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

François Bachoc  
Josselin Garnier  
Jean-Marc Martinez

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

June 2012
Two components of the PhD

- Use of Kriging model for code validation
  

- 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*. 
Context for Cross Validation

Case of a single variance parameter

Numerical studies in the general case

Conclusion
Cross Validation (Leave-One-Out)

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

Cross Validation (Leave-One-Out) principle

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

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

Can be used for Kriging verification or for covariance function selection
Virtual Cross Validation

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

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

Virtual Leave-One-Out

With

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

We have:

$$y_i - \hat{y}_{i,-i} = \left(\text{diag}(Q^-)\right)^{-1} Q^- y \quad \text{and} \quad c^2_{i,-i} = \frac{1}{(Q^-)_{i,i}}$$

If Bayesian case for $\beta$ ($\beta \sim N(\beta_{\text{prior}}, Q_{\text{prior}})$), then same formula holds replacing $Q^-$ with $(R + HQ_{\text{prior}} H^T)^{-1}$

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 \( K_\theta \) a correlation function. Let

\[
\hat{y}_{\theta,i,-i} = \mathbb{E}_{\sigma^2,\theta}(Y(X_i)|y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)
\]

\[
\sigma^2 c_{\theta,i,-i}^2 = \mathbb{E}_{\sigma^2,\theta}((Y(X_i) - \hat{y}_{\theta,i,-i})^2|y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)
\]

Leave-One-Out criteria we study

\[
\hat{\theta}_{CV} \in \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (y_i - \hat{y}_{\theta,i,-i})^2
\]

and

\[
\hat{\sigma}_{CV}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_{\hat{\theta}_{CV},i,-i})^2}{c_{\hat{\theta}_{CV},i,-i}^2}
\]
Cross Validation for covariance function estimation (2/2)

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


- 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 intuition 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

$$K = \{ \sigma^2 K_\theta, \sigma^2 \geq 0, \theta \in \Theta \}$$

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 $K = \{ \sigma^2 K_2, \sigma^2 \geq 0 \}$, 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
Context for Cross Validation

Case of a single variance parameter

Numerical studies in the general case

Conclusion
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 $x_1, \ldots, x_n$ and $x_0$ with covariance function $K_1$
- $r_2$ be the covariance vector between $x_1, \ldots, x_n$ and $x_0$ with covariance function $K_2$
- $R_1$ be the covariance matrix of $x_1, \ldots, x_n$ with covariance function $K_1$
- $R_2$ be the covariance matrix of $x_1, \ldots, x_n$ with covariance function $K_2$

$\hat{y}_0 = r_2^t R_2^{-1} y$ is the Kriging prediction

$$E [(\hat{y}_0 - Y_0)^2 | y] = (r_1^t R_1^{-1} y - r_2^t R_2^{-1} y)^2 + 1 - r_1^t R_1^{-1} r_1$$ is the conditional mean square error of the non-optimal prediction

One estimates $\sigma^2$ with $\hat{\sigma}^2$ and estimates the conditional mean square error with $\hat{\sigma}^2 c_{x_0}^2$ with $c_{x_0}^2 := 1 - r_2^t R_2^{-1} r_2$
The Risk

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

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

Formula for quadratic estimators

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

$$R_{\hat{\sigma}^2, x_0} = f(M_0, M_0) + 2c_1 \text{tr}(M_0) - 2c_2 f(M_0, M_1)$$
$$+ c_1^2 - 2c_1 c_2 \text{tr}(M_1) + c_2^2 f(M_1, M_1)$$

with

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

$$M_0 = (R_2^{-1} r_2 - R_1^{-1} r_1)(r_2^t R_2^{-1} - r_1^t R_1^{-1})R_1$$

$$M_1 = MR_1$$

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

$$c_2 = 1 - r_2^t R_2^{-1} r_2$$
CV and ML estimation

- ML estimation:
  \[ \hat{\sigma}^2_{ML} = \frac{1}{n} y^t R_2^{-1} y \]
  \( \text{var}(\hat{\sigma}^2_{ML}) \) reaches the Cramer-Rao bound \( \frac{2}{n} \)

- CV estimation:
  \[ \hat{\sigma}^2_{CV} = \frac{1}{n} y^t R_2^{-1} \left[ \text{diag}(R_2^{-1}) \right]^{-1} R_2^{-1} y \]
  \( \text{var}(\hat{\sigma}^2_{CV}) \) 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),

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

Bias-variance decomposition

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

Bias on Target Ratio (BTR) criterion

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

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

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 $\nu_2$ 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.

Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification
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 $p_2 = 1.7$.

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

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

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

Case of a single variance parameter

Numerical studies in the general case

Conclusion
Work on analytical functions

Consider a deterministic function $f$ on $[0, 1]^d$

- Ishigami function:
  \[
  f(x_1, x_2, x_3) = \sin(-\pi+2\pi x_1) + 7 \sin((-\pi+2\pi x_2))^2 + 0.1 \sin(-\pi+2\pi x_1).(-\pi+2\pi x_3)^4
  \]

- Morris function:
  \[
  f(x) = \sum_{i=1}^{10} w_i(x) + \sum_{1 \leq i < j \leq 6} w_i(x) w_j(x) + \sum_{1 \leq i < j < k \leq 5} w_i(x) w_j(x) w_k(x) \\
  + \sum_{1 \leq i < j < k < l \leq 4} w_i(x) w_j(x) w_k(x) w_l(x),
  \]
  with
  \[
  w_i(x) = \begin{cases} 
  2 \left( \frac{1.1 x_i}{x_i+0.1} - 0.5 \right), & \text{if } i = 3, 5, 7 \\
  2(x_i - 0.5), & \text{otherwise}
  \end{cases}
  \]
Comparison criteria

Learning sample $y_{a,1}, \ldots, y_{a,n}$. Test sample $y_{t,1}, \ldots, y_{t,n_t}$

Mean Square Error (MSE) criterion:

$$MSE = \frac{1}{n_t} \sum_{i=1}^{n_t} (y_{t,i} - \hat{y}_{t,i}(y_a))^2$$

Predictive Variance Adequation (PVA) criterion:

$$PVA = \log \left( \frac{1}{n_t} \sum_{i=1}^{n_t} \frac{(y_{t,i} - \hat{y}_{t,i}(y_a))^2}{\hat{\sigma}^2 c_{t,i}^2(y_a)} \right)$$

We average MSE and PVA over $n_p = 100$ LHS Maximin DoE’s. For each DoE: covariance estimation and Kriging prediction.
Results with enforced correlation

We use tensorized Exponential and Gaussian correlation functions for the Ishigami function

<table>
<thead>
<tr>
<th>Correlation model</th>
<th>Enforced hyper-parameters</th>
<th>MSE</th>
<th>PVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>[1, 1, 1]</td>
<td>2.01</td>
<td>ML : 0.50 CV : 0.20</td>
</tr>
<tr>
<td>Exponential</td>
<td>[1.3, 1.3, 1.3]</td>
<td>1.94</td>
<td>ML : 0.46 CV : 0.23</td>
</tr>
<tr>
<td>Exponential</td>
<td>[1.20, 5.03, 2.60]</td>
<td>1.70</td>
<td>ML : 0.54 CV : 0.19</td>
</tr>
<tr>
<td>Gaussian</td>
<td>[0.5, 0.5, 0.5]</td>
<td>4.19</td>
<td>ML : 0.98 CV : 0.35</td>
</tr>
<tr>
<td>Gaussian</td>
<td>[0.31, 0.31, 0.31]</td>
<td>2.03</td>
<td>ML : 0.16 CV : 0.23</td>
</tr>
<tr>
<td>Gaussian</td>
<td>[0.38, 0.32, 0.42]</td>
<td>1.32</td>
<td>ML : 0.28 CV : 0.29</td>
</tr>
</tbody>
</table>

- Misspecified cases: Exponential and Gaussian isotropic
- ML have the highest PVA in the worst misspecification cases
Setting for estimated correlation

- Work on three correlation families
  - Exponential tensorized
  - Gaussian
  - Matérn with estimated regularity parameter
- Work in the isotropic and anisotropic case
  - Case2.i: A common correlation length is estimated
  - Case2.a: $d$ different correlation lengths are estimated
### Results for estimated correlation : Ishigami

<table>
<thead>
<tr>
<th>Function</th>
<th>Correlation model</th>
<th>MSE</th>
<th>PVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ishigami</td>
<td>exponential case 2.i</td>
<td>ML : 1.99 CV : 1.97</td>
<td>ML : 0.35 CV : 0.23</td>
</tr>
<tr>
<td>Ishigami</td>
<td>exponential case 2.a</td>
<td>ML : 2.01 CV : 1.77</td>
<td>ML : 0.36 CV : 0.24</td>
</tr>
<tr>
<td>Ishigami</td>
<td>Gaussian case 2.i</td>
<td>ML : 2.06 CV : 2.11</td>
<td>ML : 0.18 CV : 0.22</td>
</tr>
<tr>
<td>Ishigami</td>
<td>Gaussian case 2.a</td>
<td>ML : 1.50 CV : 1.53</td>
<td>ML : 0.53 CV : 0.50</td>
</tr>
<tr>
<td>Ishigami</td>
<td>Matérn case 2.i</td>
<td>ML : 2.19 CV : 2.29</td>
<td>ML : 0.18 CV : 0.23</td>
</tr>
<tr>
<td>Ishigami</td>
<td>Matérn case 2.a</td>
<td>ML : 1.69 CV : 1.67</td>
<td>ML : 0.38 CV : 0.41</td>
</tr>
</tbody>
</table>

- Gaussian and Matérn are more adapted than exponential because of smoothness (→ smaller MSE)
- Estimating several correlation lengths is more adapted
- In the exponential case, CV has smaller PVA and smaller or equal MSE
- In the Gaussian and Matérn cases, ML has MSE and PVA slightly smaller
Results for estimated correlation : Morris

<table>
<thead>
<tr>
<th>Function</th>
<th>Correlation model</th>
<th>MSE</th>
<th>PVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morris</td>
<td>exponential case 2.i</td>
<td>ML : 3.07</td>
<td>CV : 2.99</td>
</tr>
<tr>
<td>Morris</td>
<td>exponential case 2.a</td>
<td>ML : 2.03</td>
<td>CV : 1.99</td>
</tr>
<tr>
<td>Morris</td>
<td>Gaussian case 2.i</td>
<td>ML : 1.33</td>
<td>CV : 1.36</td>
</tr>
<tr>
<td>Morris</td>
<td>Gaussian case 2.a</td>
<td>ML : 0.86</td>
<td>CV : 1.21</td>
</tr>
<tr>
<td>Morris</td>
<td>Matérn case 2.i</td>
<td>ML : 1.26</td>
<td>CV : 1.28</td>
</tr>
<tr>
<td>Morris</td>
<td>Matérn case 2.a</td>
<td>ML : 0.75</td>
<td>CV : 1.06</td>
</tr>
</tbody>
</table>

- Gaussian and Matérn are more adapted than exponential because of smoothness (→ smaller MSE)
- Estimating several correlation lengths is more adapted
- In the Exponential case, CV has slightly smaller MSE and smaller PVA
- For Gaussian and Matérn 2.a, ML has smaller MSE and PVA
- For Gaussian and Matérn, going from 2.a to 2.i causes much more harm to ML than CV
Context for Cross Validation

Case of a single variance parameter

Numerical studies in the general case

Conclusion
Conclusion

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 → this is globally confirmed in a case study on analytical functions.

Possible perspectives
- Quantify the incompatibility of a DoE for CV?
- Problem of the choice of the CV procedure

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