Surrogate modeling based on resampled polynomial chaos expansions

Zicheng Liu[1]

Dominique Lesselier[2], Joe Wiart[1]

w. thanks to Bruno Sudret at ETH Zürich, Switzerland

[1]Chaire C2M, LTCI, Télécom Paris, France

[2]Laboratoire des Signaux et Systèmes , UMR8506 (CNRS-CentraleSupélec-Université Paris-Sud), Université Paris-Saclay, France

Outline

- Polynomial chaos expansion (PCE)
- Sparse polynomial chaos expansion
- Resampled polynomial chaos expansion (rPCE)
- Global sensitivity analysis by Sobol' indices
- Application examples
- Conclusions and perspectives

Polynomial chaos expansion (PCE)

$$f\left(\boldsymbol{x}^{(n)}\right) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} \beta_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}^{(n)})\,, \boldsymbol{x}^{(n)} \in \mathbb{R}^M, n = 1, \dots, N$$

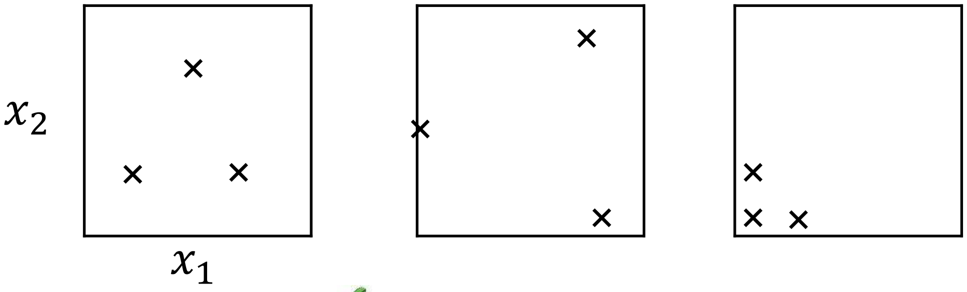
- f: function which represents for the physical system and often computed by numerical methods (e.g., FDTD, FEM) with high computational costs.
- $\psi_{\boldsymbol{\alpha}}$: basis polynomial
- $\beta_{\boldsymbol{\alpha}}$: expansion coefficient
- $\boldsymbol{\alpha}$: vector of order for multivariate polynomial (e.g., $\boldsymbol{\alpha} = (1,2)$ for $x_1x_2^2$)
- $\boldsymbol{x}^{(n)}$: sample of the input space. $\left(\boldsymbol{x}^{(n)}, f\left(\boldsymbol{x}^{(n)}\right)\right)$ composes the experimental design (ED).
- M: number of input parameters
- N: number of samples in ED

Sampling method

$$f\left(\boldsymbol{x}^{(n)}\right) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} \beta_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}\left(\boldsymbol{x}^{(n)}\right), \boxed{\boldsymbol{x}^{(n)}} \in \mathbb{R}^M, n = 1, \dots, N$$

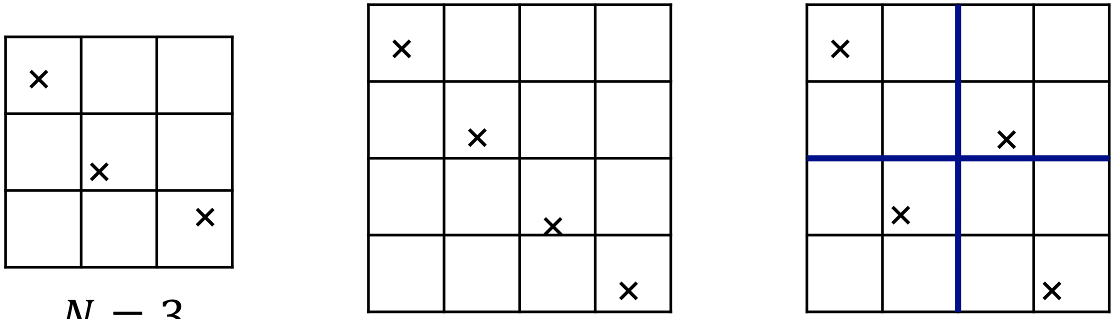
random sampling

$M = 2, N = 3$



Latin hypercube sampling (LHS) ✓

$M = 2$



$N = 3$

$N = 4$

$N = 4$

orthogonal sampling

Expansion basis

$$f\left(\boldsymbol{x}^{(n)}\right) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} \beta_{\boldsymbol{\alpha}} \boxed{\psi_{\boldsymbol{\alpha}}}(\boldsymbol{x}^{(n)}), \boldsymbol{x}^{(n)} \in \mathbb{R}^M, n = 1, \ldots, N$$

Supporting basis $\psi_{\boldsymbol{\alpha}}$ is decided by **orthogonality** and **order $\boldsymbol{\alpha}$**.

$\psi_{\boldsymbol{\alpha}}$ is a basis in a Hilbert space equipped with the inner product:

$$< f, g > = E[f(\boldsymbol{X})g(\boldsymbol{X})] = \int_{\mathbb{X}} f(\boldsymbol{x})g(\boldsymbol{x})p_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}$$

$p_{\boldsymbol{X}}$ joint probability density function (PDF) of random vector $\boldsymbol{X} = [X_1, \ldots, X_M]$.

The orthogonality of basis polynomials defined by

$$< \psi_{\boldsymbol{\alpha}}, \psi_{\boldsymbol{\gamma}} > = E\left[\psi_{\boldsymbol{\alpha}}(\boldsymbol{X})\psi_{\boldsymbol{\gamma}}(\boldsymbol{X})\right] = \delta_{\boldsymbol{\alpha},\boldsymbol{\gamma}}$$

$\delta_{\boldsymbol{\alpha},\boldsymbol{\gamma}} = 1$ if $\boldsymbol{\alpha} = \boldsymbol{\gamma}$, $= 0$ otherwise.

Expansion basis

Assuming $X_m, m = 1, \ldots, M$, are independent, i.e.,

$$p_{\boldsymbol{X}}(\boldsymbol{X}) = p_{X_1}(X_1) \times \cdots \times p_{X_M}(X_M)$$

p_{X_m} marginal PDF, $\psi_{\boldsymbol{\alpha}}$ tensor product of univariate polynomial $\pi_{\alpha_m}(X_m)$, i.e.,

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) = \pi_{\alpha_1}(X_1) \times \cdots \times \pi_{\alpha_M}(X_M)$$

Not hard to conclude that if π_{α_m} satisfies

$$< \pi_{\alpha_j}, \pi_{\alpha_k} > = \int_{\mathbb{X}_m} \pi_{\alpha_j}(x_m)\pi_{\alpha_k}(x_m)p_{X_m}(x_m)dX_m = \delta_{j,k}$$

orthogonality of $\psi_{\boldsymbol{\alpha}}$ is guaranteed.

PDF of X_m	polynomial family of π_{α_m}
Uniform distribution	Legendre polynomial
Gaussian distribution	Hermite polynomial

... ...

Expansion basis

$$f\left(\boldsymbol{x}^{(n)}\right) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} \beta_{\boldsymbol{\alpha}} \boxed{\psi_{\boldsymbol{\alpha}}}(\boldsymbol{x}^{(n)}), \boldsymbol{x}^{(n)} \in \mathbb{R}^M, n = 1, \dots, N$$

$\psi_{\boldsymbol{\alpha}}$ is decided by **orthogonality** and **order** $\boldsymbol{\alpha}$.

infinite series $\sum_{\boldsymbol{\alpha}\in\mathbb{N}^M}$ $\Longrightarrow$ truncated PCE $\sum_{\boldsymbol{\alpha}\in\mathbb{A}}$

How to decide $\mathbb{A}$? The commonly utilized **full model** follows

$$\mathbb{A}^{\text{full}} = \{\boldsymbol{\alpha}: \textstyle\sum_{m=1}^{M} \alpha_m \leq p\}$$

However, the cardinality of $\mathbb{A}^{\text{full}}$

$$\text{card}\left(\mathbb{A}^{\text{full}}\right) = \binom{p+M}{p}$$

will polynomially increases with p and M.

As a result, surrogate modeling suffers from *curse of dimensionality*, i.e., large ED required w. large M and p to avoid the overfitting phenomena.

Expansion coefficients

$$\hat{f}\left(\boldsymbol{x}^{(n)}\right) = \sum_{\boldsymbol{\alpha}\in\mathbb{A}} \boxed{\beta_{\boldsymbol{\alpha}}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}^{(n)}), \boldsymbol{x}^{(n)} \in \mathbb{R}^M, n = 1, \dots, N$$

Projection method:

Due to orthogonality of basis polynomials,

$$\beta_{\boldsymbol{\alpha}} = \int_{\mathbb{X}} f(\boldsymbol{x})\,\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}$$

Integral is numerically computed.

Regression approach: ✓

$\beta_{\boldsymbol{\alpha}}$ solution of minimization problem

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} E[(f(\boldsymbol{X}) - \boldsymbol{\Psi}(\boldsymbol{X})\boldsymbol{\beta})^2]$$

matrix $\boldsymbol{\Psi} = [\psi_{\boldsymbol{\alpha}}]$ and column vector $\boldsymbol{\beta} = [\beta_{\boldsymbol{\alpha}}]$.

Based on ED,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-\mathbf{1}}\boldsymbol{\Psi}^T\boldsymbol{y}$$ *ordinary least square* (OLS)

column vector $\boldsymbol{y} = [f(\boldsymbol{x}^{(n)})]$.

Estimation of prediction performance

Generalization error: $\mathrm{Err} = E\left[\left(f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})\right)^2\right]$ $\boldsymbol{X}$ random vector of inputs

If a large set of data is available for validation,

$$\mathrm{Err} \approx \epsilon_{\mathrm{val}} = \frac{1}{N_{\mathrm{val}}} \sum_{n=1}^{N_{\mathrm{val}}} \left(f\left(\boldsymbol{x}^{(n)}\right) - \hat{f}\left(\boldsymbol{x}^{(n)}\right)\right)^2$$

Otherwise, **cross-validation** (CV) is recommended.

Cross-validation:

Validation

Training

Leave-**o**ne-**o**ut (LOO) cross-validation: $\epsilon_{\mathrm{LOO}} = \frac{1}{N} \sum_{n=1}^{N} \left(f\left(\boldsymbol{x}^{(n)}\right) - \boxed{\hat{f}^{-(n)}}\left(\boldsymbol{x}^{(n)}\right)\right)^2$

PCE model built by leaving n-th sample out for validation

ϵ_{LOO} can be computed fast in single training process based on $\hat{f}$.

Assuming candidate models $\hat{f}_i$ are available,

$$\hat{f}^* = \arg\min_{\hat{f}_i} \epsilon_{\mathrm{LOO}}(f, \hat{f}_i)$$

Sparse polynomial chaos expansion

$$\hat{f}\left(\boldsymbol{x}^{(n)}\right) = \sum_{\boldsymbol{\alpha}\in\boxed{\mathbb{A}}} \beta_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{x}^{(n)})\,, \boldsymbol{x}^{(n)} \in \mathbb{R}^M, n = 1, \dots, N$$

Not all basis polynomials $\boldsymbol{\psi}_{\alpha}, \boldsymbol{\alpha} \in \mathbb{A}^{\text{full}}$, are influential. Greedy algorithms **LARS** (least angle regression) and **OMP** (orthogonal matching pursuit) are popularly used to select the most relevant basis polynomials.

Sparse PCE model based on OMP

```
1. Initialization: residual $\boldsymbol{R}_0 = \boldsymbol{y}$, active set $\mathbb{A}_0^a = \emptyset$,  candidate set $\mathbb{A}_0^c = \mathbb{A}^{\text{full}}$
2. For $j = 1, \dots, \min\{N-1, \mathrm{card}(\mathbb{A}^{\text{full}})\}$
        1) Find $\boldsymbol{\psi}_{\boldsymbol{\alpha}_j}$ most correlated with $\boldsymbol{R}_{j-1}$,  $\psi_{\boldsymbol{\alpha}_j} = \arg\max_{\boldsymbol{\alpha}\in\mathbb{A}_{j-1}^c} \left|\boldsymbol{R}_{j-1}^T \boldsymbol{\psi}_{\boldsymbol{\alpha}}\right|$.
        2) Update $\mathbb{A}_j^a = \mathbb{A}_{j-1}^a \cup \boldsymbol{\alpha}_j$ and $\mathbb{A}_j^c = \mathbb{A}_{j-1}^c \setminus \boldsymbol{\alpha}_j$.
        3) With $\boldsymbol{\psi}_{\mathbb{A}_j^a}$, compute $\boldsymbol{\beta}_j$ as the ordinary least square solution.
        4) Update residual $\boldsymbol{R}_j = \boldsymbol{y} - \boldsymbol{\psi}_{\mathbb{A}_j^a} \boldsymbol{\beta}_j$ and compute associated $\epsilon_{LOO}^j$.
   End
3. $\psi_{\mathbb{A}_j^a}$ with smallest $\epsilon_{LOO}$ is selected as best sparse basis.
```

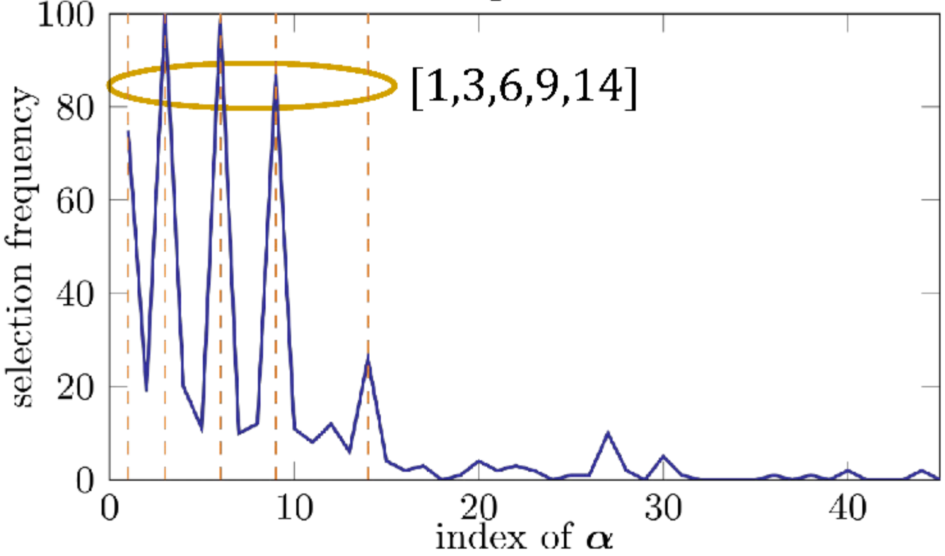
Sparse PCE model based on LARS runs similar procedure but less greedy than OMP.

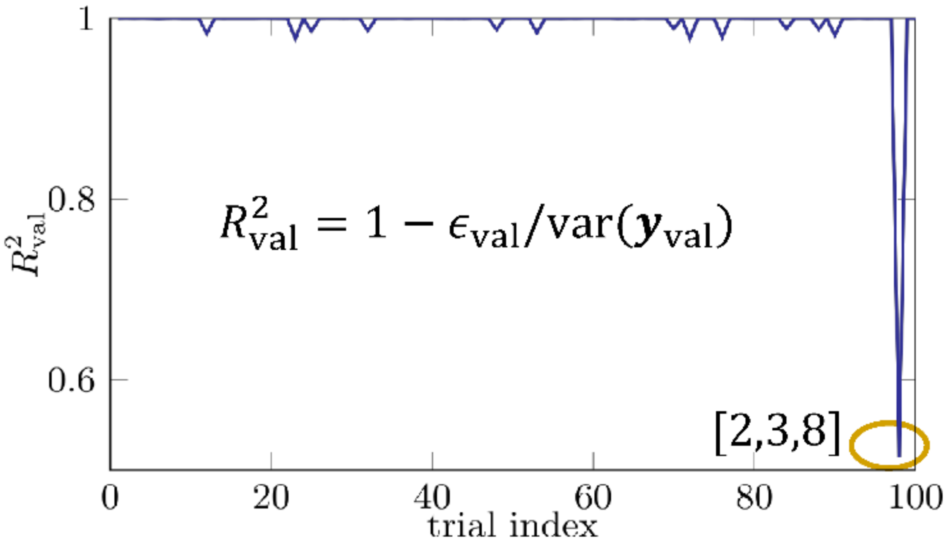
Idea of resampled PCE (rPCE)

As a result, **the basis polynomials of true model will be frequently selected during surrogate modeling with varied data,** as shown by the following example.

$$f(\boldsymbol{X}) = 1 + X_1 + X_1X_2 + X_1X_2^2 + X_1X_2^3, \qquad X_1 \sim N(0,1) \text{ and } X_2 \sim N(6,1)$$

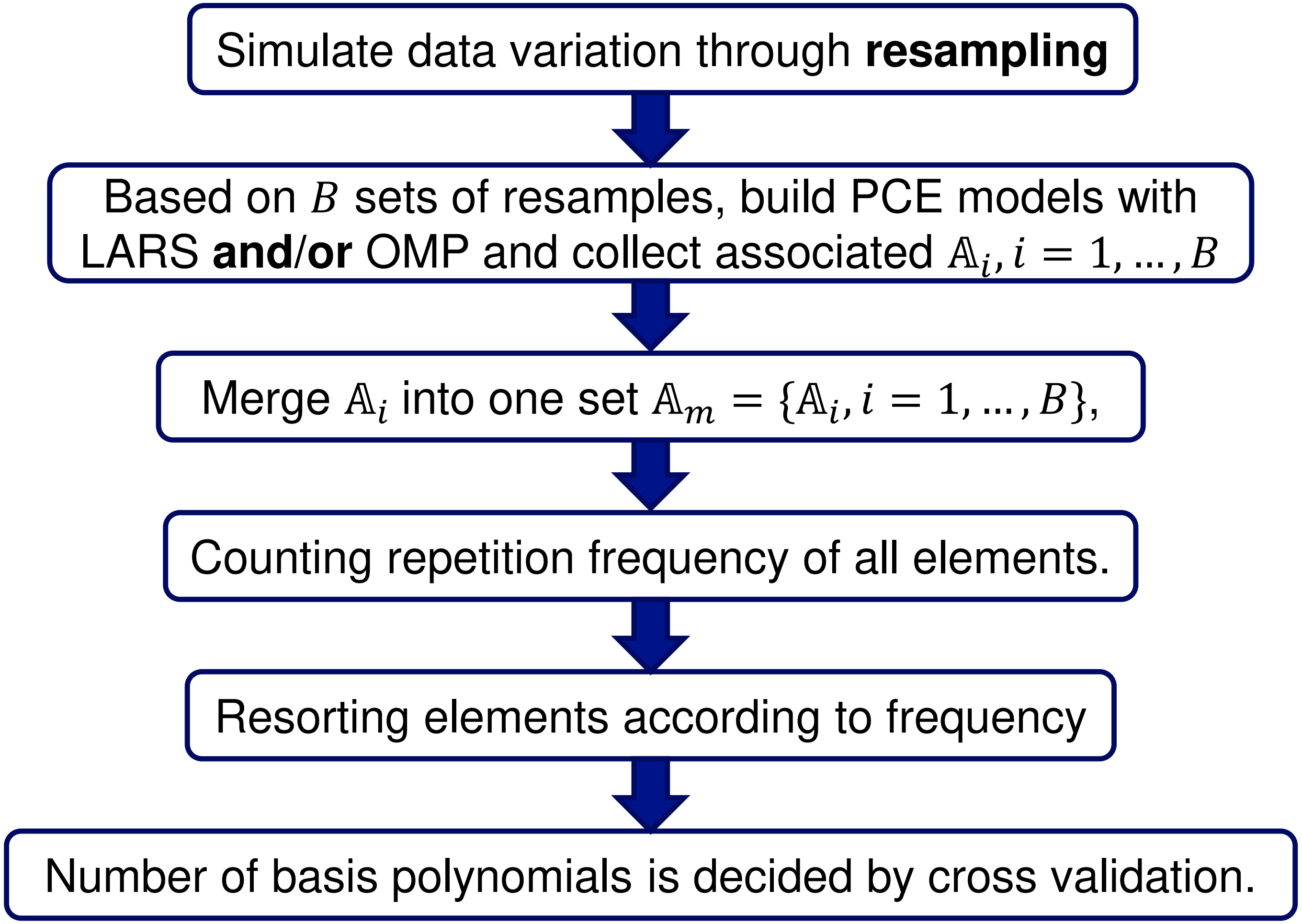
12 data for training and 10^4 data for validation, 100 trials are performed with OMP-based PCE modeling. At each trial, data are obtained through LHS method.





Selecting the optimal $\boldsymbol{\alpha}$ (associated basis polynomial) as most frequent ones might improve the performance of PCE models.

rPCE: procedure



Simulate data variation through **resampling**

Based on B sets of resamples, build PCE models with LARS **and/or** OMP and collect associated $\mathbb{A}_i, i = 1, \dots, B$

Merge $\mathbb{A}_i$ into one set $\mathbb{A}_m = \{\mathbb{A}_i, i = 1, \dots, B\}$,

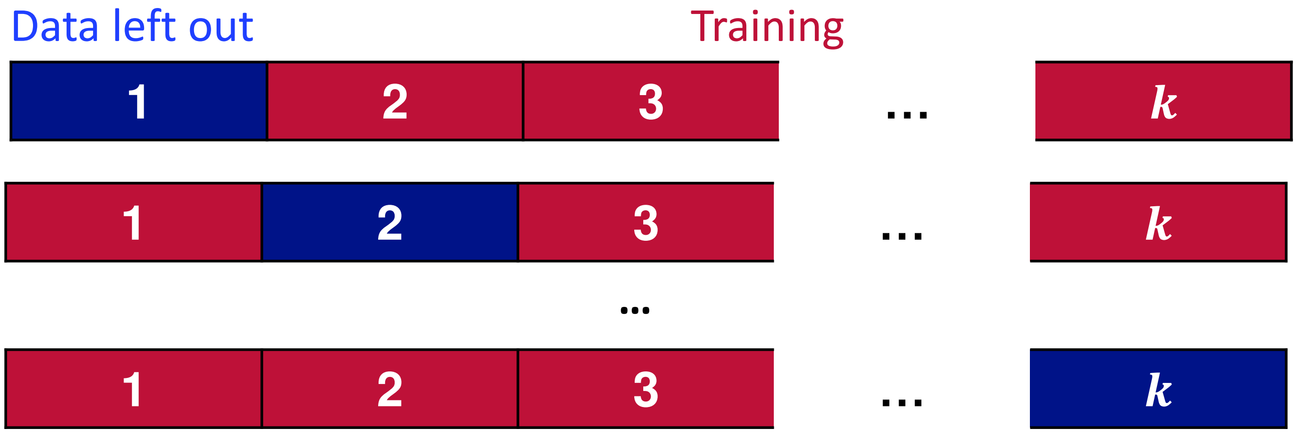
Counting repetition frequency of all elements.

Resorting elements according to frequency

Number of basis polynomials is decided by cross validation.

rPCE: data variation

Data variation is simulated by applying resampling technique - ***k*-fold division**.



How to decide the value of k?

Small k (e.g. 2) ⇒ biased basis polynomials

Large k (e.g. N) ⇒ high correlation of training data sets

The **suggested** configuration, rather than a single value,

$$k = \{3,5,10,20, N\}$$

which is a set of recommended values in literature.

rPCE: generation of candidate polynomials

Three options are available: **LARS**, **OMP**, or **their combination**, and one needs to decide which is the optimal.

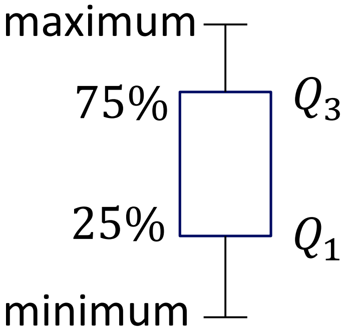
From the observation of simulations, one finds that

If LARS performs "**much better**" than OMP, one should choose LARS, and vice versa. Otherwise, the combination scheme is used.

How to properly define the criterion of "**much better**"?

Based on resamples, PCE models are constructed with LARS and OMP. Accordingly, one obtains corresponding values of R^2_{val}.

box plot of R^2_{val}



If $Q_1^{\text{LARS}} > Q_3^{\text{OMP}}$, LARS performs much better than OMP and generates candidate polynomials to rPCE, and vice versa. Otherwise, the combination is chosen.

Global sensitivity analysis by Sobol' indices

$\hat{f}(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in\mathbb{A}} \beta_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{x})$ is reformulated as

$$\beta_{\mathbf{0}} + \sum_{i=1}^{M} \sum_{\boldsymbol{\alpha}\in\mathbb{A}_{\{i\}}} \beta_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(x_i) + \sum_{1\leq i<j\leq M} \sum_{\boldsymbol{\alpha}\in\mathbb{A}_{\{i,j\}}} \beta_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(x_i, x_j) + \cdots + \sum_{\boldsymbol{\alpha}\in\mathbb{A}_{\{1,\ldots,M\}}} \beta_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(x_1, \ldots, x_M)$$

where

$$\mathbb{A}_{\{i_1,\ldots,i_s\}} = \{\boldsymbol{\alpha} \in \mathbb{A}, \alpha_k \neq 0 \text{ if } k \in \{i_1, \ldots, i_s\}; \alpha_k = 0 \text{ otherwise}\}, s \in \{1, \ldots, M\}$$

Orthogonality of basis polynomials gives the estimation of total and partial variances,

$$D = \sum_{\boldsymbol{\alpha}\in\mathbb{A}} \beta_{\boldsymbol{\alpha}}^2 - \beta_{\mathbf{0}}^2 \, , D_{i_1,\ldots,i_s} = \sum_{\boldsymbol{\alpha}\in\mathbb{A}_{\{i_1,\ldots,i_s\}}} \beta_{\boldsymbol{\alpha}}^2 - \beta_{\mathbf{0}}^2$$

and the ratio between them yields the Sobol' indices

$$S_{i_1,\ldots,i_s} = D_{i_1,\ldots,i_s}/D \qquad \boxed{\sum_{s=1}^{M} S_{i_1,\ldots,i_s} = 1}$$

Total Sobol' indices are defined as

$$S_i^T = \sum_{\mathbb{I}_i} S_{i_1,\ldots,i_s} \, , \mathbb{I}_i = \{\{i_1, \ldots, i_s\} \ni \{i\}\} \qquad \boxed{\sum_{i=1}^{M} S_i^T \geq 1}$$

Ishigami function: prediction

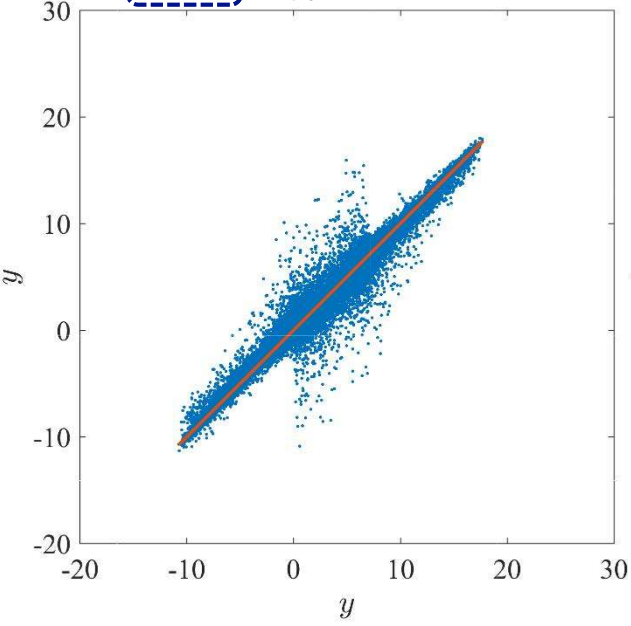
$$y(\boldsymbol{x}) = \sin(x_1) + a\sin^2(x_2) + bx_3^4\sin(x_1)$$

where $a = 7, b = 0.1, X_i$ are independent and uniform in $[-\pi, \pi], i = 1{,}2{,}3.$

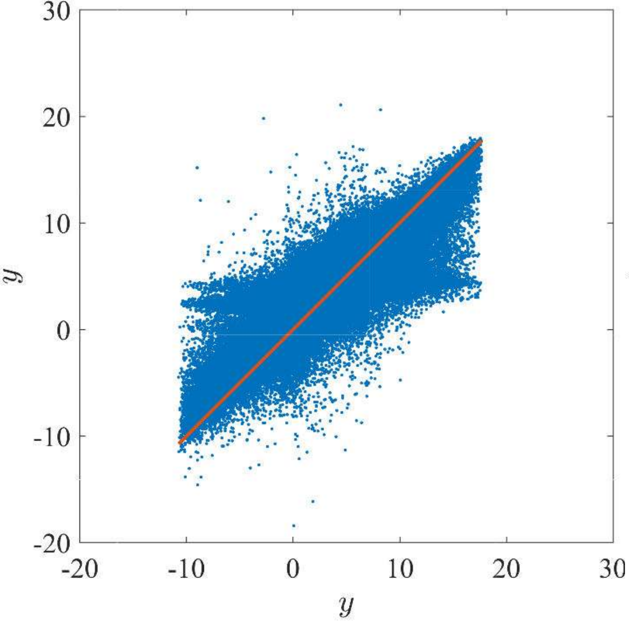
50 data for training and 10^4 data for validation, build PCE models based on LARS, OMP and rPCE. Repeating the above process 100 times, one has 10^6 prediction data.

suggested k and polynomial source at each trial

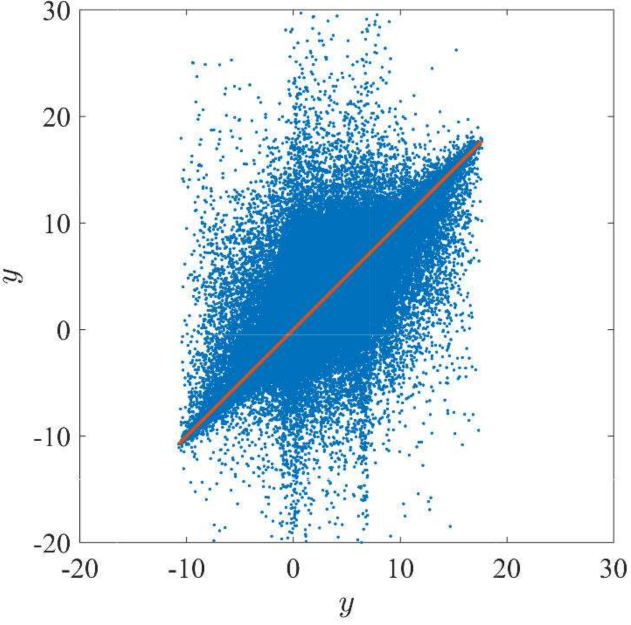
rPCE ($R^2_{\text{val}} = 0.9971$)



LARS ($R^2_{\text{val}} = 0.8724$)



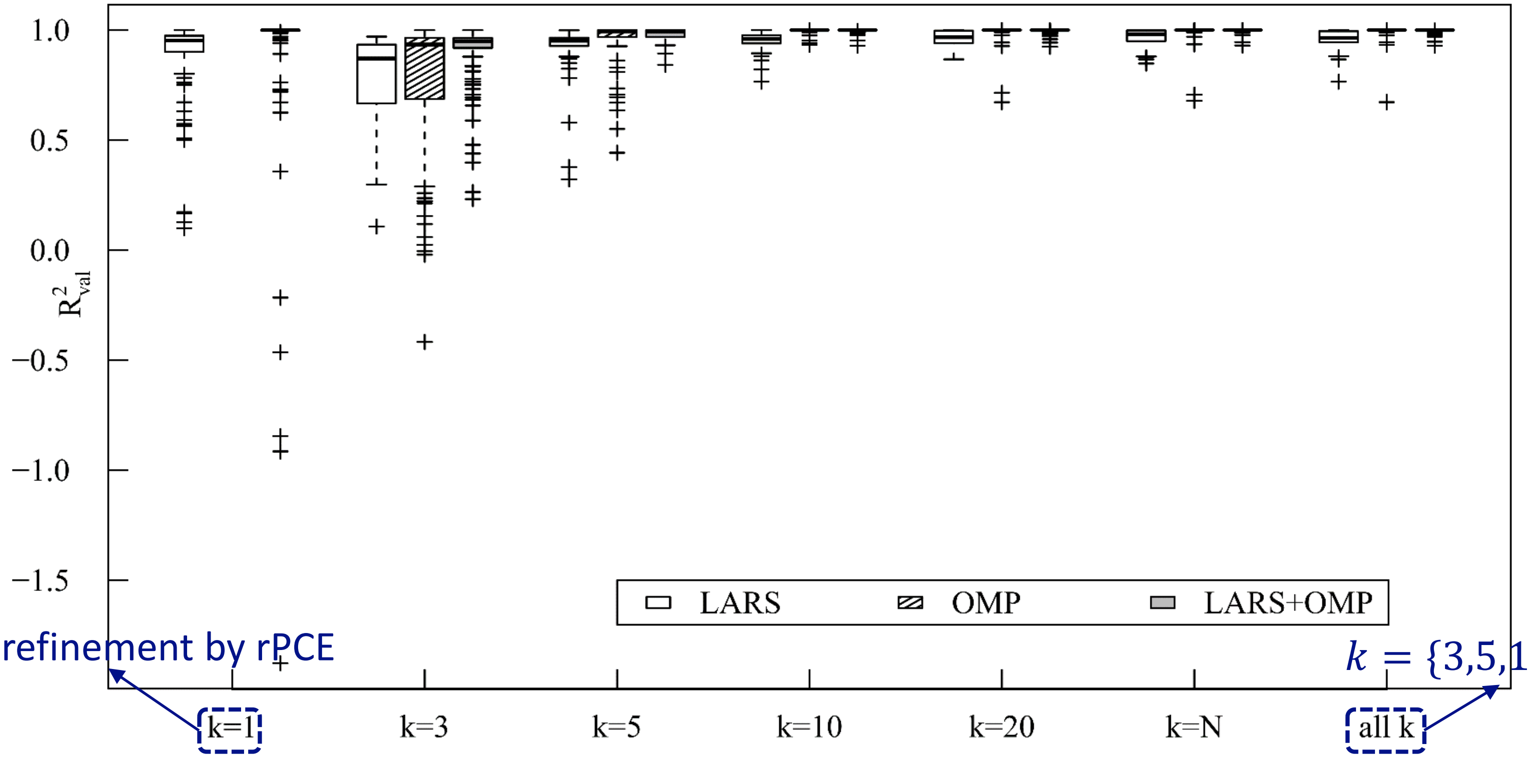
OMP ($R^2_{\text{val}} = 0.8790$)



Ishigami function: prediction

Mean of R_{val}^2 w.r.t. 100 replications

	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$	$k = N$	all k
LARS	0.8723	0.7890	0.9281	0.9542	0.9630	0.9686	0.9619
OMP	0.8788	0.7734	0.9566	0.9972	0.9919	0.9918	0.9947
LARS+OMP		0.8935	0.9817	0.9974	0.9969	0.9978	0.9971



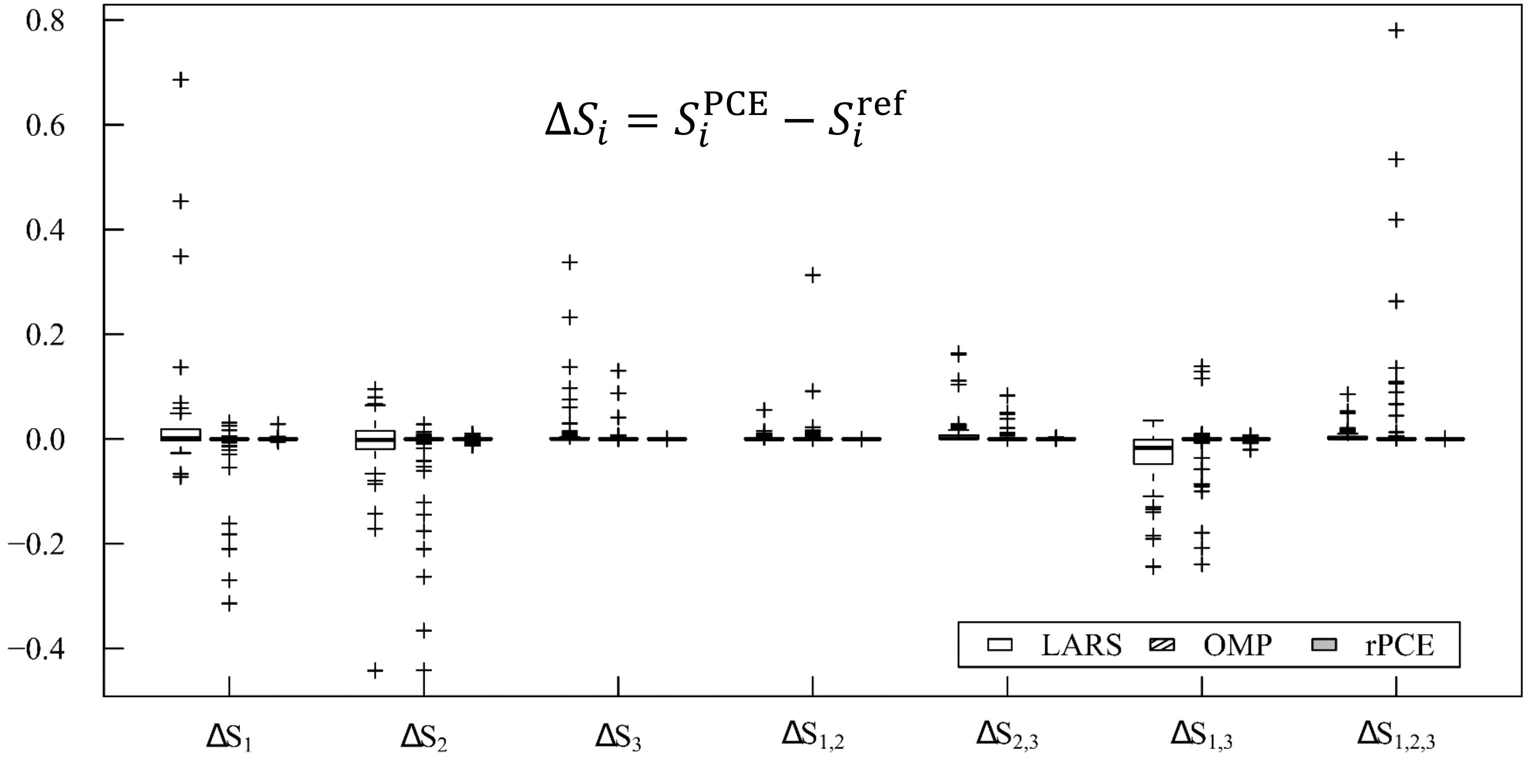
without refinement by rPCE

$k = \{3,5,10,20,N\}$

Ishigami function: Sobol' indices

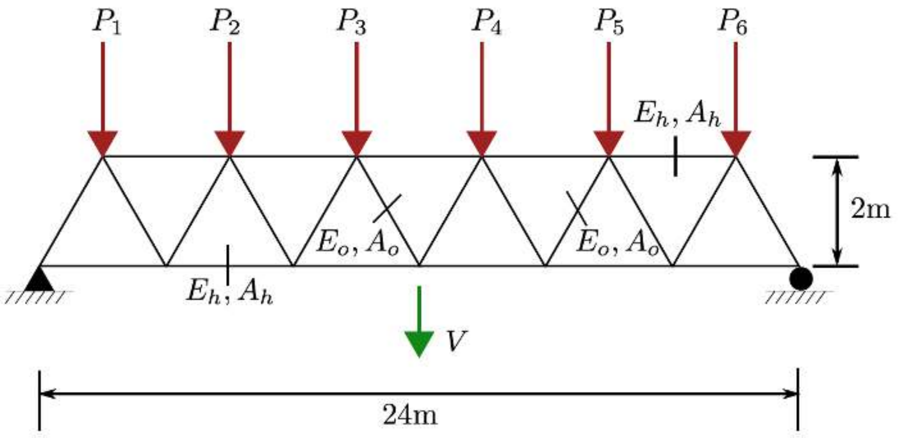
Mean of Sobol' indices w.r.t. 100 replications

	S_1	S_2	S_3	$S_{1,2}$	$S_{2,3}$	$S_{1,3}$	$S_{1,2,3}$
Reference	0.3139	0.4424	0.0000	0.0000	0.0000	0.2437	0.0000
rPCE	0.3141	0.4422	0.0000	0.0000	0.0001	0.2435	0.0001
LARS	0.3553	0.4152	0.0114	0.0017	0.0096	0.2019	0.0049
OMP	0.3017	0.4239	0.0028	0.0052	0.0042	0.2363	0.0258



$$\Delta S_i = S_i^{\text{PCE}} - S_i^{\text{ref}}$$

Maximum deflection of a truss structure



Six vertical loads denoted by $P_1 \sim P_6$ are put on a truss structure composed of 23 bars. The response quantity of interest, the mid-span deflection V, is computed with finite-element method (FEM).

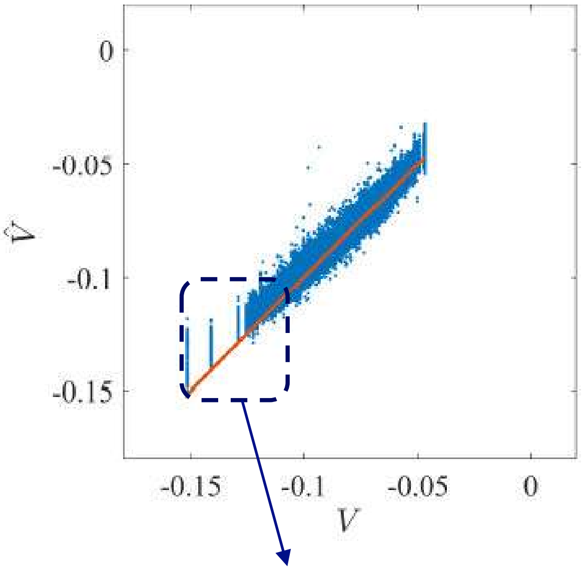
Variable	Distribution	Mean	Std	Description
E_h, E_o (Pa)	Lognormal	2.1×10^{11}	2.1×10^{10}	Young's moduli
A_h (m^2)	Lognormal	2.0×10^{-3}	2.0×10^{-4}	cross-section area of horizontal bars
A_o (m^2)	Lognormal	1.0×10^{-3}	1.0×10^{-4}	cross-section area of oblique bars
$P_1 \sim P_6$ (N)	Gumbel	5.0×10^4	7.5×10^3	vertical loads

Truss deflection: prediction

50 data for training and 10^4 data for validation, build PCE models based on LARS, OMP and rPCE. Repeating the above process 100 times, one has 10^6 prediction data.

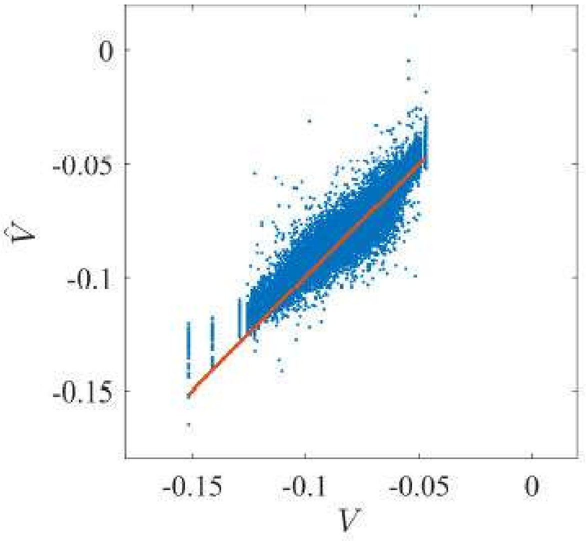
suggested k and polynomial source at each trial

rPCE ($R^2_{\mathrm{val}} = 0.9770$)

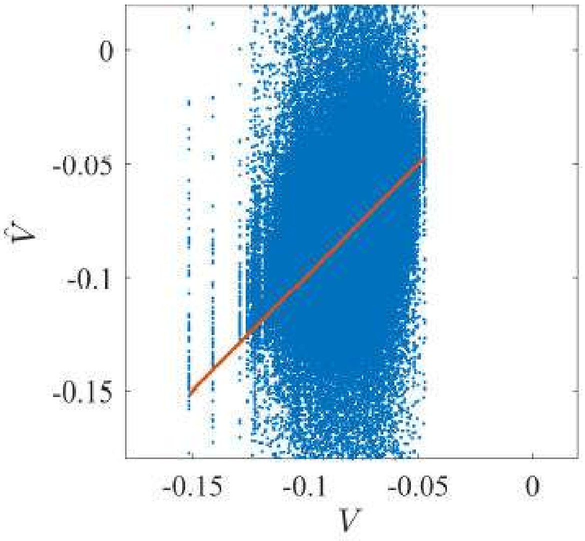


0.78% data with $V < -0.11$

LARS ($R^2_{\mathrm{val}} = 0.9631$)



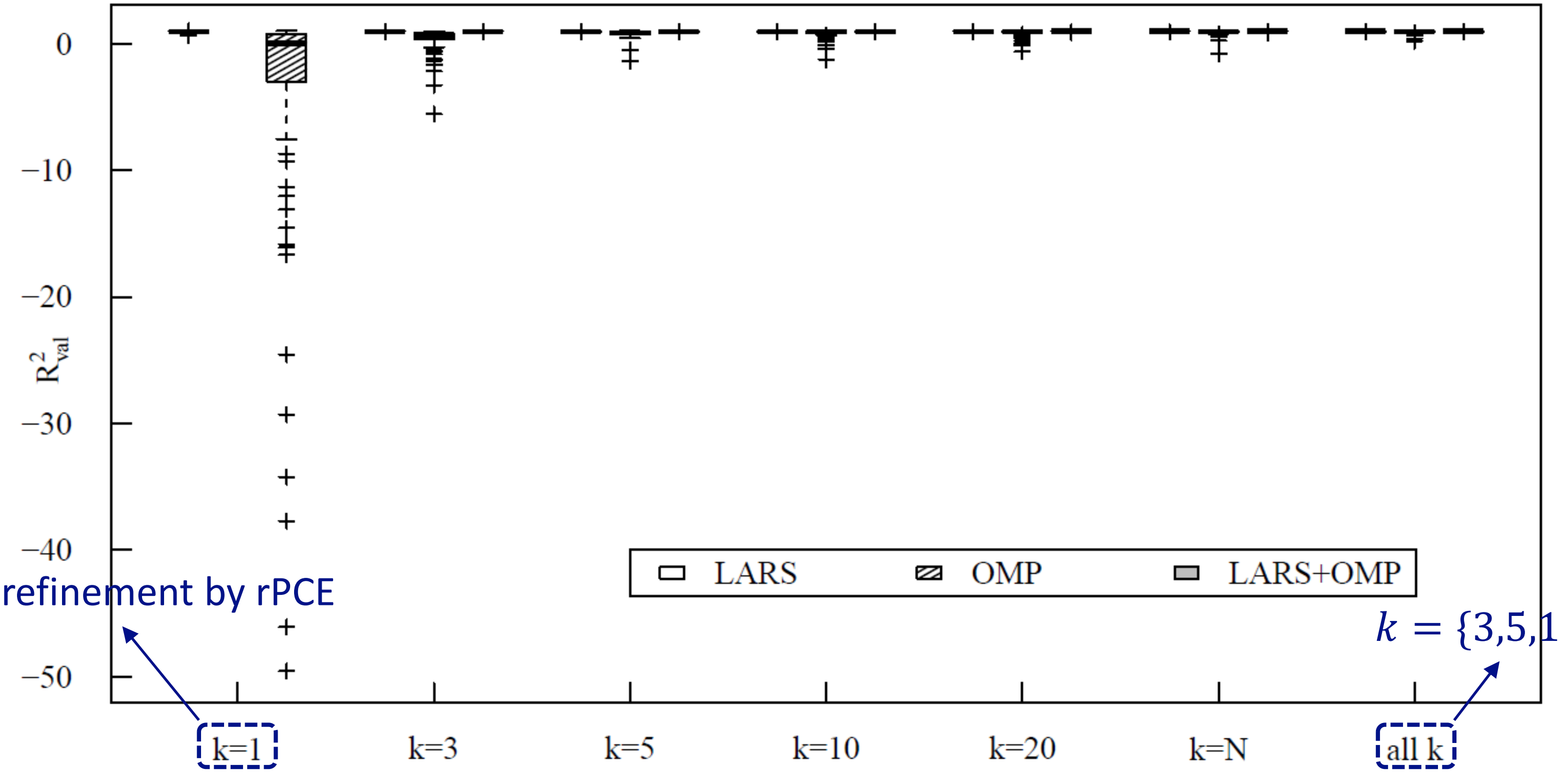
OMP ($R^2_{\mathrm{val}} = -6.2257$)



Truss deflection: prediction

Mean of R^2_{val} w.r.t. 100 replications

	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$	$k = N$	all k
LARS	0.9631	0.9651	0.9658	0.9692	0.9726	0.9735	0.9744
OMP	-6.2248	0.3873	0.7915	0.8273	0.8721	0.8974	0.9315
LARS+OMP		0.9641	0.9660	0.9693	0.9735	0.9741	0.9762



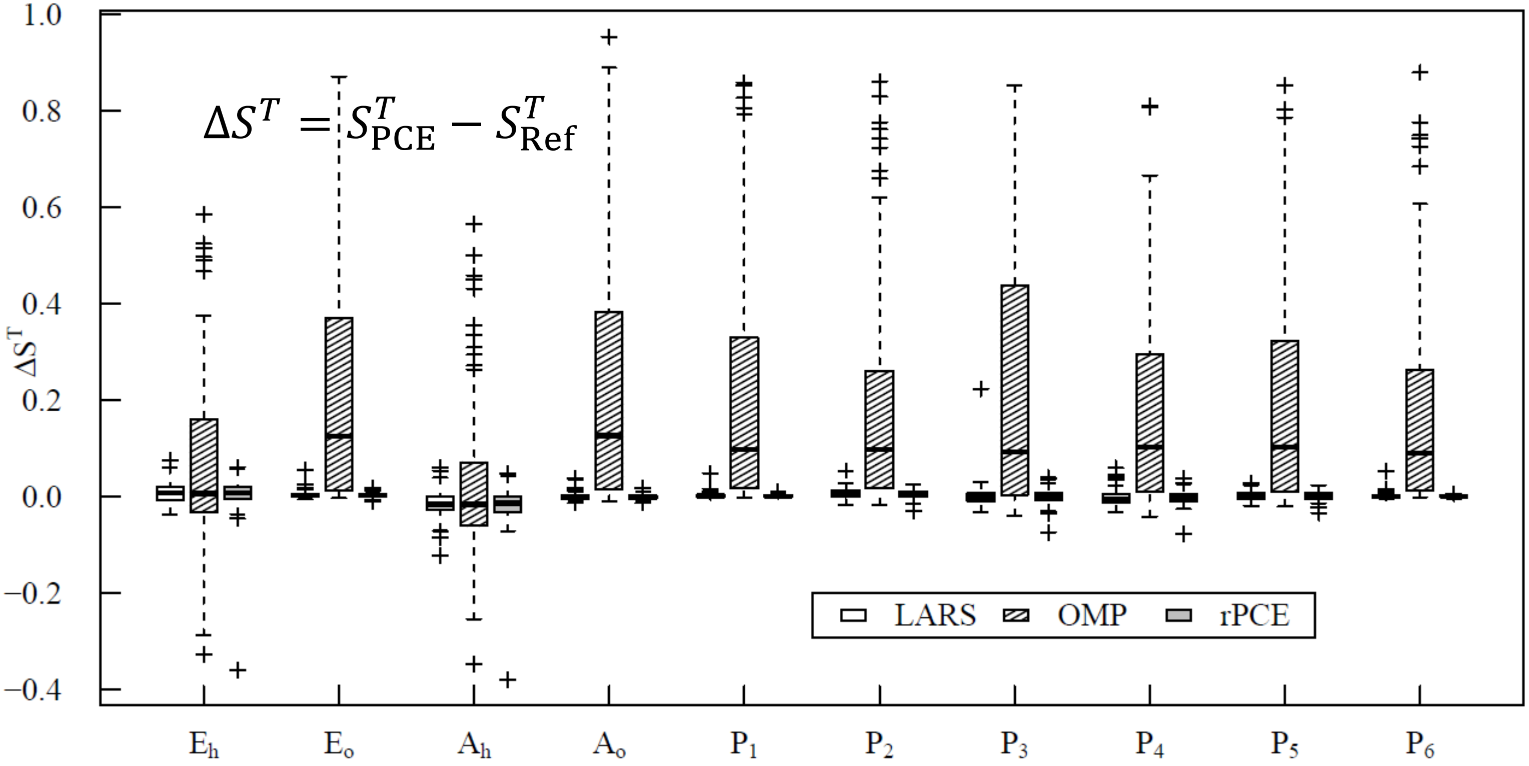
without refinement by rPCE

$k = \{3,5,10,20,N\}$

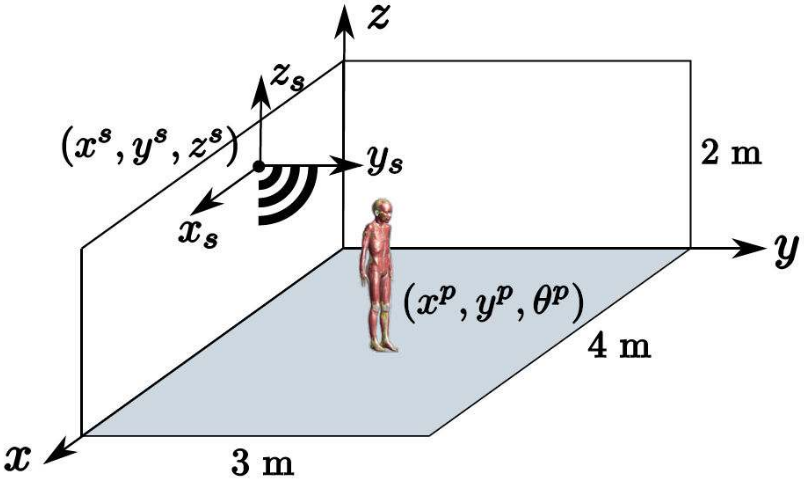
Truss deflection: total Sobol’ indices

Mean of total Sobol’ indices w.r.t. 100 replications

	E_h	E_o	A_h	A_o	P_1	P_2	P_3	P_4	P_5	P_6	$\sum$
Ref.	0.367	0.010	0.388	0.014	0.004	0.031	0.075	0.079	0.035	0.005	1.008
rPCE	0.3713	0.0121	0.3695	0.0127	0.0046	0.0359	0.0750	0.0756	0.0355	0.0048	0.9969
LARS	0.3748	0.0135	0.3715	0.0135	0.0057	0.0365	0.0759	0.0751	0.0361	0.0061	1.0086
OMP	0.4295	0.2290	0.4037	0.2291	0.2105	0.2251	0.2808	0.2557	0.2271	0.1891	2.6795



Estimation of specific absorption rate (SAR)



StarLab

Whole-body SAR (mW/kg), the ratio of the total power absorbed in the body to the mass of the human model, is computed with an in-house **FDTD** code.

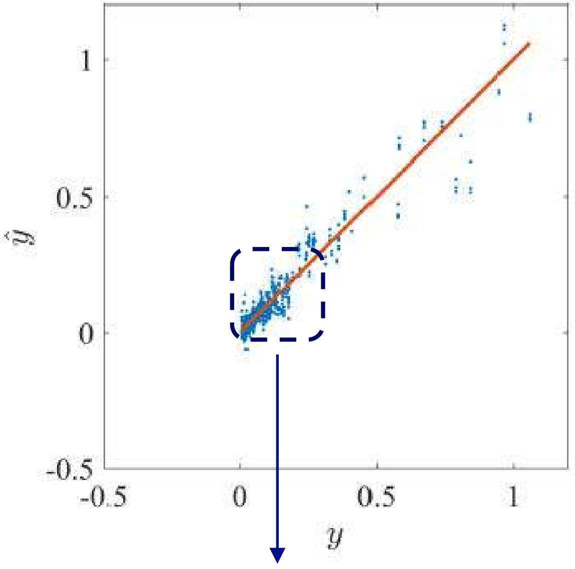
(x^s, y^s, z^s) , $(x^p, y^p, 0)$ and human orientation θ^p are inputs, which are independent and uniformly distributed in $[\mathbf{0.05}, 3.95]$, $[0.05,2.95]$, $[\mathbf{0.25}, 2]$, $[\mathbf{0.3}, 3.7]$, $[0.3,2.7]$ in meters and $[0,360)$ in degrees.

Reflection by walls, ceiling and ground is not considered. Thus, **four variables** including polar coordinates (r_s^p, ϕ_s^p), θ_s^p w.r.t. local coordinate system of WLAN source and z^s are considered finally.

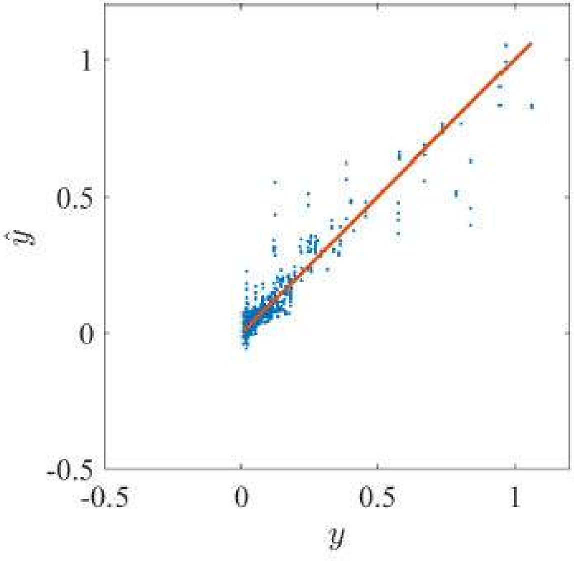
SAR estimation: prediction

340 data for training and 10 data for validation, build PCE models based on LARS, OMP and rPCE. Repeating the above process 100 times, one has 10^3 prediction data.

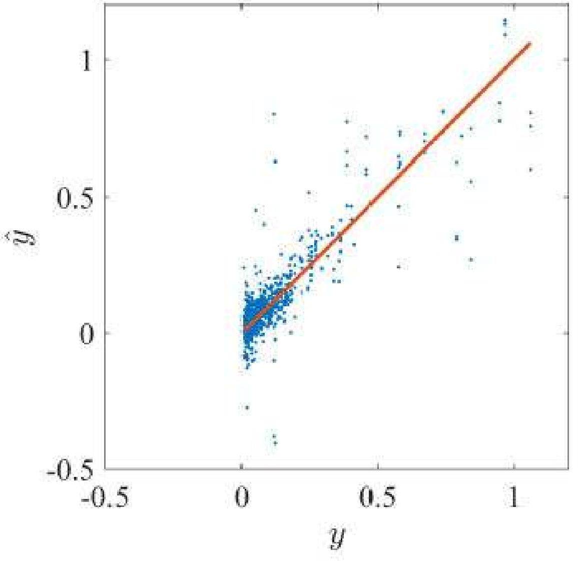
rPCE ($R^2_{\text{val}} = 0.9102$)



LARS ($R^2_{\text{val}} = 0.8688$)



OMP ($R^2_{\text{val}} = 0.7269$)

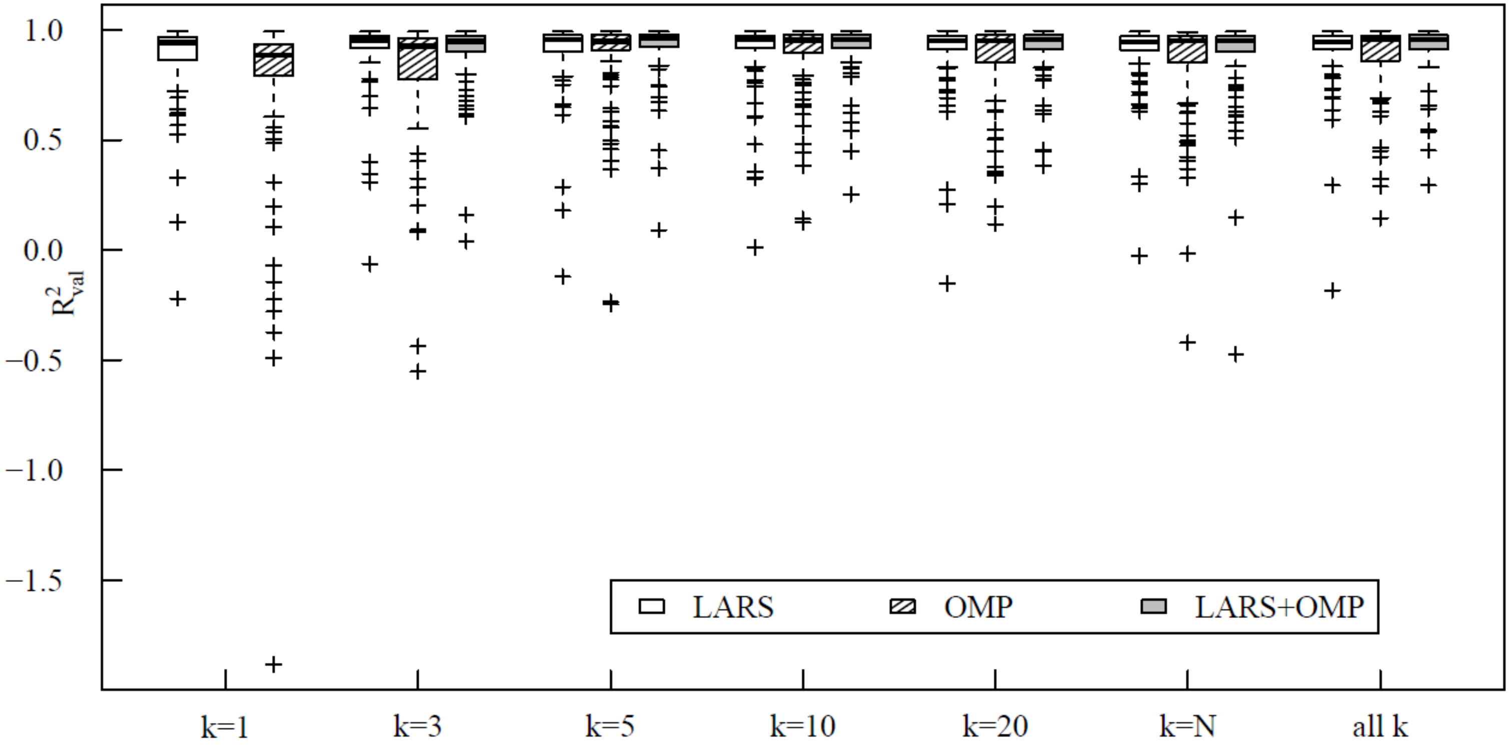


90% data with values< 0.2

SAR estimation: prediction

Mean of R^2_{val} w.r.t. 100 replications

	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$	$k = N$	all k
LARS	0.8799	0.9085	0.9067	0.8995	0.9033	0.8995	0.9068
OMP	0.7500	0.8186	0.8771	0.8854	0.8628	0.8521	0.8794
LARS+OMP		0.9046	0.9182	0.9171	0.9157	0.8893	0.9178

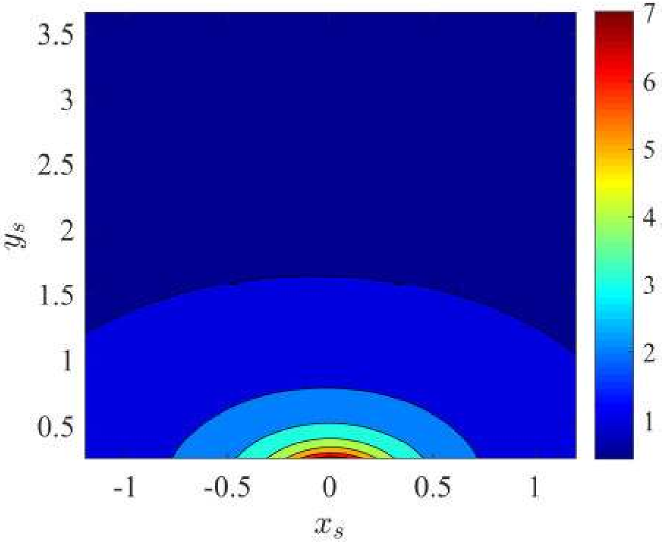


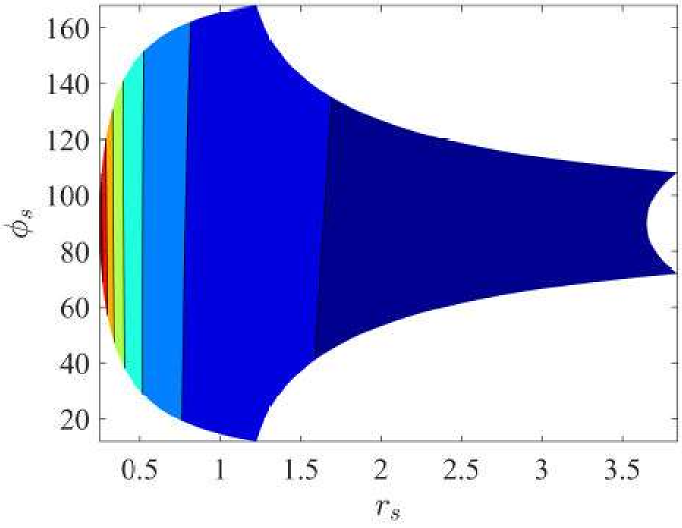
SAR estimation: global sensitivity analysis

Mean of total Sobol’ indices w.r.t. 100 replications

	r_s^p	ϕ_s^p	z^s	θ_s^p	$\sum$
rPCE	0.9809	0.0128	0.2175	0.0098	1.2210
LARS	0.9714	0.0357	0.1954	0.0316	1.2341
OMP	0.9761	0.0984	0.2925	0.0743	1.4412

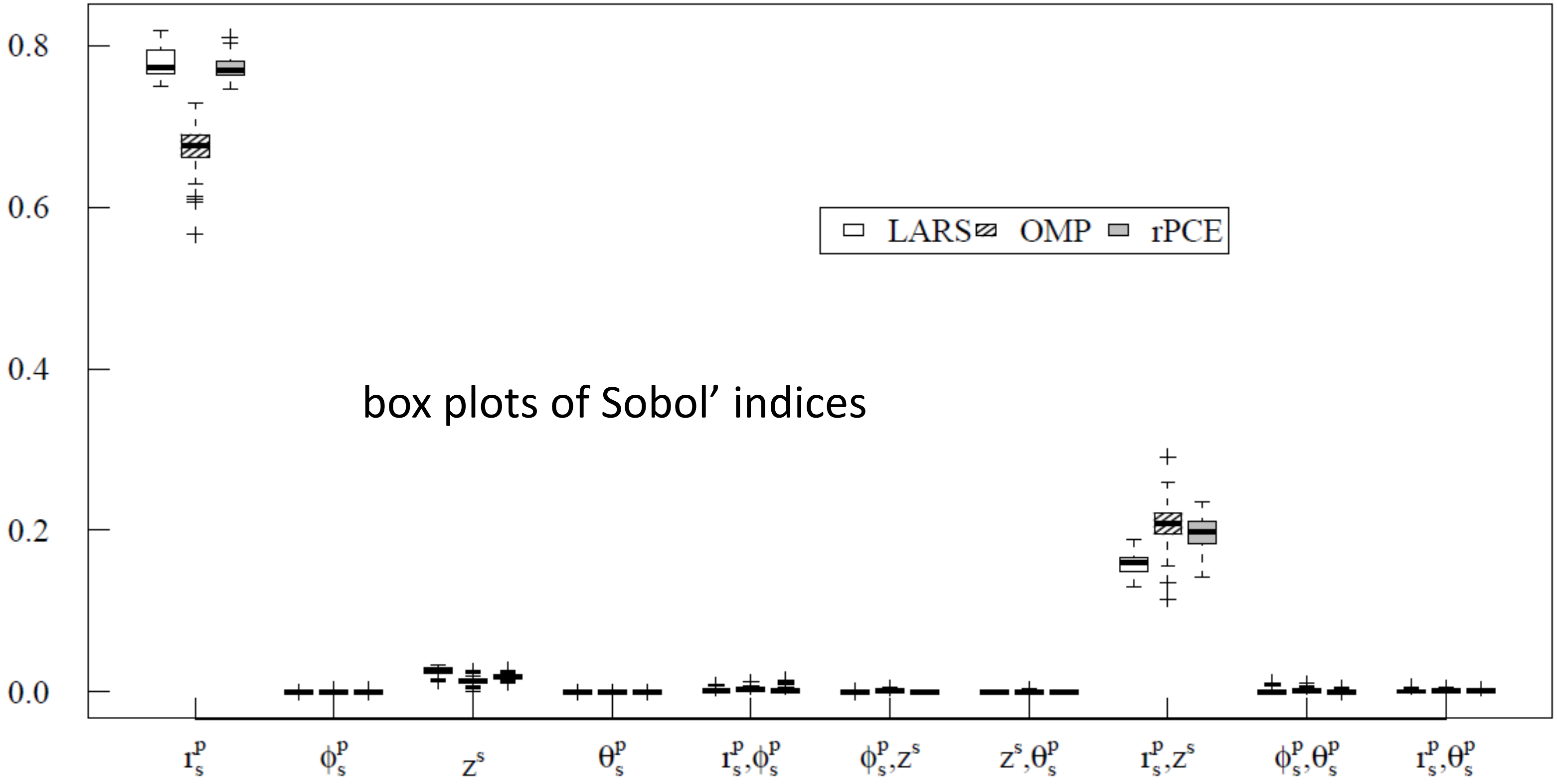
Large value of r_s^p and small value of ϕ_s^p maybe explained by observing following electric-field intensity map, where observation plane is $z_s = 0$.





SAR estimation: global sensitivity analysis

	r_s^p	ϕ_s^p	z^s	θ_s^p	Σ
rPCE	0.9809	0.0128	0.2175	0.0098	1.2210
LARS	0.9714	0.0357	0.1954	0.0316	1.2341
OMP	0.9761	0.0984	0.2925	0.0743	1.4412



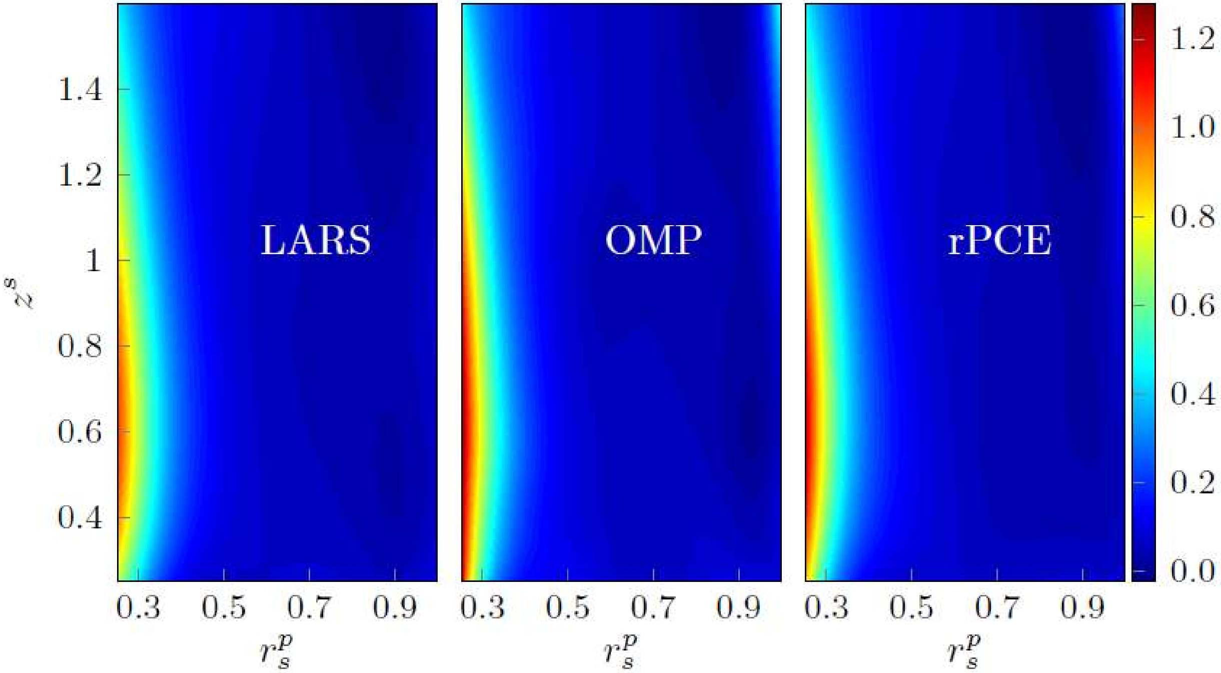
box plots of Sobol' indices

SAR estimation: global sensitivity analysis

	r_s^p	ϕ_s^p	z^s	θ_s^p	Σ
rPCE	0.9809	0.0128	0.2175	0.0098	1.2210
LARS	0.9714	0.0357	0.1954	0.0316	1.2341
OMP	0.9761	0.0984	0.2925	0.0743	1.4412

Setting $\phi_s^p = 0$ and $\theta_s^p = 0$, predict values of whole-body SAR:

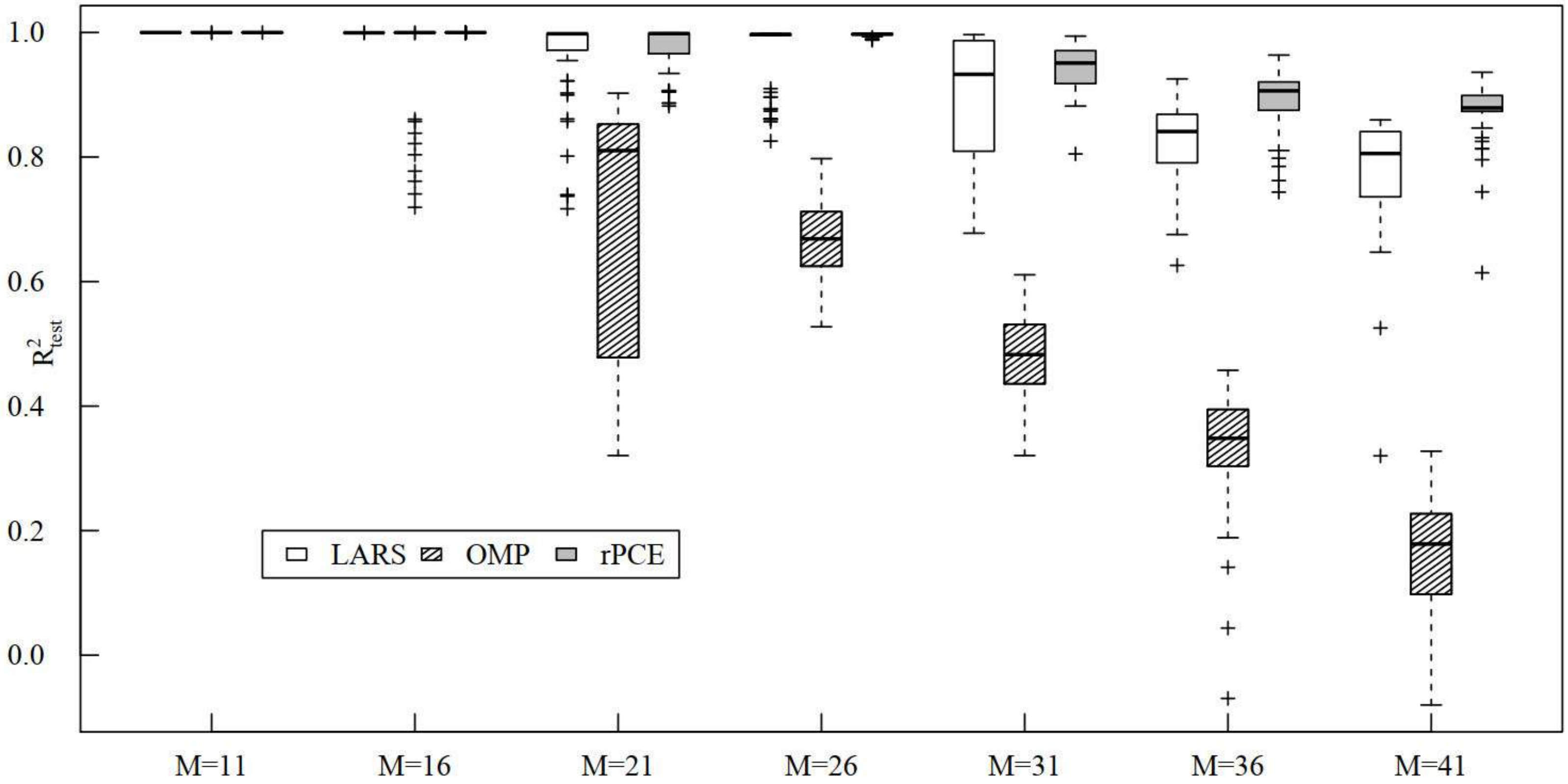
height of human model 1.36 m



Example with varied input dimension: prediction

$$y = 3 + x_1 x_2^2 - x_3 x_5 + x_2 x_4 + \frac{1}{M}\sum_{k=1}^{M} k(x_k^3 - 5x_k) + \ln\left(\frac{1}{3M}\sum_{k=1}^{M} k(x_k^2 + x_k^4)\right) + x_{M-4} + x_{M-4}x_M^2$$

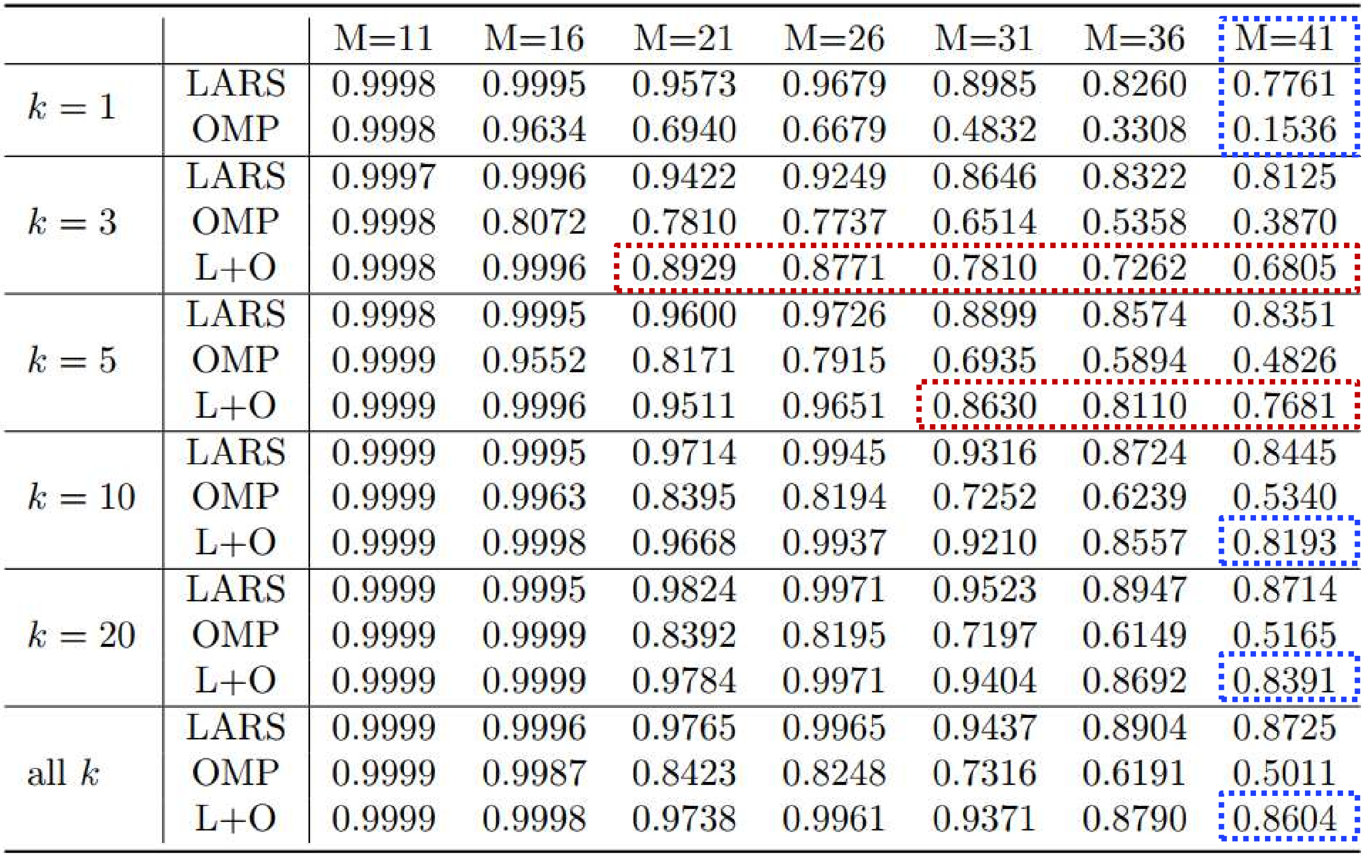
X_i are independent and uniform in $[1,2], i = 1, \ldots, M$. Range of X_{20} (when $M \geq 20$) is changed as $[1,3]$ to increase non-linearity. **200** data for training and 10^3 data for validation, repeat the construction of PCE models 50 times. A lighter setting of k, $k = \{3,5,10,20\}$, is applied.



Example with varied input dimension: prediction

$$y = 3 + x_1 x_2^2 - x_3 x_5 + x_2 x_4 + \frac{1}{M}\sum_{k=1}^{M} k(x_k^3 - 5x_k) + \ln\left(\frac{1}{3M}\sum_{k=1}^{M} k(x_k^2 + x_k^4)\right) + x_{M-4} + x_{M-4}x_M^2$$

X_i are independent and uniform in $[1,2], i = 1, \dots, M$. Range of X_{20} (when $M \geq 20$) is changed as $[1,3]$ to increase non-linearity. **200** data for training and 10^3 data for validation, repeat the construction of PCE models 50 times. A lighter setting of k, all $k = \{3,5,10,20\}$, is applied.

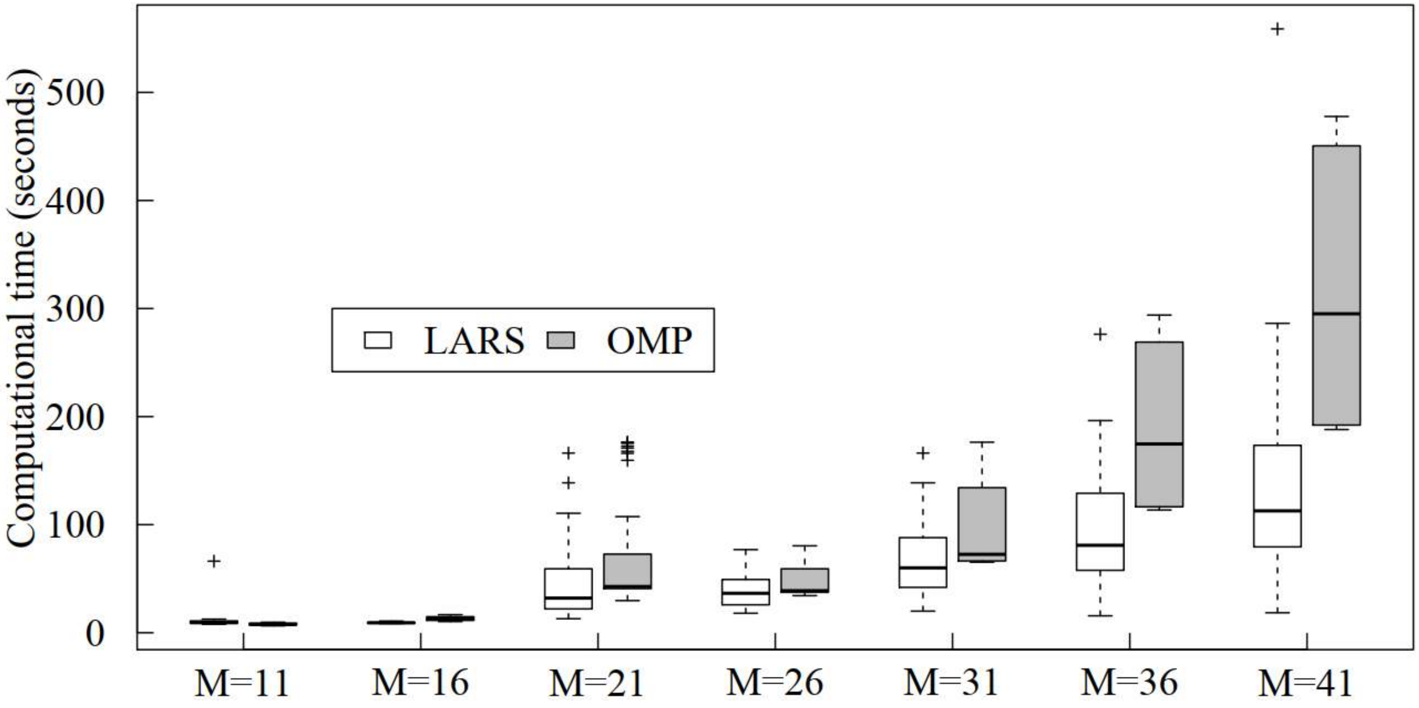


		M=11	M=16	M=21	M=26	M=31	M=36	M=41
$k = 1$	LARS	0.9998	0.9995	0.9573	0.9679	0.8985	0.8260	0.7761
	OMP	0.9998	0.9634	0.6940	0.6679	0.4832	0.3308	0.1536
$k = 3$	LARS	0.9997	0.9996	0.9422	0.9249	0.8646	0.8322	0.8125
	OMP	0.9998	0.8072	0.7810	0.7737	0.6514	0.5358	0.3870
	L+O	0.9998	0.9996	0.8929	0.8771	0.7810	0.7262	0.6805
$k = 5$	LARS	0.9998	0.9995	0.9600	0.9726	0.8899	0.8574	0.8351
	OMP	0.9999	0.9552	0.8171	0.7915	0.6935	0.5894	0.4826
	L+O	0.9999	0.9996	0.9511	0.9651	0.8630	0.8110	0.7681
$k = 10$	LARS	0.9999	0.9995	0.9714	0.9945	0.9316	0.8724	0.8445
	OMP	0.9999	0.9963	0.8395	0.8194	0.7252	0.6239	0.5340
	L+O	0.9999	0.9998	0.9668	0.9937	0.9210	0.8557	0.8193
$k = 20$	LARS	0.9999	0.9995	0.9824	0.9971	0.9523	0.8947	0.8714
	OMP	0.9999	0.9999	0.8392	0.8195	0.7197	0.6149	0.5165
	L+O	0.9999	0.9999	0.9784	0.9971	0.9404	0.8692	0.8391
all k	LARS	0.9999	0.9996	0.9765	0.9965	0.9437	0.8904	0.8725
	OMP	0.9999	0.9987	0.8423	0.8248	0.7316	0.6191	0.5011
	L+O	0.9999	0.9998	0.9738	0.9961	0.9371	0.8790	0.8604

Example with varied input dimension: time cost

$$y = 3 + x_1 x_2^2 - x_3 x_5 + x_2 x_4 + \frac{1}{M}\sum_{k=1}^{M} k(x_k^3 - 5x_k) + \ln\left(\frac{1}{3M}\sum_{k=1}^{M} k(x_k^2 + x_k^4)\right) + x_{M-4} + x_{M-4}x_M^2$$

X_i are independent and uniform in $[1,2], i = 1, \dots, M$. Range of X_{20} (when $M \geq 20$) is changed as $[1,3]$ to increase non-linearity. **200** data for training and 10^3 data for validation, repeat the construction of PCE models 50 times. A lighter setting of k, all $k = \{3,5,10,20\}$, is applied.

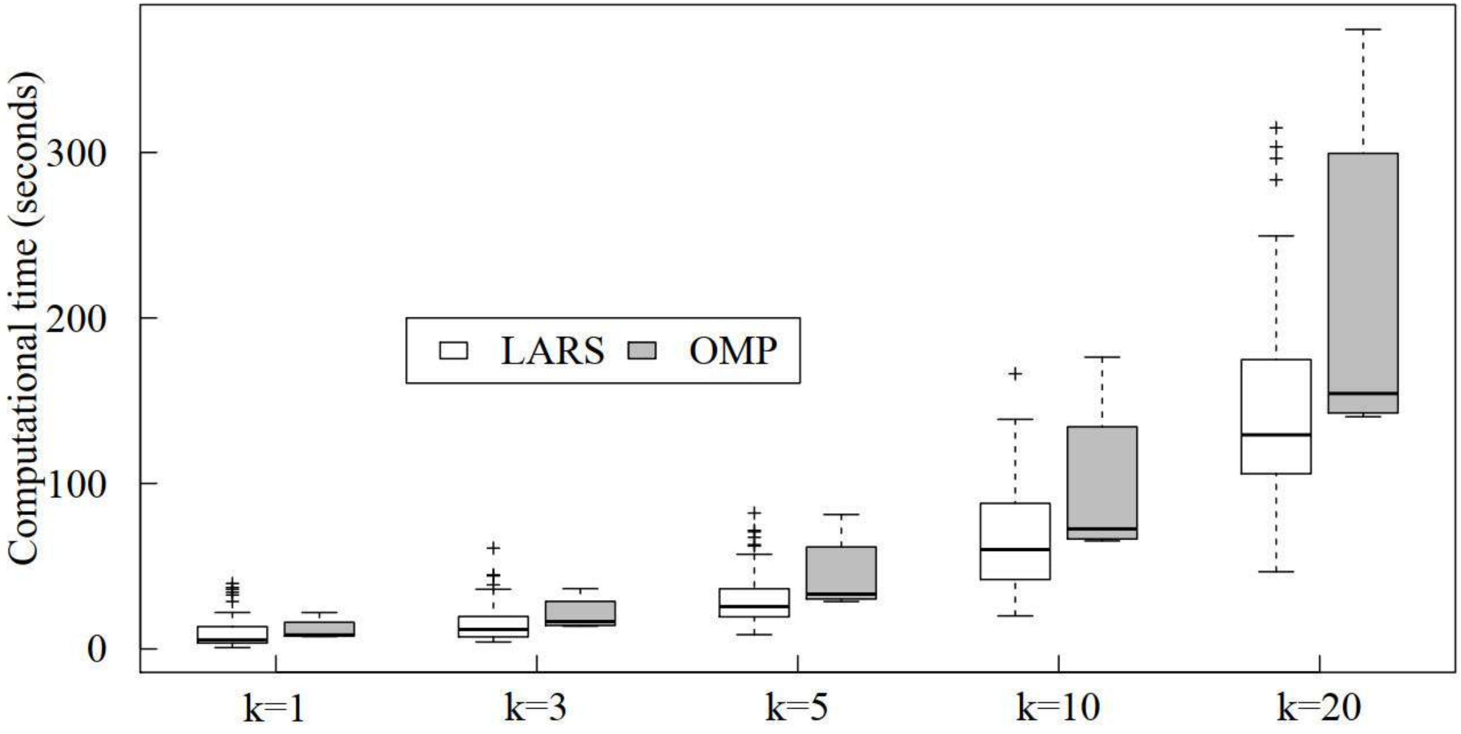


$k = 10$

Example with varied input dimension: time cost

$$y = 3 + x_1 x_2^2 - x_3 x_5 + x_2 x_4 + \frac{1}{M}\sum_{k=1}^{M} k(x_k^3 - 5x_k) + \ln\left(\frac{1}{3M}\sum_{k=1}^{M} k(x_k^2 + x_k^4)\right) + x_{M-4} + x_{M-4}x_M^2$$

X_i are independent and uniform in $[1,2], i = 1, \dots, M$. Range of X_{20} (when $M \geq 20$) is changed as $[1,3]$ to increase non-linearity. **200** data for training and 10^3 data for validation, repeat the construction of PCE models 50 times. A lighter setting of k, all $k = \{3,5,10,20\}$, is applied.



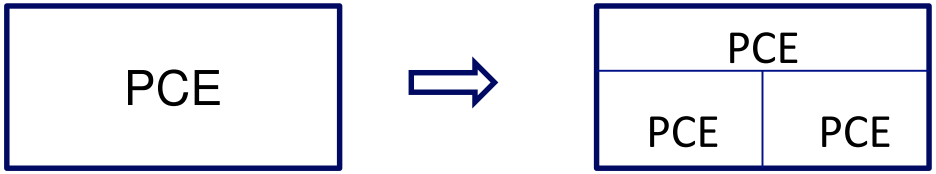
$M = 31$

Conclusions

- **Resampled PCE (rPCE)** refines the ranking of importance of candidate polynomials in the context of **sparse polynomial chaos expansions**
- **Resampling scheme** (k-fold division) and **source of candidate polynomials** (LARS, OMP or their combination) impact the performance of rPCE
- Analyse global sensitivity through the computation of **Sobol' indices**
- **Application examples** include two analytical functions, one FEM model (truss deflection) and one FDTD model (SAR estimation)
- **OMP**-based PCE modelling seems the **worst** among three methods. **LARS**-based approach generally generates a **better model** and **refinements by rPCE** are obvious in terms of prediction variance and number of outliers. **rPCE** performs as least **as well as LARS** for **global sensitivity analysis**

Perspectives

- Modelling **complicated physical scenarios** require high-order PCE models, construction of which easily sink into **overfitting problem**. Complicated scenarios are divided into **several simpler ones**.



Ranking polynomials in rPCE

Ranking polynomials by total score

$$s = s_f + s_e$$

s_f frequency score, s_e error score.

Frequency score

$$s_f = \sum_{k \in \{3,5,10,20,N\}} s_f^k \frac{\mathrm{lcm}(3,20,N)}{k}$$

"lcm" short for least common multiple.

Error score

In PCE modeling based on OMP/LARS, each basis polynomial results increment of ϵ_{LOO}

$$\Delta\epsilon_{\mathrm{LOO}}^{j} = \epsilon_{\mathrm{LOO}}^{j} - \epsilon_{\mathrm{LOO}}^{j-1}$$

and

$$s_e = \frac{1}{s_f \Delta\epsilon_{\mathrm{LOO}}^{\max}} \sum_j \Delta\epsilon_{\mathrm{LOO}}^{j}$$

Borehole function

$$Y = \frac{2\pi T_u(H_u - H_l)}{\ln(r/r_w)\left(1 + T_u/T_l\right) + 2LT_u/r_w^2K_w}$$

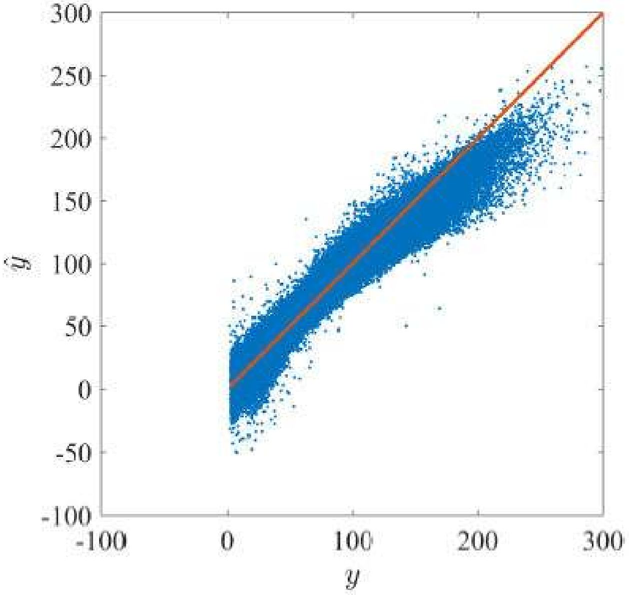
Borehole function, which is nonlinear and non-additive, models water flow through a borehole. 8 input features are independent.

Name	Distribution	Bounds	Description
r_w (m)	$\mathcal{N}(0.10, 0.0161812)$	[0.05, 0.15]	radius of borehole
r (m)	Lognormal(7.71, 1.0056)	[100, 50000]	radius of influence
T_u (m^2/yr)	Uniform	[63070, 115600]	transmissivity of upper aquifer
H_u (m)	Uniform	[990, 1110]	potentiometric head of upper aquifer
T_l (m^2/yr)	Uniform	[63.1, 116]	transmissivity of lower aquifer
H_l (m)	Uniform	[700, 820]	potentiometric head of lower aquifer
L (m)	Uniform	[1120, 1680]	length of borehole
K_w (m/yr)	Uniform	[1500, 15000]	hydraulic conductivity of borehole

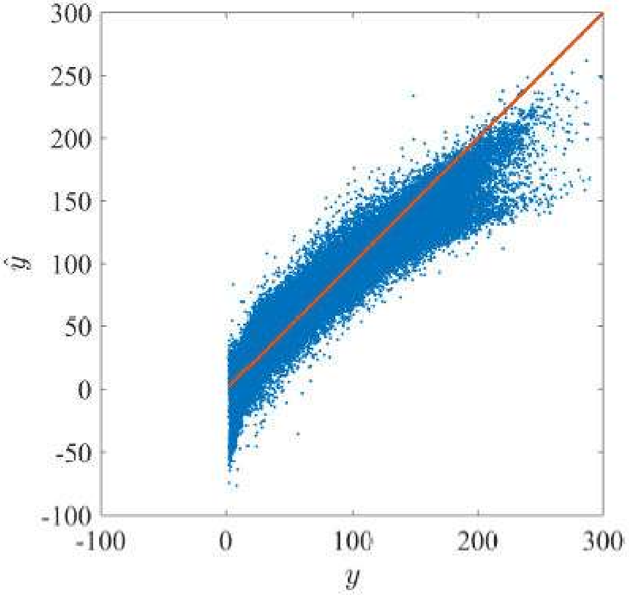
Borehole function: prediction

40 data for training and 10^4 data for validation, build PCE models based on LARS, OMP and rPCE. Repeating the above process 100 times, one has 10^6 prediction data.

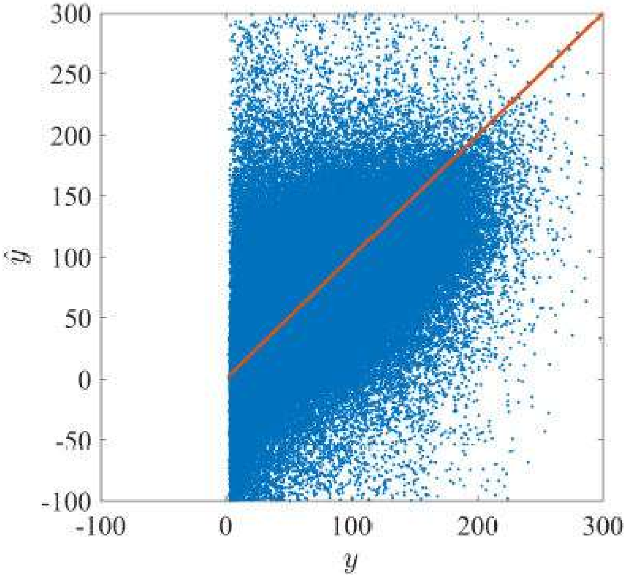
rPCE ($R^2_{\text{val}} = 0.9723$)



LARS ($R^2_{\text{val}} = 0.9517$)



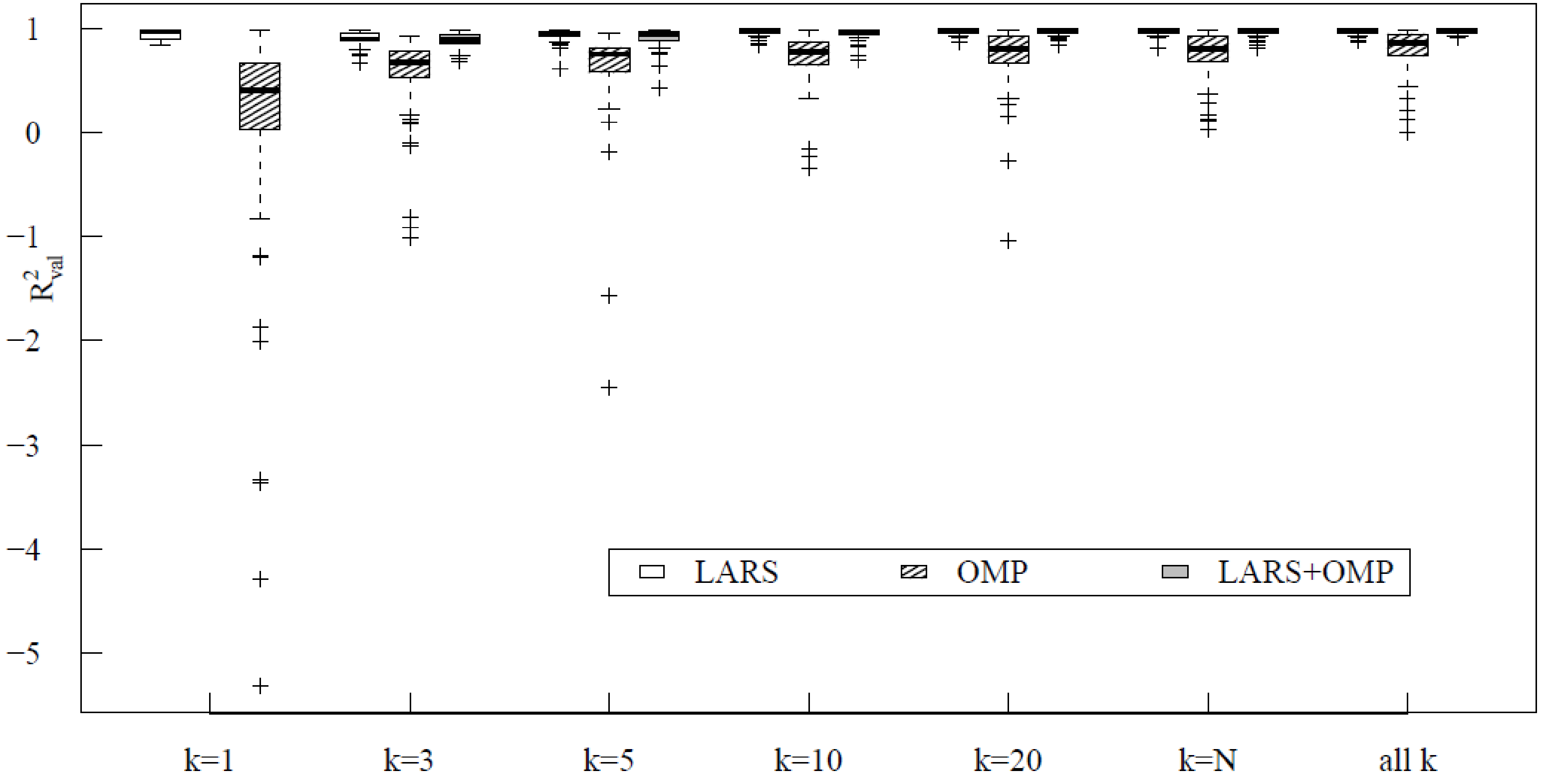
OMP ($R^2_{\text{val}} = 0.1472$)



Borehole function: prediction

Mean of R^2_{val} w.r.t. 100 replications

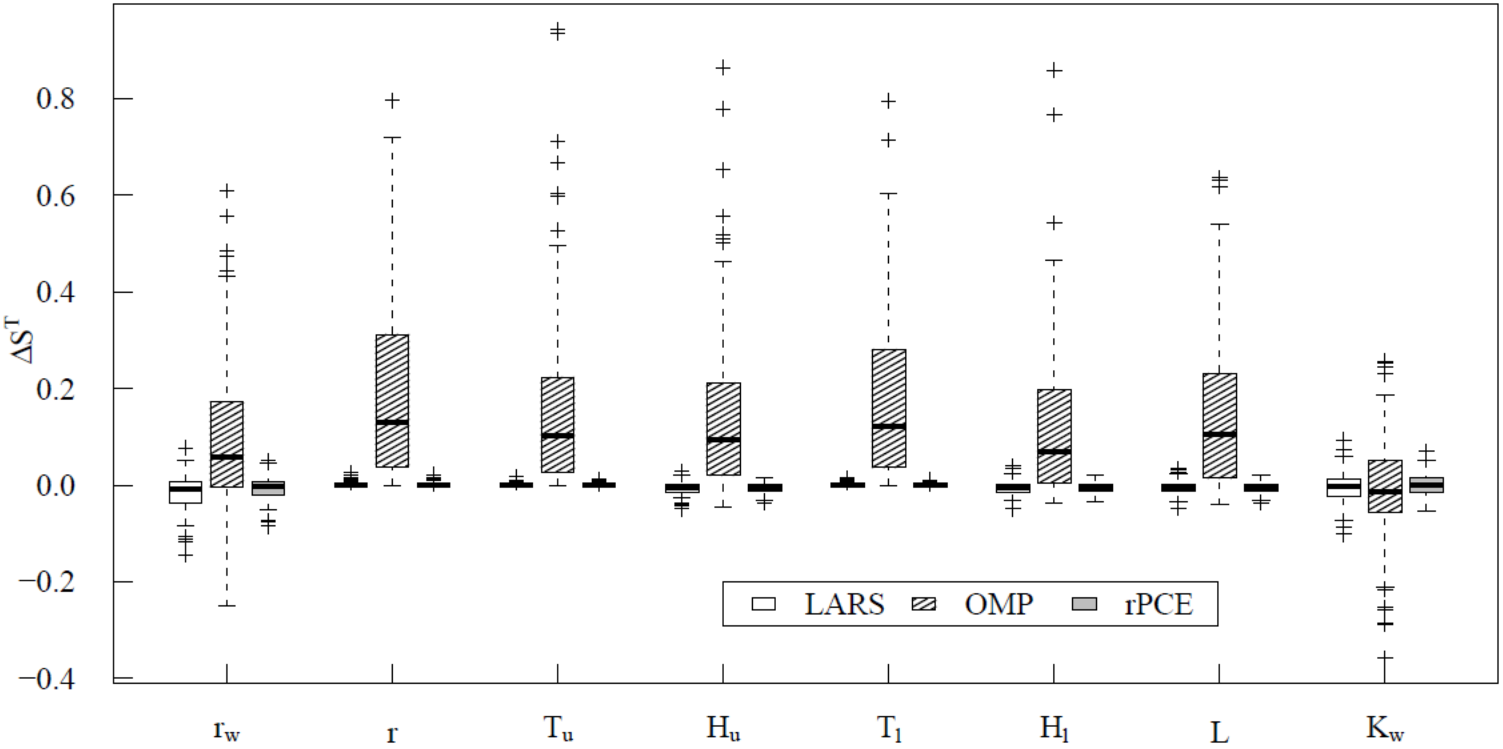
	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$	$k = N$	all k
LARS	0.9517	0.9072	0.9451	0.9673	0.9736	0.9743	0.9719
OMP	0.1467	0.5852	0.6434	0.7293	0.7506	0.7633	0.8112
LARS+OMP		0.8859	0.9239	0.9587	0.9704	0.9697	0.9723



Borehole function: total Sobol' indices

Mean of total Sobol' indices w.r.t. 100 replications

	r_w	r	T_u	H_u	T_l	H_l	L	K_w	$\sum$
Reference	0.3127	0.0000	0.0000	0.0487	0.0000	0.0487	0.0472	0.6369	1.0942
rPCE	0.3072	0.0010	0.0010	0.0418	0.0011	0.0431	0.0423	0.6376	1.0751
LARS	0.2962	0.0023	0.0015	0.0420	0.0018	0.0427	0.0427	0.6322	1.0614
OMP	0.4127	0.1967	0.1635	0.1995	0.1802	0.1751	0.2026	0.6259	2.1562


ΔS^T
LARS
OMP
rPCE
r_w
r
T_u
H_u
T_l
H_l
L
K_w