Neural Computing Research Group
Aston University
Birmingham B4 7ET
United Kingdom
Tel: +44 (0)121 333 4631
Fax: +44 (0)121 333 4586
http://www.ncrg.aston.ac.uk/

Approximately Optimal Experimental Design for Heteroscedastic Gaussian Process Models

Alexis Boukouvalas, Dan Cornford, Milan Stehlík

Unassigned Technical Report November 10, 2009

Abstract

This paper presents a greedy Bayesian experimental design criterion for heteroscedastic Gaussian process models. The criterion is based on the Fisher information and is optimal in the sense of minimizing parameter uncertainty for likelihood based estimators. We demonstrate the validity of the criterion under different noise regimes and present experimental results from a rabies simulator to demonstrate the effectiveness of the resulting approximately optimal designs.

1 Introduction

In this paper we address optimal experimental design for Gaussian Processes (GPs) with independent heteroscedastic noise. The usual assumption in experimental design and modelling is that the noise is homoscedastic. Our focus is to produce designs which minimise model parameter uncertainty, rather than predictive uncertainty (Krause et al., 2008). Zhu and Stein (2005) present an approach to experimental design for parameter estimation for the homoscedastic GP case and we extend this approach to the heteroscedastic case. In so doing we introduce a new heteroscedastic model, which simplifies previously proposed models, making the optimal experimental design problem more tractable.

Our motivation stems from the field of computer experiments and in particular how to build good statistical approximations, known as emulators, to random output simulators. Traditionally the simulators examined in the literature have been deterministic (Kennedy and O'Hagan, 2001) but computer models with a stochastic response are becoming more common in many applications, from systems biology to social modelling to climate prediction. Experimental design plays a crucial role in the building of an emulator (Sacks et al., 1989), and unlike data driven learning 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 input points at which to the simulator is run given the available *apriori* knowledge. The heteroscedastic GP emulator is then trained on the selected design set and corresponding simulator evaluations.

The paper opens with a review of the experimental design for parameter estimation in Section 2 followed by a discussion of the new heteroscedastic GP model in Section 3. The approach to experimental design is described in Section 4 followed by experimental results on synthetic data in Section 5. The new methods are applied to a random output rabies simulator in Section 6. Conclusions are given in Section 7.

2 Fisher information

In this paper we calculate experimental designs that minimize the parameter uncertainty. We accomplish this by minimizing the log determinant of the Fisher Information Matrix (FIM), a $p \times p$ symmetric matrix, where p is the number of unknown parameters θ. The FIM is defined below:

$$\mathbf{F} = \int \left(\frac{\partial^2}{\partial \theta^2} ln(f(X|\theta) \right) f(X|\theta) d\theta,$$

where $f(X|\theta)$ is the likelihood function.

In the case of multivariate normal distributions it can be computed analytically. Let $\mathbf{X}$ distributed as $N(\mu(\theta), \mathbf{\Sigma}(\theta))$, the i, j element of the FIM is:

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

where $(\cdot)^T$ denotes the transpose and tr the trace. The focus in this paper is on identifying covariance function parameters and we will assume the trend parameters are known or of no interest, thus only the second term in (1) is relevant.

In the machine learning area, the Fisher information has been used for active learning (Hoi et al., 2009) where a submodular function was found to be a good approximation to the FIM in the case of classification assuming small length scales. The submodularity allows for robust usage of a greedy optimization algorithm, guaranteed to be within a constant factor of the optimum.

We consider regularity assumptions on a covariance structure $K_r(d)$ with unknown parameter r and d the distance of two design points. Introduced in Stehlík (2009) and referred to as the ABC class, we assume for covariance K that

a) $K_r(d) \geq 0$ for all r and $0 < d < +\infty$,

b) for all r mapping $d \to K_r(d)$ is semicontinuous, non increasing on $(0, +\infty)$

c) $\lim_{d\to+\infty} K_r(d) = 0$.

Under these conditions, FIM related optimal designs are reasonably well behaved. In particular both infill domain asymptotics and increasing domain asymptotics are feasible. From the probabilistic point of view we can see the ABC class as a class of Gaussian processes which are semicontinuous extensions of the Ornstein Uhlenbeck process. The assumptions ABC are fulfilled by many covariance functions, e.g. by the power exponential and Matérn class.

The determination of optimal designs for models with a correlated errors is substantially more difficult and for this reason not so well developed. For the influential papers there is a pioneering work of Hoel (1958), who considered the weighted least square estimate, but considered mainly equidistant designs. Bickel and Herzberg (1979) considered least squares estimation and determined asymptotic optimal designs. Müller and Pázman (2003) determine an algorithm to approximate optimal designs for linear regression with correlated errors and introduced the instrument called virtual noise.

Theoretical justifications for using the Fisher information for D-optimal designing under correlation can be found in Abt and Welch (1998) where asymptotic analysis shows that in the limit the inverse of the information matrix coincides with the covariance of the limiting distribution of the maximum likelihood estimator. Pázman (2007) provides justification of the FIM for a small noise levels. An experimental justification for the use of the FIM under homoscedastic noise was given in Zhu and Stein (2005) where simulations from Matérn covariance function based GPs were used to study whether the inverse Fisher information matrix is a reasonable approximation to the empirical covariance matrix of maximum likelihood estimators, as well as a reasonable design criterion.

3 Heteroscedastic GP Models

One approach to modelling heteroscedastic noise within a GP framework is to use a system of coupled GPs modelling the mean and variance functions respectively. In Goldberg et al. (1998) a Monte Carlo approach was utilized to incorporate the uncertainty of the variance GP into the overall predictive uncertainty. The computational expense of this method however motivated an approximation whereby only the most likely value of the variance is utilized and the associated uncertainty around this estimate is discarded (Kersting et al., 2007).

Snelson and Ghahramani (2006) proposed a heteroscedastic version of the sparse pseudo-input GP method (hereafter SPGP+HS). However as was noted in Snelson and Ghahramani (2006) this method does not perform well when small numbers of observations are available due to the flexibility of the model (Snelson and Ghahramani, 2006). Large training set sizes are uncommon in the emulation context where simulator runs are typically expensive to obtain – where the simulator is very cheap, its direct use might be preferred.

The model we develop in this paper is similar to the SPGP+HS model but allows different mean and variance response structures. The log variance function is modelled as a linear in parameters regression using a set of fixed basis functions $\mathbf{h}(\mathbf{x})$. The heteroscedastic GP prior is thus:

$$p(\mathbf{t}|\theta, \mathbf{x}) = N[0, K_\mu + diag(\exp(\mathbf{h}(\mathbf{x})^T\beta))P^{-1}],$$

where *diag* denotes the diagonal matrix of the input vector, K_μ is the usual covariance matrix which depends on parameters θ_μ representing process variance and length scales, β the linear coefficients and P a diagonal matrix containing the number of replicated observations at each training point site. In this paper P is always set to the identity and $\mathbf{x}$ is the training data input matrix. The set of free parameters for this model is $\theta = \{\theta_\mu, \beta\}$.

We considered two types of basis functions, local (radial basis functions) and global (polynomial) to provide the input dependent nugget term. An advantage of local basis functions is the interpretability of priors on the β coefficients. The number of local basis functions required for domain coverage grows exponentially with the input dimension.

In high dimensional cases global basis functions may be more appropriate or a non-parametric method could be considered using an additional 'variance kernel': $p(\mathbf{t}|\theta, \mathbf{x}) = N[0, K_\mu + diag(\exp(k_\Sigma^T(K_\Sigma + \sigma_n^2)^{-1}\beta))P^{-1}]$ where K_Σ and k_Σ are the variance kernel functions, depending on parameters θ_Σ, and in this case β is a variance 'pseudo observation' vector, and σ_n^2 a nugget term. Note that sparse approaches to this parameterisation, similar to Snelson and Ghahramani (2006), are likely to be more computationally attractive. The main difference of this model from the model of Snelson and Ghahramani (2006) is that we do not entangle the mean and variance response, allowing separate kernels for each. This will be important where the complexity of the mean and variance response is different. This model also bears resemblance to the Kersting et al. (2007) model, however here we directly represent the log variance function as a non-parametric kernel regression rather than employing a Gaussian process model and then using the most likely value. This enables us to write down a simpler model, with the same flexibility as Kersting et al. (2007), for which we can evaluate the FIM.

4 Sequential Search Bayesian Design

The calculation of the FIM (Section 2) is defined for a given parameter value vector, θ_0. If a point estimate for θ is used the design is termed locally optimal, in the sense that we obtain an optimal design for that specific parameter value θ_0. In practice θ will not be known in advance so we follow the approach of Zhu and Stein (2005) using the approximate Bayesian criterion:

$$U(\mathbf{s}) = -\int \ln |\mathbf{F}(\mathbf{s}, \theta)| \; p(\theta) \; d\theta \tag{2}$$

where $p(\theta)$ the prior on the parameters, $\mathbf{s}$ the proposed design and $|\mathbf{F}(\mathbf{s}, \theta)|$ the determinant of the FIM given by (1).

The integral in (2) can be approximated using Monte Carlo:

$$U(\mathbf{s}) \approx M(\mathbf{s}) = -\frac{1}{N} \sum_{i=1}^{N} \ln |\mathbf{F}(\mathbf{s}, \theta_i)|$$

for N samples from the prior $p(\theta)$.

To complete the specification of the experimental design algorithm the method of optimization must be defined. The most commonly employed approach is to provide a large candidate design set, and select a subset of design points from this set. 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 (Zhu and Stein, 2005) 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 (Xia et al., 2006).

Another option is greedy optimization where the candidate point which maximizes the score gain at each step is included in the selected set. In Xia et al. (2006) the greedy approach is shown to be superior to simple stochastic optimization schemes. We confirm this result, providing further experimental results supporting the effectiveness of the greedy approach in Section 5.2.

One challenge with the sequential greedy optimization method is initialisation. It is necessary to have at least two points to compute the Fisher score (2), with more providing better numerical stability. A potentially useful initialisation is to evaluate the Fisher score for all point pairs. The approach utilized in the experiments is to pick a set of N points closest to a space filling equal grid for the design space. This compromise appears to have little effect on the final designs found as shown in Section 5.2.

5 Synthetic Experimental Results

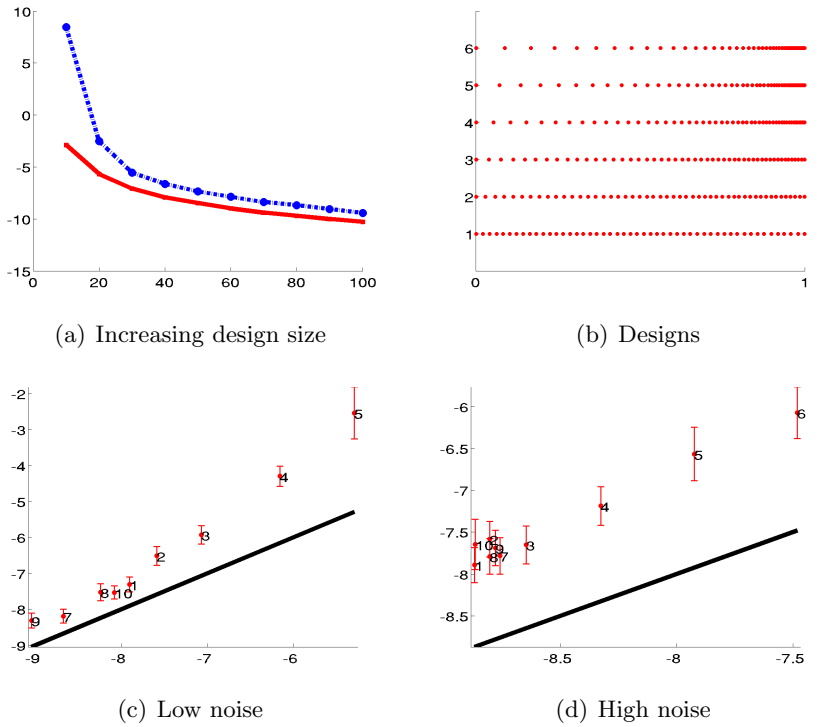
The experiments[1] on synthetic data aim to investigate the utility of the Fisher information for experimental design purposes (Section 5.1) and demonstrate the effectiveness of the greedy

[1]The code will be available online and extends the gpml library of Rasmussen and Williams (2006).

optimization method (Section 5.2). Since a local design is rarely justifiable, we present in Section 5.3 experimental results on Bayesian design.

5.1 FIM for Design

In this section we show that under different signal-to-noise ratios the Fisher score remains monotonic to the empirical parameter covariance. The inverse of the FIM provides a lower bound to the empirical parameter covariance and the bound becomes tighter as the number of samples grows. In Figure 1(a) we show for different sample sizes the approximation error.



(a) Increasing design size

(b) Designs

(c) Low noise

(d) High noise

Figure 1: Relation of the log determinant of the Fisher information to the log determinant of the empirical parameter covariance. (a) The FIM (solid) and empirical parameter covariance (dashed) for designs of size 10 to 100. (b) The non-random designs used. (c),(d) The approximation for 50 point designs under different noise levels. (c) is using a linear basis variance model and (d) an RBF variance model with two Gaussian basis functions centred equidistantly. Designs 7-9 are random and 10 is a uniform Latin hypercube.

We use the Matérn covariance function with fixed differentiability $\nu = 5/2$, length scale λ and process variance σ_p^2 and a linear model for the log variance. The empirical parameter covariance is computed by sampling from a GP with a polynomial heteroscedastic noise model. A maximum a posteriori estimate (MAP) for the GP parameters is calculated for each GP sample. The parameters of the generative GP were set to $\lambda = 0.5$, $\sigma_p^2 = 0.75$ and the linear coefficients to $\beta_0 = 0.01$ and $\beta_1 = -30$ which correspond to a high noise level in the initial part of the design space quickly reducing to low noise. The empirical parameter covariance was calculated using MAP parameter estimates from 1000 realizations of the generative GP.

The next experiment demonstrates the monotonicity of the Fisher information to the empirical parameter covariance. We generate six designs of 50 points with the distance between neighbouring points determined by the quantiles of exponential distributions with different rate parameters (Figure 1(b)). In addition three random and a Latin hypercube design were also used.

In the low noise case, a linear basis variance model was used with the parameters of the GP set to the same levels as in the previous experiment. For the high noise case a two Gaussian basis RBF model was used. The basis functions were positioned equidistantly in the design space with their variance set to the squared distance between their centres. The parameters were set to $\lambda = 0.33$, $\sigma_p^2 = 1.8$ and $\beta_0 = -3.7$, $\beta_1 = -0.8$. Finally, we calculate confidence intervals for our estimates of the log determinant of the empirical parameter covariance using 1000 bootstrap samples (see Appendix). The results are shown in Figure 1(c)-(d) where we observe that for the higher noise level case the approximation error is larger but the monotonicity still holds.

We repeat this experiment on larger designs and varying signal-to-noise ratios. We use designs of 100 points where we sample from a GP with different levels of heteroscedastic noise. Two Gaussian basis functions were used with their centres and widths set as before. Samples from the GP for the different noise scenarios are shown in Figures 2(d)-(f). The length and process variance of the Matérn covariance were unchanged. The linear coefficients for the variance model were set to $\beta_0 = -4.7$, $\beta_1 = -2.8$ for the low noise case, $\beta_0 = -3.7$, $\beta_1 = -0.8$ for the the medium noise case and $\beta_0 = -2.7$, $\beta_1 = 1.2$ for the high noise case. We see in Figures 2(a)-(c) that although the approximation of the FIM to the parameter variance gets progressively worse as the noise level increases, the monotonicity holds even for relatively high noise levels.

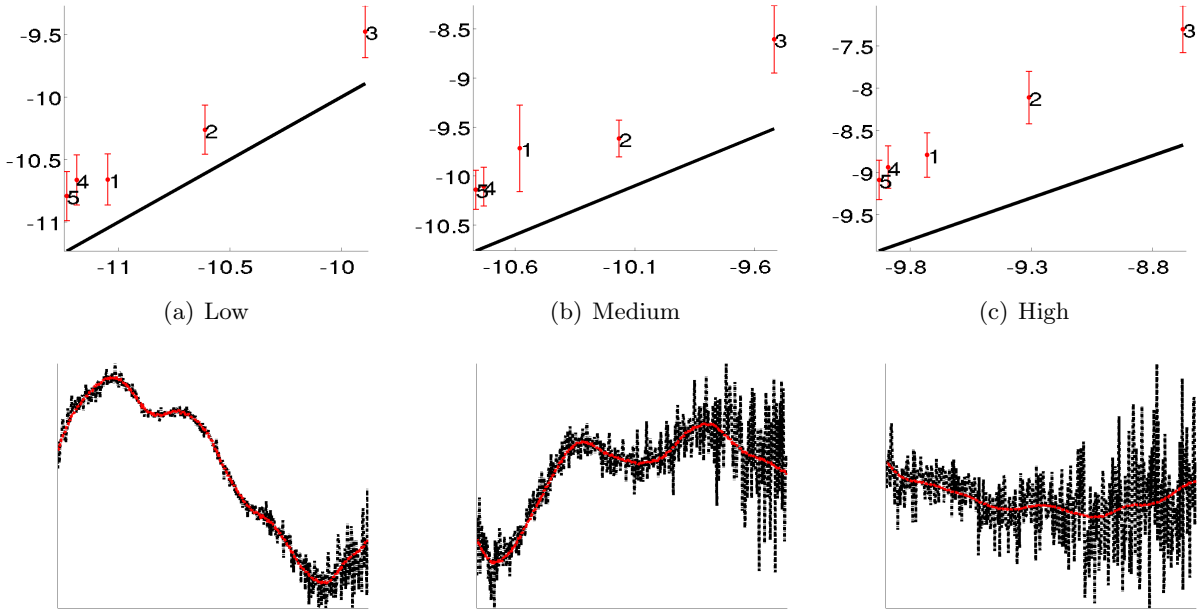
The monotone relationship between the log determinant of the FIM and the log determinant of the empirical parameter covariance holds in all scenarios tested and affirms the usage of the FIM as a design criterion for minimizing parameter uncertainty. This conclusion agrees with the findings of Zhu and Stein (2005) which showed this relationship in the homoscedastic case.

5.2 Greedy Optimization

The experiment considers the selection of 9 locations from a candidate set of 29 points in a locally optimal design. The design is given the point parameter prior $\theta_0 = (\lambda = 0.5, \sigma_p^2 = 0.7, \beta_0 = 0.1, \beta_1 = -10)$ and we compute the FIM score of all $\binom{29}{9}$ combinations. The Matérn covariance and a linear basis function for the log variance is used. We also show the Fisher scores for the solution obtained using greedy optimisation and an approximate grid design selected from the candidate set (Figure 3(b)).

In terms of Fisher score, the greedy solution is very close to the optimum while the score for the grid design is significantly worse. Additionally, even for this simple example we notice a very large number of local optima close to the optimum demonstrating the near equivalence of a large number of designs.

The optimal, greedy and grid designs are shown in Figure 3(a) along side the candidate set. The relatively long length scale of the GP means the noise signal dominates and the optimal designs place the points near the boundaries due to the log linear form of the variance function.



(a) Low (b) Medium (c) High

(d) Low (e) Medium (f) High

Figure 2: Effect of noise on the monotonicity of the FIM vs parameter uncertainty. Designs 1-3 increasing distance designs (Figure 1(b)), 4 a Latin design and 5 is random (a, b, c). Illustrative GP realisations for the various noise levels (d, e, f).

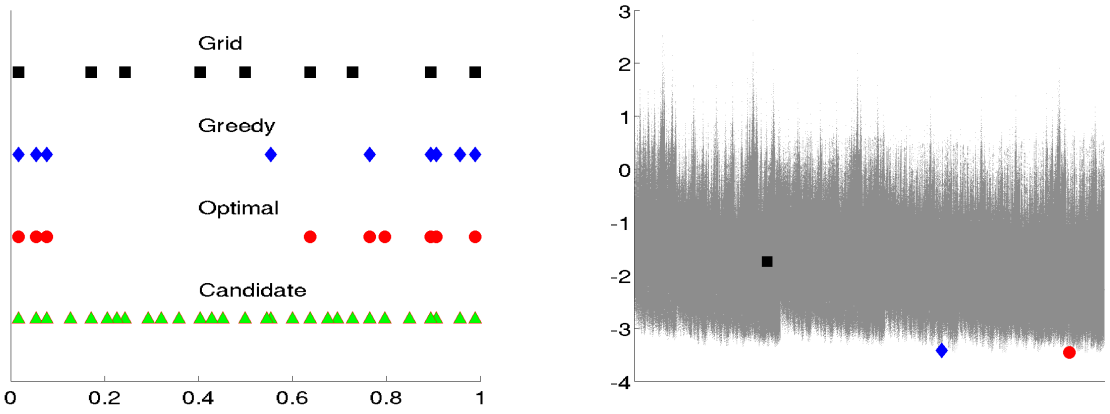
Since the motivation of using the Fisher information as a design criterion is to minimise parameter uncertainty, we expect the likelihood for the optimal designs be more informative about the optimum θ than the grid design. We demonstrate this effect by plotting the profile likelihood for each parameter (Figure 4) using a single GP sample as our training data. For all four parameters using only nine training points, the likelihood on the optimal design excludes with more certainty larger portions of the parameter domain than the grid design.

5.3 Bayesian Optimum Designs

The previous sections have demonstrated the effectiveness of the Fisher information criterion applied to the fixed basis heteroscedastic variance model. However local designs require a point estimate of the parameters which in practice is an uncertain quantity. Following the discussion in Section 4, we present experimental results demonstrating the implementation of Bayesian D-optimal design which removes this need by allowing specification of a prior belief on the distribution of parameters. The computational cost however is increased due to the intractability of integral (2) necessitating the usage of Monte Carlo.

The validation metric we use to compare the bayesian design based on a vague prior to the grid design is the relative root mean square error (rRMSE) of the parameter estimate θ_i to the true value θ_0 averaged over N samples (Zhu and Stein, 2005), i.e. $1/N \sum_{i=1}^{N} \sqrt{\theta_i - \theta_0)^2}/\theta_0$. The scaling of the error by θ_0 ensures the rRMSEs are comparable.

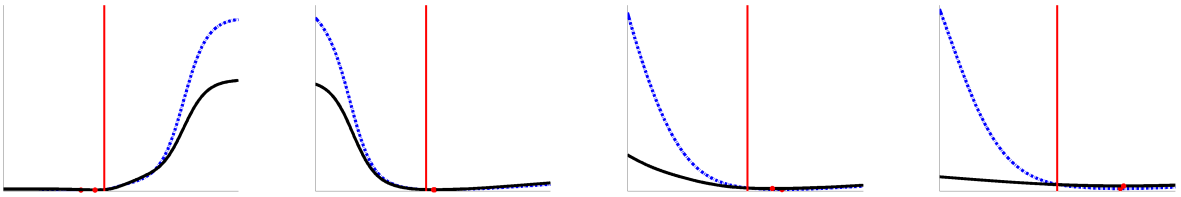
We simulate from a heteroscedastic GP with Matérn covariance and a two Gaussian basis vari-



(a) Designs

(b) FIM solutions

Figure 3: Complete enumeration for a locally optimal design. Candidate Set (green triangle), optimal design (red circle), greedy design (blue diamond) and grid design (black square).



(a) log λ

(b) log σ_p^2

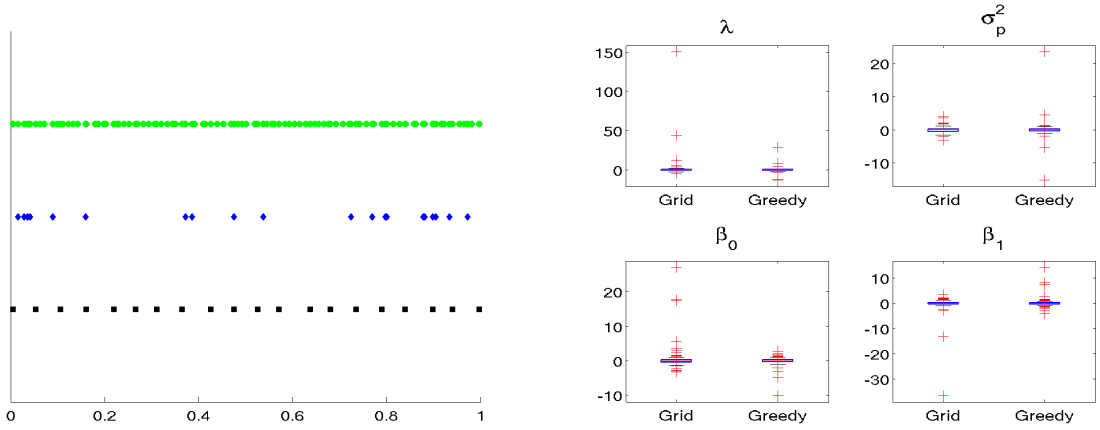
(c) β_0

(d) β_1

Figure 4: Profile likelihoods for locally optimal design (dashed blue) and a grid design (solid black). The true parameter value is also shown (vertical red line).

ance model. We place zero mean, variance 20 log normal priors on λ and σ_p^2 and normal priors $N(-2, 40)$ on the linear variance coefficients β_0, β_1. We sample from the prior 100 times and for each parameter sample we simulate from the GP 50 times providing a total of 5000 realisations of the experiments. The resulting designs and corresponding rRMSE values are shown in Figure 5.

The mean and median rRMSE for each parameter is given in Table 1 for both the greedy optimum and grid designs. We note a significant improvement for all parameters. In Zhu and Stein (2005) a similar experiment was conducted by simulating from a homoscedastic GP and a benefit in terms of average rRMSE was noted. In our case however it turns out that the average rRMSE is dominated by a few extreme values in the parameter estimation. When looking at the median rRMSE in Table 1 the metric is better for the grid design. We have repeated this experiment with different configurations with similar results. It appears the FIM based designs are more robust but further evidence is needed since these results are based on a few extreme values despite the large number of simulations.



(a) Designs (b) rRMSE

Figure 5: Bayesian Designs for 1D synthetic example: Candidate set (green circle), Greedy (blue diamond) and Grid (black square).Also shownm box whisker plot of the rRMSE values for each parameter.

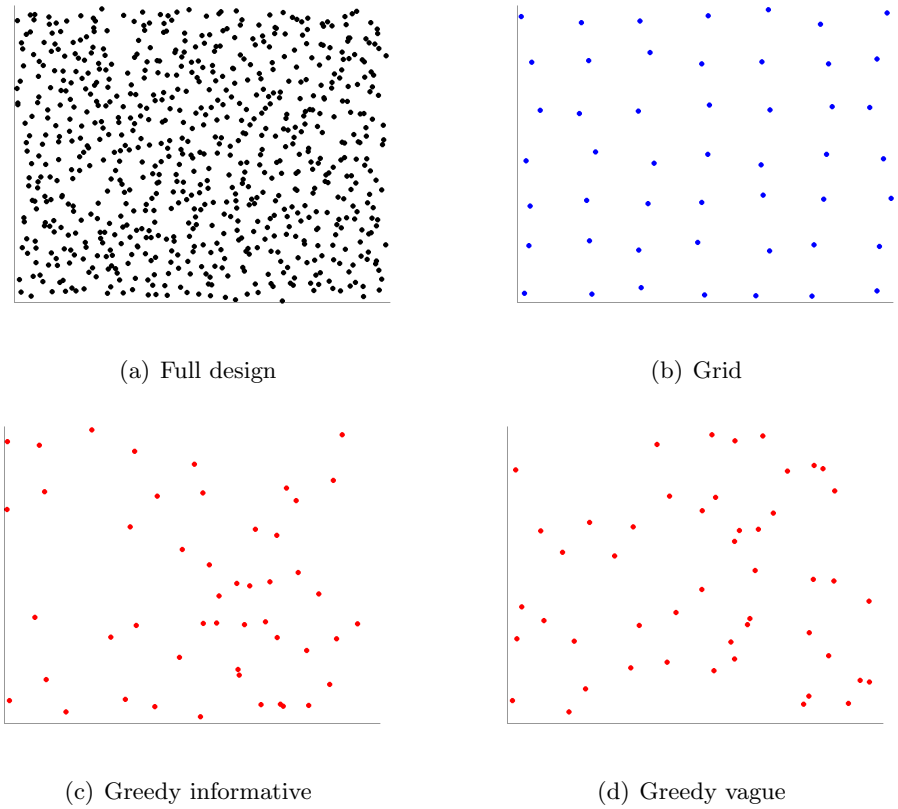
Table 1: rRMSE on Bayesian simulation for Grid and Greedy designs.

Statistic	Design	λ	σ_p^2	β_0	β_1
Mean	Grid	2.10	-0.07	0.58	-0.38
	Greedy	0.25	0.04	-0.11	0.14
Median	Grid	-0.09	0.01	-0.09	-0.09
	Greedy	0.15	0.03	-0.09	-0.10

6 Stochastic Rabies Model

Our motivating example is a stochastic simulation model developed for the analysis of the risk and strength of rabies spread in a community of raccoon dogs and foxes (Singer et al., 2008). We emulate a single output of the model, the number of time steps required for the disease to become extinct in the raccoon dog population. This output is important in deciding on the response to a potential rabies outbreak. We note this output has a rather complex, non-Gaussian, distribution; in this paper we emulate the log extinction time, which is more approximately Gaussian, as determined from visual inspection of Q-Q plots. The model normally has 14 inputs but we have fixed all but the two most relevant inputs to their nominal values to permit easy visualisation. The raccoon dog winter density and death rate parameters were identified through sensitivity analysis and discussion with the domain expert as the most relevant inputs.

The candidate set is a Latin hypercube of 961 points (Figure 6(a)) from which 49 points are selected. We used the greedy Bayesian design approach (Section 4) which allows the specification of a prior over the parameters. The computational cost is increased due to the intractability of integral (2) necessitating the usage of Monte Carlo.



(a) Full design

(b) Grid

(c) Greedy informative

(d) Greedy vague

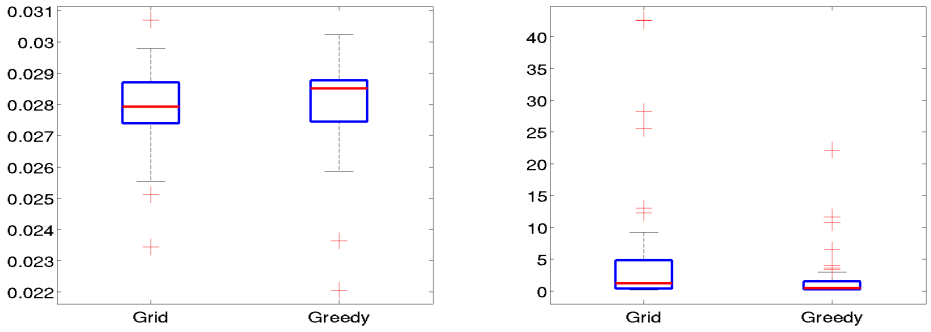
Figure 6: Resulting designs for rabies model.

Two experiments were conducted using different priors to highlight the effect of the prior on the Bayesian design. The same priors are also used during inference producing MAP estimates. Our informative prior is a zero mean GP with a Matérn isotropic covariance and fixed basis variance models consisting of 16 Gaussian basis functions. The basis functions were centred on a square grid and the width set to the squared distance between grid points. We place a Gaussian prior on the log variance linear coefficients β. The time to disease extinction is known to be correlated with the density factor and anti-correlated with the death rate (for higher death rates more individuals die before they transmit the disease). For low densities and high death rates the disease becomes extinct quickly with high certainty. We therefore set the mean of the β to low values (-11) for high death rates and low densities and a higher mean (-1) for high density, low death rate areas of the input space and place intermediate values between these extremes. To prevent the prior dominating we set the variances for the β to 20. The priors for λ and σ_p^2 are Gaussian with mean -1.6 and variance 4 in the log space, corresponding to a 95% coverage in the data space of $[0.0023, 17.5087]$. The purpose of this prior is to ensure numerical stability by excluding very small, unrealistic values.

The vague prior uses essentially the same settings but the mean for all coefficients β is set to -5 and the variance increased to 25.

The greedy algorithm was initialised with a 4 point grid to ensure numerical stability. The resulting designs using the informative and vague priors are shown in Figures 6(c) and (d) respectively. A grid design is also plotted for comparison.

To validate the designs, we maximise the parameter posterior using 50 realisations of the simulator for all designs and compute the root mean squared error (RMSE) of the mean prediction and the mean squared error (MSE) of the predictive variance (Figure 6). The MSE variance is defined as $MSEVar = 1/N \sum_{i=1}^{N} (\text{Var}[t_i] - \hat{\sigma}^2(\mathbf{x}_i))^2 / \text{Var}[\hat{\sigma}^2(\mathbf{x}_i)]$ where N is the number of design points, $\text{Var}[t_i]$ the GP predictive variance at input point x_i, $\hat{\sigma}^2(\mathbf{x}_i)$ the at-a-point sample variance calculated from 100 realisations of the simulator and $\text{Var}[\hat{\sigma}^2(\mathbf{x}_i)]$ the normalisation by the variance of the sample variance. The mean and median values of all validation measures are given in Table 2.



(a) RMSE

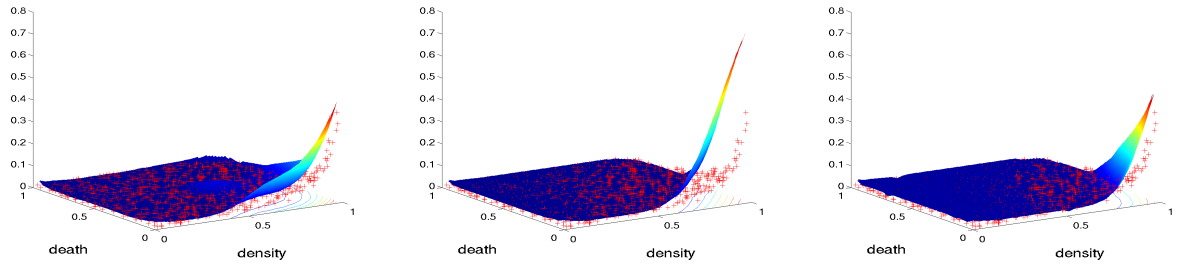
(b) MSE Variance

Figure 7: Distributions of validation measures when using the informative prior.

Comparing the grid and greedy Bayesian designs the predictions of the mean exhibit similar error as seen in terms of the RMSE (Table 2) but the greedy Bayesian design allows more accurate estimates of the variance as the MSE on the variance reveals. This is illustrated in Figure 8 where the predictive standard deviation using the grid and greedy designs with the informative prior are plotted against the sample variance calculated using 100 realisations of the simulator. The fit is better with the informative prior (Table 2).

Table 2: Design validation for the rabies model. μ_r and M_r the mean and median RMSE respectively, μ_σ and M_σ the mean and median MSE on the variance.

Prior	Design	μ_r	M_r	μ_σ	M_σ
Informative	Grid	0.028	0.027	5.16	1.15
	Greedy	0.028	0.028	1.84	0.44
Vague	Grid	0.030	0.030	2.79	0.86
	Greed	0.029	0.030	1.85	0.54



(a) Full Design (b) Grid (c) Greedy

Figure 8: Illustration of the fit of the GP standard deviation (surface) to the empirical simulator standard deviation (points).

7 Conclusions

We have presented the use of the Fisher information as an effective design criterion in the case of heteroscedastic GP noise models. Results on synthetic data have demonstrated the monotonic relationship of the FIM to the empirical parameter covariance required for its use in design. The complete enumeration of all designs in Section 5.2 demonstrated the effectiveness of the greedy algorithm in finding near optimal designs and the effect of these designs on the likelihood profiles. The application of the Bayesian design approach to the rabies model resulted in the GP capturing the simulator variance more accurately than a standard grid design. The experimental results suggest that approximately optimal Bayesian FIM designs obtained using the greedy algorithm allow for more robust parameter estimation of covariance parameters and can subsequently lead to better predictions. The approach presented in this paper can be extended to the case of replicated observations. The main difficulty is the combinatorial explosion of possible designs when replication is allowed, warranting the investigation of more efficient optimization schemes.

Appendix: The bootstrap method

We use a method suggested in Efron and Tibshirani (1993) to determine the number of bootstrap samples required to estimate the standard error in Section 5.1. As usual, bootstrap is done by random sampling with replacement.

In particular we first estimate the bias $E_{bootstrap} - E_{data}$, where $E_{bootstrap}$ the mean value across all bootstrap samples and E_{data} the estimated value from data. If the bias / standard error ratio is less then 0.25, we judge we have enough samples in our bootstrap.

Acknowledgements

We wish to thank Alexander Singer for providing the Rabies model used in Section 6. This research is funded as part of the Managing Uncertainty in Complex Models project by EPSRC grant D048893/1 and by Acciones Integradas 2008-09, Project Nr. ES 18/2008.

References

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

Bickel, P. J. and A. M. Herzberg (1979). Robustness of design against autocorrelation in time i: Asymptotic theory, optimality for location and linear regression. *The Annals of Statistics 7*(1), 77–95.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Goldberg, P. W., C. K. I. Williams, and C. M. Bishop (1998). Regression with input-dependent noise: A Gaussian process treatment. In M. I. Jordan, M. J. Kearns, and S. A. Solla (Eds.), *Advances in Neural Information Processing Systems*, Volume 10. The MIT Press.

Hoel, P. G. (1958). Efficiency problems in polynomial estimation. *The Annals of Mathematical Statistics 29*(4), 1134–1145.

Hoi, S. C. H., R. Jin, and M. R. Lyu (2009). Batch mode active learning with applications to text categorization and image retrieval. *IEEE Transactions on Knowledge and Data Engineering 99*(1).

Kennedy, M. and A. O'Hagan (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society B63*, 425–464.

Kersting, K., C. Plagemann, P. Pfaff, and W. Burgard (2007). Most likely heteroscedastic Gaussian process regression. In Z. Ghahramani (Ed.), *Proc. 24th International Conf. on Machine Learning*, pp. 393–400. Omnipress.

Krause, A., A. Singh, and C. Guestrin (2008). Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res. 9*, 235–284.

Müller, W. G. and A. Pázman (2003). Measures for designs in experiments with correlated errors. *Biometrika 90*(2), 423–434.

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

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Sacks, J., W. Welch, T. Mitchell, and H. Wynn (1989). Design and analysis of computer experiments. *Statistical Science 4*, 409–435.

Singer, A., F. Kauhala, K. Holmala, and G. Smith (2008). Rabies risk in raccoon dogs and foxes. *Biologicals, In press.*.

Snelson, E. and Z. Ghahramani (2006). Variable noise and dimensionality reduction for sparse Gaussian processes. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Arlington, Virginia. AUAI Press.

Stehlík, M. (2009). Topological regularities for regression with correlated errors. In *6th St.Petersburg Workshop on Simulation*, pp. 377–381.

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

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