Optimal design for correlated processes with input-dependent noise

A. Boukouvalas[a,*], D. Cornford[a], M. Stehlík[b]

[a]Non-Linear Complexity Research Group, Aston University, Aston Triangle, Birmingham, United Kingdom
[b]Department of Applied Statistics, Johannes Kepler University in Linz, Austria

Abstract

Optimal design for parameter estimation in Gaussian process regression models with input-dependent noise is examined. The motivation stems from the area of computer experiments, where computationally demanding simulators are approximated using Gaussian process emulators to act as statistical surrogates. In the case of stochastic simulators, which produce a random output for a given set of model inputs, repeated evaluations are useful, supporting the use of replicate observations in the experimental design. The findings are also applicable to the wider context of experimental design for Gaussian process regression and kriging. Designs are proposed with the aim of minimising the variance of the Gaussian process parameter estimates. A heteroscedastic Gaussian process model is presented which allows for an experimental design technique based on an extension of Fisher information to heteroscedastic models. It is empirically shown that the error of the approximation of the parameter variance by the inverse of the Fisher information is reduced as the number of replicated points is increased. Through a series of simulation experiments on both synthetic data and a systems biology stochastic simulator, optimal designs with replicate observations are shown to outperform space-filling designs both with and without replicate observations. Guidance is provided on best practice for optimal experimental design for stochastic response models.

Keywords: Optimal design of experiments, Correlated observations, Emulation, Gaussian Process, Heteroscedastic noise

1. Introduction

Design plays an important role in enabling effective fitting and exploitation of a wide variety of statistical models, e.g. regression models such as Gaussian

*Corresponding author at: Non-Linear Complexity Research Group, Aston University, Aston Triangle, Birmingham, United Kingdom, Tel.: +44 1212043922

Email addresses: boukouva@aston.ac.uk (A. Boukouvalas), D.Cornford@aston.ac.uk (D. Cornford), Milan.Stehlik@jku.at (M. Stehlík)

processes. The motivation for this work is a recognition that experimental design plays a crucial part in the building of an emulator [26]. The use of emulators, or surrogate statistical representations of computer simulators, provides a solution to the computational constraints that limit a full probabilistic treatment of many simulators. Experimental design is particularly relevant to emulation because we are able to choose the inputs at which the simulator is evaluated with almost complete freedom. The simulator is typically expensive to run, thus it is beneficial to optimise the design given the available *a priori* knowledge.

Most work on emulation has focused on deterministic simulators, where the outputs depend uniquely on the inputs, however it is increasingly common to encounter stochastic simulators, where the randomness is typically associated with interactions which are intrinsically unpredictable or represent some unresolved, essentially random, process within the simulator. Examples of stochastic simulators arise in microsimulation in transport modelling [24] and biochemical networks of reactions [32]. Design and emulation methods developed for deterministic computer experiments need to be extended to be applicable in the stochastic context [11].

A common feature of stochastic simulators is that the variance of the output is input dependent. This requires adaptation of the normal Gaussian Process (GP) regression model [23]. In this paper we introduce a class of heteroscedastic GP models that allow for both flexible variance modelling and tractable calculations for optimal design. Our work extends [36] which developed optimal designs for homoscedastic GPs using a Fisher information criterion. This paper expands [36] to heteroscedastic GPs with replicated observations. Our approach is general and is relevant to areas such as model-based spatial statistics [8, 28], where kriging methods are used, and more general GP regression [23].

When considering correlated processes, such as GPs, the majority of the results of traditional optimal design, such as the General Equivalence Theorem and the additivity of information matrices do not hold [16]. For an overview of classical optimal design theory see [2] or other standard textbooks. In GP regression, a parametric covariance function is used to model the variance and correlation of the unknown function. The parameters of the covariance are usually estimated by Maximum Likelihood (ML) or Bayesian inference. In this paper, we investigate design under ML estimation, with a focus on best learning the model parameters.

By utilising asymptotic results of estimators, useful approximations to finite sample properties can be constructed. Two asymptotic frameworks are considered in the literature [35, 27]: increasing domain and infill domain asymptotics. It has been found that for certain consistently estimable parameters of exponential kernels with and without a noise term, under ML estimation, approximations corresponding to these two asymptotical frameworks perform about equally well [35]. For parameters that are not consistently estimable however, the infill asymptotic framework is preferable [12]. In [14], it was shown that under increasing domain asymptotics the ML estimator, $\hat{\theta}$, converges in probability to the true parameter, θ, and standard asymptotics hold. Unfortunately no such general results exist under infill asymptotics except for specific classes of covariance functions [1]. A non-asymptotic justification is provided by [20] using a truncated function expansion, but is only valid for low process noise levels.

Recently, a ‘nearly’ universal optimality has been addressed for the case of correlated errors, see e.g. [7] and references therein, overcoming some of the difficulties in the correlated setup. Exact optimal designs for specific linear models with correlated observations have been investigated (see [12] and references therein), but even for simple models exact optimal designs are difficult to find.

Optimal design for correlated errors has also been examined under generalised least squares estimation of treatment contrasts in fixed-block effects models where correlation is assumed between treatments within the same block [30]. Within the class of equally replicated designs, designs that minimise the variance of treatment contrasts were found. It was also found that for large positive correlations unequally replicated designs could achieve lower variance values. Although the derivation was only for a specific number of treatments and units, the potential that unequally replicated designs hold for a wider class of scenarios is tantalising and is further investigated in this paper for the GP model case.

Most of the literature on optimal experimental design assumes homoscedastic noise. Optimal design under a fixed basis log-linear-in-parameters model is examined in [29]. Although stochastic processes are not considered, the variance model used is similar to the fixed basis model utilised in this work. They follow a Bayesian approach to design and demonstrate that informative priors lead to more efficient designs.

In certain cases there may exist multiple objective functions which depend upon different information matrices. Compound optimal design provides a general approach, combining multiple such objective functions such as model discrimination (T-Optimality) and parameter estimation (A- or D-optimality) via a weighted average of their information matrices [15]. Compound designs may also be used to generate designs with non-equal emphasis on the trend and covariance parameters [17]. Hybrid criteria that explicitly combine prediction and parameter estimation have also been developed [38, 37]. In [38] such a criterion is defined to minimise the maximum predictive variance as well as a summary of the ML parameter covariance. While this criterion selects observations which reduce parameter uncertainty and predictive uncertainty given the current parameter, it does not take into account the effect of parameter uncertainty on prediction error. To address this issue, [37] propose an amended criterion and derive an iterative algorithm which alternates between optimising the design for covariance estimation and spatial prediction. We note here that a space-filling design does not necessarily minimise the prediction error. For instance if one is interested in optimization of the Integrated Mean Squared Prediction Error (IMSPE), in one dimension and for an Ornstein-Uhlenbeck processes, then the space-filling, i.e. equidistant design, is optimal citeZagoraiou2010. However, this property is not generally true in a 2-dimensional design space [3]. As proven in [3], a space-filling design does not necessarily reduce the IMSPE more than a design forming a line, which they term monotonic set designs.

Geometric designs such as nested or subsampling designs have been proposed to identify hierarchically related sources of variations. They allow for the estimation of the amount of variation that is derived from each hierarchical level and the determination of the optimal allocation of sampling effort to each level [10]. Such designs place points at a variety of inter-point distances and may be used for the inference of difficult to learn GP correlation length-scale parameters [21].

Our design approach is model-based, where the assumption of a sufficiently well known model is made for the problem of interest. In geometric designs such as Latin Hypercube sampling which aim to cover the design space or nested sampling which aim to have a range of inter-point distances available, no such model is assumed. For model-based design, incorrect model assumptions may lead to arbitrarily bad performance. However, we expect model-based optimal designs using informative prior beliefs to offer superior performance to designs that arise from purely geometric grounds when the model assumptions are met. We show that this is the case via an extensive set of simulation experiments where model-based optimal designs are contrasted to space-filling maximin Latin-Hypercube [22] and grid designs with and without replication. We demonstrate the resulting gains in parameter accuracy when model-based designs are utilised.

The paper is structured in the following way. The GP model and the corresponding design criterion are described in Section 2 and Section 3 respectively. Optimisation is discussed in Section 4. A series of simulation studies is presented in Section 5 and an application of the methodology to a systems biology simulator is discussed in Section 6. We conclude with a discussion and a proposal for future work in Section 7.

2. Heteroscedastic GP model

This section describes the hetoroscedastic GP model we have developed that permits model-based optimal design. The *joint likelihood* GP model allows the optimisation of the mean and variance model parameters to proceed jointly. The assumed additive model for the simulator for each output is:

$$t(x) = f(x) + \epsilon(x) \ ,$$

where x denotes the simulator inputs, $f(x)$ is the unknown mean of the simulator response, $\epsilon(x)$ is an input dependent, zero mean, additive Gaussian random variable representing the intrinsic simulator variability and $t(x)$ represents a single realisation of the simulator output.

A zero-mean GP prior is placed on the simulator mean:

$$p(f|\theta_f) = \mathcal{GP}\left(0, K_f(\theta_f)\right), \tag{1}$$

where K_f is the input dependent covariance and θ_f the kernel hyperparameters. From now on we will omit the dependency of K_f on θ_f and just write $\mathcal{GP}\left(0, K_f\right)$.

The crucial simplification is the consideration of parametric variance models. The variance model is a parametric function $g_{\sigma^2}(x, \beta)$ with unknown parameters β. The heteroscedastic GP prior can be calculated after integrating out f (see Appendix A):

$$p(t|\theta, \mathbf{x}) = \mathcal{GP}\left(0, K_f + R\right) \ ,$$

where R is the diagonal matrix with elements $R_{ii} = \exp\left(g_{\sigma^2}(x, \beta)\right)$ representing the spatially varying noise process. To explicitly include replicate runs of the simulator we replace t with $\bar{t}$, the sample mean of the replicated runs and thus

$$p(\bar{t}|\theta, \mathbf{x}) = \mathcal{GP}\left(0, K_f + RP^{-1}\right), \tag{2}$$

where P the diagonal matrix of the number of replicated observations $P_{ii} = n_i$ at the i'th training point location x_i. For the Matérn kernel used in our experiments, θ_f includes the process variance σ_p^2 and correlation length scale λ parameters (see Section 4.2 of [23]). The set of free parameters for this model is $\theta = \{\theta_f, \beta\}$. The likelihood corresponding to this model, expressed in terms of the sample means $\bar{\mathbf{t}}$ and sample variances $\mathbf{s}^2$ of the training data, is derived in Appendix A.

The model parameters are estimated via maximum likelihood on a set of noisy observations referred to as the training set. The GP predictive equations are obtained by conditioning on the training dataset:

$$E[t_*] = {K_f}^*(K_f + RP^{-1})^{-1}\bar{\mathbf{t}}\,, \tag{3}$$

$$Var[t_*] = {K_f}^{**} + R^* - {K_f}^*(K_f + RP^{-1})^{-1}{K_f}^{*T}, \tag{4}$$

where $K_f = K(\mathbf{X}, \mathbf{X})$, ${K_f}^* = K(\mathbf{x}_*, \mathbf{x}_*)$ and ${K_f}^{**} = (\mathbf{x}_*, \mathbf{x}_*)$ are the between training, training-test and test-test input covariance functions respectively. R^* is the diagonal matrix of the variance model evaluated at the test points x_*.

We have considered two options for the variance function $g_{\sigma^2}(x, \beta)$. For the *Fixed Basis* variance model, the log variance function is represented as a log-linear-in-parameters regression:

$$g_{\sigma^2}(x, \beta) = \exp\left(H(x)^T\beta\right), \tag{5}$$

where $H(x)$ is the set of fixed basis functions with known parameters. A simple example in 2D input space is a log-linear variance model: $g_{\sigma^2}(x, \beta) = \exp\left(\beta_0 + x_1\beta_1 + x_2\beta_2\right)$ which we refer to as the *Log-Linear model.* We have considered two types of basis functions: local (e.g. radial basis functions) and global (e.g. polynomial) to provide the input dependent variance. An advantage of local basis functions is the interpretability of priors on the β coefficients as they relate to a particular region of input space. However the number of local basis functions required for domain coverage grows exponentially with the input dimension. Polynomial and other global bases are therefore better suited for higher-dimensional spaces but imply a relatively simple variance response.

In high-dimensional cases a semi-parametric model, which we refer to as the *Latent-Kernel* model, could be considered using an additional 'variance kernel':

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

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 noise term. In this case z is a vector of latent variance parameters. In principle the latent points X_z could be set to the entire training data set X_t of the GP but for quicker inference it can also be set to a much smaller set which is not necessarily a subset of X_t. The parameters of the model are X_z, z and θ_Σ. Although all could in principle be optimised, in the experiments presented herein we simplify the optimisation task by fixing X_z to a Latin Hypercube design, fixing θ_Σ to constant values and optimizing z.

3. Optimal Design under Heteroscedastic noise

The design criterion we use for the *joint likelihood* GP model (Section 2) is defined as the negative log determinant of the Fisher Information Matrix (FIM). From now on when we refer to the FIM, we are referring to log determinant of the Fisher Information Matrix and not to the matrix itself. Lower values of the FIM signify a more informative design.

The (j,p)th element of the matrix for model parameters θ_j, θ_p is:

$$\boxed{\mathcal{M}^{jp} = \sum_{i=1}^{m} \mathcal{M}_{si}^{jp} + \mathcal{M}_N^{jp},} \tag{6}$$

where m is the number of design points and $\mathcal{M}_N^{jp} = \frac{1}{2}\mathrm{tr}(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_j}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_p})$ for a zero mean GP is [19]. Inclusion of mean parameters in the criterion is straightforward (see for example [19]) but is not developed herein as our focus is on design for covariance parameter estimation. $\mathcal{M}_{si}^{jp}$ is the contribution of the uncertainty in the sample variance model parameters:

$$\mathcal{M}_{si}^{jp} = \frac{n_i - 1}{2}\frac{\partial g_{\sigma^2}}{\partial \theta_j}\frac{\partial g_{\sigma^2}}{\partial \theta_p},$$

where $\frac{\partial g_{\sigma 2}}{\partial \theta_j}$ the derivative of the variance model $g_{\sigma^2}(\theta)$ (Section 2) with respect to parameter θ_j. A complete derivation is given in Appendix B.

In the case of the fixed basis model $g_{\sigma^2}(x,\beta) = \exp(H(x)^T\beta)$ and

$$\mathcal{M}_{si}^{jp} = \frac{1}{2}(n_i - 1)H(x_i)^T J_j H(x_i)^T J_p,$$

where J_j the zero vector with j^{th} element 1. If we examine the formula, $\mathcal{M}_{si}^{jp} = 0$ unless both θ_j an θ_p are parameters of the variance model f and the number of replicates is at least 2, i.e. $n_i > 1$.

For illustrative purposes, the matrix for a fixed basis variance model is shown. For the GP prior in Equation (2) we specify a Log-Linear fixed basis variance model for a one-dimensional input space with constant nugget $g_{\sigma^2}(x, \beta) = \exp(\beta x)$. The nugget characterizes the continuity of the covariance function at the origin. In our example for $x = 0$, $g_{\sigma^2}(0, \beta) = 1$. The model specification is completed by specifying the kernel K_f with a single parameter, the length-scale λ. For this model, $\mathcal{M}$ is:

$\downarrow \theta_i, \, \theta_j \rightarrow$	λ	β
λ	$\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\frac{\partial K_f}{\partial \lambda}\right)^2$	$\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\frac{\partial K_f}{\partial \lambda}\Sigma^{-1}\frac{\partial R}{\beta}P^{-1})$
β	$\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\frac{\partial R}{\beta}P^{-1}\Sigma^{-1}\frac{\partial K_f}{\partial \lambda})$	$\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\frac{\partial R}{\beta}P^{-1}\right)^2 + \sum_{m=1}^{M}\frac{n_i - 1}{2}\beta^2$

where $\Sigma = K_f + RP^{-1}$, $\frac{\partial R}{\beta} = R \odot x$, and $\odot$ denotes element-wise matrix multiplication.

The calculation of $\mathcal{M}$ in Equation (6) is defined for a given parameter value vector, θ_0. If a point estimate for θ is used, the design is termed locally optimal since the design is optimal for a specific parameter value θ_0, see e.g. [18].

4. Optimisation

To complete the specification of the experimental design algorithm the method of optimisation must be defined. The most commonly employed approach is to select a subset of points from a large candidate design set [36]. A complete enumeration of all possible designs quickly becomes infeasible as the number of candidate points increases. Various search strategies have been proposed in the literature to address this limitation. Some authors have suggested using a stochastic algorithm like simulated annealing with multiple restarts to guarantee robustness [36] or random sampling where an information gain is estimated for each candidate point by averaging the design score over all searches in which this point was included [33].

We have implemented two optimisation methods, Simulated Annealing (SA) and a sequential greedy optimisation algorithm. Both methods are described in Algorithms 4.1 and 4.2 respectively. The fitness function minimised in both optimisation schemes is the FIM defined in the previous section. The perturbation functions used in our SA implementation are described in Algorithm 4.3. An extensive discussion of the SA algorithm and other details are given in Section 5.5 of [5]. Greedy optimisation is a sequential procedure where at each step the input point is selected from a candidate set such that the selected point maximises the score gain. In [33] the greedy approach is shown to be superior to simple stochastic optimisation schemes through a set of simulation experiments. In experiments not reported here the Greedy and SA algorithms were found to offer good performance in a complete enumeration experiment with the latter recovering the globally optimum design [6].

One challenge with the sequential greedy optimisation method is initialisation. It is necessary to have at least two points to compute the FIM. A

Algorithm 4.1 Simulated annealing design optimisation algorithm.

```
Input: Candidate points $\mathbf{X}_C$, Target Design size $p$, degree of parallelism $d$, fitness function $f_f(\mathbf{X})$, perturbation function $f_p(x)$, initial steps to determine temperature $N_t$, maximum iteration count $M$. Output: Local optimum design $\mathbf{X}_O$.
I. Initialisation. Generate $d$ Latin Hypercube designs and for each use the steps below to set the initial temperature $T_0$..
    1. Perform $N_t$ random perturbations and evaluate the average change in fitness $< \Delta E >$.
    2. Calculate initial temperature $T_0 = \frac{-<\Delta E>}{log(0.5)}$.
A. Generate Continuous Design $\mathbf{X}_O^C$. Loop until one of the termination criteria is met.
    1. Perform perturbation on current design and calculate $\Delta E$.
    2. Metropolis Acceptance Rule: if $\Delta E \leq 0$ the perturbation is accepted. If $\Delta E > 0$ perturbation is accepted with probability $\exp(-\Delta E/T)$ where $T$ is the current temperature.
    3. Check termination conditions. If any are met proceed to step B.
        (a) Has the maximum number of iterations $M$ been reached?
        (b) $12p$ perturbations accepted or $100p$ perturbations attempted (equilibrium)?.
    4. Temperature lowered according to linear schedule $T_{k+1} = 0.9T_k$.
B. Discretise Continuous Design
    1. Match optimum continuous design $\mathbf{X}_O^C$ to candidate set $\mathbf{X}_C$ by minimising the Euclidean distance of the optimum set to candidate points. Replicate points may be introduced in this process depending on the granularity of the candidate set and the clustering of the optimum design.
```

Algorithm 4.2 Greedy design optimisation algorithm.

```
Input: Target design size $p$, design fitness function $f_f(\mathbf{X})$, Candidate set design $\mathbf{X}_C$ of size $\mathcal{C}$, Initial design $\mathbf{X}_I$. Output: Optimal design $\mathbf{X}_O$.
A. Initialise current proposal design to initial design, $\mathbf{X}_O^1 = \mathbf{X}_I$.
B. Iterate p times by adding to the current proposal design $\mathbf{X}_O$ the candidate set point which maximises the fitness function $f_f(\mathbf{X})$. Denote the iteration step as $T$.
    1. Select candidate point $X_C^i$.
    2. Evaluate the criterion function on the current proposal design appended with the candidate point, $f_f\left([\mathbf{X}_O^T; \mathbf{X}_c^i]\right)$.
    3. Permanently add the point that maximises the criterion to the current proposal design $\mathbf{X}_O^{T+1} = [\mathbf{X}_O^T; \mathbf{X}_c^i]$.
```

Algorithm 4.3 Perturbation function used in the SA algorithm.

Input: Current design $\mathbf{X}_c$, current temperature T, maximum temperature T_M.
Output: Perturbed design $\mathbf{X}_O$.

A. *Generate a random number* r in $\mathrm{U}[0,1]$. *If* $r > 0.5$ *use perturbation method* P_1*, else* P_2*.*

P_1. *Shift Single Point.*

1. Pick point $\mathbf{x}_c^i$ in design $\mathbf{X}_c$ to change at random.
2. Calculate range of shift dependant on temperature ratio T/T_M and shift $\mathbf{x}_c^i$ within the feasible region. At maximum temperature the entire design space is feasible. Specifically given the upper and lower bounds for each dimension $x_i \in [l_i, u_i]$, a random value is generated by

$$\mathbf{x}_c^i = \begin{cases} \mathbf{x}_c^i + (u_i - \mathbf{x}_c^i)\frac{T}{T_M} r_{\dots D+1} + l_i & ,\mathrm{r}_1 > 0.5 \\ \mathbf{x}_c^i - (\mathbf{x}_c^i - l_i)\frac{T}{T_M} r_{\dots D+1} + l_i & ,\mathrm{r}_1 \leq 0.5 \end{cases}$$

where $r = \{r_1, r_{\dots D+1}\}$ are $D+1$ samples from the uniform distribution $U(0,1)$, where D the dimensionality of $\mathbf{X}_c$.

P_2. *Replace Points.*

1. Calculate the number of points to replace dependant on the temperature ratio T/T_M. At maximum temperature all the points are replaced. Specifically the number of points replaced for a design size M is $\mathrm{round}(M \times \frac{T}{T_M})$ where round denotes the integer rounding operation.
2. Replace the selected number of points with randomly generated points that may lie anywhere in the design domain.

potentially useful initialisation is to evaluate the FIM for all point pairs and select the pair that achieves the minimum value. Alternatively the algorithm may be initialised by selecting the centroid point of the candidate set as the initial design point. The greedy algorithm can then proceed by selecting the point in the candidate set which, in conjunction with the centroid point, minimizes the FIM.

We have also implemented a replicate only version of the algorithm referred to as the replicate greedy optimisation. In this case, two replicates at a single design point are included at each step. This approach restricts the optimisation design space to replicate only designs which we have found in some cases to offer better solutions in terms of FIM than the standard greedy approach.

5. Simulation Experiments on Synthetic data

In this section properties of optimal designs are investigated through a range of synthetic examples. The optimal designs are compared to two types of space-filling designs, maximin Latin Hypercube and grid. The designs are assessed in terms of both prediction and parameter estimation performance. A GP with known parameters is sampled in order to assess the quality of the Maximum Likelihood (ML) parameter estimates. In all the experiments presented herein, the model used in design generation is the correct model, i.e. the same model that is sampled from to generate observations. The issue of model misspecification in optimal design is discussed further in Section 7.

The following designs are compared:

1. Greedy (F) and Simulated Annealing (S). We obtain the designs using greedy and SA optimisation respectively.
2. Grid (G). A standard grid design where the distance between neighbouring points is a constant and replication is not allowed. If the design size is not a perfect root of the input dimension, the remaining points are placed randomly.
3. Maximin Latin Hypercube (L). Maximises the minimum Euclidean distance between design points by selecting from 1000 randomly generated Latin Hypercube designs.
4. Replicate Grid (*Rg*) and Replicate Maximin Latin Hypercube (R). As a Grid and Maximin Latin Hypercube design respectively, but the number of design points is halved with 'replication' giving two samples per point.

Prediction error is assessed using the standardised mean-squared-error (sMSE) [23] and the Dawid loss [4]. The sMSE is used to assess the predictive accuracy of the GP with regards to the mean only

$$\text{sMSE} = \frac{1}{N\nu_t^2}\sum_{i=1}^{N}\left(E[t_{*i}] - t_i\right)^2$$

where $E[t_{*i}]$ the GP predictive mean defined in Equation (3) for test point $i \in \{1, \ldots, N\}$, t_i the observation at that point and ν_t^2 the sample variance of the test set observations. As the sMSE ignores the predictive variance, we utilise a multivariate extension of the logarithmic score known as the Dawid loss [4], which is defined as

$$\text{Dawid} = \log|Var[t_*]| + \left(t - E[t_*]\right)^T Var[t_*]^{-1}\left(t - E[t_{*i}]\right) ,$$

where $|\cdot|$ denotes the determinant and $Var[t_*]$ the covariance matrix of the joint predictive distribution at the set of test points (Equation (4)). By incorporating the volume of the covariance ellipsoid via the log determinant, large predictive variances are penalised in the Dawid score. The Dawid loss is a more precise error measure than the average univariate logarithmic score since the full predictive covariance is utilised without assuming the errors are uncorrelated. The test set used is a 1024 point Latin Hypercube design.

In order to measure the accuracy of parameter estimation we use two measures, the parameter Mean Absolute Error (pMAE) and the Log Determinant of the ML estimator parameter covariance (LDM). The LDM is defined as the log determinant of the covariance of the ML estimates of all parameters across all realisations of the experiment under consideration. It is a measure of dispersion of the ML estimates and does not capture the error of the estimations with respect to the true parameters. However the FIM (Section 3) should approximate the Log determinant of the ML estimates and the quality of this approximation is a useful diagnostic for the performance of the design. The pMAE on the other hand is an estimate of the error of the ML estimate to the true parameter value, pMAE $= 1/N_E \sum_{i=1}^{N_E} |\hat{\theta}_i - \theta_0|/|\theta_0|$ where $|\cdot|$ the absolute value, $\hat{\theta}_i$ is the ML point estimate for realisation i, θ_0 the true parameter and N_E the number of realisations. The rescaling by θ_0 ensures the pMAEs for different parameters are comparable. To ensure robustness in the calculation of the pMAE, the maximum likelihood optimisation is restarted five times from random initial conditions for all parameters. The solution with the highest training set likelihood is selected for subsequent validation. For the multiple restarts the initial value for the log length-scale parameter was sampled from $\mathcal{N}(-2, 0.01)$, i.e. a Normal distribution centred at -2 corresponding to a length-scale of ≈ 0.1. A small variance was used to avoid numerical issues in the calculation of the model likelihood. All other parameters were initialised by sampling from $\mathcal{N}(0, 1)$. To help present the results concisely the pMAE for the variance models parameters are aggregated in a single summary. The median and interquartile range (IQR) are reported for all pMAEs based on multiple realisations of the experiments.

The design space for all experiments is set to $\mathbf{X} \in [0,1]^2$. A zero mean GP with a fixed order $\nu = 5/2$ Matérn kernel is used for both generation and fitting. A series of further experiments on other kernels and variance models is presented in Chapter 5 of [5] whose findings are consistent with the for results presented here.

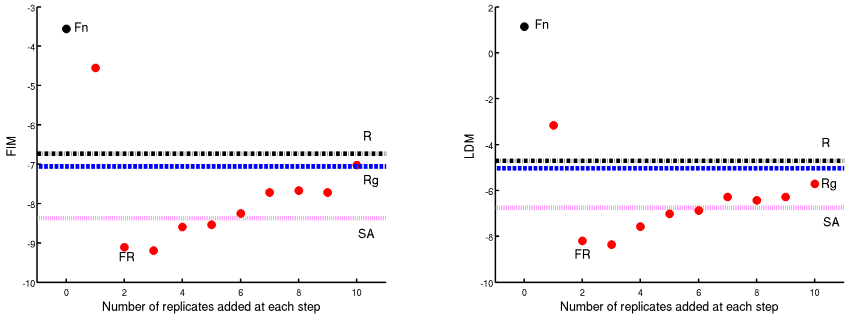
5.1. Local Design

In this section locally optimal designs are investigated using a synthetic example. To evaluate the parameter errors, 500 realisations of the experiment are performed. All designs were generated using a 1024 grid space of candidate points, picking $n = 30$ points and allowing for replication. The Greedy algorithm was initialised by computing the FIM for all possible permutations for two point designs and selecting the pair with the minimum value (Section 4).

A Fixed basis Log-Linear model variance model is used $\exp(\beta_1 + \beta_2 x_1 + \beta_3 x_2)$. The GP model length scale is set to $\lambda = 0.2$, the process variance to $\sigma_p = 1$, the variance model intercept to $\beta_1 = -4.6$ and slope to $\beta_2 = \beta_3 = -1.6$.

The Greedy algorithm can be run with a fixed number of replicates added at each step. In Figure 1 the FIM and LDM values for different Greedy designs is shown. The Greedy design where replication is not allowed (Fn) is the worst performing design both in term of FIM and LDM. Allowing for replication but adding a single point at each Greedy step improves both scores but results in a design that is still worse than the replicate Grid (Rg) and replicate Latin Hypercube (R) designs. Adding 2 replicates at each step (FR) significantly increases the scores of the resulting design outperforming the SA design, suggesting the SA optimisation could be run for longer. Adding three replicates at each step improves only modestly the design and adding more replicates impacts negatively on both scores. In this instance there is good agreement between the FIM and the actual design performance in terms of parameter error as reflected by the LDM. This approach may therefore be used in practice to judge how many replicates to add at each Greedy step.

The designs obtained using the no-replicate (Fn) and 2-replicate Greedy (FR) and SA (S) optimisation methods are shown in Figure 2; the no replicate Fn design places point along a line at the boundaries of the design space while the replicate FR and S designs place points on the corners of the space. The LDM agrees with the FIM (Figure 2(d)) with regards to separating the non-replicate designs (which show a large variability in estimated parameters) from the replicate designs. The lowest FIM and LDM values are obtained by the FR and SA designs.



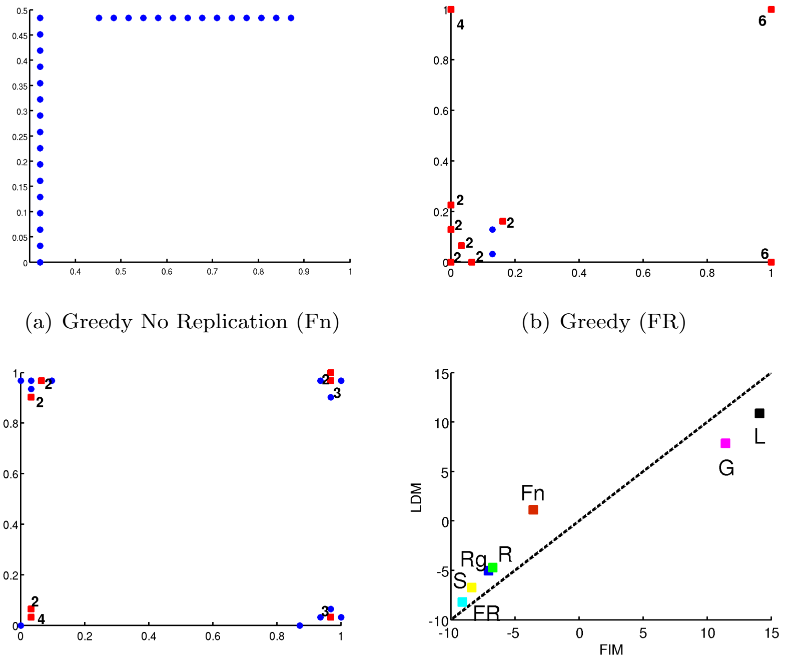
(a) Replication step size vs FIM

(b) Replication step size vs LDM

Figure 1: FIM and LDM for different replicate step sizes for Greedy algorithm. For step size 0, the FIM and LDM for the non-replicate design is shown - see Figure 2(a). For reference the FIM and LDM scores for the SA and Replicated Grid and Maximin Latin Hypercube designs are also shown as horizontal dashed lines. The non-replicate (Fn) and 2-replicate at each step Greedy designs are explicitly labelled.

In terms of parameter estimation accuracy, all variance model parameters β are better identified in the replicate designs as is shown in Table 1. Specifically the replicate designs FR, S, Rg, R achieve lower median pMAEs than the non-replicate designs Fn, G, L which incur much higher median errors and are also more variable in their performance as reflected by the increased IQR values. In terms of the length-scale parameter, the non-replicate Fn, G, L designs achieve somewhat smaller pMAEs than the corresponding replicate designs. No practically relevant differences were observed in the estimation of the process variance parameter. In this scenario the replicate designs are superior in identifying the variance model parameters without significantly sacrificing the estimation of the length-scale parameter.

In terms of predictive errors, the space-filling non-replicate designs (G, L) achieve the lowest average sMSE of 0.2, followed by the space-filling replicate designs (Rg,R) with mean sMSE 0.4 and finally the optimal non-replicate Greedy (Fn) with average sMSE 0.8, replicate Greedy (FR) and SA (S) designs with average sMSE 0.8 and 0.9 respectively. As space-filling designs cover the space more uniformly than the highly clustered optimal designs the smaller interpolation error on the mean is expected. In terms of Dawid loss (Table 1) the non-replicate Fn, G, L designs achieve significantly worse median errors than



(a) Greedy No Replication (Fn)

(b) Greedy (FR)

(c) Simulated Annealing (S)

(d) Correspondence of FIM to LDM

Figure 2: Greedy designs with replication (FR), without (Fn) and SA (S) designs for the Log-Linear model. Replicate points (red squares) are annotated with the number of replicates n_i, non-replicate points are not (blue circles). Also shown a comparison of these designs to the replicate (Rg) and non-replicate (G) Grid, replicate (R) and non-replicate (L) maximin Latin Hypercube designs. Replicate points shown as red squares, single replicate point as blue circles.

Table 1: Median and interquartile range for the pMAE and Dawid loss for the Log-Linear model.

Design	λ	$\beta_{1,2,3}$	Dawid
Greedy (FR)	0.21 ± 0.25	0.19 ± 0.33	-4170 ± 296
Simulated Annealing (S)	0.18 ± 0.23	0.26 ± 0.47	-4134 ± 333
Non-replicate Greedy (Fn)	0.17 ± 0.22	0.96 ± 2.16	-3391 ± 4001
Replicate Grid (Rg)	0.32 ± 0.50	0.31 ± 0.56	-3918 ± 990
Replicate Maximin LH (R)	0.26 ± 0.31	0.34 ± 0.61	-4065 ± 462
Grid (G)	0.18 ± 0.26	1.18 ± 2.05	21800 ± 87493
Maximin Latin Hypercube (L)	0.19 ± 0.26	1.22 ± 2.61	8195 ± 62011

the replicate FR, S, Rg, R designs. The larger interquartile values for the former are striking and point to a lack of robustness in the prediction. This is consistent with the larger IQR values for the pMAEs of the variance model parameters for these designs. In conjunction with the larger variance parameter errors and smaller sMSEs, we conclude that the non-replicate designs have higher Dawid loss mainly due to inaccurate variance prediction.

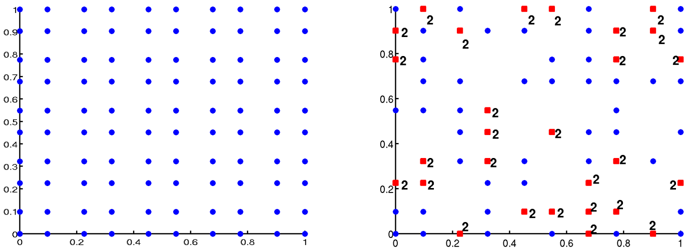
For the heteroscedastic Log-Linear model a design that is optimal for the identification of the coefficients of the log-linear variance model is required. As is well known in the case of linear regression [2], the optimal design for parameter estimation places points on the corners of the space and this is exactly the effect we observe in the SA and FR optimal designs for the Log-Linear model. The parameter estimation errors lend further credence to this conclusion as the optimal designs achieve lower errors for the variance model parameters β than the non-replicate space-filling designs. The good performance of the replicate space-filling designs is also explained by this effect since replicated design points are placed on the edges of the design space. As the noise level is quite low across the design space, design points with just two replicated observations are sufficient to capture the variance response. In the case of the non-replicate designs however, the single observation design points on the edge of the space are not as informative with regards to the variance process.

5.2. On the monotonicity of the FIM

As [36] have noted, for the FIM to be used as a design criterion, it should provide the same ordering of designs as the LDM. Based on a small simulation experiment with homoscedastic noise, they conjecture that such a monotonic relationship exists, although they note the approximation error is significant for small design sizes. In our simulation experiments under heteroscedastic noise we have found a strict monotonic relationship to be violated. However we believe an approximate ordering still holds which we empirically demonstrate.

For the first simulation experiment we consider two types of design, Grid and Latin Hypercube. For each type of design, we start with no replicates and increase the number of points with two replicates by simultaneously removing non-replicate design points to maintain a constant design size of 100 points. Examples of the designs are shown in Figure 3. A Latent-Kernel variance model with 9 latent points is used. The latter are placed on a grid in the design space. For each design the local FIM is calculated. As in [36], the experiment is performed by sampling from a GP with known parameters. The length scale prior is set to $\lambda = \{0.6\}$, the process variance is kept fixed at 0.6, and the variance model parameters to $z = \{3.5\}$. This configuration has a medium length scale process with relatively small changes in the mean response and large variability in the variance response. For all designs, the parameters are estimated using ML with 100 realisations of the experiment performed, each utilising a different GP sample.

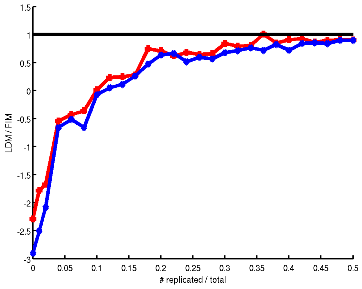
The experiment is summarised by the plot shown in Figure 4. The ratio of the LDM to the FIM is used to summarise the approximation error. The correspondence of this ratio to the ratio of replicated points in the design is



(a) Grid, no replication (b) Grid, some replicated points

Figure 3: Examples of designs considered in the FIM vs LDM consistency experiment.

plotted. The latter is defined as the ratio of design points with two replicates to the total number of points in the design (100). As the ratio of replicated points is increased, the LDM/FIM ratio approaches 1 reflecting the decrease in the approximation error. Further, when few replicate points are available in the design, the value of the LDM/FIM ratio reflects the underestimation by the FIM of the parameter variance as reflected by the LDM. But as [36] have noted, the critical property for a design criterion is the monotonicity of the FIM-LDM relationship and not the magnitude of the approximation error.



Fisher.

Figure 4: Relation of FIM to the LDM for a fixed design size of 100 points. Grid (red solid line) and Latin hypercube (blue dashed line) designs with different ratios of replicated to non-replicated points considered.

To establish whether strict monotonicity holds in the simulation experiment we compute a violation measure on the intermediary designs produced by the Simulated Annealing (SA) optimisation algorithm in Section 5.1. The final SA design is shown in Figure 2(c). The design used to initialise the SA algorithm is a Latin Hypercube with no replicated points. As the algorithm proceeds we store the design every 100th iteration giving a total of 258 designs. We split the designs into 9 categories depending on the number of replicated ($n_i > 1$) points C_r in the design (Table 2). We define a violation measure to investigate the departure within each category from strict monotonicity. The measure is defined as:

$$V(\xi) = \sum_{i \neq c}^{N_\xi} \delta_{ci} \left| \left(M(\xi) - M(\xi_i)\right) \left(L(\xi_i) - L(\xi)\right) \right|, \tag{7}$$

where ξ is the evaluated design, N_ξ the number of designs in the same category as ξ, and $M(\dots)$, $L(\dots)$ the FIM and LDM functions respectively. The indicator function $\delta_{ci} = I\left[\left(M(\xi) - M(\xi_i)\right)\left(L(\xi_i) - L(\xi)\right) > 0\right]$ returns 1 if a violation has occurred and 0 otherwise. As we see in Table 2 the violation measure is highest for designs without replicated points and is rapidly reduced when even a single replicate point is included. The approximation error of the FIM to the LDM is therefore smaller and the FIM criterion more robust when replicated points are included. For this reason we will restrict our space of candidate designs to only replicate designs in Section 6 where the systems biology application is considered.

Table 2: Number of replicated points C_r per design category, number of designs N_ξ in each category and the normalised monotonicity violation measure $V(\xi)$ for the SA intermediate designs.

C_r	N_ξ	$V(\xi)/N_\xi$	C_r	N_ξ	$V(\xi)/N_\xi$
0	61	16.52			
1	57	1.36	5	9	0.00
2	47	0.31	6	17	0.14
3	28	0.09	7	7	0.03
4	27	0.14	8	5	0.00

6. Application to prokaryotic autoregulatory network

In this section we discuss the application of the optimal design methodology to a stochastic simulator describing the autoregulatory function of prokaryotic organisms. This simulator exhibits input dependent variance requiring the use of our heteroscedastic GP model described in Section 2 when constructing an emulator.

6.1. The prokaryotic autoregulatory network

The simulator describes a simple gene expression auto-regulation mechanism often present in prokaryotic gene networks. It is composed of five reactant species, the gene g, protein P and its dimer P_2, and the mRNA molecule. The eight reactions complete the specification of the model [32]:

$$\begin{array}{rll}
g + P_2 & \xrightarrow{k_1} g.P_2 & \text{Repression} \\
g.P_2 & \xrightarrow{k_2} g + P_2 & \text{Reverse Repression} \\
g & \xrightarrow{k_3} g + r & \text{Transcription} \\
r & \xrightarrow{k_4} r + P & \text{Translation} \\
2P & \xrightarrow{k_5} P_2 & \text{Dimerisation} \\
P_2 & \xrightarrow{k_6} 2P & \text{Dissociation} \\
r & \xrightarrow{k_7} \emptyset & \text{mRNA degradation} \\
P & \xrightarrow{k_8} \emptyset & \text{Protein degradation}
\end{array}$$

Dimers of the protein P (P_2) coded for by the gene g repress their own transcription by binding to a repressive regulatory region upstream of g. This model is minimal in terms of biological detail included but contains many of the interesting features of an auto-regulatory feedback network [32]. Simulations of the network are implemented using the stochastic Gillespie algorithm [32]. The resulting model is stochastic as the simulation considers interactions for each molecule in the system under consideration, and the interaction of these molecules is inherently random [32].

Following [31], we restrict our attention to the k_6 and k_7 reaction rate parameters with range $k_6 \in [0, 7]$ and $k_7 \in [0.05, 0.4]$. The other parameters are set to reference values ($k_1 = 1, k_2 = 10, k_3 = 0.01, k_4 = 10, k_5 = 1, k_8 = 0.01$) [31]. The initial number of molecules were set to $\{g.P_2, g, r, P, P_2\} = \{100, 0, 0, 0, 0\}$. The response we have selected to emulate is the number of bound molecules $g.P_2$ at time step $T = 18$. A linear trend has been removed from the mean response using ordinary least squares regression, as we will assume a zero mean GP prior for the regression model in both the design and inference stages.

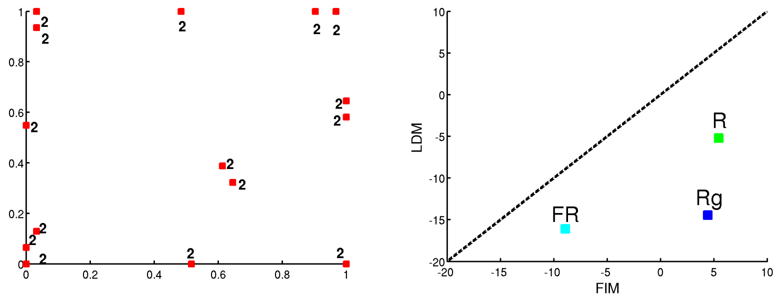
6.2. Local Design

We use the same kernel and design space as specified in Section 5. For the variance model, we utilise a nine point latent kernel structure. The latent kernel points X_z are placed on a grid in the interior of the design space. Specifically the grid is placed in the region $[0.2, 0.8]^2$. This is done to avoid placing latent basis functions on the edge of the design space where the training design is least informative. For the variance kernel a Matérn kernel with fixed differentiability $\nu = 5/2$ is used. We perform 500 realisations of the experiment.

We utilise a locally optimum design by specifying a single set of parameter values for design generation. This scenario aims to demonstrate the case where strong prior information regarding the simulator response is available. A process length-scale of $\lambda = \{0.04\}$ is assumed with the process variance set to $\sigma_p^2 = 0.36$ and the variance model coefficients to $z_{1,\ldots,9} = 2$.

The experiment consists of comparing three different design methodologies for a small design size of 30 points. Replicate-only designs are used since replicate designs allow for more robust estimation of the variance model parameters as well as reducing the approximation error of the FIM to the LDM (Section 5.2). The optimal design is produced using the 2-replicate greedy (FR) algorithm initialised using the centroid of candidate set discussed in Section 4. We compare the performance of the greedy design to a replicate Grid (Rg) and replicate Maximin Latin Hypercube (R) design. The SA algorithm was unable to produce a design with a lower FIM than that achieved by the greedy design. Further,the best SA design showed a pattern similar to the greedy design (see Section 6.3.4 of [5]).

The optimal design is shown in Figure 5(a). The design exhibits a particular structure, placing points in the centre and edges of the design space with a high number of replicates. These areas correspond to the locations of the latent points of the latent kernel variance model.



(a) Greedy (FR) design

(b) Correspondence of FIM to LDM

Figure 5: (a) Optimal design and (b) monotonicity plot in the prokaryotic autoregulatory network example. Cyan = optimal design (FR), blue = Replicate Grid design (Rg), green = replicate Maximin Latin Hypercube (R) design.

To calculate the pMAE, the 'true' parameter vector θ_0 was estimated using the entire candidate set as the training set. The median and IQR of the pMAE for each design is shown in Table 3. The variance model parameters are identified with similar accuracy and variability with one exception not apparent in Table 3. For parameter β_3 the R design has median pMAE of 0.83 (IQR 1.05), much higher than for the FR and Rg designs with values 0.35 (IQR 0.44) and 0.23 (IQR 0.29) respectively. The random nature of the R design leads to a design that misses placing points around the location corresponding to the β_3 parameter. The length scale λ and process variance σ_p^2 parameters are estimated with least error in the optimal design. However the most striking differences are in the IQR values for the length scale parameter where the optimal design exhibits the smallest variability. The Rg and R designs cannot robustly estimate the length-scale parameter for small design sizes since points are quite far apart and cannot resolve a small length-scale. The reduced variability in the estimation of the model parameters for the optimal design is summarised by the LDM measure shown in Figure 5(b) where a monotonic relationship of the FIM to the LDM is also evident.

Table 3: Median pMAE for the prokaryotic autoregulatory network model. Interquartile range in parenthesis.

Design	λ	σ_p^2	β
Greedy (FR)	1.23 ± 2.49	0.96 ± 2.74	0.28 ± 0.37
Replicate Grid (Rg)	1.78 ± 7.54	1.03 ± 2.95	0.23 ± 0.30
Replicate Maximin LH (R)	1.39 ± 4.89	1.13 ± 3.16	0.26 ± 0.38

In terms of predictive errors, all designs achieve sMSEs of 1.00 reflecting similar performance on mean prediction. The optimal design achieves the smallest Dawid loss with a median value of 6194 (1018), followed by the Rg and R median loss of 6324 (1238) and 7903 (6045) respectively. The large Dawid loss for the R design is due to the estimation error for the variance model β_3 parameter highlighted above. Overall the optimal design achieves the smallest median Dawid loss with the smallest variability (IQR) due to the more robust estimation of the length-scale and variance model parameters.

7. Summary and Discussion

In this paper we have presented a new approach to model-based optimal design for heteroscedastic regression models with correlated errors and examined empirically the performance of the optimal designs through an extensive set of simulation studies. The criterion we have used aims to minimise the estimation error of the GP covariance parameters. This can be of use for variable screening and uncertainty quantification.

In contrast to [36] we have found a strict ordering of the FIM to LDM does not hold for heteroscedastic models. However we have found that as the ratio of design points with replicates is increased, the approximation error of the FIM to the LDM is reduced and the monotonic relationship is more likely to hold. We believe this is related to the reduced inferential uncertainty on the variance model parameters when replicate points are used. We hypothesise that as the uncertainty on the parameters increases, the FIM to LDM approximation error increases. We believe a deeper theoretical understanding of this conclusion is a worthwhile direction for future research.

For both the synthetic example and prokaryotic autoregulatory network case study the predictive performance of all replicate designs was found to be superior to that of the non-replicate Grid and Maximin Latin Hypercube designs. Although the sMSE was lower for the space filling non-replicate designs, reflecting a lower error on the mean, the replicate designs achieved more accurate variance prediction through the better identification of the variance model parameters, thus producing better calibrated probabilistic GP models as evidenced by the lower Dawid losses. We suggest that under non-trivial noise regimes, employing this model-based strategy and considering replicate-only designs, i.e. designs with at least two replicates at each point, can be a very effective strategy for identifying the regression model parameters.

The methodology presented can be extended in a variety of ways. It is relatively simple to extend the locally optimal design methods to Bayesian optimal designs by integrating over the unknown parameters to compute an expected FIM [36, 5]. In work not presented in the paper for brevity these conclusions have been shown to extend to the Bayesian version of the FIM (see Chapter 5 of [5]).

We envisage the usage of our design method for approaches that linearise the correlated process using functional expansions. In [9] the GP covariance is approximated by a truncated eigenvector expansion. The approximation error of the expansion critically depends on the parameter accuracy. A Latin Hypercube is used in [34] for the initial design but a more natural choice would be a design where the parameter estimation variance is explicitly minimised.

In [13], a two stage exploration-exploitation sequential strategy is proposed. In the exploration phase a variety of designs are proposed to minimise parameter uncertainty while in the exploitation phase, the parameters are assumed to be known with sufficient accuracy to allow for the minimisation of the predictive variance. The designs we have proposed would be a natural choice to employ during the exploration phase within such a framework.

In this work the focus has been exclusively on design for identifying the covariance parameters. In practice, a non-constant mean function is specified in the GP prior as it produces an efficient and flexible model structure. It is well known in the literature (e.g. [17]) that design for trend parameters is usually antithetical to that of covariance parameters. Combining design for trend and covariance parameter estimation in the heteroscedastic emulation context is an area for future research.

Our work can also be used to motivate design strategies relying on geometric criteria. For simple stochastic responses, incorporating some replicate design points into the geometric design can substantially reduce the estimation error of the variance model parameters. For more complex noise models, a hybrid design approach, where model-based criteria such as the FIM are combined with geometric criteria, such as coverage of the input space, may also be possible. Alternatively, for specific cases a deeper understanding of the underlying geometry implied by the FIM can lead to corresponding geometric criteria which would be easier to compute. This is a promising area where only preliminary results exist, especially in the field of correlated processes. In [25] a parabola reflection transformation is used to produce space-filling designs that can identify the correlation parameter in a Chemometrics model. In more realistic setups of model-based design, an appropriate geometry derived from FIM would be non-Euclidean and geodesic lines may be pretty curved (and only in rare cases will have an easy Euclidean parametrisation). Tackling this research question in a realistic setup is challenging and we suggest it as a future research problem.

Overall our work suggests the following recommendations:

- optimal model-based designs can be very useful where there is a reasonable level of prior information on the model structure and parameter values;

- when using FIM based optimal design for noisy correlated processes replicate observations should be used;
- the FIM based design approach developed in this work produces better calibrated probabilistic models when compared to other designs, including space filling designs;
- in high input dimensions model-based design becomes challenging – geometric designs should incorporate replicate points.

Acknowledgements

The work on the paper was partially undertaken at Isaac Newton Institute for Mathematical Sciences at Cambridge, UK, during the Design and Analysis of Experiments research programme hosted by the institute. The 3rd author was partially supported by ANR project Desire. We also wish to thank Ian Vernon for useful discussions on the Prokaryotic Autoregulatory network used in Section 6. This work was supported by EPSRC / RCUK as part of the MUCM Basic Technology project (EP/D048893/1). We thank the editor and reviewers, whose insightful comments helped us to sharpen the paper considerably.

References

[1] M. Abt and W. J. Welch. Fisher information and maximum likelihood estimation of covariance parameters in Gaussian stochastic processes. *Canadian Journal of Statistics*, 26:127–137, 1998.

[2] A. C. Atkinson and A. N. Donev, editors. *Optimum Experimental Designs.* Oxford University Press, 1992.

[3] S. Baran, K. Sikolya, and M Stehlík. On the optimal designs for prediction of Ornstein-Uhlenbeck sheets. *Statistics and Probability Letters*, 83(6):1580–1587, 2013.

[4] L. S. Bastos and A. O'Hagan. Diagnostics for Gaussian process emulators. *Technometrics*, 2009.

[5] A. Boukouvalas. *Emulation of Random Output Simulators.* PhD thesis, Aston University, 2011. Available at `wiki.aston.ac.uk/foswiki/pub/AlexisBoukouvalas/WebHome/thesis.pdf`.

[6] A. Boukouvalas, D. Cornford, and M. Stehlík. Notes on optimal design for correlated processes with input-dependent noise. Technical Report `https://wiki.aston.ac.uk/AlexisBoukouvalas`, Non-Linear Complexity Group, Aston University, 2013.

[7] H. Dette, A Pepelyshev, and A Zhigljavsky. Nearly universally optimal designs for models with correlated observations. computational statistics and data analysis. *Computational Statistics and Data Analysis*, 2013.

[8] P.J. Diggle, R. A. Moyeed, and J. A. Tawn. Model-based geostatistics. *Applied Statistics*, 47:299–350, 1998.

[9] V. Fedorov and W. Müller. Optimum design for correlated fields via covariance kernel expansions. *mODa 8 - Advances in Model-Oriented Design and Analysis Contributions to Statistics*, pages 57–66, 2007.

[10] R. H. Green. *Sampling design and statistical methods for environmental biologists.* Wiley, 1979.

[11] D. A. Henderson, R. J. Boys, K. J. Krishnan, C. Lawless, and D. J. Wilkinson. Bayesian emulation and calibration of a stochastic computer model of mitochondrial dna deletions in substantia nigra neurons. *Journal of the American Statistical Association*, 104(485):76–87, 2009.

[12] J. Kiselák and M. Stehlík. Equidistant and D-optimal designs for parameters of Ornstein-Uhlenbeck process. *Statistics & Probability Letters*, 78(12):1388–1396, September 2008.

[13] A. Krause and C. Guestrin. Nonmyopic active learning of gaussian processes: an exploration-exploitation approach. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 449–456, New York, NY, USA, 2007. ACM.

[14] K. V. Mardia and R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–146, 1984.

[15] J.M. McGree, J.A. Eccleston, and S.B Duffull. Compound optimal design criteria for non-linear models. *Journal of Biopharmaceutical Statistics*, 18:646–661, 2008.

[16] W. Müller and M. Stehlík. Issues in the optimal design of computer simulation experiments. *Applied Stochastic Models in Business and Industry*, 25(2):163–177, 2009.

[17] W. G. Müller and M. Stehlík. Compound optimal spatial designs. *Environmetrics*, 21(3-4):354–364, 2010.

[18] W. G. Müller and D. L. Zimmerman. Optimal design for variogram estimation. *Environmetrics*, 10:23–37, 1993.

[19] A. Pázman. Correlated optimum design with parameterized covariance function: Justification of the fisher information matrix and of the method of virtual noise. Technical Report 5, Department of Statistics and Mathematics, Wirtschaftsuniversitat Wien, June 2004.

[20] A. Pázman. Criteria for optimal design of small-sample experiments with correlated observations. *Kybernetika*, 43(4):453–462, 2007.

[21] A. N. Pettitt and A. B. McBratney. Sampling designs for estimating spatial variance components. *Applied Statistics*, 42(1):185–209, 1993.

[22] Luc Pronzato and Werner G. Müller. Design of computer experiments: space filling and beyond. *Statistics and Computing*, 22(3):681–701, 2012.

[23] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006.

[24] Soora Rasouli and Harry Timmermans. Using emulators to approximate predicted performance indicators in complex micro-simulation and multi-agent models of travel demand. In *4th Transportation Research Board Conference on Innovations in Travel Modeling*, 2012.

[25] J.M. Rodrguez-Daz, M.T. Santos-Martn, H. Waldl, and M. Stehlík. Filling and d-optimal designs for the correlated generalized exponential models. *Chemometrics and Intelligent Laboratory Systems*, 114(0):10 – 18, 2012.

[26] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4:409–435, 1989.

[27] M. L. Stein, editor. *Interpolation of Spatial Data: Some Theory for Kriging.* New York: Springer-Verlag, 1999.

[28] M. L. Stein. *Statistical Interpolation of Spatial Data: Some Theory for Kriging.* Springer, 1999.

[29] L. Tack, P. Goos, and M. Vandebroek. Efficient bayesian designs under heteroscedasticity. *Journal of Statistical Planning and Inference*, 104(2):469 – 483, 2002.

[30] N. Uddin. Mv-optimal block designs for correlated errors. *Statistics and Probability Letters*, 78:2926–2931, 2008.

[31] I. R. Vernon and M. Goldstein. A bayes linear approach to systems biology. Mucm technical report 10/10, Durham University, 2010.

[32] D. J. Wilkinson. *Stochastic Modelling for Systems Biology.* Chapman & Hall/CRC, 1st edition, 2006.

[33] G. Xia, M. L. Miranda, and A. E. Gelfand. Approximately optimal spatial design approaches for environmental health data. *Environmetrics*, 17(4):363–385, 2006.

[34] N. Youssef. *An orthonormal function approach to optimal design for computer experiments.* PhD thesis, London School of Economics, UK, 2010.

[35] H. Zhang and D. L. Zimmerman. Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, 92(4):921–936, December 2005.

[36] Z. Zhu and M. L. Stein. Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference*, 134(2):583 – 603, 2005.

[37] Z. Zhu and M. L. Stein. Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(1):24–44, March 2006.

[38] D. L. Zimmerman. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics*, 17(6):635–652, 2006.

Appendix A. Joint likelihood Model derivation

We derive the likelihood for the model defined in Equation (2). Assuming normality, the sample variance is distributed as a scaled $\mathcal{X}^2$ distribution with $n_i - 1$ degrees of freedom:

$$s_i^2 \sim \frac{g_{\sigma^2}(x,\beta)}{n_i - 1}\mathcal{X}^2_{n_i-1} \ ,$$

where n_i the number of replicates at location x_i. This can also be expressed as a Gamma distribution:

$$p(s_i^2|\beta, x_i, n_i) \sim \Gamma\left(\frac{n_i - 1}{2}, \frac{2g_{\sigma^2}(x,\beta)}{n_i - 1}\right),$$

The joint log likelihood of the sample mean $\bar{t}$ and variance s^2 for N observations can then be derived:

$$\log p(\bar{t}, s^2|\mathbf{X}, \theta_f, \beta) = \left(\sum_{i=1}^{N} \log p(s_i^2|\beta, x_i, n_i)\right) + \log \mathcal{N}(\bar{t}|0, K_f + RP^{-1}). \quad \text{(A.1)}$$

The notation $\mathcal{N}(x|\mu, \Sigma)$ is used to denote the pdf of a normally distributed random variable x with mean μ and covariance Σ. The joint likelihood of the sample mean $\bar{t}$ and sample variance s^2 is:

$$\begin{aligned}
p(\bar{t}, s^2|\mathbf{X}, \theta_f, \beta) &= \int p(\bar{t}, s^2, f|\mathbf{X}, \theta_f, \beta) df \\
&= \int p(\bar{t}, s^2|f, \mathbf{X}, \theta_f, \beta) p(f|\theta_f) df \\
&= \left(\prod_{i=1}^{N} p(s_i^2|x_i, \beta) \right) \int p(\bar{t}|f, \theta_f, \beta, \mathbf{X}) p(f) df \\
&= \left(\prod_{i=1}^{N} p(s_i^2|x_i, \beta) \right) \mathcal{N}(\bar{t}|0, K_f + RP^{-1}) \, .
\end{aligned} \tag{A.2}$$

The last equality follows from the law of total variance. The log likelihood can then be written $\log p(\bar{t}, s^2|\mathbf{X}, \theta_f, \beta) = \left(\sum_{i=1}^{N} L_{si}\right) + L_N$ where the latter term is a GP standard likelihood with the given covariance and the former can be expanded:

$$\begin{aligned}
\log p(s_i^2|\beta, x_i) =& \frac{n_i - 1}{2} \left(\log(n_i - 1) - \log(2) - \log g_{\sigma^2}(x_i, \beta)\right) - \log \Gamma(\frac{n_i - 1}{2}) \\
&+ \frac{n_i - 3}{2} \log(s_i^2) - \frac{(n_i - 1)s_i^2}{2 g_{\sigma^2}(x_i, \beta)}.
\end{aligned} \tag{A.3}$$

Appendix B. Proof of Fisher Information for Heteroscedastic Noise Models

For the heteroscedastic GP model with parameters θ_j,$\theta_p \in \{\theta_f, \beta\}$ the corresponding element in the FIM is:

$$\mathcal{M}_{jp} = - \int \int \left(\frac{\partial^2}{\partial \theta_j \theta_p} \log p(\bar{t}, s^2|\theta_f, \beta, n) \right) p(\bar{t}, s^2|\theta_f, \beta, n) \, \mathrm{d}\bar{t} \mathrm{d}s^2,$$

where $n = \sum n_i$ the total number of replicates in the design. We omit the dependency on the inputs $\mathbf{X}$ for brevity.

The log likelihood term can be decomposed into two terms as shown in Equation (A.1), a term dependent on the distribution of the sample variances, L_{si}, and a Gaussian Process term L_N.

$$\begin{aligned}
\mathcal{M}_{jp} &= -\int \int \left[\frac{\partial^2}{\partial \theta_j \theta_p} \sum L_{si} \right] p(\bar{t}, s^2|\theta_f, \beta, n) \, \mathrm{d}\bar{t} \mathrm{d}s^2 - \int \int \left[\frac{\partial^2}{\partial \theta_j \theta_p} L_N \right] p(\bar{t}, s^2|\theta_f, \beta, n) \, \mathrm{d}\bar{t} \mathrm{d}s^2 \\
&= -\int \left[\frac{\partial^2}{\partial \theta_j \theta_p} \sum L_{si} \right] p(s|\beta, n) \, \mathrm{d}s^2 \int p(\bar{t}) \, \mathrm{d}\bar{t} - \int \left[\frac{\partial^2}{\partial \theta_j \theta_p} L_N \right] p(\bar{t}) \, \mathrm{d}\bar{t} \int p(s|\beta, n) \mathrm{d}s^2.
\end{aligned}$$

We are able to separate the sample variance integrals to the individual s_i terms due to the noise independence assumption, i.e. $p(s^2|\beta, n) = \prod_{i=1}^{N} p(s_i^2|\beta, n_i)$.

$$\begin{aligned}
\mathcal{M}_{jp} &= -\int \left[\frac{\partial^2}{\partial\theta_j\theta_p}\sum_{i=1}^{N} L_{si}\right]\prod p(s_i^2|\beta, n_i)\,\mathrm{d}s^2 + \mathcal{M}_N \\
&= -\sum_{i=1}^{N}\left(\int\left[\frac{\partial^2}{\partial\theta_j\theta_p}L_{si}\right]p(s_i^2|\beta, n_i)\,\mathrm{d}s_i^2\int\prod_{j\neq i}^{N}p(s_i^2|\beta, n_i)\,\mathrm{d}s_j\right) + \mathcal{M}_N \\
&= \sum_{i=1}^{N}\mathcal{M}_{si} + \mathcal{M}_N\,,
\end{aligned} \tag{B.1}$$

where

$$\begin{aligned}
\mathcal{M}_{si} &= -\int\left[\frac{\partial^2}{\partial\theta_j\theta_p}\log p(s_i^2|\beta, n_i)\right]p(s_i^2|\beta, n_i)\,\mathrm{d}s_i^2, \\
\mathcal{M}_N &= -\int\left[\frac{\partial^2}{\partial\theta_j\theta_p}L_N\right]p(\bar{t})\,\mathrm{d}\bar{t}\,.
\end{aligned}$$

The solution to the $\mathcal{M}_N$ integral for a zero mean GP is $\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_p})$ [19]. The $\mathcal{M}_{si}$ integral can be solved by rewriting the integral in terms of the second order derivative of the variance model $\frac{\partial^2 f}{\partial\beta_j\beta_p}$:

$$\begin{aligned}
\mathcal{M}_{si} = -\int\frac{\partial^2\log p(s_i^2|\beta, n_i)}{\partial\beta_j\beta_p}p(s_i^2|\beta, n_i)\,\mathrm{d}s_i^2 = \frac{n_i-1}{2}\frac{\partial^2 f}{\partial\beta_j\beta_p}\int p(s_i^2|\beta, n_i)\,\mathrm{d}s_i^2 \\
-\frac{(n_i-1)}{2}\left[-\exp(-f)\frac{\partial f}{\partial\beta_j}\frac{\partial f}{\partial\beta_p} + \exp(-f)\frac{\partial^2 f}{\partial\beta_j\beta_p}\right]\int s_i^2 p(s_i^2|\beta, x_i)\,\mathrm{d}s_i^2\;.
\end{aligned}$$

The integral can be analytically solved. For notational brevity let $g_{\sigma^2} = g_{\sigma^2}(x_i, \beta) = \exp(f)$.

$$\int s_i^2 p(s_i^2|\beta, x_i)\,\mathrm{d}s_i^2 = \frac{\frac{n_i-1}{2g_{\sigma^2}}^{\frac{n_i-1}{2}}}{\Gamma(\frac{n_i-1}{2})}\int s_i^2(s_i^2)^{\frac{n_i-3}{2}}\exp\left(-\frac{n_i-1}{2g_{\sigma^2}}s_i^2\right)\mathrm{d}s_i^2. \tag{B.2}$$

The last integral is the mean of Gamma distribution. Therefore the Gamma integral is $\frac{2g_{\sigma^2}}{n_i-1}\frac{n_i-1}{2} = g_{\sigma^2}$. To conclude the Fisher information contribution of the sample variance term of the log likelihood $\mathcal{M}_{si}$ is:

$$\mathcal{M}_{si} = \frac{n_i-1}{2}\left(\frac{\partial^2 f}{\partial\beta_j\beta_p} - \frac{\partial^2 f}{\partial\beta_j\beta_p} + \frac{\partial f}{\partial\beta_j}\frac{\partial f}{\partial\beta_p}\right).$$

The final result is:

$$\boxed{\mathcal{M}_{si} = \frac{n_i-1}{2}\frac{\partial f}{\partial\beta_j}\frac{\partial f}{\partial\beta_p}.}$$