

Robust Prediction Interval estimation for Gaussian Processes by Cross-Validation method

Naoufal Acharki^{a,b,*}, Antoine Bertoncello^b, Josselin Garnier^a

^a*TotalEnergies One Tech, 91120 Palaiseau, France*

^b*Centre de Mathématiques Appliquées, Ecole Polytechnique, 91128 Palaiseau, France*

Abstract

Probabilistic regression models typically use the Maximum Likelihood Estimation or Cross-Validation to fit parameters. Unfortunately, these methods may give advantage to the solutions that fit observations in average, but they do not pay attention to the coverage and the width of Prediction Intervals. In this paper, we address the question of adjusting and calibrating Prediction Intervals for Gaussian Processes Regression. First we determine the model's parameters by a standard Cross-Validation or Maximum Likelihood Estimation method then we adjust the parameters to assess the optimal type II Coverage Probability to a nominal level. We apply a relaxation method to choose parameters that minimize the Wasserstein distance between the Gaussian distribution of the initial parameters (Cross-Validation or Maximum Likelihood Estimation) and the proposed Gaussian distribution among the set of parameters that achieved the desired Coverage Probability.

Keywords: Cross-Validation, Coverage Probability, Gaussian Processes, Prediction Intervals

1. Introduction

Many approaches of supervised learning focus on point prediction by producing a single value for a new point and do not provide information about how far

*Corresponding author:

Email address: `naoufal.acharki@polytechnique.edu` (Naoufal Acharki)

those predictions may be from true response values. This may be inadmissible,
 5 especially for systems that require risk management. Indeed, an interval is
 crucial and offers valuable information that helps for better management than
 just predicting a single value.

The Prediction Intervals are well-known tools to provide more information by
 quantifying and representing the level of uncertainty associated with predictions.
 10 One existing and popular approach for prediction models without predictive
 distribution (e.g. Random Forest or Gradient Boosting models) is the bootstrap,
 starting from the Traditional bootstrap (Efron and Tibshirani [1], 1993; Heskes
 [2], 1996) to Improved Bootstrap (Li *et al.* [3], 2018). It is considered as one of
 the most used methods [1] for estimating empirical variances and for constructing
 15 Predictions Intervals, it is claimed to achieve good performance under some
 asymptotic framework.

A set of empirical methods have been proposed for these models to build
 Prediction Intervals like the Infinitesimal Jackknife (Wager *et al.* [4], 2013)
 and Jackknife-after-Bootstrap methods (Efron [5], 1992), Quantile Random
 20 Forest (Meinshausen [6], 2006) Out-Of-Bag intervals (Zhang *et al.* [7], 2018),
 Split Conformal intervals (Lei *et al.* [8], 2018). In the Deep Learning field,
 many recent methods have been also developed to quantify the uncertainty in
 Neural networks: The Delta method (Hwang and Ding [9], 1997), Mean-Variance
 Estimation (Nix and Weigned [10], 1992), the Bayesian approach (MacKay
 25 [11], 1992; Ghahramani and Gal [12], 2016), Lower Upper Bound Estimation
 (Khosravi *et al.* [13], 2011a) and Quality-Driven ensembled approach (Pearce *et al.* [14], 2018).
 Most methods estimate the Coverage Probability (CP) (Landon
 and Singpurwalla [15], 2008) and the mean Prediction Interval width (MPIW)
 (Khosravi *et al.* [16], 2011a) by using the combinational Coverage Width-based
 30 Criterion (CWC) as a metric to identify model's parameters or define a loss
 function with a Lagrangian controlling the importance of the width and coverage.
 Pang *et al.* ([17], 2018) suggest the Receiver Operating Characteristic curve
 of Prediction Interval (ROC-PI), a graphic indicator that serves as a trade-off
 between the intervals width and CP for identifying the best parameters.

35 Unlike Ensemble methods or Neural Networks, there exist several prediction
 models with probabilistic framework like the Gaussian Processes (GP) model
 (Rasmussen and Williams [18], 2006) which are able to compute an efficient
 predictor with associated uncertainty. These models are more suitable for uncer-
 tainty quantification as they provide a predictive distribution with both point
 40 prediction and interval estimation and do not require any empirical approach
 such as the bootstrap. In most cases, the Prediction Intervals are generated
 with a *plug-in* method that takes the MLE of the model’s hyperparameters and
 substitutes them into the estimated conditional probability distribution. This
 approach works well only when the model is well-specified and may fail in case of
 45 model misspecification (F. Bachoc [19], 2013). Calibration of Prediction Intervals
 is little studied in the literature. Lawless and Fredette ([20], 2005) proposed a
 frequentist approach on predictive distribution to build and calibrate the Pre-
 diction Intervals. However, to the best of our knowledge, this approach has not
 yet been extended to models with a predictive distribution and we do not have
 50 any guarantees that it can work in case of misspecified model. Furthermore, the
 full-Bayesian approach gives an estimation of the hyperparameter’s distribution
 and makes it possible to build the Prediction Intervals, but it is very complex to
 implement typically with an Markov chain Monte Carlo (MCMC) algorithm and
 it is sensitive to the choice of the *prior* distribution of the hyperparameters (J.
 55 Muré [21], 2018). In this work, we propose a method based on Cross-Validation
 (CV) on the GP model to address the problem of model misspecification and
 calibrate Prediction Intervals by adjusting the upper and lower bounds to satisfy
 the desired level of CP.

The paper is organized as follows. Section 2 formulates the problem of
 60 Prediction Intervals estimation. Section 3 introduces the Gaussian Process
 regression model and its training methods. In Section 4, we present a method
 for estimating robust Prediction Intervals supported by theoretical results. We
 show in Section 5 the application of this method to academic examples and to
 an industrial example. Finally, we present our conclusions in Section 6.

65 2. Problem formulation

We consider n observations of an empirical model or computer code f . Each observation of the output corresponds to a d -dimensional input vector $\mathbf{x} = (x_1, \dots, x_d)^\top \in \mathcal{D} \subseteq \mathbb{R}^d$. The n points corresponding to the model/code runs are called an experimental design $\mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ where $\mathbf{x}^{(i)} =$
70 $(x_1^{(i)}, \dots, x_d^{(i)})^\top \in \mathcal{D}$. The outputs are denoted by $\mathbf{y} = (y^{(1)}, \dots, y^{(n)}) \in \mathbb{R}^n$ with $y^{(i)} = f(\mathbf{x}^{(i)})$. We seek to estimate the unobserved function $\mathbf{x} \in \mathcal{D} \mapsto f(\mathbf{x})$ from the data \mathbf{y} and make accurate predictions with the associated uncertainty.

Formally, let assume that f is a realization of random process Y and let $Y(\mathbf{x})$ be the value of model output at a point $\mathbf{x} \in \mathcal{D}$, let $\alpha \in [0, 1]$ describes the nominal level of confidence. We wish to estimate the interval $\mathcal{PI}_{1-\alpha}$ with respect to the type II CP (conditional coverage given the training set) such that

$$\mathbb{P}(Y(\mathbf{x}) \in \mathcal{PI}_{1-\alpha}(\mathbf{x}) | \mathbf{X}, \mathbf{y}). \quad (1)$$

This quantity must be as close as possible to $1 - \alpha$. In most cases, $\mathcal{PI}_{1-\alpha}$ is a two-sided interval delimited by two bounds at $\mathbf{x} \in \mathcal{D}$

$$\mathcal{PI}_{1-\alpha}(\mathbf{x}) := [y_{\alpha/2}(\mathbf{x}), y_{1-\alpha/2}(\mathbf{x})], \quad (2)$$

where $y_{\alpha/2}(\mathbf{x}) = \tilde{y}(\mathbf{x}) + z_{\alpha/2} \times \tilde{\sigma}(\mathbf{x})$ is the lower bound, $y_{1-\alpha/2}(\mathbf{x}) = \tilde{y}(\mathbf{x}) +$
75 $z_{1-\alpha/2} \times \tilde{\sigma}(\mathbf{x})$ is the upper bound, $z_{\alpha/2}$ (resp. $z_{1-\alpha/2}$) is the $\alpha/2$ (resp. $1 - \alpha/2$) quantile of the normalized predictive distribution (e.g. t -distribution for regression prediction), $\tilde{y}(\mathbf{x}) = \mathbb{E}(Y(\mathbf{x}) | \mathbf{X}, \mathbf{y})$ and $\tilde{\sigma}^2(\mathbf{x}) = \text{Var}(Y(\mathbf{x}) | \mathbf{X}, \mathbf{y})$ are the predictive mean and variance respectively.

In the framework of kriging, the prior distribution of the process Y is Gaussian characterized by a mean and covariance. The Cumulative Distribution
80 Function (CDF) of the predictive variable $Y(\mathbf{x})$ given \mathbf{X} and \mathbf{y} is well-defined and continuous with the Gaussian distribution. The quantile function is defined then as the inverse of the CDF and the quantiles $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are fully characterized. Thus, estimating the interval $\mathcal{PI}_{1-\alpha}$ in equation (2) is equivalent to estimate the predictive mean $\tilde{y}(\mathbf{x})$ and variance $\tilde{\sigma}^2(\mathbf{x})$.

85 Therefore, the objective is to build a surrogate model to estimate correctly the upper and lower bounds of Prediction Intervals $\mathcal{PI}_{1-\alpha}$. This goes through the CV method with respect to the CP. In the following sections, $\|\cdot\|$ refers to the Euclidean norm $\|\cdot\|_2$ if applied to a vector and to the Frobenius norm, defined by $\|\mathbf{M}\|_F = (\text{Tr}(\mathbf{M}\mathbf{M}^\top))^{1/2}$, if applied to a matrix.

90 3. Modelling with Gaussian Processes

We use the GP model to learn the unobserved function f . It is a Bayesian non-parametric regression (see [22] for Bayesian inference) which employs GP *prior* over the regression functions. It will be converted into a *posterior* over functions once some data has been observed. In the kriging framework, Y is assumed a *priori* to be a GP with mean $\mu(\mathbf{x})$ and covariance function $\mathbf{k}(\mathbf{x}, \mathbf{x}') + \sigma_\epsilon^2 \mathbf{1}\{\mathbf{x} = \mathbf{x}'\}$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{D}$. $\sigma_\epsilon^2 \geq 0$ is the variance of measurement error, also called the nugget effect.

3.1. The mean and covariance functions

The assumption made on the existing knowledge of the model Y and the mean function μ defines three sub-cases of kriging

- The Simple Kriging : μ is assumed to be known, usually null $\mu = 0$.
- The Ordinary Kriging : μ is assumed to be constant but unknown.
- The Universal Kriging : μ is assumed to be of the form $\sum_{j=1}^p \beta_j f_{j-1}(x)$, where f_j are predefined (e.g. polynomial functions $f_0(\mathbf{x}) = 1, f_j(\mathbf{x}) = x_j, j = 1, \dots, p-1$) and unknown scalar coefficients β_j .

The covariance function \mathbf{k} is a map that is symmetric positive semi-definite, usually stationary $\mathbf{k}(\mathbf{x}, \mathbf{x}') = \mathbf{r}(\mathbf{x} - \mathbf{x}')$. The most commonly used kernel in \mathbb{R} is the Matérn kernel class

$$r_{\sigma^2, \theta}^\nu(x - y) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|x - y|}{\theta} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{|x - y|}{\theta} \right), \quad x, y \in \mathbb{R} \quad (3)$$

where $\sigma^2 > 0$ is the amplitude, $\theta > 0$ is the length-scale, Γ is the complete Gamma function and K_ν is the modified Bessel function of the second kind. (σ^2, θ) are called hyperparameters. Some particular cases of Matérn kernel are when $\nu = \frac{1}{2}$ (Exponential), $\nu = \frac{3}{2}$ (Matérn 3/2), $\nu = \frac{5}{2}$ (Matérn 5/2) and $\nu \rightarrow \infty$ (Gaussian or Squared-Exponential).
110

The choice of kernels is important in the kriging scheme and requires prior knowledge of the smoothness of the function f . For example, the choice of the Gaussian kernel assumes that the function is very smooth of class \mathcal{C}^∞ (infinitely differentiable) which is often too strict as a condition. A common alternative is
115 the functions Matérn 5/2 or Matérn 3/2 kernel

Since the product of kernels is a kernel, it is possible to build high-dimensional kernels. We can obtain more complex covariance models in \mathbb{R}^d based on classical kernels in \mathbb{R} . In this paper we consider the Matérn anisotropic geometric model (radial model),

$$\mathbf{k}_{\sigma^2, \boldsymbol{\theta}}^{\text{radial}}(\mathbf{x}, \mathbf{x}') = \mathbf{r}_{\sigma^2, \boldsymbol{\theta}} \left(\sqrt{\sum_{j=1}^d \frac{|x_j - x'_j|^2}{\theta_j^2}} \right), \quad (4)$$

120 where \mathbf{r} is a Matérn kernel \mathbb{R} as defined in (3) and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$ the length-scale vector. The described method can be applied to other forms of covariance models like the tensorized product model with d -dimensional kernels or the Power-Exponential model. In the following sections, instead of writing $\mathbf{k}_{\sigma^2, \boldsymbol{\theta}}$, we denote simply \mathbf{k} when there is no possible confusion.

125 3.2. Gaussian Process Regression Model

The *prior* distribution of Y on the learning experimental design \mathbf{X} is multivariate Gaussian

$$\mathbf{y} | \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}(\mathbf{F}\boldsymbol{\beta}, \mathbf{K}), \quad (5)$$

where

- $\mathbf{F} = (F_{ij}) \in \mathbb{R}^{n \times p}$ is the regression matrix such that $F_{ij} = f_j(\mathbf{x}^{(i)})$.
- $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_p\} \in \mathbb{R}^p$ are the regression coefficients.

130 • $\mathbf{K} = (\mathbf{k}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}))_{1 \leq i, j \leq n} + \sigma_\epsilon^2 \mathbf{I}_n \in \mathbb{R}^{n \times n}$ is the covariance matrix of the learning design \mathbf{X} .

Hypothesis \mathcal{H}_1 : In the case of ordinary or universal kriging, we assume that $n \geq p$, \mathbf{F} is a full rank matrix, and $\mathbf{e} \in \text{Im } \mathbf{F}$ where $\mathbf{e} = (1, \dots, 1)^\top$.

Remark 1. In Ordinary Kriging, the hypothesis \mathcal{H}_1 is always satisfied. In the Universal Kriging, the hypothesis $\mathbf{e} \in \text{Im } \mathbf{F}$ is satisfied as soon as the constant
135 function $f_0(\mathbf{x}) = C$ is included in the chosen predefined functions f_i .

3.3. Prediction

The Gaussian conditioning theorem is useful to deduce the *posterior* distribution. By considering a new point \mathbf{x}_{new} , it can be shown that the predictive distribution of $Y(\mathbf{x}_{\text{new}})$ conditioned on the learning sample \mathbf{X}, \mathbf{y} is also Gaussian

$$Y(\mathbf{x}_{\text{new}}) | \mathbf{X}, \mathbf{y}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}(\tilde{y}(\mathbf{x}_{\text{new}}), \tilde{\sigma}^2(\mathbf{x}_{\text{new}})), \quad (6)$$

140 where, in case of Ordinary or Universal Kriging

$$\tilde{y}_{\sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2}(\mathbf{x}_{\text{new}}) = f_{\text{trend}}(\mathbf{x}_{\text{new}})^\top \hat{\boldsymbol{\beta}} + \mathbf{k}(\mathbf{x}_{\text{new}}, \mathbf{X})^\top \mathbf{K}^{-1}(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}), \quad (7)$$

$$\begin{aligned} \tilde{\sigma}_{\sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2}^2(\mathbf{x}_{\text{new}}) &= \mathbf{k}(\mathbf{x}_{\text{new}}, \mathbf{x}_{\text{new}}) + \sigma_\epsilon^2 - \mathbf{k}(\mathbf{x}_{\text{new}}, \mathbf{X})^\top \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}_{\text{new}}, \mathbf{X}) + (f_{\text{trend}}(\mathbf{x}_{\text{new}}) - \\ &\quad \mathbf{F}\mathbf{K}^{-1} \mathbf{k}(\mathbf{x}_{\text{new}}, \mathbf{X}))^\top (\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F})^{-1} (f_{\text{trend}}(\mathbf{x}_{\text{new}}) - \mathbf{F}\mathbf{K}^{-1} \mathbf{k}(\mathbf{x}_{\text{new}}, \mathbf{X})), \end{aligned} \quad (8)$$

and

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{y}, \\ f_{\text{trend}}(\mathbf{x}_{\text{new}}) &= (f_j(\mathbf{x}_{\text{new}}))_{j=0}^{p-1}. \end{aligned} \quad (9)$$

$\tilde{y}_{\sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2}$ (resp. $\tilde{\sigma}_{\sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2}^2$) will be also denoted by \tilde{y} (resp. $\tilde{\sigma}^2$) without specifying its dependence on hyperparameters or nugget effect when there is no possible confusion.

The most outstanding advantage of GP model compared to other models relies
145 on the previous equations (7) and (8). The predictive distribution can be used for

sensitivity analysis (J. Oakley *et al.* [23], 2004) and uncertainty quantification instead of costly methods based on Monte Carlo algorithms. Other possible considerations and extensions of GP modelling are described in (C. Currin *et al.* [24], 1991; Rasmussen and Williams [18], 2006).

150 Given a GP regression model and a point $\mathbf{x}_{\text{new}} \in \mathcal{D}$, the *posterior* law of prediction in (6) can be standardized into

$$\tilde{Z}(\mathbf{x}_{\text{new}}) = \frac{Y(\mathbf{x}_{\text{new}}) - \tilde{y}(\mathbf{x}_{\text{new}})}{\tilde{\sigma}(\mathbf{x}_{\text{new}})} \mid \mathbf{X}, \mathbf{y}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}(0, 1). \quad (10)$$

By considering the standardized variable $\tilde{Z}(\mathbf{x}_{\text{new}})$, the α -quantiles z_α are those of the standard normal law : $q_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ and $q_{\alpha/2} = \Phi^{-1}(\alpha/2) = -q_{1-\alpha/2}$ where Φ is the CDF of the standard normal distribution, such that the Prediction Intervals $\mathcal{PI}_{1-\alpha}$ in (2) can be written as

155

$$\mathcal{PI}_{1-\alpha}(\mathbf{x}_{\text{new}}) = [\tilde{y}(\mathbf{x}_{\text{new}}) - q_{1-\alpha/2} \times \tilde{\sigma}(\mathbf{x}_{\text{new}}); \tilde{y}(\mathbf{x}_{\text{new}}) + q_{1-\alpha/2} \times \tilde{\sigma}(\mathbf{x}_{\text{new}})], \quad (11)$$

which gives a natural definition for $y_{\alpha/2}$ and $y_{1-\alpha/2}$

$$y_{\alpha/2}(\mathbf{x}) = \tilde{y}(\mathbf{x}) - q_{1-\alpha/2} \times \tilde{\sigma}(\mathbf{x}) ; y_{1-\alpha/2}(\mathbf{x}) = \tilde{y}(\mathbf{x}) + q_{1-\alpha/2} \times \tilde{\sigma}(\mathbf{x}). \quad (12)$$

3.4. Training model with Maximum Likelihood method

Constructing a GP model and computing the kriging mean and variance as shown in (7) and (8) implies estimating the nugget effect σ_ϵ^2 and the covariance parameters $(\sigma^2, \boldsymbol{\theta})$. Here, we assume that σ_ϵ^2 is known or has been estimated by the method proposed in [25] for instance.

160

The Maximum Likelihood Estimator (MLE) $\hat{\sigma}_{ML}^2$ and $\hat{\boldsymbol{\theta}}_{ML}$ of σ^2 and $\boldsymbol{\theta}$ is

$$(\hat{\sigma}_{ML}^2, \hat{\boldsymbol{\theta}}_{ML}) \in \operatorname{argmin}_{\sigma^2, \boldsymbol{\theta}} \mathbf{y}^\top \left(\mathbf{K}^{-1} - \mathbf{K}^{-1} \mathbf{F} (\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{K}^{-1} \right) \mathbf{y} + \log(\det \mathbf{K}). \quad (13)$$

The MLE method is optimal (F. Bachoc [26], 2013b) when the covariance function is well-specified (i.e. when the data \mathbf{y} comes from a function f that

is a realization of a GP with covariance function that belongs to the family of
 165 covariance functions in section 3.1).

However, it may not be the case as the MLE method is poorly robust with
 respect to model misspecifications. Besides, training and assessing the quality of
 a predictor should not be done on the same data (Tibshirani *et al.* [27] in chapter
 7). In particular, the MLE method does not show how well the model will do
 170 when it is asked to make new predictions for data it has not already seen. The
 CV method represents an alternative to estimate the covariance hyperparameters
 $(\sigma^2, \boldsymbol{\theta})$ for prediction purposes and has the advantage of being more efficient and
 robust when the covariance function is misspecified (F. Bachoc [26], 2013b).

3.5. Training model with Cross-Validation method for point-wise prediction

175 We consider the same learning set of n observations $\mathbf{D}_{\text{learn}} = (\mathbf{X}, \mathbf{y}) =$
 $\{(\mathbf{x}^{(i)}, y^{(i)}), i \in \{1, \dots, n\}\}$ and we assume that the value of $\sigma_\epsilon^2 \in [0, +\infty)$
 is known. The Leave-One-Out method (i.e. n -Cross-Validation) consists in
 predicting $y^{(i)}$ by building a GP model, denoted \mathcal{GP}_{-i} and trained on $\mathbf{D}_{-i} =$
 $\{(\mathbf{x}^{(j)}, y^{(j)})\}_{j \in \{1, \dots, n\} \setminus \{i\}}$. The obtained prediction mean and variance are func-
 180 tions of parameters $(\sigma^2, \boldsymbol{\theta})$ as shown in (7) and (8) and are used to assess the
 predictive capability of the global GP model.

In the case of the Leave-One-Out method, the Mean Squared prediction
 Error (MSE) is used to assess the quality of the point-wise prediction (See [28]
 for more details about this metric) of the GP model, it can be expressed as

$$\mathcal{LOO}_{MSE}(\sigma^2, \boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \left(y^{(i)} - \tilde{y}_i \right)^2, \quad (14)$$

where \tilde{y}_i and $\tilde{\sigma}_i^2$ are the Leave-One-Out predictive mean and variance of $f(\mathbf{x}^{(i)})$
 by a GP model trained on \mathbf{D}_{-i} with the hyperparameters $(\sigma^2, \boldsymbol{\theta})$.

Hypothesis \mathcal{H}_2 : Let $(\mathbf{e}_i)_{i=1}^n$ be the canonical basis. We assume that $\mathbf{e}_i \notin \text{Im}$
 185 *\mathbf{F} for all $i \in \{1, \dots, n\}$.*

Let $\bar{\mathbf{K}}$ be the matrix defined by

$$\bar{\mathbf{K}} = \mathbf{K}^{-1} - \mathbf{K}^{-1} \mathbf{F} (\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{K}^{-1}. \quad (15)$$

For all $i \in \{1, \dots, n\}$, we have $(\bar{\mathbf{K}})_{i,i} > 0$ by Lemma 2 (see Appendix A), and, in the case of Ordinary or Universal Kriging, the Virtual Cross-Validation formulas (O. Dubrule [29], 1983) of the predictive mean \tilde{y}_i and variance $\tilde{\sigma}_i^2$ are given by

$$y^{(i)} - \tilde{y}_i = \frac{(\bar{\mathbf{K}}\mathbf{y})_i}{(\bar{\mathbf{K}})_{i,i}} \quad \text{and} \quad \tilde{\sigma}_i^2 = \frac{1}{(\bar{\mathbf{K}})_{i,i}}. \quad (16)$$

With the presence of the nugget effect, the GP regressor does not interpolate the training data \mathbf{y} but approximates them as best as possible. The Leave-One-Out method looks for the best approximation by minimizing the \mathcal{LOO}_{MSE} criterion. The criterion (14) can be written with explicit quadratic forms in \mathbf{y}

$$(\hat{\sigma}_{MSE}^2, \hat{\boldsymbol{\theta}}_{MSE}) \in \operatorname{argmin}_{\sigma^2, \boldsymbol{\theta}} \mathbf{y}^\top \bar{\mathbf{K}} \operatorname{Diag}(\bar{\mathbf{K}})^{-2} \bar{\mathbf{K}} \mathbf{y}. \quad (17)$$

Note that in the absence of the nugget effect $\sigma_\epsilon^2 = 0$, $\bar{\mathbf{K}}$ is of the form $\sigma^{-2} \bar{\mathbf{R}}_\theta$ where $\bar{\mathbf{R}}_\theta$ does not depend on σ^2 , the predictive variance $\hat{\sigma}_{MSE}^2$ can be computed explicitly without joint optimization

$$\hat{\boldsymbol{\theta}}_{MSE} \in \operatorname{argmin}_{\boldsymbol{\theta}} \mathbf{y}^\top \bar{\mathbf{R}}_\theta \operatorname{Diag}(\bar{\mathbf{R}}_\theta)^{-2} \bar{\mathbf{R}}_\theta \mathbf{y}, \quad (18)$$

$$\hat{\sigma}_{MSE}^2 = \mathbf{y}^\top \bar{\mathbf{R}}_{\hat{\boldsymbol{\theta}}_{MSE}} \operatorname{Diag}(\bar{\mathbf{R}}_{\hat{\boldsymbol{\theta}}_{MSE}})^{-1} \bar{\mathbf{R}}_{\hat{\boldsymbol{\theta}}_{MSE}} \mathbf{y}. \quad (19)$$

4. Prediction Intervals estimation for Gaussian Processes

190 The MSE hyperparameters $(\hat{\sigma}_{MSE}^2, \hat{\boldsymbol{\theta}}_{MSE})$ are obtained from a point-wise prediction metric and do not focus on Prediction Intervals neither on quantifying the uncertainty of the model. For these purposes, using the CP is more appropriate.

The *Coverage Probability* (CP) is defined as the probability that the Prediction Interval procedure will produce an interval that captures what it is intended to capture (Hong *et al.* [30], 2009). In the Leave-One-Out framework, we keep the notations of \tilde{y}_i and $\tilde{\sigma}_i^2$: the predictive mean and variance on $\mathbf{x}^{(i)} \in \mathbf{X}$ using the learning set $\mathbf{D}_{-i} = \{(\mathbf{x}^{(j)}, y^{(j)})\}_{j \in \{1, \dots, n\} \setminus \{i\}}$. We define then the

Leave-One-Out CP $\tilde{\mathbb{P}}_{1-\alpha}$ as the percentage of observed values \mathbf{y} belonging to Prediction Intervals $\mathcal{PI}_{1-\alpha}$ of \tilde{y}_i for all $i \in \{1, \dots, n\}$

$$\tilde{\mathbb{P}}_{1-\alpha} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{\tilde{y}_i + q_{\alpha/2} \times \tilde{\sigma}_i < y^{(i)} \leq \tilde{y}_i + q_{1-\alpha/2} \times \tilde{\sigma}_i\}, \quad (20)$$

where q_a is the a -quantile of the standard normal law and $\mathbf{1}\{A\}$ is the indicator function of A . We introduce the Heaviside step function h

$$h(x) = \mathbf{1}\{x \geq 0\} = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (21)$$

The Leave-One-Out CP $\tilde{\mathbb{P}}_{1-\alpha}$ can be written as

$$\tilde{\mathbb{P}}_{1-\alpha} = \frac{1}{n} \sum_{i=1}^n h\left(q_{1-\alpha/2} - \frac{y^{(i)} - \tilde{y}_i}{\tilde{\sigma}_i}\right) - \frac{1}{n} \sum_{i=1}^n h\left(q_{\alpha/2} - \frac{y^{(i)} - \tilde{y}_i}{\tilde{\sigma}_i}\right). \quad (22)$$

When the model is well-specified, the coverage of the Prediction Intervals $\mathcal{PI}_{1-\alpha}$ is optimal as the predictive distribution is fully characterized by the Gaussian law (see section 3.1), each term of the right-hand side of (22) is an unbiased estimator of the probability

$$\mathbb{P}\left(\frac{Y(\mathbf{x}^{(i)}) - \tilde{y}_i}{\tilde{\sigma}_i} \leq q_{1-\alpha/2} \mid \mathbf{D}_{-i}\right) = 1 - \alpha/2, \quad (23)$$

and

$$\mathbb{P}\left(\frac{Y(\mathbf{x}^{(i)}) - \tilde{y}_i}{\tilde{\sigma}_i} \leq q_{\alpha/2} \mid \mathbf{D}_{-i}\right) = \alpha/2. \quad (24)$$

Conversely, if the model is misspecified, each predictive quantile, needs to be
 195 quantified properly with respect to the normal law quantile so that the CP as described in section 2 achieves the desired level.

Let $a \in (0, 1/2) \cup (1/2, 1)$ describe a nominal level of quantile. We define the *quasi-Gaussian* proportion ψ_a as a map from $[0, +\infty) \times (0, +\infty)^d$ to $[0, 1]$

$$\psi_a(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n h\left(q_a - \frac{y^{(i)} - \tilde{y}_i}{\tilde{\sigma}_i}\right), \quad (25)$$

where \tilde{y}_i and $\tilde{\sigma}_i$ are the predictive mean and variance on $\mathbf{x}^{(i)}$ using the learning set \mathbf{D}_{-i} and the hyperparameters $(\sigma^2, \boldsymbol{\theta})$. Given the Virtual Cross-Validation formulas (O. Dubrule [29], 1983), ψ_a can be written in terms of

covariance matrix $\bar{\mathbf{K}}$

$$\psi_a(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n h \left(q_a - \frac{(\bar{\mathbf{K}}\mathbf{y})_i}{\sqrt{(\bar{\mathbf{K}})_{i,i}}} \right). \quad (26)$$

The *quasi-Gaussian* proportion ψ_a describes how close the a -quantile q_a of the standardized predictive distribution is to the level a , if possible, it should correspond to a . Therefore, the objective is to fit the hyperparameters $(\sigma^2, \boldsymbol{\theta})$ according to the *quasi-Gaussian* proportions and to find two pairs $(\bar{\sigma}^2, \bar{\boldsymbol{\theta}})$ and $(\underline{\sigma}^2, \underline{\boldsymbol{\theta}})$ such that $\psi_{1-\alpha/2}(\bar{\sigma}^2, \bar{\boldsymbol{\theta}}) = 1 - \alpha/2$ and $\psi_{\alpha/2}(\underline{\sigma}^2, \underline{\boldsymbol{\theta}}) = \alpha/2$. This allows us to get the optimal Leave-One-Out CP by respecting the nominal confidence level $(1 - \alpha)$, that is $\tilde{\mathbb{P}}_{1-\alpha} = 1 - \alpha$.

4.1. Presence of nugget effect

In this section, we assume $\sigma_\epsilon^2 > 0$. The *quasi-Gaussian* proportion ψ_a is, however, piecewise constant and takes exactly n values in $\{k/n, k \in \{0, \dots, n\}\}$. We first need to modify the problem $\psi_a(\sigma^2, \boldsymbol{\theta}) = a$. Let $\delta > 0$, we define the continuous functions h_δ^- and h_δ^+

$$\begin{aligned} h_\delta^+(x) &= \begin{cases} 1 & \text{if } x > \delta \\ x/\delta & \text{if } 0 < x \leq \delta \\ 0 & \text{otherwise} \end{cases} \\ h_\delta^-(x) &= \begin{cases} 1 & \text{if } x \geq 0 \\ 1 + x/\delta & \text{if } -\delta \leq x < 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (27)$$

If $a > 1/2$ we define

$$\psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n h_\delta^+ \left(q_a - \frac{(\bar{\mathbf{K}}\mathbf{y})_i}{\sqrt{(\bar{\mathbf{K}})_{i,i}}} \right). \quad (28)$$

If $a < 1/2$ we define

$$\psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n h_\delta^- \left(q_a - \frac{(\bar{\mathbf{K}}\mathbf{y})_i}{\sqrt{(\bar{\mathbf{K}})_{i,i}}} \right). \quad (29)$$

Let $\delta > 0$ be a small enough so that $\delta < q_a$ if $a > 1/2$ (respectively, $\delta < q_{1-a}$ if $a < 1/2$) in such a way that $h_\delta^+(q_a) = 1$ (respectively, $h_\delta^-(q_a) = 0$). We consider the problem

$$\psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) = a, \quad (30)$$

and we denote by $\mathcal{A}_{a,\delta}$ the solution set of the problem (30)

$$\mathcal{A}_{a,\delta} := \left\{ (\sigma^2, \boldsymbol{\theta}) \in [0, +\infty) \times (0, +\infty)^d, \psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) = a \right\}. \quad (31)$$

205 *Hypothesis \mathcal{H}_3 :* We assume that $k_\epsilon = \text{Card}\{i \in \{1, \dots, n\}, \frac{(\boldsymbol{\Pi}\mathbf{y})_i}{\sqrt{(\boldsymbol{\Pi})_{ii}}} \leq \sigma_\epsilon q_a\} < na$ if $a > 1/2$ and $k_\epsilon = \text{Card}\{i \in \{1, \dots, n\}, \frac{(\boldsymbol{\Pi}\mathbf{y})_i}{\sqrt{(\boldsymbol{\Pi})_{ii}}} \leq \sigma_\epsilon q_a\} > na$ if $a < 1/2$ where $\boldsymbol{\Pi}$ is the orthogonal projection matrix on $(\text{Im } \mathbf{F})^\perp$ such that $\boldsymbol{\Pi} = \mathbf{I}_n - \mathbf{F}(\mathbf{F}^\top \mathbf{F})^{-1} \mathbf{F}^\top$.

Remark 2. The hypothesis \mathcal{H}_3 is satisfied in Ordinary and Universal Kriging, except for very special experimental designs \mathbf{X} . Moreover, $\boldsymbol{\Pi}$ is the projection on the space $(\text{Im } \mathbf{F})^\perp$ and is expected to remove the trend of the model. It is reasonable to think that $(\boldsymbol{\Pi}\mathbf{y})_i$ is centered and that

$$\text{Card}\{i \in \{1, \dots, n\}, (\boldsymbol{\Pi}\mathbf{y})_i \leq 0\} \approx \frac{n}{2}. \quad (32)$$

If σ_ϵ^2 is smaller than σ^2 , then we have also

$$\text{Card}\{i \in \{1, \dots, n\}, \frac{(\boldsymbol{\Pi}\mathbf{y})_i}{\sqrt{(\boldsymbol{\Pi})_{ii}}} \leq \sigma_\epsilon q_a\} \approx \frac{n}{2}, \quad (33)$$

so that the hypothesis \mathcal{H}_3 is not restrictive.

210

Proposition 1. Let us assume the hypotheses \mathcal{H}_1 , \mathcal{H}_2 and \mathcal{H}_3 , then $\mathcal{A}_{a,\delta}$ is non-empty.

Proof. In Appendix A. □

The challenge now is to identify and choose wisely the optimal solutions
 215 $(\sigma_{\text{opt}}^2, \boldsymbol{\theta}_{\text{opt}}) \in \mathcal{A}_{a,\delta}$. Some authors (Khosravi *et al.* [16], 2011a) suggest the mean Prediction Intervals width (MWPI) of Prediction Intervals $\mathcal{PI}_{1-\alpha}$ as an additional constraint to reduce the set of solutions. However, this constraint

may not work when dealing with quantile estimation because the lower bound of the corresponding interval is infinite.

Instead, we will compare these solutions with MLE's solution $(\hat{\sigma}_{ML}^2, \hat{\boldsymbol{\theta}}_{ML})$ (subsection 3.4) or MSE CV solution $(\hat{\sigma}_{MSE}^2, \hat{\boldsymbol{\theta}}_{MSE})$ (subsection 3.5) and we will take the closest pair $(\sigma_{\text{opt}}^2, \boldsymbol{\theta}_{\text{opt}})$ by using an appropriate notion of similarity between multivariate Gaussian distributions. Ideally, we aim to solve the following problem

$$\operatorname{argmin}_{(\sigma^2, \boldsymbol{\theta}) \in \mathcal{A}_{a, \delta}} d^2((\sigma^2, \boldsymbol{\theta}), (\sigma_0^2, \boldsymbol{\theta}_0)), \quad (34)$$

220 where d is a continuous similarity measure of hyperparameters $(\sigma^2, \boldsymbol{\theta})$ operating on the mean vector \mathbf{m} and covariance matrix \mathbf{K} , and $(\sigma_0^2, \boldsymbol{\theta}_0) \in \{(\hat{\sigma}_{ML}^2, \hat{\boldsymbol{\theta}}_{ML}), (\hat{\sigma}_{MSE}^2, \hat{\boldsymbol{\theta}}_{MSE})\}$ as described in (13) or (17).

The resolution of the problem (34) may be too costly and heavy to solve when the dimension is high, say $d \geq 10$. An alternative is to apply *the relaxation* 225 method where we redefine this optimization problem of $\boldsymbol{\theta}$ from $(0, +\infty)^d$ to $(0, +\infty)$ by shifting the length-scale vector $\boldsymbol{\theta}_0$ by a parameter $\lambda \in (0, +\infty)$.

Let $\boldsymbol{\theta}_0$ denote a solution of the problems (13) or (17) and for $\lambda \in (0, +\infty)$, let $H_\delta(\lambda)$ denote the subset

$$H_\delta(\lambda) = \{\sigma^2 \in [0, +\infty), \psi_a^{(\delta)}(\sigma^2, \lambda \boldsymbol{\theta}_0) = a\}. \quad (35)$$

Hypothesis \mathcal{H}_4 : The set-valued mapping (the so-called correspondence function) $H_\delta : (0, +\infty) \rightarrow \mathcal{P}((0, +\infty))$, where $\mathcal{P}(S)$ denotes the power set of a set S , is lower semi-continuous, that is, for all $\lambda \in (0, +\infty)$, for each open set \mathcal{U} with 230 *$H_\delta(\lambda) \cap \mathcal{U} \neq \emptyset$, there exists a neighborhood $\mathcal{O}(\lambda)$ such that if $\lambda^* \in \mathcal{O}(\lambda)$ then $H_\delta(\lambda^*) \cap \mathcal{U} \neq \emptyset$.*

In the kriging framework, σ^2 should be as small as possible to reduce the uncertainty of the model, a natural choice of σ_{opt}^2 is

$$\forall \lambda \in (0, +\infty) : \sigma_{\text{opt}}^2(\lambda) := \min\{\sigma^2 \in [0, +\infty), \psi_a^{(\delta)}(\sigma^2, \lambda \boldsymbol{\theta}_0) = a\}. \quad (36)$$

Proposition 2. *The function $\lambda \mapsto \sigma_{\text{opt}}^2(\lambda)$ is well-defined under hypotheses \mathcal{H}_1 to \mathcal{H}_3 , and continuous on $(0, +\infty)$ under hypothesis \mathcal{H}_4 .*

Proof. In Appendix A. \square

Concerning the choice of d , one known similarity measure between probability distributions is the Wasserstein distance, widely used in optimal transportation problems (see Chapter 6 of [31] for more details). In case of two Gaussian random distributions $\mathcal{N}(\mathbf{m}_1, \mathbf{K}_1)$ and $\mathcal{N}(\mathbf{m}_2, \mathbf{K}_2)$, the second Wasserstein distance is equal to

$$W_2^2(\mathcal{N}(\mathbf{m}_1, \mathbf{K}_1), \mathcal{N}(\mathbf{m}_2, \mathbf{K}_2)) = \|\mathbf{m}_1 - \mathbf{m}_2\|^2 + \text{Tr} \left(\mathbf{K}_1 + \mathbf{K}_2 - 2\sqrt{\mathbf{K}_1^{1/2} \mathbf{K}_2 \mathbf{K}_1^{1/2}} \right), \quad (37)$$

235 where, in our setting, $\mathbf{m}_1 = \mathbf{F}\hat{\boldsymbol{\beta}}_1 = (\mathbf{F}^\top \mathbf{K}_1^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{K}_1^{-1} \mathbf{y}$ and $\mathbf{m}_2 = \mathbf{F}\hat{\boldsymbol{\beta}}_2 = (\mathbf{F}^\top \mathbf{K}_2^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{K}_2^{-1} \mathbf{y}$.

Therefore, each couple $(\sigma^2, \boldsymbol{\theta})$ is associated to a Gaussian distribution $\mathcal{N}(\mathbf{m}, \mathbf{K})$ and we define the similarity measure d as

$$d^2((\sigma^2, \boldsymbol{\theta}), (\sigma_0^2, \boldsymbol{\theta}_0)) = W_2^2(\mathcal{N}(\mathbf{m}, \mathbf{K}), \mathcal{N}(\mathbf{m}_0, \mathbf{K}_0)) \quad (38)$$

The choice of the second Wasserstein distance d^2 and σ_{opt}^2 makes the Prediction Intervals $\mathcal{PI}_{1-\alpha}$ shorter without the need for an additional metric like the MPIW and without affecting significantly the point-wise prediction of the model, as we will see in section 5.

240 The *relaxed* optimisation problem in (30) for the quantile estimation is given by the problem \mathcal{P}_λ

$$\mathcal{P}_\lambda : \quad \text{argmin}_{\lambda \in (0, +\infty)} \mathcal{L}(\lambda) := d^2((\sigma_{\text{opt}}^2(\lambda), \lambda \boldsymbol{\theta}_0), (\sigma_0^2, \boldsymbol{\theta}_0)) \quad (39)$$

Proposition 3. *Under all hypotheses \mathcal{H}_1 to \mathcal{H}_4 , the function $\mathcal{L} : (0, +\infty) \rightarrow \mathbb{R}^+$ is continuous and coercive on $(0, +\infty)$. The problem \mathcal{P}_λ admits at least one global minimizer λ^* in $(0, +\infty)$.*

Proof. In Appendix A. \square

245 **Remark 3.** *The coercivity of the function \mathcal{L} is guaranteed by the hypotheses \mathcal{H}_1 to \mathcal{H}_3 (see Appendix A). The function \mathcal{L} is also upper semi-continuous (J.Zhao*

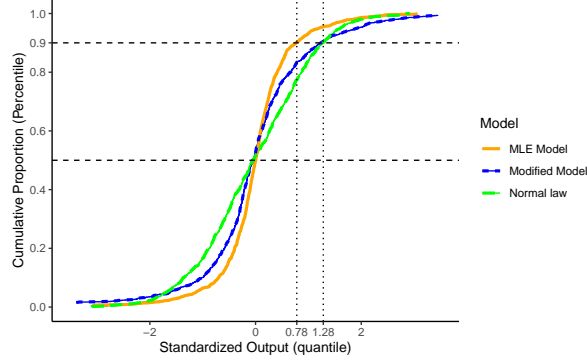


Figure 1. Illustration of the relaxation effect on the ECDF of the Leave-One-Out standardized predictive distribution on the quantile of level $a = 90\%$; The relaxed standardized predictive distribution coincides with the standard normal law distribution on point $(q_a, a) = (1.28, 0.90)$ instead of $(\psi_a, a) = (0.78, 0.90)$

Green : standard normal law - MLE standardized Predictive distribution when the model is well-specified; Orange : MLE standardized predictive distribution when the model is misspecified; Blue : standardized predictive distribution after relaxing model's hyperparameters.

[32], 1997). The hypothesis \mathcal{H}_4 insures that \mathcal{L} is continuous and that there exists a global minimizer. This hypothesis is not easy to check. If it does not hold or if it cannot be checked, then it is possible to solve the problem (39) on a discrete grid with a grid step δ for instance.

Let $\hat{\beta}_{\text{opt}}$ denote the corresponding regression parameter $\hat{\beta}_{\text{opt}} = \left(\mathbf{F}^\top \mathbf{K}_{\sigma_{\text{opt}}^2(\lambda^*), \lambda^* \theta_0}^{-1} \mathbf{F} \right)^{-1} \mathbf{F}^\top \mathbf{K}_{\sigma_{\text{opt}}^2(\lambda^*), \lambda^* \theta_0}^{-1} \mathbf{y}$. The purpose of this resolution is to create a GP model with hyperparameters $(\hat{\beta}_{\text{opt}}, \sigma_{\text{opt}}^2(\lambda^*), \lambda^* \theta_0)$ able to predict the quantile \tilde{y}_a such that a proportion a of true values are below \tilde{y}_a with respect to the constraint of *quasi-Gaussian* proportion ψ_a (see Figure 1). Finally, the Prediction Intervals $\mathcal{PI}_{1-\alpha}$ will be obtained using two GP models built with the same method, one for the upper quantile $1 - \alpha/2$ with optimal relaxation parameter $\bar{\lambda}^*$ and the other for the lower quantile of $\alpha/2$ with parameter $\underline{\lambda}^*$. The CP of $\mathcal{PI}_{1-\alpha}$ is optimal and insured by respecting the coverage of each quantile as shown in (22). In the following, we call this method *Robust Prediction Intervals Estimation* (RPIE).

4.2. Absence of nugget effect

When the nugget effect is null $\sigma_\epsilon^2 = 0$, the set of solutions $\mathcal{A}_{a,\delta}$ is still non-empty because one can show that, for $\boldsymbol{\theta}$ in the neighborhood of $\mathbf{0} \in \mathbb{R}^d$, the problem $\psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) = a$ has a solution $\sigma^2 \in (0, +\infty)$ (see Appendix B). In particular, the correspondence function H_δ is non-empty valued for $\lambda > 0$ small enough. But it may be empty valued for some large $\lambda \in (0, +\infty)$.

Let $\lambda > 0$ and let k_λ be the integer defined by

$$k_\lambda := \text{Card} \{i \in \{1, \dots, n\}, (\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0} \mathbf{y})_i \leq 0\}. \quad (40)$$

The existence of $H_\delta(\lambda)$ and $\sigma_{\text{opt}}^2(\lambda)$ depends on the condition $k_\lambda \leq na$ that may fail to hold for all $\lambda \in (0, +\infty)$. We shall assume some additional hypotheses on \mathbf{y} , indeed, when the model is well-specified, that is, there exist hyperparameters $(\boldsymbol{\beta}_*, \sigma_*^2, \boldsymbol{\theta}_*)$ such that \mathbf{y} corresponds to a realization of random vector $\mathbf{Y} \sim \mathcal{N}(\mathbf{F}\boldsymbol{\beta}_*, \sigma_*^2 \mathbf{R}_{\boldsymbol{\theta}_*})$. One has then

$$\overline{\mathbf{R}}_{\boldsymbol{\theta}_*} \mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \sigma_*^2 \overline{\mathbf{R}}_{\boldsymbol{\theta}_*}). \quad (41)$$

and, for a large number of observations n , we can anticipate that

$$\text{Card} \{i \in \{1, \dots, n\}, (\overline{\mathbf{R}}_{\boldsymbol{\theta}_*} \mathbf{y})_i \leq 0\} \approx \frac{n}{2}. \quad (42)$$

Hence, the condition $n/2 < k_\lambda \leq na$ is satisfied in a neighborhood of $\boldsymbol{\theta}_*$. One may think that, for a family of vectors $\mathbf{y} \in \mathbb{R}^n$, this condition is satisfied around $\lambda = 1$. In addition, if $\sigma_{\text{opt}}^2(\lambda)$ exists for all $\lambda \in (0, +\infty)$, its limit would diverge when $\lambda \rightarrow +\infty$. If it is not the case, the function \mathcal{L} is no more continuous, but the problem can be solved by the Grid Search method.

5. Numerical Results

5.1. Test cases with analytical functions

In this section, we give three numerical examples to illustrate Prediction Intervals estimation by the RPIE method. We show that for the *Wing-Weight* function, a well-specified model, the CP is optimal and no optimization is required.

For *Zhou* and *Morokoff & Caflisch* functions where the model is misspecified, we solve then the problem (39) using the Wasserstein distance. We build specific GP models following the RPIE method described in section 4 for some level a . The following metrics : the Leave-One-Out CP $\tilde{\mathbb{P}}_{1-\alpha}$ defined in (22), the Coverage Probability (CP), the mean (MPIW) standard-deviation (SdPIW) of Prediction Intervals width, and the accuracy Q^2 (Kleijnen and Sargent [33], 2000) are used to asses and compare GP models built either by standard methods (MLE or MSE CV) or the RPIE method:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n_{test}} \left(y_{test}^{(i)} - \tilde{y}_{i,test} \right)^2}{\sum_{i=1}^{n_{test}} \left(y_{test}^{(i)} - \bar{y} \right)^2}, \quad (43)$$

$$CP_{a,b} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbf{1}_{y_{test}^{(i)} \in \mathcal{PI}_{a,b}(\mathbf{x}_{test}^{(i)})}, \quad (44)$$

$$MPIW_{a,b} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left| \mathcal{PI}_{a,b}(\mathbf{x}_{test}^{(i)}) \right|, \quad (45)$$

$$SdPIW_{a,b} = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left[\left| \mathcal{PI}_{a,b}(\mathbf{x}_{test}^{(i)}) \right| - MPIW_{a,b} \right]^2}, \quad (46)$$

where $\mathbf{y}_{test} = \left(y_{test}^{(1)}, \dots, y_{test}^{(n_{test})} \right)$ is the vector to predict of length n_{test} at $\left(\mathbf{x}_{test}^{(1)}, \dots, \mathbf{x}_{test}^{(n_{test})} \right)$, $\mathcal{PI}_{a,b}$ is the Prediction Interval delimited by the quantiles q_a and q_b , and $|\mathcal{PI}_{a,b}|$ is the length of the interval.

In particular, $CP_{\alpha/2, 1-\alpha/2} = CP_{1-\alpha}$ denotes the CP of points belonging to the $(1 - \alpha) \times 100\%$ Prediction Interval. Analogously, $MPIW_{1-\alpha}$ and $SdPIW_{1-\alpha}$ refer to the mean and standard deviation of $(1 - \alpha)$ Prediction Intervals length. Note that the CP may be different from the Leave-One-Out CP $\tilde{\mathbb{P}}_{1-\alpha}$, this case can happen when the distributions of the training and validation sets are different. However, a Leave-One-Out CP $\tilde{\mathbb{P}}_{1-\alpha}$ close to $1 - \alpha$ insures that, if the distributions are similar, $CP_{1-\alpha}$ is also close to $1 - \alpha$.

This subsection provides results obtained on $d = 10$ -dimensional GP with constant mean function (Ordinary Kriging). The value of δ is fixed at $\delta = 10^{-2}$.

Example 1 : Well-specified model - The Wing Weight function

The Wing Weight function is a model in dimension $d = 10$ proposed by Forrester *et al.* ([34], 2008) that estimates the weight of a light aircraft wing.

300 For an input vector $\mathbf{x} \in \mathbb{R}^{10}$, the response y is :

$$f(\mathbf{x}) = 0.036x_1^{0.758}x_2^{0.0035}\left(\frac{x_3}{\cos^2(x_4)}\right)^{0.6}x_5^{0.006}x_6^{0.04}\left(\frac{100x_7}{\cos(x_4)}\right)^{-0.3}(x_8x_9)^{0.49}+x_1x_{10}. \quad (47)$$

The components x_i are assumed to vary over the ranges given in Table 1 (see [34] and [35] for details).

Table 1. The input variables x_j and their domain ranges $[a_j; b_j]$

Component	Domain	Component	Domain
x_1	[150; 200]	x_6	[0.5; 1]
x_2	[220; 300]	x_7	[0.08; 0.18]
x_3	[6, 10]	x_8	[2.5; 6]
x_4	[-10; 10]	x_9	[1700, 2500]
x_5	[16; 45]	x_{10}	[0.025; 0.08]

We create an experimental design \mathbf{X} of $n = 600$ observations and $d = 10$ variables where observations $\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_d^{(i)})$ are sampled i.i.d with uniform 305 distribution over $\bigotimes_{j=1}^d [a_j, b_j]$. We generate the response $\mathbf{y} = (y^{(1)}, \dots, y^{(n)})$ such that $y^{(i)} = f(\mathbf{x}^{(i)}) + \epsilon^{(i)}$ with f defined in (47) and $\epsilon^{(i)}$ are sampled i.i.d. with the distribution $\mathcal{N}(0, \sigma_\epsilon^2 = 16)$. Here the nugget effect is estimated with the methodology described in [25] and the covariance kernel is the Matérn 3/2.

The GP model is trained on 75% of the data (25% of data is left for validation). 310 The diagnostics of the model are presented in Table 3 with the metrics described above. The accuracy Q^2 is correct for the MLE solution but the MSE CV does much better, an expected result since the MSE CV method is more adapted

Table 3. Diagnostics of the model built by MLE and CV methods ; confidence level
 $1 - \alpha = 80\%$

	MLE (section 3.4)	MSE CV (section 3.5)
Q^2	0.583	0.769
$\tilde{\mathbb{P}}_{1-\alpha}$	80.2	95.6

Q^2 : Accuracy and $\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in %.

for point-wise prediction criterion. However, the Leave-One-Out CP $\tilde{\mathbb{P}}_{1-\alpha}$ is far from the required level, which means that they were poorly estimated with
315 point-wise prediction criterion.

Table 4. Quantile estimations for MLE GP model on the training set (*Quasi-Gaussian* proportion) and validation set (CP)

	Confidence level $1 - \alpha = 90\%$	Confidence level $1 - \alpha = 80\%$
$\psi_{1-\alpha/2}$	93.3	89.1
$\psi_{\alpha/2}$	2.89	8.89
$\text{CP}_{-\infty, 1-\alpha/2}$	94.6	94.0
$\text{CP}_{-\infty, \alpha/2}$	4.67	14.0
$\tilde{\mathbb{P}}_{1-\alpha}$	90.4	80.2

CP: Coverage Probability, ψ : *Quasi-Gaussian* proportion and $\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP; in %.

Table 4 shows in particular that the model is well-specified for Matérn 3/2 correlation kernel when applying the MLE method since the CP is optimal and close to required level, the quantiles have been correctly estimated despite a little difference from the true bound.

320 Example 1 is a case of well-specified model in which the coverage probabilities obtained by the MLE method are good and the RPIE method does not bring

additional value.

Example 2 : Misspecified model with noise - Morokoff & Caflisch function -

325 We consider the Morokoff and Caflisch ([36], 1995) function defined on $[0, 1]^d$ by

$$f(\mathbf{x}) = \frac{1}{2} \left(1 + \frac{1}{d}\right)^d \prod_{i=1}^d (x_i)^{1/d}. \quad (48)$$

In Example 2, we consider an experimental design \mathbf{X} of $n = 600$ observations and $d = 10$ correlated inputs. Each observation has the form $\mathbf{x}^{(i)} = (\Phi(z_1^{(i)}), \dots, \Phi(z_d^{(i)})) \in \mathbb{R}^d$, Φ is the CDF of the standard normal distribution, $\mathbf{z}^{(i)}$ are sampled from the multivariate distribution $\mathcal{N}(\mathbf{0}, \mathbf{C})$ and $\mathbf{C} \in \mathbb{R}^{d \times d}$ is the following covariance matrix :

$$\mathbf{C} = \begin{bmatrix} 1 & 0.90 & 0 & 0 & 0 & 0.50 & -0.30 & 0 & 0 & 0 \\ 0.90 & 1 & 0 & 0 & 0 & 0 & 0 & 0.10 & 0 & 0 \\ 0 & 0 & 1 & 0 & -0.30 & 0.10 & 0.40 & 0 & 0.05 & 0 \\ 0 & 0 & 0 & 1 & 0.40 & 0 & 0 & -0.35 & 0 & 0 \\ 0 & 0 & -0.30 & 0.40 & 1 & 0 & 0 & 0 & 0.10 & 0 \\ 0.05 & 0 & 0.10 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -0.30 & 0 & 0.40 & 0 & 0 & 0 & 1 & 0 & 0 & -0.30 \\ 0 & 0.1 & 0 & -0.35 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0.05 & 0 & 0.10 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.3 & 0 & 0 & 1 \end{bmatrix}. \quad (49)$$

The response vector \mathbf{y} is generated as $y^{(i)} = f(\mathbf{x}^{(i)}) + \epsilon^{(i)}$ with f the *Morokoff & Caflisch* function defined in (48) and $\epsilon^{(i)}$ are sampled i.i.d. with the distribution $\mathcal{N}(0, \sigma_\epsilon^2 = 10^{-4})$. We consider the Matérn anisotropic geometric correlation
330 model with smoothness 5/2 as covariance model and we study the Prediction Interval's problem with a nugget effect estimated with the methodology [25].

The model is not well-specified as example 1. Table 5 summarizes the results of MLE and MSE CV estimations. The accuracy Q^2 of both models is satisfactory

Table 5. Results for MLE and MSE CV methods for *Morokoff & Caflisch* function (48) ;
confidence level $1 - \alpha = 90\%$

	MLE (section 3.4)	MSE CV (section 3.5)
Q^2	0.936	0.949
$\tilde{\mathbb{P}}_{1-\alpha}$	94.3	99.1
$\text{CP}_{1-\alpha}$	95.3	98.7
$\text{MPIW}_{1-\alpha}$	$1.49 \cdot 10^{-1}$	$1.66 \cdot 10^{-1}$
$\text{SdPIW}_{1-\alpha}$	$7.30 \cdot 10^{-3}$	$5.01 \cdot 10^{-3}$

Q^2 : Accuracy; $\text{CP}_{1-\alpha}$: Coverage Probability in %; $\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in %.
MPIW: Mean of Prediction Interval widths and SdPIW: standard deviation of Prediction Interval widths.

and is improved when applying the MSE CV method. However, the Prediction
335 Intervals are overestimated for both models and the CP does not correspond to
the required level of 90%, the MSE CV model performs even worse.

Table 6. Performances of the models $\mathbf{P}_{\alpha/2}$ and $\mathbf{P}_{1-\alpha/2}$ for *Morokoff & Caflisch* function (48)
; confidence level $1 - \alpha = 90\%$

	$\mathbf{P}_{\alpha/2}$ model		$\mathbf{P}_{1-\alpha/2}$ model	
	MLE $\hat{\boldsymbol{\theta}}_{ML}$	MSE CV $\hat{\boldsymbol{\theta}}_{MSE}$	MLE $\hat{\boldsymbol{\theta}}_{ML}$	MSE CV $\hat{\boldsymbol{\theta}}_{MSE}$
λ^*	1.03	1.36	1.73	2.21
$\tilde{\mathbb{P}}_{1-\alpha}$	94.0	93.8	83.7	84.3
$\text{CP}_{1-\alpha}$	94.0	94.7	85.3	90.0

λ^* : Optimal relaxation parameter; CP: Coverage Probability in % and $\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in %.

We now address the problem of Prediction Intervals Estimation for each
solution of MLE $\hat{\boldsymbol{\theta}}_{ML}$ and MSE CV $\hat{\boldsymbol{\theta}}_{MSE}$. We consider the upper and lower

bounds $1 - \alpha/2 = 95\%$ and $\alpha/2 = 5\%$. We apply the RPIE method as described
 in section 4 and we solve the problem (39). With the optimal values $\bar{\lambda}^*$ and
 $\underline{\lambda}^*$, we build two GP models, denoted $\mathbf{P}_{1-\alpha/2}$ and $\mathbf{P}_{\alpha/2}$. Figure 2 shows the
 variation of the function \mathcal{L} for *Morokoff & Caflisch* function while resolving the
 problem (39) on the upper bound $1 - \alpha/2 = 95\%$, it illustrates the statement of
 Proposition 3 : \mathcal{L} is continuous and coercive on $(0, +\infty)$ and reaches a global
 minimum.

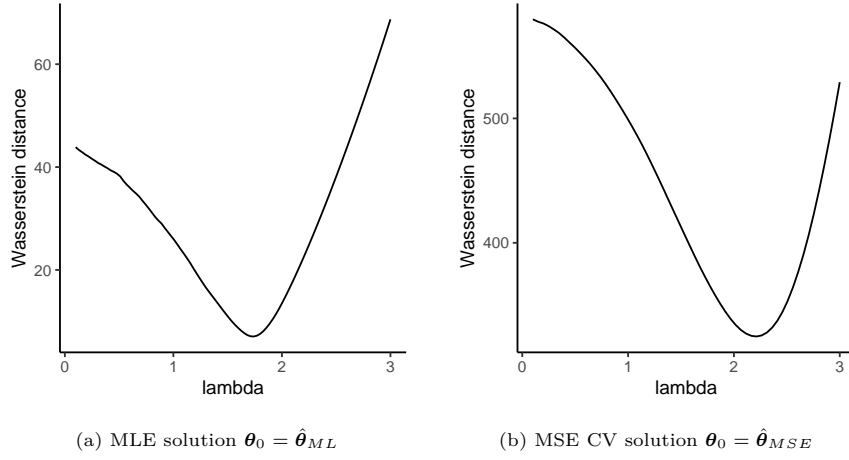


Figure 2. : The variation of the *relaxed* Wasserstein distance \mathcal{L} for *Morokoff & Caflisch*
 function; $a = 1 - \alpha/2 = 95\%$

The optimal *relaxation* parameters and the coverage probabilities of $\mathbf{P}_{\alpha/2}$
 and $\mathbf{P}_{1-\alpha/2}$ models are given in Table 6. The *relaxation* parameter λ^* is closer
 to 1 for the $\mathbf{P}_{\alpha/2}$ model than it is for the $\mathbf{P}_{1-\alpha/2}$ model. It means that the upper
 bounds of Prediction Intervals of the standards methods (MLE and MSE CV)
 were poorly estimated compared to the lower bounds. Note that each model is
 built to predict either the upper ($1 - \alpha/2 = 95\%$) or the lower bound ($\alpha/2 = 5\%$).
 As a consequence, the poor estimation of the other bound does not affect the
 RPIE method.

We consider now the Prediction Intervals whose upper and lower bounds are
 estimated respectively by $\mathbf{P}_{1-\alpha/2}$ and $\mathbf{P}_{\alpha/2}$ models. In table 7, one observes that
 these Prediction Intervals are twice shorter compared to those of MLE or CV

Table 7. Performances of Prediction Intervals obtained by the RPIE method for *Morokoff* & *Caflisch* function (48) ; confidence level $1 - \alpha = 90\%$

	MLE $\hat{\theta}_{ML}$	MSE CV $\hat{\theta}_{MSE}$
$\tilde{\mathbb{P}}_{1-\alpha}$	90.0	90.0
CP $_{1-\alpha}$	88.7	92.0
MPIW $_{1-\alpha}$	$5.57 \cdot 10^{-2}$	$6.18 \cdot 10^{-2}$
SdPIW $_{1-\alpha}$	$1.19 \cdot 10^{-2}$	$1.53 \cdot 10^{-2}$

$\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in %; CP: coverage Probability in %; MPIW: Mean of Prediction Interval widths and SdPIW: standard deviation of Prediction Interval widths.

MSE models, the coverage rate of $1 - \alpha = 90\%$ on the training set is achieved, which is the main objective of the RPIE method, and the CP on the validation set is close to this level.

360 Example 2 is a case of misspecified model with noise in which the CP obtained by MLE and MSE CV methods are not good and the RPIE method reduces Prediction Intervals length and improves the robustness of Prediction Intervals in such way that they cover as best as possible the optimal coverage rate.

Example 3 : Misspecified model without noise - Zhou function -

The Zhou ([37], 1998) function, considered initially for the numerical integration of spiky functions, is defined on $[0, 1]^d$ by

$$f(\mathbf{x}) = \frac{10^d}{2} \left[\phi\left(10\left(\mathbf{x} - \frac{1}{3}\right)\right) + \phi\left(10\left(\mathbf{x} - \frac{2}{3}\right)\right) \right], \quad (50)$$

where

$$\phi(\mathbf{x}) = (2\pi)^{-d/2} \exp(-0.5\|\mathbf{x}\|^2). \quad (51)$$

In Example 3, we create an experimental design \mathbf{X} similar to Example 1, containing $n = 600$ and $d = 10$ variables where observations $\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_d^{(i)})$ are sampled independently with uniform distribution over $[0, 1]^d$. As the *Zhou* function in (50) takes some high values, we generate the response \mathbf{y} by applying

a logarithmic transformation :

$$y^{(i)} = \log f(\mathbf{x}^{(i)}) / (d \log 10). \quad (52)$$

365 Note that there is no measurement noise here. We will address two situations
: In the first one, we assume that we know that there is no measurement noise,
we impose that there is no nugget effect in the model $\sigma_\epsilon^2 = 0$ and we consider
the Exponential anisotropic geometric correlation model ($\nu = 1/2$) as covariance
model. In the second situation, we assume that we do not know whether there is
370 measurement noise and we estimate the nugget effect of the model. We consider
consequently the Matérn 3/2 anisotropic geometric correlation model ($\nu = 3/2$),
a reasonable choice for a smooth covariance model when assuming a nugget
effect.

Table 8. Results for MLE and MSE CV methods for *Zhou* function (50) in the first setting
($\sigma_\epsilon^2 = 0$) ; confidence level $1 - \alpha = 90\%$

	MLE (section 3.4)	MSE CV (section 3.5)
Q^2	0.887	0.889
$\tilde{\mathbb{P}}_{1-\alpha}$	98.6	0
$\text{CP}_{1-\alpha}$	97.3	0.67
$\text{MPIW}_{1-\alpha}$	$8.67 \cdot 10^{-1}$	$1.27 \cdot 10^{-3}$
$\text{SdPIW}_{1-\alpha}$	$6.10 \cdot 10^{-2}$	$8.95 \cdot 10^{-5}$

Q^2 : Accuracy; $\text{CP}_{1-\alpha}$: Coverage Probability in %; $\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in %.

MPIW: Mean of Prediction Interval widths and SdPIW: standard deviation of Prediction Interval widths.

In Table 8, both models are good in terms of accuracy Q^2 with small
375 advantage for the MSE CV method but none of them satisfy the required level
of CP, especially the MSE CV model with a extremely low CP.

When proceeding similarly as Example 2 to build robust Prediction Intervals
by the RPIE model, the result is striking in Table 9 : the MSE CV solution
 $\hat{\theta}_{MSE}$ is shifted to small values (e.g. $\lambda^* = 2.41 \cdot 10^{-2}$ for the lower bound

Table 9. Performances of Prediction Intervals obtained by the RPIE method for *Zhou* function (50) in the first setting ($\sigma_\epsilon^2 = 0$); confidence level $1 - \alpha = 90\%$

	MLE $\hat{\theta}_{ML}$	MSE CV $\hat{\theta}_{MSE}$
$\tilde{\mathbb{P}}_{1-\alpha}$	90.0	90.0
CP $_{1-\alpha}$	87.3	86.7
MPIW $_{1-\alpha}$	$7.62 \cdot 10^{-1}$	$5.68 \cdot 10^{-1}$
SdPIW $_{1-\alpha}$	$5.36 \cdot 10^{-2}$	$3.98 \cdot 10^{-2}$

$\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in %; CP: coverage Probability in %; ; MPIW: Mean of Prediction Interval widths and SdPIW: standard deviation of Prediction Interval widths.

380 $\alpha/2$), meaning that the amplitude $\hat{\sigma}_{MSE}^2$ was largely underestimated and the built Prediction Intervals are now 400 times larger. Table 9 also shows that the coverage probabilities for the validation set are close to their desired value $1 - \alpha = 90\%$.

In the second setting, we consider the estimated nugget effect $\hat{\sigma}_\epsilon^2 = 1.89 \cdot 10^{-2}$ 385 using [25]. The results of MLE and MSE CV methods are shown in Table 10 : considering the nugget effect allows an increase of 4% to 5% of the accuracy Q^2 . Table 11 shows the obtained Prediction Intervals when applying the RPIE method. We reduce their width, 4 times shorter than Prediction Intervals of the MSE CV solution, and twice shorter than Prediction Intervals of the MLE 390 solution. One can notice also a decrease of 50% of the MPIW compared to the first setting, all this while maintaining an optimal coverage of $1 - \alpha = 90\%$.

Example 3 illustrates a case of misspecified model without noise where the RPIE method adjusts Prediction Intervals length and improves the robustness of Prediction Intervals so that CP is respected. One can conclude that it is 395 preferable to consider a nugget effect for shorter Prediction Intervals and for better accuracy if one is interested in making accurate predictions with standard methods.

Table 10. Results for MLE and MSE CV methods for *Zhou* function (50) in the second setting ($\hat{\sigma}_\epsilon^2 = 1.89 \cdot 10^{-2}$) ; confidence level $1 - \alpha = 90\%$

	MLE (section 3.4)	MSE CV (section 3.5)
Q^2	0.938	0.929
$\tilde{\mathbb{P}}_{1-\alpha}$	98.6	100
$\text{CP}_{1-\alpha}$	100	100
$\text{MPIW}_{1-\alpha}$	$6.64 \cdot 10^{-1}$	1.31
$\text{SdPIW}_{1-\alpha}$	$7.04 \cdot 10^{-2}$	$2.69 \cdot 10^{-1}$

Q^2 : Accuracy; $\text{CP}_{1-\alpha}$: Coverage Probability in %; $\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in %.

MPIW : Mean of Prediction Interval widths and SdPIW : standard deviation of Prediction Interval widths.

Table 11. Performances of Prediction Intervals obtained by the RPIE method for *Zhou* function (50) in the second setting ($\hat{\sigma}_\epsilon^2 = 1.89 \cdot 10^{-2}$) ; confidence level $1 - \alpha = 90\%$

	MLE $\hat{\theta}_{ML}$	MSE CV $\hat{\theta}_{MSE}$
$\tilde{\mathbb{P}}_{1-\alpha}$	90.0	90.0
$\text{CP}_{1-\alpha}$	84.7	86.7
$\text{MPIW}_{1-\alpha}$	$3.64 \cdot 10^{-1}$	$2.93 \cdot 10^{-1}$
$\text{SdPIW}_{1-\alpha}$	$6.61 \cdot 10^{-2}$	$5.21 \cdot 10^{-2}$

$\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in %; CP: coverage Probability in %; MPIW: Mean of Prediction Interval widths and SdPIW: standard deviation of Prediction Interval widths.

5.2. Application to Oil & Gas production for future wells

In this section, we focus on a production forecast use-case for unconventional
400 oil and gas wells and we show that the RPIE method can be used to build two
models estimating the quantiles of level $\alpha/2 = 10\%$ and $1 - \alpha/2 = 90\%$.

A fundamental challenge of oil and gas companies is to predict how much
oil and gas they will produce in the future. It drives both their exploration

and development strategy. However, forecasting a well future production is
405 challenging because subsurface reservoirs properties are never fully known. This
makes estimating well production with their associated uncertainty a crucial
task. The agencies **PRMS** and **SEC** ([38], [39]) define specific rules **1P/2P/3P**
for reserves estimates based on quantile estimate :

- **1P**: 90% of wells produce more than **1P** predictions (**proven**).
- 410 • **2P**: 50% of wells produce more than **2P** predictions (**probable**).
- **3P**: 10% of wells produce more than **3P** predictions (**possible**).

These rules are to be disclosed to security investors for publicly traded oil
and gas companies and aim to provide investors with consistent information and
associated value assessments. Many Machine Learning algorithms have shown
415 their efficiency in estimating the median **2P** but failed to estimate **1P** and **3P**.
Thus, the objective of this study is to build a proper estimation of the quantiles
 $p_{0.90}$ and $p_{0.10}$ by applying the RPIE method described in section 4.

Our dataset, *field data*, is derived from unconventional wells localized in
the *Utica* shale oil reservoir, located in the north-east of the United States.
420 It contains approximately $n = 1850$ wells and $d = 12$ variables, including
localization, Cumulative Production over 12 months, completion design and
exploitation conditions. The raw dataset can be found at the Ohio Oil & Gas
well locator of the Ohio Department of Natural Resources [40].

We standardized the data (\mathbf{X}, \mathbf{y}) and we divided into a 60% – 20% – 20%
425 partition of three datasets : a training set and two validation sets. The response
 \mathbf{y} (Cumulative Production over 12 months) is noisy due to the uncertainty of the
reservoir parameters in the field. The nugget effect σ_ϵ^2 is unknown but estimated
to $\hat{\sigma}_\epsilon^2 = 0.16$ using the method of [25].

Table 12 shows the performances of the GP models trained by MLE and
430 MSE CV methods compared with two other statistical models : Random Forest
and Gradient Boosting whose Prediction Intervals are built using the Bootstrap
method. Here we consider the Prediction Intervals of level $1 - \alpha = 80\%$: the

lower bound is the 10% quantile ($p_{0.10}$) and the upper bound the 90% quantile ($p_{0.90}$) of the predictive distribution.

Table 12. Accuracy, CP and Computing time results obtained for GP model, Random Forest and Gradient Boosting ; confidence level $1 - \alpha = 80\%$

	MLE	MSE CV	Random Forest	Gradient Boosting
Q^2	0.88	0.86	0.87	0.88
$CP_{1-\alpha}$	92.8	93.2	98.1	49.8
Ct	207.5	507.2	2.33	96.3
$MPIW_{1-\alpha}$	1.18	1.20	1.52	0.48
$SdPIW_{1-\alpha}$	0.21	0.16	0.29	0.22

Q^2 : Accuracy; CP: Coverage Probability in % and Ct: Computing time (s).

MPIW: Mean of Prediction Interval widths and SdPIW: standard deviation of Prediction Interval widths.

435 The accuracy of the MLE model is 88% against an accuracy of 86% for
MSE CV model, both models have approximately the same accuracy as other
models like Random Forest or Gradient Boosting. Furthermore, the CP of the
Prediction Intervals of $1 - \alpha = 80\%$ is not satisfactory but it is quite *reasonable*
for MLE model compared to Random Forest (overestimated Prediction Intervals)
440 or Gradient Boosting (underestimated Prediction Intervals). Finally, it appears
that the GP model requires some computing resources to be built and to estimate
its hyperparameters by MLE or MSE CV methods.

The MLE's solution is defined as reference $\theta_0 = \hat{\theta}_{ML}$ in the optimization
problem (39) for the quantiles $\alpha/2 = 10\%$ and $1 - \alpha/2 = 90\%$ and we build
445 robust Prediction Intervals confidence level $1 - \alpha = 80\%$ with the RPIE method
as described in section 4.

In Table 13, the Relaxation parameter λ^* is larger than 1 and is up to 6.28
for \mathbf{P}_{10} model, this means that the initial vector length $\hat{\theta}_{ML}$ is underestimated
especially for the lower bound of 10%. A reason could be related to the fact

Table 13. Performances of Prediction Intervals estimated by \mathbf{P}_{90} and \mathbf{P}_{10} models and the RPIE method on production data ; confidence level $1 - \alpha = 80\%$.

	\mathbf{P}_{90} model	\mathbf{P}_{10} model	RPIE
λ^*	4.50	6.88	-
$\tilde{\mathbb{P}}_{1-\alpha}$	82.2	76.9	79.9
$\text{CP}_{1-\alpha}^{\text{Val},1}$	81.2	78.2	81.0
$\text{CP}_{1-\alpha}^{\text{Val},2}$	85.6	80.3	83.2

λ^* : Relaxation parameter; $\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in % and CP: Coverage Probability in %.

Table 14. Comparison of the mean and standard-deviation Prediction Interval width of MLE and RPIE methods ; confidence level $1 - \alpha = 80\%$.

	MLE method	RPIE method
$\text{MPIW}_{1-\alpha}^{\text{Val},1}$	1.18	1.06
$\text{SdPIW}_{1-\alpha}^{\text{Val},1}$	0.21	0.01
$\text{MPIW}_{1-\alpha}^{\text{Val},2}$	1.17	1.06
$\text{SdPIW}_{1-\alpha}^{\text{Val},2}$	0.17	0.01

MPIW : Mean Prediction Intervals width and SdPIW : standard deviation Prediction Intervals width.

450 the MLE model predicts accurately "*poor*" wells and overfits to predict "*good*" wells which makes the lower bound overestimated as it can be seen in figure 3. Another reason may be the non stationarity of some well characteristics that strongly depend on the field (e.g. Localization), it may be recommended to take more complex and tensorized covariance functions for these variables.

455 When considering Prediction Intervals, the CP is optimal among the training set and is close to $1 - \alpha = 80\%$ for both validation sets. Therefore, we fulfill the objective of estimating the upper bound and lower bound, the obtained quantiles

$p_{0.90}$ and $p_{0.10}$ respect **1P** and **3P** rules as mentioned above. Finally, one can also notice that Prediction Intervals in (3b) and (3d) are not homogeneous and some of them are longer, the RPIE method makes them less wider and more homogeneous as it can be seen in the evolution of the standard deviation width SdPIW between the two models. Note that in sub-figures (3b) and (3d), MeanPred correspond to the centers of Prediction Intervals obtained by RPIE method unlike sub-figures (3a) and (3c) where they represent the mean predictions of the MLE model.

6. Conclusion

In this paper, we have introduced a new approach for Prediction Intervals estimation based on the Cross-Validation method. We used the Gaussian Processes model because the predictive distribution at a new point is completely characterized by Gaussian law. We addressed an optimization problem for model's hyperparameters estimation by considering the notion of Coverage Probability. The optimal hyperparameters were identified by minimizing the Wasserstein distance between the Gaussian distribution with the hyperparameters determined by Cross Validation , and the Gaussian distribution with hyperparameters achieving the desired Coverage Probability. This method is relevant when the model is misspecified. It insures an optimal Leave-One-Out coverage probability for the training set. It also achieves a reasonable coverage probability for the validation set when it is available. It can be also extended to other statistical models with a predictive distribution but more detailed work is needed to consider the influence of hyperparameters on Prediction Interval's coverage and solve the optimization problem more efficiently in these cases.

Acknowledgements

The author⁽¹⁾ would like to thank Achraf Ourir, Zinyat Agharzayeva (TotalEnergies SE - La Défense, France), Daniel Busby (TotalEnergies SE, CSTJF -

485 Pau, France) and the SINCLAIR Lab (IA Commun Lab - Saclay, France) for
useful discussions and helpful suggestions.

References

References

- [1] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, Chapman
490 & Hall/CRC Monographs on Statistics & Applied Probability, Taylor &
Francis, 1994.
URL <https://books.google.fr/books?id=gLlpIUxRntoC>
- [2] T. Heskes, Practical confidence and prediction intervals, in: Advances in
Neural Information Processing Systems 9, MIT press, 1997, pp. 176–182.
- 495 [3] K. Li, R. Wang, H. Lei, T. Zhang, Y. Liu, X. Zheng, Interval prediction of
solar power using an improved bootstrap method, Solar Energy 159 (2018)
97 – 112. doi:<https://doi.org/10.1016/j.solener.2017.10.051>.
URL <http://www.sciencedirect.com/science/article/pii/S0038092X17309313>
- 500 [4] S. Wager, T. Hastie, B. Efron, Confidence intervals for random forests:
the jackknife and the infinitesimal jackknife, Journal of machine learning
research : JMLR 15 1 (2014) 1625–1651.
- [5] B. Efron, Jackknife-after-bootstrap standard errors and influence functions,
Journal of the Royal Statistical Society. Series B (Methodological) 54 (1)
505 (1992) 83–127.
URL <http://www.jstor.org/stable/2345949>
- [6] N. Meinshausen, Quantile regression forests, J. Mach. Learn. Res. 7 (2006)
983–999.
- [7] H. Zhang, J. Zimmerman, D. Nettleton, D. Nordman, Random forest
510 prediction intervals, The American Statistician 74 (2020) 392 – 406.

- [8] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, *Journal of the American Statistical Association* 113 (523) (2018) 1094–1111. doi:10.1080/01621459.2017.1307116.
- 515 [9] J. T. G. Hwang, A. A. Ding, Prediction intervals for artificial neural networks, *Journal of the American Statistical Association* 92 (438) (1997) 748–757.
- [10] D. A. Nix, A. S. Weigend, Estimating the mean and variance of the target probability distribution, in: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, Vol. 1, 1994, pp. 55–60 vol.1.
- 520 [11] D. J. C. MacKay, A practical bayesian framework for backpropagation networks, *Neural Comput.* 4 (3) (1992) 448–472. doi:10.1162/neco.1992.4.3.448.
- [12] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: M. F. Balcan, K. Q. Weinberger (Eds.), *Proceedings of The 33rd International Conference on Machine Learning*, Vol. 48 of *Proceedings of Machine Learning Research*, PMLR, New York, New York, USA, 2016, pp. 1050–1059.
- 525 [13] A. Khosravi, S. Nahavandi, D. Creighton, A. F. Atiya, Lower upper bound estimation method for construction of neural network-based prediction intervals, *IEEE Transactions on Neural Networks* 22 (3) (2011) 337–346.
- 530 [14] T. Pearce, A. Brintrup, M. Zaki, A. Neely, High-quality prediction intervals for deep learning: A distribution-free, ensembled approach, in: J. Dy, A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, PMLR, Stockholmsmässan, Stockholm Sweden, 2018, pp. 4075–4084.
- 535 [15] J. Landon, N. Singpurwalla, Choosing a coverage probability for prediction intervals, *The American Statistician* 62 (2008) 120–124. doi:10.1198/000313008X304062.

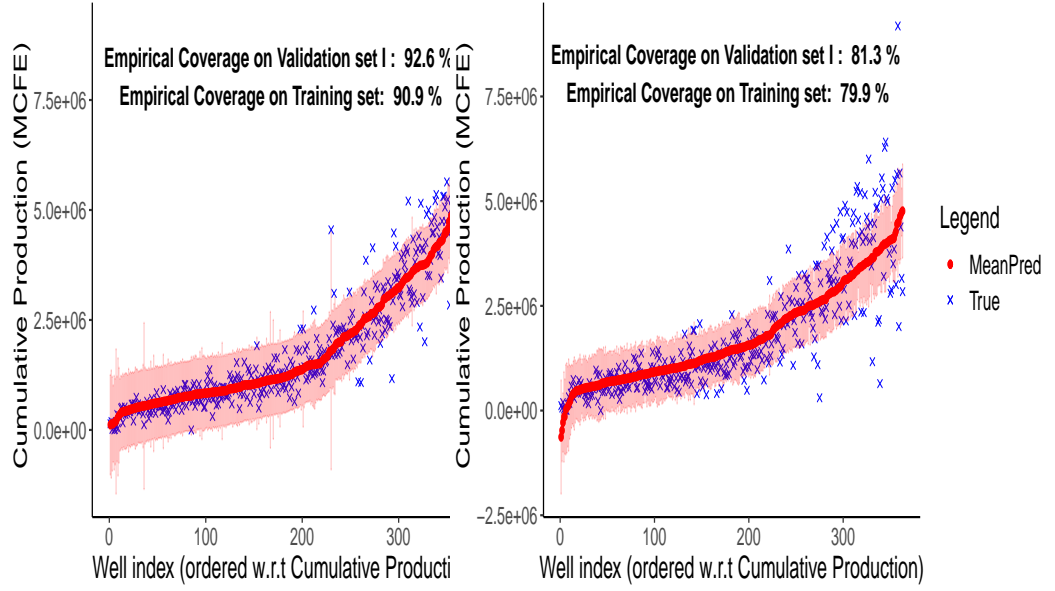
- [16] A. Khosravi, S. Nahavandi, D. Creighton, A prediction interval-based
 540 approach to determine optimal structures of neural network metamodels,
 Expert Syst. Appl. 37 (2010) 2377–2387. doi:10.1016/j.eswa.2009.07.
 059.
- [17] J. Pang, D. Liu, Y. Peng, X. Peng, Optimize the coverage probability of
 prediction interval for anomaly detection of sensor-based monitoring series,
 545 Sensors (Basel, Switzerland) 18. doi:10.3390/s18040967.
- [18] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learn-
 ing (Adaptive Computation and Machine Learning), The MIT Press, 2005.
- [19] F. Bachoc, Estimation paramétrique de la fonction de covariance dans le
 modèle de krigeage par processus gaussiens : application à la quantification
 550 des incertitudes en simulation numérique, Ph.D. thesis, University Paris 7
 (2013).
 URL <http://www.theses.fr/2013PA077111>
- [20] J. F. Lawless, M. Fredette, Frequentist prediction intervals and predictive
 distributions, Biometrika 92 (3) (2005) 529–542.
- [21] J. Muré, Objective bayesian analysis of kriging models with anisotropic
 555 correlation kernel, Ph.D. thesis, Sorbonne Paris Cité (2018).
 URL <http://www.theses.fr/2018USPCC069>
- [22] M. E. Tipping, Bayesian Inference: An Introduction to Principles and
 Practice in Machine Learning, Springer Berlin Heidelberg, Berlin, Heidelberg,
 560 2004, pp. 41–62.
- [23] J. Oakley, A. Hagan, A. O’Hagan, Probabilistic sensitivity analysis of
 complex models: A bayesian approach, Journal of the Royal Statistical
 Society: Series B (Statistical Methodology) 66 (2004) 751 – 769.
- [24] C. Currin, T. J. Mitchell, M. D. Morris, D. Ylvisaker, Bayesian prediction
 565 of deterministic functions, with applications to the design and analysis

of computer experiments, *Journal of the American Statistical Association* 86 (416) (1991) 953–963.

- [25] B. Iooss, A. Marrel, An efficient methodology for the analysis and modeling of computer experiments with large number of inputs, in: UNCECOMP 2017 2nd ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Sciences and Engineering, Rhodes Island, Greece, 2017, pp. 187–197.
- [26] F. Bachoc, Cross validation and maximum likelihood estimations of hyperparameters of gaussian processes with model misspecification, *Computational Statistics & Data Analysis* 66 (2013) 55–69.
- [27] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer Series in Statistics, Springer New York, 2009.
- [28] D. Wallach, B. Goffinet, Mean squared error of prediction as a criterion for evaluating and comparing system models, *Ecological Modelling* 44 (3) (1989) 299 – 306.
- [29] O. Dubrule, Cross validation of kriging in a unique neighborhood, *Journal of the International Association for Mathematical Geology* 15 (6) (1983) 687–699.
- [30] Y. Hong, W. Q. Meeker, J. D. McCalley, Prediction of remaining life of power transformers based on left truncated and right censored lifetime data, *Ann. Appl. Stat.* 3 (2) (2009) 857–879. doi:10.1214/00-AOAS231.
- [31] C. Villani, *The Wasserstein distances*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 93–111.
- [32] J. Zhao, The lower semicontinuity of optimal solution sets, *Journal of Mathematical Analysis and Applications* 207 (1) (1997) 240 – 254.

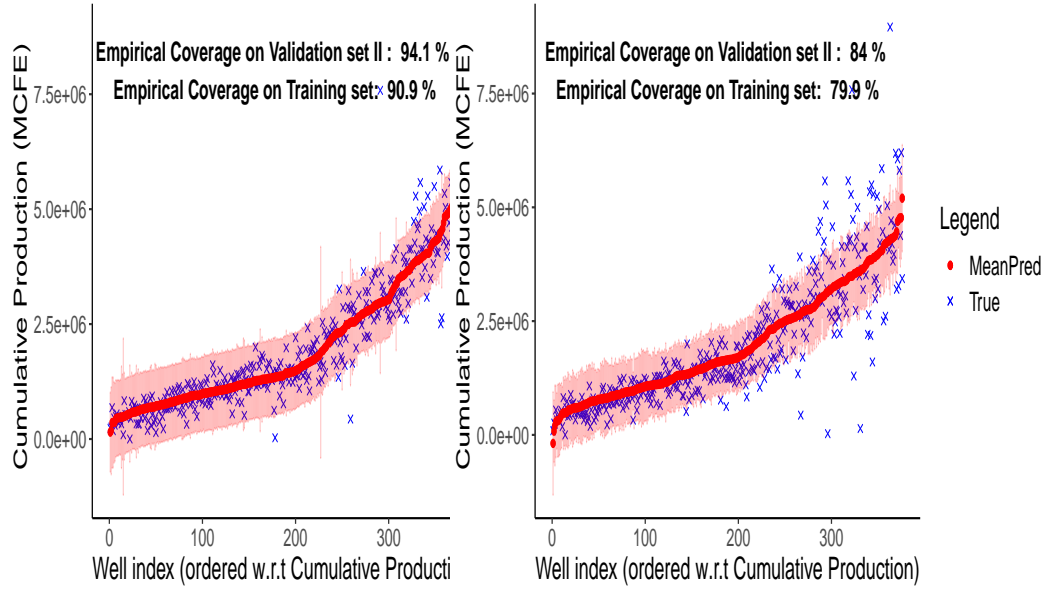
- [33] J. P. C. Kleijnen, R. G. Sargent, A methodology for fitting and validating metamodels in simulation, *European Journal of Operational Research* 120 (2000) 14–29.
- 595 [34] A. I. J. Forrester, A. Sóbester, A. J. Keane, *Engineering Design Via Surrogate Modelling: A Practical Guide*, Progress in Astronautics and Aeronautics, American Institute of Aeronautics and Astronautics, 2008.
- [35] H. Moon, Design and analysis of computer experiments for screening input variables, Ph.D. thesis, The Ohio State University (01 2010).
- 600 [36] W. J. Morokoff, R. E. Caflisch, Quasi-monte carlo integration, *Journal of computational physics* 122 (1995) 218–230.
- [37] Y. Zhou, Adaptive importance sampling for integration, Ph.D. thesis, Stanford University (1998).
- [38] P. Reserves, R. Definitions, Petroleum reserves and resources definitions,
605 <https://www.spe.org/en/industry/reserves/>.
- [39] Securities, E. Commission, Modernization of oil and gas reporting, revisions and additions to the definition section in rule 4-10 of regulation s-x, <https://www.sec.gov/rules/final/2008/33-8995.pdf> (Jan. 2010).
- [40] T. O. D. of Natural Resources. Division of Oil & Gas Resources, The Ohio
610 Oil & Gas well locator, <https://ohiodnr.gov/wps/portal/gov/odnr/discover-and-learn/safety-conservation/about-odnr/oil-gas/oil-gas-resources/well-locator>.
- [41] V. De Oliveira, Objective bayesian analysis of spatial data with measurement error, *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 35 (2) (2007) 283–301.
615
- [42] C. Berge, *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces and Convexity*, Oliver & Boyd, 1963.

- [43] V. De Oliveira, J. O. Berger, B. Sansó, Objective Bayesian analysis of spatially correlated data, *Journal of the American Statistical Association* 96 (456) (2001) 1361–1374.
- [44] M. Rosenblatt, Review: A. M. Yaglom, correlation theory of stationary and random functions vol. i; basic results, vol. ii, supplementary notes and references, *Bulletin (New Series) of the American Mathematical Society* 20 (2) (1989) 207–211.
- URL <https://projecteuclid.org:443/euclid.bams/1183555023>
- [45] C. Ren, D. Sun, C. Z. He, Objective bayesian analysis for a spatial model with nugget effects, *Journal of Statistical Planning and Inference* 142 (2012) 1933–1946.
- [46] J. Muré, Propriety of the reference posterior distribution in gaussian process modeling, available at arXiv: 1805.08992 (2020 in press, *Annals of Stats*).
arXiv:1805.08992.



(a) Fitted vs Predicted values for MLE model on Validation set I

(b) Fitted vs Predicted values when applying RPIE method on Validation set I



(c) Fitted vs Predicted values for MLE model on Validation set II

(d) Fitted vs Predicted values when applying RPIE method on Validation set II

Figure 3. Production use-case after re-scaling : True values in blue vs predicted values in red with the associated $(1 - \alpha) \times 100\%$ Prediction Intervals

APPENDIX A. PROOFS OF PROPOSITIONS 1 - 3

PRELIMINARY LEMMAS

Lemma 1. *Let \mathbf{F} be a full rank matrix (hypothesis \mathcal{H}_1), let \mathbf{K} be a positive definite matrix and let $\bar{\mathbf{K}}$ defined by $\bar{\mathbf{K}} = \mathbf{K}^{-1} \left(\mathbf{I}_n - \mathbf{F} (\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{K}^{-1} \right)$ then $\text{Ker } \bar{\mathbf{K}} = \text{Im } \mathbf{F}$ and $\bar{\mathbf{K}}$ is singular.*

Proof. Let $\bar{\mathbf{K}}$ be the matrix defined above. Suppose that $\mathbf{x} \in \text{Im } \mathbf{F}$, then there exists \mathbf{y} such that $\mathbf{x} = \mathbf{F}\mathbf{y}$, and $\bar{\mathbf{K}}\mathbf{x} = \mathbf{K}^{-1} \left(\mathbf{F}\mathbf{y} - \mathbf{F} (\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\mathbf{y} \right) = \mathbf{K}^{-1} (\mathbf{F}\mathbf{y} - \mathbf{F}\mathbf{y}) = \mathbf{0}$. Thus $\mathbf{x} \in \text{Ker } \bar{\mathbf{K}}$.

If $\mathbf{x} \in \text{Ker } \bar{\mathbf{K}}$, then $\mathbf{K} \bar{\mathbf{K}}\mathbf{x} = \mathbf{0}$, and $\mathbf{x} = \mathbf{F} (\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{x} = \mathbf{F}\mathbf{x}' \in \text{Im } \mathbf{F}$.

In case of Ordinary or Universal kriging, $p = \text{rank}(\mathbf{F}) = \dim(\text{Ker } \bar{\mathbf{K}}) \geq 1$ which means that $\bar{\mathbf{K}}$ is not invertible. \square

Lemma 2 (De Oliveira [41], 2007). *Under the hypotheses of Lemma 1 and given the full rank regression matrix \mathbf{F} , there exists a matrix $\mathbf{W} \in \mathbb{R}^{n \times (n-p)}$ satisfying :*

$$\mathbf{W}^\top \mathbf{W} = \mathbf{I}_{n-p} \text{ and } \mathbf{F}^\top \mathbf{W} = \mathbf{O}_{p \times (n-p)}, \quad (.1)$$

$$\bar{\mathbf{K}} = \mathbf{W} (\mathbf{W}^\top \mathbf{K} \mathbf{W})^{-1} \mathbf{W}^\top. \quad (.2)$$

Lemma 3. *Under the hypotheses of Lemma 1, if additionally hypothesis \mathcal{H}_2 holds true, then $\bar{\mathbf{K}}_{ii} > 0$ for all $i \in \{1, \dots, n\}$.*

Proof. $\bar{\mathbf{K}}$ is a positive semi-definite matrix by Lemma 2 and we can write

$$\bar{\mathbf{K}} = \sum_{j=1}^n \lambda_j \mathbf{u}_j \mathbf{u}_j^\top, \quad (.3)$$

with $\lambda_j \geq 0$ the eigenvalues of $\bar{\mathbf{K}}$ and $(\mathbf{u}_j)_{j=1}^n$ the orthonormal basis of the eigenvectors. We have

$$\bar{\mathbf{K}}_{ii} = \mathbf{e}_i^\top \bar{\mathbf{K}} \mathbf{e}_i = \sum_{j=1}^n \lambda_j (\mathbf{u}_j^\top \mathbf{e}_i)^2. \quad (.4)$$

If $\bar{\mathbf{K}}_{ii} = 0$, then $\mathbf{u}_j^\top \mathbf{e}_i = 0$ for all j such that $\lambda_j > 0$. Therefore

$$\bar{\mathbf{K}}\mathbf{e}_i = \sum_{j=1}^n \lambda_j (\mathbf{u}_j^\top \mathbf{e}_i) \mathbf{u}_j = \mathbf{0}, \quad (.5)$$

which shows that $\mathbf{e}_i \in \text{Ker } \bar{\mathbf{K}}$, that is, $\mathbf{e}_i \in \text{Im } \mathbf{F}$ by Lemma 1. \square

Lemma 4. Let $\mathbf{\Pi} = \mathbf{W}\mathbf{W}^\top = \mathbf{I}_n - \mathbf{F}(\mathbf{F}^\top \mathbf{F})^{-1} \mathbf{F}^\top$ the orthogonal projection
 650 matrix on $(\text{Im } \mathbf{F})^\perp$ then, with the hypothesis \mathcal{H}_2 , $(\mathbf{\Pi})_{i,i} \neq 0$ for all $i \in \{1, \dots, n\}$.

Proof. This lemma is a direct application of Lemma 3 by choosing $\mathbf{K} = \mathbf{I}_n$. \square

PROOF OF PROPOSITION 1

From the preliminary lemmas, we show now the stronger result (stronger than Proposition 1):

655 **Lemma 5.** Under the hypotheses $\mathcal{H}_1 - \mathcal{H}_3$, for any $\boldsymbol{\theta} \in (0, +\infty)^d$, there exists $\sigma^2 \in (0, +\infty)$ such that $(\sigma^2, \boldsymbol{\theta}) \in \mathcal{A}_{a,\delta}$.

Proof. Here $\sigma_\epsilon^2 > 0$. Let us assume that $a > 1/2$ (i.e. $q_a > 0$), then for $\boldsymbol{\theta}$ fixed in $(0, +\infty)^d$, the limit of $\bar{\mathbf{K}}$ when $\sigma^2 \rightarrow 0$ is well defined and is equal to

$$\lim_{\sigma^2 \rightarrow 0} \bar{\mathbf{K}} = \sigma_\epsilon^{-2} \mathbf{W}\mathbf{W}^\top = \sigma_\epsilon^{-2} \mathbf{\Pi}. \quad (.6)$$

By the hypothesis \mathcal{H}_2 and from Lemma 4 we can write for all $i \in \{1, \dots, n\}$

$$\frac{(\bar{\mathbf{K}}\mathbf{y})_i}{\sqrt{(\bar{\mathbf{K}})_{i,i}}} \xrightarrow{\sigma^2 \rightarrow 0} \frac{1}{\sigma_\epsilon} \frac{(\mathbf{\Pi}\mathbf{y})_i}{\sqrt{(\mathbf{\Pi})_{i,i}}}. \quad (.7)$$

Since $h_\delta^+ \leq h$ for all $\delta > 0$, then

$$\lim_{\sigma^2 \rightarrow 0} \psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) \leq \lim_{\sigma^2 \rightarrow 0} \psi_a(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n h \left(q_a - \frac{1}{\sigma_\epsilon} \frac{(\mathbf{\Pi}\mathbf{y})_i}{\sqrt{(\mathbf{\Pi})_{i,i}}} \right) = \frac{k_\epsilon}{n} \quad (.8)$$

When $\sigma^2 \rightarrow +\infty$, we have

$$\bar{\mathbf{K}} \stackrel{\sigma^2 \rightarrow +\infty}{\sim} \sigma^{-2} \bar{\mathbf{R}}_\theta, \quad (.9)$$

where

$$\bar{\mathbf{R}}_\theta = \mathbf{W}(\mathbf{W}^\top \mathbf{R}_\theta \mathbf{W})^{-1} \mathbf{W}^\top. \quad (.10)$$

By lemma 3, we have $(\overline{\mathbf{R}}_{\boldsymbol{\theta}})_{i,i} > 0$ for all $i \in \{1, \dots, n\}$ and we obtain that

$$\frac{1}{\sigma} \frac{(\overline{\mathbf{R}}_{\boldsymbol{\theta}} \mathbf{y})_i}{\sqrt{(\overline{\mathbf{R}}_{\boldsymbol{\theta}})_{i,i}}} \xrightarrow{\sigma^2 \rightarrow +\infty} 0. \quad (.11)$$

With δ small enough satisfying $\delta < q_a$, we obtain

$$\psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) \xrightarrow{\sigma^2 \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^n h_{\delta}^+(q_a) = 1. \quad (.12)$$

Since $k_{\epsilon} < an < n$ by hypothesis \mathcal{H}_3 and since $\psi_a^{(\delta)}$ is continuous, the Intermediate Value Theorem gives the existence of $\sigma_{\delta}^2 \in (0, +\infty)$ such that

$$\psi_a^{(\delta)}(\sigma_{\delta}^2, \boldsymbol{\theta}) = a, \quad (.13)$$

which gives the desired result.

Similarly, if $a < a/2$ then $q_a < 0$ and

$$\lim_{\sigma^2 \rightarrow 0} \psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) \geq \lim_{\sigma^2 \rightarrow 0} \psi_a(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n h \left(q_a - \frac{1}{\sigma_{\epsilon}} \frac{(\boldsymbol{\Pi} \mathbf{y})_i}{\sqrt{(\boldsymbol{\Pi})_{i,i}}} \right) = \frac{k_{\epsilon}}{n} > a. \quad (.14)$$

When $\delta < q_{1-a}$, one obtains

$$\psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) \xrightarrow{\sigma^2 \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^n h_{\delta}^-(q_a) = 0. \quad (.15)$$

By the hypothesis \mathcal{H}_3 , one has the existence of $\sigma_{\delta}^2 \in (0, +\infty)$ such that

$$\psi_a^{(\delta)}(\sigma_{\delta}^2, \boldsymbol{\theta}) = a, \quad (.16)$$

which completes the proof of the lemma. \square

PROOF OF PROPOSITION 2

660 The existence of $\sigma_{\text{opt}}^2(\lambda)$ for all $\lambda \in (0, +\infty)$ results directly from the following lemma 6 :

Lemma 6. *For all $\lambda \in (0, +\infty)$, $H_{\delta}(\lambda)$ is a non-empty and compact subset of \mathbb{R}^+ i.e. H_{δ} is compact-valued.*

Proof. By Lemma 5, $H_\delta(\lambda)$ is non-empty for all $\lambda \in (0, +\infty)$.

665 $H_\delta(\lambda)$ is closed since the functions h_δ^+, h_δ^- are continuous and the map $(\sigma^2, \boldsymbol{\theta}) \mapsto \bar{\mathbf{K}}$ is also continuous for all $(\sigma^2, \boldsymbol{\theta})$ by the continuity of the kernel function $\mathbf{k}_{\cdot, \cdot}^\nu(\mathbf{x}, \mathbf{x}')$ for any $\nu > 0$ and $\mathbf{x}, \mathbf{x}' \in \mathcal{D}$.

We now prove that $H_\delta(\lambda)$ is bounded. Let us assume that $a \in (1/2, 1)$. If $H_\delta(\lambda)$ is not bounded then there exists a sequence $(\sigma_m^2)_{m \in \mathbb{N}}$ of $H_\delta(\lambda)$ such that $\lim_{m \rightarrow +\infty} \sigma_m^2 = +\infty$ and, by continuity of $\psi_a^{(\delta)}$

$$a = \lim_{m \rightarrow +\infty} \psi_a^{(\delta)}(\sigma_m^2, \lambda \boldsymbol{\theta}_0) = \frac{1}{n} \sum_{i=1}^n h_\delta^+(q_a) = 1, \quad (.17)$$

which is a contradiction. Therefore, $H_\delta(\lambda)$ is closed and bounded, $H_\delta(\lambda)$ is compact. \square

670 $\sigma_{\text{opt}}^2(\lambda)$ can be seen the solution of a constrained maximization problem

$$\sigma_{\text{opt}}^2(\lambda) = - \max_{\sigma^2 \in H_\delta(\lambda)} u(\sigma^2, \lambda), \quad \lambda \in (0, +\infty), \quad (.18)$$

where $u(\sigma^2, \lambda) = -\sigma^2$ is a continuous function. H_δ is non-empty-valued and compact-valued by Lemma 6, upper semi-continuous since $\psi_a^{(\delta)}$ is continuous on $[0, +\infty) \times (0, +\infty)^d$, and continuous if the hypothesis \mathcal{H}_4 is satisfied, the Maximum theorem (C. Berge [42], 1963, p. 116) provides the continuity of σ_{opt}^2 on $(0, +\infty)$.
675

PROOF OF PROPOSITION 3

Let $\boldsymbol{\theta}_0$ be a solution of one of the problems described in (13) or (17). The continuity of \mathcal{L} on $(0, +\infty)$ follows from the continuity of the trace function $\text{Tr}(\cdot)$, the continuity of the map $(\sigma^2, \boldsymbol{\theta}) \mapsto \bar{\mathbf{K}}$ and the continuity of σ_{opt}^2 by proposition

680 2.

Assume that $\lim_{\lambda \rightarrow +\infty} \sigma_{\text{opt}}^2(\lambda) \neq +\infty$, then there exists $M > 0$ such that for all $\lambda > 0$ there exists $\lambda' \geq \lambda$ and $\sigma_{\text{opt}}^2(\lambda') \leq M$. Hence, we can recursively build a sequence $(\lambda_m)_{m \in \mathbb{N}}$ of integers such that $\lambda_{m+1} \geq \lambda_m + 1$ and $\sigma_{\text{opt}}^2(\lambda_m) \leq M$ for all $m \in \mathbb{N}$.

By the Bolzano-Weierstrass theorem, we extract a convergent sub-sequence $(\lambda_{\phi(m)})_{m \in \mathbb{N}}$ where $\phi : \mathbb{N} \rightarrow \mathbb{N}$ such that $\sigma_{\text{opt}}^2(\lambda_{\phi(m)}) \xrightarrow{m \rightarrow +\infty} \sigma_{\infty}^2 < +\infty$ and

$$\mathbf{K}_{\sigma_{\text{opt}}^2(\lambda_{\phi(m)}), \lambda_{\phi(m)} \boldsymbol{\theta}_0} \xrightarrow{m \rightarrow +\infty} \sigma_{\infty}^2 \mathbf{J} + \sigma_{\epsilon}^2 \mathbf{I}_n = \mathbf{K}_{\infty}. \quad (.19)$$

When there is a nugget effect $\sigma_{\epsilon}^2 > 0$, the limit of $\bar{\mathbf{K}}_m := \bar{\mathbf{K}}_{\sigma_{\text{opt}}^2(\lambda_{\phi(m)}), \lambda_{\phi(m)} \boldsymbol{\theta}_0}$ when $m \rightarrow +\infty$ exists because the matrix \mathbf{K}_{∞} is nonsingular by the auxiliary fact 1 of Berger *et al.* ([43], 2001)

$$\det \mathbf{K}_{\infty} = \left(\frac{\sigma_{\epsilon}^2}{\sigma_{\infty}^2} \right)^n \left(1 + \frac{\sigma_{\epsilon}^2}{\sigma_{\infty}^2} \mathbf{e}^{\top} \mathbf{I}_n \mathbf{e} \right) = \left(\frac{\sigma_{\epsilon}^2}{\sigma_{\infty}^2} \right)^n \left(1 + n \frac{\sigma_{\epsilon}^2}{\sigma_{\infty}^2} \right) > 0. \quad (.20)$$

From hypothesis \mathcal{H}_1 , \mathbf{e} is a column of \mathbf{F} and we can prove that

$$\begin{aligned} \bar{\mathbf{K}}_m \xrightarrow{m \rightarrow +\infty} \bar{\mathbf{K}}_{\infty} &:= \mathbf{W} (\mathbf{W}^{\top} (\sigma_{\infty}^2 \mathbf{J} + \sigma_{\epsilon}^2 \mathbf{I}_n) \mathbf{W})^{-1} \mathbf{W}^{\top} \\ &= \sigma_{\epsilon}^{-2} \mathbf{W} (\mathbf{W}^{\top} \mathbf{W})^{-1} \mathbf{W}^{\top} = \sigma_{\epsilon}^{-2} \boldsymbol{\Pi}. \end{aligned} \quad (.21)$$

By hypothesis \mathcal{H}_2 , the Leave-One-Out formulas (16) give for all $i \in \{1, \dots, n\}$

$$\frac{(\bar{\mathbf{K}}_m \mathbf{y})_i}{\sqrt{(\bar{\mathbf{K}}_m)_{i,i}}} \xrightarrow{m \rightarrow +\infty} \frac{1}{\sigma_{\epsilon}} \frac{(\boldsymbol{\Pi} \mathbf{y})_i}{\sqrt{(\boldsymbol{\Pi})_{i,i}}}. \quad (.22)$$

If $a > 1/2$ for example and by definition of $\sigma_{\text{opt}}^2(\lambda_{\phi(m)})$, one obtains

$$\begin{aligned} a &= \frac{1}{n} \sum_{i=1}^n h_{\delta}^+ \left(q_a - \frac{(\bar{\mathbf{K}}_m \mathbf{y})_i}{\sqrt{(\bar{\mathbf{K}}_m)_{i,i}}} \right) \\ &\xrightarrow{m \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^n h_{\delta}^+ \left(q_a - \frac{(\bar{\mathbf{K}}_{\infty} \mathbf{y})_i}{\sqrt{(\bar{\mathbf{K}}_{\infty})_{i,i}}} \right) \\ &= \frac{1}{n} \sum_{i=1}^n h_{\delta}^+ \left(q_a - \frac{1}{\sigma_{\epsilon}} \frac{(\boldsymbol{\Pi} \mathbf{y})_i}{\sqrt{(\boldsymbol{\Pi})_{i,i}}} \right) = \frac{k_{\epsilon}}{n} < a, \end{aligned} \quad (.23)$$

which is contradictory. Therefore, $\lim_{\lambda \rightarrow +\infty} \sigma_{\text{opt}}^2(\lambda) = +\infty$ and \mathcal{L} is coercive. The case $a < 1/2$ can be addressed in the same way.

APPENDIX B. THE NO-NUGGET CASE

690 *PROOF OF THE EXISTENCE OF A SOLUTION TO THE PROBLEM (30)*

In the absence of $\sigma_\epsilon^2 = 0$, it follows from the Leave-One-Out formulas that, for all $i \in \{1, \dots, n\}$

$$\frac{(\overline{\mathbf{K}}\mathbf{y})_i}{\sqrt{(\overline{\mathbf{K}})_{i,i}}} = \frac{1}{\sigma} \frac{(\overline{\mathbf{R}}_\theta \mathbf{y})_i}{\sqrt{(\overline{\mathbf{R}}_\theta)_{i,i}}}, \quad (.24)$$

which is a monotonic function in σ^2 when θ is fixed in $(0, \infty)^d$.

Let θ be fixed in $(0, +\infty)^d$ and let $a > 1/2$. The proportion $\psi_a^{(\delta)}(\sigma^2, \theta)$ has the limit

$$\lim_{\sigma^2 \rightarrow +\infty} \psi_a^{(\delta)}(\sigma^2, \theta) = \frac{1}{n} \sum_{i=1}^n h_\delta^+(q_a) = 1, \quad (.25)$$

and, if $\sigma^2 \rightarrow 0$, it has the limit

$$\lim_{\sigma^2 \rightarrow 0} \psi_a^{(\delta)}(\sigma^2, \theta) = \frac{1}{n} \text{Card} \{i \in \{1, \dots, n\}, (\overline{\mathbf{R}}_\theta \mathbf{y})_i \leq 0\} = \frac{k_\theta}{n}. \quad (.26)$$

Let θ denote the norm of θ (i.e. $\theta = \|\theta\|$) and consider the set $\mathcal{J} = \{i \in \{1, \dots, n\}, (\Pi \mathbf{y})_i \leq 0\}$. For $i \in \mathcal{J}^c$, one has $(\Pi \mathbf{y})_i > 0$, and, since $\overline{\mathbf{R}}_\theta$ converges to Π when $\theta \rightarrow 0$

$$\forall i \in \mathcal{J}^c : (\overline{\mathbf{R}}_\theta \mathbf{y})_i \xrightarrow{\theta \rightarrow 0} (\Pi \mathbf{y})_i > 0 \quad (.27)$$

It results that, there exists $\theta_c > 0$ such that if $\theta \in \mathcal{B}_r(\mathbf{0}, \theta_c)$ (the open ball of radius θ_c centered at $\mathbf{0}$) then $(\overline{\mathbf{R}}_\theta \mathbf{y})_i > 0$ for any $i \in \mathcal{J}^c$. Consequently, one gets for any $\theta \in \mathcal{B}_r(\mathbf{0}, \theta_c)$

$$\text{Card} \{i \in \{1, \dots, n\}, (\overline{\mathbf{R}}_\theta \mathbf{y})_i > 0\} \geq \text{Card}(\mathcal{J}^c) = n - k_\epsilon. \quad (.28)$$

Hence

$$k_\theta = \text{Card} \{i \in \{1, \dots, n\}, (\overline{\mathbf{R}}_\theta \mathbf{y})_i \leq 0\} \leq k_\epsilon. \quad (.29)$$

Therefore, if θ belongs to a neighborhood of $\mathbf{0}$, the condition $k_\theta \leq k_\epsilon$ is satisfied and, under the hypothesis \mathcal{H}_3 , the set of solutions $\mathcal{A}_{a,\delta}$ is also non-empty.

PROOF OF THE COERCIVITY

Let assume that, under some conditions on \mathbf{y} , $\lambda \mapsto \sigma_{\text{opt}}^2(\lambda)$ is well-defined for all $\lambda \in (0, +\infty)$. In the absence of nugget effect $\sigma_\epsilon^2 = 0$, the limit of $\bar{\mathbf{R}}_{\lambda\theta_0}$ does not exist when $\lambda \rightarrow +\infty$. Still, we can assume that the correlation matrix $\mathbf{R}_{\lambda\theta_0}$ satisfies (Berger *et al.* [43], 2001)

$$\mathbf{R}_{\lambda\theta_0} = \mathbf{J} + g_\lambda (\mathbf{D}_0 + o(1)), \quad (.30)$$

695 where

- $\lambda \mapsto g_\lambda$ is a continuous function such that $\lim_{\lambda \rightarrow +\infty} g_\lambda = 0$.
- \mathbf{D}_0 and $\mathbf{J} = \mathbf{e}\mathbf{e}^\top$ are fixed symmetric matrices.

\mathbf{D}_0 can be singular or nonsingular depending on the chosen kernel \mathbf{k} . A review of Yagloom's book ([44], 1989) shows that \mathbf{D}_0 is nonsingular only for
700 Power-Exponential ($q < 2$) and Matérn kernels with smoothness parameter $\nu < 1$ like the Exponential kernel ($\nu = 1/2$ in (3)). For the rest of Matérn kernels with smoothness parameter $\nu \geq 1$ \mathbf{D}_0 becomes singular.

Case 1 : \mathbf{D}_0 is nonsingular .

In this case, let $\mathbf{D}_\lambda = g_\lambda \mathbf{D}_0 (1 + o(1))$ such that

$$\mathbf{R}_{\lambda\theta_0} = \mathbf{J} + \mathbf{D}_\lambda. \quad (.31)$$

We consider the matrix $\bar{\mathbf{R}}_{\lambda\theta_0}$ in $\bar{\mathbf{K}} = \sigma^{-2} \bar{\mathbf{R}}_{\lambda\theta_0}$, we have

$$\bar{\mathbf{R}}_{\lambda\theta_0} = \mathbf{R}_{\lambda\theta_0}^{-1} \left[\mathbf{I}_n - \mathbf{F} (\mathbf{F}^\top \mathbf{R}_{\lambda\theta_0}^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{R}_{\lambda\theta_0}^{-1} \right]. \quad (.32)$$

By using Lemma 4, Appendix B3 in (Berger *et al.* [43], 2001) and under assumption that $\mathbf{e} \in \text{Im } \mathbf{F}$ (hypothesis \mathcal{H}_1), we have

$$\bar{\mathbf{R}}_{\lambda\theta_0} = \mathbf{D}_\lambda^{-1} \left[\mathbf{I}_n - \mathbf{F} (\mathbf{F}^\top \mathbf{D}_\lambda^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{D}_\lambda^{-1} \right]. \quad (.33)$$

Then we get

$$\bar{\mathbf{R}}_{\lambda\theta_0} = g_\lambda^{-1} \left[\mathbf{D}_0^{-1} \left(\mathbf{I}_n - \mathbf{F} (\mathbf{F}^\top \mathbf{D}_0^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{D}_0^{-1} \right) + o(1) \right]. \quad (.34)$$

Finally

$$\overline{\mathbf{R}}_{\lambda\theta_0} \stackrel{\lambda \rightarrow +\infty}{\sim} g_\lambda^{-1} \mathbf{A}, \quad (.35)$$

where

$$\mathbf{A} = \mathbf{D}_0^{-1} \left(\mathbf{I}_n - \mathbf{F} (\mathbf{F}^\top \mathbf{D}_0^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{D}_0^{-1} \right). \quad (.36)$$

Hypothesis \mathcal{H}_5 : Let \mathbf{A} be the matrix defined in (.36). We assume that \mathbf{y} does not belong to a family of vectors such that $(\mathbf{A}\mathbf{y})_i = 0$ for all $i \in \{1, \dots, n\}$ and that $\text{Card}\{i \in \{1, \dots, n\}, (\mathbf{A}\mathbf{y})_i \leq 0\} \neq na$.

By applying Lemmas 1 and 2 on \mathbf{D}_0 , we show that $(\mathbf{A})_{ii} \neq 0$ and we can write for all i in $\{1, \dots, n\}$

$$\frac{(\overline{\mathbf{R}}_{\lambda\theta_0}\mathbf{y})_i}{\sqrt{(\overline{\mathbf{R}}_{\lambda\theta_0})_{ii}}} \stackrel{\lambda \rightarrow +\infty}{\sim} g_\lambda^{-1/2} \frac{(\mathbf{A}\mathbf{y})_i}{\sqrt{(\mathbf{A})_{ii}}}. \quad (.37)$$

Analogously to the proof of Proposition 3, if we assume that $\lim_{\lambda \rightarrow +\infty} \sigma_{\text{opt}}^2(\lambda) \neq +\infty$ and by taking a sub-sequence $(\sigma_{\text{opt}}^2(\lambda_{\psi(m)}))_{m \in \mathbb{N}}$ converging to σ_∞^2

$$\frac{1}{\sigma_\infty} g_{\lambda_{\psi(m)}}^{-1/2} \frac{(\mathbf{A}\mathbf{y})_i}{\sqrt{(\mathbf{A})_{ii}}} \xrightarrow{m \rightarrow +\infty} \begin{cases} +\infty & \text{if } (\mathbf{A}\mathbf{y})_i > 0 \\ -\infty & \text{otherwise} \end{cases} \quad (.38)$$

The limit $\psi_a^{(\delta)}(\sigma_{\text{opt}}^2(\lambda_{\psi(m)}), \lambda_{\psi(m)}\theta_0)$ when $m \rightarrow +\infty$ exists and is equal to

$$a = \lim_{m \rightarrow +\infty} \psi_a^{(\delta)}(\sigma_{\text{opt}}^2(\lambda_{\psi(m)}), \lambda_{\psi(m)}\theta_0) = \frac{1}{n} \text{Card}\{i \in \{1, \dots, n\}, (\mathbf{A}\mathbf{y})_i \leq 0\}, \quad (.39)$$

which is contradictory and completes the proof.

Case 2 : \mathbf{D}_0 is singular .

In this case, one needs to go further in the Taylor expansion of $\overline{\mathbf{R}}_{\lambda\theta_0}$. We consider the matrix \mathbf{W} in Lemma 3, by Lemma 6 of Ren *et al.* ([45], 2012)

$$\overline{\mathbf{R}}_{\lambda\theta_0} = \mathbf{W} (\mathbf{W}^\top \mathbf{R}_{\lambda\theta_0} \mathbf{W})^{-1} \mathbf{W}^\top. \quad (.40)$$

By setting $\Sigma_\lambda = \mathbf{W}^\top \mathbf{R}_{\lambda\theta_0} \mathbf{W}$, the asymptotic study of $\overline{\mathbf{R}}_{\lambda\theta_0}$ is equivalent to the asymptotic study of Σ_λ . In case of Matérn kernel with noninteger smoothness $\nu \geq 1$, the matrix Σ_λ can be written as (J. Muré [46], 2020)

$$\Sigma_\lambda = g_\lambda (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W} + g_\lambda^* \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda)), \quad (.41)$$

where

- 715 – Either $g_\lambda = c\lambda^{-2k_1}$ with k_1 a nonnegative integer, or $g_\lambda = c\lambda^{-2\nu}$.
- $g_\lambda^* = c^*\lambda^{-2l}$ with $l \in (0, +\infty)$.
- \mathbf{R}_g is a differentiable mapping from $[0, +\infty)$ to \mathcal{M}_n such that $\|\mathbf{R}_g(\lambda)\| = o(\lambda^{-2l})$.
- \mathbf{D}_1 and \mathbf{D}_1^* are both fixed symmetric matrices with elements $\|x_i - x_j\|^{2k}$
- 720 where $k \in k_1 \cup \nu$ for \mathbf{D}_1 and $k = l$ for \mathbf{D}_1^* .

The matrix $\mathbf{W}^\top \mathbf{D}_1 \mathbf{W} + g_\lambda^* \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W}$ is nonsingular when $\lambda \rightarrow +\infty$, whether if $\mathbf{W}^\top \mathbf{D}_1 \mathbf{W}$ is nonsingular or if it is singular.

The case where $\mathbf{W}^\top \mathbf{D}_1 \mathbf{W}$ is nonsingular happens for Matérn kernels with smoothness $1 \leq \nu < 2$ [46], whereas the other case occurs for regular and
725 smooth Matérn kernels with $\nu \geq 2$. These kernels are however less robust in uncertainty quantification so we will give only the proof for less smooth kernels with $1 \leq \nu < 2$ in particular the Matérn 3/2 kernel.

In this case, we write Σ_λ in (.41) as

$$\Sigma_\lambda = g_\lambda \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \left(\mathbf{I}_n + g_\lambda^* (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} (\mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda)) \right). \quad (.42)$$

As \mathbf{W} is full rank matrix, Σ_λ is non-singular and

$$\Sigma_\lambda^{-1} = g_\lambda^{-1} \left(\mathbf{I}_n + g_\lambda^* (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} (\mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda)) \right)^{-1} (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1}. \quad (.43)$$

Let $\mathbf{M}_\lambda = g_\lambda^* (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} (\mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda))$, since $\|\mathbf{M}_\lambda\| \xrightarrow{\lambda \rightarrow +\infty} 0$, we can assume that $\|\mathbf{M}_\lambda\| < 1$ when λ is large enough and apply the Taylor series expansion at order 1

$$\begin{aligned} \left[\mathbf{I}_n + g_\lambda^* (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} (\mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda)) \right]^{-1} &= \mathbf{I}_n - g_\lambda^* (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} \\ &\quad \times (\mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda) + o(g_\lambda^*)). \end{aligned} \quad (.44)$$

Then, we plug this quantity into the equation (.43)

$$\begin{aligned}\Sigma_\lambda^{-1} &= g_\lambda^{-1} \left(\mathbf{I}_n - g_\lambda^* (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} (\mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda)) \right) (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} \\ &= g_\lambda^{-1} \left[(\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} - g_\lambda^* (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} (\mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda)) (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} \right].\end{aligned}\quad (.45)$$

Finally, we can write the matrix $\overline{\mathbf{R}}_{\lambda\theta_0}$ as

$$\overline{\mathbf{R}}_{\lambda\theta_0} = g_\lambda^{-1} \mathbf{W} \left[(\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} - g_\lambda^* (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} (\mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda)) (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} \right] \mathbf{W}^\top. \quad (.46)$$

We can also simplify the previous expression into

$$\overline{\mathbf{R}}_{\lambda\theta_0} = g_\lambda^{-1} (\mathbf{A} - \mathbf{B}_\lambda), \quad (.47)$$

where

$$\mathbf{A} = \mathbf{W} (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} \mathbf{W}^\top \text{ a fixed matrix.} \quad (.48)$$

$$\mathbf{B}_\lambda = g_\lambda^* \mathbf{W} (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} (\mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda)) (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} \mathbf{W}^\top \text{ such that } \mathbf{B}_\lambda \stackrel{\lambda \rightarrow +\infty}{\equiv} o(1). \quad (.49)$$

Or, equivalently, $\overline{\mathbf{R}}_{\lambda\theta_0} \stackrel{\lambda \rightarrow +\infty}{\sim} g_\lambda^{-1} \mathbf{A}$.

Lemma 7. *Let \mathbf{A} be the matrix defined in (.48) , then $\mathbf{A}_{ii} \neq 0$ for all $i \in \{1, \dots, n\}$.*

Proof. \mathbf{A} is non-singular because

$$\det \mathbf{A} = \det \mathbf{W} (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} \mathbf{W}^\top = \det (\mathbf{W}^\top \mathbf{D}_1 \mathbf{W})^{-1} \neq 0. \quad (.50)$$

\mathbf{A} is then a positive definite matrix

$$\mathbf{A}_{ii} = \mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i > 0. \quad (.51)$$

□

Hypothesis \mathcal{H}_6 : Let \mathbf{A} be the matrix defined in (.48). We assume that \mathbf{y} does not belong to a family of vectors such that $(\mathbf{A}\mathbf{y})_i = 0$ for all $i \in \{1, \dots, n\}$ and that $\text{Card}\{i \in \{1, \dots, n\}, (\mathbf{A}\mathbf{y})_i \leq 0\} \neq na$.

With Lemma 6 and Hypothesis \mathcal{H}_6 , the proof of the divergence of $\sigma_{\text{opt}}^2(\lambda)$ when $\lambda \rightarrow +\infty$ is similar to the previous case when \mathbf{D}_0 is nonsingular.

Remark 4. *The hypotheses \mathcal{H}_5 and \mathcal{H}_6 are not restrictive, one can verify numerically, that each component of $\mathbf{A}\mathbf{y}$ is not null where \mathbf{A} is one of the*
740 *matrices defined in (.36) or (.48).*