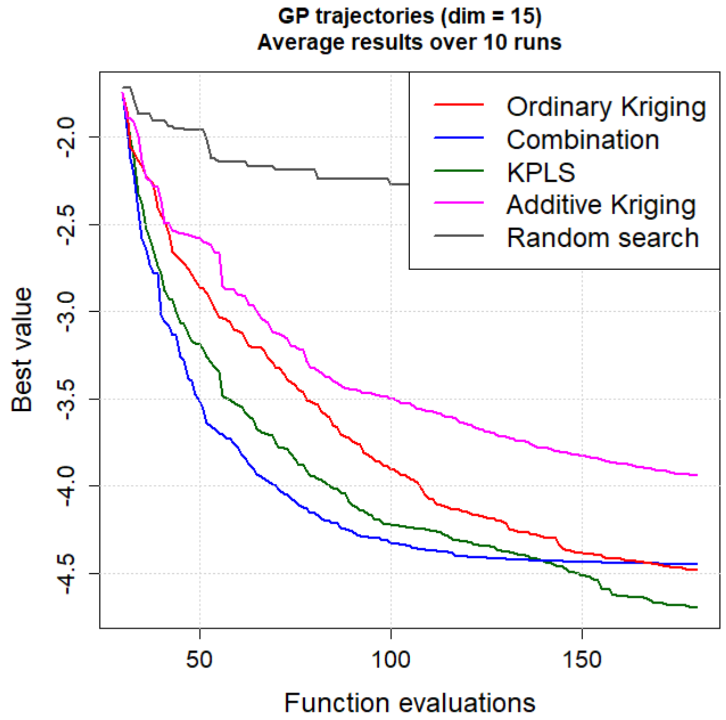


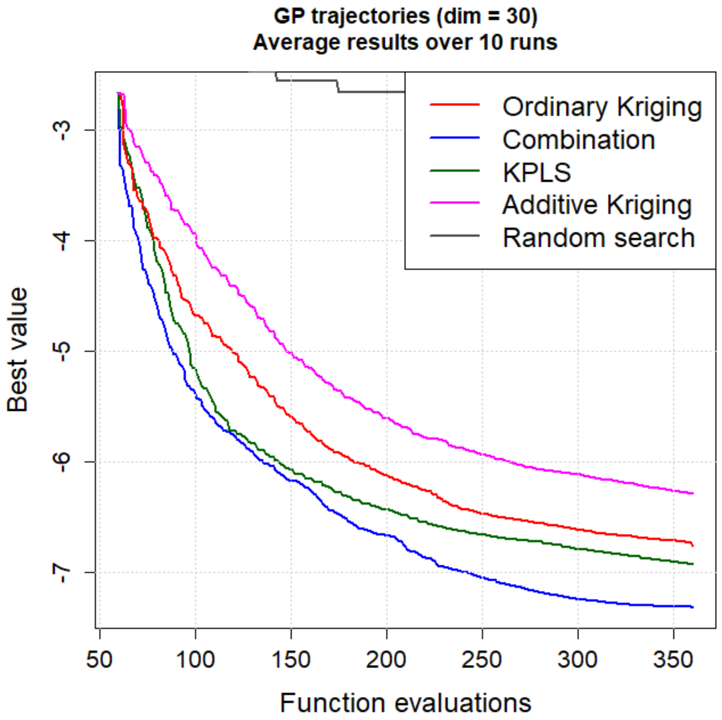
- Trajectoires de GPs : $f_{GP}(.) \sim GP(\mathbf{0}, k_\theta(.,.))$,

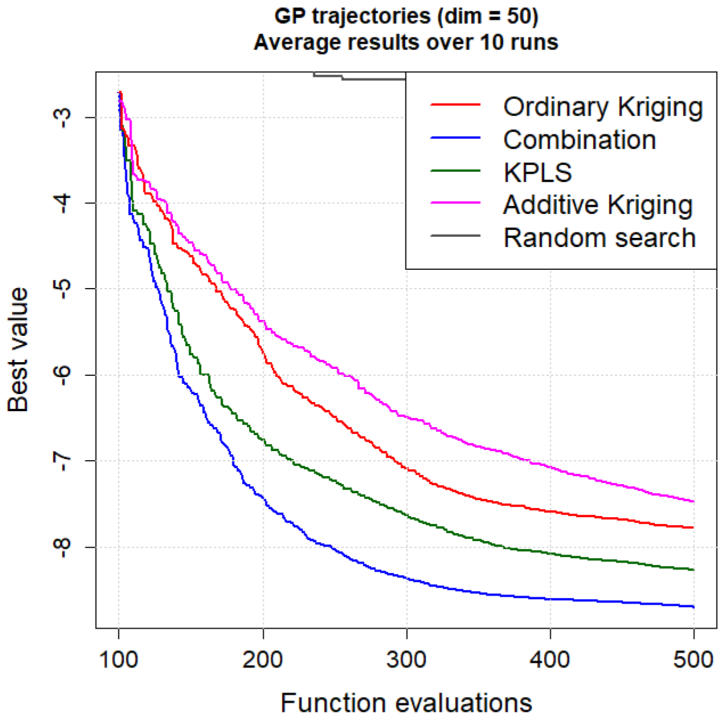
k_θ est un noyau Matérn 5/2 isotrope de portée $\theta = \sqrt{\frac{d}{12}}$

→ Plus difficile à optimiser (multimodale) et plus représentatif des fonctions rencontrées en ptratique.

→ Cas où l'hypothèse de Krigeage est vérifiée.

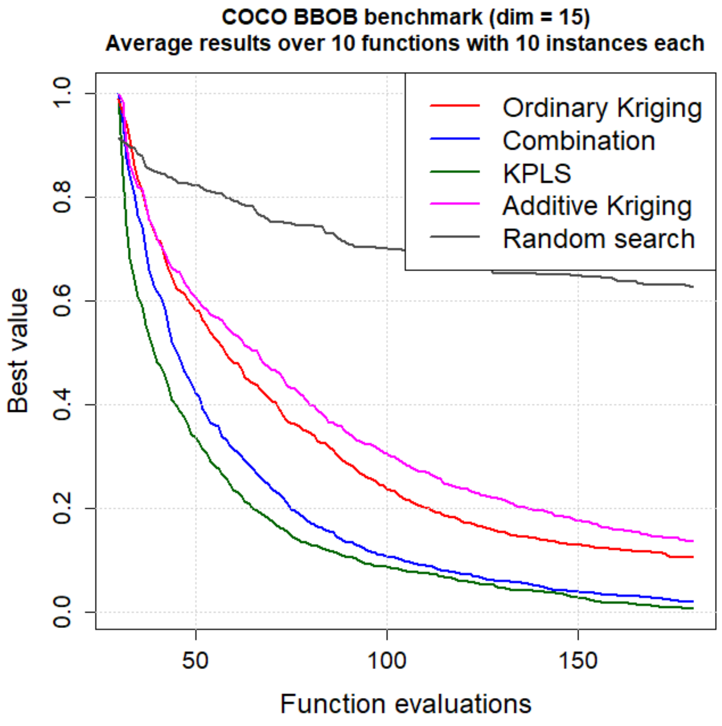


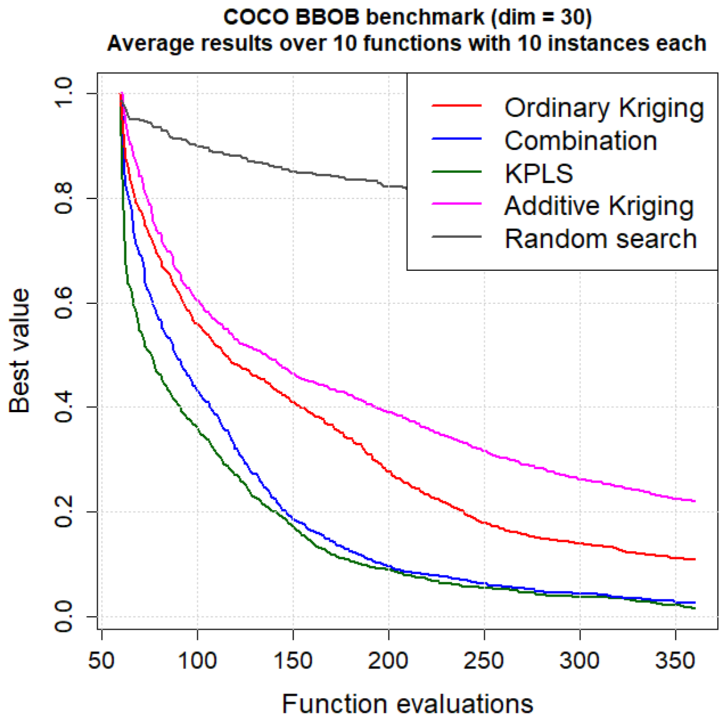


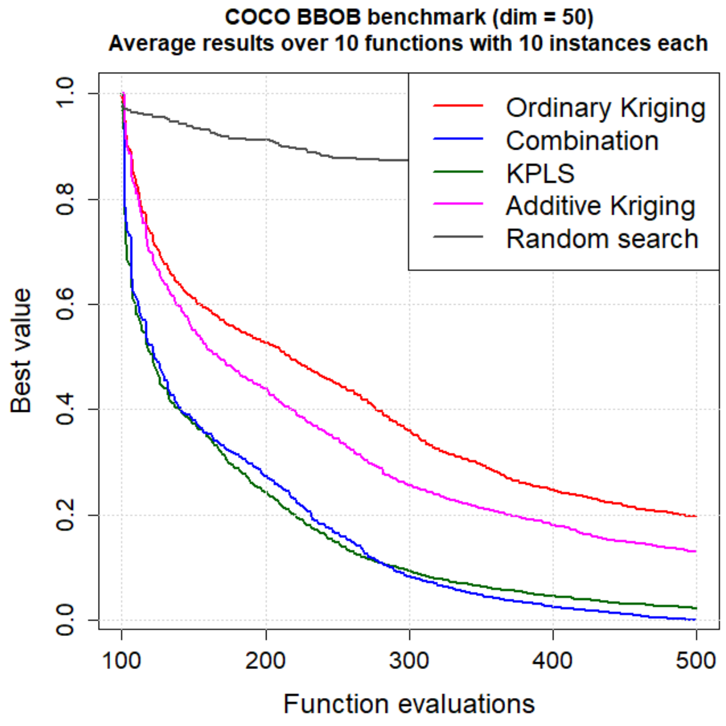


- COCO BBOB benchmark : 9 fonctions multimodales du benchmark (fonctions f15 à f22 et f24)

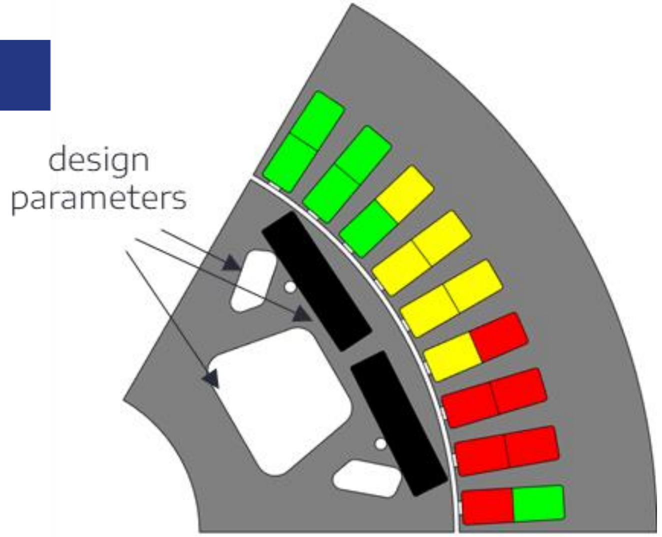
→ Difficile à optimiser (fonctions très multimodales).



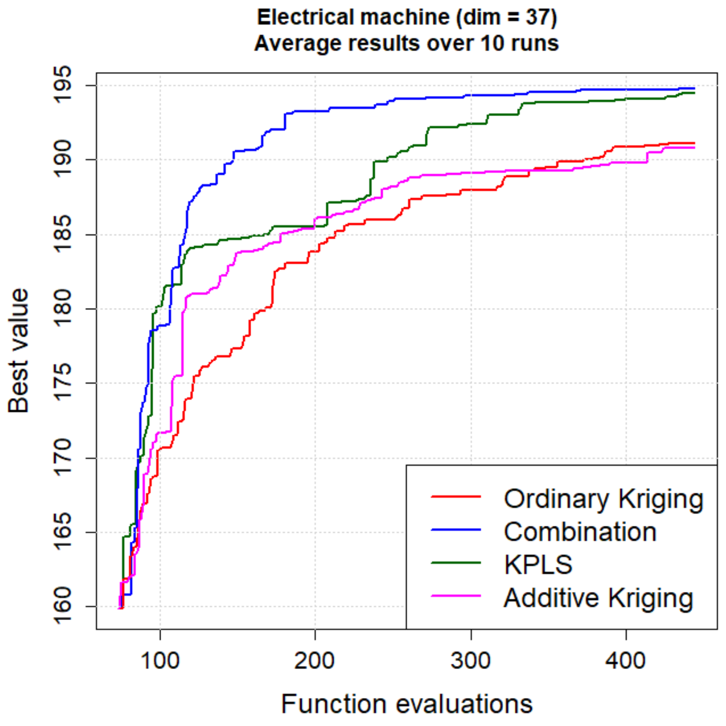




RÉSULTATS NUMÉRIQUES



- Machine électrique (dimension 37)
- Le problème complet comporte 2 objectifs et 10 contraintes. Ici on optimise simplement sur une contrainte (vitesse maximale du véhicule) qui est suffisamment complexe (multimodale) dont de fortes valeurs correspondent globalement à de bonnes machines.



Thank you for your attention !

Contact :

Tanguy APPRIOU
tanguy.appriou@stellantis.com