Gaussian Process Regression with Kernels Learned from Data

Éric Savin
ONERA & CentraleSupélec, Université Paris-Saclay, France
eric.savin@centralesupelec.fr

Joint work with:
Jean-Luc Akian (ONERA)
Luc Bonnet (Sandia National Laboratories)
Houman Owhadi (Caltech)

J. Comput. Phys. **470**:111595 (2022)

ONERA
THE FRENCH AEROSPACE LAB

CentraleSupélec

université
PARIS-SACLAY

SIAM CSE'23, February 27th, 2023

Metamodeling

Regression setting

Let $F : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ be a smooth function. Given I observations of the function F, denoted by $(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{X}_i, Y_i)_{i=1,\dots,I}$, approximate F.

- One needs to construct a surrogate model quick to evaluate and the most accurate possible;
- There exists many different methods depending on the available information: (generalized) Polynomial Chaos, Gaussian Process Regression/Kriging, Support Vector Machine, Artificial Neural Network, *etc.*;
- In the following, we will focus on the Gaussian Process Regression metamodeling method.

Kernel Ridge Regression solution

- Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. Let $\lambda > 0$. The Kernel Ridge Regression (KRR) solution G_λ is

$$G_\lambda := \arg \min_{G \in \mathcal{H}_K} \sum_{i=1}^{I} (Y_i - G(\boldsymbol{X}_i))^2 + \lambda \| G \|_K^2 \,, \tag{1}$$

  where $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$ is the Reproducing Kernel Hilbert Space (RKHS) associated with the kernel K defined by $\mathcal{H}_K = \{ G : \mathcal{X} \to \mathbb{R}; \forall \boldsymbol{x} \in \mathcal{X}, G(\boldsymbol{x}) = \langle G, K(\boldsymbol{x}, \cdot) \rangle_K \}$;
- The solution of the KRR approximation at an unobserved point $\boldsymbol{x}$ is:

$$\boxed{F(\boldsymbol{x}) \simeq G_\lambda(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X}) \left(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda \boldsymbol{I}_I \right)^{-1} \boldsymbol{Y}} \,,$$

  such that:

$$\| G_\lambda \|_K^2 = \boldsymbol{Y}^\mathsf{T} \left(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda \boldsymbol{I}_I \right)^{-1} \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) \left(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda \boldsymbol{I}_I \right)^{-1} \boldsymbol{Y} \,.$$

  Equivalent view: Kriging and Gaussian Process Regression.
- G_λ depends on the **choices** of the kernel K and nugget λ.

Choice of the kernel K

Two different methods have been used together to determine a "best" kernel K:

- Kernel Flow (KF) algorithms. Originally developed in a classification context but it can be extended to a regression context. Two different versions:
  - Parametric, including its sparse version;
  - Non-parametric.
- Spectral Kernel Ridge Regression (SKRR) algorithms:
  - Sparse SKRR algorithm;
  - Non-Sparse SKRR algorithm.

H. Owhadi, G. R. Yoo, *J. Comput. Phys.* **389**:22-47 (2019)
J.-L. Akian, L. Bonnet, H. Owhadi, É. Savin, *J. Comput. Phys.* **470**:111595 (2022)
L. Yang, X. Sun, B. Hamzi, H. Owhadi, N. Xie, arXiv:2301.10321 (2023)

Spectral Kernel Ridge Regression (SKRR) algorithms

- Assume that $F \in \mathcal{H}_K$, then for any $\boldsymbol{x} \in \mathcal{X}$:

$$|F(\boldsymbol{x}) - G_\lambda(\boldsymbol{x})| \leq \sigma(\boldsymbol{x}) \, \|F\|_K \ ,$$

where $\sigma^2(\boldsymbol{x}) = K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{K}(\boldsymbol{x},\boldsymbol{X}) \, (\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \lambda \boldsymbol{I}_l)^{-1} \, \boldsymbol{K}(\boldsymbol{X},\boldsymbol{x})$;

- How to find a "best" kernel? We focus on:

$$\min_K \|F\|_K \ .$$

- Use of Mercer's theorem to decompose K into the eigenvalues/eigenfunctions associated with its integral operator.

B. Schölkopf, R. Herbrich, A. J. Smola, *Lecture Notes in Computer Science* **2111**, pp. 416-426, Springer (2001)
J.-L. Akian, L. Bonnet, H. Owhadi, É. Savin, *J. Comput. Phys.* **470**:111595 (2022)
H. Owhadi, *Physica D* **444**:133592 (2023)

Spectral Kernel Ridge Regression (SKRR) algorithms

- **Mercer's theorem**: for $\mathcal{X}$ compact, K continuous, symmetric, and semi-definite positive,

$$K(\boldsymbol{x},\boldsymbol{y}) = \sum_{j=1}^{\infty} \sigma_j e_j(\boldsymbol{x}) \otimes e_j(\boldsymbol{y}) \,,$$

  where $\{e_j\}_{j=1}^{\infty}$ is a Hilbertian basis of $L^2(\mathcal{X})$, and $\sum_{j=1}^{\infty} \sigma_j < +\infty$.
- Consequently:

$$\mathcal{H}_K = \left\{ G \in L^2(\mathcal{X});\ \|G\|_K^2 = \sum_{j=1}^{\infty} \frac{\langle G, e_j \rangle_{L^2}^2}{\sigma_j} < +\infty \right\} .$$

- Let $\kappa > 0$, let $\{c_j\}_{j=1}^{\infty}$ such that $\sum_{j=1}^{\infty} c_j^2 < +\infty$, and consider:

$$\min_{\{\sigma_j\}} \sum_j \frac{c_j^2}{\sigma_j} \quad \text{such that} \quad \sum_j \sigma_j = \kappa \,;$$

  then:

$$\boxed{\sigma_k = \frac{\kappa\, |c_k|}{\sum_j |c_j|}} \,.$$

Sparse Spectral Kernel Ridge Regression (SSKRR) algorithm

Sparse Spectral Kernel Ridge Regression (SSKRR) algorithm

1. Let $\{e_k\}_{k\in\mathcal{K}} \subset \{e_i\}_{i=1}^{\infty}$ with $\#\mathcal{K} = R$ be orthonormal vectors in $L^2(\mathcal{X})$ on which $F \in L^2(\mathcal{X})$ is expected to be S-sparse;
2. Let $0 < \epsilon \ll 1$; compute the expansion coefficients $\boldsymbol{c}^\star = (c_{k_1}, \dots c_{k_R})$ by say ℓ_1-minimization (Basis Pursuit Denoise):

$$\boldsymbol{c}^\star = \arg\min_{\boldsymbol{c}\in\mathbb{R}^R} \|\boldsymbol{c}\|_1 \quad \text{such that} \quad \left| Y_i - \sum_{k\in\mathcal{K}} c_k e_k(\boldsymbol{X}_i) \right|^2 \leq \epsilon\,, \quad i = 1, 2 \dots I\,;$$

3. Compute:

$$\sigma_k^\star = \frac{\kappa\,|c_k^\star|}{\sum_{j\in\mathcal{K}} |c_j^\star|}\,;$$

4. The "best" kriging surrogate is defined as:

$$F(\boldsymbol{x}) \simeq G_\lambda^\star(\boldsymbol{x}) = \boldsymbol{K}^\star(\boldsymbol{x}, \boldsymbol{X})\left(\boldsymbol{K}^\star(\boldsymbol{X}, \boldsymbol{X}) + \lambda \boldsymbol{I}_I\right)^{-1} \boldsymbol{Y}\,,$$

where $K^\star = \sum_{k\in\mathcal{K}} \sigma_k^\star e_k \otimes e_k$.

SSKRR algorithm

In the SSKRR algorithm, two parameters need to be determined:

- The trace κ of the integral operator associated with K:

$$
\begin{aligned}
\kappa &= \sum_{j=1}^{\infty} \sigma_j = \mathrm{Tr}\, T_K = \int_{\mathcal{X}} K(\boldsymbol{x},\boldsymbol{x}) d\boldsymbol{x} \\
&= \mathbb{V}\, F \\
&\simeq \mathbb{V}\, \boldsymbol{Y}\, ;
\end{aligned}
$$

- The nugget λ of the approximation:

$$
F(\boldsymbol{x}) \simeq G^{\star}_{\lambda}(\boldsymbol{x}) = \boldsymbol{K}^{\star}(\boldsymbol{x},\boldsymbol{X})\left(\boldsymbol{K}^{\star}(\boldsymbol{X},\boldsymbol{X}) + \lambda \boldsymbol{I}_l\right)^{-1} \boldsymbol{Y}
$$

  can be estimated through: KF algorithm, grid search, cross-validation, marginal likelihood, *etc.*

SSKRR algorithm: Remarks

- $R = \#\mathcal{K}$ depends on the dimension d of the input set $\mathcal{X}$. In high-dimensional sets, finding $\boldsymbol{c}^\star$ can be numerically costly;
- The nugget λ allows us to improve the condition number of $\boldsymbol{K}^\star(\boldsymbol{X}, \boldsymbol{X})$;
- The $l \times l$ matrix $\boldsymbol{K}^\star(\boldsymbol{X}, \boldsymbol{X})$ is difficult to store and inverse for l large. This issue has been addressed recently;
- The prediction variance $\sigma(\boldsymbol{x})$ is independent of the observations $\boldsymbol{Y}$. Here SKRR puts some flavour of F into $\sigma(\boldsymbol{x})$ through $K^\star$.

A. G. Wilson, Z. Hu, R. Salakhutdinov, E. P. Xing, *Proc. Mach. Learn. Res.* **51**: 370-378 (2016)
F. Schäfer, T. J. Sullivan, H. Owhadi, *SIAM Multiscale Model. Simul.* **19**(2):688-730 (2021)

Non-sparse Spectral Kernel Ridge Regression (NSKRR) algorithm

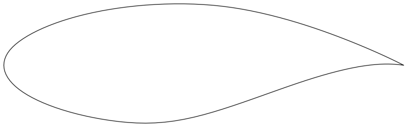
Non-sparse Spectral Kernel Ridge Regression (NSKRR) algorithm

1. Let $\{e_k\}_{k\in\mathcal{K}} \subset \{e_i\}_{i=1}^{\infty}$ be orthonormal vectors in $L^2(\mathcal{X})$;
   Let $K^{(0)}(\boldsymbol{x},\boldsymbol{y}) = \sum_{k\in\mathcal{K}} \sigma_k^{(0)} e_k(\boldsymbol{x}) \otimes e_k(\boldsymbol{y})$ be the initial kernel;
2. **for** $n \leftarrow 1$ **to** N **do**
   - Approach F by its NSKRR approximation $F(\boldsymbol{x}) \simeq G_\lambda^{(n-1)}(\boldsymbol{x}) = \boldsymbol{K}^{(n-1)}(\boldsymbol{x},\boldsymbol{X})\left(\boldsymbol{K}^{(n-1)}(\boldsymbol{X},\boldsymbol{X}) + \lambda \boldsymbol{I}_l\right)^{-1}\boldsymbol{Y}$;
   - **for** $k \in \mathcal{K}$ **do**
     - Compute $c_k^{(n-1)} = \langle G_\lambda^{(n-1)}, e_k\rangle_{L^2}$;
   - **end**
   - Compute $\sigma_k^{(n)} = \dfrac{\kappa\,|c_k^{(n-1)}|}{\sum_{j\in\mathcal{K}} |c_j^{(n-1)}|}$;
   - Form the new kernel as $K^{(n)}(\boldsymbol{x},\boldsymbol{y}) = \sum_{k\in\mathcal{K}} \sigma_k^{(n)} e_k(\boldsymbol{x}) \otimes e_k(\boldsymbol{y})$.

   end

Application: Lift coefficient C_L of RAE2822

- **Performance measure**: $\boldsymbol{X} \mapsto F(\boldsymbol{X})$ is the lift coefficient C_L of a RAE2822 wing profile, with $\boldsymbol{X} = (r, M, \alpha)$, where r is the thickness-to-chord ratio, M is Mach number, and α is the angle of attack of the wing profile following $\beta_I(4, 4)$ laws.



RAE 2822 wing profile.

	X_{lb}	X_{ub}
$X_1 = r$	$0.97 \times \underline{r}$	$1.03 \times \underline{r}$
$X_2 = M$	$0.95 \times \underline{M}$	$1.05 \times \underline{M}$
$X_3 = \alpha$	$0.98 \times \underline{\alpha}$	$1.02 \times \underline{\alpha}$

Range of each input parameter.

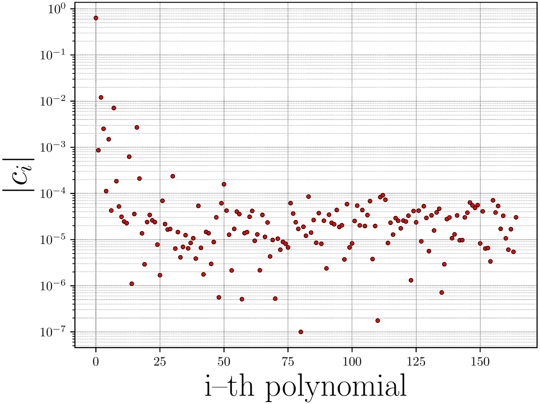
É. Savin, A. Resmini, J. Peter, AIAA Paper #2016-0433 (2016)
L. Cambier, S. Heib, S. Plot, *Mechanics & Industry* **14**(3):159-174 (2013)

Application: Lift coefficient C_L of RAE2822

- $\{e_k\}_{k\in\mathcal{K}}$ is chosen as a Jacobi polynomial basis;
- `SPGL1` (Spectral Projected Gradient Algorithm) is used to compute $\boldsymbol{c}^\star$;
- $I_{\mathrm{Tot}} = I + I_{\mathrm{V}} + I_{\mathrm{T}} = 80 + 15 + 25$ observations of F computed using `elsA`. I is the learning set, I_{V} is the validation set, I_{T} is the test set;
- $\kappa = \mathbb{V}\boldsymbol{Y}$ and λ is determined by the parametric KF algorithm using the learning and validation sets.

E. van den Berg, M. P. Friedlander, *SIAM J. Optim.* **21**(4):1201-1229 (2011)
L. Cambier, S. Heib, S. Plot, *Mechanics & Industry* **14**(3):159-174 (2013)

Application: Lift coefficient C_L



Expansion coefficients $\mathbf{c}^\star$ with $I = 80$ observations of the lift coefficient C_L.

Application: Lift coefficient C_L of RAE2822

Lift coefficient C_L			
	SSKRR	Sparse gPC	Fully tensorized gPC
e_{RMSE}	7.574×10^{-5}	3.715×10^{-4}	1.159×10^{-4}
e_{NRMSE}	1.040×10^{-4}	5.103×10^{-4}	8.437×10^{-5}
e_{MRE}	0.0319%	0.232%	0.0368%
Q^2	0.99996	0.99911	0.99995

Comparison of errors between surrogate models for the lift coefficient C_L with $I = 80$ and $I_{\mathrm{T}} = 25$.

$$e_{\mathrm{MRE}} = \max_{1 \leq i \leq I_{\mathrm{T}}} \frac{|Y_i - G_\lambda(\boldsymbol{X}_i)|}{|Y_i|} \,, \quad e^2_{\mathrm{RMSE}} = \frac{1}{I_{\mathrm{T}}} \sum_{i=1}^{I_{\mathrm{T}}} |Y_i - G_\lambda(\boldsymbol{X}_i)|^2 \,,$$

$$e^2_{\mathrm{NRMSE}} = \frac{e^2_{\mathrm{RMSE}}}{(\mathbb{E}\boldsymbol{Y}_{\mathrm{T}})^2 + \mathbb{V}\boldsymbol{Y}_{\mathrm{T}}} \,, \quad \mathrm{Q}^2 = 1 - \frac{e^2_{\mathrm{RMSE}}}{\mathbb{V}\boldsymbol{Y}_{\mathrm{T}}} \,,$$

$$\mathbb{E}\boldsymbol{Y}_{\mathrm{T}} = \frac{1}{I_{\mathrm{T}}} \sum_{i=1}^{I_{\mathrm{T}}} Y_i \,, \quad \mathbb{V}\boldsymbol{Y}_{\mathrm{T}} = \frac{1}{I_{\mathrm{T}}} \sum_{i=1}^{I_{\mathrm{T}}} (Y_i - \mathbb{E}\boldsymbol{Y}_{\mathrm{T}})^2 \,.$$

Outlook

Sparse and non-sparse SKRR algorithms:

- Choices of the basis $\{e_j\}_{j=1}^{\infty}$;
- Performances of the sparse and non-sparse SKRR algorithms;
- Time series: approximation of space-time fields;
- Solve PDEs using GPR.

Thank You!

Reproducing Kernel Hilbert Space (RKHS)

- We denote by $\mathfrak{F}(\mathcal{X},\mathbb{R})$ the set of functions from $\mathcal{X}$ to $\mathbb{R}$.

Definition (RKHS)

Let $\mathcal{X}$ be a non-empty set. We will call a subset $\mathcal{H} \subseteq \mathfrak{F}(\mathcal{X},\mathbb{R})$ a Reproducing Kernel Hilbert Space (RKHS) on $\mathcal{X}$ if

- $\mathcal{H}$ is a vector subspace of $\mathfrak{F}(\mathcal{X},\mathbb{R})$;
- $\mathcal{H}$ is endowed with an inner product $\langle \cdot,\cdot \rangle_{\mathcal{H}}$, with respect to which $\mathcal{H}$ is a Hilbert space;
- for every $\boldsymbol{x} \in \mathcal{X}$, the linear evaluation functional $\delta_{\boldsymbol{x}} : \mathcal{H} \to \mathbb{R}$ defined by $\delta_{\boldsymbol{x}}(f) = f(\boldsymbol{x})$ is bounded: $\exists C_{\boldsymbol{x}} > 0$, $\forall f, g \in \mathcal{H}$, $|\delta_{\boldsymbol{x}}(f-g)| = |f(\boldsymbol{x}) - g(\boldsymbol{x})| \leq C_{\boldsymbol{x}} \|f-g\|_{\mathcal{H}}$, where $\|f\|^2_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}$.

This means in particular that for $(f_n) \in \mathcal{H}$ such that $\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}} = 0$ then:

$$\lim_{n\to\infty} \delta_{\boldsymbol{x}}(f_n) = \delta_{\boldsymbol{x}}(f) \quad \text{or} \quad \lim_{n\to\infty} f_n(\boldsymbol{x}) = f(\boldsymbol{x}) \,, \quad \forall \boldsymbol{x} \in \mathcal{X} \,.$$

V. I. Paulsen, M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, Cambridge University Press (2016)

Reproducing Kernel Hilbert Space (RKHS)

- The Riesz representation theorem shows that the linear evaluation functional $\delta_{\boldsymbol{x}}$ is given by the inner product with a unique vector in $\mathcal{H}$.

Reproducing kernel

A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if

- $\forall \boldsymbol{x} \in \mathcal{X},\ K(\boldsymbol{x}, \cdot) \in \mathcal{H}$;
- $\forall \boldsymbol{x} \in \mathcal{X},\ \forall f \in \mathcal{H},\ f(\boldsymbol{x}) = \langle f, K(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}}$ (reproducing property).

Kernel function

Let $\mathcal{X}$ be a non-empty set and let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function. K is called a kernel function if it is positive semi-definite that is, for any $m \geq 1$, for every $(a_1, \dots, a_m) \in \mathbb{R}^m$, for any distinct $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_m) \in \mathcal{X}^m$,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0\,.$$

V. I. Paulsen, M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, Cambridge University Press (2016)

Reproducing Kernel Hilbert Space (RKHS)

- It can be shown that there is a one-to-one correspondence between RKHS on a set and kernel functions on this set.

One-to-one correspondence

Given a kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\mathcal{H}_K$ denotes the unique RKHS with reproducing kernel K.

Positive definite

Let $\mathcal{X}$ be a non-empty set and let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function. K is assumed positive definite, or non-degenerate, that is, for any $m \geq 1$, for any $\mathbf{a} = (a_1, \dots, a_m) \in \mathbb{R}^m$, $\mathbf{a} \neq \mathbf{0}$, for any distinct $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_m) \in \mathcal{X}^m$,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) > 0 \,.$$

V. I. Paulsen, M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, Cambridge University Press (2016)