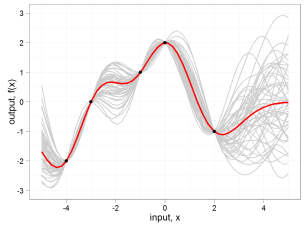
Design for screening, emulation and calibration



MANCHESTER 1824
The University of Manchester

http://personalpages.manchester.ac.uk/staff/alexis.boukouvalas/

Joint work with Dan Cornford, Milan Stehlík, J.P. Gosling and H. Maruri-Aguilar

Overview

Design questions permeate many aspects of computer experiments.

- Sequential screening
  $\rightarrow$ Identifying important simulator parameters.
- Parameter estimation with Heteroscedastic Gaussian Process models.
  $\rightarrow$ Learning the parameters of the surrogate model.
- Calibration.
  $\rightarrow$ How well can the simulator describe the observed system?

What design to use in each case to minimize number of simulator evaluations needed?

Introduction

Screening

Identify important inputs to the simulator.

Emulation

- A statistical model to the computer code simulator.
- Typically modelled as a Gaussian Process.

Design

- Where to evaluate the simulator?
- Use criterion to minimize predictive variance, parameter uncertainty?
- Standard optimal design theory does not apply to GPs as it assumes independent homoscedastic errors.

Calibration

Understanding the simulator parameter space.

Screening: Overview of sequential approach

- High-dimensional input spaces may require more simulator runs to train and validate the emulator.
- Screening the simulator's inputs for non-linear effects on the output rather than distinguishing between negligible and active effects.
- Based on the elementary effects method for screening (Morris, 1991).
- Utilises a threshold value to separate the inputs with linear and non-linear effects.
- Sequential to keep the number of simulator runs down to a minimum.

Screening review

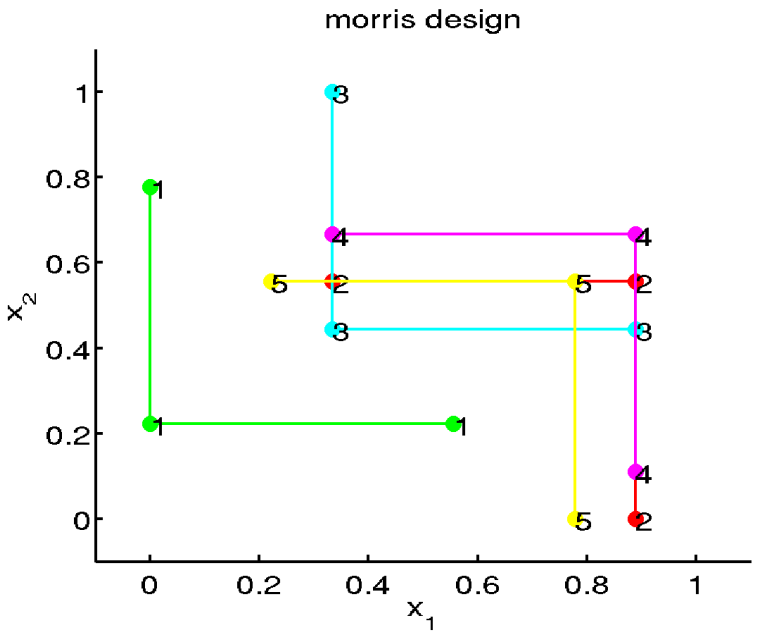
Cheap simulators

- Simulator-based functional ANOVA.
- e.g. Sobol' indices where an additive model of first and higher order effects used.

Expensive simulators

- Supersaturated design: use fewer model runs than input dimensions by making assumptions on the number of active inputs or the type of effects on the response
- e.g. monotonicity of the model output with respect to the inputs $\rightarrow$ sequential bifurcation (Kleijnen, 2009).
- Response surface methods $\rightarrow$ a surrogate model is utilised to approximate the simulator response.
- e.g. Savitsky et al. (2011) uses a Gaussian-process with a screening prior to encapsulate the assumptions of effect sparsity.

Elementary effects method: Morris (1991)



Morris Design with five trajectories

1. Based on calculation of elementary effects (EE) for each input variable:

$$EE_i(\mathbf{x}) = \frac{f(\mathbf{x} + \Delta \mathbf{e}_i) - f(\mathbf{x})}{\Delta}.$$

2. Use One-At-a-Time (OAT) trajectory designs.

Elementary effects method

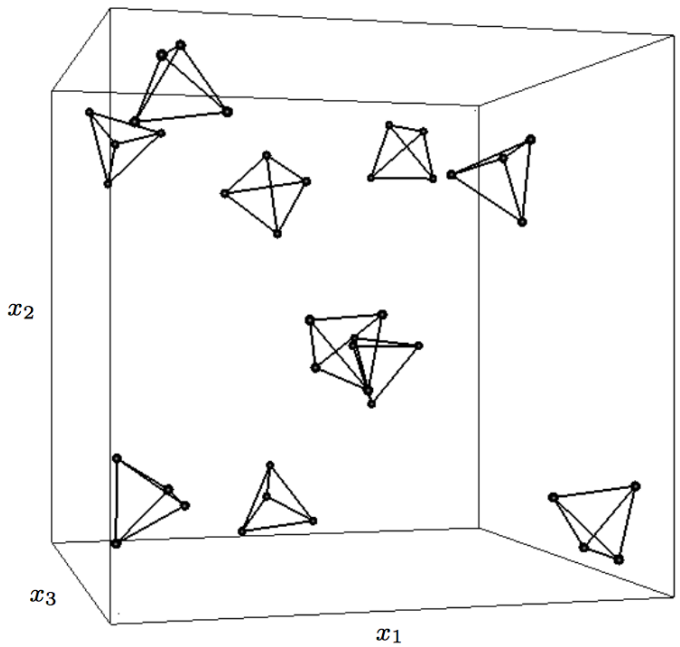
1. Compute sample moments for each input factor

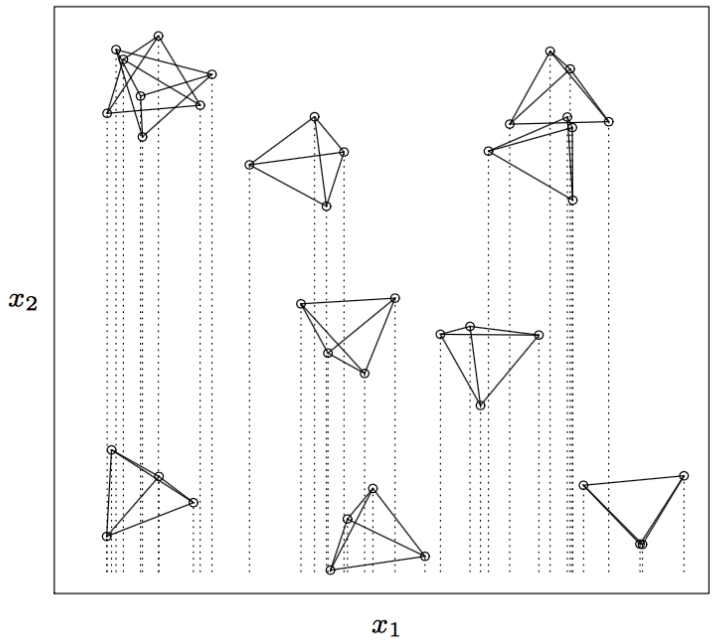
$$\mu_i = \frac{1}{R}\sum_{r=1}^{R} EE_i(\mathbf{x}_r),\ \mu_i^* = \frac{1}{R}\sum_{r=1}^{R} |EE_i(\mathbf{x}_r)| \text{ and } \sigma_i = \sqrt{\sum_{r=1}^{R} \frac{(EE_i(\mathbf{x}_r) - \mu_i)^2}{R-1}}.$$

2. Constant or linear effect will have constant EE_i and the σ_i zero.
3. Linear scaling: Number of simulator runs $R(K+1)$ for K factors.
4. Number of trajectories typically small to minimize simulator evaluations, e.g. Morris (1991) used $R = 3$ although values between 10 and 50 also used (Campolongo et al., 2004, 2007).
5. Initial points for trajectories random from the input space grid (Morris, 1991) or space-filling criterion, maximising the minimum distance between many trajectories (Campolongo et al., 2007).

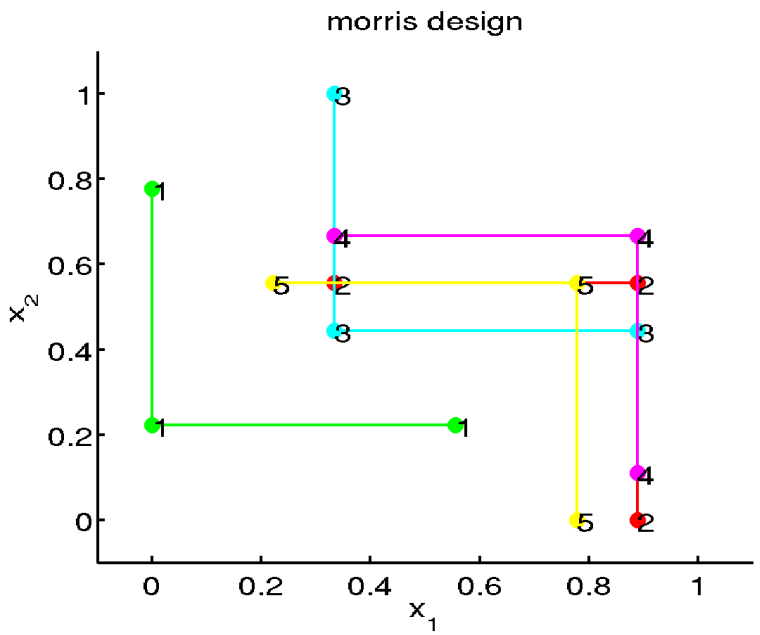
Randomly rotated simplices

1. Caveat: design points may fall on top of each other when projected into lower dimensions. Reduces number of available runs after screening out unimportant factors.
2. Screen with randomly rotated simplices (Pujol, 2009) but suboptimal effect calculation.





Sequential Elementary effects method



Morris Design with five trajectories

1. Based on calculation of elementary effects (EE) for each input variable:

$$EE_i(\mathbf{x}) = \frac{f(\mathbf{x} + \Delta \mathbf{e}_i) - f(\mathbf{x})}{\Delta}.$$

2. Use One-At-a-Time (OAT) trajectory designs, increasing by one trajectory at each stage.
3. If variance of EE exceeds threshold, remove factor from consideration.

Sequential Elementary effects method

1. Create space-filling design of M starting points, ordering according to the biggest distance between points.
2. Create one-at-a-time for current active factors and run the simulator at those points.
3. Compute elementary effects and their sample moments.
4. Remove factors $\sigma_i > \sigma_0$. These have non-linear effects and should be kept for downstream analysis.
5. Got to step 2 unless all factors have been removed as non-linear or reach maximum number of trajectories.
6. All factors remaining have constant or linear effects.

Thresholding solely on the variance of the elementary effects σ_i

1. Linear effects of factors may be removed from the simulator output at a preprocessing stage or during the emulation phase.
2. Linear effects may be incorporated in the mean function of a Gaussian Process emulator while omitted from the covariance specification.

Heuristic selection of variance threshold σ_0

1. No natural units for the selection of the variance threshold.
2. Specify the threshold in terms of deviation for each factor from a simple regression line.

$$Y(x_i) = ax_i + b + \epsilon_i$$

where $\epsilon_i \sim (0, \gamma_i)$.

3. Then threshold is :

$$\sigma_0 = \sqrt{\chi^2_{0.99,R-1} \frac{2\gamma}{\Delta^2(R-1)}}$$

where $\chi^2_{0.99,R-1}$ is the 99% quantile of a chi-squared distribution with $R-1$ degrees of freedom.

4. Adaptive threshold as it depends on number of trajectories used.

Synthetic data

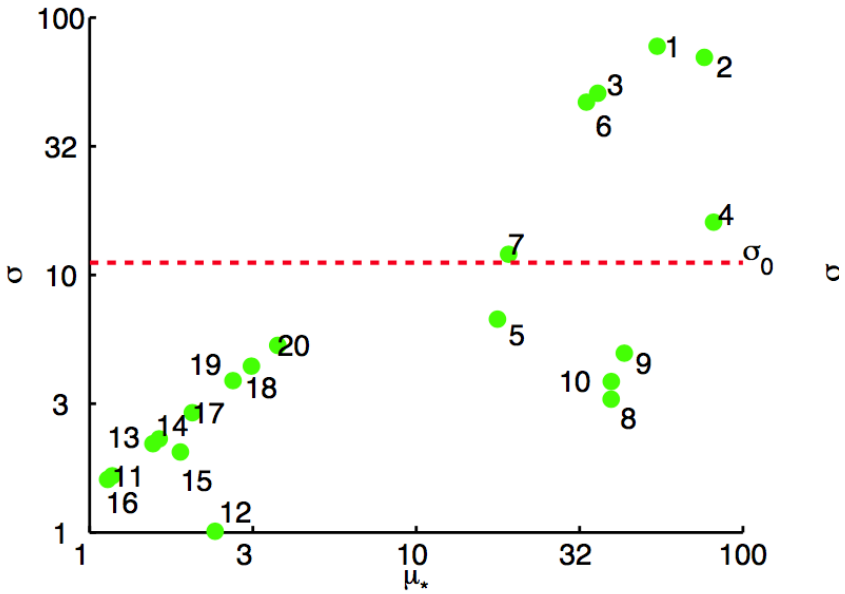
1. Synthetic test function introduced in Morris (1991)

$$y = \beta_0 + \sum_{i=1}^{20} \beta_i w_i + \sum_{i<j}^{20} \beta_{ij} w_i w_j + \sum_{i<j<l}^{20} \beta_{ijl} w_i w_j w_l + \sum_{i<j<l<s}^{20} \beta_{ijls} w_i w_j w_l w_s,$$

2. Factors $x_1, \ldots, x_7$ have a non-linear effect on the function output while factors x_8, x_9, x_{10} have a linear effect and factors $x_{11}, \ldots, x_{20}$ have negligible effect.
3. A threshold value of $\gamma = 2.6$ was used, corresponds to around 0.005% of the range of the response y.
4. Results: 210 function evaluations $R = 10$ for the batch EE procedure, 150 function evaluations required on average for the sequential approach.
5. That is an average savings of 28% of simulator runs.

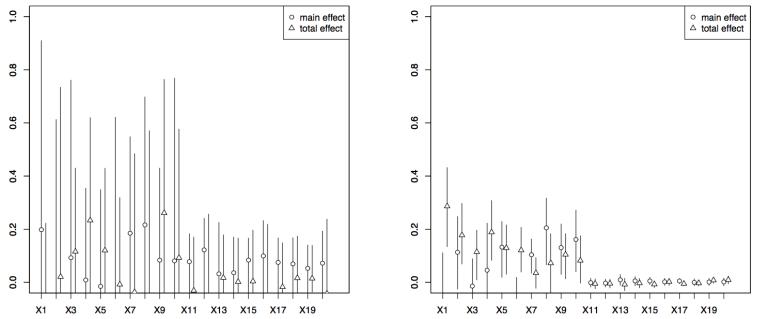
Synthetic data results

1. Six of the seven factors with non-linear effects are identified at the first iteration:



Synthetic data results: Sobol'

1. Sobol' sensitivity analysis method with 220 runs to compute first-order and total indices results in large 95% confidence intervals.
2. Many more runs required to reduce confidence interval.



(a) 220 runs

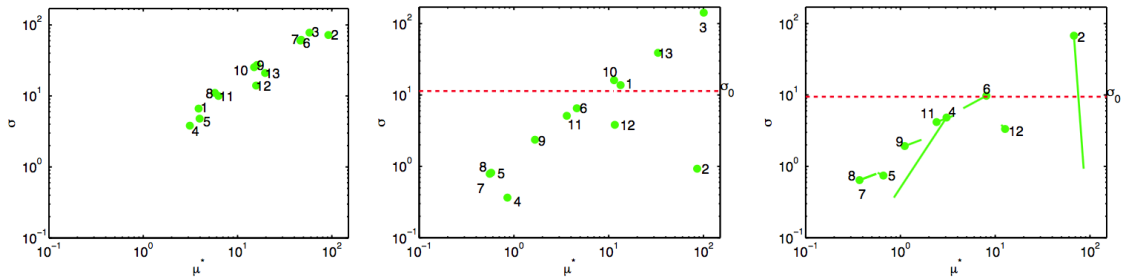
(b) 4378 runs.

3. Screening methods such as the EE method, can be utilised prior to a more detailed sensitivity analysis in order to minimise the number of model runs.

Rabies model

1. Quantify risk of disease re-introduction by modelling the raccoon dog vector species interactions with red fox.
2. Non-spatial and disease propagation is calculated solely with respect to population dynamics.
3. $\gamma =$ a factor has near-linear effect if the output varies no more than $\sim 5\%$ from linear.
4. Encapsulates both the internal variability of the stochastic model and our prior definition of a near-linear effect.
5. Model has 13 free parameters and $R = 20 \rightarrow 280$ simulator evaluations for EE design was previously used (Singer and Kennedy, 2008) .
6. Sequential approach required 102 runs $\sim 61\%$ savings.

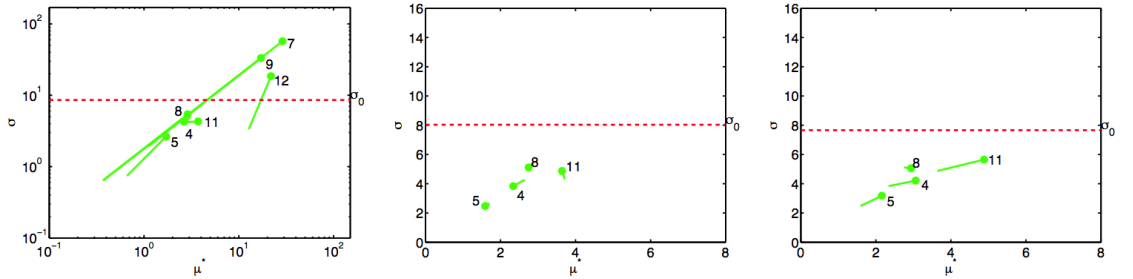
Rabies model results



(a) Batch Morris using R=20.

(b) Sequential iteration 1

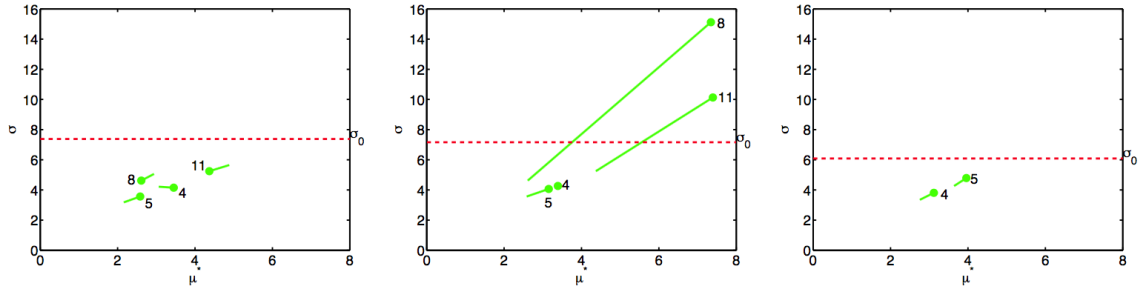
(c) Sequential iteration 2



(d) Sequential iteration 3

(e) Sequential iteration 4

(f) Sequential iteration 5



(g) Sequential iteration 6

(h) Sequential iteration 7

(i) Sequential iteration 19

Two factors found to have linear effect. Solid line denotes path from previous value of EE samples moments for each factor.

Screening summary

1. Identify inputs with non-linear effects with a minimum number of trajectories.
2. Significant computational savings compared to the batch approach on both synthetic data and a real-world simulator.
3. Utilise an easily interpretable variance value γ specified on the simulator output space.
4. Use space-filling design for starting trajectory points with a maximum size or use a low discrepancy space filling sequence.

Random Output Simulators

Stochastic Simulator

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

Gaussian Process Approximation

In addition to having a finite number of simulator runs, uncertainty due to stochastic simulator. Our assumed observational model is:

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

Approaches to Heteroscedastic Modelling

Coupled system of GPs

- Model heteroscedastic variance using a coupled system of GPs.
- MCMC inference, Goldberg et al (1998)
- Most Likely value, Kersting et al (2007),
- Variational, Lázaro and Titsias (2011).
- Extended to utilise repeated observations (replicates).

Joint Likelihood Model

- Coupled model too complex for design calculations.
- Use parametric deterministic variance model.
- Optimisation of the mean and variance model parameters proceeds jointly $\rightarrow$ tractable optimal design calculations.
- Efficient inference with replicated observations.

Motivation: Why replicate observations?

- Simulator evaluations at input locations much closer than length scale can cause numeric difficulties.
- For moderate number of simulator evaluations, inference time can become impractical. Utilising replicate observations allows for much quicker inference.
- Learning the noise model more accurately as we will show.

Joint Likelihood Model

Crucial simplification: consideration of only deterministic variance models. The heteroscedastic GP prior is thus:

$$p(\mu|\theta,\beta) = \mathcal{GP}\left(0, K_\theta + \mathrm{diag}(\exp(f_{\sigma^2}(x,\beta))P^{-1}\right),$$

where $f_{\sigma^2}(x,\beta)$ is the deterministic variance model.
The joint log likelihood of the sample mean $\hat{\mu}$ and variance s^2 for N observations:

$$\log p(\hat{\mu}, s^2|\mathbf{X},\theta,\beta) = \left(\sum_{i=1}^{N} \log p(s_i^2|\beta, x_i, n_i)\right) + \log N\left(\hat{\mu}|0, K_\theta + RP^{-1}\right),$$

where K_θ the GP covariance function with parameters θ, R the diagonal matrix with elements $\exp\left(f_{\sigma^2}(x_i,\beta)\right)$.

Deterministic Variance Models

Fixed Basis

Fixed Basis variance model, the log variance function is modelled as a linear in parameters regression using a set of fixed basis functions:

$$f_{\sigma^2}(x, \beta) = \exp\left(H(x)^T \beta\right),$$

where $H(x)$ is the set of fixed basis functions with known parameters.

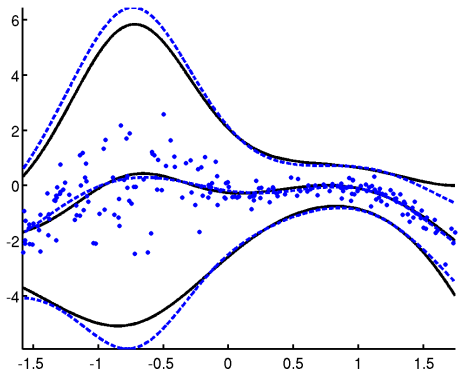
Latent Kernel

In high dimensional cases a non-parametric method could be considered using an additional 'variance kernel'.

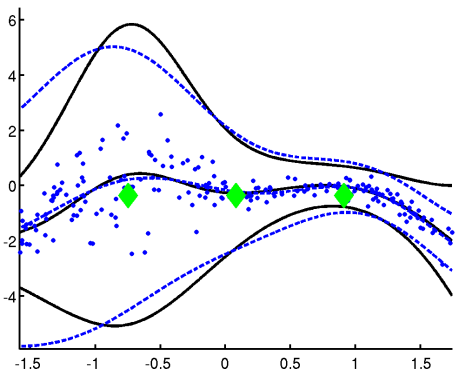
$$f_{\sigma^2}(x, \beta) = k_\Sigma^T (K_\Sigma + \sigma_n^2)^{-1} \beta,$$

where $K_\Sigma = k(X_z, X_z)$ and $k_\Sigma = k(X_z, X_t)$ are the variance kernel functions, depending on parameters θ_Σ and σ_n^2 a nugget term.

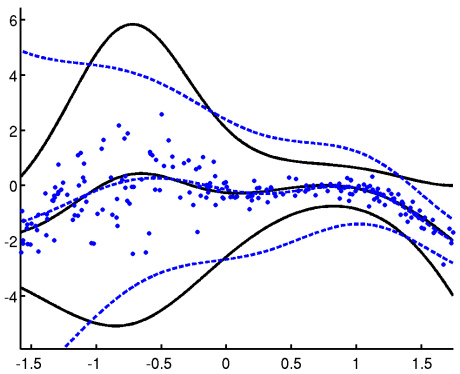
Example of three variance models



(a) Coupled Model



(b) Latent Kernel



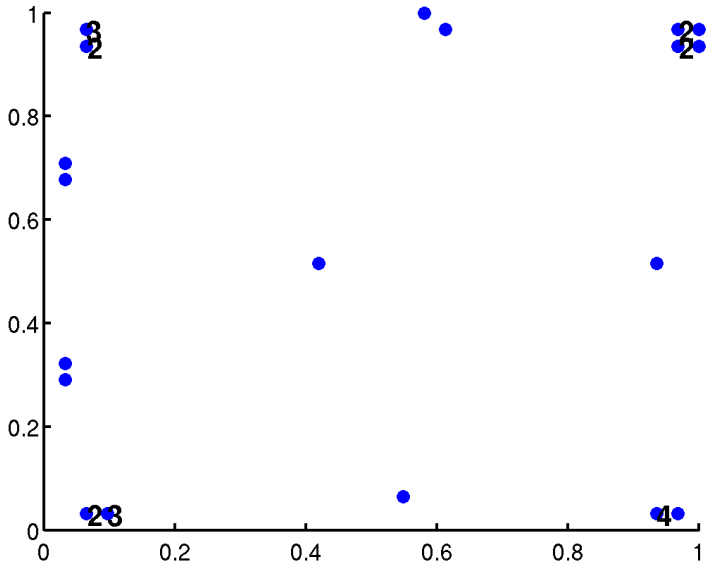
(c) Quadratic Polynomial

Comparison of the Coupled, Latent Kernel and Quadratic polynomial variance models. Training set consists of using 200 design points with 4 replicate observations at each site. Dots are the empirical means of the samples. The black solid lines are the true function mean and standard deviation and the blue dashed lines the GP predictions.

Experimental Design For Correlated Processes

- Design to minimise kernel parameter uncertainty $\rightarrow$ D-optimality.
- Why not minimise predictive variance instead?
- Ans: All such methods either assume parameters known or use approximate values.
- D-optimal design used as preliminary design or as part of hybrid criterion (e.g. see Zhu and Stein (2005)).

Optimal Design for Heteroscedastic Gaussian Process Regression with replicated observations



i

- Design to minimise parameter uncertainty $\rightarrow$ D-optimality
- Maximize Fisher information of design ξ:

$$\mathcal{F}(\xi) = E\left[\frac{\partial^2}{\partial\theta^2}\ln L(X|\theta,\xi)\right]$$

- Analytic solution derived for GP with parametric variance model.

Joint Likelihood: Fisher Information

The FIM for a design ξ is defined as:

$$\mathcal{F}(\xi) = \int \left(\frac{\partial^2}{\partial \theta^2} \ln \left[L(X|\theta, \xi) \right] \right) L(X|\theta, \xi) \, \mathrm{d}X,$$

where $L(X|\theta, \xi)$ is the likelihood function.
For Joint Likelihood model FIM can be calculated analytically:

$$\boxed{\mathcal{F}_{ij} = \sum_{m=1}^{M} F_{ij}^{s} + F_{ij}^{N},} \tag{1}$$

where

- M the number of design points.
- $F_{ij}^{s} = \frac{n_i - 1}{2} \frac{\partial f}{\partial \theta_i} \frac{\partial f}{\partial \theta_j}$ where n_i the number of replicate observations at design point i and $\frac{\partial f}{\partial \theta_j}$ the derivative of the variance model $f(\theta)$ with respect to parameter θ_j.
- $F_{ij}^{N} = \frac{1}{2} \mathrm{tr}(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j})$.

Experimental Setup

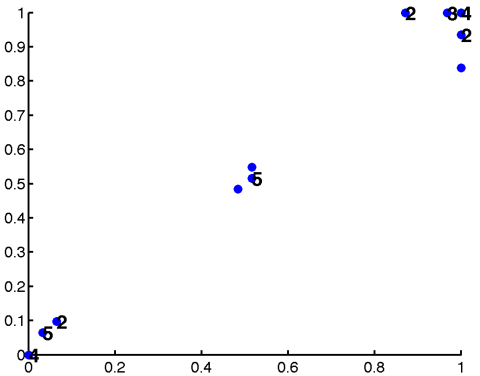
Synthetic Experiment

- Sample from GP with known parameters.
- GP Maximum Likelihood Inference with same covariance using different designs.
- Compute parameter errors.
- 500 realisations.

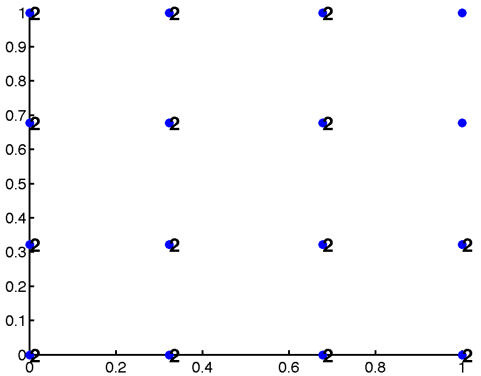
Optimisation

- Greedy: add point from candidate set that maximises $\mathcal{F}$.
- Simulated annealing: Global optimisation across candidate space.

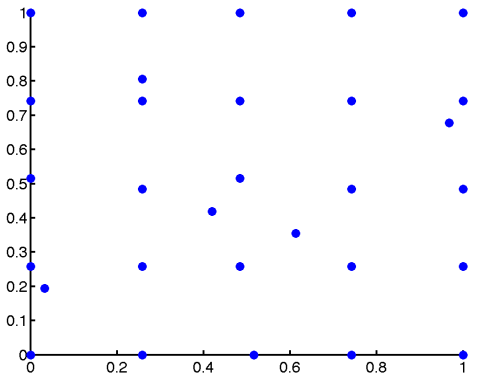
Local Design: Latent Kernel Variance model



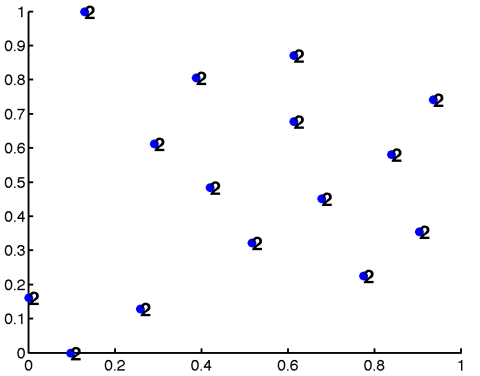
(a) Greedy



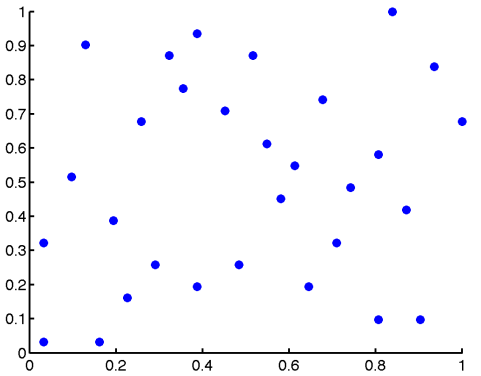
(b) Replicate Grid



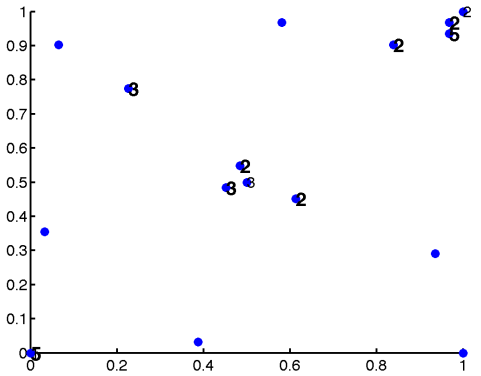
(c) Grid



(d) Latin Hypercube Rep

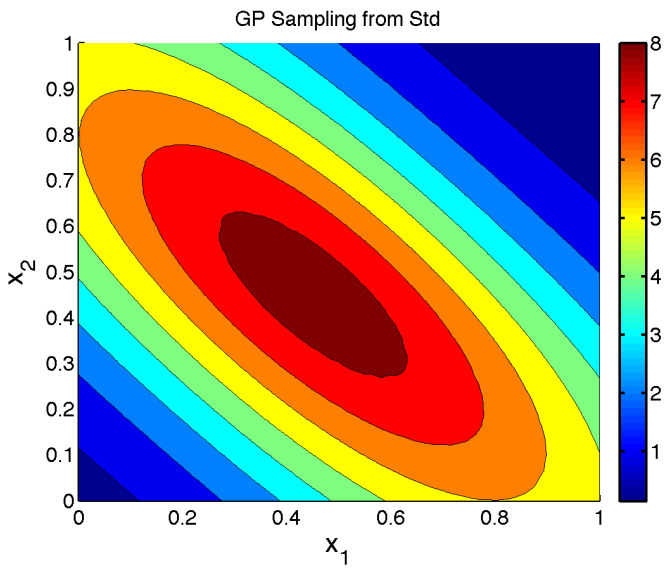


(e) Latin Hypercube

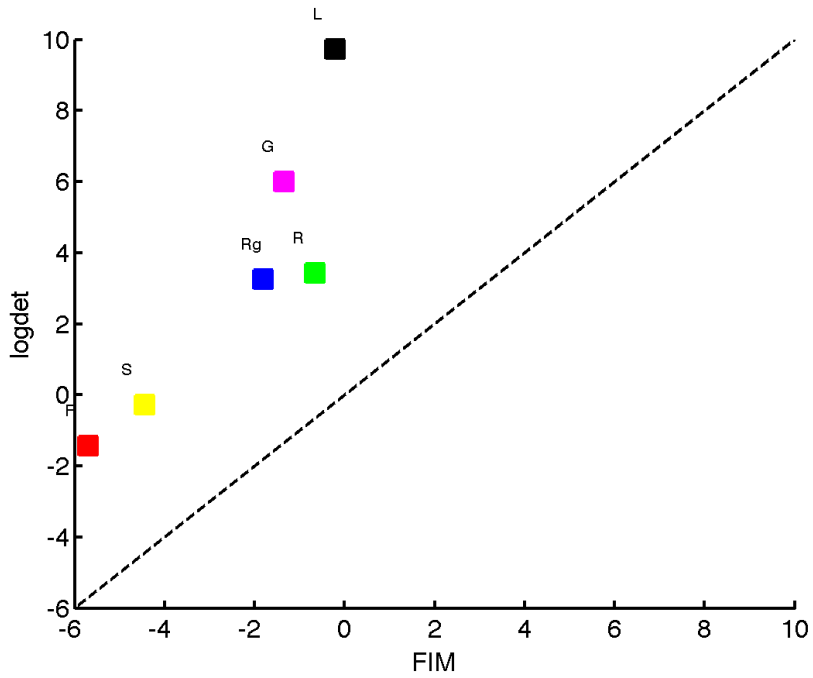


(f) Sim Annealing

Parameter errors for Latent Kernel variance model



Variance surface.



FIM (x axis) and LDM (y axis).

Variance model parameter errors

Greedy	Replicate Grid	Grid	Latin Rep	Latin	Sim Ann
0.22	0.46	0.66	0.49	0.82	0.25

Calibration - A History Matching approach

Calibration

Under what parametrisation, if any, does the computer model fit the noisy observations?

Emulation

Use the heteroscedastic GP emulator to quickly eliminate implausible regions of parameter space reducing the need for simulator evaluations.

Design

For each calibration iteration, need to efficiently generate a design for the non-implausible region.

Calibration - A History Matching approach

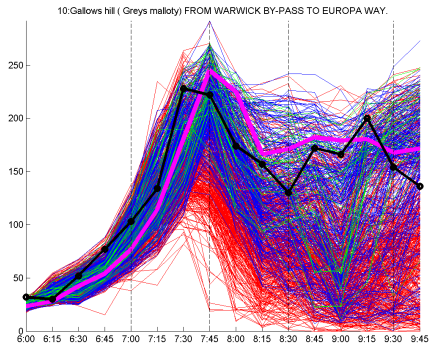
Implausibility measure

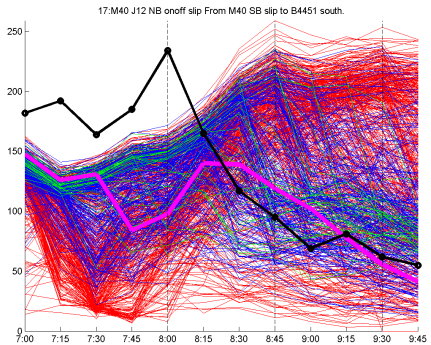
$$\boxed{I(x) = (E[f] - z)^2 / (V[f] + V_o + V_{MD} + V_E)}$$

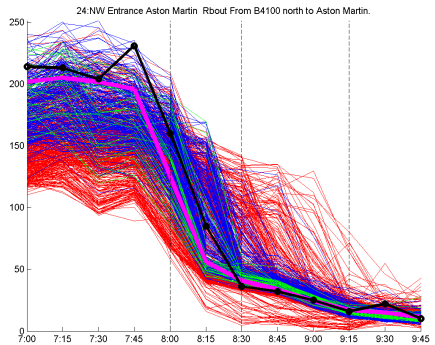
where

- z the observed turn count.
- $E[f]$, $V[f]$ the mean and variance of the replicated runs for a given parameter setting.
- V_o the observation error.
- V_{MD} model discrepancy.
- V_E is the sample error for the mean.

Learning about the simulator



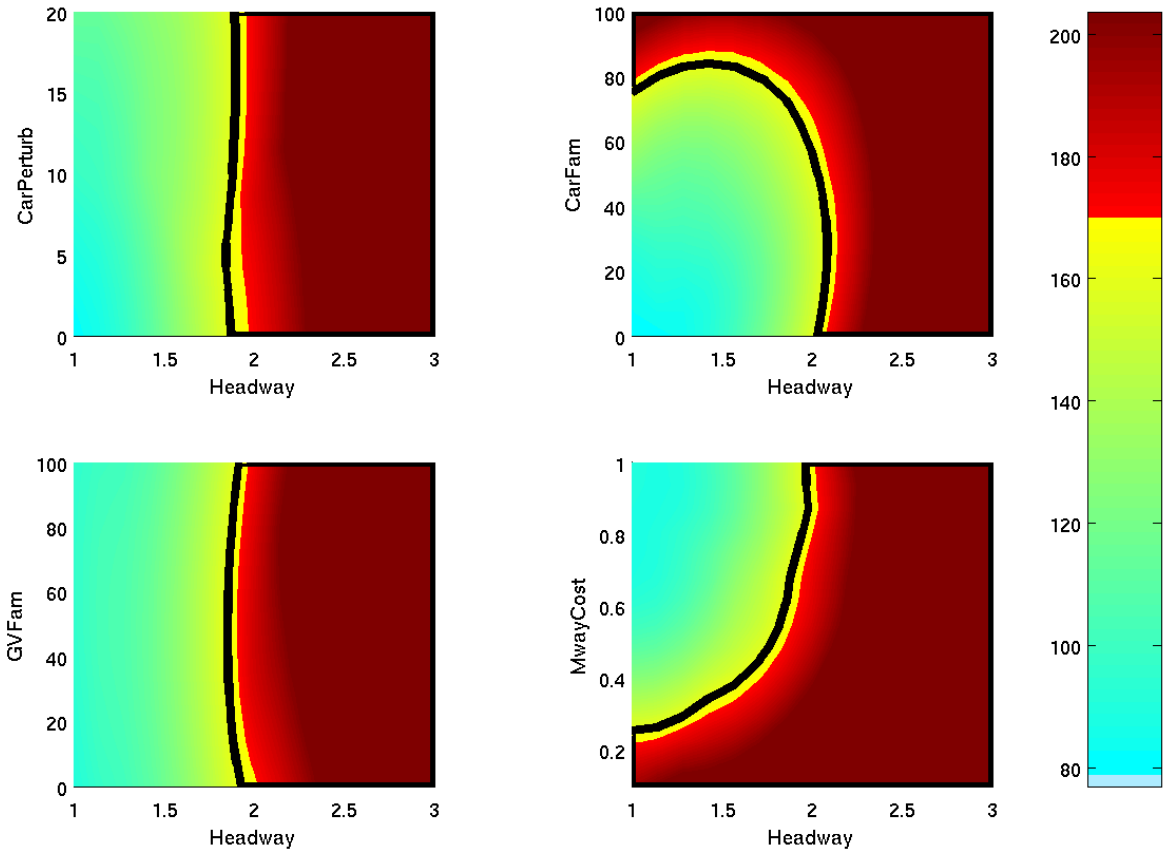




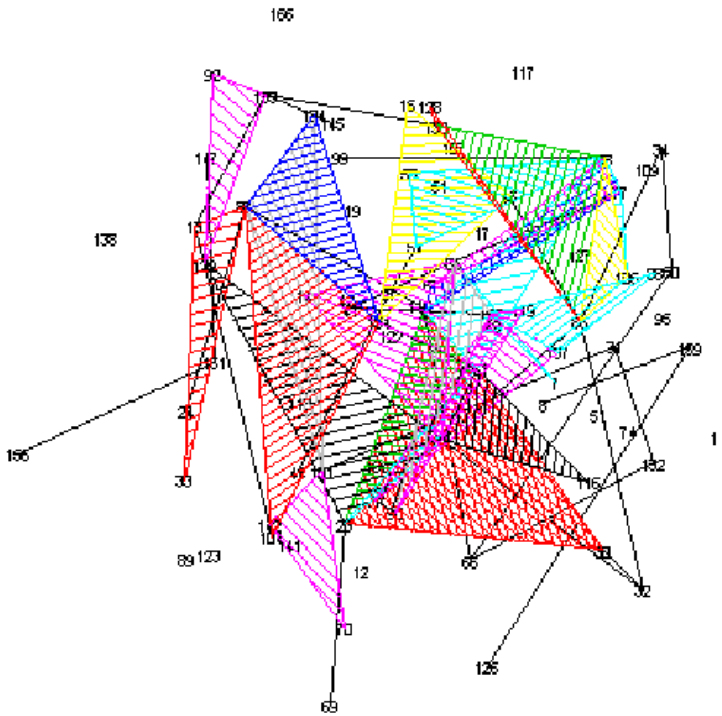
- Varying 25 input parameters (from the many hundreds present, guided by our experts) we could find simulator evaluations not inconsistent with the observations.
- Biggest challenge was eliciting discrepancies and uncertainties from experts.

We used a multiple wave approach, which means emulators need to work in smaller and smaller parts of the input space.

Bivariate Implausibility Plots



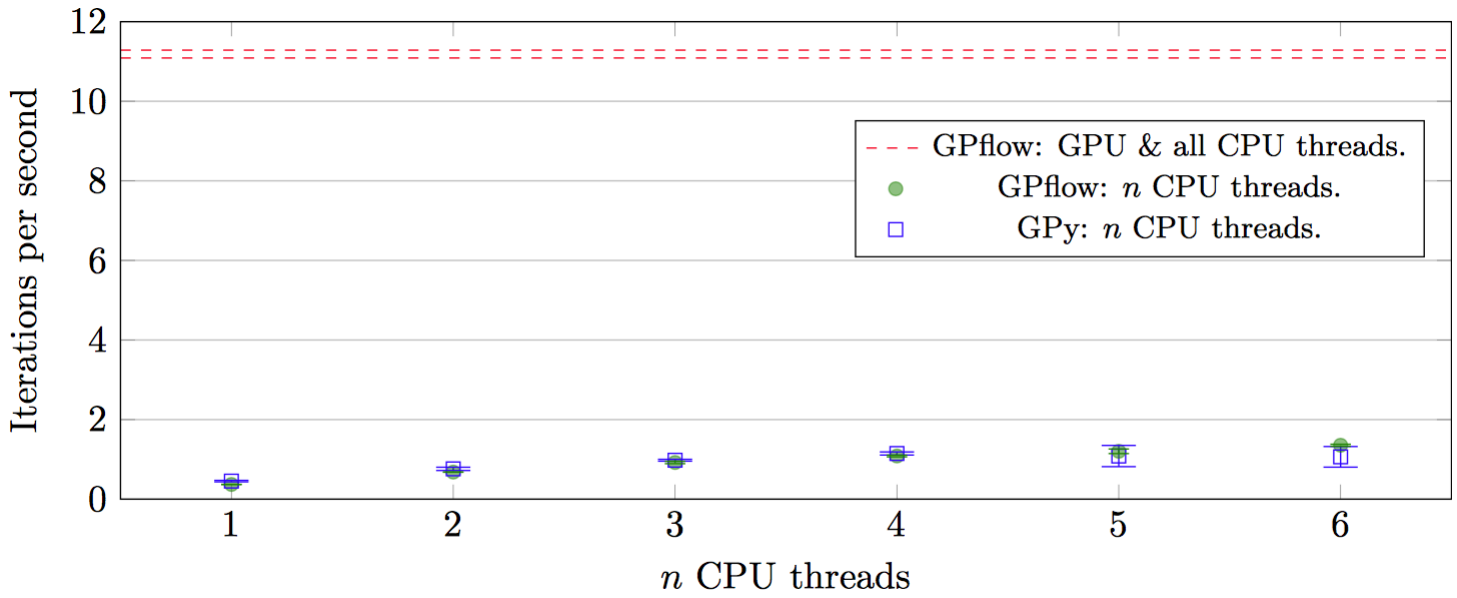
Using persistent homology for visualisation and design



Projection of original data coordinates using multidimensional scaling.

1. Persistent homology is a natural extension of cluster analysis.
2. Taking clusters as its elemental building element, the analysis identifies topological features such as two-, three- and higher-dimensional cycles in the data.
3. Use convex bounding regions to efficiently generate designs in the high-dimensional implausibility space. Convex hull method 70%, vs LH 2%.

A new software framework for GPs: GPflow



Iterations per second for GPy and GPflow on a large classification task.

- Use Google Tensorflow to scale up computation.
- Adding a GPU, results in significant performance gains.
- Task using 6 CPUs $\sim$ 2 days, GPU $\sim$ 5 hours.

Conclusions

1. Sequential screening to quickly eliminate irrelevant dimensions.
2. Simple heteroscedastic GP model allows for optimal design calculation.
3. Fisher Designs minimise kernel parameter estimation variance.
4. Utilising Replicated observations beneficial for stochastic emulation,
5. . . . particularly to identify parameters in the heteroscedastic covariance terms.
6. Leverage HGP for efficient calibration.
7. Calibration poses new design questions.
8. Good software helps!

Papers

Sequential screening with one-at-a-time designs

Boukouvalas, A., Gosling, J. P., Maruri-Aguilar, H., An efficient screening method for computer experiments, Technometrics, 2013.

Calibration of stochastic simulator

Boukouvalas, A., Sykes, P., Cornford, C., Maruri-Aguilar H., Bayesian Precalibration of a Large Stochastic Microsimulation Model, IEEE Transactions on Intelligent Transportation Systems, 2014.

Optimal design learning Gaussian process parameters

Boukouvalas, A., Cornford, D. Stehlík, M., Optimal design for correlated processes with input-dependent noise, Computational Statistics & Data Analysis, 2013.

GPflow: A Gaussian Process Library using TensorFlow

Matthews et al, JMLR, 2017.

References

- Morris, Max D. "Factorial sampling plans for preliminary computational experiments." Technometrics 33.2 (1991): 161-174.
- Vernon, Ian, Michael Goldstein, and Richard G. Bower. "Galaxy formation: a Bayesian uncertainty analysis." Bayesian Analysis 5.4 (2010): 619-669.
- Zhu, Zhengyuan, and Michael L. Stein. "Spatial sampling design for parameter estimation of the covariance function." Journal of Statistical Planning and Inference 134.2 (2005): 583-603.
- Saltelli, Andrea, Karen Chan, and E. Marian Scott, eds. Sensitivity analysis. Vol. 1. New York: Wiley, 2000. Good review of sensitivity analysis methods.
- Kersting, Kristian, et al. "Most likely heteroscedastic Gaussian process regression." Proceedings of the 24th international conference on Machine learning. ACM, 2007.
- Titsias, Michalis K., and Miguel Lázaro-Gredilla. "Variational heteroscedastic Gaussian process regression." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.
- Goldberg, Paul W., Christopher KI Williams, and Christopher M. Bishop. "Regression with input-dependent noise: A Gaussian process treatment." Advances in neural information processing systems (1998): 493-499.