7. Kriging – Gaussian Process Model (Ch.11.5 Experiments with Computer Models)

Hae-Jin Choi
School of Mechanical Engineering, Chung-Ang University
Introduction to Metamodel

- Metamodel is ‘model of a model’
- Computer simulation model (such as FEM) is a model of a physical system.
- Metamodeling is modeling the computer simulation model with a mathematical construct. (i.e., model of a simulation model)
- Why? Metamodel
  - Computer simulation tends to require a large amount of computing resources (i.e., computationally expensive)
  - Design optimization, which is an iterative process to find an optimum, cannot be performed with computationally expensive simulation model
  - A metamodel reduces the computing expenses significantly, since it is an mathematical equation.
Metamodelling of computer experiments

- Two types of computer simulation models
  - Stochastic simulation model: output (response) is random variable
  - Deterministic simulation model: output (response) is deterministic by mathematic behind the simulation. (i.e., NO CHANGE in INPUT then NO CHANGE in OUTPUT)

- For the optimization with Stochastic simulation model, traditional DOE techniques and RSM methodology may be applied.

- For the optimization with Deterministic simulation model, those are not available since it does not include randomness in response.
Metamodeling of computer experiments

- Design of computer experiments: Space-filling methods
  - Latin Hypercube Design
  - Sphere-Packing Design
  - Uniform Design
  - Maximum Entropy Design
- Modeling of computer experiments
  - Kriging – Gaussian Process Model
Kriging – a metamodeling method

- **Kriging** (Gaussian Process Model) is a representative metamodeling technique originated from Geostatistics. (D.G. Krige, South African geologist)

Original elevation data and sample points: 300 randomly placed points where elevation data was sampled.

Predicted elevation with 300 randomly placed points where elevation data was sampled.

(Source: http://casoilresource.lawr.ucdavis.edu/drupal/node/442)
Kriging Theory

- Assumption:
  - **Response function** is composed of a regression model and stochastic process.

\[
Y(x) = f(x)^T \beta + Z(x)
\]

where \( f(x) = [f_1(x), f_2(x), \ldots, f_p(x)]^T \) : \((p \times 1)\) vector of regression functions

\[
\beta = [\beta_1, \beta_2, \ldots, \beta_p]^T \) : \((p \times 1)\) vector of unknown coefficients
Kriging Theory – continued..

- In many cases, \( f(x)^T \beta \) is assumed to be a constant (simple or ordinary kriging), or it may be first or second order polynomial.
- \( Z(x) \) is assumed to be a Gaussian process with zero mean, \( E[Z(x)]=0 \), and covariance, \( \text{Cov}[Z(x^i), Z(x^j)] \).
- While \( f(x)^T \beta \) globally approximates the design space, \( Z(x) \) creates “localized” deviations so that the kriging model interpolate \( n \) sampled data.
- The covariance matrix of the samples is
  \[
  \text{Cov}[Z(s^i), Z(s^j)] = \sigma^2 R([R(s^i, s^j)])
  \]
  where
  \( R \) is an \((n_s \times n_s)\) correlation matrix with ones along the diagonal
  \([R(s^i, s^j)]\) is correlation function
Kriging Theory – correlation function

- Correlation function, $R(s^i, s^j)$, in the correlation matrix, $R$, is user defined function and a variety of correlation function exist.
- Gaussian correlation function (the mostly preferred correlation function) is

$$R(s^i, s^j) = \exp\left[-\sum_{k=1}^{n} \theta_k \left| s^i_k - s^j_k \right|^2 \right]$$

where

- $\theta_k$ : unknown correlation parameters
- $n$ : the number of design variables
- $s^i_k$ and $s^j_k$ : $k^{th}$ components of the samples, $s^i$ and $s^j$
Kriging Theory – correlation function

- Various correlation functions

<table>
<thead>
<tr>
<th>Name</th>
<th>$R(\theta_k, d_{ij}^k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$e^{-\theta_k</td>
</tr>
<tr>
<td>Exponential-Gauss</td>
<td>$e^{-\theta_k</td>
</tr>
<tr>
<td>Gauss</td>
<td>$e^{-\theta_k d_{ij}^2}$</td>
</tr>
<tr>
<td>Linear</td>
<td>$\max{0, 1 - \theta_k</td>
</tr>
<tr>
<td>Spherical</td>
<td>$1 - 1.5\xi_{ij}^k + 0.5\xi_{ij}^{k^3}$, $\xi_{ij}^k = \min{1, \theta_k</td>
</tr>
<tr>
<td>Cubic</td>
<td>$1 - 3\xi_{ij}^{k^2} + 2\xi_{ij}^{k^3}$, $\xi_{ij}^k = \min{1, \theta_k</td>
</tr>
</tbody>
</table>

*where*, $d_{ij}^k = (s_i^k - s_j^k)$
Kriging Theory – correlation function

- Various correlation functions for $0 \leq d_{ij}^k \leq 2$
- Dash, full, and dashed-dotted line for $\theta_k = 0.2, 1, 5$
Kriging for Prediction

- Predicted estimates at the untried point $x$ is

$$
\hat{Y}(x) = f(x)^T \hat{\beta} + r(x)^T R^{-1} (Y - F\hat{\beta})
$$

where

$$
\hat{\beta} = (F^T R^{-1} F)^{-1} F^T R^{-1} Y
$$

Here,

- $R$: correlation matrix

- $F = [f(s_1)^T, f(s_2)^T, \ldots, f(s_{ns})]^T$

- $r(x) = [R(x, s_1), R(x, s_2), \ldots, R(x, s_{ns})]$

- $Y = [y_{s_1}, y_{s_2}, \ldots, y_{s_{ns}}]^T$
Predicted estimates at the untried point $x$ is

$$\hat{Y}(x) = f(x)^T \hat{\beta} + r(x)^T R^{-1} (Y - F\hat{\beta})$$
Kriging Theory – correlation parameter

- How to determine the correlation parameters, $\theta = [\theta_1, \theta_2, ..., \theta_n]$?
- Maximizing Maximum Likelihood Estimator (MLE) to find the optimum correlation parameters.

$$\text{maximize } \theta \left[-\frac{n_s \ln \hat{\sigma}_z^2 + \ln |R|}{2}\right]$$

\[ \text{where} \]
\[ \hat{\sigma}_z^2 \text{(estimated process variance)} = (Y - F\hat{\beta})^T R^{-1} (Y - F\hat{\beta}) / n_s \]
Kriging – different types of Kriging

- **Simple kriging** assumes a known constant trend.
- **Ordinary kriging** assumes an unknown constant trend.
- **Universal kriging** assumes a general polynomial trend model, such as a linear trend model.

\[
\hat{Y}(x) = f(x)^T \hat{\beta} + r(x)^T R^{-1} (Y - F\hat{\beta})
\]
Kriging Example

Ordinary Kriging Example

To clearly understand kriging, we illustrate the process of kriging modeling for 2-dimensional problem as follows:

\[ y = x_1 \sin x_2, \quad x_1, x_2 \in [1,8] \]  

(41)

As shown in Fig. 6, test function is slightly nonlinear in only one dimension.

Fig. 6 Sample points and original function of test function
Location of the nine sample points and the corresponding responses is given in Table 2. Sample points and responses are normalized by each means and standard deviations before using the data. This is helpful to alleviate the dimension effect of each design variable and prevent an erratic prediction of the kriging model.

<table>
<thead>
<tr>
<th>No</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$N(x_1)^*$</th>
<th>$N(x_2)^*$</th>
<th>$Y$</th>
<th>$N(Y)^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1.1547</td>
<td>-1.1547</td>
<td>0.8415</td>
<td>-0.0855</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4.5</td>
<td>-1.1547</td>
<td>0</td>
<td>-0.9775</td>
<td>-0.4401</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>8</td>
<td>-1.11547</td>
<td>1.1547</td>
<td>0.9894</td>
<td>-0.0566</td>
</tr>
<tr>
<td>4</td>
<td>4.5</td>
<td>1</td>
<td>0</td>
<td>-1.11547</td>
<td>3.7866</td>
<td>0.4886</td>
</tr>
<tr>
<td>5</td>
<td>4.5</td>
<td>4.5</td>
<td>0</td>
<td>0</td>
<td>-4.3989</td>
<td>-1.1070</td>
</tr>
<tr>
<td>6</td>
<td>4.5</td>
<td>8</td>
<td>0</td>
<td>1.1547</td>
<td>4.4521</td>
<td>0.6184</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>1</td>
<td>1.1547</td>
<td>-1.1547</td>
<td>6.7318</td>
<td>1.0628</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>4.5</td>
<td>1.1547</td>
<td>0</td>
<td>-7.8202</td>
<td>-1.7740</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>8</td>
<td>1.1547</td>
<td>1.1547</td>
<td>7.9149</td>
<td>1.2934</td>
</tr>
</tbody>
</table>

*Note: $N(\cdot)=$ normalized($\cdot$)*
Kriging Example

In order to fit a kriging model, we first consider maximum likelihood estimation. Two terms, \( \sigma_z^2 \) and \( \mathbf{R} \), that consist of maximum likelihood function, must be constructed. First, correlation matrix with Gaussian correlation function is computed as follows:

\[
\mathbf{R} = \begin{bmatrix}
1 & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} \\
1 & 1 & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} \\
1 & 1 & 1 & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} \\
1 & 1 & 1 & 1 & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} \\
1 & 1 & 1 & 1 & 1 & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} \\
1 & 1 & 1 & 1 & 1 & 1 & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & e^{-1.3336} & e^{-1.3336} & e^{-1.3336} \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & e^{-1.3336} & e^{-1.3336} \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & e^{-1.3336} \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}
\]  

(42)

Note that the distance is calculated by the normalized data. If we select a constant underlying global model, \( \mathbf{F} \) becomes simply a column vector of ones:

\[
\mathbf{F} = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]^T
\]

(43)

The estimation of global model is now constructed as follows:

\[
\hat{\mathbf{\beta}} = (\mathbf{F}^T \mathbf{R}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{R}^{-1} \mathbf{Y}
\]

(44)

Note that \( \mathbf{Y} \) is also normalized response. Estimation of process variance is given as
Kriging Example

\[ \hat{\sigma}_z^2 = \frac{(Y - F\hat{\beta})^T R^{-1} (Y - F\hat{\beta})}{9} \] (45)

Maximum likelihood estimation for \( \theta_1 \) and \( \theta_2 \) is found by the following expression:

\[ \min_{\theta_1, \theta_2} \hat{\sigma}_z^2 |R|^{1/9} \] (46)

Fig. 7 A plot of likelihood function
Kriging Example

Note that the maximum likelihood function is only a function of $\theta_1$ and $\theta_2$. When optimization range of correlation parameters is given in $\theta_1, \theta_2 \in [0.1, 5]$, the “best” kriging model to fit these sample points is obtained as $\theta_1^* = 0.1$ and $\theta_2^* = 5$. These values are the lower-limit and upper-limit, respectively. A plot of maximum likelihood function is given in Fig. 7. As shown in Fig. 7, you can find that the optimum solution exists on the lower left of the plot.

Now, we can predict the response at an untried point within design range. We can substitute the optimum solution into correlation matrix as follows:

$$
\begin{bmatrix}
1 & 1.273 \times 10^{-3} & 2.623 \times 10^{-12} & 8.752 \times 10^{-4} & 1.114 \times 10^{-3} & 2.296 \times 10^{-13} & 5.867 \times 10^{-1} & 7.466 \times 10^{-4} & 1.539 \times 10^{-12} \\
1.273 \times 10^{-3} & 1 & 1.114 \times 10^{-3} & 8.752 \times 10^{-4} & 1.114 \times 10^{-3} & 7.466 \times 10^{-4} & 5.867 \times 10^{-1} & 7.466 \times 10^{-12} \\
2.296 \times 10^{-12} & 1.114 \times 10^{-3} & 1 & 8.752 \times 10^{-4} & 1.114 \times 10^{-3} & 7.466 \times 10^{-4} & 5.867 \times 10^{-1} & 7.466 \times 10^{-12} \\
8.752 \times 10^{-4} & 8.752 \times 10^{-4} & 8.752 \times 10^{-4} & 1 & 1.114 \times 10^{-3} & 8.752 \times 10^{-1} & 1.114 \times 10^{-3} & 2.296 \times 10^{-12} \\
1.114 \times 10^{-3} & 1.114 \times 10^{-3} & 1.114 \times 10^{-3} & 1.114 \times 10^{-3} & 1 & 1.114 \times 10^{-3} & 8.752 \times 10^{-4} & 1.273 \times 10^{-12} \\
2.296 \times 10^{-13} & 7.466 \times 10^{-4} & 7.466 \times 10^{-4} & 8.752 \times 10^{-4} & 1.114 \times 10^{-3} & 1 & 1.114 \times 10^{-3} & 1.273 \times 10^{-12} \\
5.867 \times 10^{-1} & 5.867 \times 10^{-1} & 5.867 \times 10^{-1} & 8.752 \times 10^{-4} & 8.752 \times 10^{-4} & 1.114 \times 10^{-3} & 1 & 1.273 \times 10^{-12} \\
7.466 \times 10^{-4} & 7.466 \times 10^{-4} & 7.466 \times 10^{-4} & 7.466 \times 10^{-4} & 7.466 \times 10^{-4} & 1.114 \times 10^{-3} & 1.114 \times 10^{-3} & 1 \\
1.539 \times 10^{-12} & 7.466 \times 10^{-12} & 7.466 \times 10^{-12} & 2.296 \times 10^{-12} & 2.296 \times 10^{-12} & 1.273 \times 10^{-12} & 1.273 \times 10^{-12} & 1
\end{bmatrix}
$$
We need to know the form of the correlation vector as follows:

\[ r(x) = \begin{bmatrix} R(x, s_1), R(x, s_2), R(x, s_3), R(x, s_4), \cdots \\ R(x, s_5), R(x, s_6), R(x, s_7), R(x, s_8), R(x, s_9) \end{bmatrix}^T \]  

(48)

It is important to note that the prediction point should be normalized. Correlation vector is also only function of \( x \).

Finally, we can obtain the predictor of kriging as follows:

\[ \hat{Y}(x)_{\text{normalized}} = \hat{\beta} + r(x)^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F} \hat{\beta}) \]  

(49)

Because the predictor is normalized, we need to transform the predicted value into real value as follows:

\[ \hat{Y}(x) = \sigma_y \hat{Y}(x)_{\text{normalized}} + \overline{Y} \]  

(50)

where \( \sigma_y \) and \( \overline{Y} \) are standard deviation and mean used to normalize responses.
Kriging Example

Fig. 8 shows that kriging predict the behavior of the true function accurately. In addition, predictor reveals that response along the axis of $x_2$ is more nonlinear than along the axis of $x_1$ because the estimation of correlation parameter $\theta^*_2 = 5 > \theta^*_1 = 0.1$.

Fig. 8 Contour plots of true function and kriging predictor
Latin Hypercube Design

- A space-filling sampling technique with random permutation (McKay, Conover, and Beckman, 1979)
- $X_1 = [2, 5, 1, 4, 3, 6]$, $X_2 = [4, 5, 6, 3, 2, 1]$

Random field with two variables (uniform distribution)

Random field with two variables (normal distribution)
Maximum Entropy Design

- Popular space-filling sampling technique (Shewry and Wynn 1987)
- Entropy: a measure of the amount of information contained in the distribution of a data set
- The maximum entropy is achieved by maximizing the determinant of $R(\theta)$

![A 10-run maximum entropy design](image)
References

• Kriging

• Latin Hypercube Design

• Maximum Entropy Design