Surrogate Models

Hans Bruun Nielsen

IMM, Numerical Analysis Section

- Alternatives to physically based mathematical models and local Taylor expansions
- Metamodels, Surface Response, Neural Networks, ...
- Space Mapping
- Radial Basis Functions (RBF)
- Kriging, "Design and Analysis of Computer Experiments" (DACE)

Approximation tools. Interested in applications in

- Data representation (fitting)
- Optimization

Data Fitting

Given $\{(x_i, y_i)\}_{i=1}^m, \quad y_i = Y(x_i) + e_i$

Seek (an approximation to) $Y(x)$

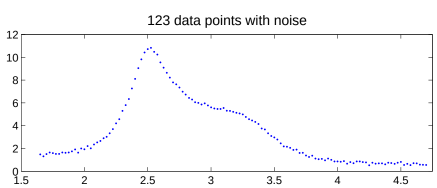
May have a mathematical model

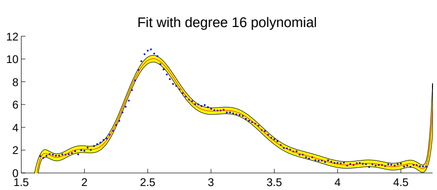
$$Y(x) \simeq M(p, x)$$

Parameters p *eg* determined by minimizing

$$\varrho(p) = \sum w_i^2 \left(y_i - M(p, x_i)\right)^2$$

In lack of a proper model we may use a polynomial as *"surrogate model"*.





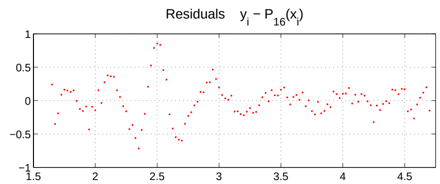
Poor approximation.

Polynomials have too "long memory".

The Taylor expansion

$$P_n(x+h) = P_n(x) + \sum_{k=1}^{n} \frac{1}{k!} h^k P_n^{(k)}(x)$$

is exact for any h



Trends for $2.2 \lesssim x \lesssim 3$: Increase n

Degree too high for $x \lesssim 2$ and $x \gtrsim 3.5$

Cubic Splines

Information that should be carried by 3rd and higher derivatives is lost.

Local nature. Put knots where they are needed.

M.J.D. Powell: *Curve fitting by splines in one variable*

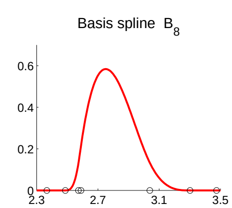
pp 65 –83 in J.G. Hayes (ed): *Numerical approximation to functions and data,* 1970.

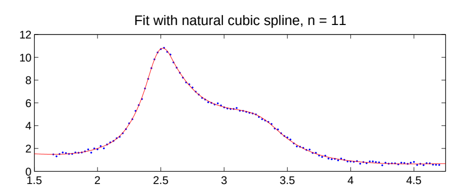
Knots $\kappa_0, \kappa_1, \ldots, \kappa_n$

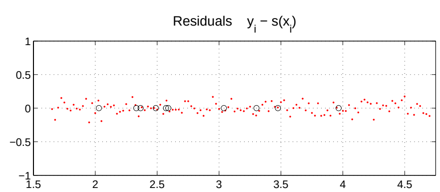
$$s(x) = \sum_{j=1}^{n+3} c_j B_j(x)$$

Basis spline B_j is nonzero only in four consecutive knot intervals. Local support.

c_j has influence only in $[\kappa_{j-4}, \kappa_j]$







$x \in \mathbb{R}^d$. Polynomials and splines generalize.

Curse of dimensionality

Interpolation or fitting,

$$B\,c \simeq f$$

Serious risk of rank deficient B.

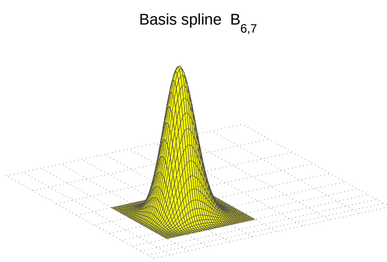
Number of basis functions for P_n

d	1	2	3	5	10	50
$n=1$	2	3	4	6	11	51
2	3	6	10	21	66	1326
3	4	10	20	56	286	23426
4	5	15	35	126	1001	316251
5	6	21	56	252	3003	3478761

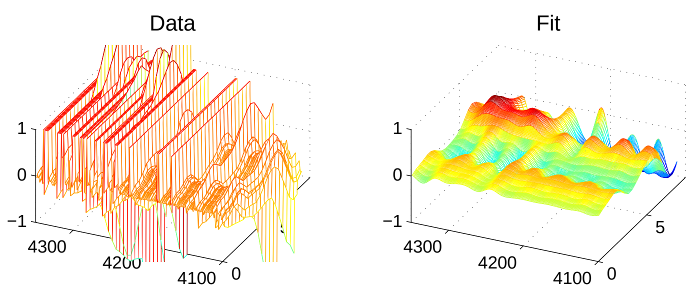
Bicubic splines $(x = (u,v) \in \mathbb{R}^2)$.

$$s(x) = \sum_{i=1}^{n_1+3} \sum_{j=1}^{n_2+3} c_{ij} B_i(\xi, u) B_j(\eta, v)$$

ξ and η : knots in u and v-directions, resp.



Example. *Apnea.* Measurements of pressure in throat as function of distance (22 values in $]0, 10[$ cm) and time (every 0.1 second).

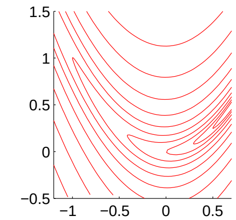


Missing data and "wild points".

$n_1 = 8$, sliding window with one knot per 20 profiles (2 seconds).

Byproduct: Data compression.

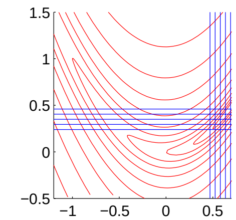
Level curves of a function that we want to approximate show *eg*, that we need close knots in both directions at $(0.6, 0.5)$.



Also close where it is not needed. The system,

$$A\,c \simeq f$$

is either rank deficient or needs many "superfluous" data points.



Alternative approximating function.
Given data points (x_i, y_i), $i = 1, 2, \ldots, m$
with distinct $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$

Surrogate model

$$s(x) = c^T \underline{\phi}(x) + \beta^T \underline{\psi}(x) = \sum_{j=1}^{m} c_j \phi_j(x) + \sum_{j=1}^{n} \beta_j \psi_j(x)$$

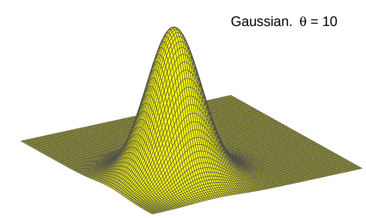
where the ψ_j are basis functions *eg* for a low order polynomial that models a "global trend", and

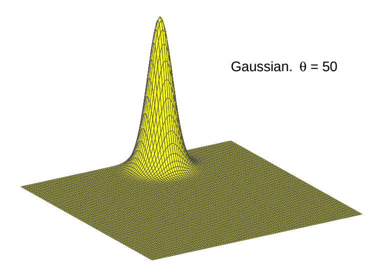
$$\phi_j(x) = \phi(\|x - x_j\|_2)$$

Eg a Gaussian

$$\phi(r) = e^{-\theta\, r^2}$$

The figures show $x \in [-1, 1]^2$





Surrogate models based on Kriging and Radial Basis Functions (RBF) both have the form

$$s(x) = c^T \underline{\phi}(x) + \beta^T \underline{\psi}(x) \; .$$

Different derivation, but (under certain conditions on ϕ): same model.
We consider interpolation, *ie* $s(x_i) = y_i,\ i{=}1,\ldots,m$.

Let $\Phi \in \mathbb{R}^{m\times m}, \quad \Psi \in \mathbb{R}^{m\times n}$ be the matrices defined by

$$\Phi_{ij} = \phi(\|x_i - x_j\|_2), \qquad \Psi_{ij} = \psi_j(x_i)$$

The interpolation condition can be expressed as

$$\Phi\, c + \Psi\, \beta = y.$$

In the case of RBF this is combined with the condition that c should be orthogonal to the range of Ψ, and we get the linear system of equations

$$\begin{pmatrix} \Phi & \Psi \\ \Psi^T & 0 \end{pmatrix} \begin{pmatrix} c \\ \beta \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix} \quad \Leftrightarrow \quad \begin{pmatrix} \Phi & \Psi \\ 0 & -\Psi^T\Phi^{-1}\Psi \end{pmatrix} \begin{pmatrix} c \\ \beta \end{pmatrix} = \begin{pmatrix} y \\ -\Psi^T\Phi^{-1}y \end{pmatrix}.$$

Solution: $\quad \beta = \left(\Psi^T\Phi^{-1}\Psi\right)^{-1} Psi^T \Phi^{-1} y, \quad c = \Phi^{-1}\left(y - \Psi\beta\right)$.

Kriging is a statistical approach.

$$s(x) = z(x) + \beta^T \underline{\psi}(x) \; ,$$

where z is stochastic with mean 0 and covariances $\mathrm{E}\big[z(x), z(x_j)\big] = \sigma^2 \phi_j(x)$. Process variance σ^2

Apply this model to the given data to get $y = Z + \Psi\beta, \qquad \mathrm{E}\big[Z\, Z^T\big] = \sigma^2\, \Phi$.

Express $s(x)$ as a linear predictor, $s(x) = y^T\gamma(x)$. The error is

$$s(x) - y(x) = (Z + \Psi\beta)^T\gamma(x) - \big(z(x) + \beta^T\underline{\psi}(x)\big) = Z^T\gamma(x) - z(x) + \beta^T\left(\Psi^T\gamma(x) - \underline{\psi}(x)\right)$$

Unbiased predictor. Constraint: $\quad \Psi^T\gamma(x) - \underline{\psi}(x) = 0$.

Mean squared error (MSE)

$$\Omega(x) = \mathrm{E}\big[(s(x) - y(x))\big] = \mathrm{E}\big[z^2 + \gamma^T Z\, Z^T\gamma - 2\gamma^T Z\, x\big] = \sigma^2\left(1 + \gamma^T\Phi\,\gamma - 2\gamma^T\underline{\psi}\right)$$

Minimize Ω with respect to γ and subject to the constraint:

$$\begin{pmatrix} \Phi & \Psi \\ \Psi^T & 0 \end{pmatrix} \begin{pmatrix} \gamma \\ \lambda \end{pmatrix} = \begin{pmatrix} \underline{\phi} \\ \underline{\psi} \end{pmatrix}$$

The solution can be reformulated to

$$s(x) = c^T \underline{\phi}(x) + \beta^T \underline{\psi}(x) \; .$$

β is the "generalized least squares" solution to $\Psi\beta \simeq y$ and c is found from the corresponding residual,

$$\left(\Psi^T\Phi^{-1}\Psi\right)\beta = \Psi^T\Phi^{-1}y, \qquad c = \Phi^{-1}(y - \Psi\beta). \qquad \text{Same solution!}$$

The process variance is estimated by $\quad \sigma^2 = \dfrac{1}{m}(y - \Psi\beta)^T\Phi^{-1}(y - \Psi\beta)$,
and the MSE can be computed by

$$\Omega(x) = \sigma^2\left(1 + u^T\left(\Psi^T\Phi^{-1}\Psi\right)^{-1}u - \underline{\phi}^T\Phi^{-1}\underline{\phi}\right), \qquad u = \Psi^T\Phi^{-1}\underline{\phi} - \underline{\psi}\; .$$

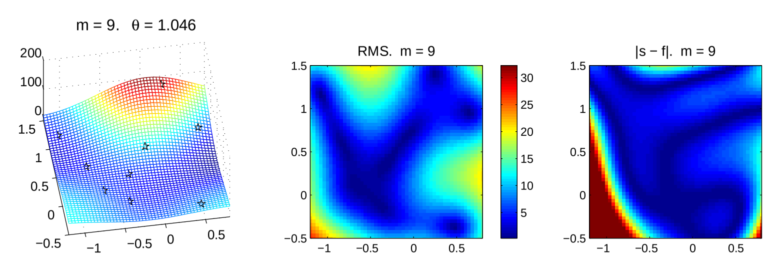
In classical Kriging the correlation function ϕ is estimated from a variogram of the data, or it may be taken from a list. We may, *eg*, use a Gaussian: $\phi_j(x) = e^{-\theta\,\|x - x_j\|^2}$.

The results depend on θ, and assuming a Gaussian process the optimal choice is

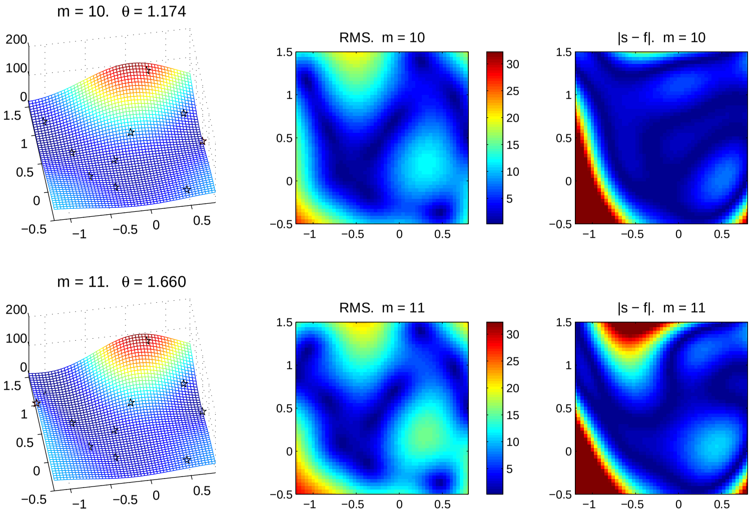
$$\theta_{\mathrm{opt}} = \mathrm{argmin}\{\ |\Phi(\theta)|^{1/m}\cdot\sigma^2(\theta)\ \} \qquad (\,|\cdot| = \det(\cdot)\,)\; .$$

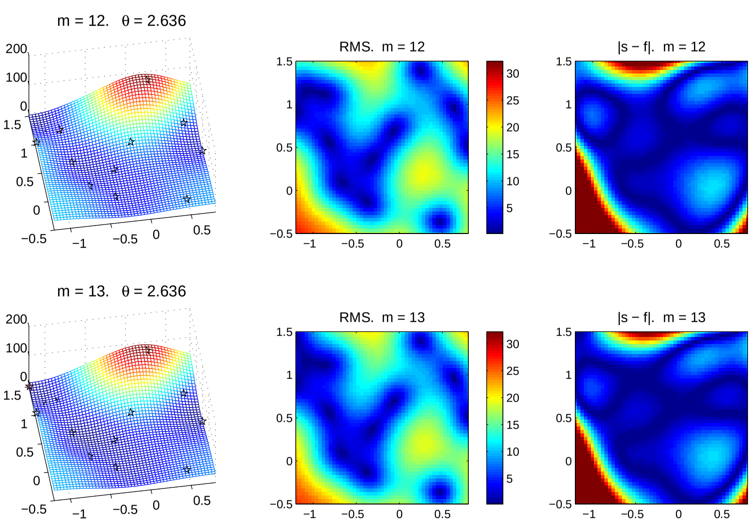
MATLAB toolbox DACE (with *Søren N. Lophaven* and *Jacob Søndergaard*)
Available at http://www.imm.dtu.dk/∼hbn/dace/

Example. *Rosenbrock's function.* $\quad n = 1, \quad \psi(x) = 1$
Start with 9 points. Best $\theta \in [0.1, 100]$. $\quad RMS = \sqrt{\Omega}$

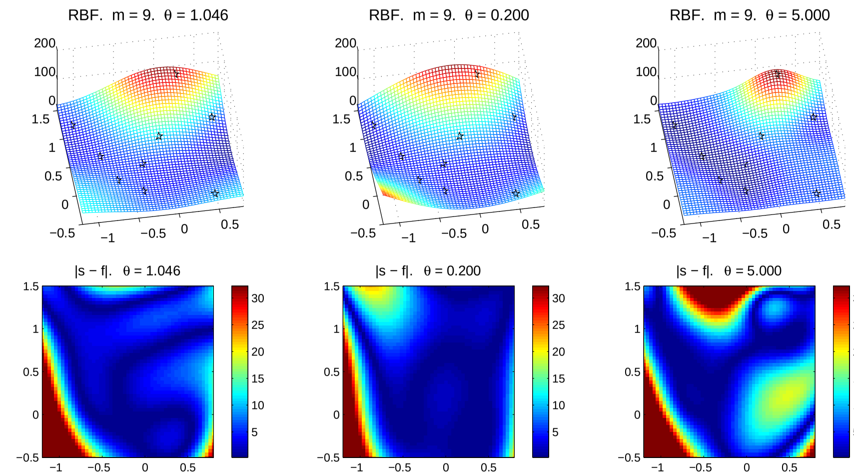


Successively insert new data points where $s(x) - .05\sqrt{\Omega(x)}$ is minimal.

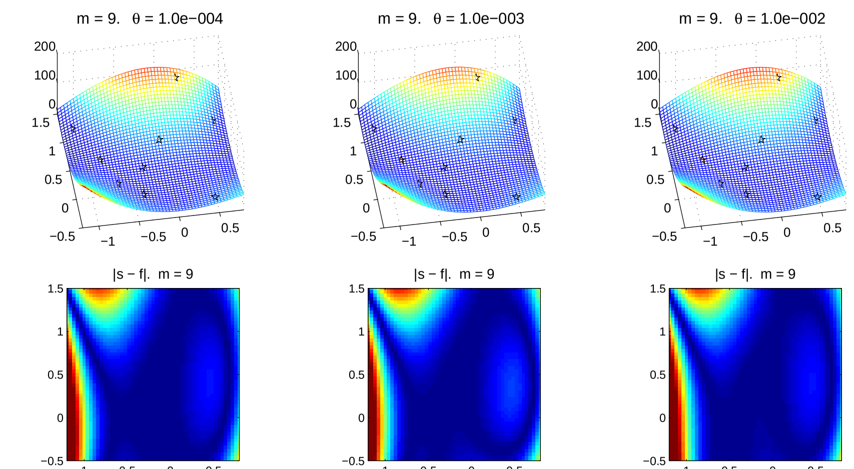




RBF. Gaussian. Choose θ



RBF. Inverse multiquadric: $\phi_j(x) = \left(\theta\|x - x_j\|^2 + 1\right)^{-1/2}$



Plans for future work

- Better error estimation for Kriging
- Extend DACE to cope with errors in data
- Strategy for use in optimization
- Choice of θ in RBF
- Extend DACE to cope with other RBFs
- $\cdots$

With *Kristine Frisenfeldt Thuesen*