The interplay between random field models for Bayesian spatial prediction and the design of computer experiments

Part II: Design of Computer Experiments Using Additive Gaussian Process Models

Jürgen Pilz

Dept. of Statistics
Alpen-Adria University of Klagenfurt
9020 Klagenfurt, Austria
juergen.pilz@aau.at

WIAS and School of Business and Economics
Humboldt-Universität zu Berlin
November 14, 2018

Introduction

- **Experiments**
  - physical experiments
  - Computer experiments (Computer-based simulations like FEM)

Which simulations to run?
Main difference: Computer Models are **deterministic**
Modification of classical DOE $\Rightarrow$ DOCE

Math. model: $y = \tilde{f}(x_1 \ldots, x_k)$, e.g. solution of ODE/PDE system

$$\mathbf{x} = (x_1, \ldots, x_k)^T \in \mathcal{X} \text{ = experimental domain}$$

replaced by *meta-model*

$$\mathbb{E}Y(\mathbf{x}) = f(x_1, \ldots, x_k), f \text{ "close" to } \tilde{f}$$

Introduction

Requirements for good designs:

- space filling property
- projective property
- computational efficiency

Compromise: **LHD**= Latin Hypercube Designs

w.l.o.g. experimental domain $\mathcal{X} = [0,1]^k$

Designs: $d_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{X}^n$,

n runs, k factors

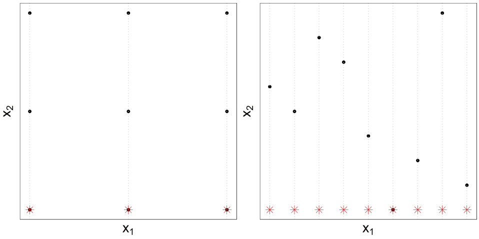


Figure : Regular (left) and latin hypercube design (right)

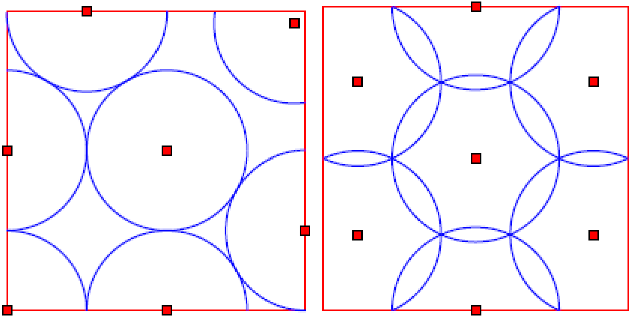


Figure : Maximin (left) and Minimax (right) designs

Introduction

Common uniformity measure:

$$\mathcal{D}(d_n) = \left\{ \int_{\mathcal{X}} \left\| \frac{1}{n} \#(d_n, [\mathbf{0}, \mathbf{x})) - Vol([0, \mathbf{x})) \right\|^p d\mathbf{x} \right\}^{1/p}$$

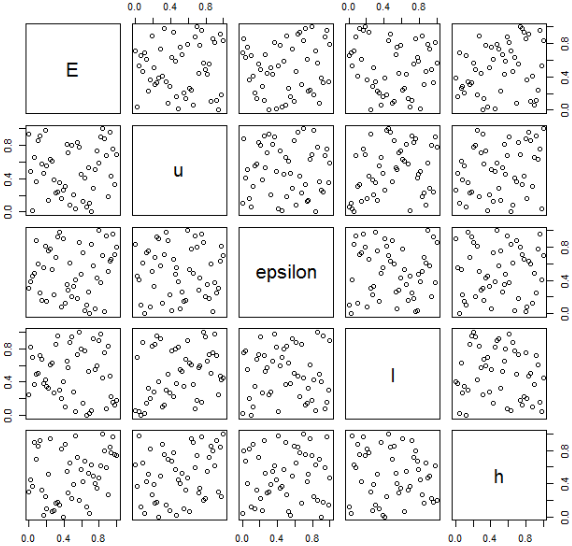
star L_p-**discrepancy**

Optimality critera: e.g. based on sample distribution

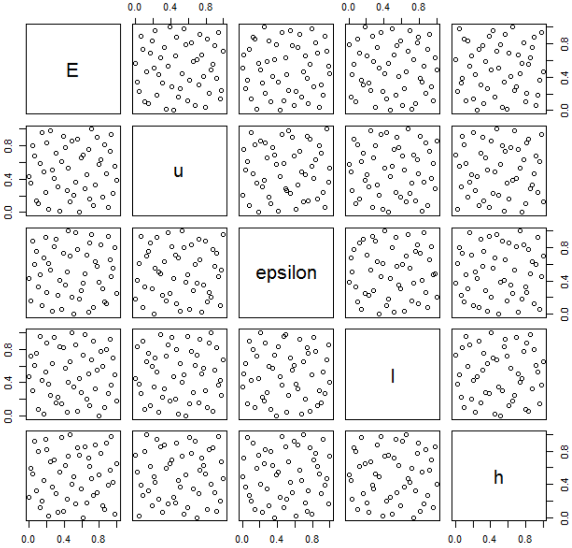
$$\mathbb{E}\left(- \ln \underbrace{p(\mathbf{y}(d_n))}_{\text{posterior d.}} \right) \longrightarrow Max_{d_n}$$

max. entropy criterion

Start design



Optimal design for outeri=500



RSM modelling

Classical approach: Regression (response surface) modelling $\Rightarrow$

prediction reduces to interpolation problem

e.g. quadratic RSM

$$y(\mathbf{x}) = \beta_0 + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j}^{k} \sum_{j=1}^{k} \beta_{ij} x_i x_j$$

For complex responses, LSE $\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}_d$

yields bad interpolations

Remedy

Kriging models

$\Rightarrow$ response = realization of stochastic process

$$Y(\mathbf{x}) = \underbrace{\mu(\mathbf{x})}_{\text{trend}} + \underbrace{Z(\mathbf{x})}_{\substack{\text{Gaussian Process} \\ \text{(zero mean)}}}$$

Effect: good approx. over a wide range of different designs and sample sizes and well-defined basis for statistical framework

$$Y(\cdot) \sim GP(\mu(\mathbf{x}), \sigma^2 R(\cdot, \cdot))$$

Main difference to geostatistical settings:
– $\mathbf{x}$ is not a spatial coordinate vector
– usually, higher dimensional settings: $k > 3$

Covariance structure

covariance function: $Cov(Z(\mathbf{x}_i), Z(\mathbf{x}_j)) = \underset{\substack{\downarrow \\ \text{process variance}}}{\sigma^2} \quad \underbrace{R(\mathbf{x}_i, \mathbf{x}_j)}_{\text{correlation function}}$

Common assumptions:

1. covariance-stationarity, i.e.

$$R(\mathbf{x}_i, \mathbf{x}_j) = R(\mathbf{x}_i - \mathbf{x}_j)$$

2. (tensor-)product correlation structure

$$R(\mathbf{x}_i, \mathbf{x}_j) = \prod_{m=1}^{k} \underbrace{R_m(|x_{im} - x_{jm}|)}_{\text{univariate Matérn c.f.}}$$

Correlation functions

e.g. $R_m(|x_{im} - x_{jm}|) = \exp(-|x_{im} - x_{mj}|^2/\theta_m^2)$

Gaussian correlation function

Flexibilization: $R_m(d) = \frac{(d/\theta_m)^\nu}{2^{\nu-1}\Gamma(\nu)} \underset{\substack{\downarrow \\ \text{Bessel function of order } \nu}}{\mathcal{K}_\nu}(d/\theta_m), d = |x_{im} - x_{jm}|$

ν = smoothness parameter

Correlation functions

Special cases:

$$\nu = \frac{1}{2} : \text{exponential c.f. } R_m(d) = \exp(-d/\theta_m)$$

$$\nu = \infty : \text{Gaussian c.f. } R_m(d) = \exp(-d^2/\theta_m^2)$$

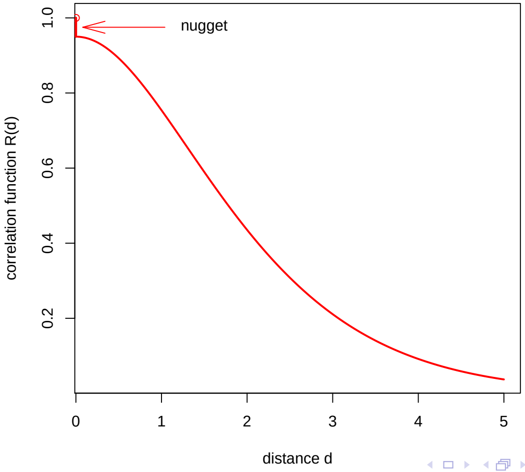
$$\nu = s + \frac{1}{2} : R_m(d) = P_s(d/\theta_m) * \exp(-d/\theta_m)$$

Common choices in DOCE: $\nu = \frac{3}{2}$ or $\nu = \frac{5}{2}$

+ small nugget (discontinuity) at the origin

Matérn c.f. $\nu = \frac{5}{2}$

$$R(d) = \left(1 - \frac{\tau^2}{\sigma^2}\right) * \left(1 + \frac{\sqrt{5}d}{\theta} + \frac{5d^2}{3\theta^2}\right) \exp\left(-\frac{\sqrt{5}d}{\theta}\right),\ d > 0$$



Estimation and Prediction

MLE: available for β and σ^2

$\hat{\beta} = (X^T R_n^{-1}(\theta) X)^{-1} X^T R_n^{-1}(\theta) \mathbf{y}_d$

$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y}_d - X\hat{\beta})^T R_n^{-1}(\theta)(\mathbf{y}_d - X\hat{\beta})$

$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)^T$ Gauss-Newton (or genetic optimiz.)

Optimal prediction:

$$\begin{aligned} \hat{Y}(\mathbf{x}_0) &= \mathbf{f}(\mathbf{x}_0)^T \hat{\beta} + \mathbf{r}_0^T R_n^{-1}(\mathbf{y}_d - X\hat{\beta}) \\ &= \text{GLSE + smoothed residual} \end{aligned}$$

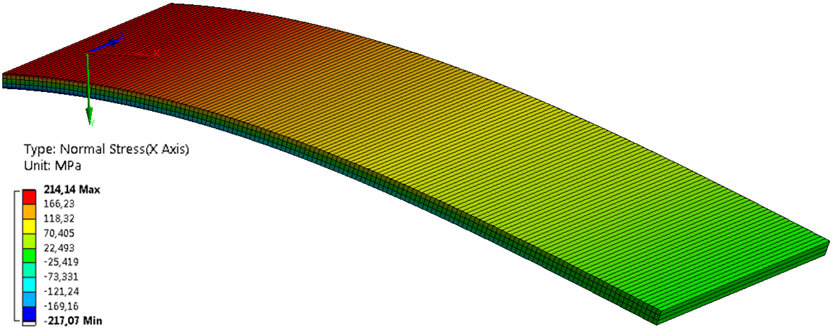
where $\mathbf{r}_0^T = (R(\mathbf{x}_0 - \mathbf{x}_1), \ldots, R(\mathbf{x}_0 - \mathbf{x}_n))$, $R_n =$ correl. matrix

Implementation in R: **DiceKriging**

Experiences

Stress testing in semiconductor processing for **thin wafers** (thickness$\leq 40\mu m$)

Kriging metamodel for stress prediction validated against Ramann spectroscopy measurements, FEM simulations



Robust Additive Gaussian Processes

Aims

- higher flexibility in meta-modelling
- numerical stability: robustness of parameter estimates, esp. for correlation parameters

Solution: Bayesian approach using additive models and (objective) reference priors

Side effect: high-dimensional optimization problems reduced to a few sub-routines of ≤ 3 dimensions

Additive model:

$$\mathbb{E}Y(\mathbf{x}) = f_0 + \sum_{i=1}^{k} f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \ldots + f_{12\ldots k}(x_{1,\ldots}, x_k)$$

Functional ANOVA Representation

Special case: first order GAM

$$\mathbb{E}Y(x_1, \ldots, x_k) = f_0 + \sum_{i=1}^{k} f_i(x_i)$$

$f_1, ..., f_k$: smooth basis functions

$\Rightarrow$ non-parametric modelling of main effects

Goal: Extension of classical GAM regression

Fo a good overview of the advantages of additive structures compared to fully parametric GP models in high dimensions see Dourante, Ginsbourger, Roustant (2012)

Robust Additive Gaussian Processes

Novelty of our recently proposed concept: Combination of AGP with robust reference priors proposed by Gu, Wang and Berger (submitted to AS 2017) + new sampling design scheme

Our new model: Second order Kriging AGP with

$$f_i \sim N(\mu_i, \sigma^2 R_i)$$

$$f_{ij} \sim N(\mu_{ij}, \sigma^2 R_i R_j)$$

Result: AGP $Y(\mathbf{x}) \sim N(\mu, \sigma^2 R(\cdot, \cdot))$, locally constant trend

and $R(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{k} R_i(x_i, x_i') + \sum_{i=1}^{k} \sum_{j=i+1}^{k} R_i(x_i, x_i') R_j(x_j, x_j') + \delta_{\mathbf{x}\mathbf{x}'} \tau^2$

Renormalization such that

$f^* := \frac{1}{\sqrt{m}} f \sim N\left(\frac{1}{\sqrt{m}}\mu, \sigma^2 R^*\right)$

$R^* := \frac{1}{m} R$ valid correlation matrix, $m = \#$ correlation components

Each component function has parameters

$\Theta_i = (\mu, \sigma^2, \theta_i, \tau_i^2)$ for 1st order terms

$\Theta_{ij} = (\mu, \sigma^2, \theta_i, \theta_j, \tau_{ij}^2)$ for 2nd order terms

Profile likelihood approach often fails, results in estimates $\hat{\theta}$ for which

$(*)$ $\quad R \approx I_n$ $\quad$ or $\quad R \approx \mathbf{1}_n \mathbf{1}_n{}^T$

$(\hat{\theta} \approx 0)$ $\quad$ singular corr.m.

$\downarrow$

bad ("impulse") prediction

Robust Additive Gaussian Processes

Remedy: robust Bayes prediction using reference priors of the form

$$\underset{\substack{\downarrow\\ \text{correl. parameters}}}{\pi^R(\mu, \sigma^2, \theta^*)} = \frac{\pi^R(\theta^*)}{\sigma^2}$$

where $\pi^R(\theta^*) \propto (det \;\; \underset{\substack{\downarrow\\ \text{exp. Fisher information}}}{I_F(\theta^*)})^{1/2}$

Explicit representations for $I_F(\theta^*)$ available in Kazianka & Pilz (2012)

Result: proper posteriors $p(\theta^*|\mathbf{y}_d)$

Simplified estimate: $\hat{\theta}^* = arg \;\; \max_{\theta^*} \;\; p(\theta^*|\mathbf{y}_d)$

posterior mode (to avoid MCMC)

Bayes Prediction

Bayes predictor of $Y(\mathbf{x}_0)$ for untried input $\mathbf{x}_0$ is based on the predictive distribution

$$p(Y_0|\mathbf{y}_d) = \int \underbrace{p(Y_0|\mathbf{y}_d, \theta^*)}_{Student-t} p(\theta^*|\mathbf{y}_d) d\theta^*$$

Simplification: Use plug-in predictor

$$\begin{aligned} \mu^* &:= E(Y_0|\mathbf{y}_d, \hat{\theta}^*) \\ &= \hat{\mu} + \mathbf{r}_0^T R_{\hat{\theta}^*}^{-1}(\mathbf{y}_d - \hat{\mu}\mathbf{1}_n) \\ \text{where } \hat{\mu} &= (\mathbf{1}_n{}^T R_{\hat{\theta}^*}^{-1}\mathbf{1}_n)^{-1}\mathbf{1}_n{}^T R_{\hat{\theta}^*}^{-1}\mathbf{y}_d \text{ GLSE} \end{aligned}$$

R-implementation fully described in Vollert, Ortner & Pilz (2017) is based on an iterative estimation scheme, using reparametrizations

Note: increasing nugget with increasing dimension k of input space

Sampling design

Due to additive structure, space-filling is important (for all variable projections)

Need **compromise** between LHD and regular grid designs: **Cut-FD** combines *HDMR Designs* based on a cut-center with *Factorial Designs*

$\Rightarrow n_0 = 2^k + 2k$ boundary points $+1$ cut-point

$< 10 \cdot k = n^*$ (recommended `min.size` for DOCE)

whenever $k \leq 5$

Add $(n^* - n_0)$ points along (ij)-planes of cut point $\mathbf{x}_c$

For $k > 5$ we recommend to use *Fractional Factorial Designs* instead of Full Factorials.

Example

3 commonly used test functions

Pepelyshev function: $x_i \in [0,1]; i = 1,2,3$

$$f_1(\mathbf{x}) = 4(x_1 - 2 + 8x_2 - 8x_2^2)^2 + (3 - 4x_2)^2 + 16\sqrt{x_3 + 1}(2x_3 - 1)^2$$

$n_1 = 31$ samples

Park function: $x_i \in [0,1]; i = 1,\ldots,4$

$$f_2(\mathbf{x}) = \tfrac{2}{3}\exp(x_1 + x_2) - x_4 \sin(x_3) + x_3$$

$n_2 = 41$ samples

Friedman function: $x_i \in [0,1]; i = 1,\ldots,5$

$$f_3(\mathbf{x}) = 10\sin(x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$$

$n_3 = 47$ samples

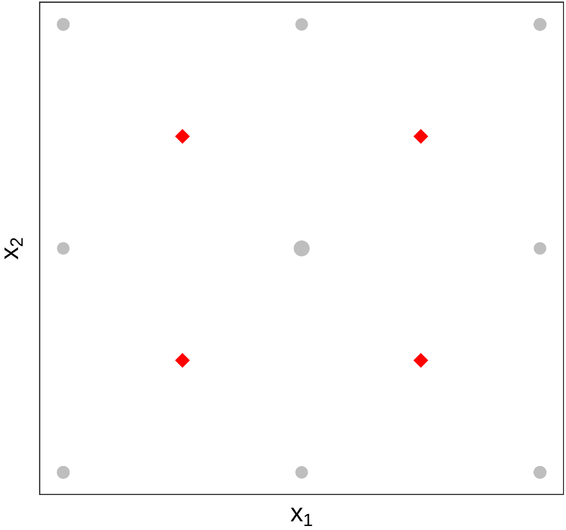


Figure : Initial size $n = 43$ points (one cut-point in the centre). Four points (red) added when the interaction f_{12} enters the model (final size $n = 47$)

Implementation

All calculations in R,
lhs package for constructing a maximin LHD,
DiceKriging package for constructing GP models

Setup:

- Matérn correlation with $\nu = 5/2$ for all components
- comparison for 3 designs: random LHD, maximin LHD and Cut-FD

Criterion for comparisons: MAPE = mean absolute prediction error, measured (in %) at 25000 positions (generated by 50 random designs each containing 500 points)

Results

1. Cut-FD can better determine the actual structure of the test functions than maximin and other LHDs (found exact set of components for Pepelyshev and Friedman functions, maximin did not)
2. Maximin LHD design was best with regard to MAPE (pred. power): For 5D-Friedman function MAPE $< 4\%$ (based on only $n_3 = 47$ sample points!)
3. **Robust AGP** model **outperforms** commonly used **GP** models for all three test functions
4. Simple random LHDs are least appropriate for approximation, getting even worse with increasing dimension

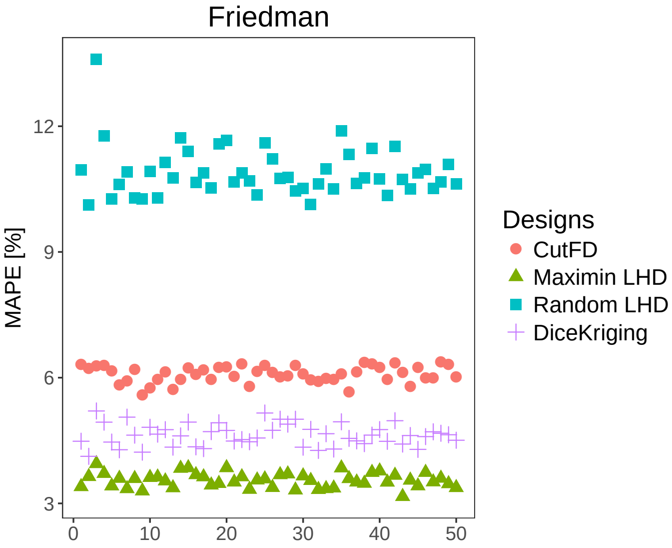
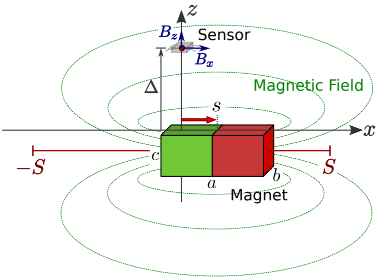


Figure : MAPE values of 50 validation LHDs for 5D-Friedman test function

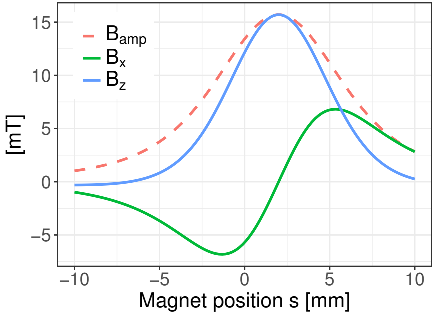
Work in progress:

AGP modelling for real DOCE applications based on FEMs for geometric and material parameter optimization problems, e.g. Magnetic field shaping for position and orientation detection systems

a)



b`

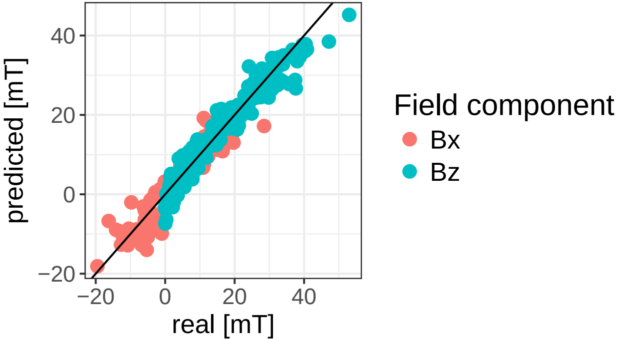


Model Performance based on random LHD_{250}

Component functions chosen by our algorithm:

– for B_x : f_s, f_b, f_{cs}, f_{as}

– for B_z : $f_s, f_a, f_b, f_c, f_{as}, f_{bM}, f_{ab}, f_{bs}, f_{cs}$



References

O. Roustant, D. Ginsbourger, Y. Deville, DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization, Journal of Statistical Software 51 (2012) 1, 1–55

H. Kazianka, J. Pilz, Objective Bayesian analysis of spatial data with uncertain nugget and range parameters, The Canadian Journal of Statistics 40 (2012), 304–327

M. Gu, X. Wang, J. O. Berger, Robust Gaussian stochastic process emulation, submitted to the Annals of Statistics (2017)

N. Vollert, M. Ortner and J. Pilz, Robust Additive Gaussian Process Models Using Reference Priors and Cut-Off-Designs, J. Applied Mathematical Modelling 65 (2019) 586-596, https://doi.org/10.1016/j.apm.2018.07.050

J. Mure, Propriety of the reference posterior distribution in Gaussian Process regression, submitted to the Annals of Statistics (2018)