Cross Validation and Maximum Likelihood estimations of hyperparameters of Gaussian processes with model misspecification

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

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

‡Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VII

Introduction

- Main Context: Estimation of the variance hyperparameter in Kriging in case of Model misspecification.
- Goals:
  - Comparison of Maximum Likelihood (ML) and Cross Validation (CV) in case of model misspecification
  - Comparison based on Predictive Variance reliability

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:

$$\mathcal{C} = \left\{\sigma^2 C_2, \sigma \in \mathbb{R}^+\right\}$$

with C_2 stationary correlation function. $C_2 \neq C_1$: model misspecification
- Quantity of interest for $\hat{\sigma}$: the Risk at x_0:

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

Analytical expression of the risk for an estimator $\hat{\sigma}^2$ of the form $y^t M y$:

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

With:

$$f(A, B) = tr(A)tr(B) + 2tr(AB) \quad \text{for } A, B \; n \times n \text{ real symmetric matrices},$$

$$M_0 = R_1^{\frac{1}{2}}(R_2^{-1} r_2 - R_1^{-1} r_1)(r_2^t R_2^{-1} - r_1^t R_1^{-1}) R_1^{\frac{1}{2}},$$

$$M_1 = R_1^{\frac{1}{2}} M R_1^{\frac{1}{2}},$$

$$c_1 = 1 - r_1^t R_1^{-1} r_1,$$

$$c_2 = 1 - r_2^t R_2^{-1} r_2,$$

CV and ML estimation

- Maximum Likelihood (ML) estimator:

$$\hat{\sigma}^2_{ML} = \frac{1}{n} y^t R_2^{-1} y.$$

- Cross Validation (CV) estimator:

$$\hat{\sigma}^2_{CV} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_{i,-i})^2}{\hat{\sigma}^2_{i,-i}}.$$

with $\hat{y}_{i,-i}, \hat{\sigma}^2_{i,-i}$ the Kriging predictive mean and variance of y_i based on $(y_1, ..., y_{i-1}, y_{i+1}, y_n)$ with covariance function C_2.
- Thanks to the virtual Leave One Out formulas [Dub83] we have:

$$\hat{\sigma}^2_{CV} = \frac{1}{n} y^t R_2^{-1} \left[diag(R_2^{-1})\right]^{-1} R_2^{-1} y.$$

- Case $C_1 = C_2$: ML reaches the Cramer Rao bound $\left(\frac{2}{n}\right)$
- Case $C_1 \neq C_2$: numerical evaluation of the risk formulas

Numerical results

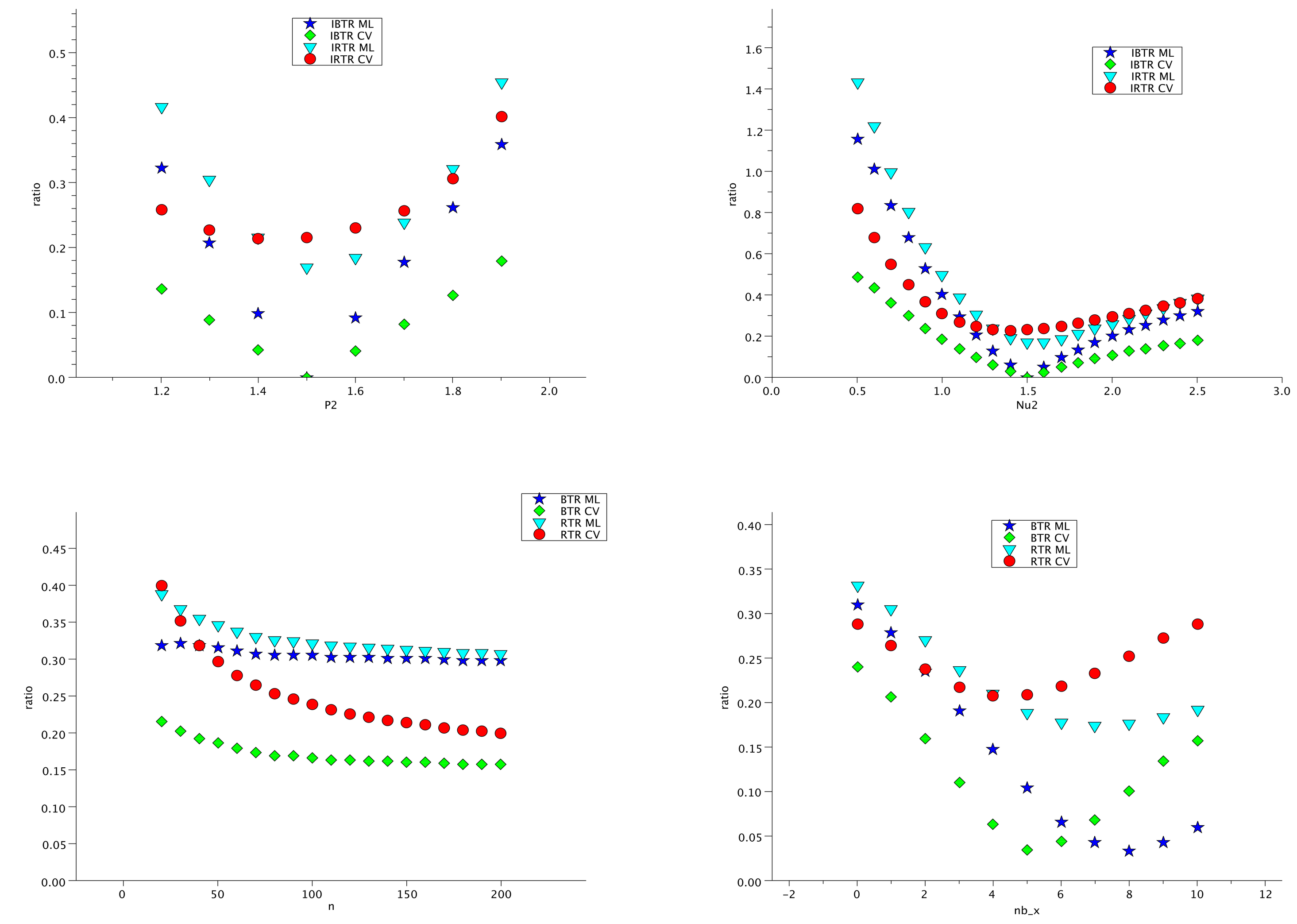
- Quantities of interest for an estimator $\hat{\sigma}$:

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

- 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 formultation 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 (Power Exponential)	Isotropic Power Exponential (l_c, p) $l_{c,1} = l_{c,2} = 1.2$, $p_1 = 1.5$, p_2 varying	70	5	50	Integration
Influence model error (Matern)	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 n	Isotropic Power Exponential (p, l_c) $l_{c,1} = l_{c,2} = 1.2$, $p_1 = 1.5$, $p_2 = 1.7$	varying	5	500	center
Influence x_0	Isotropic Power Exponential (p, l_c) $l_{c,1} = l_{c,2} = 1.2$, $p_1 = 1.5$, $p_2 = 1.7$	70	10	500	varying

- Plot of the Quantities of interest:



Top left: Influence model error (Power Exponential). Top right: Influence model error (Matern) Bot left: influence n. Bot right: influcence x_0, x_0 has nb_x component at 0.1 and $10 - nb_x$ at 0.5, plot of BTR and RTR as a function of nb_x.

Study on analytical functions

Procedure

- Function f on $[0, 1]^d$
- Building of a Kriging Model with training sample $(x_{a,1}, ..., x_{a,n})$ with two different covariance estimation procedures:
  - Case 1: covariance model $\left\{\sigma^2 C, \sigma^2 \geq 0\right\}$, with C a given correlation function. Estimation of σ by ML and CV as before.
  - Case 2: covariance model $\left\{\sigma^2 C_\theta, \sigma^2 \geq 0, \theta \in \Theta\right\}$. Classical ML estimators $\hat{\theta}_{ML}$ and $\hat{\sigma}_{ML}$ of θ and σ. CV estimation of θ minimizing the Leave One Out Mean Square Error and, given $\hat{\theta}_{CV}$, CV estimation of σ as before.
- Quantities of interest on a Monte Carlo test sample $(x_{t,1}, ..., x_{t,n_t})$:
  - 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): $\left|\log\left(\frac{1}{n_t} \sum_{i=1}^{n_t} \frac{(y_{t,i} - \hat{y}_{t,i}(y_a))^2}{\hat{\sigma}^2_{t,i}(y_a)}\right)\right|$

  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.
- Quantities of interest are averaged over n_p LHS Maximin designs.

Results

- We consider the Ishigami ($d = 3$) and Morris ($d = 10$) functions:
  - Ishigami: $\sin(-\pi + 2\pi x_1) + 7\sin((-\pi + 2\pi x_2)^2) + 0.1\sin(-\pi + 2\pi x_1).(-\pi + 2\pi x_3)^4$
  - Morris: see [GCI06] p. 17 for the exact expression.
- For Case 1 we take C isotropic exponential with fixed unit correlation length. For Case 2, C_θ is isotropic exponential (case 2.i) or anisotropic exponential (case 2.a).
- Results:

Function	n, n_p	Case for correlation model	MSE	PVA
Ishigami	$n = 70$ $n_p = 100$	1	ML: 2.75 CV: 2.75	ML: 0.41 CV: 0.35
Ishigami	$n = 70$ $n_p = 100$	$2.i$	ML: 2.87 CV: 2.86	ML: 0.36 CV: 0.42
Ishigami	$n = 70$ $n_p = 100$	$2.a$	ML: 3.21 CV: 2.84	ML: 0.39 CV: 0.40
Morris	$n = 100$ $n_p = 100$	1	ML: 6.23 CV: 6.23	ML: 0.51 CV: 0.25
Morris	$n = 100$ $n_p = 100$	$2.i$	ML: 3.14 CV: 3.05	ML: 0.27 CV: 0.25
Morris	$n = 100$ $n_p = 100$	$2.a$	ML: 2.17 CV: 2.18	ML: 0.28 CV: 0.33

- When model misspecification (high MSE) CV performs better than ML. Model misspecification can come from a too large model (over parametrization) for Ishigami and a too small model (under parametrization) for Morris.

Conclusion

- In our studies: when the model misspecification becomes important, the CV performs better than ML.
- Other Cross Validation criteria?

References

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

[GCI06] J. Garnier, C. Cannamela, and B. Iooss. Estimation de quantiles. Technical report, 2006.