Validation de codes de calcul - Validation croisée pour l'estimation d'hyper-paramètres

François Bachoc
Josselin Garnier
Jean-Marc Martinez

CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-Sur-Yvette, France
LPMA, Université Paris 7

Septembre 2012

The PhD

Two components of the PhD

- Use of Kriging model for code validation
  - Bachoc F, Bois G, and Martinez J.M, Gaussian process computer model validation method, *Submitted*.
- Work on the problem of the covariance function estimation
  - Bachoc F, Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification, *Submitted*.

Kriging for code validation
Statistical model and method studied
Conclusion

Cross Validation for hyper-parameters estimation
Context on Cross Validation
Case of a single variance parameter estimation
Conclusion

Current work and prospects

Numerical code and physical system

A numerical code, or parametric numerical model, is represented by a function f :

$$\begin{aligned} f \quad : \mathbb{R}^d \times \mathbb{R}^m &\rightarrow \mathbb{R} \\ (x, \beta) &\rightarrow f(x, \beta) \end{aligned}$$

Observations can be made of a physical system Y_{real}

$$x_i \quad \rightarrow \boxed{Y_{real}} \rightarrow y_i$$

- The inputs x are the experimental conditions
- The inputs β are the calibration parameters of the numerical code
- The outputs $f(x_i, \beta)$ and y_i are the variable of interest

A numerical code modelizes (gives an approximation of) a physical system

Statistical model (1/2)

Statistical model followed in the phd

- The physical system Y_{real} is a deterministic function
- There exist a correct parameter β and a model error function Z so that

$$Y_{real}(x) = f(x, \beta) + Z(x)$$

- The observations are noised $Y_i = Y_{real}(x_i) + \epsilon_i$, where ϵ_i are *i.i.d* centered Gaussian variables

The model is different from other classical models for inverse problems where

- $Y_i = f(x_i, \beta_i) + \epsilon_i$. The β_i are *i.i.d* realizations of the random vector β
- Hence the physical system is random

Fu S, An adaptive kriging method for characterizing uncertainty in inverse problems, *Journée du GdR MASCOT-NUM - 23 mars 2011*.

de Crécy A, A methodology to quantify the uncertainty of the physical models of a code, *Rapport CEA DEN/DANS/DM2S/STMF*.

Statistical model (2/2)

- Z is modeled as the realization of a centered Gaussian process
- A Bayesian modelling is possible for the correct parameter β
  - no prior information case. β is constant and unknown
  - prior information case. β is the realization of a Gaussian vector $\mathcal{N}(\beta_{prior}, Q_{prior})$
- Linear approximation of the code w.r.t β

$$f(x, \beta) = \sum_{j=1}^{m} h_j(x)\beta_j$$

Eventually, the statistical model is a Kriging model

$$\boxed{Y_i = \sum_{j=1}^{m} h_j(x_i)\beta_j + Z(x_i) + \epsilon_i}$$

Hence calibration and prediction are carried out within the Kriging framework

Results with the thermal-hydraulic code Flica IV (1/2)

The experiment consists in pressurized and possibly heated water passing through a cylinder. We measure the pressure drop between the two ends of the cylinder.

Quantity of interest : The part of the pressure drop due to friction : ΔP_{fro}

Two kinds of experimental conditions :

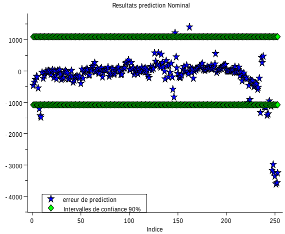
- System parameters : Hydraulic diameter D_h, Friction height H_f, Channel width e
- Environment variables : Output pressure P_s, Flowrate G_e, Parietal heat flux Φ_p, Liquid enthalpy h_e^l, Thermodynamic title X_{th}^e, Input temperature T_e

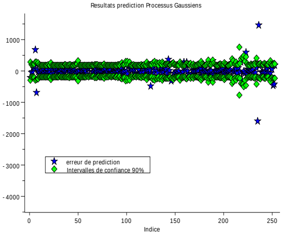
We dispose of 253 experimental results

Important : Among the 253 experimental results, only 8 different system parameters $\rightarrow$ Not enough to use the Gaussian processes model for prediction for new system parameters $\rightarrow$ We predict for new environment variables only

Results with the thermal-hydraulic code Flica IV (2/2)

	RMSE	90% Confidence Intervals
Nominal code	$661\,Pa$	$234/253 \approx 0.925$
Gaussian Processes	$189\,Pa$	$235/253 \approx 0.93$





Conclusion on code Validation

- We can improve the prediction capability of the code by completing the physical representation with a statistical model
- Number of experimental results needs to be sufficient. No extrapolation

Further questions

- Assumptions of this statistical model. A statistical question or an engineering/physical question ?

Kriging for code validation
Statistical model and method studied
Conclusion

Cross Validation for hyper-parameters estimation
Context on Cross Validation
Case of a single variance parameter estimation
Conclusion

Current work and prospects

Cross Validation (Leave-One-Out)

Gaussian Process Y observed at $x_1, ..., x_n$ with values $y = (y_1, ..., y_n)^t$

Cross Validation (Leave-One-Out) principle

- $\hat{y}_{i,-i} = \mathbb{E}(Y(x_i)|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$
- $c_{i,-i}^2 = \mathbb{E}((Y(x_i) - \hat{y}_{i,-i})^2|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$

Can be used for Kriging verification or for covariance function selection

Remark on Kriging model verification

When a new test sample is available, one can decorrelates the prediction errors :

L.S. Bastos and A. O'Hagan Diagnostics for Gaussian Process Emulators, *Technometrics, 51 (4), 425-438*.

However, decorrelating the LOO errors is equivalent to decorrelating the original observations

Virtual Cross Validation

When mean of Y is parametric : $\mathbb{E}(Y(x)) = \sum_{i=1}^{p} \beta_i h_i(x)$. Let

- $\mathbf{H}$ the $n \times p$ matrix with $\mathbf{H}_{i,j} = h_j(x_i)$
- $\mathbf{R}$ the covariance matrix of $y = (y_1, ..., y_n)$

Virtual Leave-One-Out

With

$$\mathbf{Q}^{-} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}$$

We have :

$$y_i - \hat{y}_{i,-i} = \frac{1}{(\mathbf{Q}^{-})_{i,i}}(\mathbf{Q}^{-}y)_i \quad \text{and} \quad c_{i,-i}^2 = \frac{1}{(\mathbf{Q}^{-})_{i,i}}$$

If Bayesian case for β ($\beta \sim \mathcal{N}(\beta_{prior}, \mathbf{Q}_{prior})$), then same formula holds replacing $\mathbf{Q}^{-}$ with $(\mathbf{R} + \mathbf{H}\mathbf{Q}_{prior}\mathbf{H}^t)^{-1}$

O. Dubrule, Cross Validation of Kriging in a Unique Neighborhood, *Mathematical Geology*, 1983.

Cross Validation for covariance function estimation (1/2)

Let $\{\sigma^2 K_\theta, \sigma^2 \geq 0, \theta \in \Theta\}$ be a set of covariance function for Y, with K_θ a correlation function. Let

- $\hat{y}_{\theta,i,-i} = \mathbb{E}_{\sigma^2,\theta}(Y(x_i)|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$
- $\sigma^2 c^2_{\theta,i,-i} = \mathbb{E}_{\sigma^2,\theta}((Y(x_i) - \hat{y}_{\theta,i,-i})^2|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$

Leave-One-Out criteria we study

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

and

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

Cross Validation for covariance function estimation (2/2)

- Leave-One-Out estimation is tractable
- Other Cross-Validation criteria exist

  C.E. Rasmussen and C.K.I. Williams, Gaussian Processes for Machine Learning, *The MIT Press, Cambridge*, 2006.
- To the best of our knowledge : problems of the choice of the cross validation criterion and of the cross validation procedure are not fully solved for Kriging
- It is our experience that when one is primarily interested in prediction mean square error and point-wise estimation of the prediction mean square error, the Leave-One-Out criteria presented are reasonable

Objectives

We want to study the cases of model misspecification, that is to say the cases when the true covariance function K_1 of Y is far from $\mathcal{K} = \left\{\sigma^2 K_\theta, \sigma^2 \geq 0, \theta \in \Theta\right\}$

In this context we want to compare Leave-One-Out and Maximum Likelihood estimators from the point of view of prediction mean square error and point-wise estimation of the prediction mean square error

We proceed in two steps

- When $\mathcal{K} = \left\{\sigma^2 K_2, \sigma^2 \geq 0\right\}$, with K_2 a correlation function, and K_1 is the true covariance function : Theoretical formula and numerical tests
- In the general case : Numerical studies

Kriging for code validation
Statistical model and method studied
Conclusion

Cross Validation for hyper-parameters estimation
Context on Cross Validation
Case of a single variance parameter estimation
Conclusion

Current work and prospects

Setting

Let x_0 be a new point and assume the mean of Y is zero and K_1 is unit-variance stationary. Let

- r_1 be the covariance vector between $x_1, ..., x_n$ and x_0 with covariance function K_1
- r_2 be the covariance vector between $x_1, ..., x_n$ and x_0 with covariance function K_2
- $\mathbf{R}_1$ be the covariance matrix of $x_1, ..., x_n$.with covariance function K_1
- $\mathbf{R}_2$ be the covariance matrix of $x_1, ..., x_n$.with covariance function K_2

$\hat{y}_0 = r_2^t \mathbf{R}_2^{-1} y$ is the Kriging prediction

$\mathbb{E}\left[(\hat{y}_0 - Y_0)^2 | y\right] = (r_1^t \mathbf{R}_1^{-1} y - r_2^t \mathbf{R}_2^{-1} y)^2 + 1 - r_1^t \mathbf{R}_1^{-1} r_1$ is the conditional mean square error of the non-optimal prediction

One estimates σ^2 with $\hat{\sigma}^2$ and estimates the conditional mean square error with $\hat{\sigma}^2 c_{x_0}^2$ with $c_{x_0}^2 := 1 - r_2^t \mathbf{R}_2^{-1} r_2$

The Risk

The Risk

We study the Risk criterion for an estimator $\hat{\sigma}^2$ of σ^2

$$R_{\hat{\sigma}^2,x_0} = \mathbb{E}\left[\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2|y\right] - \hat{\sigma}^2 c_{x_0}^2\right)^2\right]$$

Formula for quadratic estimators

When $\hat{\sigma}^2 = y^t\mathbf{M}y$, we have

$$\begin{aligned} R_{\hat{\sigma}^2,x_0} &= f(\mathbf{M}_0,\mathbf{M}_0) + 2c_1 tr(\mathbf{M}_0) - 2c_2 f(\mathbf{M}_0,\mathbf{M}_1) \\ &\quad + c_1^2 - 2c_1 c_2 tr(\mathbf{M}_1) + c_2^2 f(\mathbf{M}_1,\mathbf{M}_1) \end{aligned}$$

with

$$f(\mathbf{A},\mathbf{B}) = tr(\mathbf{A})tr(\mathbf{B}) + 2tr(\mathbf{AB})$$

$$\mathbf{M}_0 = (\mathbf{R}_2^{-1}r_2 - \mathbf{R}_1^{-1}r_1)(r_2^t\mathbf{R}_2^{-1} - r_1^t\mathbf{R}_1^{-1})\mathbf{R}_1$$

$$\mathbf{M}_1 = \mathbf{M}\mathbf{R}_1$$

$$c_1 = 1 - r_1^t\mathbf{R}_1^{-1}r_1$$

$$c_2 = 1 - r_2^t\mathbf{R}_2^{-1}r_2$$

CV and ML estimation

- ML estimation :

$$\hat{\sigma}^2_{ML} = \frac{1}{n} y^t \mathbf{R}_2^{-1} y$$

$var(\hat{\sigma}^2_{ML})$ reaches the Cramer-Rao bound $\frac{2}{n}$

- CV estimation :

$$\hat{\sigma}^2_{CV} = \frac{1}{n} y^t \mathbf{R}_2^{-1} \left[diag(\mathbf{R}_2^{-1}) \right]^{-1} \mathbf{R}_2^{-1} y$$

$var(\hat{\sigma}^2_{CV})$ can reach 2

- When $K_2 = K_1$, ML is best. Numerical study when $K_2 \neq K_1$

Criteria for numerical studies (1/2)

Risk on Target Ratio (RTR),

$$RTR(x_0) = \frac{\sqrt{R_{\hat{\sigma}^2, x_0}}}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]} = \frac{\sqrt{\mathbb{E}\left[\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2 | y\right] - \hat{\sigma}^2 c_{x_0}^2\right)^2\right]}}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]}$$

Bias-variance decomposition

$$R_{\hat{\sigma}^2, x_0} = \left(\underbrace{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right] - \mathbb{E}\left(\hat{\sigma}^2 c_{x_0}^2\right)}_{\text{bias}}\right)^2 \underbrace{+ var\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2 | y\right] - \hat{\sigma}^2 c_{x_0}^2\right)}_{\text{variance}}$$

Bias on Target Ratio (BTR) criterion

$$BTR(x_0) = \frac{\left|\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right] - \mathbb{E}\left(\hat{\sigma}^2 c_{x_0}^2\right)\right|}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]}$$

Criteria for numerical studies (2/2)

$$\Big(\underbrace{RTR}_{\text{relative error}}\Big)^2 = \Big(\underbrace{BTR}_{\text{relative bias}}\Big)^2 + \underbrace{\frac{var\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2|y\right] - \hat{\sigma}^2 c_{x_0}^2\right)}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]^2}}_{\text{relative variance}}$$

Integrated criteria on the prediction domain $\mathcal{X}$

$$IRTR = \sqrt{\int_{\mathcal{X}} RTR^2(x_0)d\mu(x_0)}$$

and

$$IBTR = \sqrt{\int_{\mathcal{X}} BTR^2(x_0)d\mu(x_0)}$$

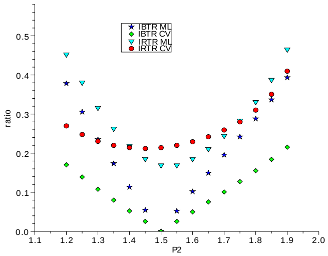
Numerical results

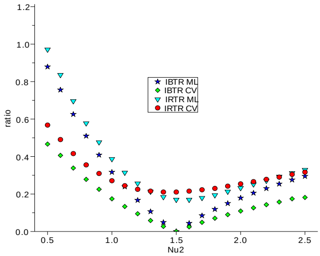
70 observations on $[0,1]^5$. Mean over LHS-Maximin DoE's.

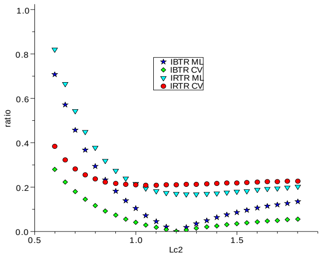
Top : K_1 and K_2 are power-exponential, with $l_{c,1} = l_{c,2} = 1.2$, $p_1 = 1.5$, and p_2 varying.

Bot left : K_1 and K_2 are Matérn (non-tensorized), with $l_{c,1} = l_{c,2} = 1.2$, $\nu_1 = 1.5$, and ν_2 varying.

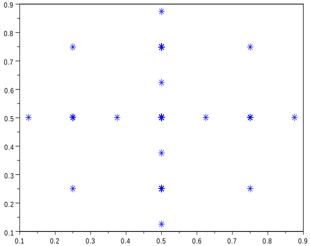
Bot right : K_1 and K_2 are Matérn $\frac{3}{2}$ (non-tensorized), with $l_{c,1} = 1.2$, and $l_{c,2}$ varying.

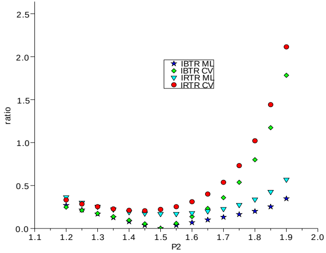


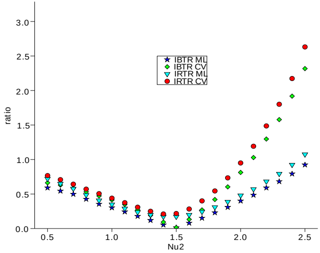


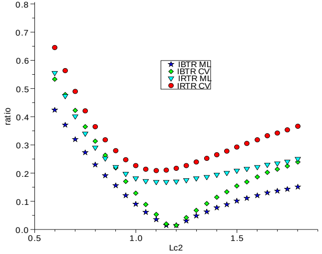


Case of a regular grid (Smolyak construction)









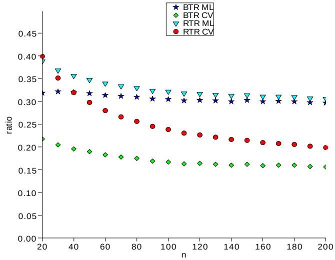
Influence of the number of points

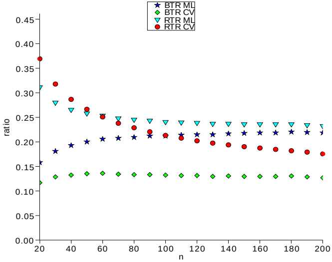
n observations on $[0,1]^5$. Pointwise prediction (center).

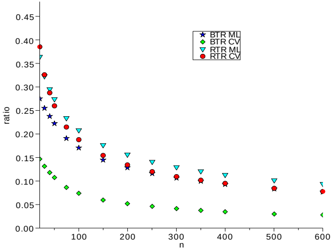
Top : K_1 and K_2 are power-exponential, with $l_{c,1} = l_{c,2} = 1.2$, $p_1 = 1.5$, and $p_2 = 1.7$.

Bot left : K_1 and K_2 are Matérn (non-tensorized), with $l_{c,1} = l_{c,2} = 1.2$, $\nu_1 = 1.5$, and $\nu_2 = 1.8$.

Bot right : K_1 and K_2 are Matérn $\frac{3}{2}$ (non-tensorized), with $l_{c,1} = 1.2$, and $l_{c,2} = 1.8$.







Conclusion on Cross Validation

- We study robustness relatively to prediction mean square errors and point-wise mean square error estimation
- For the variance estimation, CV is more robust than ML to correlation function misspecification
- This is not true for the Smolyak construction we tested
- In the general case of correlation function estimation $\rightarrow$ this is globally confirmed in a case study on analytical functions

Possible perspectives

- Quantify the suitability of CV given a DoE ?
- Problem of the choice of the CV procedure

Kriging for code validation

Statistical model and method studied

Conclusion

Cross Validation for hyper-parameters estimation

Context on Cross Validation

Case of a single variance parameter estimation

Conclusion

Current work and prospects

Current work and prospects

Current work :

- In an expansion asymptotic context, is the regular grid a local optimum for covariance function estimation ?
- Work on ML and CV estimators

Possible prospect : For the context of model misspecification : analytical study of an AR model of the kind $X(t) = \alpha X(t-1) + \epsilon$