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

François Bachoc1, Josselein Garnier1 and Jean-Marc Martinez1

1CEA-Saclay, DEN, DM25, STMP, LGIS, F-91919 Gif-sur-Yvette, France.
2Laboratoire 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 \( X \) with covariance function \( C \).

• Vector \( y \) of observations on \( x_1, \ldots, x_n \in X \).

• Kriging metamodel \( \hat{y}(x) = \mathbf{k}^T(x) \hat{\theta} \mathbf{p} \) of the form \( \mathbb{E}[Y(x)] \) and \( \text{Var}(Y(x)) \).

• With \( \hat{C} \) a stationary correlation function, \( \mathbb{E}[Y(x)] \) and \( \text{Var}(Y(x)) \) are defined by

\[
C(p) = \mathbb{E}[Y(x)p] = \mathbb{E}[\mathbf{k}^T(x) \mathbf{p}] = \mathbf{k}^T(x) \mathbf{p}
\]

with \( \mathbf{p} \) a vector of \( n \) parameters and \( \mathbf{C} \) the covariance matrix.

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

Model misspecification

8

Isotropic Matern (\( \nu \))

Observation of a centered, unit variance, stationary Gaussian process \( C \)

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

Maximum Likelihood:

\[
\hat{\theta}_{MV} \in \text{argmin}_{\theta} \sum_{i=1}^{n} \left( y_i - \text{Kriging}(\mathbf{x}_i; \theta) \right)^2
\]

Cross-Validation:

\[
\hat{\theta}_{CV} \in \text{argmin}_{\theta} \sum_{i=1}^{n} \left( y_i - \text{Kriging}(\mathbf{x}_i; \theta) \right)^2
\]

Thanks to the virtual Leave One Out formulas \([Dub83]\) we have:

\[
\hat{\theta}_{CV} \in \text{argmin}_{\theta} \sum_{i=1}^{n} \left( y_i - \text{Kriging}(\mathbf{x}_i; \theta) \right)^2
\]

Outline:

• First step: Case of the estimation of the hyper-parameter \( \theta \). Observations of the hyper-parameters of the form \( \mathbb{E}[Y(x)] \) and \( \text{Var}(Y(x)) \).

• Second step: Case of the estimation of the hyper-parameter \( \theta \). Observations of the hyper-parameters of the form \( \mathbb{E}[Y(x)] \) and \( \text{Var}(Y(x)) \).

Step 1: Estimation of the variance hyper-parameter

In this case \( C_0 = C_1 \) if \( C_2 \in \mathcal{C} \).

Quantity of interest for \( \theta_2 \), the Risk at \( x_3 \):

\[
\text{Risk}(x_3) = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \left( y_i - \text{Kriging}(\mathbf{x}_i; \theta) \right)^2\right]
\]

The risk increases when the predictive variance is wrong.

Analytical expression of the risk for an estimator \( \theta_2 \) of the form \( \mathbb{E}[Y(x)] \) and \( \text{Var}(Y(x)) \):

\[
\text{Risk}(x_3) = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \left( y_i - \text{Kriging}(\mathbf{x}_i; \theta_2) \right)^2\right] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \left( y_i - \mathbb{E}[Y(x_i)] \right)^2\right]
\]

With:

\[
f(A; B) = \text{tr}(AB) + 2\text{tr}(AB) \quad \text{for} \quad A, B \in \mathbb{R}^{n \times n}
\]

Influence model error (regularity parameter):

\[
|I_{\text{model error}}| = \left| \hat{\theta} - \theta \right|
\]

Influence model error (correlation length):

\[
|I_{\text{model error}}| = \left| \hat{\theta} - \theta \right|
\]

Influence of the regularity parameter:

\[
|I_{\text{model error}}| = \left| \hat{\theta} - \theta \right|
\]

Influence of the correlation length:

\[
|I_{\text{model error}}| = \left| \hat{\theta} - \theta \right|
\]

References


Graphical representation of the quantities of interest:

Top left: Influence model error (regularity parameter). Top right: Influence model error (correlation length) but left: Influence \( \theta \).

Results:

• We consider the Ishigami function \( d = 3 \) and Morris \( d = 10 \) functions.

• Ishigami: \( \sin(2x_1x_2x_3) + 7\sin^2(x_1) + 0.1(x_2 - 1)^2 + 0.1\sin(2x_1x_3) - (2x_3 - 1)^4 \)

• Morris: An autostistic function.

• Results:

<table>
<thead>
<tr>
<th>Function</th>
<th>Correlation model</th>
<th>MSE</th>
<th>PVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ishigami exponential case 2a</td>
<td>ML</td>
<td>2.01 CSV</td>
<td>0.18 CV</td>
</tr>
<tr>
<td>Morris exponential case 2a</td>
<td>ML</td>
<td>2.01 CSV</td>
<td>0.18 CV</td>
</tr>
<tr>
<td>Ishigami exponential case 2b</td>
<td>ML</td>
<td>2.01 CSV</td>
<td>0.18 CV</td>
</tr>
<tr>
<td>Morris exponential case 2b</td>
<td>ML</td>
<td>2.01 CSV</td>
<td>0.18 CV</td>
</tr>
</tbody>
</table>

• With inappropriate non-smooth correlation functions family, CV performs better than ML, but MSE performs better when the correlation functions family is well-specified. Enhancing an isotropic correlation functions family has more positive 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.

Step 2: Estimation of the correlation hyper-parameters

Procedure:

• Function \( f \) on \([0,1]^d\).

• Building of a Kriging Model with training sample \((x_{1,1}, \ldots, x_{1,n})\) with the exponential, Gaussian and Matérn covariance function, and with two different cases for the hyper-parameters estimation:

• Case 2.1: Estimation of an isotropic correlation length, and of the regularity parameter for the Matérn case.

• Case 2.2: Estimation of a correlation length, and of the regularity parameter for the Matérn case.

• Quantities of interest on a Monte Carlo test sample \((x_{1,1}, \ldots, x_{1,n})\): with \( \hat{\theta}_0(x_i) \) and \( \hat{\sigma}_2(y_i) \) the predictive mean and variance at \( x_{1,i} \) of the bulk Kriging model:

\[
\text{Predicitive Variance Adecuation (PVA)}, \quad \text{Risk on Target Ratio (RTR)}
\]

• Quantities of interest are averaged over \( n \) LHS Maxim design.

Results:

• We consider the Ishigami \((d = 3)\) and Morris \((d = 10)\) functions.

• Ishigami: \( \sin(2x_1x_2x_3) + 7\sin^2(x_1) + 0.1(x_2 - 1)^2 + 0.1\sin(2x_1x_3) - (2x_3 - 1)^4 \)

• Morris: An autostistic function.

• Results:

<table>
<thead>
<tr>
<th>Function</th>
<th>Correlation model</th>
<th>MSE</th>
<th>PVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ishigami exponential case 2a</td>
<td>ML</td>
<td>2.01 CSV</td>
<td>0.18 CV</td>
</tr>
<tr>
<td>Morris exponential case 2a</td>
<td>ML</td>
<td>2.01 CSV</td>
<td>0.18 CV</td>
</tr>
<tr>
<td>Ishigami exponential case 2b</td>
<td>ML</td>
<td>2.01 CSV</td>
<td>0.18 CV</td>
</tr>
<tr>
<td>Morris exponential case 2b</td>
<td>ML</td>
<td>2.01 CSV</td>
<td>0.18 CV</td>
</tr>
</tbody>
</table>

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