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.
  - Formalized by French engineer, Georges Matheron in 1960.
  - 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.

Kriging philosophy

- 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*

$$\mathrm{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X \mu_Y$$

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

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

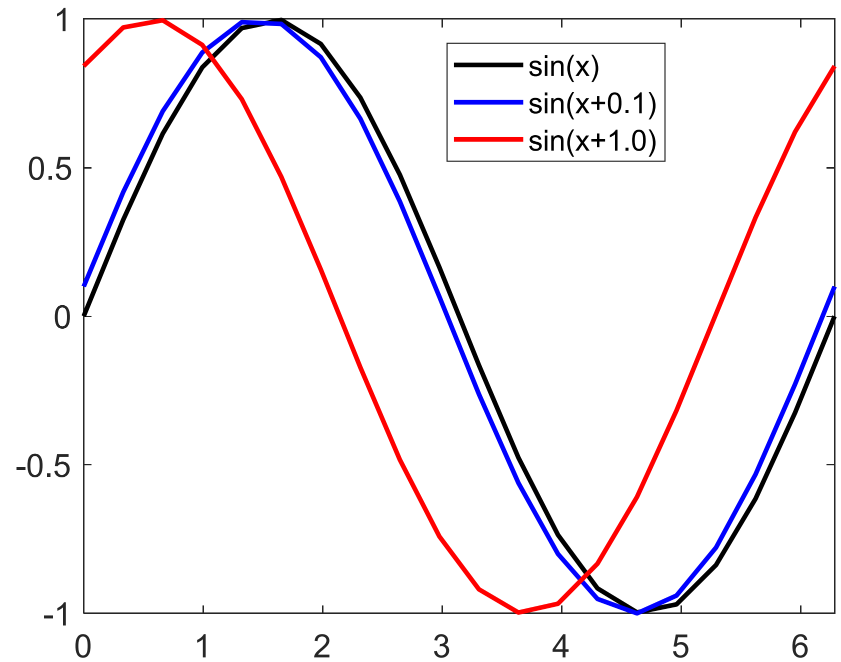
- Correlation $cor(X,Y) = \dfrac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$ $\quad -1 \le cor(X,Y) \le 1$

- The correlation matrix has 1 on the diagonal.

Correlation between functions at near and far points

- Generate 10 random samples, translate them by a bit (0.1), and by more (1.0)

```
x=10*rand(1,10);
xnear=x+0.1; xfar=x+1;
ynear=sin(xnear);
y=sin(x);
yfar=sin(xfar);
```



- Compare corelations:

```
r=corrcoef(y,ynear)         0.9894;
rfar=corrcoef(y,yfar)       0.4229;
```

High correlation
Low correlation

- Decay to about 0.4 over one sixth of the wavelength.
  - Wavelength on sine function is 2π~6

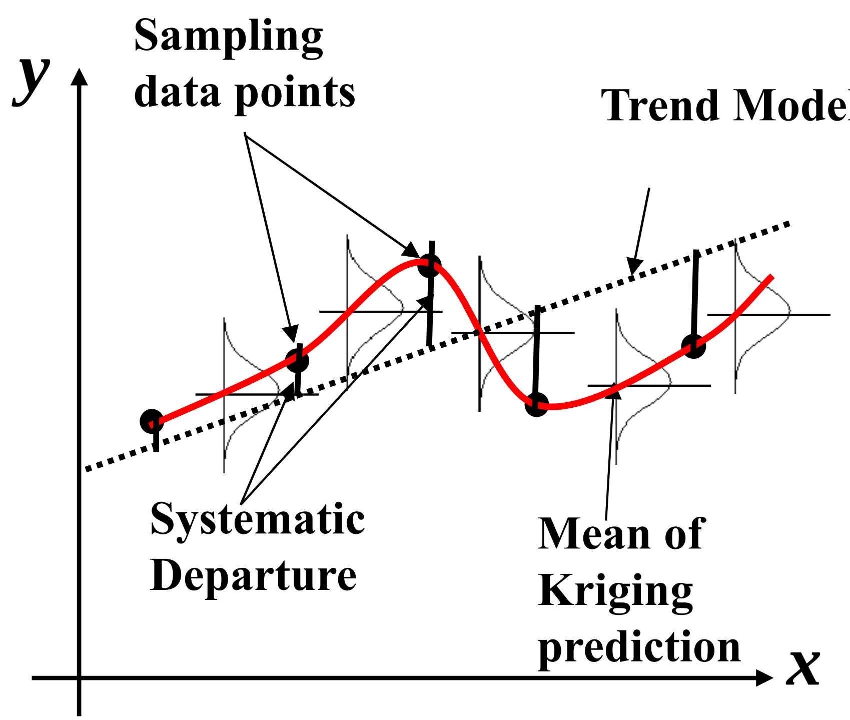
Universal Kriging approximation

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

$$\hat{y}(x) = \sum_{i=1}^{n_\beta} \beta_i \xi_i(x) + z(x)$$

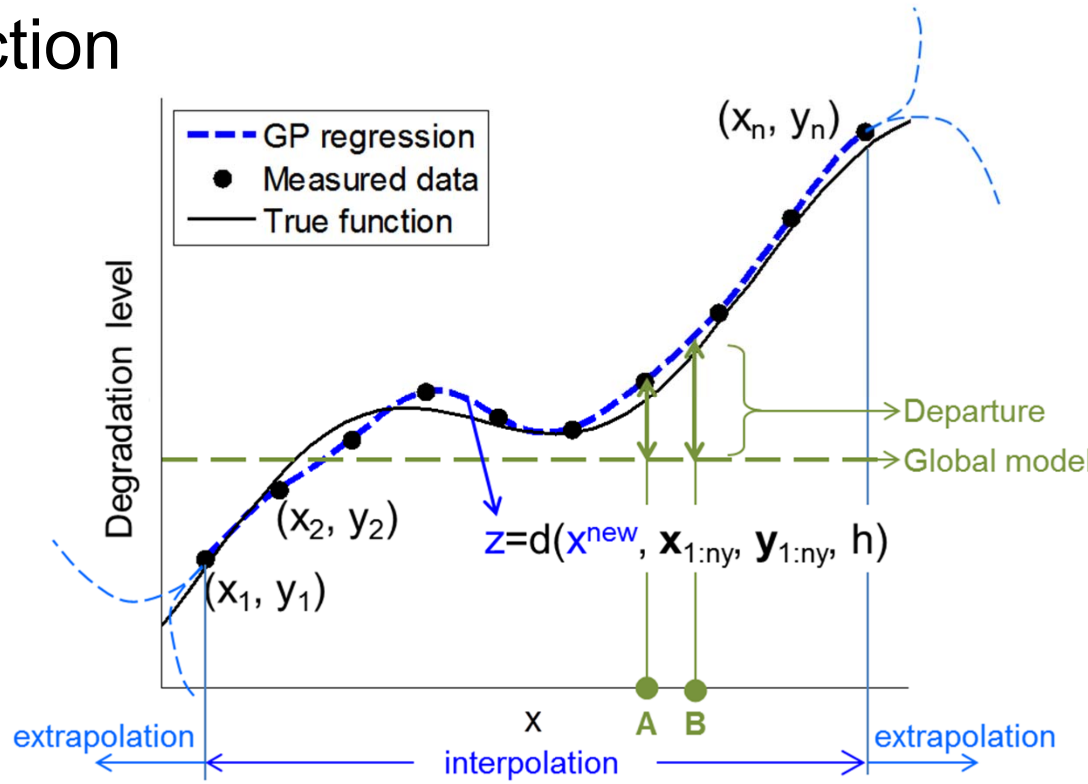
Systematic departure (random process) → $z(x)$

Trend function, Global function (low-order polynomials) → $\beta_i \xi_i(x)$



Ordinary and simple Kriging

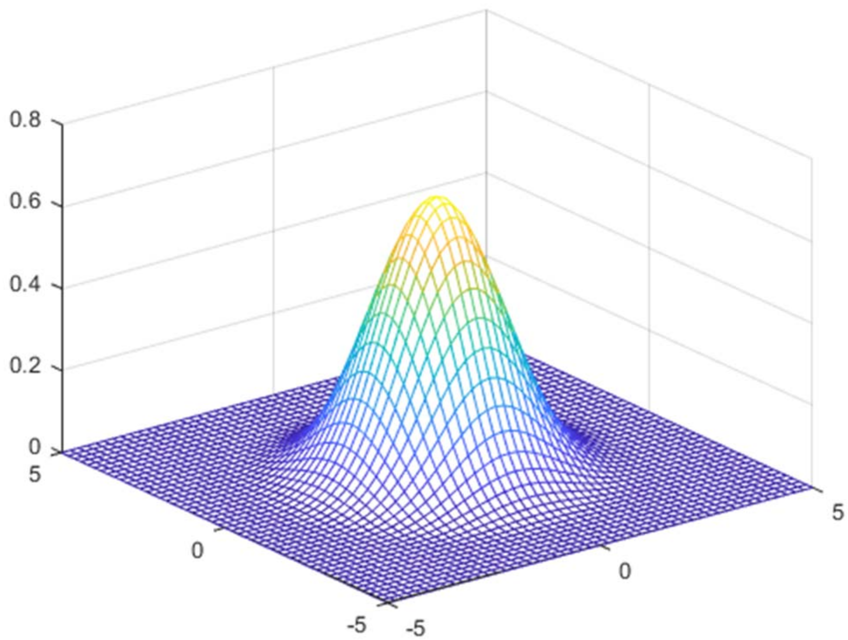
- Ordinary Kriging: constant trend function
- Simple Kriging: constant trend function is known (often 0)
- Assumption: Systematic departures $z(x)$ are correlated.
- Kriging prediction comes with a normal distribution of the uncertainty in the prediction
- At the sample points, the uncertainty is zero



Correlation model

- 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: $\text{cov}[z(\mathbf{x}^{(i)}), z(\mathbf{x}^{(j)})] = \sigma^2 \phi(\theta, \mathbf{x}^{(i)}, \mathbf{x}^{(j)})$
  - Variance of function: $\sigma^2 = \text{cov}[z(x), z(x)]$
- Isotropic correlation: $\phi(\theta, \mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \prod_{k=1}^{n} \phi(\theta, \left|x_k^{(i)} - x_k^{(j)}\right|)$
- Anisotropic correlation: $\phi(\theta, \mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \prod_{k=1}^{n} \phi_j(\theta_j, \left|x_k^{(i)} - x_k^{(j)}\right|)$

Isotropic vs. anisotropic correlation functions



(a) Isotropic correlation (b) Anisotropic correlation

Gaussian correlation function

- Correlation between point x and point s

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

- θ_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 $(l_i/6\theta_\mathrm{k})^2 = 1$ or $\theta_i = l_i/6$
- For the function $y = \sin(x_1) * \sin(5x_2)$ we would like to estimate $\theta_1 \approx 1$, $\theta_2 \approx 0.2$

Notation

- n_y sample points $\left(\mathbf{x}^{(i)}, y_i\right)$, with n-dimension of input $x_k^{(i)}, k = 1, \ldots, n$ and $y_i = y(\mathbf{x}^{(i)})$
- Given decay rates θ_k, the covariance matrix of the data

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

- The correlation matrix **R** is formed from the covariance matrix, assuming a constant standard deviation σ, which measures the uncertainty in function values (stationary covariance)
- Small σ for dense data, large σ 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 $\epsilon \sim N(0, \sigma^2)$ that are statistically independent
- Kriging assumes that data are accurate, but the model form is uncertain → 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, $\boldsymbol{\beta}$, and variance of data, σ^2
- Error b/w data and global function

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_y} \end{Bmatrix} - \begin{bmatrix} -\boldsymbol{\xi}(\mathbf{x}_1)- \\ -\boldsymbol{\xi}(\mathbf{x}_2)- \\ \vdots \\ -\boldsymbol{\xi}(\mathbf{x}_{n_y})- \end{bmatrix} \begin{Bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n_p} \end{Bmatrix}$$

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

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

Maximum likelihood estimate (MLE)

- Logarithmic likelihood (ignore $\boldsymbol{\theta}$ for now)

$$\ln\left[f\left(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2\right)\right] = -\frac{n_y}{2}\ln(2\pi) - \frac{n_y}{2}\ln(\sigma^2) - \frac{1}{2}\ln|\mathbf{R}| - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}$$

- Stationary condition

$$\frac{\partial \ln f}{\partial \boldsymbol{\beta}} = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} = 0$$

$$\frac{\partial \ln f}{\partial \sigma^2} = -\frac{n_y}{2}\frac{1}{\sigma^2} + \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^4} = 0$$

- Solve for $\boldsymbol{\beta}$ and σ^2

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\left\{\mathbf{X}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y}\right\}$$

$$\hat{\sigma}^2 = \frac{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n_y - n_\beta}$$

PRS linear regression

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\hat{\sigma}^2 = \frac{SSe}{n_y - n_\beta}$$

For unbiased estimate

Local departure

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

- Minimizing mean squared error (MSE)

$$\varepsilon(\mathbf{x}) = \hat{y}(\mathbf{x}) - y(\mathbf{x}) = \mathbf{w}(\mathbf{x})^{\mathrm{T}}\mathbf{y} - y(\mathbf{x})$$

Weight function

  – Data: $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{z}$

  – True function: $y(\mathbf{x}) = \boldsymbol{\xi}(\mathbf{x})\hat{\boldsymbol{\beta}} + z(\mathbf{x})$

$$\Rightarrow \varepsilon(\mathbf{x}) = \mathbf{w}(\mathbf{x})^{\mathrm{T}}\{\mathbf{X}\hat{\beta} + \mathbf{z}\} - \left(\xi(\mathbf{x})\hat{\beta} + z(\mathbf{x})\right)$$

$$= \underbrace{\left(\mathbf{w}(\mathbf{x})^{\mathrm{T}}\mathbf{X} - \xi(\mathbf{x})\right)\hat{\beta}}_{\text{Global error}} + \underbrace{\mathbf{w}(\mathbf{x})^{\mathrm{T}}\mathbf{z} - z(\mathbf{x})}_{\text{Departure error}}$$

Global error　　Departure error

Minimize MSE

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

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

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

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

$$\text{cov}\left[z(\mathbf{x}), z(\mathbf{x})\right] = \sigma^2$$

$$\text{cov}\left[z(\mathbf{x}_k), z(\mathbf{x})\right] = \sigma^2\mathbf{r}$$

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

$$\mathbf{r}(\mathbf{x}) = \left[R\left(\mathbf{x}_k, \mathbf{x}\right)\right]$$

- MSE is the variance (uncertainty) in Kriging prediction

- Goal: find **w**(**x**) that minimizes MSE while satisfying the unbiased constraint

Lagrange function for constrained optimization

- Lagrange function (min. MSE with constraint)

$$L(\mathbf{w},\lambda)=\sigma^2\left(\mathbf{w}^T\mathbf{R}\mathbf{w}-2\mathbf{w}^T\mathbf{r}+1\right)-\lambda\left(\mathbf{X}^T\mathbf{w}-\xi^T\right)$$

Lagrange multiplier

- Stationary conditions (KKT)

$$\begin{cases}\dfrac{\partial L(\mathbf{w},\lambda)}{\partial \mathbf{w}}=2\sigma^2\left(\mathbf{R}\mathbf{w}-\mathbf{r}\right)-\mathbf{X}\lambda^T=\mathbf{0}\\[2ex]\dfrac{\partial L(\mathbf{w},\lambda)}{\partial \lambda}=\mathbf{X}^T\mathbf{w}-\xi^T=\mathbf{0}\end{cases}$$

$$\Rightarrow\begin{cases}\mathbf{w}=\mathbf{R}^{-1}\mathbf{r}+\mathbf{R}^{-1}\mathbf{X}\dfrac{\lambda^T}{2\sigma^2}\\[2ex]\dfrac{\lambda^T}{2\sigma^2}=(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}\left\{\xi^T-\mathbf{X}^T\mathbf{R}^{-1}\mathbf{r}\right\}\end{cases}$$

$$\Rightarrow \hat{y}(\mathbf{x})=\mathbf{w}(\mathbf{x})^T\mathbf{y}$$

Kriging prediction

- Kriging as linear combination of data and weight functions

$$\hat{y}(\mathbf{x}) = \mathbf{w}(\mathbf{x})^T \mathbf{y}$$

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

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

$$= \mathbf{r}^T\mathbf{R}^{-1}\mathbf{y} + \left(\xi - \mathbf{r}^T\mathbf{R}^{-1}\mathbf{X}\right)\underbrace{\left(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\left\{\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}\right\}}_{\hat{\beta}}$$

$$\Rightarrow \hat{y}(\mathbf{x}) = \xi(\mathbf{x})\hat{\beta} + \mathbf{r}(\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$$

Trend function Local departure

– Local departure term is the weighted sum of the trend function error $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ based on the correlation term $\mathbf{r}(\mathbf{x})^T\mathbf{R}^{-1}$

Simplification for ordinary Kriging

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

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

- Linear in **r**(**x**) that the radial basis can be viewed as basis functions

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

- The prediction is linear in the data **y**, in common with linear regression, but **b** is not calculated by minimizing MSE.

- Note that far away from data, $\hat{y}(\mathbf{x}) \sim \hat{\mu}$ (not good for substantial extrapolation)

Estimating hyperparameters θ

- Estimating $\hat{\beta}$ and $\hat{\sigma}$ depends on hyperparameter $\boldsymbol{\theta}$ in **R**
- Maximizing the log-likelihood that the data comes from a Gaussian process defined by θ_k.

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

$\mathbf{R}^{-1}$ will be canceled with σ^2

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

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

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

Estimating hyperparameters θ (ordinary Kriging)

- Once $\boldsymbol{\theta}$ is found, the estimate of the mean and standard deviation is obtained as (ordinary Kriging)

$$\hat{\mu} = \frac{\mathbf{1}^T \mathbf{R}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1}}, \quad \hat{\sigma}^2 = \frac{\left(\mathbf{y} - \mathbf{1}\hat{\mu}\right)^T \mathbf{R}^{-1} \left(\mathbf{y} - \mathbf{1}\hat{\mu}\right)}{n_y - n_\beta}$$

Prediction uncertainty

- Kriging prediction $\hat{y}(\mathbf{x}) = \xi(\mathbf{x})\hat{\beta} + \mathbf{r}(\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$ is the mean prediction and MSE is the variance

- Kriging prediction is Gaussian distribution

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

  – This is an estimated uncertainty using data

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

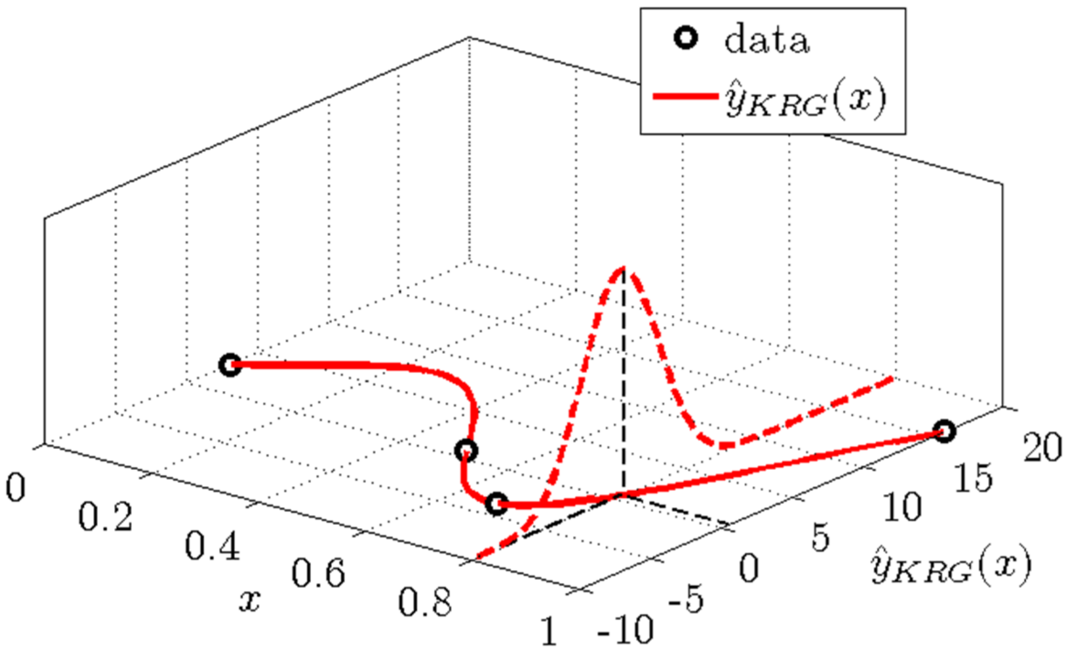
$$\hat{Y}(\mathbf{x}) \sim \xi\hat{\beta} + \mathbf{r}^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) + t_{n_y - n_\beta} \cdot \hat{\sigma}\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{R}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathbf{r} + 1}$$

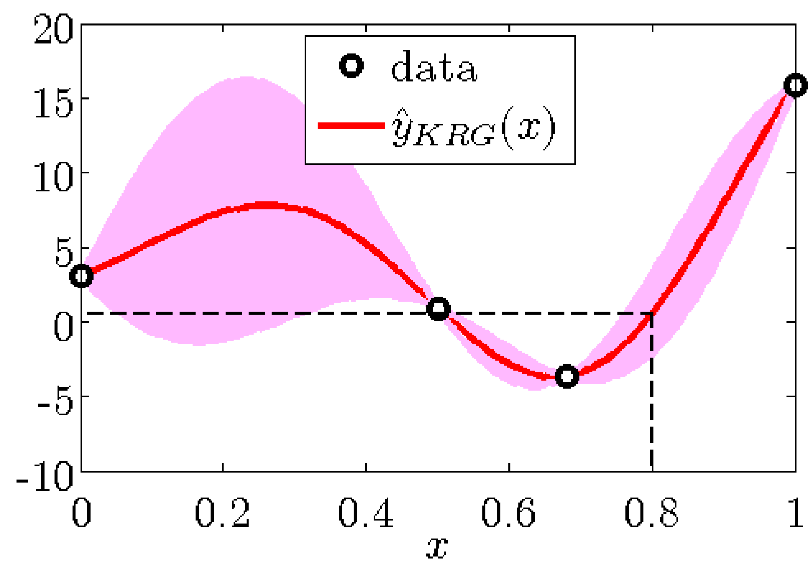
- Ordinary Kriging

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

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.

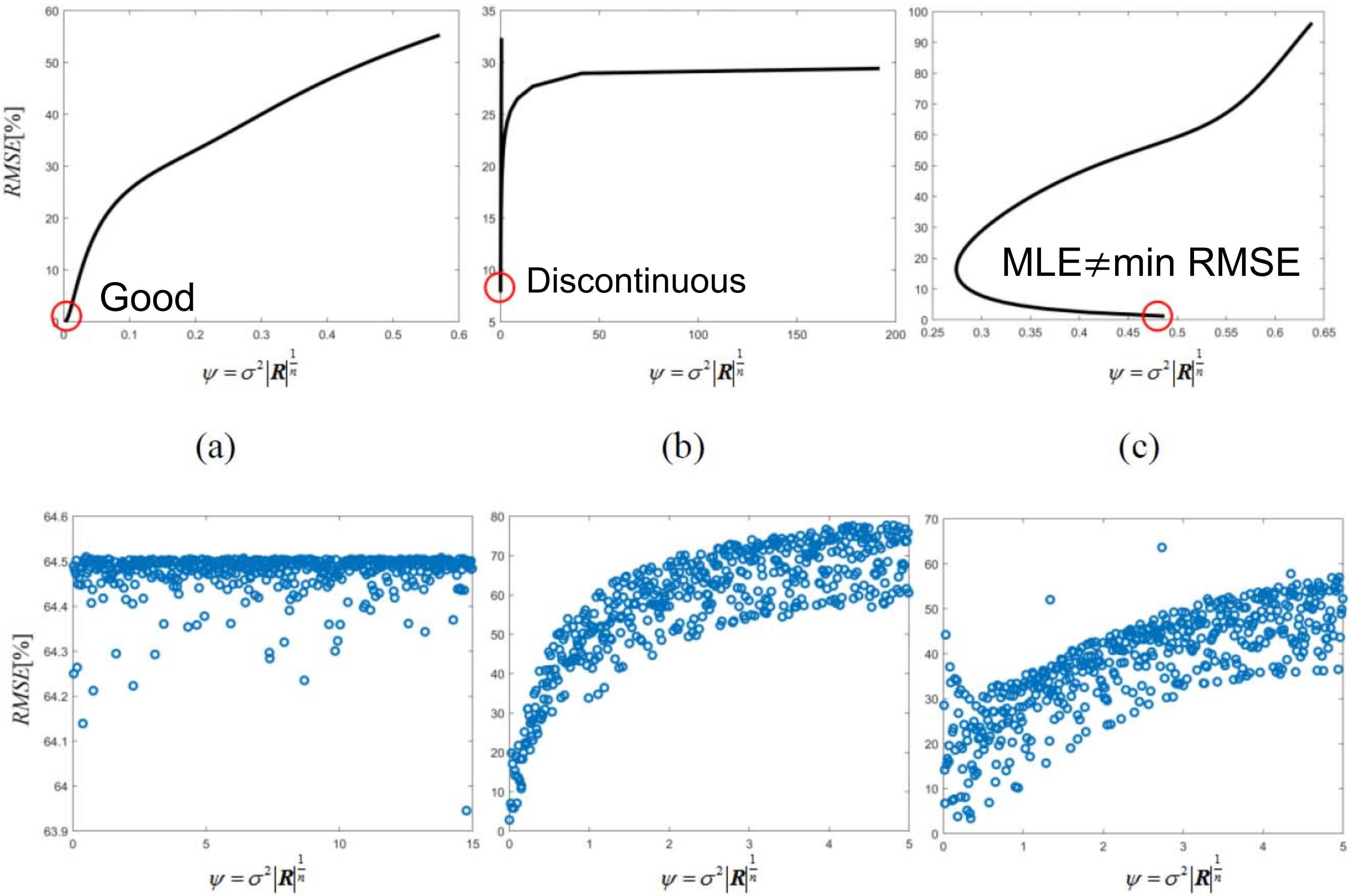




Kriging fitting issues

- 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

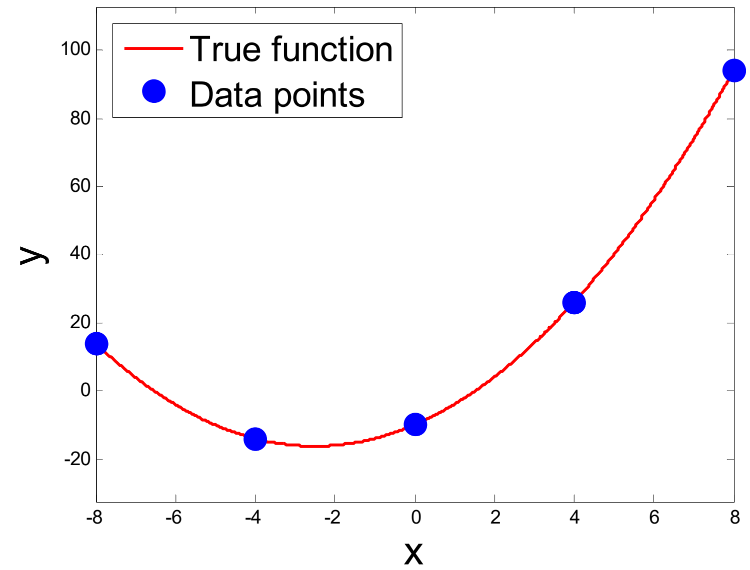


(a) (b) (c)

(d) More than 2 hyperparameters (f)

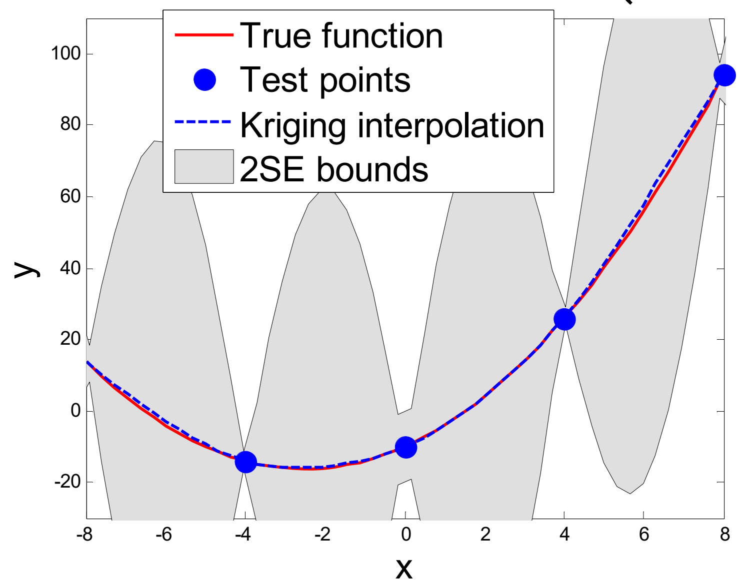
Ex) Quadratic function fit

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

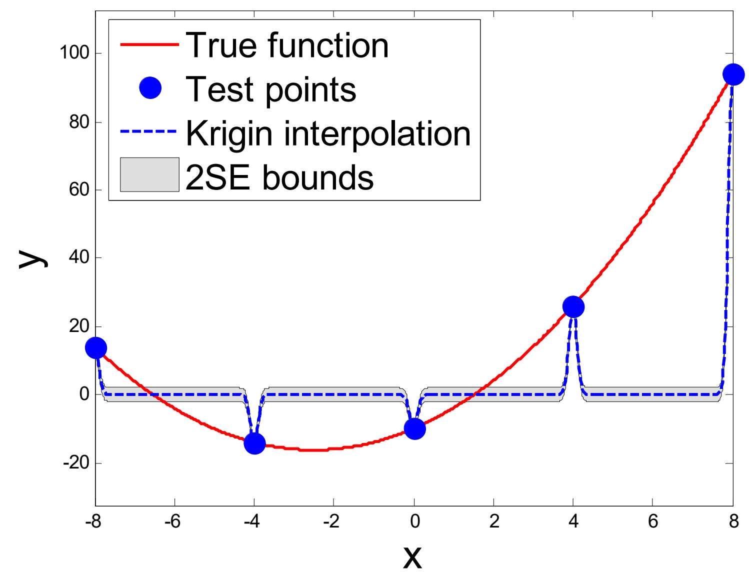


- Covariance

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



Too large θ
Good fit with poor variance



Too small θ
Bad fit with poor variance

Ex) Kriging fit

- Fit data $\mathbf{x} = \{0, 5, 10, 15, 20\}^{\mathrm{T}}$, $\mathbf{y} = \{1, 0.99, 0.99, 0.94, 0.95\}^{\mathrm{T}}$ for the global function using ordinary Kriging with $\theta = 5.2$

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

```
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(-(norm(x(k,:)-x(l,:))/h)^2);
end; end;
```

$$\mathbf{R} = \begin{bmatrix} 1 & 0.3967 & 0.0248 & 0.0002 & 0 \\ 0.3967 & 1 & 0.3967 & 0.0248 & 0.0002 \\ 0.0248 & 0.3967 & 1 & 0.3967 & 0.0248 \\ 0.0002 & 0.0248 & 0.3967 & 1 & 0.3967 \\ 0 & 0.0002 & 0.0248 & 0.3967 & 1 \end{bmatrix}$$

Ex) Kriging fit *cont.*

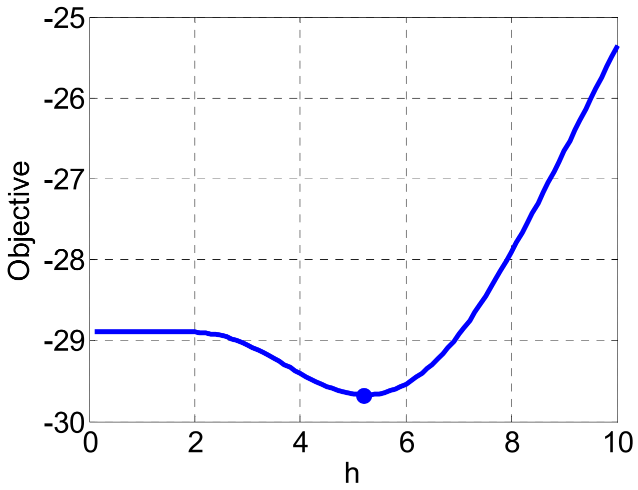
- Global function parameters

```
Rinv=inv(R);
thetaH=(X'*Rinv*X)\(X'* Rinv*y);
sigmaH=sqrt(1/(ny-np)*((y-X*thetaH)'*Rinv*(y-X*thetaH)));
```

$$\hat{\beta} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\left\{\mathbf{X}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{y}\right\} = 3.0989^{-1} \times 3.0226 = 0.9754$$

$$\hat{\sigma}^2 = \frac{(\mathbf{y}-\mathbf{X}\hat{\beta})^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta})}{n_y - n_\beta} = 7.28 \times 10^{-4},\ \hat{\sigma} = 0.0270$$

- Estimate the optimum hyperparameter
  - Instead of optimization the hyperparameter, we calculate it graphically
  - $\theta_{opt} = 5.2$
  - We used this value in calculating $\hat{\beta}$ and $\hat{\sigma}^2$



Ex) Kriging fit *cont.*

- Matlab code for the graph

```
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 $x = 10$

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

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

– Exact at the sample point !

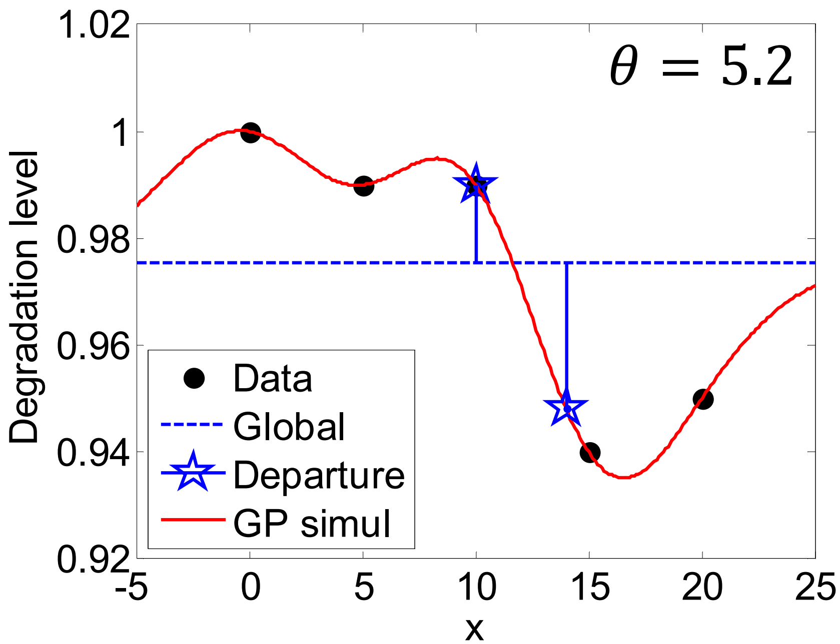
Ex) Kriging fit *cont.*

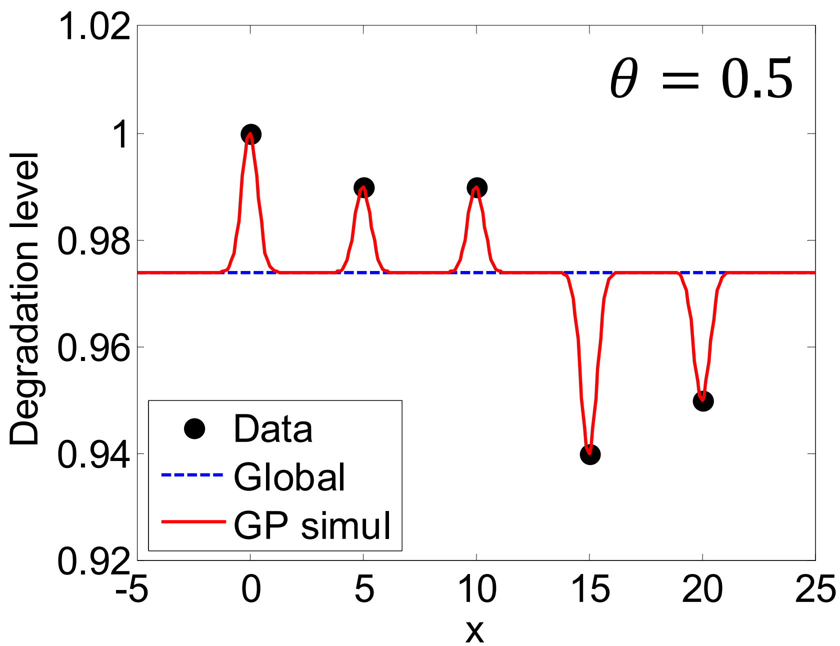
- Prediction at $x = 14$

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

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

```
xNew=10; %or xNew=14
for k=1:ny; r(k,1)=exp(-(norm(x(k,:)-xNew)/h)^2); end;
gpDepar=r'*Rinv*(y-X*thetaH);
```



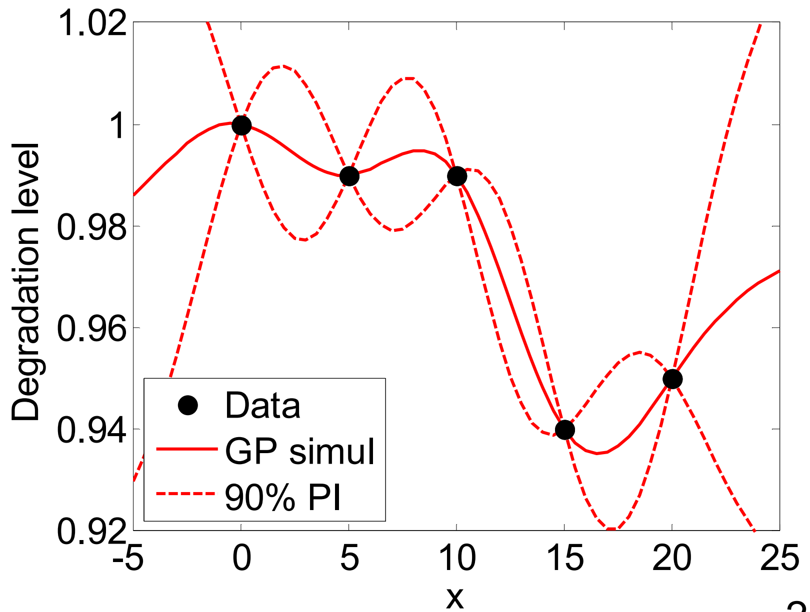


Ex) Kriging fit *cont.*

- 90% confidence intervals
  - Standard error: $s_y = \hat{\sigma}\sqrt{\mathbf{w}^T\mathbf{R}\mathbf{w} - 2\mathbf{w}^T\mathbf{r} + 1}$

```
xi=1;
w=Rinv*r+Rinv*X*((X'*Rinv*X)\(xi'-X'*Rinv*r));
zSigmaH=sigmaH*sqrt(w'*R*w-2*w'*r+1);
% using the inverse calculation
gpMean=0.9482; % from Example 5.2
PI=[ gpMean + tinv(0.05,ny-np)*zSigmaH, ...
          gpMean + tinv(0.95,ny-np)*zSigmaH]
% using the random samples
ns=5e3;
tDist=trnd(ny-np,1,ns);
yHat=gpMean+tDist*zSigmaH;
PI=prctile(yHat,[5 95])
```

x_{new}	5 percentile	95 percentile	90% P.I.
x = 10	0.99	0.99	0
x = 14	0.9394	0.9570	0.0176



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).