Kriging and Alternatives in Computer Experiments

C. F. Jeff Wu
ISyE, Georgia Institute of Technology

• Use kriging to build meta-models in computer experiments, a brief review
• Numerical problems with kriging
• Alternatives to kriging:
  - Regularized kriging, Hybrid kriging
  - Overcomplete basis surrogate model (OBSM)
Why computer experiments?

- No need for expensive lab equipments and materials, less costly than physical experiments.

- Not affected by human and environmental factors.

- Study dangerous or infeasible physical experiments, such as ammunition detonation.
Some examples

- Mechanical: machining, assembling...
- Chemical & Biology: nanoparticle and polymer synthesis...
- Aerospace: aircraft design, dynamics...
- Computer Experiments/Simulations
Statistical Meta-Modeling of Computer Experiments

- Robustness, optimization
- Surrogate model (Kriging)
  - More FEA runs
  - Computer modeling (finite-element simulation)
  - Noise simulations, error propagation
  - Physical experiment or observations
Kriging Models

• Ordinary Kriging

\[ Y(x) = \mu + Z(x) \]

\[ Z(x) \sim N_n\left(0, \sigma^2 \varphi(h)\right) \equiv GP\left(0, \sigma^2 \varphi(h)\right) \]

• Correlation function

\( \varphi(0) = 1 \)

\( \varphi(h) = \varphi(-h), \) (symmetric function)

\( \varphi \) is a positive semi-definite function
Correlation function

• Matérn

\[ \phi(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left( 2\sqrt{\nu} \theta |h| \right)^\nu K_\nu \left( 2\sqrt{\nu} \theta |h| \right) \]

where \( K_\nu \) is the modified Bessel function of order \( \nu \)

\[ \nu \to \infty, \quad \phi(h) \to \exp(-\theta h^2) \]

• **Power exponential correlation**

\[ \phi(h) = \exp\left( -\theta |h|^q \right), \quad 0 < q \leq 2, \quad 0 < \theta \]

  – \( q = 2 \)  **Gaussian correlation function** (infinitely differentiate)
  – \( q = 1 \)  Ornstein-Uhlenbeck process (\( \nu = 1 \) in Matern)

• Linear, Cubic correlation
Kriging predictor

• Best Linear Unbiased Predictor (BLUP)

\[
\hat{y}(x) = \hat{\mu} + r(x)'R^{-1}(y - \hat{\mu}1),
\]

\[
r(x)' = (\varphi(x - x_1), \ldots, \varphi(x - x_n)),
\]

\[
\hat{\mu} = 1'R^{-1}y/1'R^{-1}1,
\]

\[
\hat{y}(x_i) = y_i \quad \text{an interpolating property*}
\]

*required for deterministic simulations
Recent work in kriging

• Calibration of computer model, Kriging with *calibration* parameters (Kennedy-O’Hagan, 2001), with *tuning* parameters (Santner et al., 2009)

• Computer simulations with different levels of accuracy (Kennedy-O’Hagan, 2000; Qian et al., 2006; Qian-Wu, 2008)

construction of nested space-filling (e.g., Latin hypercube) designs (Qian-Ai-Wu, 2009, various papers by Qian and others, 2009-date)
Recent work in kriging (cont.)

• Kriging for multiple outputs and functional response (Conti et al., 2009; Conti and O’Hagan, 2010)

• Treed Gaussian Process model (Gramacy and Lee, 2008).

• Kriging (i.e., GP model) with quantitative and qualitative factors (Qian-Wu-Wu, 2008, Han et al., 2009) construction of sliced space-filling (e.g., Latin hypercube) designs (Qian-Wu, 2009, Qian, 2010)
Maximum Likelihood Estimation

- Profile log-likelihood approach

\[ Q(\theta) = n \log(\sigma^2(\theta)) + \log|R(\theta)| \]

where \( \sigma^2(\theta) = \frac{\{y - \hat{\mu}(\theta)1\}'R^{-1}(\theta)\{y - \hat{\mu}(\theta)1\}}{n} \)

\[ \hat{\mu}(\theta) = \frac{1'R^{-1}(\theta)y}{1'R^{-1}(\theta)1} \]
Numerical Instability in $R^{-1}(\theta)$

- $R(\theta)$ is an $nxn$ matrix, $n=$sample size
- Its condition number (max e.v./min e.v.) $\uparrow$ as
  
  I. Sample size $n$ $\uparrow$
  
  II. Dimension of input vectors $\uparrow$

(Peng-Wu, 2010)
Branin function

(Andre, Siarry and Dognon, 2001)

$$f(x_1, x_2) = \left( x_2 - \frac{5.1}{4\pi^2} x_1^2 + \frac{5}{\pi} x_1 - 6 \right)^2 + 10 \left( 1 - \frac{1}{8\pi} \right) \cos(x_1) + 10$$
Log-likelihood function
(Regular grid: $n = 7^2$)
Regularized Kriging

- Introducing a regularizing constant $\lambda$ into the predictor

$$\hat{y}_\lambda(x) = \hat{\mu}_\lambda + r(x)'(R + \lambda I)^{-1}(y - \hat{\mu}_\lambda 1)$$

where $\hat{\mu}_\lambda = 1'(R + \lambda I)^{-1}y/1'(R + \lambda I)^{-1}1$

Peng and Wu (2010, submitted)

- Similar modification in estimation: maximizing a regularized likelihood
Kriging with nugget effects

• Model from spatial statistics

\[ Y(x) = \mu + Z(x) + \delta \varepsilon \]

• BLUP

\[ \hat{y}_\delta(x) = \hat{\mu}_\delta + r(x)' \left( R + \frac{\delta^2}{\sigma^2} I \right)^{-1} (y - \hat{\mu}_\delta 1) \]

\[ \hat{\mu}_\delta = \frac{1' \left( R + \frac{\delta^2}{\sigma^2} I \right)^{-1} y}{1' \left( R + \frac{\delta^2}{\sigma^2} I \right)^{-1} 1} \]
Algorithm (Ridge Trace)

• **Root Mean Squared Prediction Error (RMSPE)**

\[
\text{RMSPE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y(x_i) - \hat{Y}(x_i))^2}
\]

1. Set \( \lambda^* \) as the lower bound and choose a grid point set for \( \lambda \), say, \( (\lambda_1, \ldots, \lambda_k) \), and let \( i = 1 \).

2. Use \( \lambda_i \) in regularized kriging to estimate \( \theta_{\lambda} \).

3. Compute the RMSPE. Let \( i = i + 1 \).

4. Repeat steps 2 and 3 until all \( k \) grid points are exhausted.

5. The final estimator \( \hat{\theta}_{\hat{\lambda}} \) is the one with the lowest RMSPE with \( \hat{\lambda} \).
Log-likelihood function, Branin function, (Regular grid: $n = 7^2$)
\[
\lambda^* = \frac{n^{1/2}}{\Delta^{1/2} - 1} = 3.3 \times 10^{-7}
\]

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>RMSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-1}$</td>
<td>6.1842</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>3.2085</td>
</tr>
<tr>
<td>$10^{-3}$</td>
<td>1.5605</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>1.0706</td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td>1.7069</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>2.5104</td>
</tr>
<tr>
<td>$10^{-7}$</td>
<td>3.2348</td>
</tr>
</tbody>
</table>
Overcomplete Basis Surrogate Model

- Use an overcomplete dictionary of basis functions
- Use linear combinations of basis functions to approximate unknown functions
- Use Matching Pursuit for fast (i.e. greedy) computations
- Choice of basis functions to “mimic” the shape of the surface

Chen, Wang, and Wu (2010, *IIE Tran. Q&R*)
Surrogate Representation

• Surrogate model: use a linear combination of pre-specified basis functions, i.e.,

\[ f(\mathbf{x}) = \sum_j c_j \phi_j(\mathbf{x}), \mathbf{x} \in \chi \]

- unitary norm \( \| \phi_j \| = 1 \)
- basis dictionary, \( \{ \phi_j, j = 1, \ldots, M \} \)
- no unknown parameter in \( \phi_j \), only unknown
  are the linear \( c_j \)

• Overcomplete: \( M \) much larger than data size
Surrogate Model (continued)

• Explored point set: \( P_{\text{exp}} = \{x_1, ..., x_p\} \).

• Current responses:
  \[
  V_{P_{\text{exp}}} = \left( f(x_1), ..., f(x_p) \right)^T
  \]

• Use
  \[
  \sum_j c_j \tilde{\phi}_j
  \]
  to approximate \( V_{P_{\text{exp}}} \),
  \[
  \tilde{\phi}_j = \left( \phi_j(x_1), ..., \phi_j(x_p) \right)^T
  \]

• Two interesting questions:
  • *Choice of* the basis functions?
  • Estimation of the linear coefficients \( C_i \)?
Coefficient Inference

- Matching Pursuit Algorithm (Mallat and Zhang, 1993):
  - Infer coefficients by minimizing $\| V_{p_{\text{exp}}} - \sum_j c_j \phi_j \|$.
  - A greedy algorithm: at the $j$th iteration,
    
    Let $R^{(j-1)}$ be the current residual vector.
    
    Selected a basis by $\tilde{\phi}_{(j)} = \arg \max_i \langle R^{(j-1)}, \phi_i \rangle$:
    
    $c_{(j)} = c_{(j)} + \langle R^{(j-1)}, \tilde{\phi}_{(j)} \rangle$,
    
    $R^{(j)} = R^{(j-1)} - \langle R^{(j-1)}, \tilde{\phi}_{(j)} \rangle \tilde{\phi}_{(j)}$. 
Response Surface for Bistable Laser Diodes

• The true surface over a pre-specified grid:

• Search all positive Lyapunov exponents (PLE) (red area)
• PLE corresponds to chaotic light output.
Gabor Functions

- Basis functions:
  - $n$-dimensional Gabor function
    \[ g(x) = \exp\left(-\frac{x^T M x}{2}\right) \exp(2\pi i A x), \quad x = (x_1, \ldots, x_n)^T \]
  - Two-dimensional Gabor function, i.e. $n = 2$
    \[ g(u, v) = \frac{1}{Z} \exp \left[-\frac{1}{2} (\sigma_u u^2 + \sigma_v v^2)\right] \cos \left[\frac{2\pi u}{\lambda} + \varphi\right], \]
    \[ u = u_0 + x_1 \cos \theta - x_2 \sin \theta \]
    \[ v = v_0 + x_1 \sin \theta - x_2 \cos \theta \]
Plots of 2-D Gabor Function
Overall Comparisons

Figure 6: Cumulative numbers of PLEs found by using different explored points in $21 \times 21$, $81 \times 81$, and $161 \times 161$ grids.
Summary

• Computer experiments/simulations have become popular in engineering and science

• Kriging is the most common method for statistical meta-model building but is more limited for large or complex problems

• Alternatives to kriging are being sought:
  – Tweaking of kriging to achieve stability (regularized, hybrid, tapering, reduced rank)
  – Approximations with fast computations (OBSM, RIDW): but lacking inferential capability
Covariance Matrix Tapering

- Covariance tapering (Kaufman et al., 2008)
  - Covariance matrix is “tapered” or multiplied elementwise by a sparse matrix, to approximate the likelihood.

- Advantages:
  - Significant computational gains/stability.
  - Retain interpolating property.
  - Asymptotic convergence of the tamper estimator.

- But:
  - The tapering function is isotropic: OK for spatial statistic problems, but not applicable to engineering problems.
  - The tapering radius needs to be determined.
Rank Reduction

• Fixed rank kriging (Cressie-Johannesson, 2008)
  ✓ A flexible family of non-stationary covariance function is defined by using a set of basis functions that are fixed in number (smaller than the data size $n$).

• Advantage:
  ✓ Reduce the computational cost of kriging to $O(n)$.

• But:
  ✓ How to choose the appropriate basis functions.
  ✓ Not an interpolator.
Upper bound

- Upper bound

\[
\kappa_2^p \left( R(\theta) + \lambda I_n; X \right) \leq \prod_{k=1}^{p} \kappa_2^1 \left( R(0) + \lambda I_{n_k}; D_k \right)
\]

- The worst case of a correlation matrix
Inverse Distance Weighting (IDW)

- Inverse Distance Weighting (Shepard, 1968):

\[
\hat{y}(\mathbf{x}) = \frac{\sum_{k=1}^{n} w_k(\mathbf{x})y_k}{\sum_{i=1}^{n} w_i(\mathbf{x})}.
\]

- \( w_i(\mathbf{x}) = 1/d(\mathbf{x}, \mathbf{x}_i)^2 \).
- \( d(\mathbf{x}, \mathbf{x}_i) = \left\{ \sum_{j=1}^{p} (x_j - x_{i,j})^2 \right\}^{1/2} \).
- Simple computation but poor prediction.
Regression-Based Inverse Distance Weighting (RIDW)

- Add regression part to IDW (Joseph and Kang, 2009):

\[
\hat{y}(\mathbf{x}) = \mu(\mathbf{x}; \beta) + \frac{\sum_{k=1}^{n} w_k(\mathbf{x}) e_k}{\sum_{i=1}^{n} w_i(\mathbf{x})}
\]

- \( \mu(\mathbf{x}_k; \beta) \) = mean part: can be linear, nonlinear, nonparametric.

- \( e_k = y_k - \mu(\mathbf{x}_k; \beta) = y_k - \mu_k \).

- \( w_i(\mathbf{x}) = \exp\left\{-\frac{d^2(\mathbf{x}, \mathbf{x}_i)}{2\theta^2}\right\} \). (faster convergence than IDW)

- \( d(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{p} \theta_j (x_j - x_{i,j})^2} \).
Comparisons Between RIDW and Kriging

Standardized RMSPE

CPU time in simulation
Lower bound on $\lambda$

- Lower bound

$$\lambda^* = \inf \left\{ \lambda \middle| \prod_{k=1}^{p} (1 + n_k/\lambda) < \Delta \right\}$$

where

$$\epsilon = 2^{-52} \approx 2.22 \times 10^{-16}$$

$$\Delta = 1/(10\epsilon) = 4.5 \times 10^{14}$$

Machine accuracy (or unit round-off)