Kriging Surrogate Model
Kriging and cost of surrogates

- In linear regression, the process of fitting involves solving a set of linear equations once.

- Moving least squares performs the fit for each function evaluation, using only nearby points.

- Radial basis surrogates use shape functions that are based around data points and decay away from them, so that nearby data have more influence on prediction.

- Kriging, is even more expensive, we have a spread constant in every direction and we have to perform optimization to calculate the best set of constants (hyperparameters).

  - With many hundreds of data points this can become significant computational burden.
Introduction to Kriging

• Method invented in the 1950s by South African geologist Daniel G. Krige (1919-2013) for predicting distribution of minerals.
  – Statisticians refer to a more general Gaussian Process regression.

• Became very popular for fitting surrogates to expensive computer simulations in the 21st century.

• It is one of the best surrogates available.

• It probably became popular late mostly because of the high computer cost of fitting it to data.
• We assume that the data is sampled from an unknown function that obeys simple correlation rules.

• The value of the function at a point is correlated to the values at neighboring points based on their separation in different directions.

• The correlation is strong to nearby points and weak with far away points, but strength does not change based on location, only separation between points.

• Normally Kriging is used for noise free data so that it interpolates exactly the function values.
Reminder: Covariance and Correlation

- Covariance of two random variables $X$ and $Y$
  \[
  \text{cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X\mu_Y
  \]

- The covariance of a random variable with itself is the square of the standard deviation. $\text{Var}(X) = [\sigma(X)]^2$

- Covariance matrix
  \[
  \Sigma_{XY} = \begin{bmatrix}
  \text{Var}(X) & \text{cov}(X,Y) \\
  \text{cov}(X,Y) & \text{Var}(Y)
  \end{bmatrix}
  \]

- Correlation
  \[
  \text{cor}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y} \quad -1 \leq \text{cor}(X,Y) \leq 1
  \]

- The correlation matrix has 1 on the diagonal.
• Generate 10 random samples, translate them by a bit (0.1), and by more (1.0)

\[
x=10*\text{rand}(1,10); \\
x_{\text{near}}=x+0.1; \ x_{\text{far}}=x+1; \\
y_{\text{near}}=\sin(x_{\text{near}}); \\
y=\sin(x); \\
y_{\text{far}}=\sin(x_{\text{far}});
\]

• Compare corelations:

\[
r=\text{corrcoef}(y,y_{\text{near}}) \quad 0.9894; \quad \text{High correlation} \\
r_{\text{far}}=\text{corrcoef}(y,y_{\text{far}}) \quad 0.4229; \quad \text{Low correlation}
\]

• Decay to about 0.4 over one sixth of the wavelength.

– Wavelength on sine function is \(2\pi \sim 6\)
Universal Kriging approximation

- Kriging is similar to RBF, but starting from statistical view

\[ \hat{y}(x) = \sum_{i=1}^{n_B} \beta_i \xi_i(x) + z(x) \]

Systematic departure (random process)

Trend function

Global function (low-order polynomials)

Mean of Kriging prediction

Sampling data points

Systematic Departure

Trend Model

\( y \)

\( x \)
Ordinary and simple Kriging

• Ordinary Kriging: constant trend function

• Simple Kriging: constant trend function is known (often 0)

• Assumption: Systematic departures $z(x)$ are correlated.

• Kriging prediction comes with a normal distribution of the uncertainty in the prediction

• At the sample points, the uncertainty is zero
• Kriging assumes that predictions are correlated inversely proportional to the distance

• Systematic departure captures this correlation
  – Zero mean: $E[z(x)] = 0$
  – Covariance of data: $\text{cov}[z(x^{(i)}), z(x^{(j)})] = \sigma^2 \phi(\theta, x^{(i)}, x^{(j)})$
  – Variance of function: $\sigma^2 = \text{cov}[z(x), z(x)]$

• Isotropic correlation: $\phi(\theta, x^{(i)}, x^{(j)}) = \prod_{k=1}^{n} \phi(\theta, |x_k^{(i)} - x_k^{(j)}|)$

• Anisotropic correlation: $\phi(\theta, x^{(i)}, x^{(j)}) = \prod_{k=1}^{n} \phi_j(\theta_j, |x_k^{(i)} - x_k^{(j)}|)$
Isotropic vs. anisotropic correlation functions

(a) Isotropic correlation

(b) Anisotropic correlation
Gaussian correlation function

- Correlation between point x and point s

\[ C(z(x), z(s), \theta) = \prod_{k=1}^{n} \exp \left[ - \left( \frac{x_k - s_k}{\theta_k} \right)^2 \right] \]

- \( \theta_k \): hyperparameter, decaying rate

- The correlation should decay to about 0.4 at one sixth of the wavelength \( l_i \) and \( e^{-1} = 0.37 \approx 0.4 \).

- Approximately \((l_i/6\theta_k)^2 = 1\) or \( \theta_i = l_i/6 \)

- For the function \( y = \sin(x_1) \times \sin(5x_2) \) we would like to estimate \( \theta_1 \approx 1, \theta_2 \approx 0.2 \)
\( n_y \) sample points \((x^{(i)}, y_i)\), with \( n \)-dimension of input \( x_k^{(i)}, k = 1, \ldots, n \) and \( y_i = y(x^{(i)})\)

- Given decay rates \( \theta_k \), the covariance matrix of the data

\[
\operatorname{cov}(y_i, y_j) = \sigma^2 \exp\left[-\sum_{k=1}^{n} \left(\frac{x_k^{(i)} - x_k^{(j)}}{\theta_k}\right)^2\right] = \sigma^2 R_{ij}, \quad i,j = 1,\ldots,n_y
\]

- The **correlation matrix** \( R \) is formed from the covariance matrix, assuming a constant standard deviation \( \sigma \), which measures the uncertainty in function values (stationary covariance)

- Small \( \sigma \) for dense data, large \( \sigma \) for sparse data
  - How do you decide whether the data is sparse or dense?
**Kriging vs. PRS**

- **Kriging**
  \[
  \hat{y}(x) = \sum_{i=1}^{n_\beta} \beta_i \xi_i(x) + z(x)
  \]

- **PRS**
  \[
  \hat{y}(x) = \sum_{i=1}^{n_\beta} \beta_i \xi_i(x) + \varepsilon(x)
  \]

- **PRS** assumes that \( \hat{y} = \sum \beta_i \xi_i(x) \) is a correct form, but data have error \( \varepsilon \sim N(0, \sigma^2) \) that are statistically independent.

- **Kriging** assumes that data are accurate, but the model form is uncertain \( \rightarrow \) Kriging fits data

  \[
  y_k = \hat{y}(x_k) = \sum_{i=1}^{n_\beta} \beta_i \xi_i(x_k) + z(x_k)
  \]

- At prediction points, error in Kriging is described by local departure \( z(x) \sim N(0, \sigma^2) \).
Determining the global function

- Global function coefficients, $\mathbf{\beta}$, and variance of data, $\sigma^2$
- Error b/w data and global function

$$ e = y - X\beta = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_y} \end{pmatrix} - \begin{pmatrix} -\xi(x_1) - \beta_1 \\ -\xi(x_2) - \beta_2 \\ \vdots \\ -\xi(x_{n_y}) - \beta_{n_p} \end{pmatrix} $$

- Assumption: error $e \sim N(0, \sigma^2)$ and correlation between data
- Maximum likelihood estimate (MSE)
  - Likelihood: PDF of getting data $y$ for given parameters, $\mathbf{\beta}$ and $\sigma^2$

$$ f(y | \theta, \mathbf{\beta}, \sigma^2) = \frac{1}{\sqrt{(2\pi)^{n_y} (\sigma^2)^{n_y} | R |}} \exp \left( -\frac{(y - X\beta)^T R^{-1} (y - X\beta)}{2\sigma^2} \right) $$
Maximum likelihood estimate (MLE)

- Logarithmic likelihood (ignore $\theta$ for now)
\[
\ln \left[ f \left( y \mid \beta, \sigma^2 \right) \right] = -\frac{n_y}{2} \ln(2\pi) - \frac{n_y}{2} \ln(\sigma^2) - \frac{1}{2} \ln |R| - \frac{(y - X\beta)^T R^{-1}(y - X\beta)}{2\sigma^2}
\]

- Stationary condition
\[
\frac{\partial \ln f}{\partial \beta} = \frac{X^T R^{-1}(y - X\beta)}{\sigma^2} = 0
\]
\[
\frac{\partial \ln f}{\partial \sigma^2} = -\frac{n_y}{2} \frac{1}{\sigma^2} + \frac{(y - X\beta)^T R^{-1}(y - X\beta)}{2\sigma^4} = 0
\]

- Solve for $\beta$ and $\sigma^2$
\[
\hat{\beta} = \left( X^T R^{-1}X \right)^{-1} \left\{ X^T R^{-1} y \right\}
\]
\[
\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^T R^{-1}(y - X\hat{\beta})}{n_y - n_{\beta}}
\]

PRS linear regression
\[
b = (X^T X)^{-1} X^T y
\]
\[
\hat{\sigma}^2 = \frac{SSE}{n_y - n_{\beta}}
\]
For unbiased estimate
Local departure

- Kriging passes data point \(\rightarrow\) Kriging can be expressed by a linear combination of data and weights: \(\hat{y}(x) = w(x)^T y\)

- Minimizing mean squared error (MSE)

\[
\varepsilon(x) = \hat{y}(x) - y(x) = w(x)^T y - y(x)
\]

- Data: \(y = X\hat{\beta} + z\)

- True function: \(y(x) = \xi(x)\hat{\beta} + z(x)\)

\[
\varepsilon(x) = w(x)^T \{X\hat{\beta} + z\} - \left(\xi(x)\hat{\beta} + z(x)\right)
\]

\[
= \left(w(x)^T X - \xi(x)\right)\hat{\beta} + w(x)^T z - z(x)
\]

  - Global error
  - Departure error
• To keep the global function unbiased, a constraint of global error being zero is added

\[ w(\mathbf{x})^T \mathbf{X} - \xi(\mathbf{x}) = 0 \]

• MSE = \( E[\varepsilon(\mathbf{x})^2] = E[(\mathbf{w}^T \mathbf{z} - z)^2] = E[\mathbf{w}^T \mathbf{z} \mathbf{z}^T \mathbf{w} - 2\mathbf{w}^T \mathbf{z} \mathbf{z} + z^2] \)

\[ \text{MSE} = \sigma^2 \left( \mathbf{w}^T \mathbf{R} \mathbf{w} - 2\mathbf{w}^T \mathbf{r} + 1 \right) \]

\( \text{cov}[\mathbf{z}(\mathbf{x}), \mathbf{z}(\mathbf{x})] = \sigma^2 \)
\( \text{cov}[\mathbf{z}(\mathbf{x}_k), \mathbf{z}(\mathbf{x})] = \sigma^2 \mathbf{r} \)
\( \text{cov}[\mathbf{z}(\mathbf{x}_k), \mathbf{z}(\mathbf{x}_l)] = \sigma^2 \mathbf{R} \)
\( \mathbf{r}(\mathbf{x}) = [R(\mathbf{x}_k, \mathbf{x})] \)

• MSE is the variance (uncertainty) in Kriging prediction

• Goal: find \( \mathbf{w}(\mathbf{x}) \) that minimizes MSE while satisfying the unbiased constraint
Lagrange function for constrained optimization

- Lagrange function (min. MSE with constraint)
  \[ L(w, \lambda) = \sigma^2 (w^T R w - 2w^T r + 1) - \lambda (X^T w - \xi^T) \]

- Stationary conditions (KKT)
  \[
  \begin{align*}
  \frac{\partial L(w, \lambda)}{\partial w} &= 2\sigma^2 (R w - r) - X \lambda^T = 0 \\
  \frac{\partial L(w, \lambda)}{\partial \lambda} &= X^T w - \xi^T = 0 \\
  w &= R^{-1} r + R^{-1} X \frac{\lambda^T}{2\sigma^2} \\
  \frac{\lambda^T}{2\sigma^2} &= (X^T R^{-1} X)^{-1} \left\{ \xi^T - X^T R^{-1} r \right\}
  \end{align*}
  \]

  \[ \hat{y}(x) = w(x)^T y \]
Kriging prediction

- Kriging as linear combination of data and weight functions

\[ \hat{y}(x) = w(x)^T y \]

\[ = \left( R^{-1}r + R^{-1}X \frac{\lambda^T}{2\sigma^2} \right)^T y \]

\[ = r^T R^{-1} y + \frac{\lambda}{2\sigma^2} X^T R^{-1} y \]

\[ = r^T R^{-1} y + \left( \xi - r^T R^{-1} X \right) \left( X^T R^{-1} X \right)^{-1} \left( X^T R^{-1} y \right) \]

\[ \hat{y}(x) = \xi(x) \hat{\beta} + r(x)^T R^{-1} (y - X \hat{\beta}) \]

Trend function  Local departure

- Local departure term is the weighted sum of the trend function error

\[ (y - X \hat{\beta}) \] based on the correlation term \[ r(x)^T R^{-1} \]
Simplification for ordinary Kriging

• For ordinary Kriging, $\mathbf{X} = [1]$ and $\hat{\beta} = \hat{\mu}$ (mean of data)

$$\hat{y}(x) = \hat{\mu} + \mathbf{r}(x)^T \mathbf{R}^{-1}(y - \mathbf{1}\hat{\mu}) = \hat{\mu} + \mathbf{b}^T \mathbf{r}(x)$$

• Linear in $\mathbf{r}(x)$ that the radial basis can be viewed as basis functions

$$r_i(x) = \exp \left[ - \sum_{k=1}^{n} \left( \frac{x_k^{(i)} - x_k}{\theta_k} \right)^2 \right]$$

• The prediction is linear in the data $\mathbf{y}$, in common with linear regression, but $\mathbf{b}$ is not calculated by minimizing MSE.

• Note that far away from data, $\hat{y}(x) \sim \hat{\mu}$ (not good for substantial extrapolation)
Estimating hyperparameters $\theta$

- Estimating $\hat{\beta}$ and $\hat{\sigma}$ depends on hyperparameter $\theta$ in $\mathbb{R}$

- Maximizing the log-likelihood that the data comes from a Gaussian process defined by $\theta_k$.

\[
\ln[f(y | \theta, \beta, \sigma^2)] = -\frac{n_y}{2} \ln(2\pi) - \frac{n_y}{2} \ln(\sigma^2) - \frac{1}{2} \ln|\mathbf{R}| - \frac{(y - \mathbf{X}\beta)^\top \mathbf{R}^{-1}(y - \mathbf{X}\beta)}{2\sigma^2}
\]

\[
\theta = \arg \max \left[ \ln[f(y | \theta, \beta, \sigma^2)] = -\frac{n_y}{2} \ln(\sigma^2) - \frac{1}{2} \ln|\mathbf{R}| \right]
\]

\[
\theta = \arg \min \left[ \ln\left(\hat{\sigma}^{2(n_y-n_\beta)} \times |\mathbf{R}|\right) \right] \quad \text{Equivalent}
\]

- Maximum likelihood is a tough optimization problem
  - the likelihood often varies slowly in a wide range of argument
  - Some Kriging codes minimize the cross-validation error instead
Once $\theta$ is found, the estimate of the mean and standard deviation is obtained as (ordinary Kriging)

\[
\hat{\mu} = \frac{1^T R^{-1} y}{1^T R^{-1} 1}, \quad \hat{\sigma}^2 = \frac{(y - 1\hat{\mu})^T R^{-1} (y - 1\hat{\mu})}{n_y - n_\beta}
\]
• Kriging prediction $\hat{y}(x) = \xi(x)\hat{\beta} + r(x)^{T}R^{-1}(y - X\hat{\beta})$ is the mean prediction and MSE is the variance

• Kriging prediction is Gaussian distribution

$$\hat{Y}(x) \sim N(\xi\hat{\beta} + r^{T}R^{-1}(y - X\hat{\beta}), \sigma^{2}(w^{T}Rw - 2w^{T}r + 1))$$

– This is an estimated uncertainty using data

– When the # of data is small, use t-distribution

$$\hat{Y}(x) \sim \xi\hat{\beta} + r^{T}R^{-1}(y - X\hat{\beta}) + t_{n_{y} - n_{\beta}} \cdot \hat{\sigma}\sqrt{w^{T}Rw - 2w^{T}r + 1}$$

• Ordinary Kriging

$$V[\hat{Y}(x)] = \sigma^{2}\left[1 - r^{T}R^{-1}r + \frac{(1 - 1^{T}R^{-1}r)^{2}}{1^{T}R^{-1}1}\right]$$
Prediction variance

- Square root of variance is called standard error.
- The uncertainty at any \( x \) is normally distributed.
- \( \hat{y}(x) \) represents the mean of Kriging prediction.
Kriging fitting issues

- MLE or cross-validation optimization problem solved to obtain the kriging fit is often ill-conditioned leading to poor fit, or poor estimate of the prediction variance.

- Poor estimate of the prediction variance can be checked by comparing it to the cross validation error.

- Poor fits are often characterized by the kriging surrogate having large curvature near data points.

- It is recommended to visualize by plotting the kriging fit and its standard error.
Comparison b/w RMSE and MLE

Good

Discontinuous

MLE ≠ min RMSE

More than 2 hyperparameters
Ex) Quadratic function fit

- Use 9 data and a constant global function $\xi(x) = 0$ to fit a quadratic function $y(x) = x^2 + 5x - 10$

- Covariance

$$cov(x_i, x_j) = \exp\left(-\left(\frac{x_i - x_j}{\theta}\right)^2\right)$$

Too large $\theta$
Good fit with poor variance

Too small $\theta$
Bad fit with poor variance
• Fit data $\mathbf{x} = \{0, 5, 10, 15, 20\}^T$, $\mathbf{y} = \{1, 0.99, 0.99, 0.94, 0.95\}^T$ for the global function using ordinary Kriging with $\theta = 5.2$

- For ordinary Kriging, $\mathbf{X} = [1, 1, 1, 1, 1]^T$ and $\xi(x) = [1]$

  ```matlab
  y=[1 0.99 0.99 0.94 0.95]'; % measurement data
  x=[0 5 10 15 20]'; % input variable
  X=ones(5,1); % design matrix
  ny=length(y); np=size(X,2);
  
  - Correlation matrix $\mathbf{R}$
    
    $h=5.2$;
    
    for k=1:ny; for l=1:ny;
    
    $R(k,l)=\exp\left(-\left(\frac{\text{norm}(x(k,:)-x(l,:))/h}{h}\right)^2\right)$;
    
    end; end;

  $$
  \mathbf{R} = \begin{bmatrix}
  1 & 0.3967 & 0.0248 & 0.0002 & 0 \\
  0.3967 & 1 & 0.3967 & 0.0248 & 0.0002 \\
  0.0248 & 0.3967 & 1 & 0.3967 & 0.0248 \\
  0.0002 & 0.0248 & 0.3967 & 1 & 0.3967 \\
  0 & 0.0002 & 0.0248 & 0.3967 & 1
  \end{bmatrix}
  $$
  $$
• Global function parameters
  \[
  R_{	ext{inv}} = \text{inv}(R);
  \theta_{\text{H}} = (X' \cdot R_{	ext{inv}} \cdot X) \backslash (X' \cdot R_{	ext{inv}} \cdot y);
  \sigma_{\text{H}} = \sqrt{\frac{1}{n_y - n_p} \cdot ((y - X \cdot \theta_{\text{H}})' \cdot R_{	ext{inv}} \cdot (y - X \cdot \theta_{\text{H}}))};
  \]
  \[
  \hat{\beta} = (X^T R^{-1} X)^{-1} \{X^T R^{-1} y\} = 3.0989^{-1} \times 3.0226 = 0.9754
  \]
  \[
  \hat{\sigma}^2 = \frac{(y - X \hat{\beta})^T R^{-1} (y - X \hat{\beta})}{n_y - n_\beta} = 7.28 \times 10^{-4}, \hat{\sigma} = 0.0270
  \]

• Estimate the optimum hyperparameter
  
  – Instead of optimization the hyperparameter, we calculate it graphically
  
  – \( \theta_{\text{opt}} = 5.2 \)
  
  – We used this value in calculating \( \hat{\beta} \) and \( \hat{\sigma}^2 \)
Ex) Kriging fit cont.

- Matlab code for the graph

```matlab
h=zeros(20,1); Obj=zeros(20,1);
for i=1:20
    h(i)=0.5*i;
    for k=1:ny; for l=1:ny;
        R(k,l)=exp(-(norm(x(k,:)-x(l,:))/h(i))^2);
    end; end;
    Rinv=inv(R);
    thetaH=(X'*Rinv*X)\(X'* Rinv*y);
    sigmaH=sqrt(1/(ny-np)*((y-X*thetaH)'*Rinv*(y-X*thetaH)));
    Obj(i)=log(sigmaH^(2*(ny-np))*det(R));
end
plot(h,Obj,'linewidth',2); grid on;
```

- Prediction at $x = 10$

\[
r = \{R(x_k,x)\} = \begin{bmatrix} 0.0248 & 0.3967 & 1 & 0.3967 & 0.0248 \end{bmatrix}^T
\]

\[
\hat{y}(10) = \xi \hat{\beta} + r^T R^{-1} (y - X \hat{\beta}) = 0.9754 + 0.0146 = 0.99
\]

- Exact at the sample point!
Ex) Kriging fit cont.

• Prediction at $x = 14$

\[ r = \begin{bmatrix} 0.0007 & 0.05 & 0.5534 & 0.9637 & 0.2641 \end{bmatrix}^T \]

\[ \hat{y}(14) = \xi \hat{\beta} + r^T R^{-1}(y - X \hat{\beta}) = 0.9754 - 0.0272 = 0.9482 \]

$x_{\text{New}} = 10; \ %$or $x_{\text{New}} = 14$

for $k = 1:ny;$

\[ r(k,1) = \exp\left(-\frac{\text{norm}(x(k,:)-x_{\text{New}})/h)^2}{2}\right); \]

end;

gpDepar = r' \times R^{-1} \times (y - X \times \theta H);$
• 90% confidence intervals
  
  – Standard error: \( s_y = \hat{\sigma} \sqrt{w^T R w - 2 w^T r + 1} \)
  
  \[
  \begin{align*}
  x_i &= 1; \\
  w &= R_{inv} \ast r + R_{inv} \ast X \ast((X' \ast R_{inv} \ast X) \backslash (x_i' - X' \ast R_{inv} \ast r)); \\
  z_{\text{Sigma}} &= \sigma_{H} \ast \sqrt{w' \ast R \ast w - 2 \ast w' \ast r + 1};
  \end{align*}
  \]

  % using the inverse calculation
  
  \[
  \begin{align*}
  \text{gpMean} &= 0.9482; \quad \% \text{ from Example 5.2} \\
  \text{PI} &= [ \text{gpMean} + \text{tinv}(0.05, ny-np) \ast z_{\text{Sigma}}, \ldots \\
  & \quad \text{gpMean} + \text{tinv}(0.95, ny-np) \ast z_{\text{Sigma}} ];
  \end{align*}
  \]

  % using the random samples
  
  \[
  \begin{align*}
  \text{ns} &= 5e3; \\
  \text{tDist} &= \text{trnd}(ny-np,1,ns); \\
  \text{yHat} &= \text{gpMean} + \text{tDist} \ast z_{\text{Sigma}}; \\
  \text{PI} &= \text{prctile(yHat, [5 95])}
  \end{align*}
  \]

  \[
  \begin{array}{|c|c|c|c|}
  \hline
  x_{new} & 5 \text{ percentile} & 95 \text{ percentile} & 90\% \text{ P.I.} \\
  \hline
  x = 10 & 0.99 & 0.99 & 0 \\
  \hline
  x = 14 & 0.9394 & 0.9570 & 0.0176 \\
  \hline
  \end{array}
  \]

  Ex) Kriging fit cont.
Kriging with nuggets

- Nuggets – refers to the inclusion of noise at data points.
- The more general Gaussian Process surrogates or Kriging with nuggets can handle data with noise (e.g. experimental results with noise).