Smooth polynomial interpolators

Hugo Maruri, Henry Wynn

London School of Economics

Statistische Modellbildung

5th Autumn Symposium “Statistical modelling”
Universitätskolleg Bommerholz, 29st November 2008

Smooth polynomial interpolators

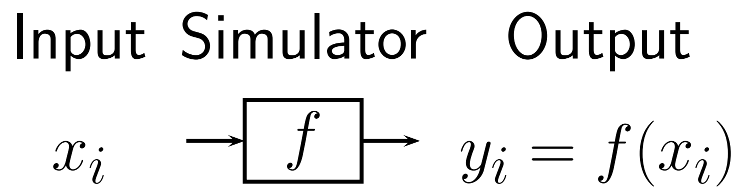
Splines are piecewise polynomial functions, known to minimise smoothness properties in one and higher dimensions. However, splines are not always continuously differentiable and sometimes it might be difficult to construct splines over irregular design regions.

Smooth saturated interpolators can be constructed by first extending the algebraic monomial basis and then minimising over a given region a measure of smoothness with respect to the free parameters in the extended basis. This method allows a polynomial approximation to splines and allows for a flexible smoothing region. The resulting model shares all the advantages of polynomial models (linearity in parameters and in observations), while at the same time is smooth and thus close to a spline model.

As an example, this technique is applied to sensitivity analysis of computer simulations. Future work pointing to smooth Hermite interpolation is also discussed.

Computer experiments

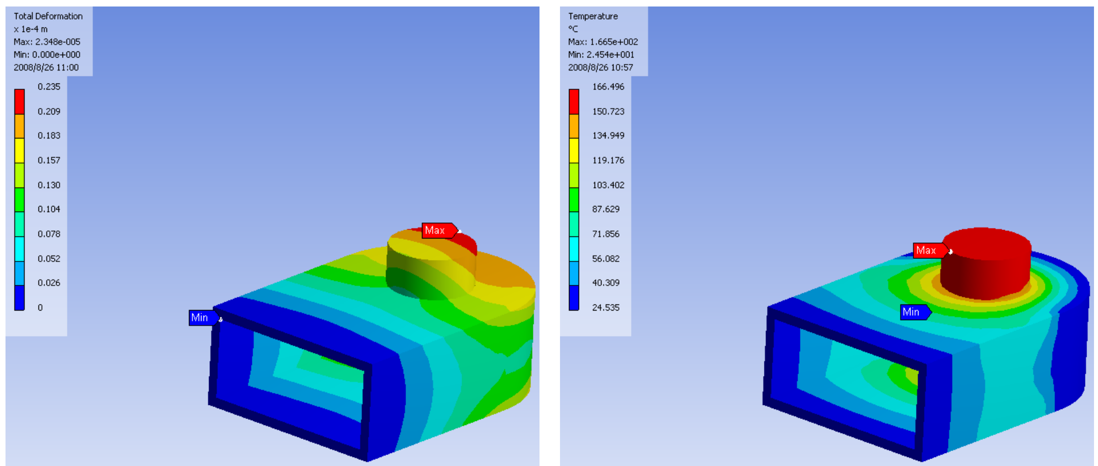
A computer simulator takes inputs x and produces outputs $y = f(x)$.



Example 1: Engine emissions simulator (Bates *et al.* 2003). Five input variables N, C, A, B, M and one output Y. The code output is deterministic, i.e. replications give same value and thus models must interpolate the response.

Example 2: Manufacturing design

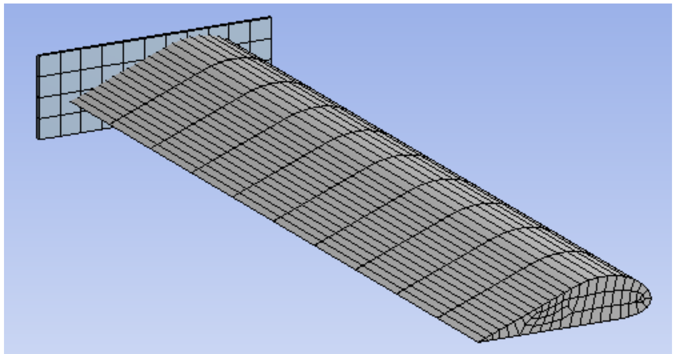
Responses: Total deformation of the piece, temperature. These are simulated under given operating conditions and material properties (*i.e.* design treatment or run).



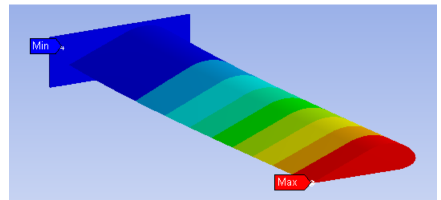
Graphics from ANSYS ED®

Although it is possible to propose models for the whole response values, usually the analysed response is a single value, representative for the whole assembly, *e.g.* maximum temperature, critical stress before break.

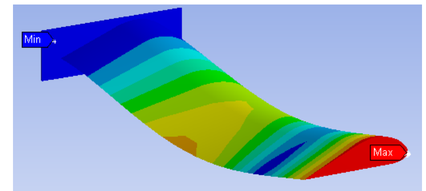
Example 3: Stress analysis of a plane wing



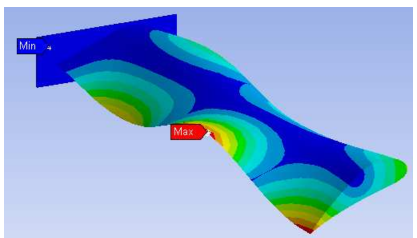
Plane wing mesh



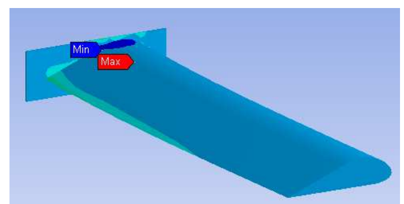
1st harmonic



3rd harmonic



10th harmonic



Total stress

Graphics from ANSYS ED®

The number of factors can be quite large and thus an important task is to screen out unimportant factors.

Modeling alternatives: Gaussian processes, splines, polynomials.

Contents of the talk

- Polynomial interpolators with algebra
  - Generalised confounding
- Modelling the response, the smooth polynomial approach
  - Smoothing the univariate interpolator
  - Smooth supersaturated models (SSM)
  - Comparison with kriging and splines by example:
    - Bidimensional example
    - Engine emissions dataset
  - Computing sensitivity analysis
- Smooth Hermite interpolation
- References

Saturated model I

- Design $\mathcal{D}$, n points, d factors.
- Identification of polynomial interpolator, e.g. with 2^2 design

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

- Saturated model, perfect fit
- No random error
- Suitable for computer experiments
- Submodels identifiable, random error, e.g.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- Model selection applicable

Saturated model II

One factor design : Lagrange interpolation

Polynomial division: Any polynomial passing through the values observed at n distinct x points is equivalent (on the design) to a polynomial of degree $n-1$.

Multifactor design : Polynomial interpolation

Multivariate polynomial division

$$\overbrace{\begin{pmatrix} x_1^2x_2+x_1x_2 \\ +x_1-x_2+1 \end{pmatrix}}^{\text{Polynomial}} \quad = \quad \overbrace{\begin{pmatrix} x_2(x_1^2-1) \\ +0(x_2^2-1) \end{pmatrix}}^{\text{Result}} \quad + \quad \overbrace{\begin{pmatrix} 1+x_1 \\ +x_1x_2 \end{pmatrix}}^{\text{Remainder}}$$

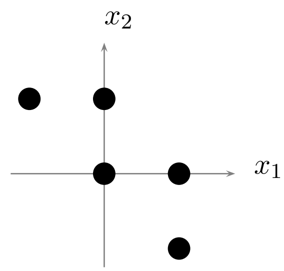
with $f = \sum_{i=1}^m h_i g_i + r$ (Polynomial = Result + Remainder).

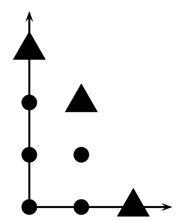
This division is the division of f by the *Ideal* of the design. The remainder r is created by l.i. monomials$\Rightarrow$model.

Generalised confounding (PW, 1996)

- Design $\mathcal{D}$, n points, d factors
- Study the $\mathcal{D}$ through the design ideal $I(\mathcal{D}) \subset \mathbb{R}[x]$
- The support for a model is given by those monomials not divisible by the leading terms of the Gröbner basis $G_\tau \subset I(\mathcal{D})$

Design $\mathcal{D}$





term order $\tau : x_1 \succ x_2$

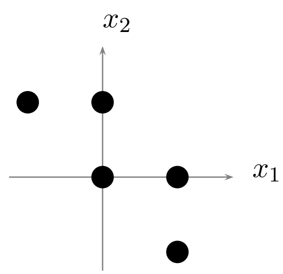
Model $= \{1, x_1, x_2, x_1x_2, x_2^2\}$

$$G_\tau = \{\underline{x_1^2} + 2x_1x_2 + x_2^2 - x_1 - x_2, \underline{x_2^3} - x_2, \underline{x_1x_2^2} - x_1x_2 - x_2^2 + x_2\}$$

- Exact polynomial interpolator $=$ saturated regression model
- Hierarchical polynomial model
- Link with aliasing/confounding $f(x) = g(x), x \in \mathcal{D}$

An indentifiability algorithm

- Take design points, construct design-model matrix
- Sort out columns (terms) in the design-model matrix
- Identify the first n linearly independent columns



Design $\mathcal{D}$

Point\term	1	x_2	x_1	x_2^2	x_1x_2	x_1^2	x_2^3	$\cdots$
$(0,0)$	1	0	0	0	0	0	0	$\cdots$
$(1,0)$	1	0	1	0	0	1	0	
$(0,1)$	1	1	0	1	0	0	1	$\cdots$
$(1,-1)$	1	-1	1	1	-1	1	-1	
$(-1,1)$	1	1	-1	1	-1	1	1	$\cdots$

Linear independence with a term ordering!
Technique applicable to a wide range of designs: LHS, factorial, optimal...

Modelling the response, the smooth approach

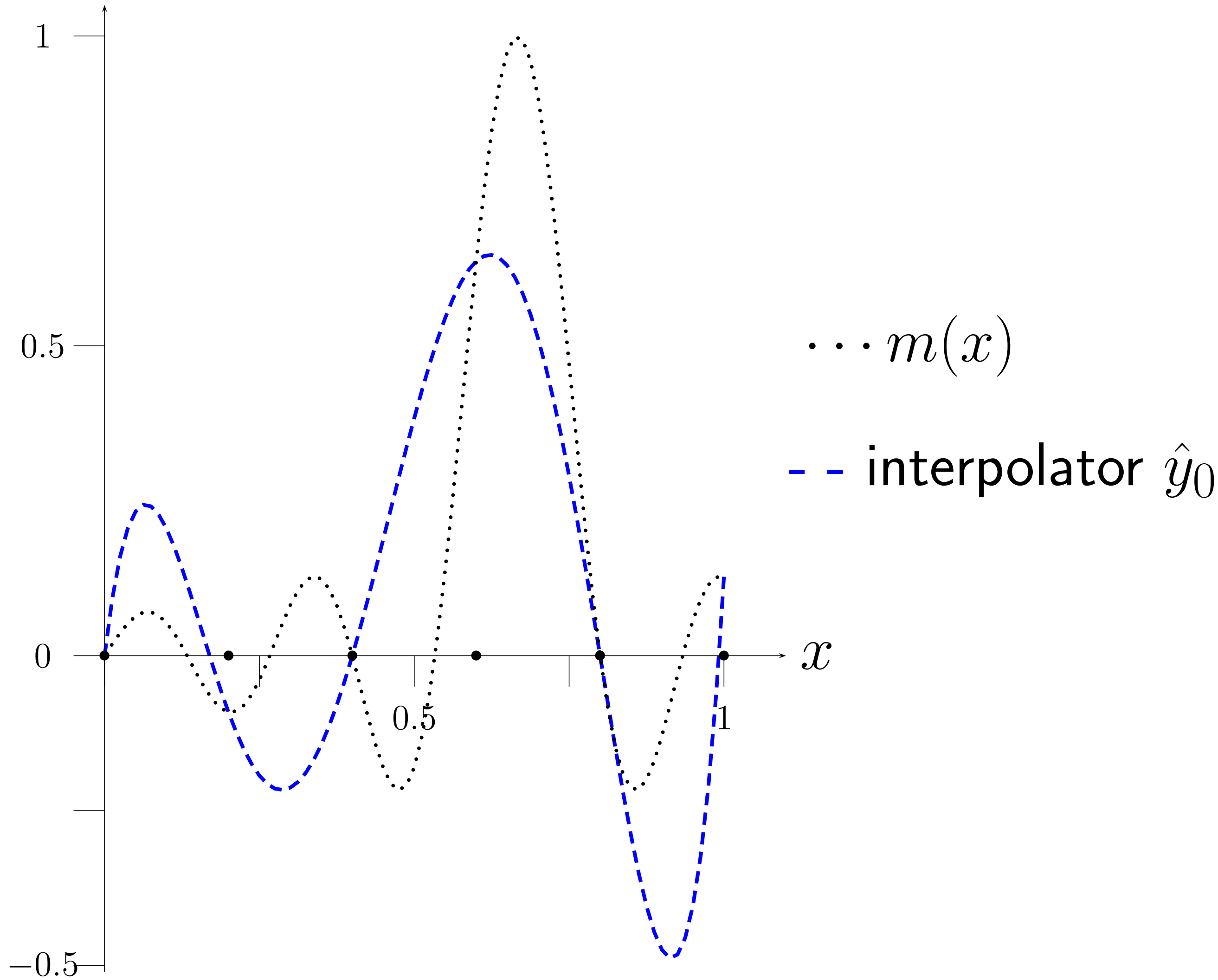
Once the experiments has been performed, the response (computer code output) is to be modelled.

For example, suppose the simulation model is

$$m(x) = \mathsf{sinc}(ax + b) = \sin(ax + b)/(ax + b)$$

with $a = 15\pi/2$ and $b = -10\pi/2$. The design is a uniform set of 6 points in $[0, 1]$.

A simple model is to just construct a polynomial interpolator for the response at the design points. $\hat{y}_0(x) = h(x)^T \hat{\beta}$ is a polynomial (fifth degree) which interpolates the response.



1
0.5
0
−0.5
0.5
1
x
$\cdots\, m(x)$
$- -$ interpolator $\hat{y}_0$

Smoothing the univariate interpolator

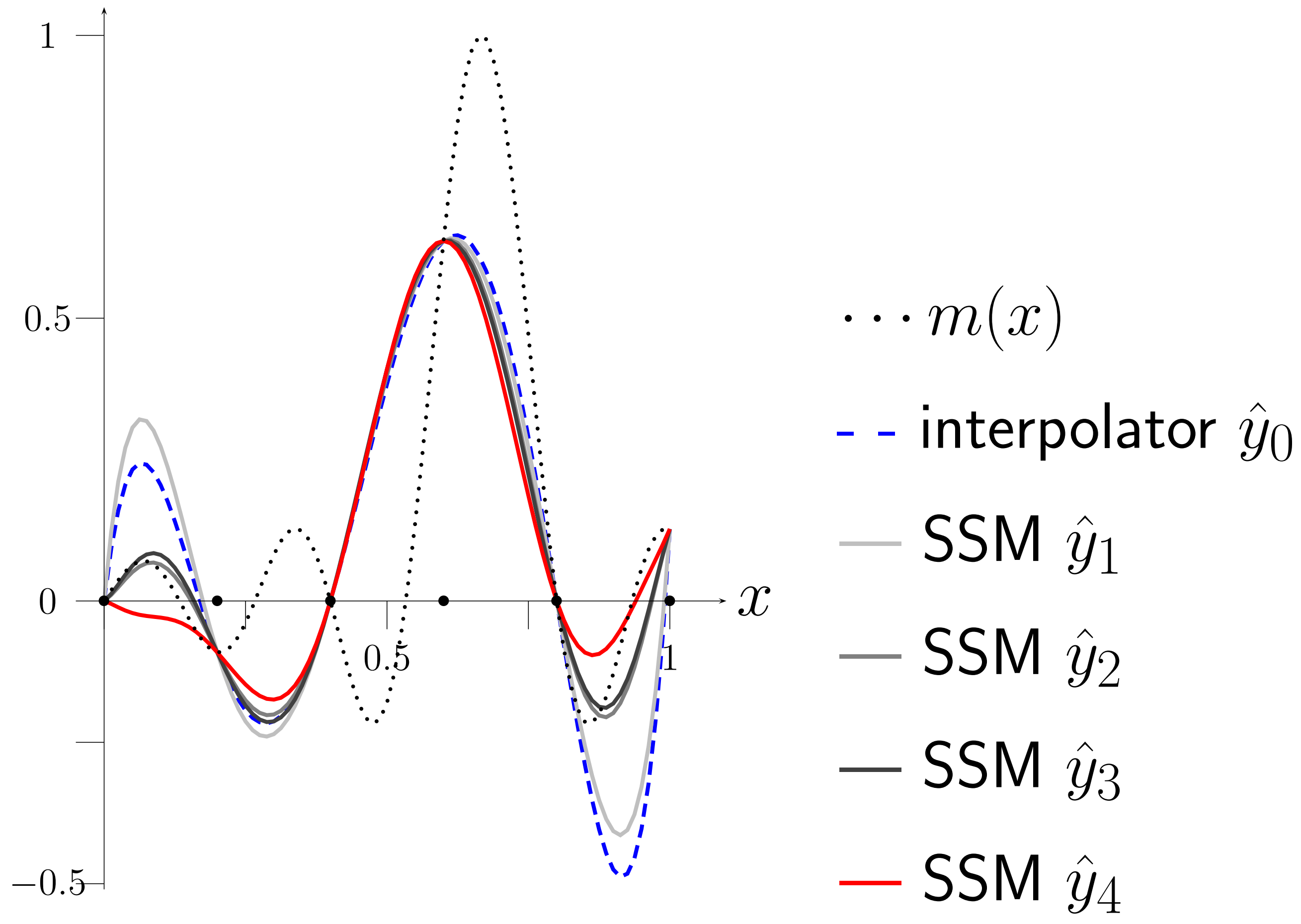
Proposal: smooth $\hat{y}_0(x)$ by adding a vector of extra terms $g(x)$, i.e. construct a supersaturated model $y(x) = h(x)^T\beta + g(x)^T\gamma$. Then select β, γ to minimise a measure of smoothness

$$\Psi_2(y) = \int_{\mathcal{X}} |y(x)''|^2 dx,$$

while keeping $y(x)$ an interpolator.

- $\mathcal{X}$ is the region over which the supersaturated interpolators will be "smoothed", which allows for flexibility.
- Quadratic problem in the parameters (β, γ), unique solution
- Smooth supersaturated model (SSM), linear in parameters and in observations
- Polynomial approximation to cubic spline

Model	$\hat{y}_0$	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$	$\hat{y}_5$	Spline
Ψ_2^*	**76.54**	**74.69**	**33.15**	**33.02**	**27.77**	**27.75**	**26.74**



Multivariate constrained optimisation

Supersaturated model $y(x) = f(x)^T \theta$, criterion

$$\Psi_2(y) = \int_{\mathcal{X}} \mathrm{tr}(H(y)^2) dx = \theta^T K \theta$$

with $K = \int_{\mathcal{X}} \left(\sum_{i,j=1}^{d} f^{(ij)} f^{(ij)^T} \right) dx$.

$$\min \Psi_2(y(x)) \text{ subject to } X\theta = y$$

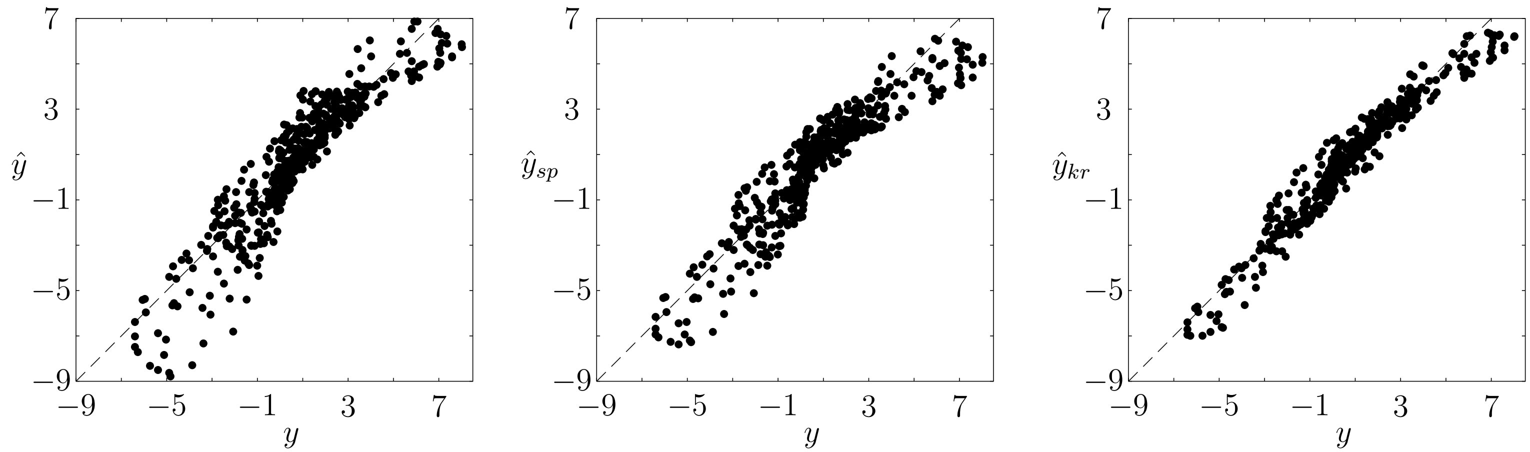
If the model has constant and linear terms, then K has *structural zeros* i.e. $K = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{K} \end{bmatrix}$. General closed form solution:

$$\begin{bmatrix} X_0 & X_1 & 0 \\ 0 & \tilde{K} & -X_1^T \\ 0 & 0 & X_0^T \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \lambda \end{bmatrix} = \begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} \tag{1}$$

Bidimensional example

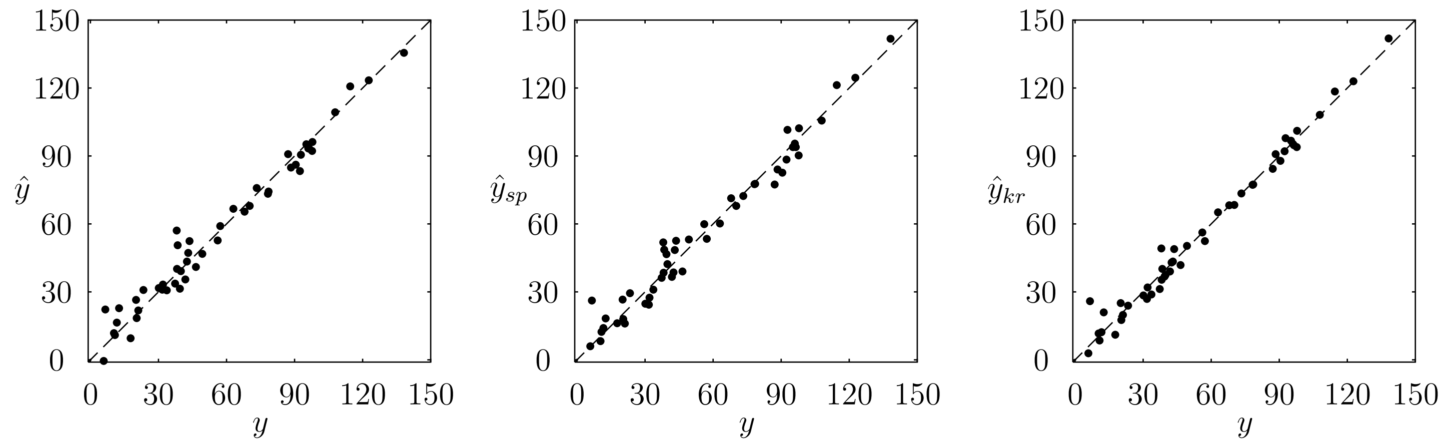
The function `peaks` (MatLab) was simulated at a space filling (LP_τ) design with $n = 24$. Three interpolation schemes were fitted: $\hat{y}$ (SSM); $\hat{y}_{sp}$ (thin plate spline); $\hat{y}_{kr}$ (kriging with $\mathrm{Cov}(Z(r), Z(s)) = \exp(\sum_{i=1}^{2} \theta_i |r_i - s_i|^{p_i})$)

500 further observations were generated to validate and compare.



Engine emissions data set (Bates et al. (2003))

48 observations at a LHS in factors N, C, A, B, M, smoothed with 75 extra terms. An extra set of 49 observations was available for validation.



Model	$\hat{y}$	$\hat{y}_{sp}$	$\hat{y}_{kr}$
RMSE	5.844	5.896	4.450
% response range	4.4	4.5	3.4

Sensitivity analysis: Sobol' indices

An integrable function f defined on $[0,1]^d$ can be decomposed as the sum of univariate effects, bivariate interactions, triple interactions...

Example (trivariate f):

$$f = f_0 + f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} + f_{123}$$

with $f_0 = \int f dx$, $f_1 = \int f dx_{23} - f_0$, $f_2 = \int f dx_{13} - f_0$,
$f_3 = \int f dx_{12} - f_0$, $f_{12} = \int f dx_3 - f_0 - f_1 - f_2$,
$f_{23} = \int f dx_1 - f_0 - f_2 - f_3, \ldots$

- Total variance $D = \int f^2 dx - f_0^2$, and $D_1 = \int f_1^2 dx_1$, $D_2 = \int f_2^2 dx_2$, $D_3 = \int f_3^2 dx_3$, $D_{12} = \int f_{12}^2 dx12, \ldots$

$$D = D_1 + D_2 + D_3 + D_{12} + D_{13} + D_{23} + D_{123}$$

- Global sensitivity indices $S_1 = D_1/D$, $S_2 = D_2/D$, $S_3 = D_3/D$, $S_{12} = D_{12}/D, \ldots$ and (for the trivariate example)

$$S_1 + S_2 + S_3 + S_{12} + S_{13} + S_{23} + S_{123} = 1$$

- Uses:
  - Ranking of effects,
  - screening (fixing unessential variables, deleting high order terms),
  - assessing importance of interaction terms

Global sensitivity analysis using SSM

- A SSM can be constructed using data from the simulator
- Global sensitivity analysis computed with the SSM
- SSM is a polynomial function, linear on parameters and on observations and thus the computations are quickly performed
- Sensitivity indices, plots of main effects, screening, ranking of variables

Engine emissions data set (Bates et al. (2003)

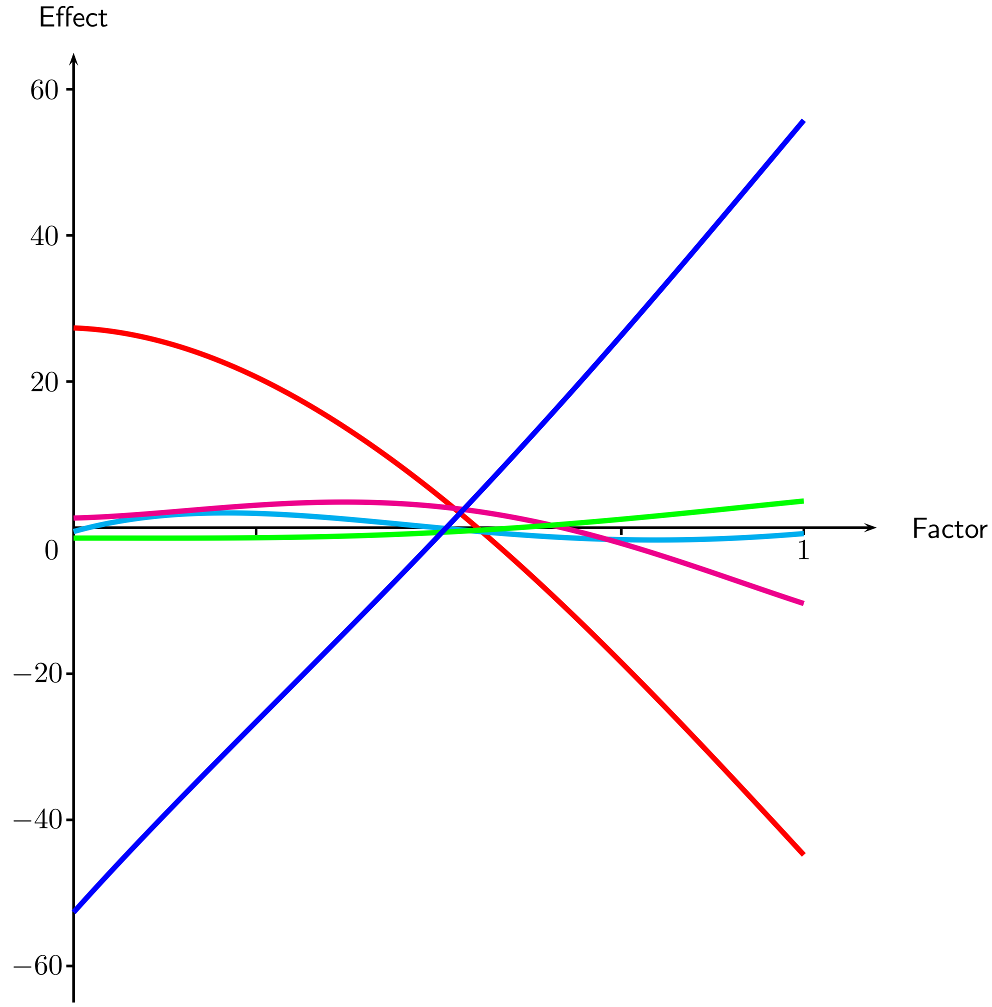
- Total variance $D = 1494.54$.
- Global sensitivity indices $(\%)$

Univariate Total $= 97.93$					Double interactions Total $= 1.58$									
S_N	S_C	S_A	S_B	S_M	S_{NC}	S_{NA}	S_{NB}	S_{NM}	S_{CA}	S_{CB}	S_{CM}	S_{AB}	S_{AM}	S_{BM}
32.68	0.12	1.04	0.16	63.90	0.06	0.13	0.36	0.56	0.07	0.01	0.03	0.02	0.09	0.20

- Univariate plus double interactions account for 99.51 $(\%)$ of D.
- The remaining variation (0.485%) is explained by triple interactions (10), quadruple interactions (5) and a quintuple interaction.
- Main effects: M (linear) and N, also A has a (small) effect. There is an interaction effect MN (small).

Univariate effects plot

Key: N, C, A, B, M



Extensions: Smooth Hermite interpolation

Often, derivatives of the output are available as part of the simulation process (use of adjoint).

We can also use derivative information, i.e. to interpolate derivative values: Hermite interpolation.

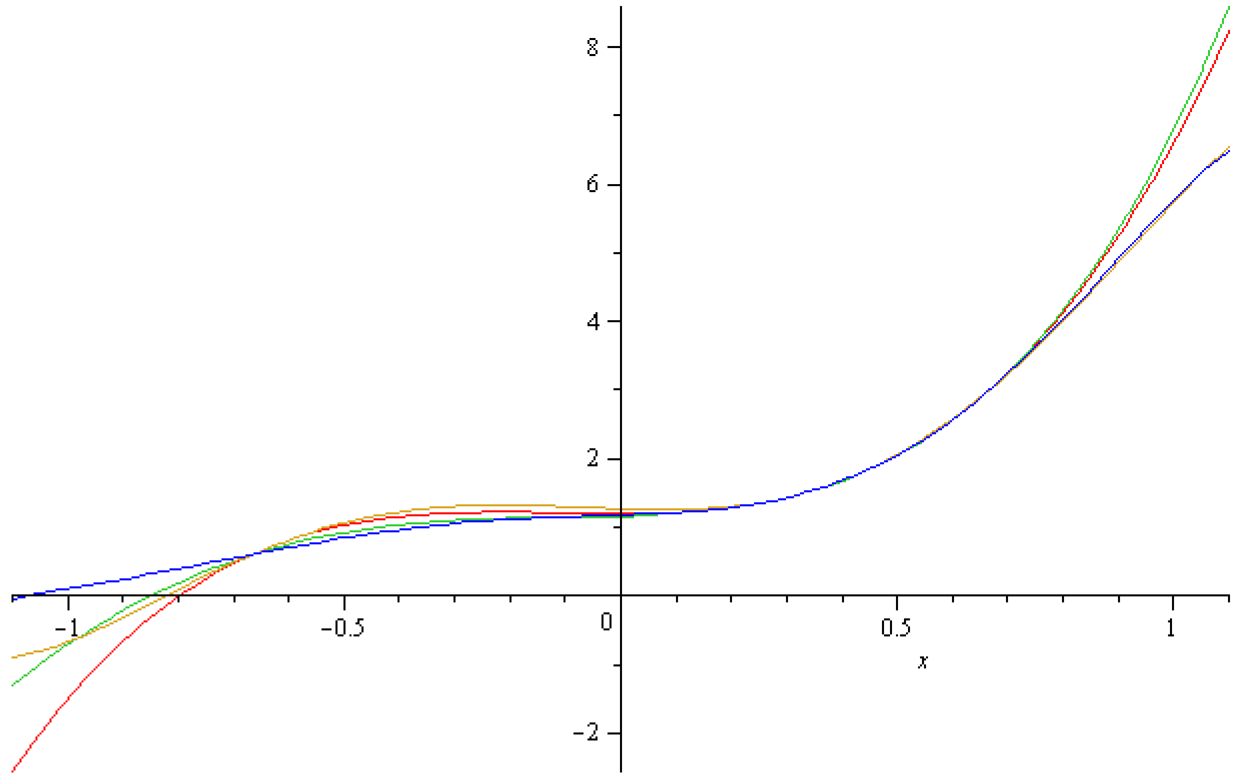
- Supersaturate the interpolator
- Quadratic problem with restrictions

$$\underbrace{\min \Psi_2(y(x))}_{\text{smoothness}} \text{ subject to } \underbrace{X\theta = y}_{\text{interpolation}} \text{ and } \underbrace{X'\theta = y'}_{\text{derivatives}}$$

- Smooth Hermite interpolator
- Not restricted to only first derivative, interpolation of higher derivatives

Example: Smooth supersaturated Hermite

$f(x) = \frac{1}{1-x}$; Observations at $\pm\frac{2}{3}$, "fat point" at $\frac{1}{3}$.



	m_0	m_1	m_2	m_3
IMSE	0.429	0.178	0.281	0.056

Final remarks on SSM

- High degree polynomial models which are tailored to be smooth (contrary to intuition that high degree models tend to oscillate)
- Methodology directly extended into higher dimensions.
- The smoothing region $\mathcal{X}$ can be placed where desired, e.g. experiment in a different region than the region for smoothing the model.
- One can get arbitrarily close to thin plate splines: further work to prove this property.
- SSMs are linear in the parameters and in the observations, e.g. easy for differentiation on sensitivity analysis and turning points.
- Use of derivative information: Smooth supersaturated Hermite interpolation.
- *Smooth supersaturated models*. MUCM internal report 08/01.

References

Becker, Weispfenning (1991). *ISSAC '91*, 64-69.

Bates *et al.* (2003). *Techno*. 45(3), 246-255.

Bates *et al.* (2008). *MUCM IR* 08/01.

Holladay (1957). *MTAC* 11(3), 233-243.

Kimeldorf, Wabba (1970). *Ann. Stat.* 41,495-502.

Micula (2002). *Gen. Math.* 10(1-2),21-50.

Pistone and Wynn (1996). *Biometrika* **83**(3), 653-666.

Pistone *et al.* (2000). *Algebraic Statistics*.

Sobol' (2001). *Maths. Comp. Sim.* **55**, 271-280.

Thanks!

Contact details

H.Maruri-Aguilar@lse.ac.uk

Managing Uncertainty in Computer Models (MUCM project): Modelling computer simulations

- Computer simulations are increasingly used to study scientific problems (e.g. climate forecast, engineering problems). There is growing interest in modeling computer simulations to answer scientific questions.
- MUCM project (`http://www.mucm.group.shef.ac.uk/`) aims to develop state of the art technology to understand how models perform.
- Five UK universities: Aston, Sheffield, Durham, NOC and LSE. 5 leading researchers, 7 RAs and 4 PhD students.
- Within LSE, we work on the experimental design component of MUCM. We both develop new technology and make available already existant technology.