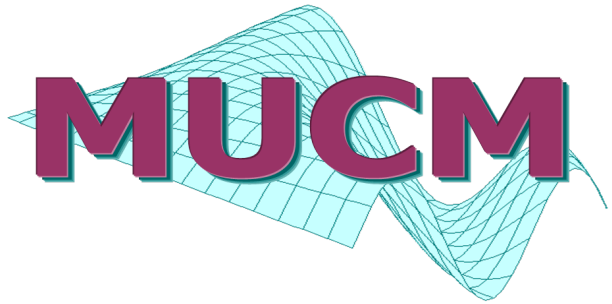
Experimental Design for the Heteroscedastic Model

Alexis Boukouvalas, Dan Cornford

Neural Computing Research Group, Aston University



July 9rd, 2009

Overview

- Refresher of Heteroscedastic Gaussian Process Emulator.
  - Experimental results on Yuhba and Rabies model.
- Experimental Design using the Fisher Information Matrix.
  - Motivation.
  - Derivation.
  - Experimental results:
    - Monotonicity.
    - Submodularity.
    - Optimization space using Exhaustive Search.
    - Design Criterion Test.
- Open Questions and Conclusions.

Aston University MUCM

Random Output Simulators

Stochastic simulator

A mapping that produces random output given a fixed set of inputs.

Observational model

$$y_i(x_i) = t_i(x_i) + \varepsilon(x_i) \tag{1}$$

Aston University MUCM

Algorithm

Main idea

Use a coupled system of GPs to evaluate the mean and variance.

- We estimate a standard homoscedastic GP: $\mathbf{G_H}$ by maximum likelihood on the two sets of observations $t = \{t_r, t_s\}$.
- We train a GP on the log(variance) $\mathbf{G_S}$.
  - For set r we use the corrected sample variance.
  - For set s we sample from $\mathbf{G_H}$ to estimate the variance at that point.
- Estimate the heteroscedastic GP $\mathbf{G_M}$ to jointly predict the mean and variance (next slide).
- If s non empty, set $\mathbf{G_H} = \mathbf{G_M}$ and repeat from step 2 until convergence.

Aston University MUCM

Algorithm

Main idea

Use a coupled system of GPs to evaluate the mean and variance.

- We estimate a standard homoscedastic GP: $\mathbf{G_H}$ by maximum likelihood on the two sets of observations $t = \{t_r, t_s\}$.
- We train a GP on the log(variance) $\mathbf{G_S}$.
  - For set r we use the corrected sample variance.
  - For set s we sample from $\mathbf{G_H}$ to estimate the variance at that point.
- Estimate the heteroscedastic GP $\mathbf{G_M}$ to jointly predict the mean and variance (next slide).
- If s non empty, set $\mathbf{G_H} = \mathbf{G_M}$ and repeat from step 2 until convergence.

Aston University MUCM

Algorithm

Main idea

Use a coupled system of GPs to evaluate the mean and variance.

- We estimate a standard homoscedastic GP: $\mathbf{G_H}$ by maximum likelihood on the two sets of observations $t = \{t_r, t_s\}$.
- We train a GP on the log(variance) $\mathbf{G_S}$.
  - For set r we use the corrected sample variance.
  - For set s we sample from $\mathbf{G_H}$ to estimate the variance at that point.
- Estimate the heteroscedastic GP $\mathbf{G_M}$ to jointly predict the mean and variance (next slide).
- If s non empty, set $\mathbf{G_H} = \mathbf{G_M}$ and repeat from step 2 until convergence.

Aston University MUCM

Algorithm

Main idea

Use a coupled system of GPs to evaluate the mean and variance.

- We estimate a standard homoscedastic GP: $\mathbf{G_H}$ by maximum likelihood on the two sets of observations $t = \{t_r, t_s\}$.
- We train a GP on the log(variance) $\mathbf{G_S}$.
  - For set r we use the corrected sample variance.
  - For set s we sample from $\mathbf{G_H}$ to estimate the variance at that point.
- Estimate the heteroscedastic GP $\mathbf{G_M}$ to jointly predict the mean and variance (next slide).
- If s non empty, set $\mathbf{G_H} = \mathbf{G_M}$ and repeat from step 2 until convergence.

Aston University MUCM

$\mathbf{G_M}$ estimation

- For set r the target values are the sample means, not the random individual samples of the underlying process. Since mean is distributed as $N(m, \sigma^2/n)$ we have to divide by number of realizations when predicting the mean of the training points.
- The predictive distribution equations are[1]:

$$\begin{aligned} \mu_* &= K^*(K + RN^{-1})^{-1} t \\ \Sigma_* &= K^{**} + R^* - K^{*^T}(K + RN^{-1})^{-1} K^* \end{aligned}$$

  where
  - $K = c(.,.)$ the training point covariance.
  - $R = diag[r(x_1) \dots r(x_N)]$ the variance estimate from $\mathbf{G_S}$. Diagonal since we assume independent noise.
  - $N = diag(n_1 \dots n_N)$ the number of samples at each training point.
  - K^*, K^{**} and R^* the corresponding test point matrices.
- We use the most likely value of the variance from $\mathbf{G_S}$. Another option would be Monte Carlo.

[1] We omit the mean function although inclusion is straightforward.

Aston University MUCM

$\mathbf{G_M}$ estimation

- For set r the target values are the sample means, not the random individual samples of the underlying process. Since mean is distributed as $N(m, \sigma^2/n)$ we have to divide by number of realizations when predicting the mean of the training points.
- The predictive distribution equations are[1]:

$$\begin{aligned} \mu_* &= K^*(K + RN^{-1})^{-1} t \\ \Sigma_* &= K^{**} + R^* - K^{*^T}(K + RN^{-1})^{-1} K^* \end{aligned}$$

where

  - $K = c(.,.)$ the training point covariance.
  - $R = diag[r(x_1) \ldots r(x_N)]$ the variance estimate from $\mathbf{G_S}$. Diagonal since we assume independent noise.
  - $N = diag(n_1 \ldots n_N)$ the number of samples at each training point.
  - K^*, K^{**} and R^* the corresponding test point matrices.

- We use the most likely value of the variance from $\mathbf{G_S}$. Another option would be Monte Carlo.

[1] We omit the mean function although inclusion is straightforward.

Aston University MUCM

$\mathbf{G_M}$ estimation

- For set r the target values are the sample means, not the random individual samples of the underlying process. Since mean is distributed as $N(m, \sigma^2/n)$ we have to divide by number of realizations when predicting the mean of the training points.
- The predictive distribution equations are[1]:

$$\begin{aligned} \mu_* &= K^*(K + RN^{-1})^{-1} t \\ \Sigma_* &= K^{**} + R^* - K^{*^T}(K + RN^{-1})^{-1} K^* \end{aligned}$$

where

  - $K = c(.,.)$ the training point covariance.
  - $R = diag[r(x_1) \ldots r(x_N)]$ the variance estimate from $\mathbf{G_S}$. Diagonal since we assume independent noise.
  - $N = diag(n_1 \ldots n_N)$ the number of samples at each training point.
  - K^*, K^{**} and R^* the corresponding test point matrices.

- We use the most likely value of the variance from $\mathbf{G_S}$. Another option would be Monte Carlo.

[1] We omit the mean function although inclusion is straightforward.

Aston University MUCM

Heteroscedastic Model: Joint Process Derivation

Our observation equation at a given design point x_i is:

$$t_i(x_i) = y_i(x_i) + \varepsilon(x_i)$$

Likelihood:

$$p(\bar{t}_i|\bar{y}_i) = p(\bar{t}_i|y_i) = N(\bar{t}_i|y_i, \frac{\sigma^2(x_i)}{n_i}),$$

where n_i the number of replicate observations and $\sigma^2(x_i)$ the true variance at location x_i. We estimate the true variance by the predictive mean of $\mathbf{G_S}$.

Due to independence of the noise we can write the likelihood in matrix form for all observations $1 \ldots N$:

$$p(\bar{\mathbf{t}}|\bar{\mathbf{y}}) = p(\bar{\mathbf{t}}|\mathbf{y}) = N(\bar{\mathbf{t}}|\mathbf{y}, RP^{-1}),$$

where $R = diag(\sigma^2(x_i))_{i=1}^N$ and $P = diag(n_i)_{i=1}^N$.

Our zero mean GP prior is:

$$p(\mathbf{y}) = N(\mathbf{y}|0, K).$$

Aston University MUCM

Heteroscedastic Model: Joint Process Derivation Continued

The marginal observation density can then be calculated:

$$\begin{aligned} p(\bar{\mathbf{t}}) &= \int p(\bar{\mathbf{t}}|\mathbf{y})p(\mathbf{y})dy = \int N(\bar{\mathbf{t}}|\mathbf{y}, RP^{-1})N(\mathbf{y}|0, K)dy \\ &= N(\bar{\mathbf{t}}|0, C_\mu = K + RP^{-1}). \end{aligned}$$

Condition on the known sites to obtain predictive distribution:

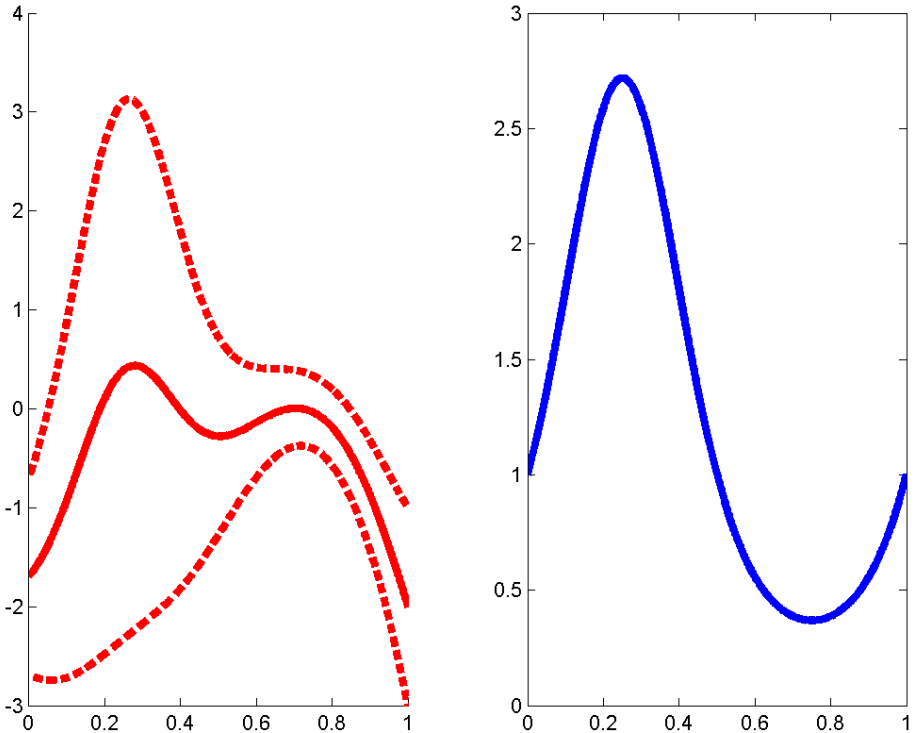
$$\begin{aligned} p(\bar{\mathbf{t}}_*|\bar{\mathbf{t}}) &= N(K(x_*, x)^T(K(x, x) + R(x)P(x)^{-1})^{-1}\bar{\mathbf{t}} \\ &, \quad K(x_*, x_*) + R(x_*)P(x_*)^{-1} \\ &+ \quad K(x_*, x)^T(K(x, x) + R(x)P(x)^{-1})^{-1}K(x_*, x)). \end{aligned}$$

Aston University MUCM

Heteroscedastic Simulated Example

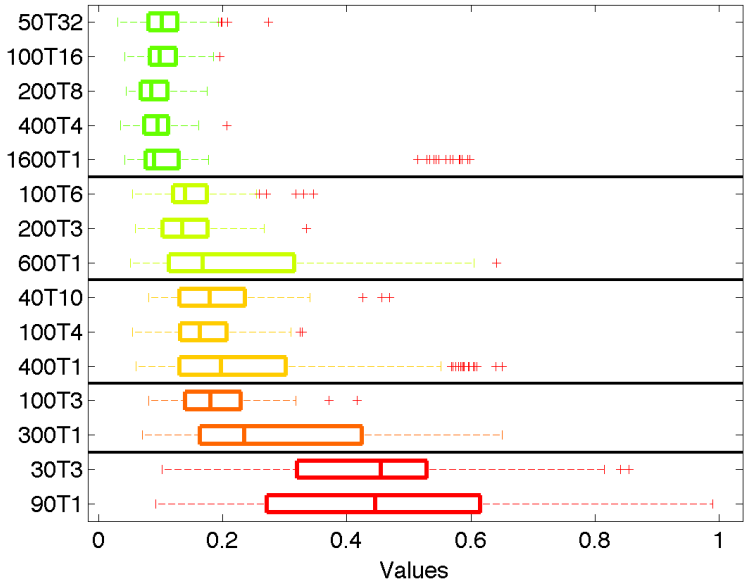
'Yuhba' function

$y = 2(e^{-30(x-0.25)^2} + \sin(\pi x^2)) - 2 + exp(sin(2\pi x))N(0,1)$, where $N(0,1)$ is the standard normal distribution.

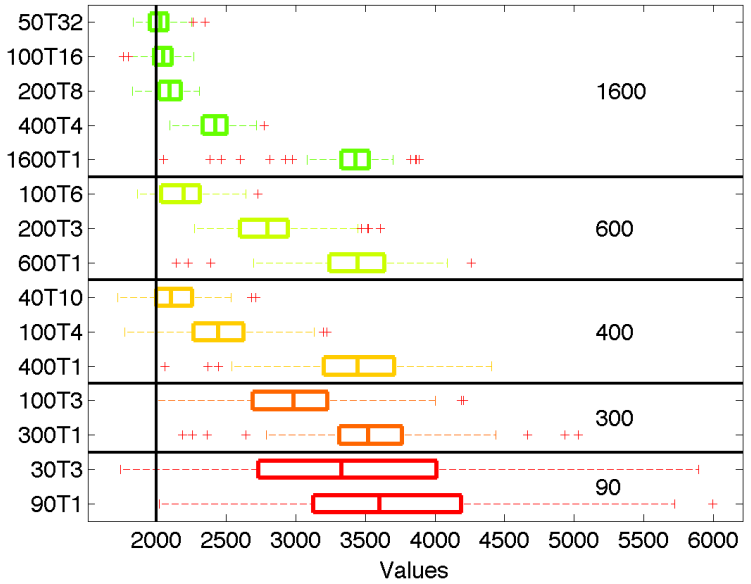


Aston University MUCM

Yuhba function: Total number of evaluations fixed



(b) Root Mean Squared Error



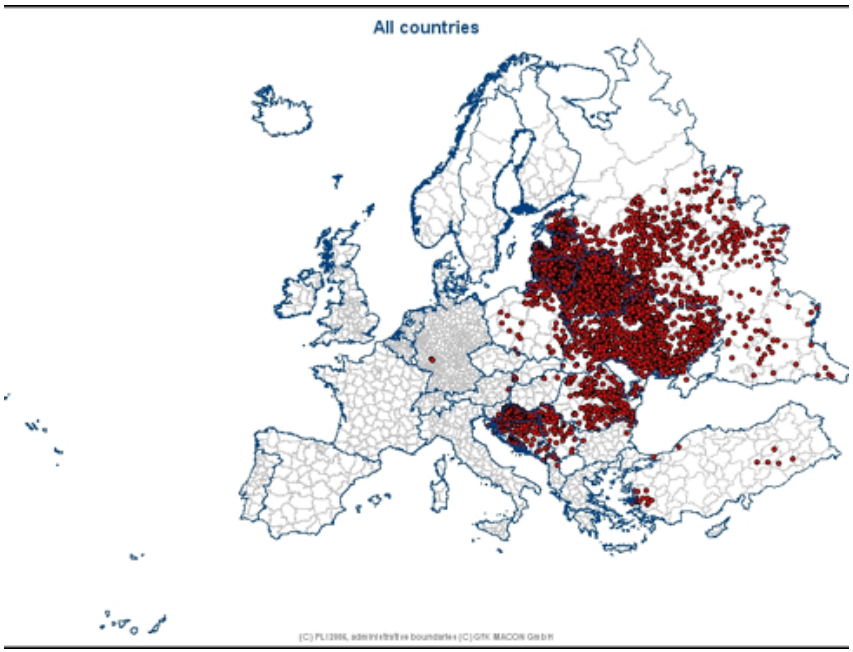
(c) Mahalanobis

Comparison of emulator fit where the total number of simulator evaluations is fixed at different levels. Notation: 30T3 = 30 design points each with 3 replicates. Results shown for a total of 90, 300, 400, 600 and 1600 total number of simulator evaluations.

Aston University MUCM

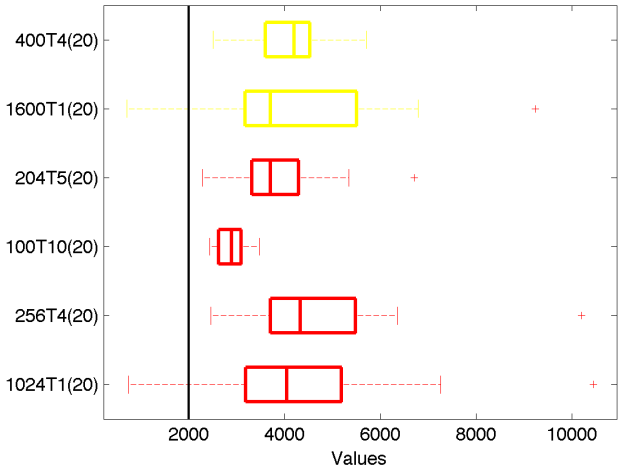
CSL Simulator Overview

- Rabies disease propagation simulator with two vector species: raccoon dogs and foxes.
- Two types of output: time series and summary statistics for each run.
- Stochastic simulator. Output is stochastic but not normally distributed.
- 14 inputs, 1 output (Time To Disease Extinction)

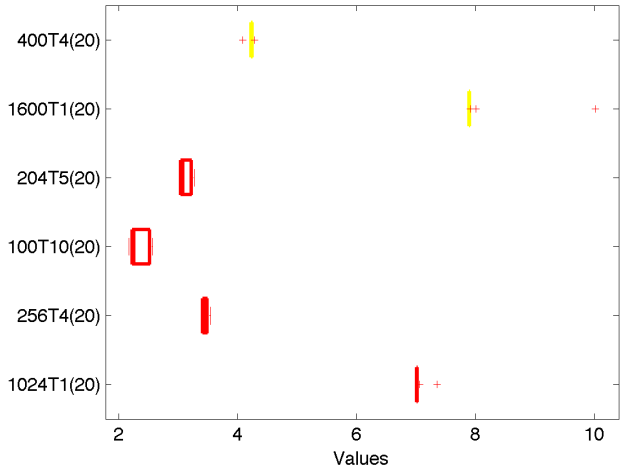


Aston University MUCM

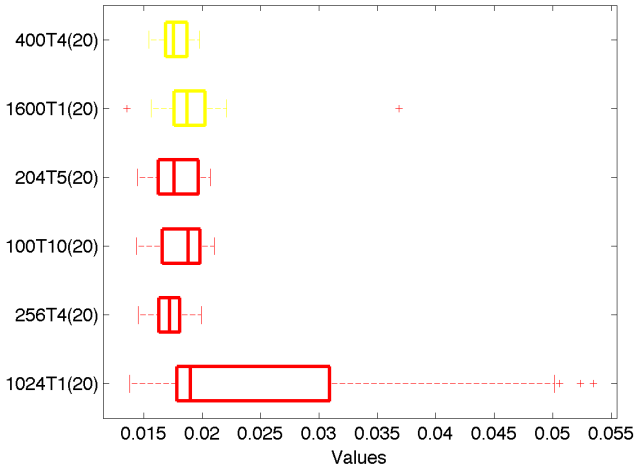
Performance



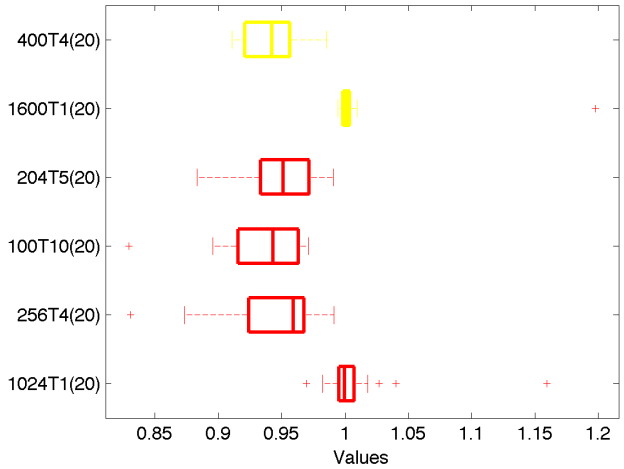
(a) Mahalanobis



(b) Elapsed Time



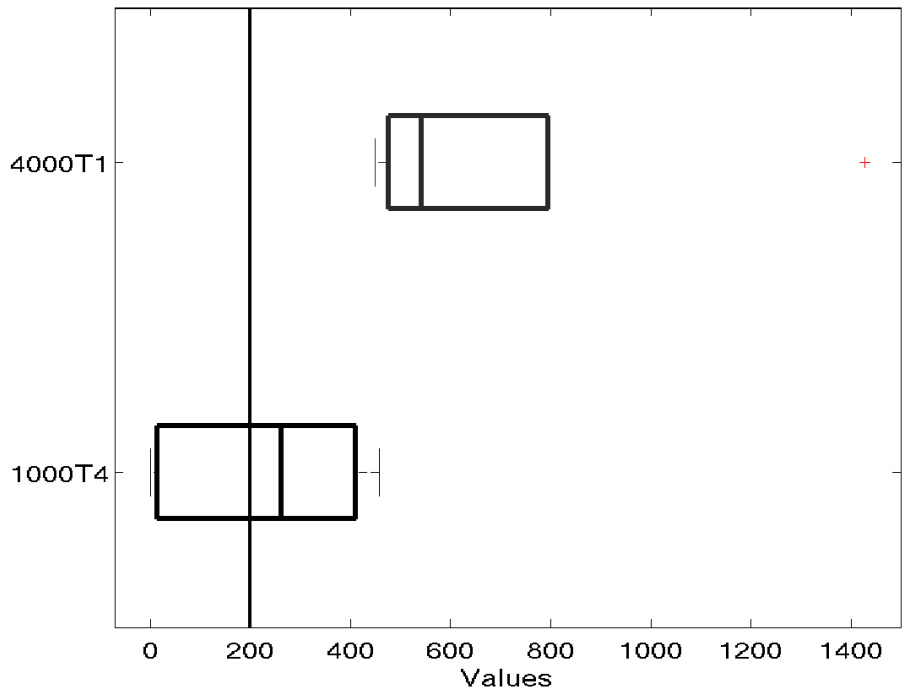
(c) RMSE



(d) MSE Variance

Aston University MUCM

Comparing sparse approximation methods to replicate design



Mahalanobis Error of Projected process 'Kersting' (4000 design points using 1000 support points) vs replicated design (1000 design points $\times$ 4 replicates).

Aston University MUCM

Interpretation of Gaussian Process

- All input factors have been sphered so length scales can be used for importance ranking.
- With mean function length scales apply to residual process only.

Interpreting the variance emulator ($\mathbf{G_S}$) by looking at the regression coefficients (Coeff) and correlation length scales (Scale).

Factor	Coeff	Factor	Scale
Rac Density	0.1608	Rac Rabid	1.4281
Rac Death	0.0633	Fox Inf	1.4594
Rac Birth	0.0200	Fox Rabid	1.5047

Aston University MUCM

Fisher Information Matrix

The FIM is $p \times p$ symmetric matrix where p the number of unknown parameters:

$$\mathbf{F} = -\int [\frac{\partial^2}{\partial \theta_i \theta_j} [ln(f(\mathbf{X};\theta)] f(\mathbf{X};\theta) d\mathbf{X}$$

Given $\mathbf{X}$ distributed as $N(\mu(\theta), \Sigma(\theta)$, the i,j element of the FIM is:

$$\mathbf{F}_{ij} = \frac{\partial \mu^T}{\partial \theta_i} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j} + \frac{1}{2} tr(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j}) \quad (2)$$

Aston University MUCM

Derivatives of the joint mean function

The derivatives of the joint mean function with respect to the covariance parameters of $\mathbf{G_M}$ and $\mathbf{G_S}$ are:

$$\frac{\partial \mu_{\mathbf{G_M}*}}{\partial \theta_\mu} = \frac{\partial K_{\mu*}^T}{\partial \theta_\mu} C_\mu^{-1} \bar{\mathbf{t}} - K_{\mu*}^T C_\mu^{-1} \frac{\partial C_\mu}{\partial \theta_\mu} C_\mu^{-1} \bar{\mathbf{t}}, \tag{3}$$

$$\frac{\partial \mu_{\mathbf{G_M}*}}{\partial \theta_\Sigma} = -K_{\mu*}^T C_\mu^{-1} \frac{\partial R}{\partial \theta_\Sigma} P^{-1} C_\mu^{-1} \bar{\mathbf{t}}, \tag{4}$$

The R matrix is a diagonal with elements $R_{ii} = exp(r(x_i))$ and hence the derivative is $\frac{\partial R_{ii}}{\partial \theta_\Sigma} = exp(r(x_i)) \frac{\partial r(x_i)}{\partial \theta_\Sigma}$. $r(x_i)$ is the most likely prediction of the variance from $\mathbf{G_S}$ at point x_i. Hence the derivative is

$$\frac{\partial r(x_i)}{\partial \theta_\Sigma} = \frac{\partial K_{\Sigma*}}{\partial \theta_\Sigma}^T C_\Sigma^{-1} \lambda^2 - K_{\Sigma*}^T C_\Sigma^{-1} \frac{\partial C_\Sigma}{\partial \theta_\Sigma} C_\Sigma^{-1} \lambda^2. \tag{5}$$

Aston University MUCM

Derivatives of the joint variance

The derivatives of the joint variance function with respect to the covariance parameters of $\mathbf{G_M}$ and $\mathbf{G_S}$ are:

$$\frac{\partial \Sigma_{\mathbf{G_M}*}}{\partial \theta_\mu} = \frac{\partial K_{\mu**}}{\partial \theta_\mu} - \Xi - \Xi^T + K_{\mu*}^T C_\mu^{-1} \frac{\partial K_\mu}{\partial \theta_\mu} C_\mu^{-1} K_{\mu*}, \tag{6}$$

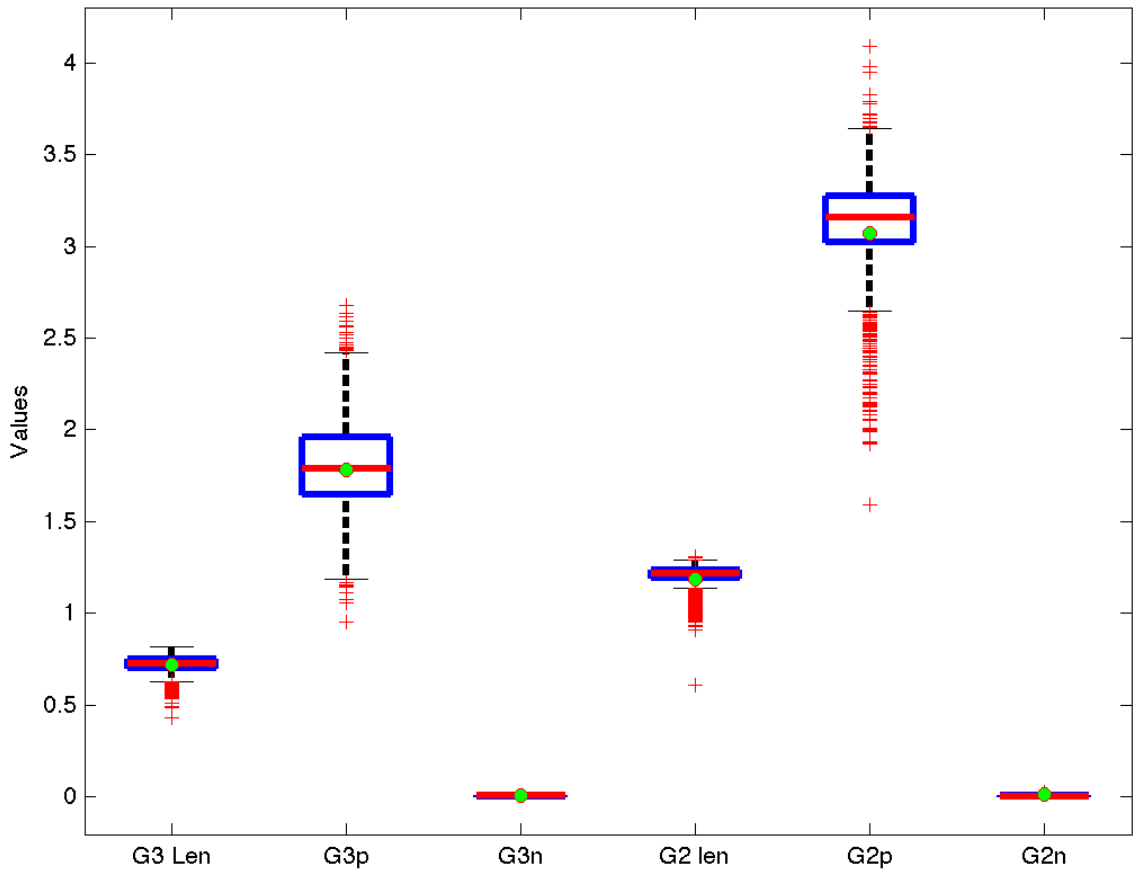
$$\frac{\partial \Sigma_{\mathbf{G_M}*}}{\partial \theta_\Sigma} = \frac{\partial R_*}{\partial \theta_\Sigma} P_*^{-1} + K_{\mu*}^T C_\mu^{-1} \frac{\partial R}{\partial \theta_\Sigma} P^{-1} C_\mu^{-1} K_{\mu*}, \tag{7}$$

where $\Xi = \frac{\partial K_{\mu*}^T}{\partial \theta_\mu} C_\mu^{-1} K_{\mu*}$.

Aston University MUCM

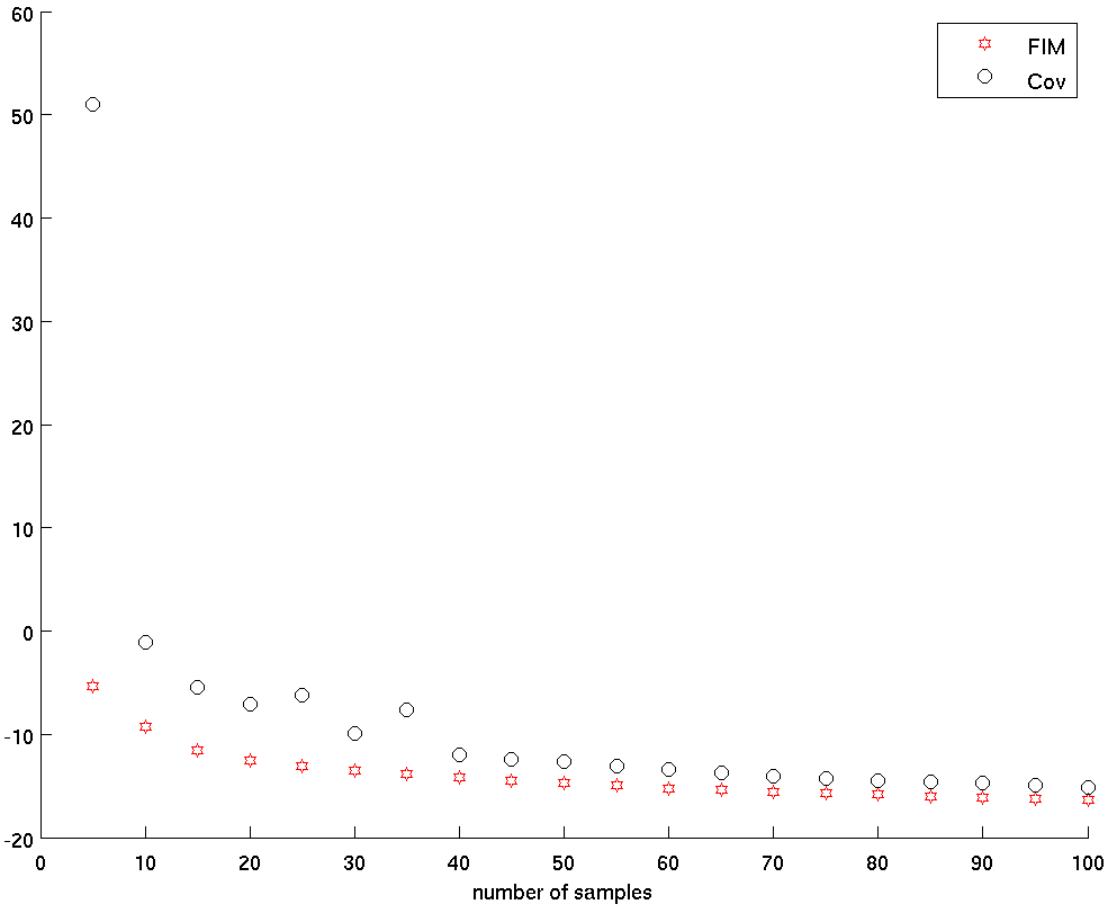
Empirical Parameter Covariance

We compute the empirical covariance of squared exponential kernel using 2000 samples of a Heteroscedastic GP.



Aston University MUCM

Monotonicity



Aston University MUCM

Submodularity

Definition

A function F is submodular iff for $A \subset B$ and $\varepsilon \setminus B$:

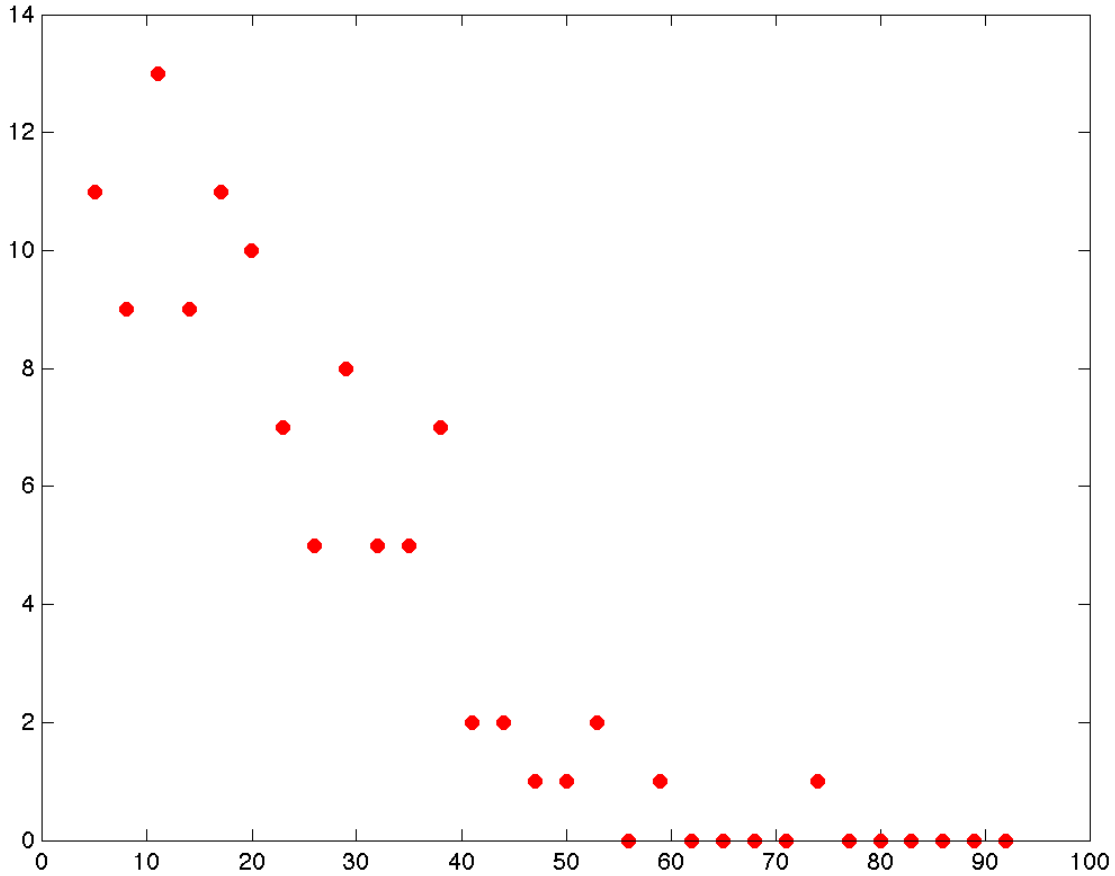
$$F(A \cup \varepsilon) - F(A) \geq F(B \cup \varepsilon) - F(B)$$

Nemhauser et al, 1978 Theorem

If F is a monotone submodular function over a finite ground set with $F(0) = 0$ then the greedy optimization algorithm is within $(1 - (\frac{k-1}{k})^k)$ constant factor of the optimal strategy for k design points.

Aston University MUCM

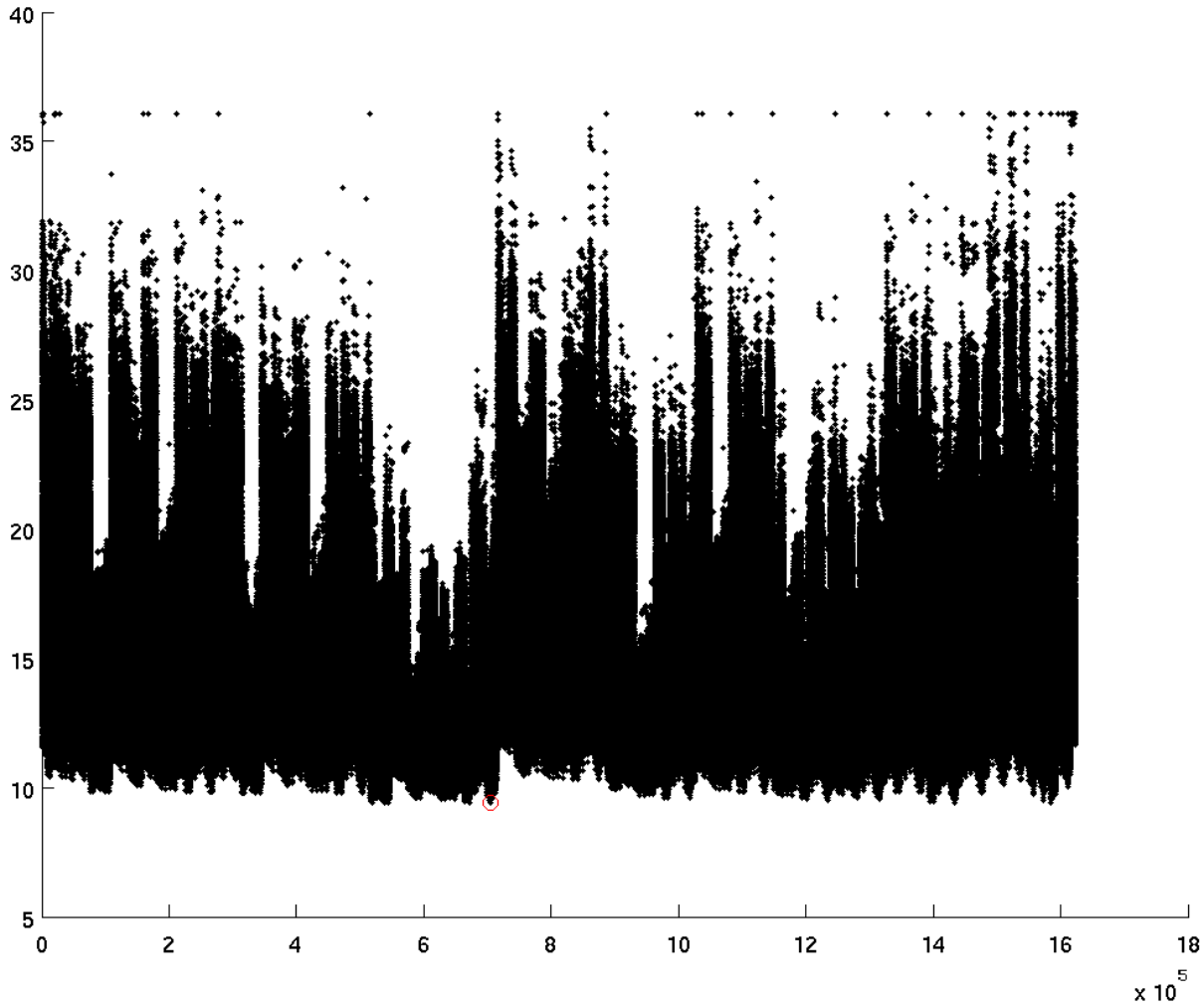
Is the FIM a submodular function?



X-axis: Design Size, Y Axis: number of violations over 100 realizations.

Aston University MUCM

Optimization space using Exhaustive Search

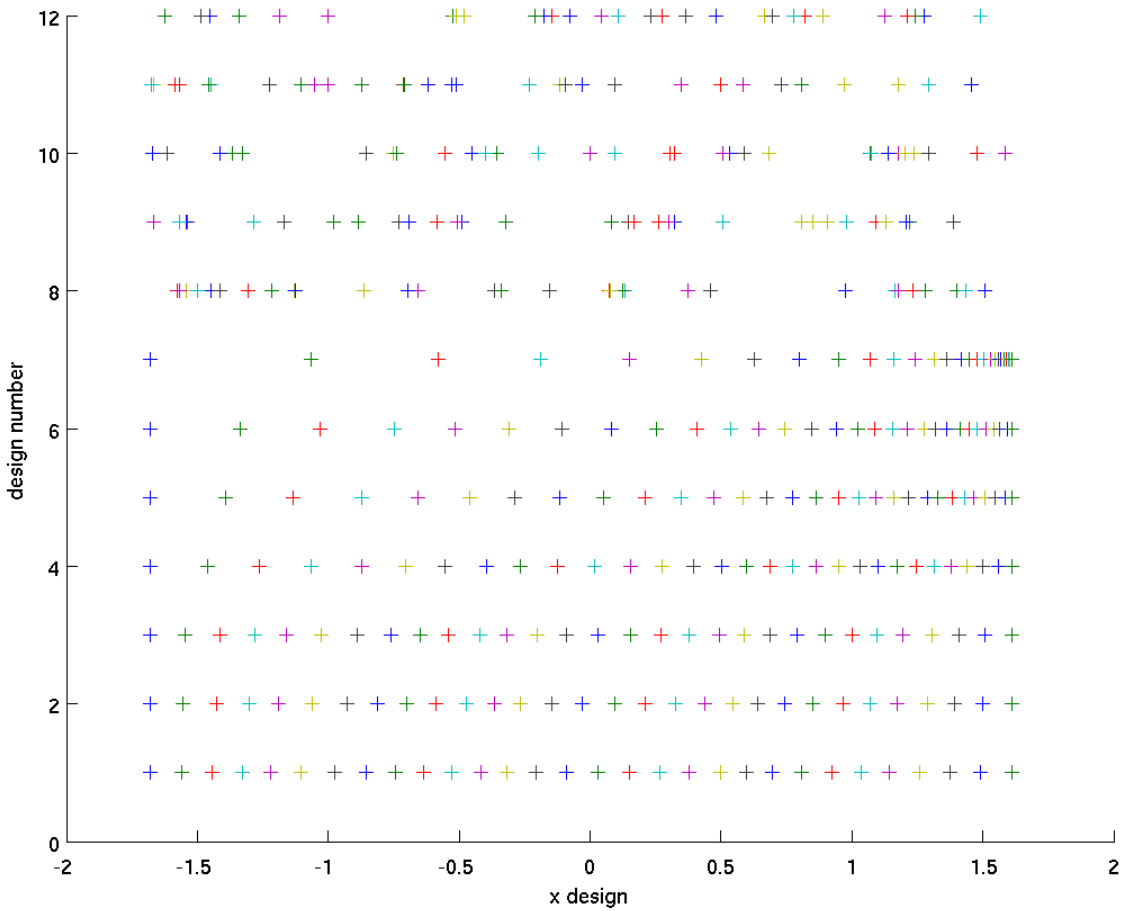


Using FIM pick 6 points from 30 point candidate set (1,623,160).

Aston University MUCM

Design Criterion Experiment: Designs Used

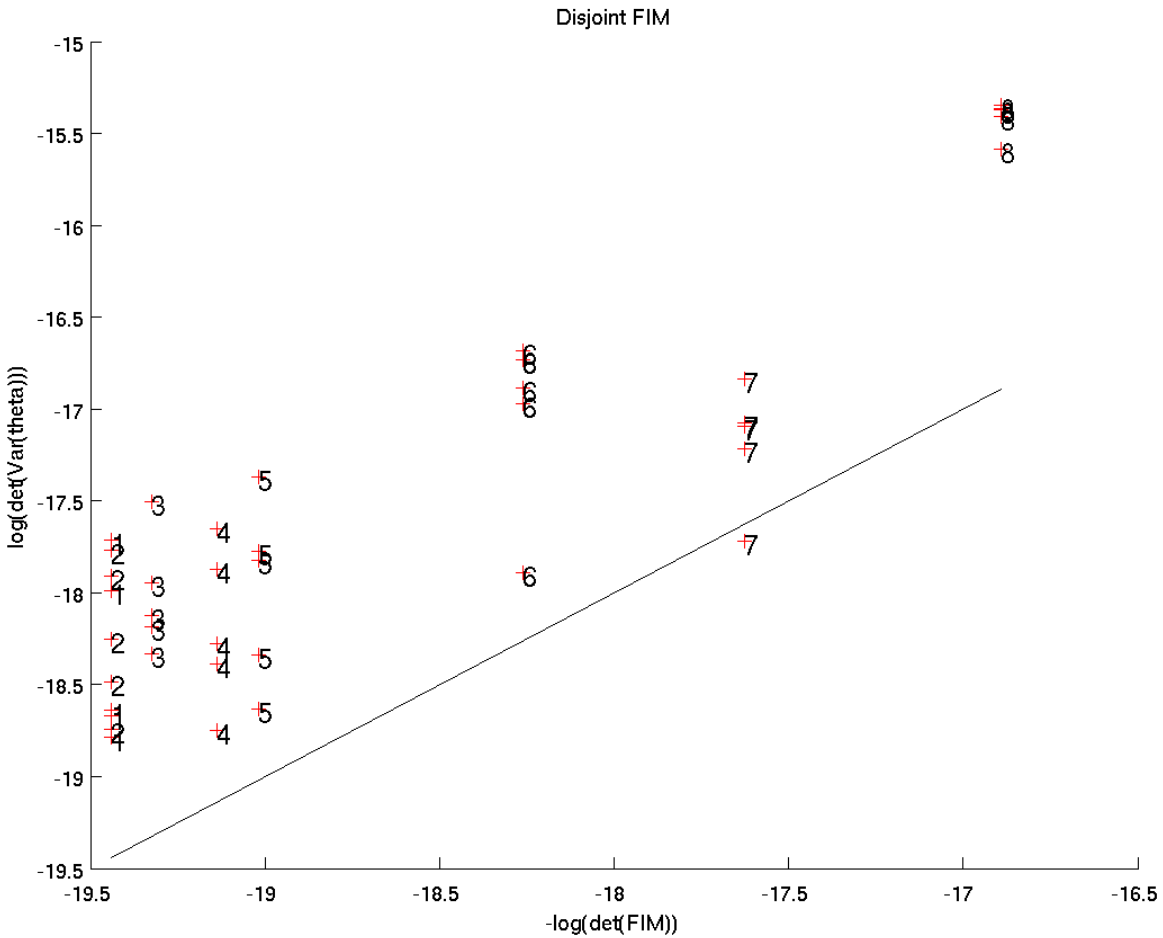
Designs considered



Aston University MUCM

Design Criterion Test

Empirical Parameter Variance vs FIM



Aston University MUCM

Summary

Heteroscedastic Framework

- Approach improves upon existing methods both in terms of accuracy and computational efficiency, in terms of inference and prediction time.
- In combination with a discrepancy model and real-world observations, this method could facilitate the efficient statistical calibration of stochastic simulators.

Open Questions

- Is the FIM a good design criterion for correlated non-linear models? In Zhu and Stein monotonicity of Fisher Information matrix to empirical parameter variance is shown but this is empirical evidence only.
- When doing maximum likelihood for the GP parameters, are they identifiable and consistently estimable? In some cases (see Stehlnik) Matern parameters are not. Under what conditions are they consistently estimable for the squared exponential?

References

- K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard. ”Most likely heteroscedastic gaussian process regression”. In Proc. 24th International Conf. on Machine Learning, 2007.
- Paul W. Goldberg and Christopher K. I. Williams and Christopher M. Bishop. ”Regression with Input-dependent Noise: A Gaussian Process Treatment”. Advances in Neural Information Processing Systems. The MIT Press, 1998.
- Werner Muller and Milan Stehlik. ”Issues in the Optimal Design of Computer Experiments.”. IFAS Research Paper Series, July 2007.
- Werner Muller and Milan Stehlik. ”Issues in the Optimal Design of Computer Experiments.”. IFAS Research Paper Series, July 2007.
- Zhengyuan Zhu and Michael L. Stein. ”Spatial sampling design for parameter estimation of the covariance function”. Journal of Statistical Planning and Inference 2005

Aston University MUCM