Treating Computer Experiments: What Matters, What Doesn't, What Evidence

Jerome Sacks

National Institute of Statistical Sciences

Work with J. Jakeman (Sandia), J. Loeppky (UBC, Okanagan), H. Chen and W. Welch (UBC, Vancouver)

SIAM UQ April 1, 2014

INTRODUCTION

25 years ago in Statistics:
Bayesian Gaussian Process models (GaSP) introduced for ***design***
and ***analysis*** of expensive-to-run computer experiments

About the same time in Polynomial Chaos (PC) methods
introduced.

Since then: a multitude of varieties of design, GaSPs and PCs.
What to choose and on what basis?

Why Bother

Two reasons (apart from avoiding embarrassment):

- "UQ is the end-to-end study of the reliability of scientific inferences" so apply UQ to reliability of UQ methods
- In the absence of proof what constitutes adequate evidence?

How to Address

(1) Focus: use of computer experiment to build a surrogate to the code output. (With a good surrogate "everything" can be done.)

(2) Accuracy of surrogate prediction at untried inputs.

(3) Apply to codes that reflect what modelers might face.

(4) Present evidence to a Court for "judgment"

Accuracy of Prediction

Notation: **x** in *d*-cube, input to code; $y(\mathbf{x})$ the scalar output; $\hat{y}$ the surrogate

For fast codes: large (N) set of holdout points i.e., inputs not used in the experiment, and compute

$$e_{\text{rmse,ho}} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}(\mathbf{x}_{\text{ho}}^{(i)}) - y(\mathbf{x}_{\text{ho}}^{(i)})\right)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\bar{y} - y(\mathbf{x}_{\text{ho}}^{(i)})\right)^2}}.$$

$\bar{y}$ = average of the experimental output, the training data.

Benchmark: $e_{\text{rmse,ho}} \leq .10$

Docket

Case 1: Designs

Case 2: GaSP v. PC

Case 3: Parameters of GaSP

Case 4: Stationary v. Non-stationary GaSP

Case 5: MLE v. Bayes

Argument for Fast Codes

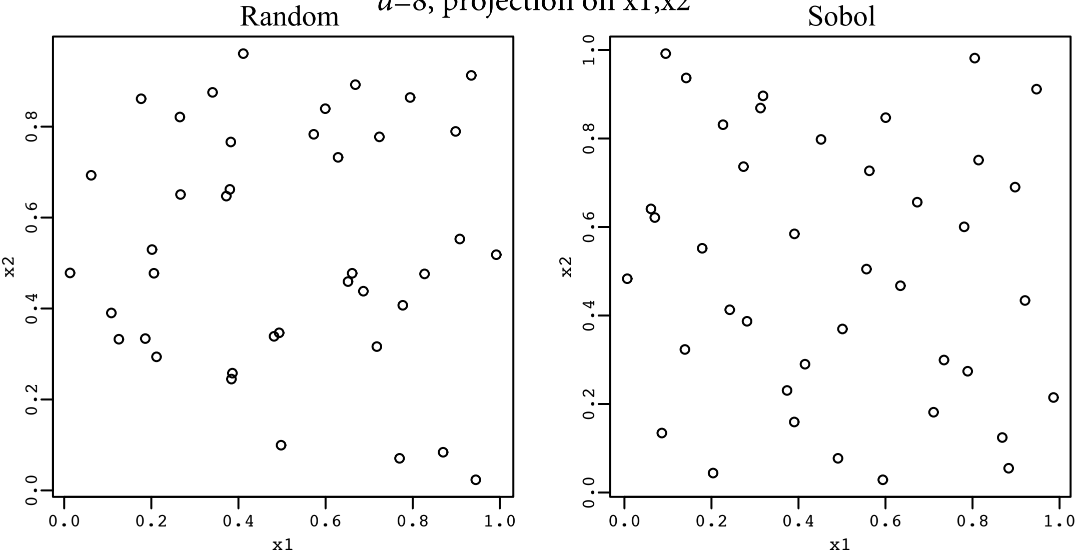
- Select (base) design
- Choose GaSP or PC
- Select test function and run code
- Compute $e_{rmse,ho}$
- Repeat for (20-25) designs by permuting coordinates of base design; holdout set remains fixed

Varieties of Design

(1) Completely **Random**

(2) Discrepancy Sequence (**Sobol**)

d=8, projection on x1,x2

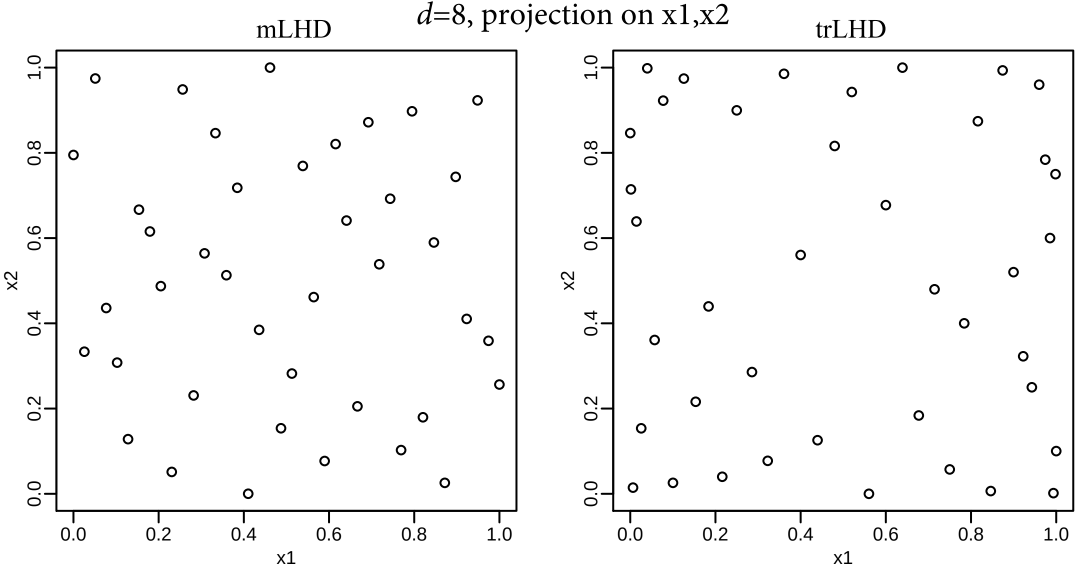


(3) “Maximin” Latin Hypercube Design (**mLHD**):

maximize minimum distance between points in the class of LHDs
-- "approximately" and with “nicer” 2-d projections

(4) Transform mLHD (**trLHD**) (Dette&Pepelyshev, 2009):

$$x_i \longrightarrow (1 - \cos(\pi x_i))/2$$



d=8, projection on x1,x2
mLHD
trLHD
x1
x2
0.0
0.2
0.4
0.6
0.8
1.0

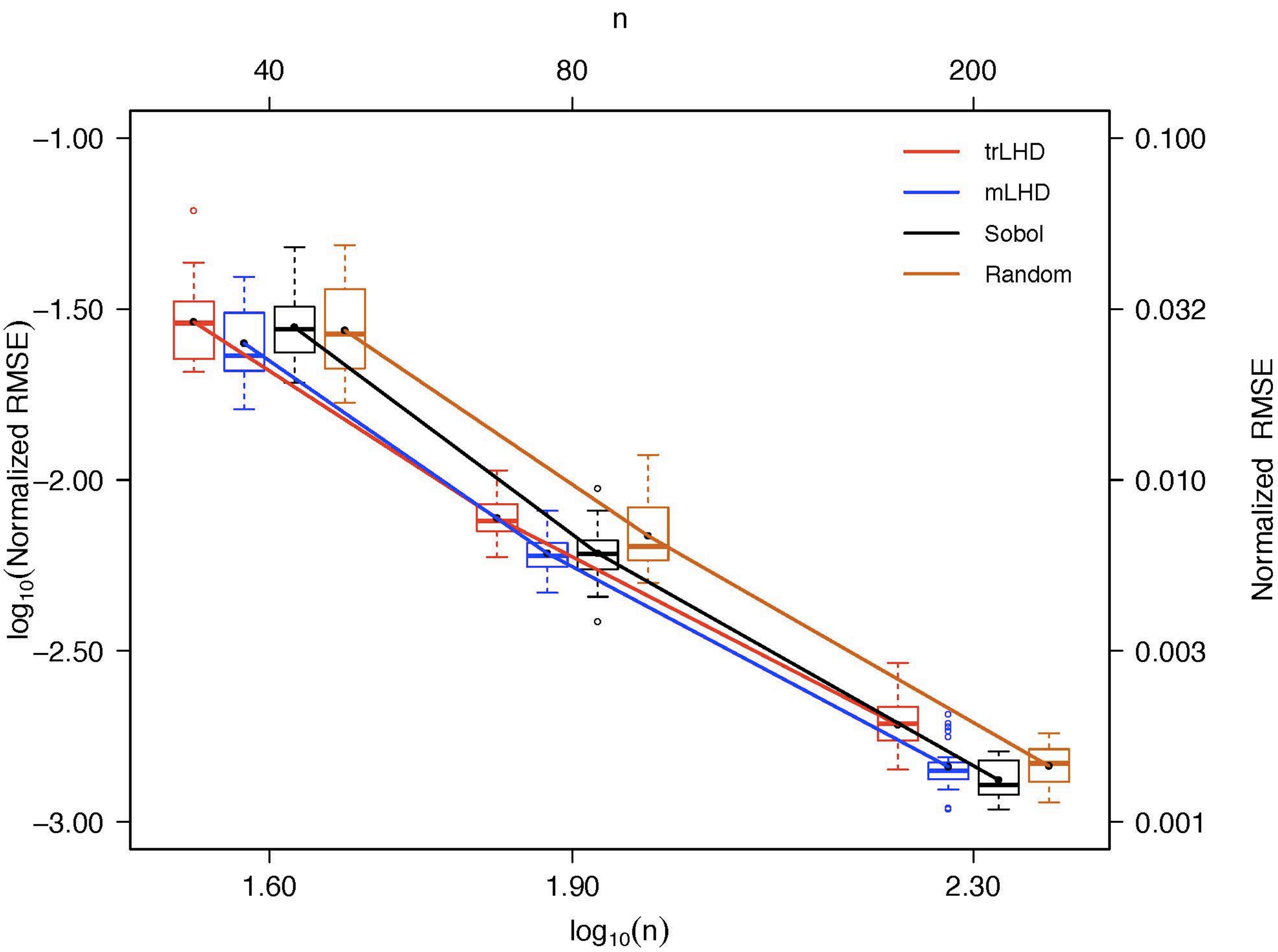
Case 1: trLHD v Others

Use GaSP for prediction.
Test Function:

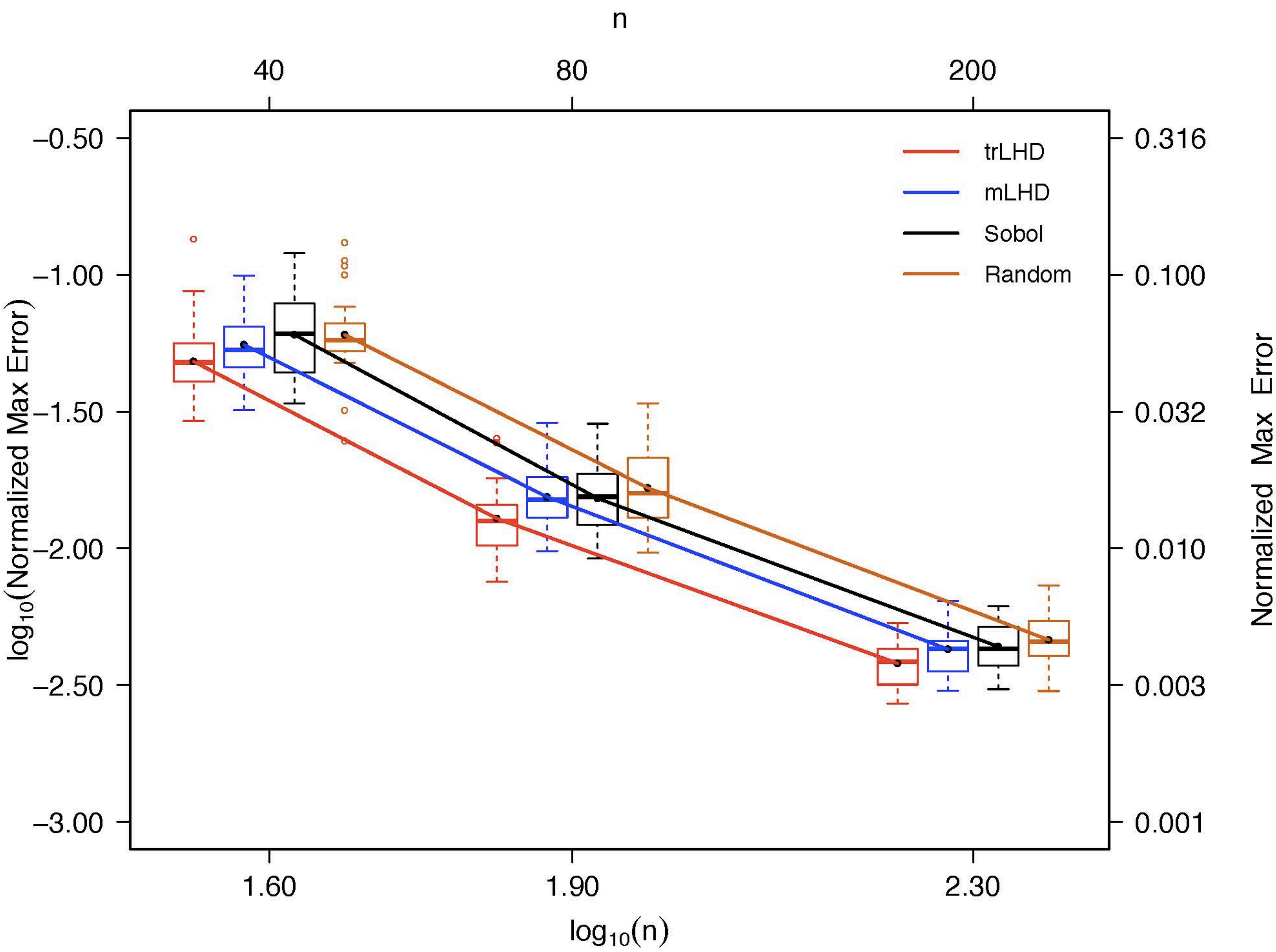
Borehole: *d*=8

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

Borehole (*d*=8); $e_{rmse,ho}$



Borehole (d=8); max error



Decision – Case 1: Designs

1. On basis of RMSE: disadvantage for trLHD unless extreme behavior at boundary. Advantage to trLHD under max error.

2. Minor differences among other choices.

3. Use easy to generate random LHD or Orthogonal LHD (nice 2-dimensional projections)

Don't sweat the design

Case 2: GaSP v PC

Test Functions:

(1) Corner Peak: d =10

$$y(\mathbf{x}) = \left(1 + \sum_{i=1}^{d} c_i x_i\right)^{-(d+1)} ; \quad c_i = C(1/i^2); \sum c_i = .25$$

(2) Resistor Network d=40

(3) Random Oscillator: d=6

$$\frac{d^2\mathrm{y}}{dt^2}(t, \boldsymbol{\xi}) + \gamma \frac{d\mathrm{y}}{dt} + kx = f \cos(\omega t),$$

$$\mathrm{y}(0) = \mathrm{y}_0, \quad \dot{\mathrm{y}}(0) = \mathrm{y}_1,$$

Polynomial Ch(oices)

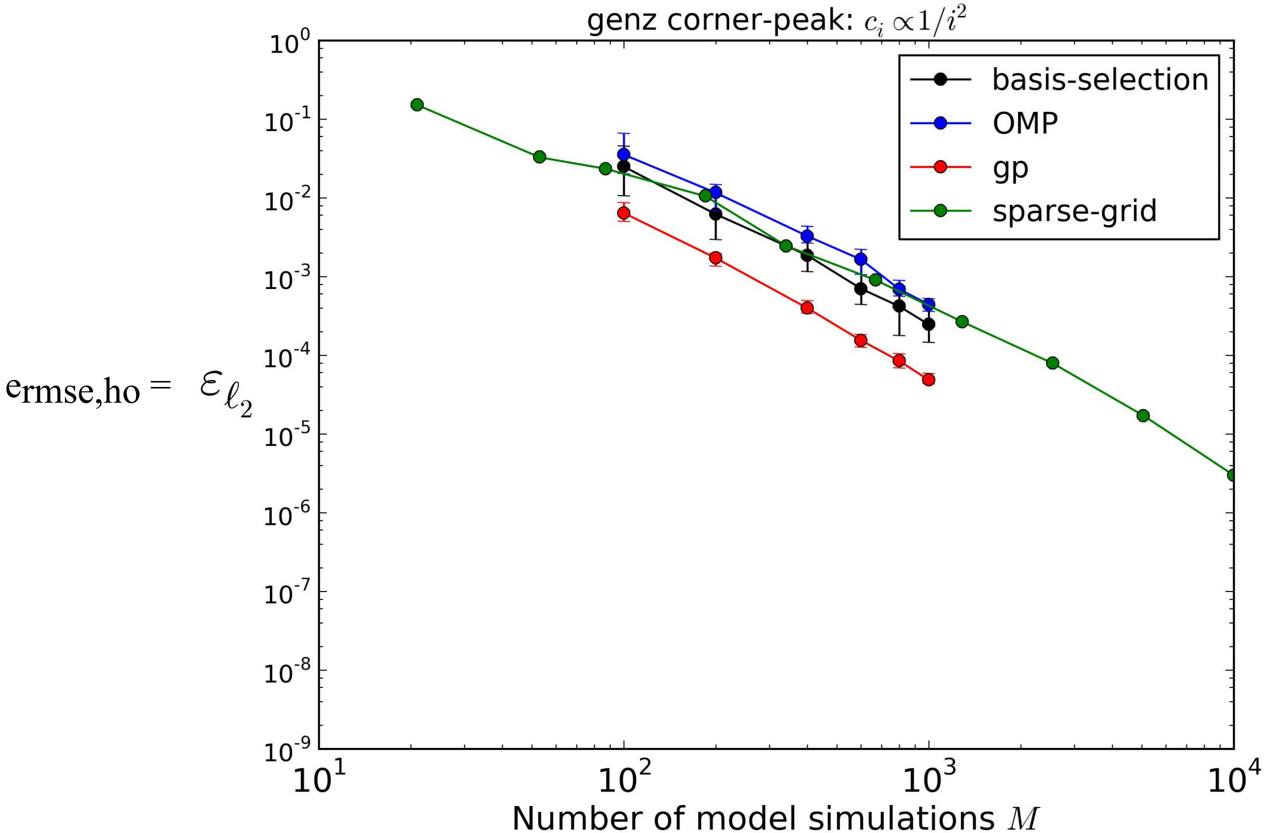
Relies on orthogonal polynomials

1. Adaptive sparse-grid
2. Compressed Sensing (OMP) — dual of Lasso:

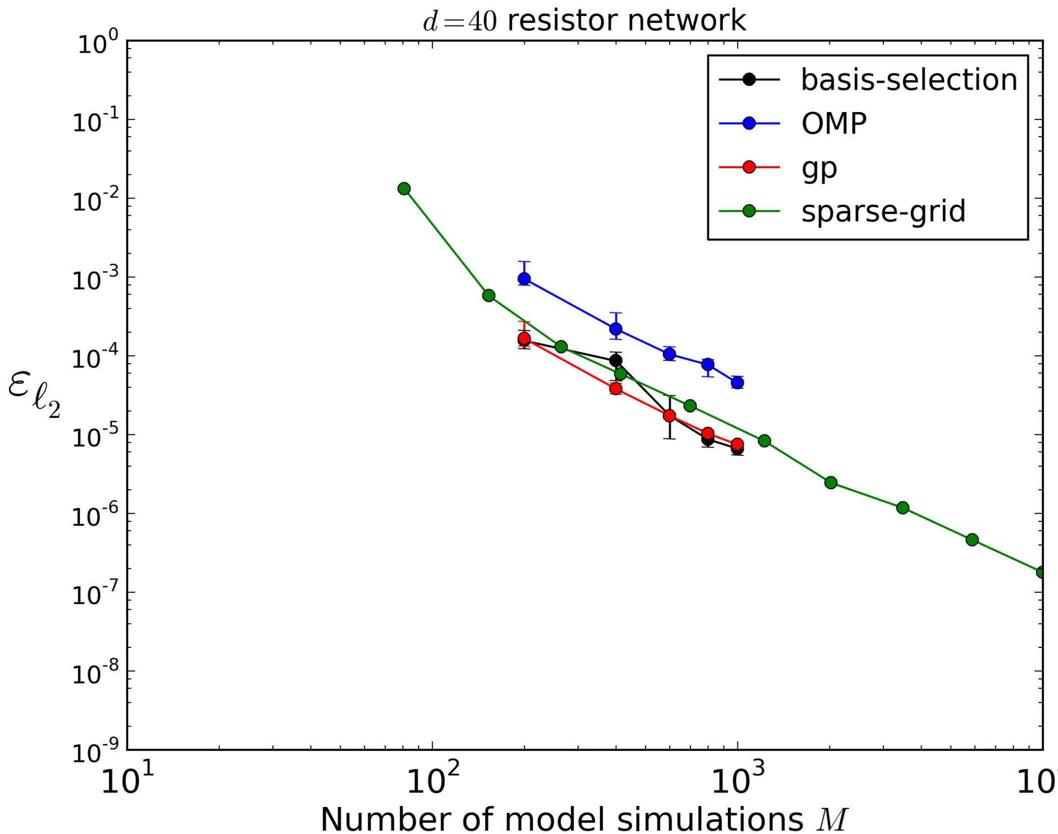
$$\min_{\mathbf{c}} \|\mathbf{c}\|_1 \;\; \text{s.t.} \;\; \|\boldsymbol{y} - \mathbf{Ac}\|_2 < \varepsilon$$

   Choose ε,p (p the degree of the polynomials) via cross-validation

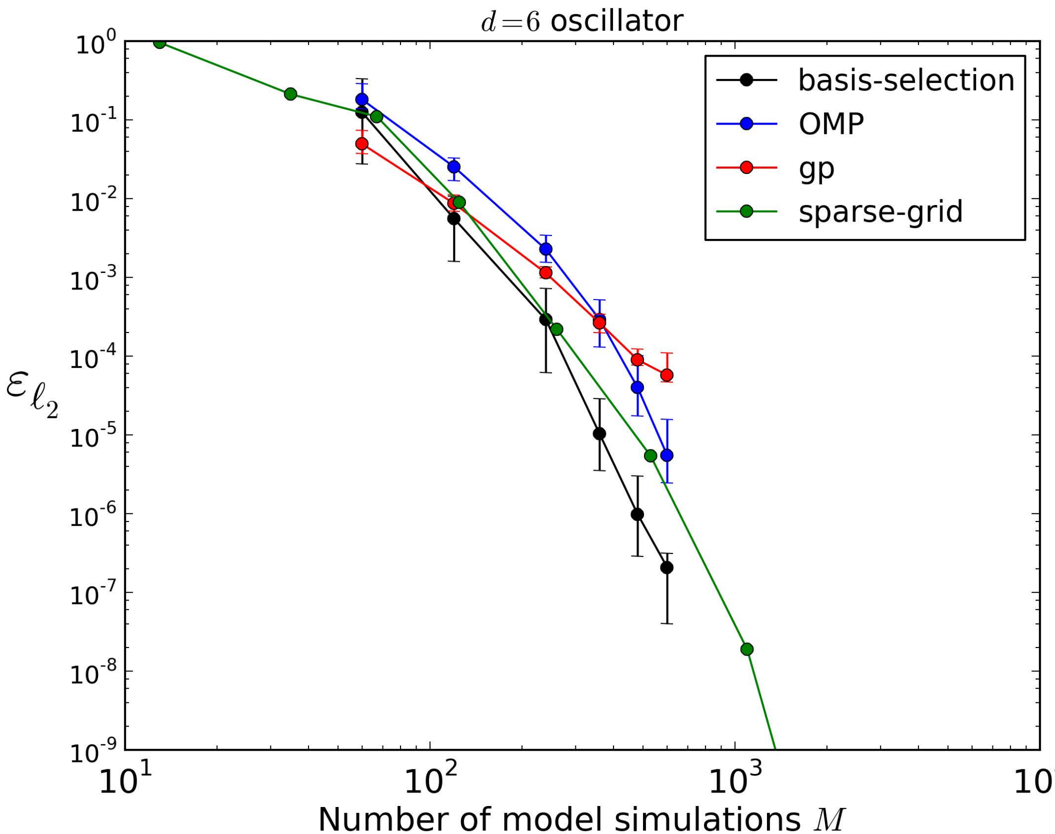
3. Expand #2 by choosing better basis terms e.g., by adding higher degree terms where called for (Jakeman, Eldred, Sargsyan, 2014)



genz corner-peak: $c_i \propto 1/i^2$
$e_{\mathrm{rmse,ho}} = \varepsilon_{\ell_2}$
basis-selection
OMP
gp
sparse-grid
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^1
10^2
10^3
10^4
Number of model simulations M



$d = 40$ resistor network
basis-selection
OMP
gp
sparse-grid
ε_{ℓ_2}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{1}
10^{2}
10^{3}
10^{4}
Number of model simulations M



$d = 6$ oscillator
basis-selection
OMP
gp
sparse-grid
ε_{ℓ_2}
Number of model simulations M

Decision – Case 2: GaSP v PC

1. For "modest" n (e.g., n=10d) preponderance of evidence favors GaSP.

2. For "large" n not clear.

Things go better with GaSP

GaSP

Model y as a random function, a Gaussian stochastic process with mean function $\mu(\mathbf{x})$ and covariance function σ^2R.

For any $\{(\mathbf{x}^{(1)}),\ldots,(\mathbf{x}^{(n)})\}$ let $\boldsymbol{\mu}$ = the vector of $\mu(\mathbf{x}^{(i)})$ and $\mathbf{R}$ the covariance matrix of $\boldsymbol{y} = \{y(\mathbf{x}^{(1)}),\ldots,y(\mathbf{x}^{(n)})\}$.

The likelihood, or density, of $\boldsymbol{y}$, L($\boldsymbol{y}$ | parameters μ,σ^2,R), is

$$\frac{1}{(2\pi\sigma^2)^{n/2}\det^{1/2}(\mathbf{R})}\exp\left(-\frac{1}{2\sigma^2}\left((\boldsymbol{y}\text{-}\boldsymbol{\mu})^T\mathbf{R}^{-1}(\boldsymbol{y}\text{-}\boldsymbol{\mu})\right)\right)$$

GaSP Prediction

Given data $\boldsymbol{y}$ the conditional (posterior) distribution of $y(\mathbf{x}^{\text{new}})$,

$$y(\mathbf{x}^{\text{new}}) \;|\{y(\mathbf{x}^{(1)}), \ldots, y(\mathbf{x}^{(n)})\} \text{ is } \mathrm{N}(\hat{y}(\mathbf{x}^{\text{new}}), \mathrm{v}(\mathbf{x}^{\text{new}})).$$

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

is the predictor of $y(\mathbf{x}^{\text{new}})$.

$(\mathbf{r}^T(\mathbf{x}) = (\mathrm{R}(\mathbf{x}, \mathbf{x}^{(1)}), \ldots, \mathrm{R}(\mathbf{x}, \mathbf{x}^{(n)}))$

GaSP - Choices of μ

Constant (Con)

Linear in **x** (FL)

Select linear: linear in "significant" coordinates (SL)

GaSP - Choices of R

R(**x**,**x**') = R(**x**–**x**') (stationary):

$$\mathrm{R}(\mathbf{x},\mathbf{x}') = \prod_{j=1}^{d} \exp(-\theta_j |x_j - x_j'|^{p_j}) \quad 1 \le p_j \le 2 \quad \textbf{PowExp}$$

When $\boldsymbol{p} = \mathbf{2}$, R = **Gauss**

Other choices - **Matérn**

Matérn Class

$\mathrm{R}(\mathbf{x},\mathbf{w})$ has its j^{th} factor one of

Matérn-1: $\left(1+\theta_j|x_j-w_j|\right)\exp\left(-\theta_j|x_j-w_j|\right)$

Matérn-2: $\left(1+\theta_j|x_j-w_j|+\frac{1}{3}\theta_j^2|x_j-w_j|^2\right)\exp\left(-\theta_j|x_j-w_j|\right)$

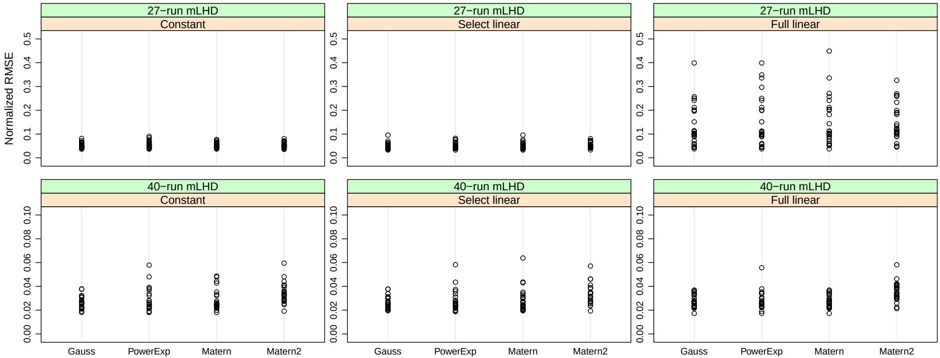
Matérn-0: Same as PowExp with $p_j=1$

Matérn-∞: Same as Gauss

Case 3: μ=Con, R=PowExp v Others

Test Function: Borehole (d=8)

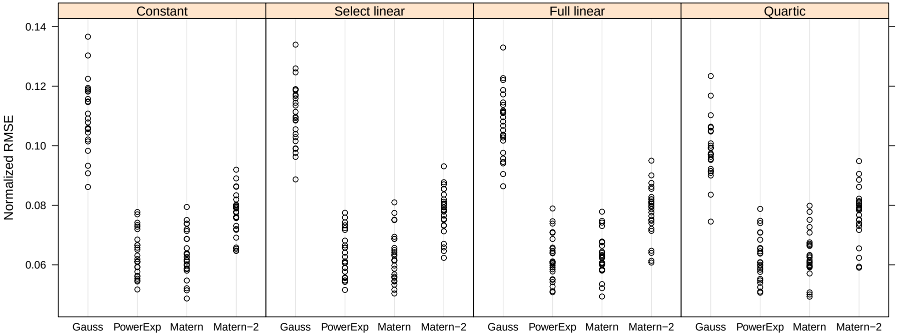
$$y = \frac{2\pi T_u (H_u - H_l)}{\log(r/r_w)\left(1 + \frac{2LT_u}{\log(r/r_w) r_w^2 K_w} + T_u/T_l\right)};$$



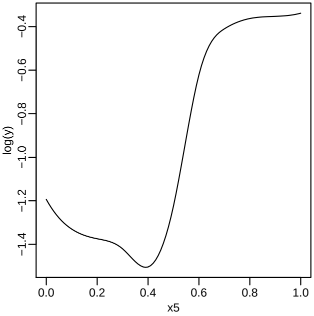
Case 3: μ,R comparisons; Borehole Function

Argument for Slow Codes

- Code, design, runs have already been made
- Use GaSP for prediction
- Select 20% of the runs at random for a holdout set
- Compute $e_{rmse,ho}$ with the remaining 80% using GaSP
- Repeat 25 times each with a new randomly selected holdout set



Case 3: Nilson-Kuusk (1989) 5-*d* reflectance plant canopy; *n*=250, so 25 holdout sets of 50 points each. Quartic: refers to quartic in x_5, linear in others (used by Bastos&O'Hagan (2009)



Nilson-Kuusk: Estimated main effect of x_5.

Decision – Case 3: μ, R

1. μ = Con: clear and convincing
2. No difference between R = PowExp or Matérn-opt
3. Do <u>not</u> rely solely on R = Gauss

Case 4: GaSP(Con,PowExp) v CGP

CGP (Ba & Joseph, 2012): $R = \text{Gauss}_1 + q(\mathbf{x})\text{Gauss}_2 q(\mathbf{x}')$;

Gauss_1 has correlation parameters (θ_1) bounded above to capture smooth global trend;

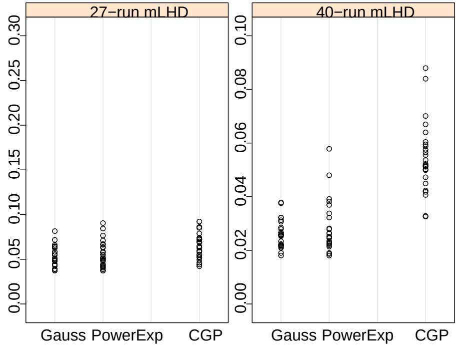
Gauss_2 has correlation parameters (θ_2) bounded below to capture short range volatility;

$q(\mathbf{x})$ allows non-stationary behavior for second term.

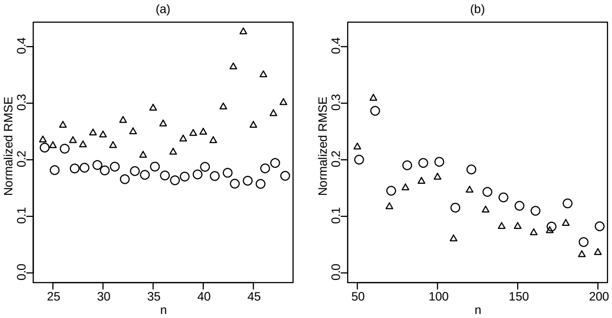
Test Functions:

(1) Borehole; (2) $\sin(1/(x_1 x_2))$ on $[0.3,1.0]^2$;

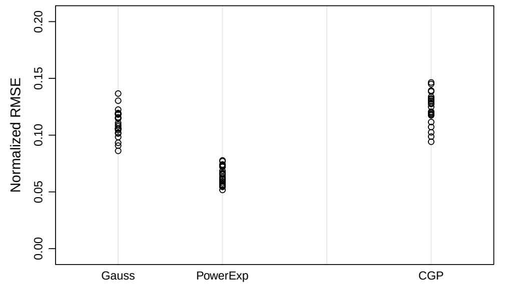
(3) Nilson-Kuusk; (4) Volcano (d=2) code



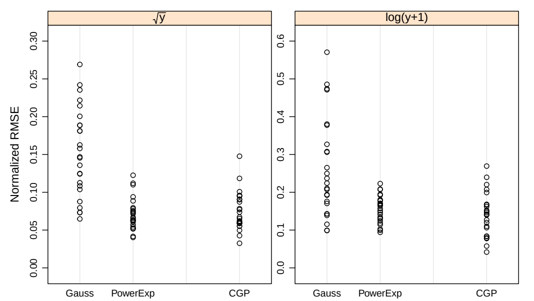
Case 4: PowExp v CGP; Test Function: Borehole



Case 4: GaSP(Con,PowExp)(triangles) v. CGP(circles); y=sin(1/(x_1x_2)) on $[0.3,1.0]^2$.



Case 4 : Nilson-Kuusk d=5, n=250 leading to 25 designs each of 200 runs with 50 holdout points.



Case 4 : Volcano (Bayarri et al, 2010) d=2, n=32 leading to 25 designs each of 27 runs with 5 holdout points.

Decision – Case 4: CGP?

1. Evidence is unclear when R=CGP is preferred to R= PowExp.

2. Plausible that non-stationary R is useful but what R, when and where is unclear.

3. Diagnostics indicating R=PowExp is inadequate point to follow-up strategies, but which ones?

Case stayed

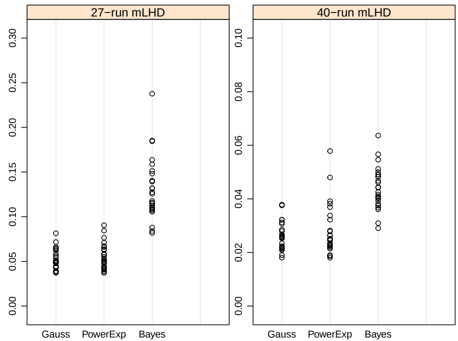
Flavors of Bayes

Empirical Bayes (**MLE**): $\max_{\mu,\sigma,\boldsymbol{\theta},\boldsymbol{p}}\ L(\boldsymbol{y}|\mu,\sigma,\boldsymbol{\theta},\boldsymbol{p})$

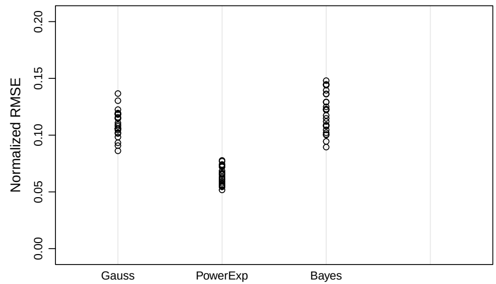
Bayes GEM-SA: $\boldsymbol{p}=\mathbf{2};\ \pi(\mu)=1,\ \pi(\sigma^2)=1/\sigma^2,\ \pi(\theta_j)=\exp(-.01\theta_j)$

Other Bayes: $\boldsymbol{p}=\mathbf{2};$ different priors for $\pi(\theta_j)$

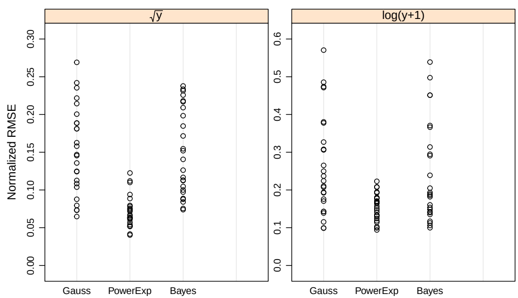
Hybrid Bayes: Get $\boldsymbol{\theta},\boldsymbol{p}$ via MLE, "plug-in", Bayes for μ,σ



Case 5: MLE v. Bayes (GEM-SA). Test function: Borehole



Case 5: Nilson-Kuusk *d*=5, *n*=250 leading to 25 designs each of 200 runs with 50 holdout points.



Case 5: Volcano (Bayarri et al, 2010) *d*=2, *n*=32 leading to 25 designs each of 27 runs with 5 holdout points.

Decision – Case 5: Bayes v MLE

1. Bayes with R=Gauss not better, sometimes worse than MLE with R = PowExp

2. Extend Bayes to allow $\boldsymbol{p} < \mathbf{2}$

Conclusion

Bertrand Russell, upon being asked what he would reply if, after dying, he were brought into the presence of God and asked why he had not been a believer :

"Not enough evidence God! Not enough evidence!"