Active subspace embedding for parsimonious gaussian process surrogates

Amandine MARREL*, Bertrand IOOSS‡

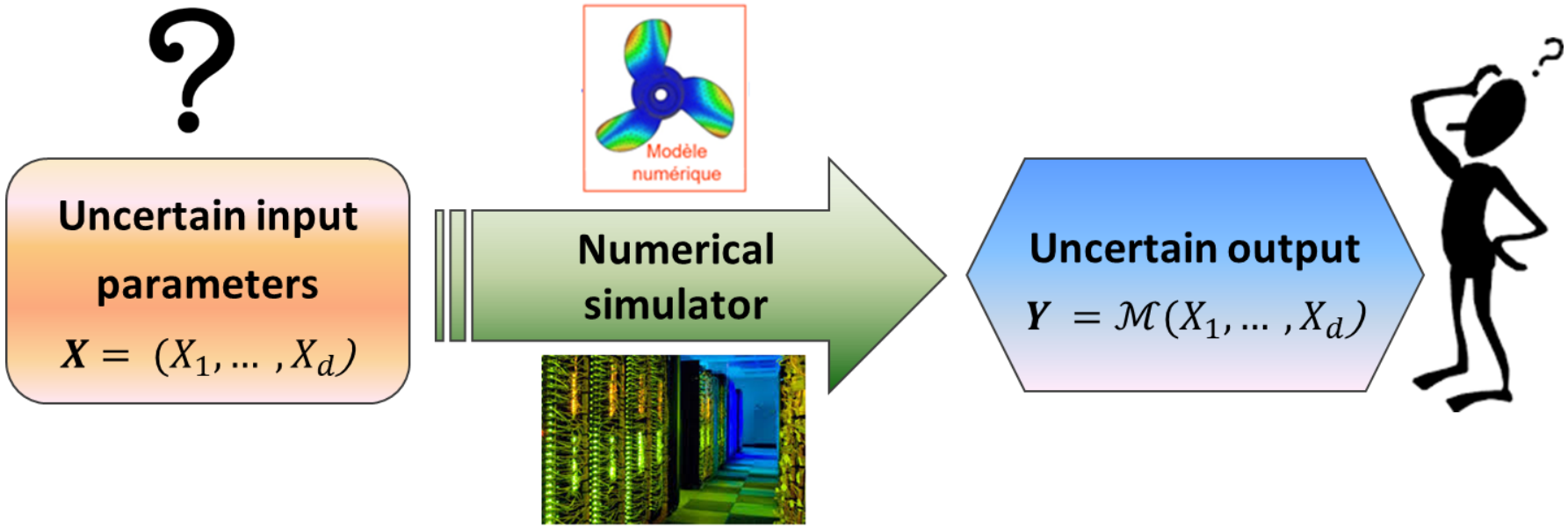
* CEA Energy Division, IRESNE, DER, Cadarache, France

‡ EDF R&D, Chatou, France

ENBIS Conference 2026, Florence, Italy.

Risk assessment in nuclear accident analysis

- **Safety studies:** compute a failure risk (margins, rare events) with validated computer/numerical models
- **Numerical simulators:** fundamental tools to understand, model & predict physical phenomena
- **Large number of input parameters,** related to physical and numerical modelling
- **Uncertainty on some inputs → uncertainty on output & safety margins**



Risk assessment in nuclear accident analysis

- **How to deal with uncertainties in numerical simulation?**

→ Probabilistic framework and Monte Carlo-based methods

→ **CPU-expensive simulator** ⇒ Use of a **surrogate model*** **to propagate input uncertainties**

→ **Applicative constraints/framework:**

✓ **Given data**: a single i.i.d. inputs/output **sample** $\left(\boldsymbol{x}^{(i)}, \boldsymbol{y}_i\right)_{1 \leq j \leq n}$ **where** $\boldsymbol{y}_i = \mathcal{M}\left(\boldsymbol{x}^{(i)}\right)$ → to be used for multi-purpose *(sensitivity analysis, uncertainty propagation... And **training a metamodel**)*

→ n-dataset denoted: $\mathcal{D}_n = (\mathbf{X}_s, \mathbf{Y}_s)$

✓ **Small sample size**: $n \approx$ 100 to 1000 simulations

✓ **Large number of uncertain inputs**: $d \approx$ 10 to 100 inputs

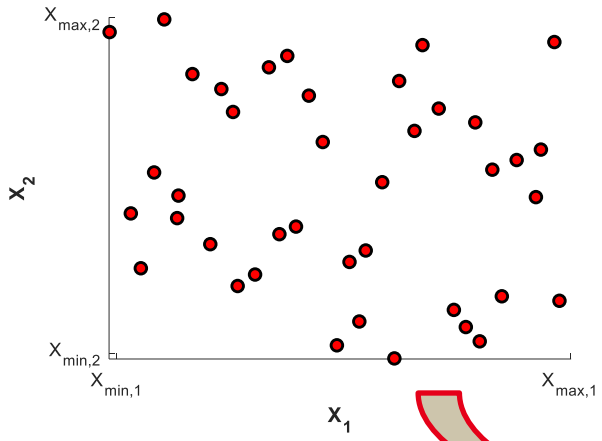
✓ **Required predictive uncertainty associated to each metamodel prediction**

→ Gaussian Process Regression (GPR): particularly well-suited tool ⇒ Very popular

*Equivalently referred to as a metamodel, and based on machine learning techniques.

Crucial use of GPR metamodel

Design of numerical experiments



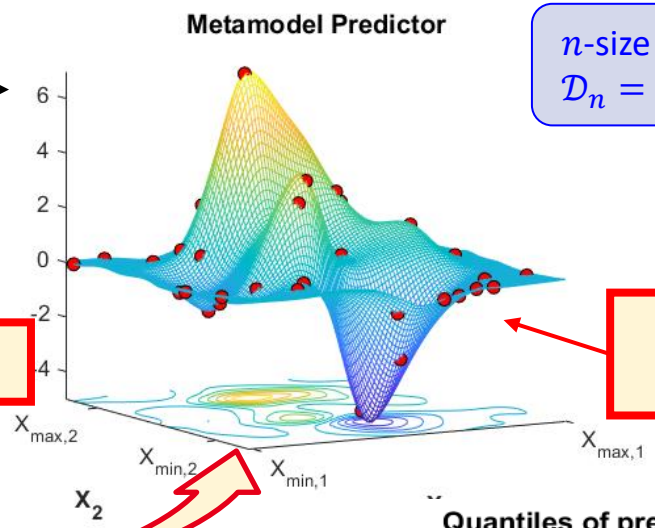
Incertain inputs domain

Simulator CPU++
$Y = \mathcal{M}(X_1, \ldots, X_d)$

Approximation with GPR

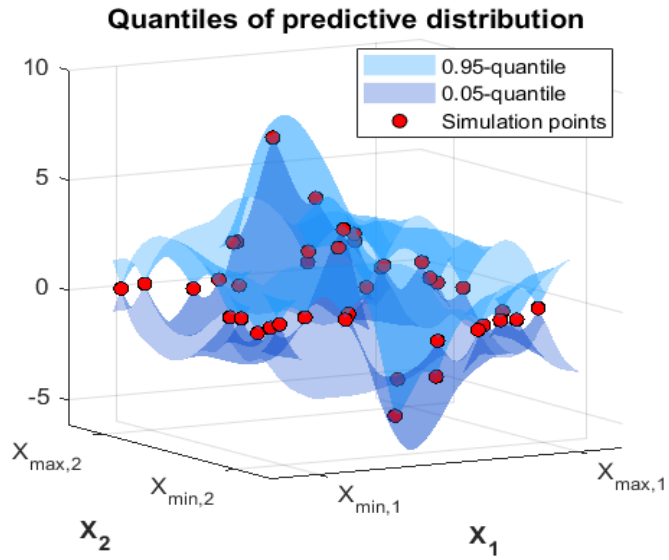
Metamodel: $Y_{app} = \widehat{\mathcal{M}}(X) \approx \mathcal{M}(X)$

Analysis of simulator outputs



n-size Dataset
$\mathcal{D}_n = (\mathbf{X}_s, \mathbf{Y}_s)$

Probabilistic metamodel



✓ Build from the n-size dataset $\mathcal{D}_n = (\mathbf{X}_s, \mathbf{Y}_s)$, **GPR mimics the true model $\mathcal{M}$**, providing a **GP predictive distribution** for each new evaluation point

⇒ **How to deal with large number of inputs *d*?**

How to build parsimonious GPR in large dimension?

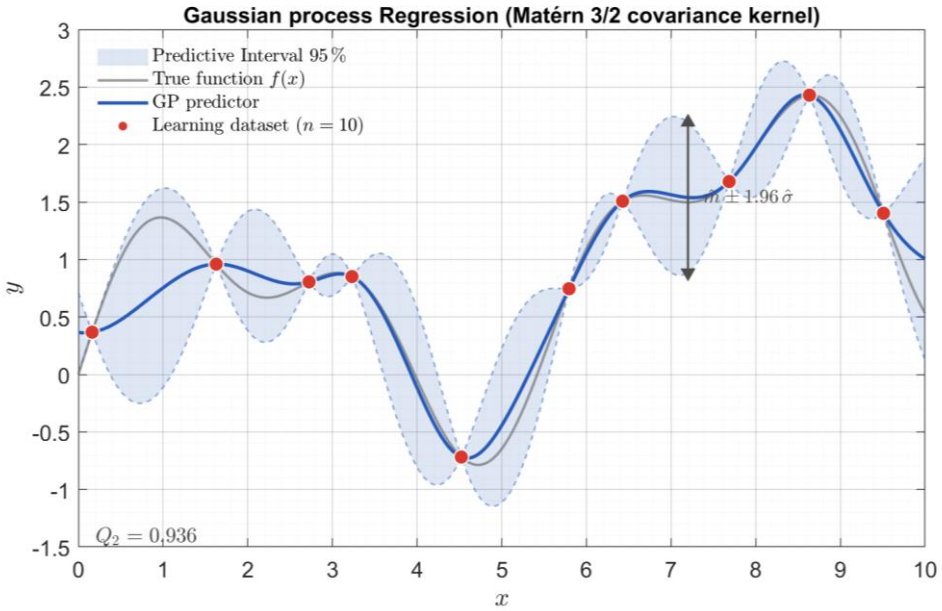
1. Reminds on Gaussian Process Regression

Reminders on GPR

▶ **Probabilistic surrogate model**: response considered as a realization of a random GP [RW06,Gra21]

$$Y(\boldsymbol{x}) \sim GP(\mu(\boldsymbol{x}), k(\boldsymbol{x}', \boldsymbol{x}))$$

With $\mu(\boldsymbol{x})$ the mean and $k(\boldsymbol{x}', \boldsymbol{x})$ the covariance function.



⇒ **<u>Predictive</u> GP** is the GP conditioned by the dataset $\mathcal{D}_n$:

$$Y(\boldsymbol{x}^*)_{|\mathcal{D}_n} \sim GP\left(\hat{y}(\boldsymbol{x}^*), \hat{s}(\boldsymbol{x}', \boldsymbol{x}^*)\right)$$

With <u>analytical formulations</u> for $\hat{y}(\boldsymbol{x}^*)$ *and* $\hat{s}(\boldsymbol{x}', \boldsymbol{x}^*)$

⇒ Conditional mean $\hat{y}(\boldsymbol{x}^*)$ serves as the **predictor** at location $\boldsymbol{x}^*$

⇒ Conditional covariance $\hat{s}(\boldsymbol{x}^*, \boldsymbol{x}^*)$ gives the predictive variance at location $\boldsymbol{x}^*$

⇒ Predictive interval of any level α can be built at any location $\boldsymbol{x}^*$

Recommendations for parametric choices

▶ **In practice:** standard parametric choices for trend function μ and covariance function k

$$Y(\boldsymbol{x}) \sim GP(\mu(\boldsymbol{x}), k(\boldsymbol{x}', \boldsymbol{x}))$$

$\Rightarrow$ For μ: either **constant** or linear basis

$\Rightarrow$ For k: stationary covariance built-upon tensorized 1-D covariance functions of ν-Matérn

1-Dim ⟶ $k_{\sigma,\nu,\theta}(x,\tilde{x}) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}h}{\theta}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}h}{\theta}\right)$ ⟶ **3/2 or 5/2 Matérn covariances** offer good properties and « intermediate » regularity

d-Dim ⟶ $k_{\sigma,\nu,\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}}) = \sigma^2 \prod_{i=1}^{d} k_{1,\nu,\theta_i}(x_i - \tilde{x}_i)$ with $h = |x - \tilde{x}|$ ⇒ Hyperparameters $\boldsymbol{\theta} \in \mathbb{R}^{+,d}$

	$v = \frac{1}{2}$	$v = \frac{3}{2}$	$v = \frac{5}{2}$	$v = +\infty$
Usual name	exponential	3/2-Matérn	5/2-Matérn	Gaussian
$k_{\sigma,\nu,\theta}(x,\tilde{x})$	$\sigma^2 e^{-\frac{h}{\theta}}$	$\sigma^2(1+\sqrt{3}\frac{h}{\theta})e^{-\sqrt{3}\frac{h}{\theta}}$	$\sigma^2\left(1+\sqrt{5}\frac{h}{\theta}+\frac{5}{3}\left(\frac{h}{\theta}\right)^2\right)e^{-\sqrt{5}\frac{h}{\theta}}$	$\sigma^2 e^{-\frac{1}{2}\left(\frac{h}{\theta}\right)^2}$
Differentiability of GP trajectories	$\mathcal{C}^0$	$\mathcal{C}^1$	$\mathcal{C}^2$	$\mathcal{C}^\infty$

$\Rightarrow$ Additional nugget effect : $\tilde{k}(\boldsymbol{x}', \boldsymbol{x}) = k(\boldsymbol{x}', \boldsymbol{x}) + \tau^2 \delta_{x=x'}$ ($\rightarrow$ nugget hyperparameter τ)

Estimation of GP hyperparameters (θ, τ)

$\Rightarrow$ How to robustly estimate the **hyperparameters** $(\boldsymbol{\theta}, \boldsymbol{\tau}) \in \mathbb{R}^{+,d+1}$ from the learning sample?

For good **predicitivity** + **reliable prediction intervals** $\Rightarrow$ Crucial for safety applications

- **Usual estimation methods** [Review in MI24a]
  - $\rightarrow$ **Maximum likelihood**
  - $\rightarrow$ **Cross-validation Mean Squared Error**
  - $\rightarrow$ **Bayesian approaches**

New multi-objective estimation algorithm for more reliable GP in [MI24b]

$\Rightarrow$ Reliable results observed in practice **up to a max dimension of** $\boldsymbol{d = 20}$

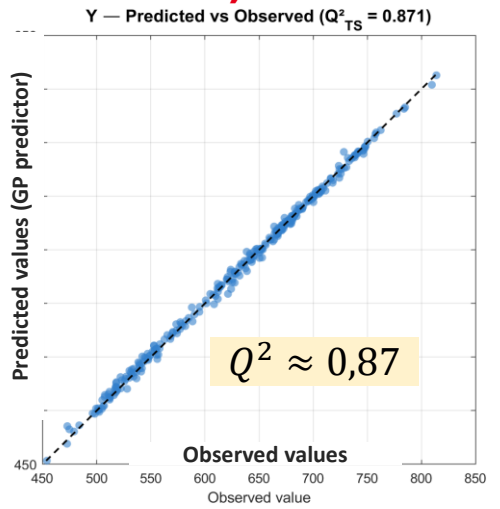
Validation metrics → by cross-validation (LOO or K-fold CV)

Accuracy of the GPR predictor (only) [DIG+22]

$$Q^2 = 1 - \frac{MSE_{LOO}}{n^{-1}\sum_{i=1}^{n}\left(y_i - n^{-1}\sum_{i=1}^{n} y_i\right)^2}$$

with $\mathrm{MSE}_{LOO} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{-i}(\mathbf{x}^{(i)})\right)^2$

where $\hat{y}_{-i}(\mathbf{x}^{(i)})$ is the GPR predictor in $\mathbf{x}^{(i)}$ when $(\mathbf{x}^{(i)}, y_i)$ is removed from the learning sample $\mathcal{D}_n$



→ ***The closer to 1 the Q^2, the better the accuracy of the predictor.***

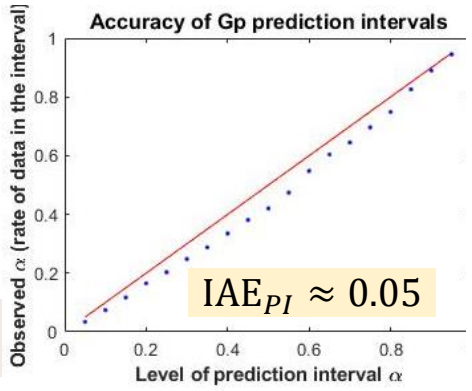
Accuracy of the whole GP predictive distribution [DIG+22, MI24a]

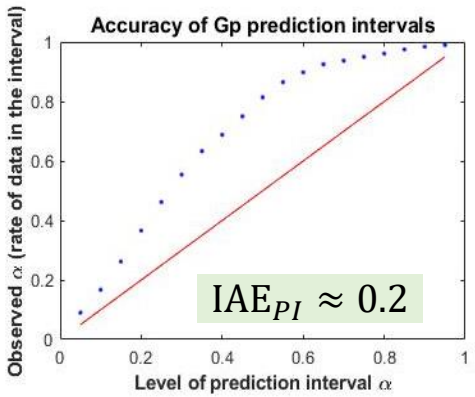
From empirical coverage function for $\alpha \in [0,1]$: $\hat{\Delta}(\alpha) = n^{-1}\sum_{i=1,\ldots,n} \mathbf{1}\{y_i \in PI_{\alpha,-i}(\mathbf{x}^{(i)})\}$

with $PI_{\alpha,-i}(\mathbf{x}^{(i)})$ the α-level GP predictive interval for $\mathbf{x}^{(i)}$ when $(\mathbf{x}^{(i)}, y_i)$ is removed from $\mathcal{D}_n$

⇒ α-PI plot and integrated absolute error on $\hat{\Delta}(\alpha)$:

$$IAE_{PI} = \int_0^1 \left|\hat{\Delta}(\alpha) - \alpha\right|$$





→ ***The closer to 0 the IAE_{PI}, the better calibrated the predictive intervals.***

Why and how to reduce dimension?

Objectives :

- Simplify and **improve reliability of GPR** hyperparameter estimation
- **Reduce computational burden** of Bayesian inference with GPR (UQ, calibration...)
- Find a **lower-dimensional input space** preserving most output variability, and potentially better suited to GPR stationarity assumptions

Under our applicative constraints

- ✓ Given data: single i.i.d. sample $\mathcal{D}_n$
- ✓ Small sample size: $n \approx 100$ to 1000
- ✓ Large input dimension: d ≈ 10 to 100
- ✓ **Gradient-free setting:** black-box simulator

⇒ Proposed approach: HSIC screening + Active Subspaces (AS) + GPR [MI26]

How to build parsimonious GPR in large dimension?

1. Reminds on Gaussian Process Regression

2. Proposed 4-step methodology → **HSIC + AS + GPR**

Proposed 4-step methodology

1st dim. reduction : $d \to d_s$

2nd dim. reduction : $d_s \to q^*$

Total dim. reduction: $d \to d_s \to q^*$

INPUT

- **Learning sample** $\mathcal{D}_n$ (n obs., d inputs)
- Independent test sample $\mathcal{D}_{\text{test}}$

d **inputs**

STEP 1: HSIC-based screening

Select the d_s most influential inputs via HSIC independence tests (test level to have $d_s \approx 15$–20)

$\longrightarrow$ Screened dataset: $\mathcal{D}_n^{\text{screen}} = (\mathbf{X}_s^{\text{screen}}, \mathbf{Y}_s)$ with $\mathbf{X}_s^{\text{screen}} \in \mathbb{R}^{n \times d_s}$

$d \to d_s$

STEP 2: Initial GPR and estimation of $\hat{\mathbf{C}}$

Fit a GP on screened data: estimate GP hyperparameters and obtain GP posterior $Y \mid \mathcal{D}_n^{\text{screen}}$

Estimate gradient covariance matrix $\mathbf{C}$ via Monte Carlo integration $\longrightarrow \hat{\mathbf{C}}$

STEP 3: Active Subspace identification

3a. Eigendecomposition of $\hat{\mathbf{C}}$:

Solve $\hat{\mathbf{C}}\,\mathbf{w}_i = \lambda_i\,\mathbf{w}_i$ and order $\lambda_1 \geq \cdots \geq \lambda_{d_s}$

Obtain active direction matrix:

$\longrightarrow \mathbf{W}_1 = [\mathbf{w}_1, \ldots, \mathbf{w}_{d_s}] \in \mathbb{R}^{d_s \times d_s}$

3b. Select optimal reduced dimension q^*

Sequential GPR fits on first j active directions; compute CV-$Q^2(j)$

$\longrightarrow$ **Select q^* via CV-Q^2 stagnation**

$d_s \to q^*$

STEP 4: Final reduced GPR on q^*-dimensional active subspace

- **Project inputs** onto optimal active subspace: $\tilde{\mathbf{X}}_s = \mathbf{X}_s^{\text{screen}}\,\mathbf{W}_{1,*}$ with $\mathbf{W}_{1,*} = [\mathbf{w}_1, \ldots, \mathbf{w}_{q^*}]$
- **Fit final GPR** on $\tilde{\mathcal{D}}_n = (\tilde{\mathbf{X}}_s, \mathbf{Y}_s)$ and validate on test set (Q^2, IAE_{PI})

OUTPUT

GP posterior $Y \mid \tilde{\mathcal{D}}_n$ on active subspace $\mathbf{W}_{1,*}^\top \mathbf{x}^{\text{screen}}$ $\Rightarrow$ **Dim. reduction:** $d \to d_s \to q^*$

Step 1: HSIC-based screening

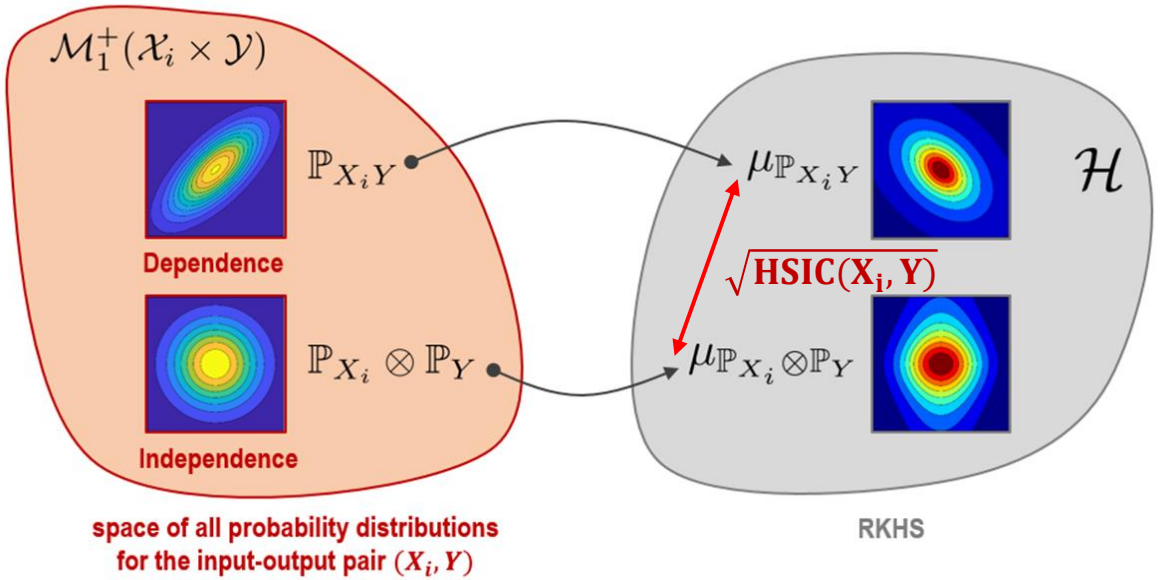
1st dimension reduction to obtain a reduced **subset of $d_s \ll d$ inputs**, to fit an initial GPR (step2)
$\rightarrow$ **Typically $d_s \lesssim 20$**

HSIC-based SCREENING: Surrogate-free + gradient-free + given-data and small-data context

▶ **HSIC*** [GFT+07, Dav15] measures dependence between X_i and Y via kernel embeddings in RKHS

$$HSIC(X_i, Y) = \left\| \mu_{\mathbb{P}_{X_i Y}} - \mu_{\mathbb{P}_{X_i} \otimes \mathbb{P}_Y} \right\|^2$$

- ✓ **Vanishes iff $X_i \perp\!\!\!\perp Y$** with characteristic kernels
  $\rightarrow$ **Independence test** for screening
- ✓ No distributional assumption, any input type
- ✓ Estimation from a unique random sample, efficient in practice from $n \sim 100$ (U-stat and V-stat estimators)



Extract from a presentation by G. Sarazin (CEA)

***Standard kernel choice** for real variables: **Gaussian kernels** with bandwidth = empirical standard deviation $\rightarrow$ characteristic kernels*

Hilbert-Schmidt **I**ndependence **C**riterion

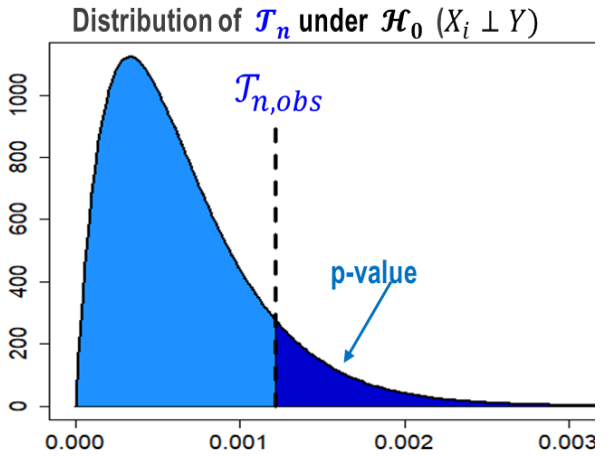
Step 1: HSIC-based screening

▶ **HSIC-based independence tests** [GFT+07]: $HSIC(X_i, Y) = 0 \Leftrightarrow X_i \perp\!\!\!\perp Y$ (*with characteristic kernels!)*

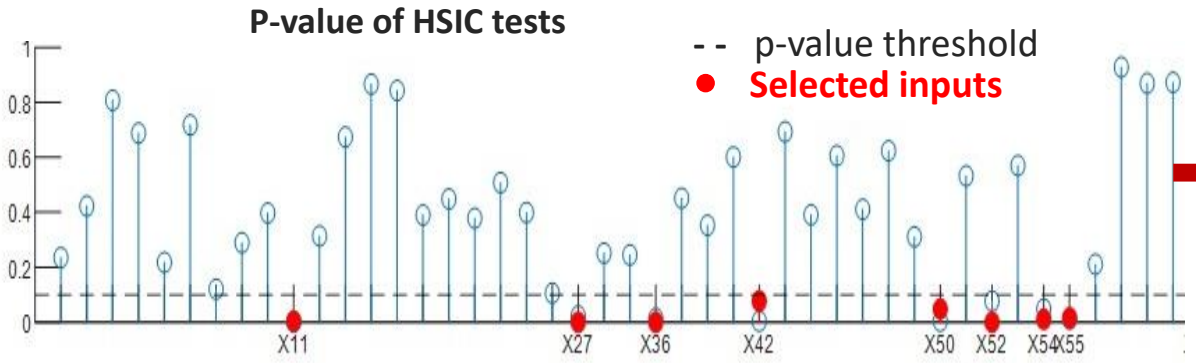
- ✓ **Null hypothesis**: $\mathcal{H}_0 : X_i \perp\!\!\!\perp Y$ against $\mathcal{H}_1 : X_i \not\perp\!\!\!\perp Y$
- ✓ **Test statistics**: $\mathcal{T}_n := n\widehat{\mathrm{HSIC}}(X_i, Y)$
- ✓ **Decision rule**: $\mathcal{H}_0$ rejected iff $\mathcal{T}_n > q_{1-\alpha}$ with $q_{1-\alpha}$ the quantile of $\mathcal{T}_n$ under $\mathcal{H}_0$

  $\Leftrightarrow$ iff *p-value* $\leq \alpha$

  $\rightarrow$ Procedures to approximate the distribution of $\mathcal{T}_n$ under $\mathcal{H}_0$ [GFT+07, EM24]



⇒ Use for screening of inputs



Select inputs with *p-value* $\leq \alpha$ (e.g. $\alpha \approx 0.10$–0.15, or adaptive to retain $d_s \lesssim 20$)

$\rightarrow$ ***d_s* explicative inputs** of GPR in Step 2

Step 2: build an initial GPR for AS estimation

- **Fit a GPR on the reduced subset of d_s inputs** (selected by HSIC-test)

  - ✓ **Covariance kernel choice for stable gradient estimation: Matérn-5/2 or Gaussian kernel**

    → The objective is to target average gradient structure, not local variation

    → $\mathcal{C}^1$or $\mathcal{C}^2$-diff kernel k required, depending on estimator of gradient cov. matrix $\hat{C}$ (→ next slide)

  - ✓ **Nugget regularization:** for training matrix only ($\widetilde{\boldsymbol{K}} = \boldsymbol{K} + \tau^2 \mathrm{I_n}$), NOT in base kernel k

    → Remove interpolator property (justified as *(d-d_s)* non-selected inputs)

    → Smoother predictor: reduces sensitivity to local fluctuations

    → Preserves differentiability as it does not modify base kernel $k(\cdot,\cdot)$

Numerical illustration in [MI26]: **high-regularity kernels + nugget yield stable estimation of C**

Step 2: build an initial GPR for AS estimation

- **Estimate the gradient covariance matrix C** [Con15] defined for a model $\mathcal{M}$ as:

$$C = \mathbb{E}_X[\nabla\mathcal{M}(X)\nabla\mathcal{M}(X)^\top] \quad \text{where } \nabla\mathcal{M}(\mathbf{x}) = \left(\frac{\partial\mathcal{M}}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial\mathcal{M}}{\partial x_d}(\mathbf{x})\right)^\top \in \mathbb{R}^d$$

Estimation from GPR: **2 closed-form estimators** from posterior distribution of $\nabla Y|\mathcal{D}_n$ (which remains a GP)

→ **Plug-in estimator** $\hat{C}_{plug-in}$ (requiring $k \in C^1$): only based on the posterior mean of $\nabla Y|\mathcal{D}_n$

$$\hat{C}_{\text{plug-in}} = \mathbb{E}_X[\nabla\hat{y}(X)\nabla\hat{y}(X)^\top]$$

→ **Bayesian estimator** $\hat{C}_{Bayes}$ (requiring $k \in C^2$): based on the full posterior distribution of $\nabla Y|\mathcal{D}_n$

$$\hat{C}_{\text{Bayes}} = \mathbb{E}\left[\mathbb{E}_X[(\nabla Y|\mathcal{D}_n)(\nabla Y|\mathcal{D}_n)^\top]\right]$$

$$\rightarrow \hat{C}_{\text{Bayes}} = \hat{C}_{\text{plug-in}} + \mathbb{E}_X\left[\text{Cov}[\nabla Y|\mathcal{D}_n]\right]$$

Tests in [MI26]: negligible accuracy gain of $\hat{C}_{Bayes}$, $\hat{C}_{plug-in}$ is even more robust ⇒ $\widehat{C}_{plug-in}$ **default choice**

Step 3: identify Active Subspace

- **Eigenvalue decomposition of $\widehat{C} \in \mathbb{R}^{d_s \times d_s}$: $\widehat{C} = W \Lambda W^T$**

  with W the matrix of eigenvectors and Λ the diagonal matrix of ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{d_s} \geq 0$

- **Sequential GPR construction:**

  1. Project input data onto the *j first active directions*: $\boldsymbol{u^{(j)}} = \left(\boldsymbol{W_1^{(j)}}\right)^{\boldsymbol{T}} \boldsymbol{x}$ with $\boldsymbol{W_1^{(j)}} \in \mathbb{R}^{d_s \times j}$
  2. Fit a GPR surrogate on corresponding training dataset $\mathcal{D}_n^{(j)}$ (*see details for kernel below**)
  3. Estimate the predictivity coefficient $Q^2(j)$ via K-fold CV

- **Selection of optimal dimension *q** of active subspace** → *Stagnation criterion of* $Q^2(j)$

  → **Balance parsimony against accuracy of the final reduced GP**

See [MI26] for heuristics

***Kernel notes for sequential GPR:**

- **Nugget** is part of the kernel $\tilde{k} = k + \tau^2 \delta$ (unlike Step 2) to absorb approximation error from restricting to $j < d_s$ directions
- **No constraint here on covariance regularity,** freely chosen to optimize predictive performance

Step 4: Final reduced GPR

- **Final GPR construction on active subspace of dimension *q**** (q^* first active directions)
- **Validation by computation of Q^2 and IAE_{PI}**

Large dataset → Test set available

✓ Compute directly Q^2 and IAE_{PI} on test set

→ Unbiased estimate of generalization performance

Small dataset → outer cross-validation ↺

✓ Fix d_s (Step 1) and q^* (Step 3) from full dataset

✓ Each fold: re-estimate Steps 2–3 on training portion, compute residuals on test portion

✓ Aggregate residuals → final CV Q^2 and IAE_{PI}

→ Slightly biased estimate of the error

Analytical benchmarks in [MI26]: diverse variable subset embeddings, projections & intrinsic dimensionalities

⇒ Robustness of methodology confirmed

⇒ At equal reduced dim., AS+GPR identifies more performant subspaces than variable selection

How to build parsimonious GPR in large dimension?

1. Reminds on Gaussian Process Regression

2. Proposed 4-step methodology → **HSIC + AS + GPR**

3. Illustration on thermal-hydraulic simulation testcase

Application: IB-LOCA Thermal-Hydraulic simulation

Pressurized Water Reactor scenario:

Loss of primary coolant accident due to a break in cold leg → **IB-LOCA accident**

LSTF mock-up (1/48 scale) simulated with **thermal-hydraulic CATHARE2 code**

→ CPU cost for one code run > 1 hour

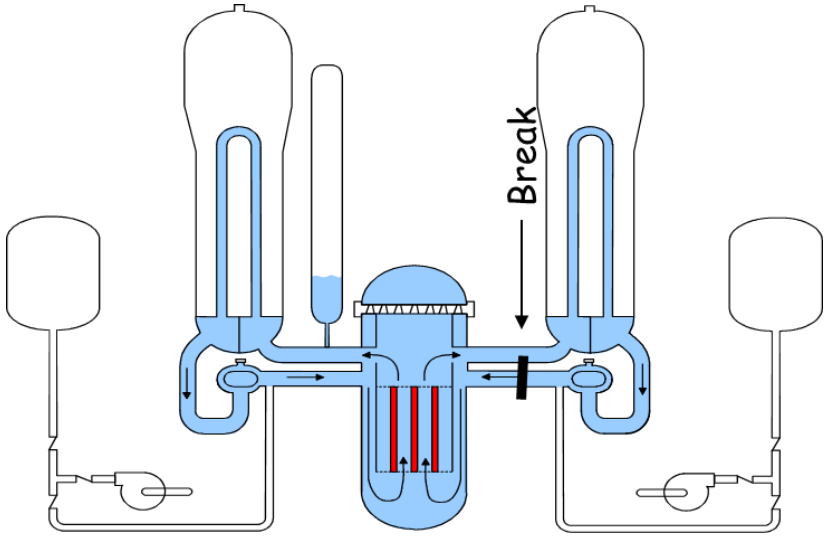
***d* = 27 uncertain input variables *X*:**

Critical flowrates, initial/boundary conditions, physical equation coefficients, scenario parameters…

Output of interest

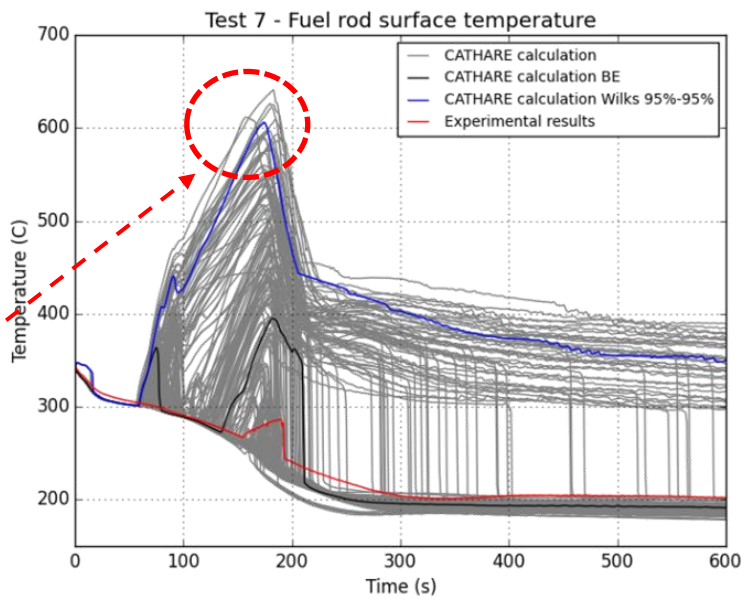
2nd peak of cladding temperature (PCT)

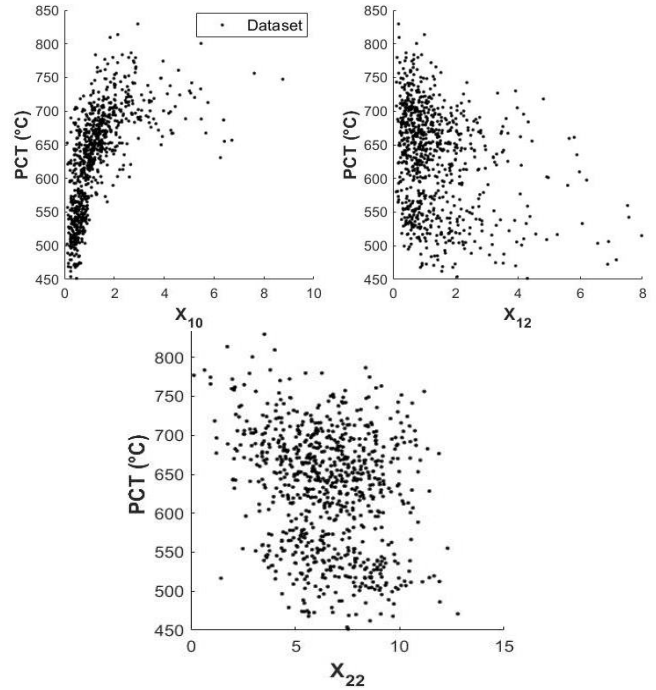
→ Critical safety indicator for licensing constraint



Available dataset $\mathcal{D}_n$:

***n* = 1000** Monte Carlo simulations





Application: IB-LOCA Thermal-Hydraulic simulation

3 approaches compared : constant mean + anisotropic stationary Matérn-5/2 for all GPRs

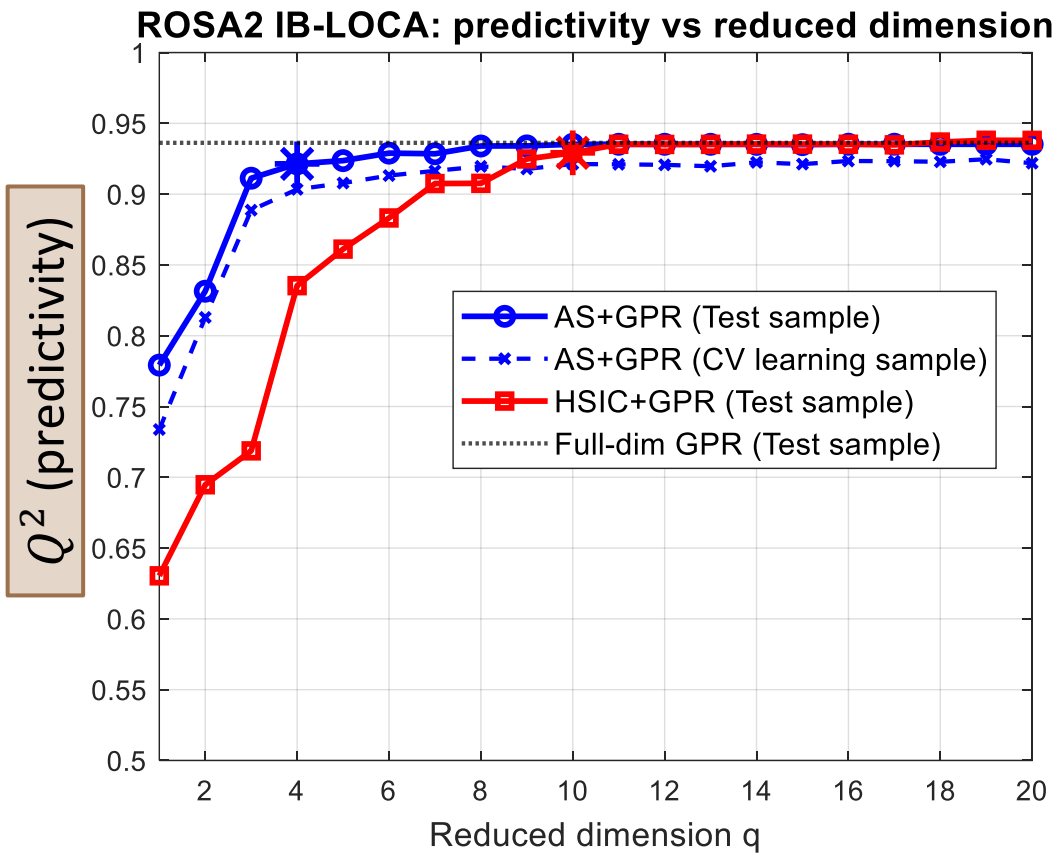
1. **Full-dim GPR** baseline: full d=27 inputs

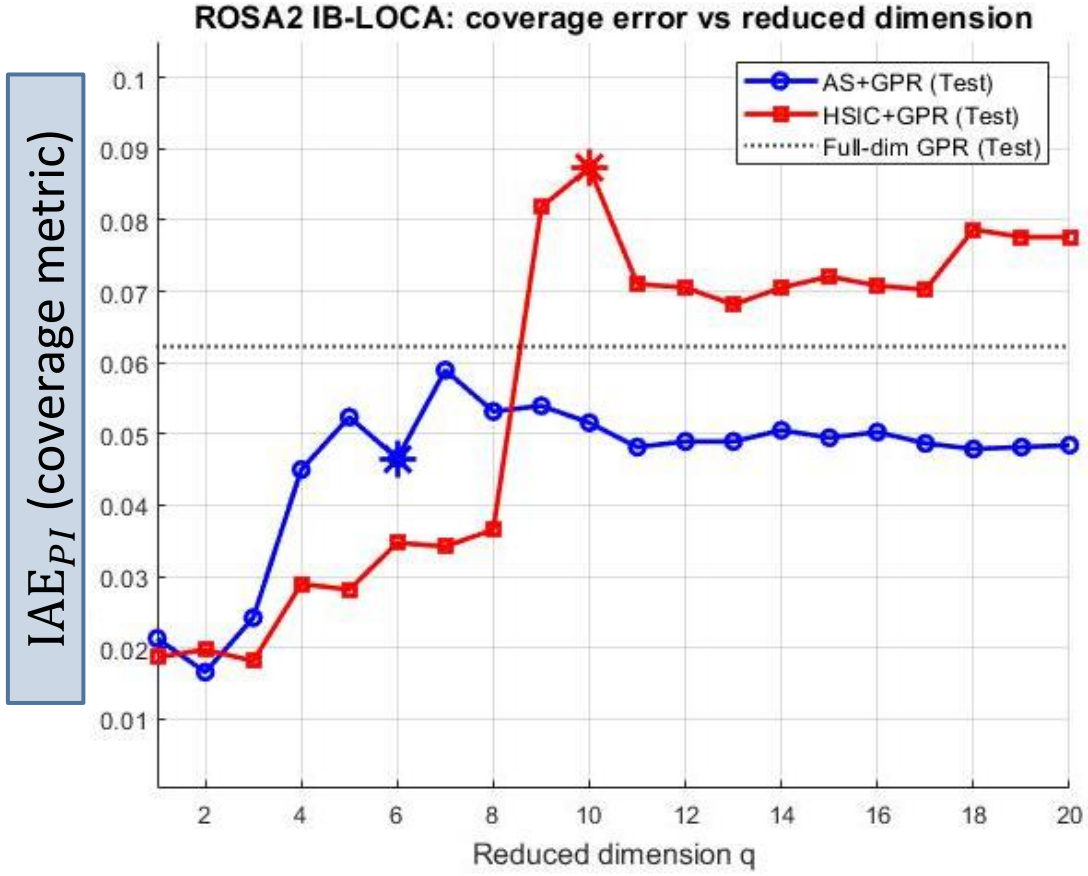
2. **HSIC+GPR:**
   - ✓ Standard HSIC screening with $\alpha = 0.1$ + GPR on selected inputs (kernel with nugget)

3. **AS+GPR [our 4-step methodology]** with:
   - ✓ Step1: α adaptive → d_s=20
   - ✓ Step2: nugget for training kernel matrix only, plug-in estimator $\hat{C}_{plug-in}$
   - ✓ Step3: Q^2-stagnation criterion of to select q^* (CV with 8 folds)

- Training set: $n_{train} = 700$ simulations
- Test set: $n_{test} = 300$

Application: IB-LOCA Thermal-Hydraulic simulation





Full-dim GPR : dim $d = 27$

Q^2 = 0.936 IAE_{PI} = 0.062

→ *High dimension, poor coverage*

HSIC + GPR: dim $d^* = 10$

Q^2 = 0.929 IAE_{PI} = 0.087

→ *63% dim. reduction, predictivity preserved but degraded coverage*

AS + GPR: dim $q^* = 6$

Q^2 = 0.934 IAE_{PI} = 0.047

→ 78% dim. reduction, no Q^2 loss and better coverage

Conclusions

- **Methodology for an efficient and parsimonious GPR from Active Subspace:**
  - ✓ GPR benefits greatly from **preliminary HSIC-based screening:** it allows to fit a **1st GPR in medium dimension** (dim <20) to obtain a correct estimation of the **gradients** → **Reliable AS estimation**
  - ✓ **Optimal AS reduction**: **heuristic approaches** proposed to select the optimal number of eigenvectors to keep, to balance GPR predictivity and dimension reduction
  - ✓ **Final estimation of a low-dimensional GPR**: in reduced space derived from AS linear embedding
    - ⇒ Handles **high-dimensional**, **gradient-free** problems effectively
    - ⇒ More cost-effective use of **low-dim GPR for UQ** tasks in the Bayesian framework
- Extension to **modified active subspaces** [Lee19]
- Extension to **non-linear embedding** combined with selection or sparsity algorithm

[MI26] A. Marrel and B. Iooss (2026), Gaussian process and active subspace recipe for parsimonious metamodeling. Preprint ⟨cea-05658731⟩.

References 1/2

[BW22] Binois, M., Wycoff, N., 2022. A survey on high-dimensional gaussian process modeling with application to Bayesian optimization. ACM Trans. Evol. Learn. Optim. 2.

[Con15] Constantine, P. G. (2015). Active Subspaces. Philadelphia, PA: SIAM.

[DL13] Damianou, A., Lawrence, N.D., 2013. Deep gaussian processes, in: Carvalho, C., Ravikumar, P. (Eds.), Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics, Scottsdale, AZ, USA. pp. 207–215.

[Dav15] Da Veiga (2015). Global sensitivity analysis with dependence measures, Journal of Statistical Computation and Simulation, 85:1283-1305, 2015.

[DIG+22] Demay, C., Iooss, B., Gratiet, L., and Marrel, A. (2022). Model selection for GP regression: an application with highlights on the model variance validation. QREI Journal, 38:1482-1500.

[DRC13] Durrande, N., O., D.G., Roustant, Carraro, L., 2013. ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis. Journal of Multivariate Analysis 155, 57–67.

[DNR11] Duvenaud, D.K., Nickisch, H., Rasmussen, C.E., 2011. Additive gaussian processes, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 226–234.

[EM24] El Amri and. Marrel (2024). More powerful HSIC based independence tests, extension to space filling designs and functional data. International Journal for Uncertainty Quantification14(2): 69-98.

[Gra21] B. Gramacy (2021) Gaussian Process Modeling, Design, and Optimization for the Applied Sciences. Chapman and Hall/CRC.

[GFT+07] Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B. & Smola, A. (2007). A kernel statistical test of independence. Advances in Neural Information Processing Systems, 2007.

[Lee19] Lee, M.R., 2019. Modified active subspaces using the average of gradients. SIAM/ASA Journal on Uncertainty Quantification 7, 53–66. doi:10.1137/17M1140662.

References 2/2

[MI24a] A. Marrel and B. Iooss (2024), Probabilistic surrogate modeling by Gaussian process: A review on recent insights in estimation and validation, Reliability Engineering and System Safety, 247:110120.

[MI24b] A. Marrel and B. Iooss (2024), Probabilistic surrogate modeling by Gaussian process: A new estimation algorithm for more robust prediction, Reliability Engineering and System Safety, 247:110094.

[RW06] C.E. Rasmussen and C.K.I. Williams (2006). Gaussian processes for machine learning. MIT Press.

[RWB+20] Raponi, E.,Wang, H., Bujny, M., Boria, S., Doerr, C., 2020. High dimensional bayesian optimization assisted by principal component analysis, in: Parallel Problem Solving from Nature – PPSN XVI, Springer. pp. 169–183.

[SMD+23] Sarazin, G., Marrel, A., Da Veiga, S. & Chabridon (2023). New insights into the feature maps of Sobolev kernels: application in global sensitivity analysis. https://cea.hal.science/cea-04320711.

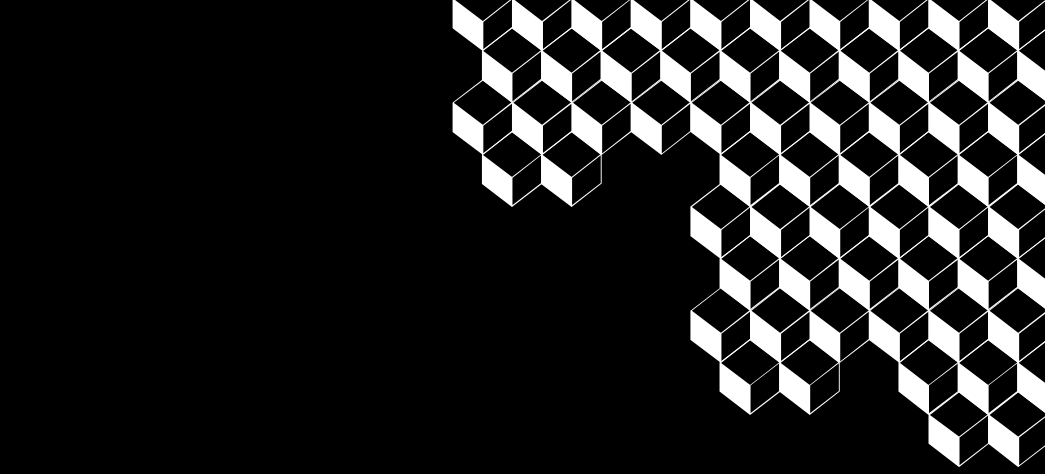
[TBG16] Tripathy, R., Bilionis, I., Gonzalez, M., 2016. Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation. Journal of Computational Physics 32, 191–223.

[TB19] Tripathy, R., Bilionis, I., 2019. Deep active subspaces: A scalable method for high-dimensional uncertainty propagation, in: ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers. p. V001T02A074.

[WBW21] N. Wycoff, M. Binois, S. M. Wild (2021). Sequential Learning of Active Subspaces. Journal of Computational and Graphical Statistics, 30(4), 1224–1237.

[WHS+17] Wilson, A.G., Hu, Z., Salakhutdinov, R., Xing, E.P., 2017. Deep kernel learning. Journal of Machine Learning Research 18, 1–33. Available at: https://arxiv.org/pdf/1511.02222

[WHZ+16] Wang, Z., Hutter, F., Zoghi, M., Matheson, D., De Freitas, N., 2016. Bayesian optimization in a billion dimensions via random embeddings. Journal of Artificial Intelligence Research 55, 361–387.



Appendix

Why choosing Active Subspaces (AS)?

Approach	Advantages ✔	Limitations ✘
Linear embeddings *AS, PCA, random projections, linear manifolds [Con15, TBG16, WHZ+16, RWB+20]*	Simple, interpretable, small data-friendly, theoretical guarantees	May degrade with strong nonlinearities
Non-linear embeddings coupled with GPR *Deep kernel learning, deep GP, VAE... [WHS+17, DL13, TB19]*	Expressive for complex functions	Large data needed Expensive Low interpretability
Additive / ANOVA-type GP *Additive GP, sparse ANOVA kernels... [DRC13, DNR11] + Review of [BW22]*	Effective for sparse additive structures	Complex model selection Not general-purpose

→ Adapted to our applicative settings and objectives

→ Our contribution [MI26]: **a practical 4-step methodology combining GPR with AS in gradient-free settings,** robustly estimating AS and selecting the optimal number of active directions

Estimation of GP hyperparameters (θ, τ)

$\Rightarrow$ How to robustly estimate the **hyperparameters** $(\boldsymbol{\theta}, \boldsymbol{\tau}) \in \mathbb{R}^{+,d+1}$ from the learning sample?

→ For good **predicitivity** + **reliable prediction intervals** $\Rightarrow$ Crucial for safety applications

► **Usual estimation methods** [Review in MI24a]

→ **Maximum likelihood**

→ **Cross-validation Mean Squared Error**

Ill-posedness, problem of **flatness** of functions to be minmized

→ **Bayesian approaches**

CPU ++, delicate choice of priors
Except RobustGAsp method

→ **New multi-objective estimation algorithm** for more reliable GP in [MI24b]

$\Rightarrow$ Reliable results observed in practice **up to a max dimension of** $\boldsymbol{d = 20}$

Step 1: HSIC-based screening

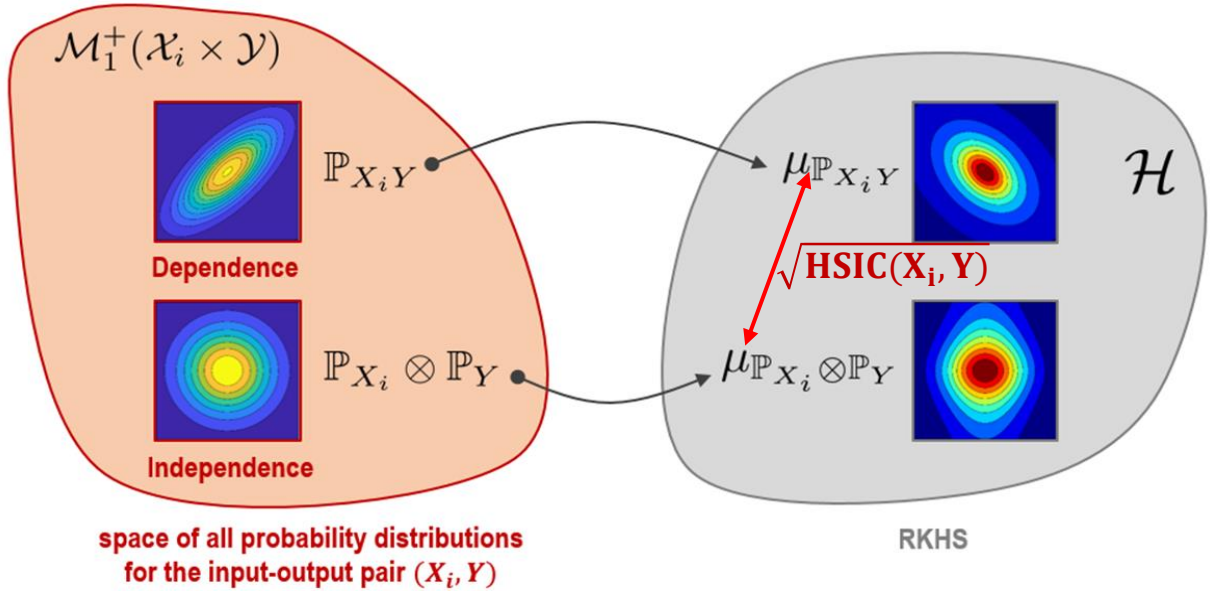
1st dimension reduction to obtain a reduced subset of $d_s \ll d$ inputs, enabling reliable initial GPR estimation in Step 2 → **Typically $d_s \lesssim 20$**

HSIC-based SCREENING: Surrogate-free + Gradient-free + Adapted to given-data and small-data context

- **HSIC*** ([GFT+07]) measure dependence between X_i and Y via kernel embeddings in RKHS

$$HSIC(X_i, Y) = MMD^2\left(P_{X_i Y}, P_{X_i} \otimes P_Y\right) = \left\| \mu_{\mathbb{P}_{X_i Y}} - \mu_{\mathbb{P}_{X_i} \otimes \mathbb{P}_Y} \right\|^2$$

- ✓ $\widehat{HSIC}$: Estimation of from **a unique random sample**, efficient in practice from $n \sim 100$
- ✓ HSIC "summarizes" the cross-cov between feature maps ⇒ Large panel of input-output dependency captured [Dav15]



Extract from a presentation by G. Sarazin (CEA)

* *Hilbert-Schmidt Independence Criterion*
MMD: Maximum Mean Discrepancy

Dealing with the large input dimension

- **HSIC-based ranking** [Dav15] **:**

$$R^2_{HSIC} = \frac{HSIC\,(X,Y)}{\sqrt{HSIC(X,X)HSIC\,(Y,Y)}}$$

⇨ $R^2_{HSIC} \in [0,1]$ for easier interpretation

$$\text{Influence}(X_{[1]}) > \text{Influence}(X_{[2]}) > \cdots > \text{Influence}(X_{[d]})$$

Where order $[\cdot]$ is such that $\widehat{R}^2_{H,X_{[1]}} > \widehat{R}^2_{H,X_{[2]}} > \cdots > \widehat{R}^2_{H,X_{[d]}}$

⇒ Use for ranking of inputs

Inputs ordered by degree of influence

Can be used for **more robust sequential GPR estimation**

⇒ “forward” estimation of GPR hyperparameters: successive inclusion of ordered inputs

HSIC review: a kernel-based GSA method

▶ MMD² applied between $P_{X_i Y}$ and $P_{X_i} \otimes P_Y \Rightarrow HSIC(X_i, Y)_{\mathcal{H}_{X_i}, \mathcal{H}_Y}$

$\mathcal{H}_{X_i}$ and $\mathcal{H}_Y$ **RKHS** associated to X_i and Y, resp :

Kernel $k_{X_i}: \mathcal{X}_i \times \mathcal{X}_i \to \mathbb{R}$ with feature space $\mathcal{H}_{X_i}$ and feature map φ_{X_i}

Kernel $k_Y: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ with feature space $\mathcal{H}_Y$ and feature map φ_Y

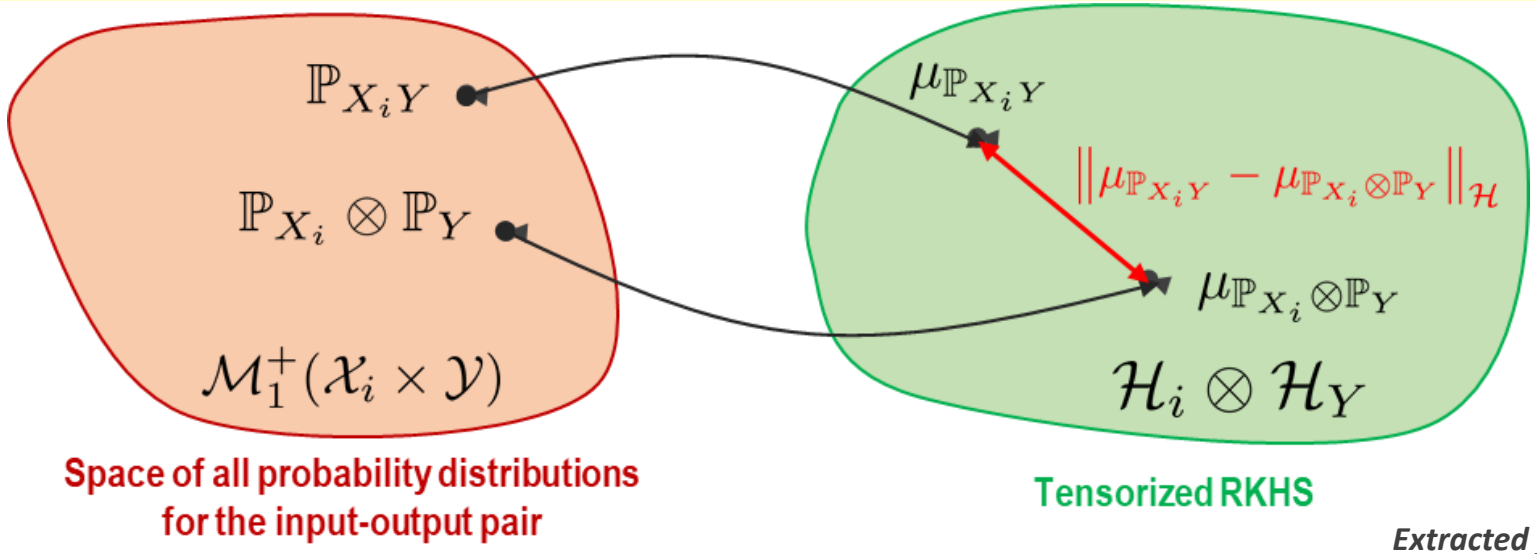
$$K_{X_i}(x, x') = \langle \varphi_{X_i}(x), \varphi_{X_i}(x') \rangle_{\mathcal{H}_{X_i}} \text{ and } K_Y(y, y') = \langle \varphi_Y(y), \varphi_Y(y') \rangle_{\mathcal{H}_Y}$$

kernel defines the inner product in the RKHS

HSIC = distance in the RKHS between the images of the two distributions of interest

Gretton et al. [2005]

$$\Rightarrow HSIC(X_i, Y)_{\mathcal{H}_{X_i}, \mathcal{H}_Y} = MMD^2_{\mathcal{H}_{X_i}, \mathcal{H}_Y}\left(P_{X_i Y}, P_{X_i} \otimes P_Y\right) = \left\| \mu_{\mathbb{P}_{X_i Y}} - \mu_{\mathbb{P}_{X_i} \otimes \mathbb{P}_Y} \right\|^2_{\mathcal{H}_{X_i}, \mathcal{H}_Y}$$



Space of all probability distributions for the input-output pair

Tensorized RKHS

Extracted from G. Sarazin's (CEA) slides

HSIC review: a kernel-based GSA method

▶ MMD² applied between $P_{X_i\,Y}$ and $P_{X_i} \otimes P_Y \Rightarrow HSIC(X_i, Y)_{\mathcal{H}_{X_i}, \mathcal{H}_Y}$

$\mathcal{H}_{X_i}$ and $\mathcal{H}_Y$ **RKHS** associated to X_i and Y, resp :

Kernel $k_{X_i}: \mathcal{X}_i \times \mathcal{X}_i \to \mathbb{R}$ with feature space $\mathcal{H}_{X_i}$ and feature map φ_{X_i}

Kernel $k_Y: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ with feature space $\mathcal{H}_Y$ and feature map φ_Y

$$K_{X_i}(x, x') = \left\langle \varphi_{X_i}(x), \varphi_{X_i}(x') \right\rangle_{\mathcal{H}_{X_i}} \text{ and } K_Y(y, y') = \langle \varphi_Y(y), \varphi_Y(y') \rangle_{\mathcal{H}_Y}$$

kernel defines the inner product in the RKHS

HSIC = distance in the RKHS between the images of the two distributions of interest

Gretton et al. [2005]

$$\Rightarrow HSIC(X_i, Y)_{\mathcal{H}_{X_i}, \mathcal{H}_Y} = MMD^2_{\mathcal{H}_{X_i}, \mathcal{H}_Y}\left(P_{X_i\,Y}, P_{X_i} \otimes P_Y\right) = \left\| C_{X,Y} \right\|^2_{HS}$$

With $C_{X,Y}$ **the covariance operator** between features maps:

$$C_{X,Y} = \mathbb{E}_{X,Y}\left[\varphi_X(X) \otimes \varphi_Y(Y)\right] - \mathbb{E}_X[\varphi_X(X)] \otimes \mathbb{E}_Y[\varphi_Y(Y)]$$

HSIC ***"summarizes" the cross-cov*** *between* ***feature maps***

$\Rightarrow$ ***Large panel of input-output dependency*** *can be captured.*

HSIC review: a kernel-based GSA method

► ***Characteristic kernels and RKHS*** ⇨ *Injective canonical feature map*

⇨ ***Equivalence to independence:*** $\boldsymbol{HSIC(X,Y) = 0 \Leftrightarrow X \perp Y}$

Ex: Gaussian Kernel

$$k(x_i, x_i') = exp\left(-\frac{(x_i - x_i')^2}{2\lambda^2}\right)$$

► ***Estimation:*** ***Kernel Trick*** ⇒ *Feature map linked to kernel function*

Very simple M-C estimator from a *n*-sample of simulations $\left(X_i^{(j)}, Y^{(j)}\right)_{1 \le j \le n}$

$$\widehat{\mathbf{HSIC}}(X_i, Y) = \frac{1}{n-1} Tr(K_i H L H)$$

where $H = I_n - \frac{1}{n}$, $K_i = \left(k_i\left(X_i^{(j)}, X_i^{(j')}\right)\right)_{1 \le j, j' \le n}$ and $L = \left(k\left(Y^{(j)}, Y^{(j')}\right)\right)_{1 \le j, j' \le n}$

► ***Statistical Properties:***

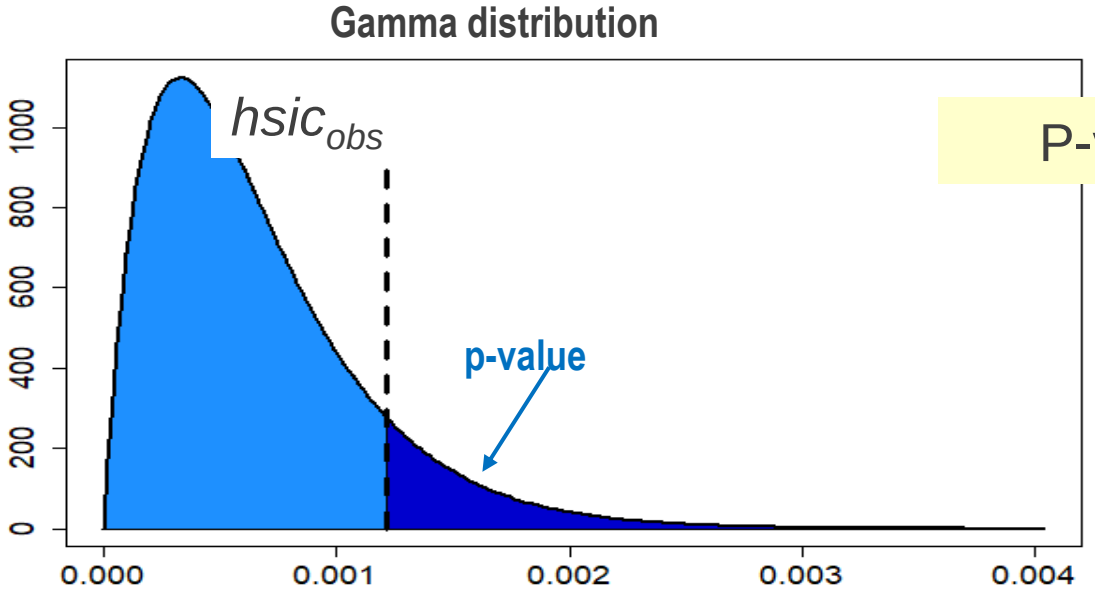
- Asymptotically unbiased, variance of order *O*(1/n)
- If $X \perp Y$, $n\widehat{\mathbf{HSIC}}(X, Y)$ converges asymptotically to a Gamma distribution

HSIC review: a kernel-based GSA method

HSIC-based independence tests for screening

How to have the distribution $n\widehat{\mathrm{HSIC}}(\mathrm{X_i}, \mathrm{Y})$ under $\mathcal{H}_0$ to compute *p-value*?

- ▶ **If *n* large: asymptotic test based** on approximation with Gamma law *(Gretton et al. (2008])*
- ▶ **If *n* small: Permutation-based** approximation *(De Lozzo & Marrel [2016a], Meynaoui [2019], El Amri & Marrel [2021a])*



P-value = *Pr* [$\widehat{\mathrm{HSIC}}(\boldsymbol{X_i}, \boldsymbol{Y}) > hsic_{obs}$]

Interpretation of *p-value* for a level α ($\alpha = 5\%$ or 10%) for screening:

- ➢ **pval < α** ⇒ H_0 (Independence) rejected ⇒ **X_i is significantly influential**

GPR validation

▶ **Validation criteria computed by cross-validation (LOO or K-fold CV)** [DIG+21]

→ **Accuracy of the whole GP conditional distribution** [DIG+21,ABG23]

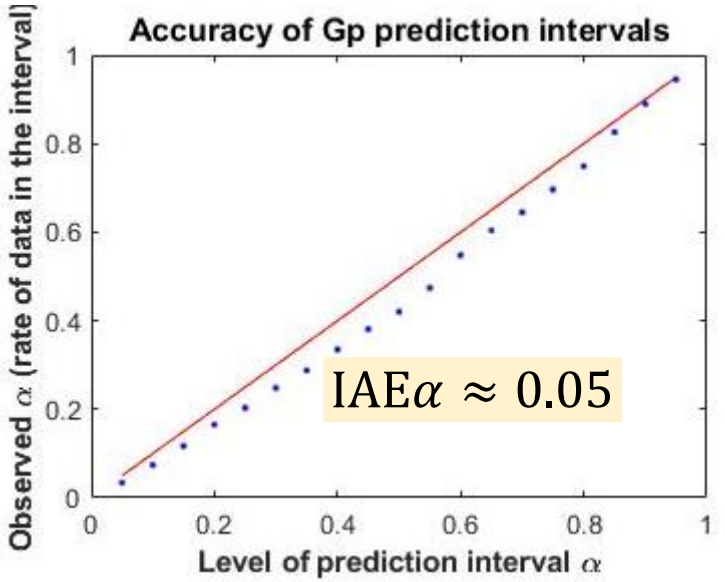
From empirical coverage function for α∈[0,1]: $\widehat{\Delta}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y_i \in PI_{\alpha,-i}(\mathbf{x}^{(i)})\}$

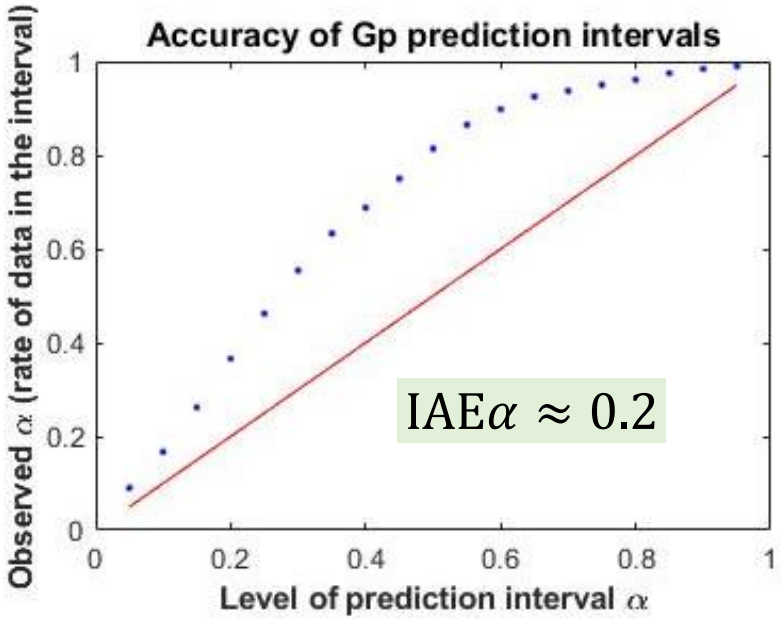
with $PI_{\alpha,-i}(\mathbf{x}^{(i)})$ the α-level GP prediction interval for $\mathbf{x}^{(i)}$ when $(\mathbf{x}^{(i)}, y_i)$ is removed from learning sample

⇒ **α-PI Plot**

⇒ **Integrated Absolute Error on $\widehat{\Delta}(\alpha)$** [MI24a]

$$\text{IAE}\alpha = \int_0^1 \left|\widehat{\Delta}(\alpha) - \alpha\right|$$



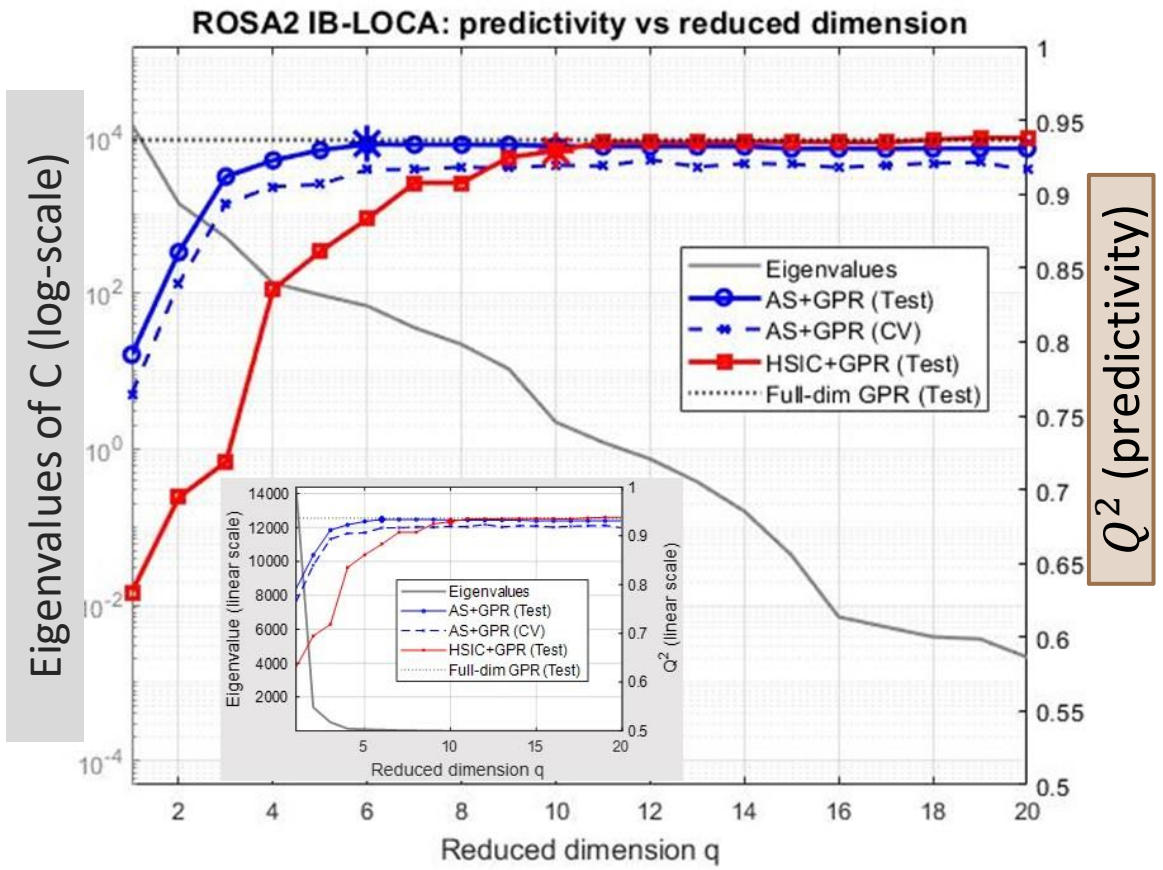


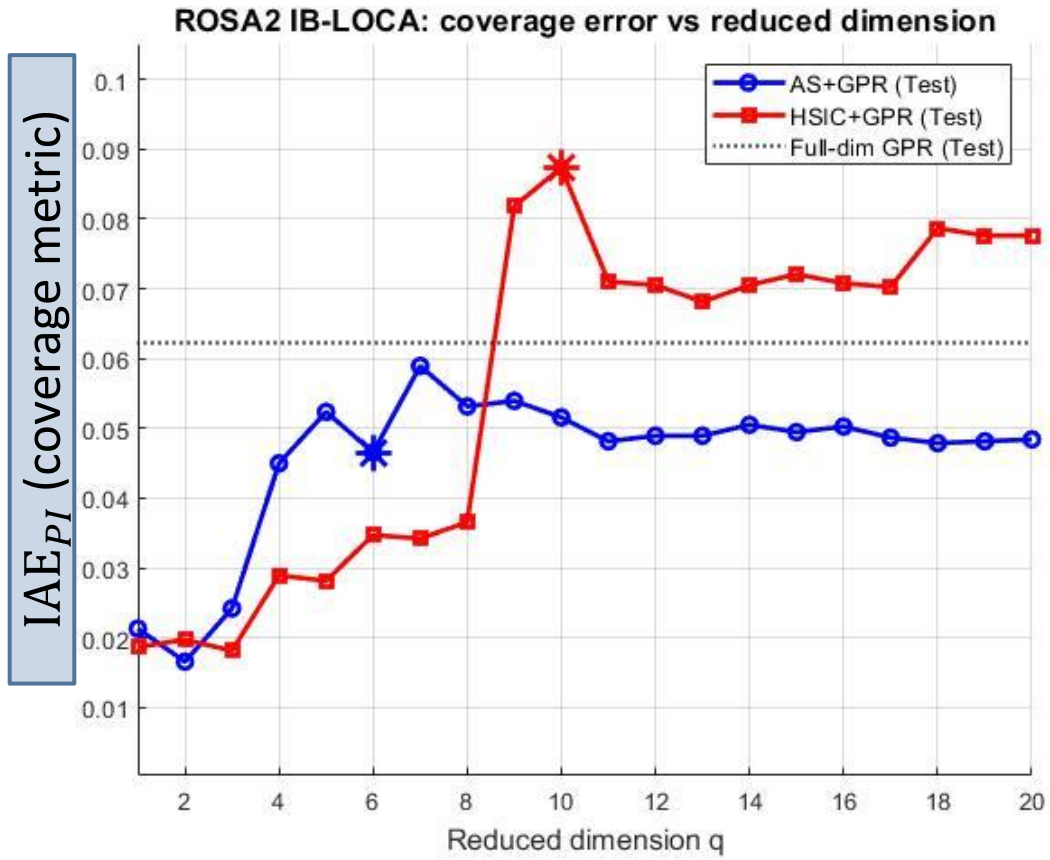
GPR validation

► **One message: Joint interpretation of Q^2 and IAEα !** [MI24a]

Criterion	Values	Interpretation
IAEα	**Value close to 0**	**Only if Q^2 is also high**, reliable predicted confidence intervals
	High values, close to 0.5 e.g.	Unreliable predicted prediction intervals ("underconfident" or "overconfident" model) **⇒ Explanation from cross-interpretation with Q^2 and α-PI plot**

Application: IB-LOCA Thermal-Hydraulic simulation





Full-dim GPR : dim $d = 27$

Q^2 = 0.936 IAE_{PI} = 0.062

→ *High dimension, poor coverage*

HSIC + GPR: dim $d^* = 10$

Q^2 = 0.929 IAE_{PI} = 0.087

→ *63% dim. reduction, predictivity preserved but degraded coverage*

AS + GPR: dim $q^* = 6$

Q^2 = 0.934 IAE_{PI} = 0.047

→ ***78% dim. reduction, no Q^2 loss and better coverage***

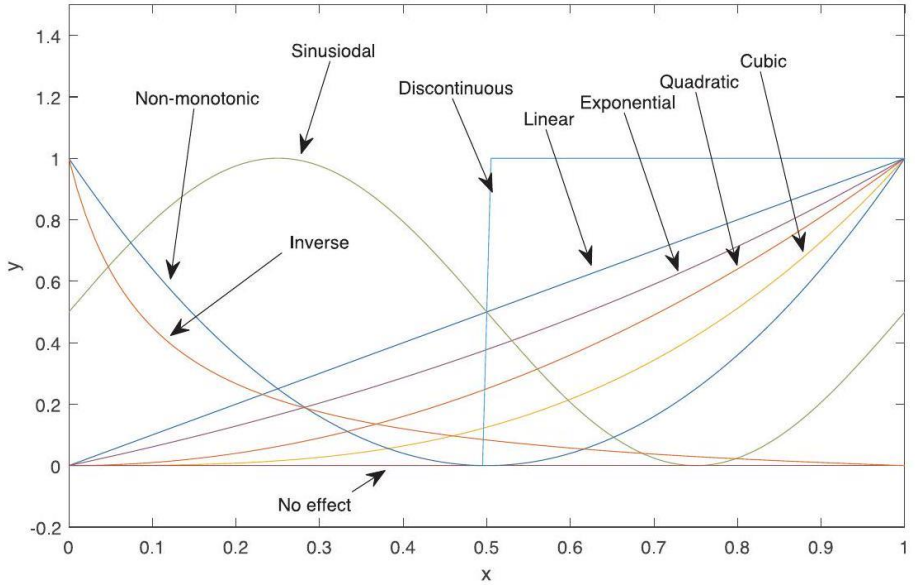
Illustration of methodology for parsimonious GPR

Becker Model $\mathcal{M}_{\text{Becker}}$ [Bec20] **with order 2 interaction effects:**

$$\mathcal{M}_{\text{Becker}}(\boldsymbol{x}, \boldsymbol{u}, \mathbf{V}, \Theta) = \sum_{i=1}^{d} a_i f^{u_i}(x_i) + \sum_{i=1}^{d_2} b_i f^{u_{V_{i,1}}}\left(x_{V_{i,1}}\right) f^{u_{V_{i,2}}}\left(x_{V_{i,2}}\right) \text{ with } \Theta = \{\boldsymbol{a}, \boldsymbol{b}\}$$

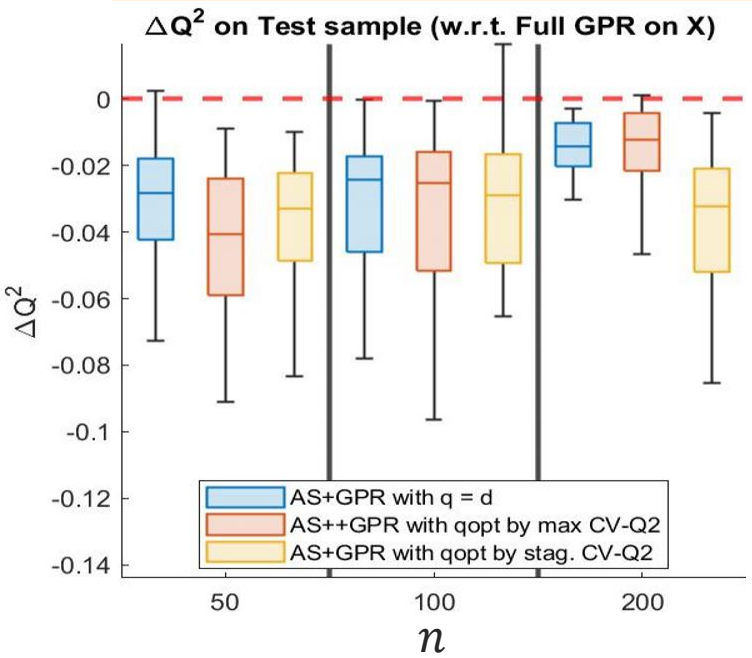
→ Built-upon a collection of 9 elementary functions f^k

→ Random sampling of parameters $\{u, V, \Theta\}$ → Randomized function $\mathcal{M}_{\text{Becker}}$

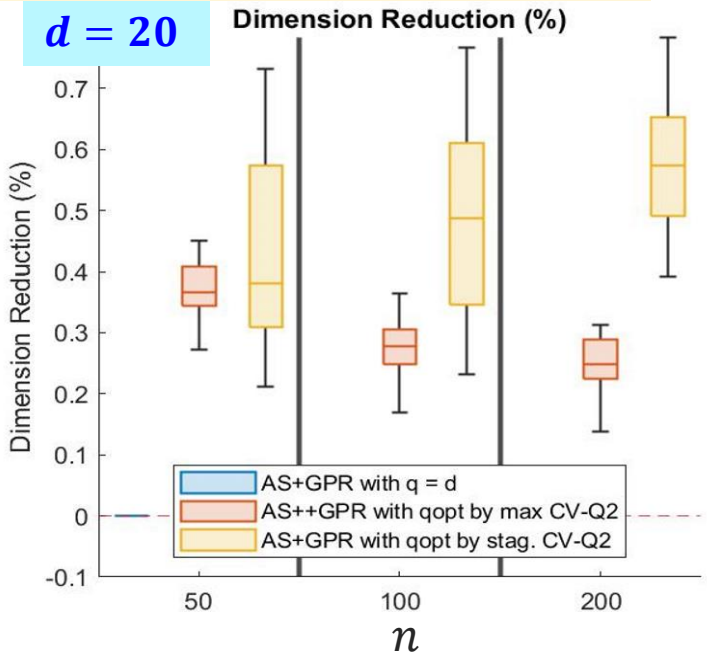


$d = 20$ independent uniform inputs: $\forall i \in [\![1,20]\!],\ X_i \sim \mathcal{U}([0,1])$

For each size n, 20 random draws of $\mathcal{M}_{\text{Becker}}$ functions.
(On average 70% of primary effects + 30% of interaction effects)



$d = 20$



Results for GPR with Matern 5/2 covariance

- **Predictivity Preserved :** With q selected by maximizing or stagnating Q^2 (estimated by CV*) → reduced GPR explains **less than 5% less variance** compared to the high-dimensional GPR
- **Significant Dimension Reduction**: On average, 25% (respectively 50%) with heuristic based on maximization (respectively stagnation) of Q^2 .

CV : cross-validation*