Gaussian process model selection for computer experiments

Sébastien Petit, Julien Bect, Paul Feliot, Emmanuel Vazquez

Gaussian process model selection for computer experiments

Sébastien J. Petit[1,2] & Julien Bect[1] & Paul Feliot[2] & Emmanuel Vazquez[1]

[1]Université Paris-Saclay, CentraleSupélec, Laboratoire des Signaux et Systèmes, Gif-sur-Yvette, France.

[2] Safran Aircraft Engines, Moissy-Cramayel, France

email: sebastien.petit@centralesupelec.fr

Context

- Exploration of black-box numerical simulators $f : \mathbb{X} \subset \mathbb{R}^d \to \mathbb{R}$ with Gaussian processes
- Given data $D_n = (\mathcal{X}_n, f|_{\mathcal{X}_n})$, a Gaussian process ξ can be used to make probabilistic predictions of f

$$\xi(x)|D_n \sim \mathcal{N}(\hat{\xi}_\theta(x), \hat{\sigma}^2_\theta(x)) \tag{1}$$

- ξ is a prior over functions
- The choice of ξ is critical for good predictions and design-of-experiments techniques

The prior ξ is often chosen within a parametric family.

- Very often: the Matérn covariance functions family is used [7]
- Many procedures have been proposed in the literature for selecting the parameters of a covariance function
- Little is known about their relative benefits

What are the most useful procedures to select the parameters of a Matérn covariance function (including or not regularity) ?

1 Maximum-likelihood [5]

- A very popular technique
- Choose the parameters that yield the highest value of the probability density for the observations, or equivalently, minimize

$$z^T K_\theta z + \ln(\det(K_\theta)),$$

where K_θ is the covariance matrix of ξ at points $\mathcal{X}_n = (x_1, ..., x_n)$ for parameters θ and $z = (z_1, ..., z_n)^T$ denotes the values of f at $\mathcal{X}_n$

2 Cross-validation

Leave-one-out (LOO) [3] is a second very popular technique

- Consists in averaging *losses* for predicting one observation using the others
- We suggest using *negatively-oriented scoring rules* [4] for the loss functions
- A (negatively-oriented) scoring rule is a mapping $S : (\mathcal{P}, \mathbb{R}) \to \mathbb{R}$ where $\mathcal{P}$ is a class of probability distributions, with $S(P, z)$ representing a loss for observing z while predicting P
- Given a scoring rule S the corresponding LOO criterion is

$$L_S^{\mathrm{LOO}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} S(\mathcal{N}(\hat{\xi}_{\theta,-i}, \hat{\sigma}^2_{\theta,-i}), z_i), \tag{2}$$

where $\mathcal{N}(\hat{\xi}_{\theta,-i}, \hat{\sigma}^2_{\theta,-i})$ denotes LOO predictive distributions

In this work we consider the following scoring rules [4]:

- $S_{\mathrm{MSPE}}(P, z) = (\mathbb{E}_{Z\sim P}(Z) - z)^2$
- $S_{\mathrm{NLPD}}(P, z) = -\ln(p(z))$, with p the pdf of P
- $S_{\mathrm{CRPS}}(P, z) = ||F - \mathbf{1}_{z\leq .}||^2_{L^2(\mathbb{R})}$, with F the cdf of P

We shall denote the resulting selection procedures by LOO-MSPE, LOO-NLPD and LOO-CRPS respectively.

3 Generalized cross-validation [1]

- A version of LOO-MSPE that takes the heterogeneity of the design into account

4 Kernel alignment [2]

- Aligns the eigenvector related to the highest eigenvalue of K_θ with the data
- Can also be seen as a similarity between K_θ and the covariance matrix obtained from the kernel $(x, y) \to f(x)f(y)$

5 Numerical study

We use a set of 36 problems:

- Goldstein-Price ($d \in \{1, 2\}$)
- Mystery ($d = 2$)
- Toms 829 C^k ($k \in \{0, 1, 2\}$, $d \in \{2, 5\}$)
- Rotated Rosenbrock ($d \in \{2, 5\}$)
- Borehole ($d = 8$)

with space-filing designs $\mathcal{X}_n$, $n \in \{10d, 20d, 50d\}$. For each case:

- We compare model selection procedures using predictions evaluated on a dense test grid
- In particular, we study the influence of the regularity parameter p of a Matérn covariance function, by setting $p \in \{0, 1, 2, 3, 4, d, 2d, +\infty\}$ or automatically selecting its value using the selection criteria
- We present averaged results through repetitions (using random $\mathcal{X}_n$ for instance)

An example

Observe that

- Procedures ML, LOO-MSPE, LOO-NLPD and LOO-CRPS give similar accuracies
- The influence of p on the accuracy is strong

MSPE	ML	LOO MSPE	LOO NLPD	LOO CRPS	KA	GCV	"Best"
0	$1.08 \cdot 10^{-2}$	$1.12 \cdot 10^{-2}$	$1.07 \cdot 10^{-2}$	$1.06 \cdot 10^{-2}$	$2.16 \cdot 10^{-1}$	$1.04 \cdot 10^{-2}$	$9.74 \cdot 10^{-3}$
1	$3.27 \cdot 10^{-5}$	$3.11 \cdot 10^{-5}$	$2.78 \cdot 10^{-5}$	$2.85 \cdot 10^{-5}$	$1.71 \cdot 10^{-1}$	$2.79 \cdot 10^{-5}$	$2.45 \cdot 10^{-5}$
2	$1.17 \cdot 10^{-5}$	$1.23 \cdot 10^{-5}$	$1.29 \cdot 10^{-5}$	$\mathbf{1.14 \cdot 10^{-5}}$	$1.59 \cdot 10^{-1}$	$1.66 \cdot 10^{-5}$	$8.68 \cdot 10^{-6}$
3	$1.54 \cdot 10^{-5}$	$1.80 \cdot 10^{-5}$	$1.81 \cdot 10^{-5}$	$1.62 \cdot 10^{-5}$	$1.36 \cdot 10^{-1}$	$2.32 \cdot 10^{-5}$	$1.26 \cdot 10^{-5}$
4	$1.90 \cdot 10^{-5}$	$2.32 \cdot 10^{-5}$	$2.30 \cdot 10^{-5}$	$2.11 \cdot 10^{-5}$	$1.23 \cdot 10^{-1}$	$3.13 \cdot 10^{-5}$	$1.60 \cdot 10^{-5}$
6	$2.36 \cdot 10^{-5}$	$3.07 \cdot 10^{-5}$	$2.90 \cdot 10^{-5}$	$2.68 \cdot 10^{-5}$	$1.12 \cdot 10^{-1}$	$4.00 \cdot 10^{-5}$	$1.99 \cdot 10^{-5}$
12	$2.60 \cdot 10^{-5}$	$3.30 \cdot 10^{-5}$	$3.04 \cdot 10^{-5}$	$2.97 \cdot 10^{-5}$	$1.03 \cdot 10^{-1}$	$4.11 \cdot 10^{-5}$	$2.16 \cdot 10^{-5}$
$+\infty$	$2.94 \cdot 10^{-5}$	$3.77 \cdot 10^{-5}$	$3.33 \cdot 10^{-5}$	$3.18 \cdot 10^{-5}$	$9.23 \cdot 10^{-2}$	$4.31 \cdot 10^{-5}$	$2.43 \cdot 10^{-5}$
$\hat{p}$	$1.17 \cdot 10^{-5}$	$1.27 \cdot 10^{-5}$	$1.29 \cdot 10^{-5}$	$\mathbf{1.15 \cdot 10^{-5}}$	$9.23 \cdot 10^{-2}$	$1.74 \cdot 10^{-5}$	$8.68 \cdot 10^{-6}$

Table: Average MSPE on the validation sets for the different selection procedures and regularity choices.

Influence of the regularity

We focus on two subsets of problems with different smoothness. Fig. 1: 5-dimensional Toms 829 problems; Fig. 2: 5-dimensional Rosenbrock and Borehole.

We compare log S_{MSPE} normalized by "Best" values both with automatically selected or fixed-$p \in \{0, 1, 2, 3, 4, d, 2d, +\infty\}$.

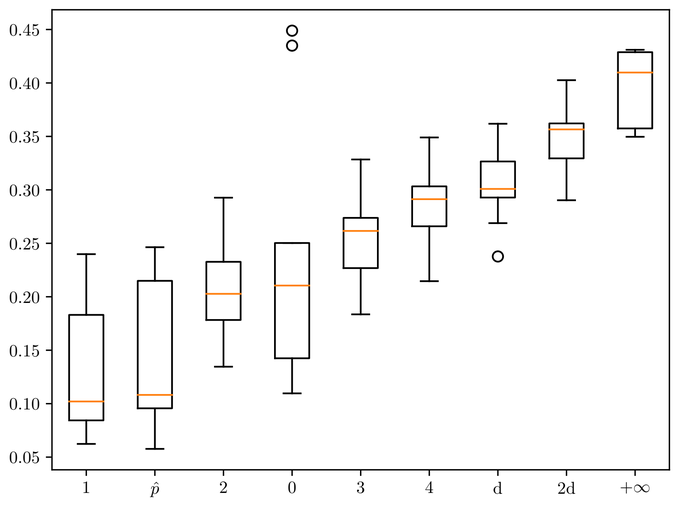

Fig. 1: Influence of the regularity on the loss for non-smooth problems.

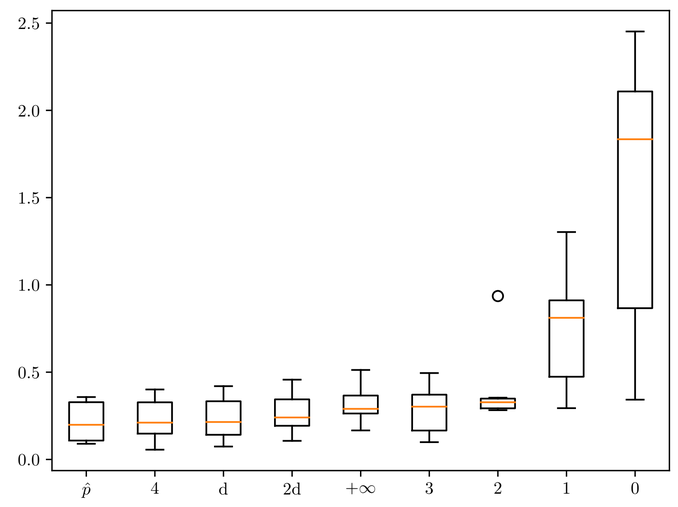

Fig. 2: Influence of the regularity on the loss for smooth problems.

Influence of the selection criteria

We compare the selection procedures with automatically selected ps. Fig. 3: $\log(S_{\mathrm{MSPE}})$ normalized by "Best" values; Fig. 4: interval score [4] defined by

$$S^{\mathrm{IS}}_\alpha(l, u, x) = (u - l) + \frac{2}{\alpha}(l - x)\mathbf{1}_{x\leq l} + \frac{2}{\alpha}(x - u)\mathbf{1}_{x>u}. \tag{3}$$

6 Conclusions

- The regularity parameter has a strong impact on the goodness of fit
- We recommend selecting the regularity from data instead of fixing it to a "standard" value
- The choice of a reasonable selection procedure has second-order impact but ML and LOO CRPS seem to give the best performances
- All procedures have the same numerical complexity, using appropriate computations of the selection criteria and their gradients [6]

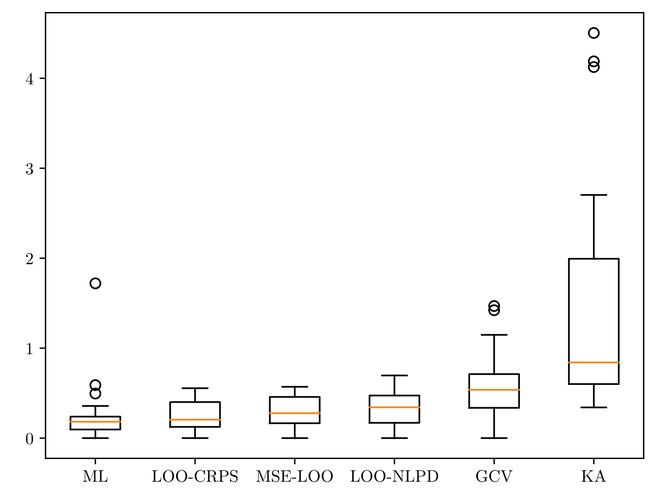

Fig. 3: Influence of the selection criteria on the MSPE.

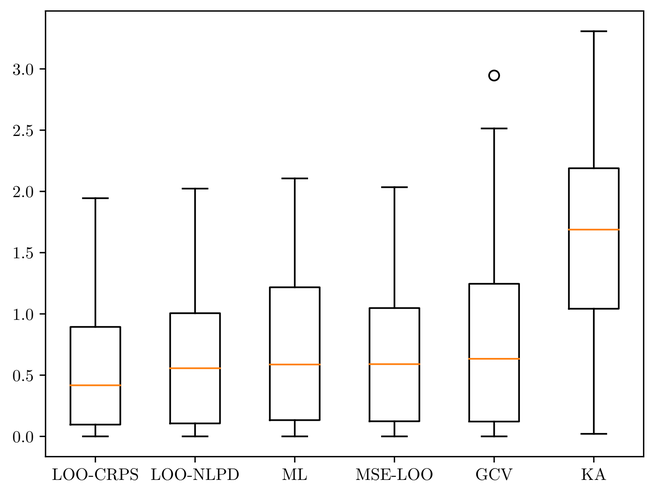

Fig. 4:Influence of the selection criteria on the interval score.

References

[1] P. Craven and G. Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 1979.

[2] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On kernel-target alignment. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, Cambridge, MA, USA, 2001. MIT Press.

[3] C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. A Bayesian approach to the design and analysis of computer experiments. Technical report, Oak Ridge National Lab., TN (USA), 1988.

[4] T. Gneiting and A. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

[5] P. K. Kitanidis. Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, 19(4):909–921, 1983.

[6] S. J. Petit, J. Bect, S. Da Veiga, P. Feliot, and E. Vazquez. Towards new cross-validation-based estimators for Gaussian process regression: efficient adjoint computation of gradients, 2020.

[7] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer Series in Statistics. Springer New York, 1999.

[1]Université Paris-Saclay, CentraleSupélec, Laboratoire des Signaux et Systèmes, Gif-sur-Yvette, France.

[2] Safran Aircraft Engines, Moissy-Cramayel, France