Towards new cross-validation-based estimators for Gaussian process regression: efficient adjoint computation of gradients

Sébastien Petit, Julien Bect, Sébastien da Veiga, Paul Feliot, Emmanuel Vazquez

Towards new cross-validation-based estimators for Gaussian process regression: efficient adjoint computation of gradients

Sébastien J. Petit[1,2,⋆] & Julien Bect[1] & Sébastien Da Veiga[3]
& Paul Feliot[2] & Emmanuel Vazquez[1]

[1] *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, Gif-sur-Yvette, France.* ⋆*E-mail :* sebastien.petit@centralesupelec.fr.
[2] *Safran Aircraft Engines, Moissy-Cramayel, France*
[3] *Safran Tech, Châteaufort, France*

Résumé. Nous nous intéressons à l'estimation par validation croisée des paramètres d'une fonction de covariance d'un processus gaussien. Nous suggérons l'utilisation de nouveaux critères de validation croisée dérivés de la littérature des *scoring rules*. Nous proposons de plus une méthode efficace pour le calcul du gradient d'un critère de validation. Cette méthode est plus efficace que ce qui est présenté dans la littérature à notre connaissance, et permet en particulier de réduire la complexité de l'évaluation jointe des critères de validation croisée et des gradients associés.

Mots-clés. Processus gaussien, validation croisée, score de prédiction probabiliste, Mode adjoint

Abstract. We consider the problem of estimating the parameters of the covariance function of a Gaussian process by cross-validation. We suggest using new cross-validation criteria derived from the literature of *scoring rules*. We also provide an efficient method for computing the gradient of a cross-validation criterion. To the best of our knowledge, our method is more efficient than what has been proposed in the literature so far. It makes it possible to lower the complexity of jointly evaluating leave-one-out criteria and their gradients.

Keywords. Gaussian process, cross-validation, scoring rule, reverse-mode differentiation

1 Introduction

Let ξ be a zero-mean Gaussian process indexed by $\mathbb{R}^d$ and denote by k the covariance function of ξ, which is assumed to belong to a parametrized family $\{k_\theta; \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^q$ denotes a q-dimensional space of parameters. We can safely say that the most popular methods for estimating k from data are maximum likelihood and related techniques.

In this article we focus instead on cross-validation methods. Classical cross-validation methods for estimating k are based on the leave-one-out mean squared prediction error or PRESS (Allen, 1974; Bachoc, 2013), and leave-one-out log predictive density (see, e.g., Rasmussen and Williams, 2006). These leave-one-out goodness-of-fit criteria can be computed using closed-form formulas (Dubrule, 1983).

The contribution of this work is twofold. First, we suggest extending the range of classical cross-validation criteria available in the literature of Gaussian processes by using the broad variety of *scoring rules* (see Gneiting and Raftery, 2007, for a survey), such as the continuous ranked probability score (CRPS). Second, we provide an efficient way for computing the gradient of any cross-validation criterion, which can then be used in gradient-based optimization algorithms. The only requirement is for the criterion to be differentiable in closed form with respect to leave-one-out posterior predictive means and variances. The new procedure has a $\mathcal{O}(n^3 + qn^2)$ complexity, against the $\mathcal{O}(qn^3)$ that was deemed "unavoidable" by Rasmussen and Williams (2006).

The article is organized as follows. Section 2 introduces scoring rules and how they can be used for estimating k. Section 3 presents the details of our contribution to the computation of gradients of a cross-validation criterion and Section 4 presents our conclusions and perspectives.

2 Scoring rules and cross-validation criteria

Let $Z_i = \xi(x_i) + \varepsilon_i$, $1 \leq i \leq n$, be some observations of ξ, at points $x_i \in \mathbb{R}^d$, where the ε_is are assumed independent and identically $\mathcal{N}(0, \sigma_\varepsilon^2)$-distributed, with $\sigma_\varepsilon^2 \geq 0$.

The classical framework of Gaussian process regression allows one to build a predictive distribution for an unobserved $\xi(x)$ at $x \in \mathbb{R}^d$ from the Z_is. Criteria for assessing the quality of probabilistic predictions have been studied in depth under the name of *scoring rules* in the seminal article of Gneiting and Raftery (2007). A scoring rule for real variable prediction is a function $S : \mathcal{P} \times \mathbb{R} \to [-\infty, +\infty]$, where $\mathcal{P}$ is a class of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For $P \in \mathcal{P}$ and $z \in \mathbb{R}$, $S(P, z)$ measures the goodness of prediction P for z.

Assume that we want to use a scoring rule S for estimating the parameters of a covariance function. Gneiting and Raftery (2007) suggest building a leave-one-out cross-validation criterion L defined as

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} S(P_{-i}^{\theta}, Z_i), \tag{1}$$

where P_{-i}^{θ} is the conditional distribution of Z_i given the Z_js, for $1 \leq j \leq n,\ j \neq i$.

In our Gaussian process regression framework, it is well known that P_{-i}^{θ} is a Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i^2)$. Let $K = (k_\theta(x_i, x_j))_{1 \leq i,j \leq n}$ be the covariance matrix of $(\xi(x_1), \ldots, \xi(x_n))^{\mathrm{T}}$, then (Dubrule, 1983; Sundararajan and Keerthi, 2001; Craven and

Wahba, 1979) show that

$$\mu_i = Z_i - \frac{(BZ)_i}{B_{i,i}} \quad \text{and} \quad \sigma_i^2 = \frac{1}{B_{i,i}}, \tag{2}$$

where $B = (K + \sigma_\varepsilon^2 I)^{-1}$ and $Z = (Z_1, \ldots, Z_n)^{\mathrm{T}}$. Note that (2) still stands true if $\sigma_\varepsilon^2 = 0$.

Remark 1. Craven and Wahba (1979, Lemma 3.1 and 3.2) show that (2) could be generalized to other types of linear predictors, beyond the particular Gaussian process regression framework considered in this article.

The mean squared prediction error and log-predictive density criteria mentioned in Section 1 correspond respectively to the scoring rules $S_1(P, z) = -(\mathbb{E}_{Z\sim P}(Z) - z)^2$ and $S_2(P, z) = \log(f(z))$, where f denotes the density of P with respect to some reference measure. A scoring rule is said *strictly proper* if $\mathbb{E}_{Z\sim P}(S(P, Z)) > \mathbb{E}_{Z\sim P}(S(Q, Z))$ for all $P,\ Q \in \mathcal{P}$ with $P \neq Q$. Strict propriety can be viewed as a sanity condition for performing estimation by maximizing (1). Note that S_1 is not strictly proper relative to the class of Gaussian measures whereas S_2 is. A large variety of scoring rules is surveyed by Gneiting and Raftery (2007). We shall use the CRPS in Section 4 for illustration.

3 Efficient computation of the gradient of a leave-one-out criterion

In this section we present our contribution for computing the gradient $\nabla_\theta L$ of (1). Let[1]

$$\begin{cases} \Gamma : \theta \in \mathbb{R}^q \mapsto K \in \mathbb{R}^{n^2}, \\ \varrho : K \in \mathbb{R}^{n^2} \mapsto (\mu, \sigma^2) \in \mathbb{R}^{2n} \text{ according to (2)}, \\ \varphi : (\mu, \sigma^2) \in \mathbb{R}^{2n} \mapsto L \in \mathbb{R} \text{ according to (1)}, \end{cases} \tag{3}$$

where $\mu = (\mu_1, ..., \mu_n)^{\mathrm{T}}$ and $\sigma^2 = (\sigma_1^2, ..., \sigma_n^2)^{\mathrm{T}}$, in such a way that $L(\theta) = (\varphi \circ \varrho \circ \Gamma)(\theta)$. Write $w = (\mu, \sigma^2)$ for simplicity. Let $J_{\varphi,w}$, $J_{\varrho,K}$ and $J_{\Gamma,\theta}$ be the $1 \times 2n$, $2n \times n^2$, $n^2 \times q$ Jacobian matrices of φ, ϱ and Γ at w, K and θ respectively. Using the chain rule for derivation we have

$$\nabla_\theta L^{\mathrm{T}} = J_{\varphi,w}\, J_{\varrho,K}\, J_{\Gamma,\theta}. \tag{4}$$

Rasmussen and Williams (2006) propose an algorithm in $\mathcal{O}(qn^3)$ time for computing $\nabla_\theta L$ from $J_{\Gamma,\theta}$.

Suppose that these Jacobian matrices are already built and stored. Then, computing (4) by multiplying those matrices from the right to the left costs about $2n \cdot n^2 \cdot q + 1 \cdot 2n \cdot q = \mathcal{O}(qn^3)$ additions and multiplications, corresponding to the complexity announced by Rasmussen and Williams (2006). On the other hand, proceeding from the left to the

[1] We identify the space of $n \times n$ matrices with $\mathbb{R}^{n^2}$ and $(\mathbb{R}^n)^2$ with $\mathbb{R}^{2n}$ with a slight abuse of notation.

right costs about $1 \cdot 2n \cdot n^2 + 1 \cdot n^2 \cdot q = 2n^3 + qn^2$ additions and multiplications. (This kind of consideration is a basic illustration of what has been studied in depth in the literature as the *matrix chain multiplication* problem for variable length products of matrices; see, e.g., Hu and Shing, 1982, and references therein.)

Let us now investigate the price paid for building $J_{\varphi,w}$ and $J_{\varrho,K}$. First of all, the computation of $B = (K + \sigma_\varepsilon^2 I)^{-1}$ and then $w = (\mu, \sigma^2)$ from K can be performed in $\mathcal{O}(n^3)$ operations using (2). Moreover, knowing w, L and $J_{\varphi,w}$ can be computed in $\mathcal{O}(n)$ time. In addition, equations used by Sundararajan and Keerthi (2001) show that $J_{\varrho,K}$ can be build from B in $\mathcal{O}(n^3)$ elementary operations. Thus, previous arguments show that it is indeed possible to compute L and $\nabla_\theta L$ from $J_{\Gamma,\theta}$ and K for $\mathcal{O}(n^3 + qn^2)$ elementary operations, thereby avoiding the $\mathcal{O}(qn^3)$ complexity mentioned by Rasmussen and Williams (2006).

Furthermore, available implementations (see, e.g., Bect et al., 2019) show that it is possible build K and $J_{\Gamma,\theta}$ from θ in a $\mathcal{O}(qn^2)$ complexity for the case of an anisotropic stationary covariance with $q = d + 1$ parameters (one variance parameter and q length scales). We see then that our contribution allows us in this case to keep the evaluation of L and $\nabla_\theta L$ from θ in $\mathcal{O}(n^3 + qn^2)$, rather than $\mathcal{O}(qn^3)$.

The main drawback of this scheme is the $2n \times n^2$ storage of $J_{\varrho,K}$. We propose to circumvent this cost by directly implementing the adjoint operators of the differentials of ϱ:

$$\mathcal{L}_\varrho^* : (K, \delta_w) \mapsto J_{\varrho,K}^{\mathrm{T}} \, \delta_w. \tag{5}$$

This can be used to compute $\mathcal{L}_\varrho^*(K, J_{\varphi,w}^{\mathrm{T}}) = J_{\varrho,K}^{\mathrm{T}} J_{\varphi,w}^{\mathrm{T}}$ and then $\nabla_\theta L$ from (4). This way of implementing chain rule derivatives is well known and has been studied under the name of *reverse-mode differentiation*[2] or *backpropagation*, and its paternity can be traced back at least to Linnainmaa (1970).

We propose Algorithm 1 to implement this operator. This algorithm only requires $\mathcal{O}(n^2)$ storage capacity and about $2n^3$ additions and multiplications, thus reducing the burden of storage, while maintaining the global $\mathcal{O}(n^3 + qn^2)$ complexity. (Note that $2n^3$ already corresponds to the cost of matrix multiplication in a "direct" approach that would first build $J_{\varrho,K}$ and then compute $\mathcal{L}_\varrho^*(K, \delta_w)$ by matrix multiplication.)

Remark 2. The algorithm can easily be adapted, through a suitable modification of the matrix B used in Step 6, to any type of linear model for which (2) holds (see Remark 1).

4 Conclusion and perspectives

We suggested using the scoring rules referenced by Gneiting and Raftery (2007) for the estimation of the parameters of a Gaussian process by leave-one-out cross-validation. We also proposed an efficient procedure for computing gradients of cross-validation criteria that is more efficient than what was available in the literature to our knowledge.

[2] In the context of Gaussian process regression, a reverse-mode differentiation approach has been proposed by Toal et al. (2009) for the computation of the likelihood function and its gradient.

Algorithm 1: Implementation of $\mathcal{L}_\varrho^*$ for computing $\delta_K = J_{\varrho,K}^{\mathrm{T}} \delta_w$, from K and δ_w. Inputs at first step refer to what has already been computed for evaluating μ and σ^2. For vectors a and b, $a \oslash b$ and $a \circ b$ denote the Hadamard element-wise division and multiplication respectively.

Input:
K, Z, $B = (K + \sigma_\varepsilon^2 I)^{-1}$, $\alpha = By$, $\kappa = (B_{i,i})_{1 \leq i \leq n}$, $\kappa^{-1} = \mathbb{1} \oslash \kappa$, $\chi = \alpha \circ \kappa^{-1}$,
$\delta_w = (\delta_\mu, \delta_{\sigma^2})$
Output: $\delta_K = J_{\varrho,K}^{\mathrm{T}} \delta_w$

$\kappa^{-2} = \kappa^{-1} \circ \kappa^{-1}$
$\delta_\chi = -\delta_\mu$
$\delta_\alpha = \delta_\chi \circ \kappa^{-1}$
$\delta_\kappa = -\delta_\chi \circ \alpha \circ \kappa^{-2} - \delta_{\sigma^2} \circ \kappa^{-2}$
$\delta_B = \delta_\alpha Z^{\mathrm{T}} + \mathrm{diag}(\delta_\kappa)$
$\delta_K = -B^{\mathrm{T}} \, \delta_B \, B^{\mathrm{T}}$

Further work will consist in investigating the properties of these estimators for several scoring rules. For instance, one can choose to use the continous rank probability score (CRPS) defined as $\mathrm{CRPS}(F, z) = -\int_{-\infty}^{+\infty} \left(F(u) - \mathbb{1}_{z \leq u}\right)^2 \mathrm{d}u$, where F is a cumulative distribution function. The CRPS is strictly proper relative to the class of Gaussian measures[3]. An empirical comparison with maximum likelihood for estimating the length scales is presented in Figure 1. Our contribution for computing gradients makes it possible to maintain the same complexity, both in terms of storage and calculation, for the two methods.

References

D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

F. Bachoc. Cross validation and maximum likelihood estimation of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics and Data Analysis*, 66:55–69, 2013.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, 2006.

O. Dubrule. Cross validation of kriging in a unique neighborhood. *Journal of the International Association for Mathematical Geology*, 15:687–699, 1983.

[3] and more generally with respect to the class of all probability measures with finite first order moment (see, e.g. Gneiting and Raftery, 2007, Section 4.2)

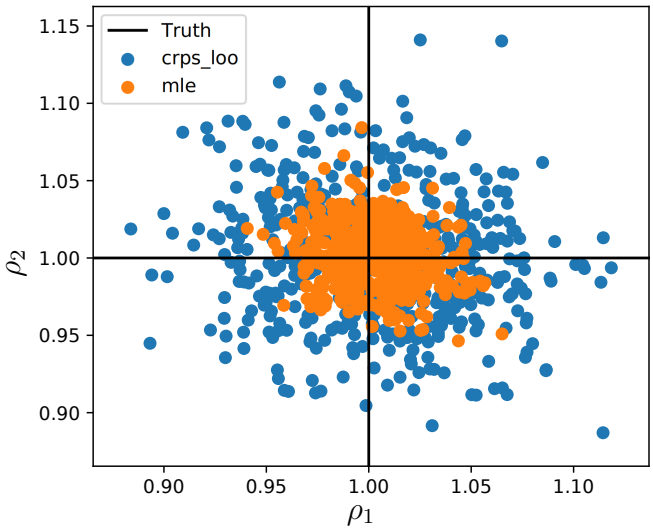

Figure 1: Scatterplots of the estimates of the two length scales of a Gaussian process on $\mathbb{R}^2$. Blue points correspond to CRPS-based cross-validation estimates; orange points correspond to maximum likelihood estimates. True length scales are represented by black lines. Each scatterplot consists of 500 estimations obtained from a space filing design of size $n = 500$. The criteria were optimized using a quasi-Newton type algorithm.

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

S. Sundararajan and S. S. Keerthi. Predictive approaches for choosing hyperparameters in Gaussian processes. *Neural Computation*, 13(5):1103–1118, 2001.

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*, 31:377–404, 1979.

T. C. Hu and M. T. Shing. Computation of matrix chain products. Part I. *SIAM Journal on Computing*, 11(2):362–373, 1982.

J. Bect, E. Vazquez, et al. STK: a Small (Matlab/Octave) Toolbox for Kriging. Release 2.6, 2019. URL http://kriging.sourceforge.net.

D. J. J. Toal, A. I. J. Forrester, N. W. Bressloff, A .J. Keane, and C. Holden. An adjoint for likelihood maximization. *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 465, 2009.

S. Linnainmaa. The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's thesis, Univ. Helsinki, 1970.