Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification

François Bachoc[†‡], Josselin Garnier[‡] and Jean-Marc Martinez[†]

†CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-sur-Yvette, France.

‡Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VII

Introduction

- Estimation of hyper-parameters in Kriging in case of Model misspecification.
- Goal: Comparison of Maximum Likelihood (ML) and Cross Validation (CV).

Framework

- Observation of a centered, unit variance, stationary Gaussian process Y on $\mathcal{X}$ with covariance function C_1.
- Vector y of observations on $x_1, ..., x_n \in \mathcal{X}$.
- Kriging metamodel $x_0 \to \left(\hat{y}_0, \hat{\sigma}^2(y)\sigma^2_{x_O}\right)$ given by the set $\mathcal{C}$ of covariance functions:

$$\mathcal{C} = \left\{\sigma^2 C_\theta, \sigma \in \mathbb{R}^+, \theta \in \Theta\right\}$$

with C_θ a stationary correlation function. $C_1 \notin \mathcal{C}$: model misspecification

- Maximum Likelihood:

$$\hat{\theta}_{ML} \in \underset{\theta\in\Theta}{\operatorname{argmin}} |R_\theta|^{\frac{1}{n}} \frac{1}{n} y^t R_\theta^{-1} y \quad \text{and} \quad \hat{\sigma}^2_{ML} = \frac{1}{n} y^t R^{-1}_{\hat{\theta}_{ML}} y.$$

- Cross-Validation, with $\hat{y}_{i,-i,\theta}, \hat{\sigma}^2_{i,-i,\theta}$ the Kriging predictive mean and variance of y_i, with covariance function C_θ, based on $(y_1, ..., y_{i-1}, y_{i+1}, y_n)$:

$$\hat{\theta}_{CV} \in \underset{\theta\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i,\hat{\theta}_{CV}})^2 \quad \text{and} \quad \hat{\sigma}^2_{CV} = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{y}_{i,-i,\hat{\theta}_{CV}})^2}{\hat{\sigma}^2_{i,-i,\hat{\theta}_{CV}}}.$$

- Thanks to the virtual Leave One Out formulas [Dub83] we have:

$$\hat{\theta}_{CV} \in \underset{\theta\in\Theta}{\operatorname{argmin}}\, y^t R_\theta^{-1} diag(R_\theta^{-1})^{-2} R_\theta^{-1} y \quad \text{and} \quad \hat{\sigma}^2_{CV} = \frac{1}{n} y^t R^{-1}_{\hat{\theta}_{CV}} \left[diag(R^{-1}_{\hat{\theta}_{CV}})\right]^{-1} R^{-1}_{\hat{\theta}_{CV}} y.$$

- Outline:
  - First step Case of the estimation of the variance hyper-parameter σ^2. Closed form expression of $\hat{\sigma}^2$: allows for a detailed and quantitative finite sample comparison.
  - Second step General case of the estimation of the hyper-parameter θ. Numerical studies on analytical functions.

Step 1: Estimation of the variance hyper-parameter

- In this case $C_\theta = C_2$, $C_2 \neq C_1$.
- Quantity of interest for $\hat{\sigma}^2$: The Risk at x_0:

$$R_{\hat{\sigma}^2, x_0} = \mathbb{E}\left[\left(\mathbb{E}\left[(\hat{y}_0 - y_0)^2 | y\right] - \hat{\sigma}^2(y)\sigma^2_{x_O}\right)^2\right].$$

  - The risk increases when the predictive variance is wrong
- Analytical expression of the risk for an estimator $\hat{\sigma}^2$ of the form $y^t M y$:

$$\boxed{\begin{aligned} R_{\hat{\sigma}^2, x_0} = & f(M_0, M_0) + 2c_1 tr(M_0) - 2c_2 f(M_0, M_1) \\ & + c_1^2 - 2c_1c_2 tr(M_1) + c_2^2 f(M_1, M_1)\end{aligned}}$$

With:

$$\begin{aligned} f(A,B) &= tr(A)tr(B) + 2tr(AB) \quad \text{for } A, B\ n\times n \text{ real matrices},\\ M_0 &= (R_2^{-1} r_2 - R_1^{-1} r_1)(r_2^t R_2^{-1} - r_1^t R_1^{-1}) R_1,\\ M_1 &= M R_1,\\ c_1 &= 1 - r_1^t R_1^{-1} r_1,\\ c_2 &= 1 - r_2^t R_2^{-1} r_2.\end{aligned}$$

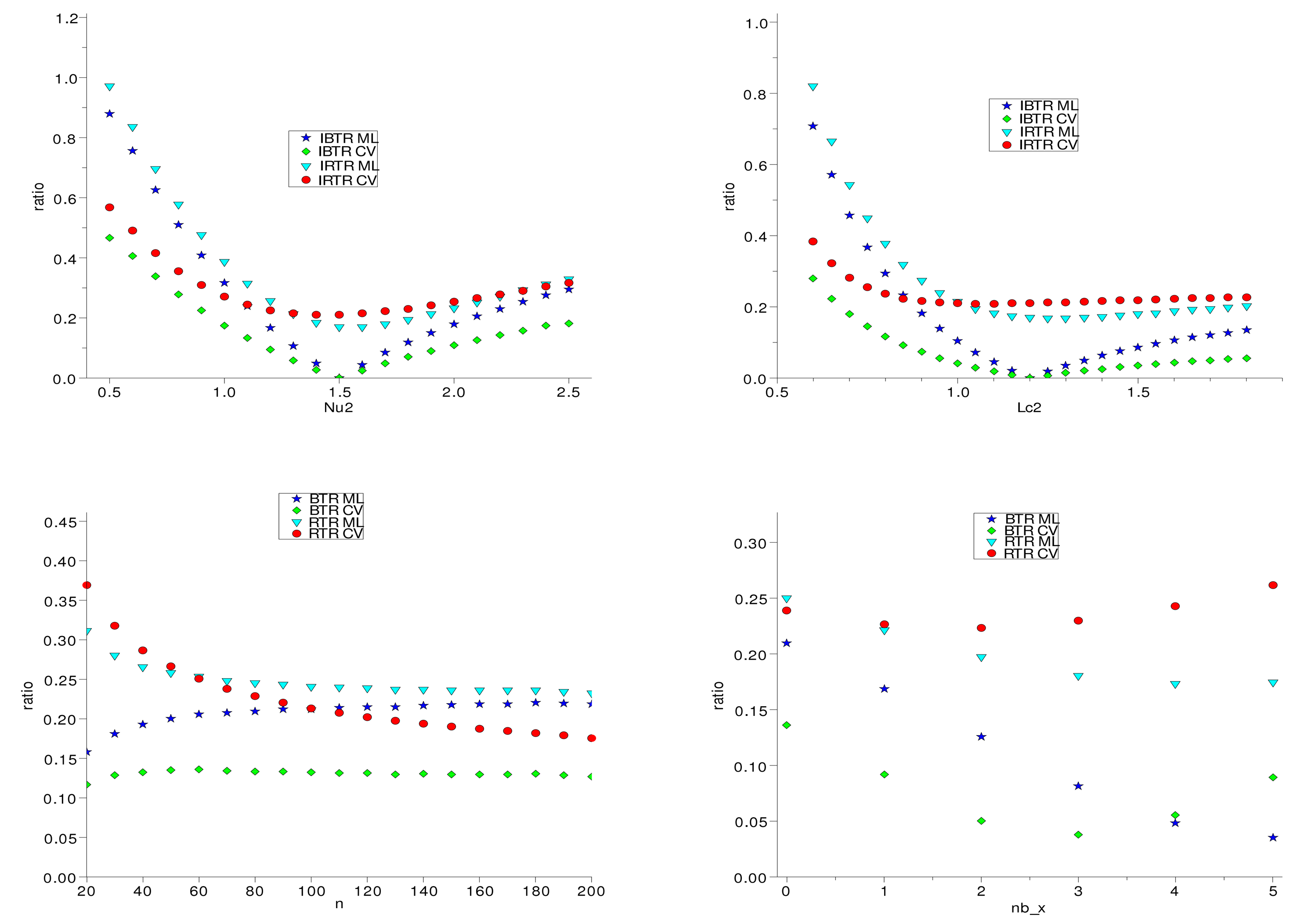
- Case $C_1 = C_2$: ML reaches the Cramer Rao bound $(\frac{2}{n})$
- Case $C_1 \neq C_2$: Numerical evaluation of the risk formulas
- Quantities of interest for an estimator $\hat{\sigma}^2$:

Quantity of interest	Expression
Risk on Target Ratio (RTR)	$RTR(x_0) = \frac{\sqrt{R_{\hat{\sigma}^2, x_0}}}{\mathbb{E}[(\hat{y}_0 - y_0)^2]}$
Integrated Risk on Target Ratio (IRTR)	$IRTR = \sqrt{\int_{\mathcal{X}} RTR^2(x_0) d\mu(x_0)}$
Bias on Target Ratio (BTR)	$BTR(x_0) = \frac{\lvert\mathbb{E}[(\hat{y}_0 - y_0)^2] - \mathbb{E}(\hat{\sigma}^2(y))\sigma^2_{x_0}\rvert}{\mathbb{E}[(\hat{y}_0 - y_0)^2]}$
Integrated Bias on Target Ratio (IBTR)	$IBTR = \sqrt{\int_{\mathcal{X}} BTR^2(x_0) d\mu(x_0)}$

- Procedure: We take $\mathcal{X} = [0,1]^d$ with uniform measure. We generate n_p designs $(x_1, ..., x_n)$ using the LHS-Maximin technique, compute each time the four criteria above (analytical formulation and Monte Carlo for integration) and plot the average.
- Setting for the figures:

Figure	C_1, C_2	n	d	n_p	x_0
Influence model error (regularity parameter)	Isotropic Matern (l_c, ν) $l_{c,1} = l_{c,2} = 1.2$, $\nu_1 = 1.5$, ν_2 varying	70	5	50	Integration
Influence model error (correlation length)	Isotropic Matern (l_c, ν) $\nu_1 = \nu_2 = 1.5$, $l_{c,1} = 1.2$, $l_{c,2}$ varying	70	5	50	Integration
Influence n	Isotropic Matern (p, l_c) $l_{c,1} = l_{c,2} = 1.2$, $\nu_1 = 1.5$, $\nu_2 = 1.8$	varying	5	1600	center
Influence x_0	Isotropic Matern (p, l_c) $l_{c,1} = l_{c,2} = 1.2$, $\nu_1 = 1.5$, $\nu_2 = 1.8$	70	5	500	varying

- Plot of the Quantities of interest:



Top left: Influence model error (regularity parameter). Top right: Influence model error (correlation length) Bot left: Influence n. Bot right: Influence x_0, x_0 has nb_x component at 0.1 and $5 - nb_x$ at 0.5, plot of BTR and RTR as a function of nb_x.

Step 2: Estimation of the correlation hyper-parameters

Procedure

- Function f on $[0,1]^d$
- Building of a Kriging Model with training sample $(x_{a,1}, ..., x_{a,n})$, with the exponential, Gaussian and Matern covariance function, and with two different cases for the hyper-parameters estimation:
  - Case 2.i: Estimation of an isotropic correlation length, and of the regularity parameter for the Matern case.
  - Case 2.a: Estimation of d correlation lengths, and of the regularity parameter for the Matern case.
- Quantities of interest on a Monte Carlo test sample $(x_{t,1}, ..., x_{t,n_t})$, with $\hat{y}_{t,i}(y_a)$ and $\sigma^2_{t,i}(y_a)$ the predictive mean and variance at $x_{t,i}$ of the built Kriging model:
  - Mean Square Error (MSE): $\frac{1}{n_t}\sum_{i=1}^{n_t}(y_{t,i} - \hat{y}_{t,i}(y_a))^2$
  - Predictive Variance Adequation (PVA): $\left|\log\left(\frac{1}{n_t}\sum_{i=1}^{n_t}\frac{(y_{t,i} - \hat{y}_{t,i}(y_a))^2}{\sigma^2_{t,i}(y_a)}\right)\right|$. The PVA increases when the predictive variance is wrong.
- Quantities of interest are averaged over n_p LHS Maximin designs.

Results

- We consider the Ishigami ($d = 3$) and Morris ($d = 10$) functions:
  - Ishigami: $\sin(\pi(2x_1 - 1)) + 7\sin(\pi(2x_2 - 1))^2 + 0.1\sin(\pi(2x_1 - 1)).(\pi(2x_3 - 1))^4$
  - Morris: An anisotropic function.
- Results:

Function	Correlation model	MSE	PVA
Ishigami	exponential case 2.i	ML: 1.99 CV: 1.97	ML: 0.35 CV: 0.23
Ishigami	exponential case 2.a	ML: 2.01 CV: 1.77	ML: 0.36 CV: 0.24
Ishigami	Gaussian case 2.i	ML: 2.06 CV: 2.11	ML: 0.18 CV: 0.22
Ishigami	Gaussian case 2.a	ML: 1.50 CV: 1.53	ML: 0.53 CV: 0.50
Ishigami	Matern case 2.i	ML: 2.19 CV: 2.29	ML: 0.18 CV: 0.23
Ishigami	Matern case 2.a	ML: 1.69 CV: 1.67	ML: 0.38 CV: 0.41
Morris	exponential case 2.i	ML: 3.07 CV: 2.99	ML: 0.31 CV: 0.24
Morris	exponential case 2.a	ML: 2.03 CV: 1.99	ML: 0.29 CV: 0.21
Morris	Gaussian case 2.i	ML: 1.33 CV: 1.36	ML: 0.26 CV: 0.26
Morris	Gaussian case 2.a	ML: 0.86 CV: 1.21	ML: 0.79 CV: 1.56
Morris	Matern case 2.i	ML: 1.26 CV: 1.28	ML: 0.24 CV: 0.25
Morris	Matern case 2.a	ML: 0.75 CV: 1.06	ML: 0.65 CV: 1.43

- With inappropriate non-smooth correlation functions family: CV performs better than ML. ML performs better when the correlation functions family is well-specified. Enforcing an isotropic correlation functions family has more negative influence on ML when the real function is anisotropic.

Conclusion

- In our studies: When the model misspecification becomes important, CV performs better than ML.
- Possible extension: Studying other Cross Validation estimation methods.

References

[Dub83] O. Dubrule. Cross validation of kriging in a unique neighborhood. *Mathematical Geology*, 15, 1983.