UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Mathematical Sciences

Bayesian design for calibration of physical models

by

Yiolanda Englezou

Thesis submitted for the degree of Doctor of Philosophy

July 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Mathematical Sciences

Doctor of Philosophy

BAYESIAN DESIGN FOR CALIBRATION OF PHYSICAL MODELS

by Yiolanda Englezou

We often want to learn about physical processes that are described by complex nonlinear mathematical models implemented as computer simulators. To use a simulator to make predictions about the real physical process, it is necessary to first perform calibration; that is, to use data obtained from a physical experiment to make inference about unknown parameters whilst acknowledging discrepancies between the simulator and reality. The computational expense of many simulators makes calibration challenging. Thus, usually in calibration, we use a computationally cheaper approximation to the simulator, often referred to as an emulator, constructed by fitting a statistical model to the results of a relatively small computer experiment. Although there is a substantial literature on the choice of the design of the computer experiment, the problem of designing the physical experiment in calibration is much less well-studied. This thesis is concerned with methodology for Bayesian optimal designs for the physical experiment when the aim is estimation of the unknown parameters in the simulator.

Optimal Bayesian design for most realistic statistical models, including those incorporating expensive computer simulators, is complicated by the need to numerically approximate an analytically intractable expected utility; for example, the expected gain in Shannon information from the prior to posterior distribution. The standard approximation method is "double-loop" Monte Carlo integration using nested sampling from the prior distribution. Although this method is easy to implement, it produces biased approximations and is computationally expensive. For the Shannon information gain utility, we propose new approximation methods which combine features of importance sampling and Laplace approximations.

These approximations are then used within an optimisation algorithm to find optimal designs for three problems: (i) estimation of the parameters in a nonlinear regression model; (ii) parameter estimation for a misspecified regression model subject to discrepancy; and (iii) estimation of the calibration parameters for a computational expensive simulator. Through examples, we demonstrate the advantages of this combination of methodology over existing methods.

Contents

List of Figures

List of Tables

List of Algorithms

Declaration of Authorship

I, Yiolanda Englezou, declare that the thesis entitled “Bayesian design for calibration of physical models” and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed:

Date:

Acknowledgements

First and foremost I want to thank my supervisors Professor Dave Woods and Dr Tim Waite for all their contributions of time and ideas, their advice and assistance. They are the reason I have gained a variety of specialist research skills and they have strengthened my confidence and aptitude to tackle new research problems. I am also thankful for the excellent example they have provided as successful academics.

I thankfully acknowledge the Engineering and Physical Sciences Research Council and the Atomic Weapons Establishment for the funding opportunities that made my PhD work possible. I am also especially grateful to all the members of the Design Group at the University of Southampton who have been a source of good advice.

Last, a big thank you to my family for all their love and support. My parents Panikos and Charalambia who supported me in all my pursuits. My sister Christina for her presence in the UK throughout my time as a PhD student. For his consistent motivation, love and patience I want to thank Erricos Michaelides. I also feel the need to thank my friends, those who were in the UK for their daily support, and of course, those who have been away but still gave their help and encouragement. A special thank you to Andreas Papallas and Maria Adamou.

Without these people the completion of this PhD work and the furthering of my academic and professional career would not be possible. Thank you!

Chapter 1

Introduction

Engineers and scientists increasingly use deterministic computer models, referred to here as *simulators*, to study actual or theoretical physical processes that would otherwise be very difficult to analyse. There are many examples of scientific and technological developments that use simulators to reduce or replace costly or infeasible physical experimentation. Two examples are:

- When designing an aircraft wing, computational fluid dynamics models are used in order to calculate the air flow over a wing (Forrester, 2010).

- In drug development, molecular modelling is an important part of exploring, describing and predicting properties of potential drug candidates (Norrisa et al., 2000).

A simulator is often an implementation of a complex mathematical model that maps several input variables to a (possibly multivariate) output. The resulting computer code is typically expensive in terms of computer time to run. Hence, only small number of runs can be performed at particular combinations of values of the input variables. Sacks et al. (1989) proposed the construction of an *emulator* or a surrogate model, specifically a Gaussian process model, which approximates the simulator but is much faster to run. This approach is now commonly used to predict the output of the simulator at untried input combinations. It is often described as a 'black box' method, meaning that it makes no use of information about the mathematical model, except knowledge of the outputs for the simulator runs that have been performed. In this thesis we focus on simulators of a process for which some limited physical experimentation is also possible.

To use a simulator to make predictions, it may be necessary to first perform *calibration*, that is, to use physical data to estimate the values of any unknown simulator parameters, whilst also acknowledging possible discrepancies between the simulator and reality (Kennedy and O'Hagan, 2001). In order to combine simulator evaluations with the physical data, the simulator is considered as a biased version of the true mean response (Brynjarsdóttir and O'Hagan, 2014).

The computational expense of the simulator makes calibration challenging. The simulator output is only known for the few combinations of input values that have been run, and for other input combinations we are uncertain about the value of the simulator output. We are also uncertain about the form of the (unobservable) model discrepancy. In their seminal paper, Kennedy and O'Hagan (2001) proposed that both these sources of uncertainty may be represented by independent Gaussian processes.

The Kennedy-O'Hagan approach has received considerable attention in the literature (e.g. Higdon et al., 2004; Reese et al., 2004; Bayarri et al., 2007b; Bayarri et al., 2007a; Gramacy and Lee, 2008; Wang et al., 2009; Wilkinson, 2010; Gramacy et al., 2015; Storlie et al., 2015; Arendt et al., 2016). It has been used in a variety of applications including hydrology, radiological protection, cylinder implosion and climate prediction (see Williams et al., 2006; Murphy et al., 2007; Higdon et al., 2008; Han et al., 2009; Goh et al., 2013; Leatherman et al., 2014). In this thesis, we also follow the Kennedy-O'Hagan approach.

1.1 A statistical model for calibration

In the calibration problem we have two groups of inputs to the simulator; the controllable variables, or inputs, $\mathbf{x} = (x_1, \ldots, x_{q_1})^{\mathrm{T}}$, $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{q_1}$ and the calibration parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p_\theta})^{\mathrm{T}}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The controllable variables, $\mathbf{x}$, can be controlled and set both when observing the physical process and when running the simulator. The calibration parameters, $\boldsymbol{\theta}$, can only be set when running the simulator, whose output is denoted as $\eta(\mathbf{x}, \boldsymbol{\theta})$. The calibration parameters are assumed to take fixed but unknown values $\boldsymbol{\theta}^p \in \boldsymbol{\Theta}$ for all physical observations (Higdon et al., 2008). The aim of calibration is to learn $\boldsymbol{\theta}^p$ to describe the physical observations.

For the physical experiment, let y_i denote the response from the ith run made under settings $\mathbf{x}_i$ of the controllable variables ($i = 1, \ldots, n$). We consider the following statistical model:

$$y_i = \zeta(\mathbf{x}_i) + \varepsilon_i = \rho\, \eta(\mathbf{x}_i, \boldsymbol{\theta}^p) + \delta_{\boldsymbol{\theta}^p}(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1.1}$$

Above, $\rho \in \mathbb{R}$ is an unknown regression parameter, and the discrepancy function, $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$, encodes the difference between the simulator evaluated at the 'true' $\boldsymbol{\theta}^p$, $\eta(\mathbf{x}, \boldsymbol{\theta}^p)$, and the mean, $\zeta(\mathbf{x})$, of the physical process. We assume $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ are independent. Usually $\delta_{\boldsymbol{\theta}^p}(\cdot)$ is a nonzero function because the simulator is built under certain assumptions that might not be true in real life and hence the physical observations might differ from the simulator output.

In addition to the physical data, we have a limited number of simulator runs from a computer experiment:

$$z_j = \eta(\mathbf{x}_j^c, \boldsymbol{\theta}_j^c), \;\; j = 1, \ldots, m, \tag{1.2}$$

where for the jth run of the computer experiment $\mathbf{x}_j^c$ denotes the vector of settings of the controllable variables and $\boldsymbol{\theta}_j^c$ denotes the vector of settings of the calibration parameters. In the Bayesian framework of Kennedy and O'Hagan (2001), calibration is performed using (1.1) and (1.2) by first placing appropriate prior distributions on $\boldsymbol{\theta}^p$, $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$, and $\eta(\mathbf{x}, \boldsymbol{\theta})$ (Gaussian process priors are typically used for $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$ and $\eta(\mathbf{x}, \boldsymbol{\theta})$; see Section 2.2). A posterior distribution for $\boldsymbol{\theta}^p$ is then formed by conditioning on $y_1, \ldots, y_n$ and $z_1, \ldots, z_m$. For more details see Chapter 6.

The Kennedy-O'Hagan framework has long been known to suffer from an identifiability problem, which we now explain. Underpinning (1.1) is the idea of writing the mean of the physical process, $\zeta(\mathbf{x})$, as

$$\zeta(\mathbf{x}) = \rho\, \eta(\mathbf{x}, \boldsymbol{\theta}^p) + \delta_{\boldsymbol{\theta}^p}(\mathbf{x}). \tag{1.3}$$

At first glance, given $\zeta(\mathbf{x})$, (1.3) appears to define both $\boldsymbol{\theta}^p$ and the discrepancy function $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$. However, in fact this is not the case. Given a different arbitrary choice of calibration parameters, $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$, we can find a corresponding function

$$\delta_{\boldsymbol{\theta}'}(\mathbf{x}) = \zeta(\mathbf{x}) - \rho\, \eta(\mathbf{x}, \boldsymbol{\theta}'),$$

that satisfies

$$\zeta(\mathbf{x}) = \rho\, \eta(\mathbf{x}, \boldsymbol{\theta}') + \delta_{\boldsymbol{\theta}'}(\mathbf{x}),$$

and so (1.3) does not define $\boldsymbol{\theta}^p$ and $\delta_{\boldsymbol{\theta}^p}(\cdot)$ uniquely. In other words, without further conditions, $\boldsymbol{\theta}^p$ and $\delta_{\boldsymbol{\theta}^p}(\cdot)$ are not identifiable from the physical process.

Bayarri et al. (2007b) argue that despite this identifiability problem the Kennedy-O'Hagan approach is still effective because the prior distributions on $\boldsymbol{\theta}^p$ and $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$ ensure that the posterior distributions are well-defined, and so predictions can still be made. Nonetheless, more recently several authors have continued to seek a resolution of the identifiability problem by more carefully considering how to define the 'true' values of the calibration parameters. We discuss some of these ideas below.

First, note that if the assumed simulator were true then there would be no discrepancy, $\delta_{\boldsymbol{\theta}^p}(\mathbf{x}) = 0$, and hence there would exist a 'true' $\boldsymbol{\theta}^p$ for which

$$\zeta(\mathbf{x}) = \eta(\mathbf{x}, \boldsymbol{\theta}^p), \tag{1.4}$$

for all $\mathbf{x} \in \mathscr{X}$. However, if the simulator is not correct, which is usually the case, Equation (1.4) cannot hold.

When $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$ is not equal to zero for all $\mathbf{x} \in \mathscr{X}$, a common approach to achieve identifiability of the calibration parameters is to redefine the 'true' parameter values, $\boldsymbol{\theta}^p$, as those that minimise the 'distance' between the mean of the physical process and the simulator output. For calibration, typically the L_2 norm has been used (Tuo and Wu, 2016), giving

$$\boldsymbol{\theta}^p = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_{\mathcal{X}} [\zeta(\mathbf{x}) - \eta(\mathbf{x},\boldsymbol{\theta})]^2 d\mathbf{x}. \tag{1.5}$$

For a deterministic physical process, Tuo and Wu (2016) defined the following estimator for these L_2-best calibration parameters:

$$\hat{\boldsymbol{\theta}}^p = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_{\mathcal{X}} [\hat{\zeta}(\mathbf{x}) - \eta(\mathbf{x},\boldsymbol{\theta})]^2 d\mathbf{x},$$

where $\hat{\zeta}(\mathbf{x})$ is the mean of a Gaussian process fitted to the physical data (Chapter 2). The estimator $\hat{\boldsymbol{\theta}}^p$ is consistent for $\boldsymbol{\theta}^p$ in the sense that as the number of physical runs n becomes larger and the design becomes more dense on $\mathcal{X}$, $\hat{\boldsymbol{\theta}}^p$ tends to $\boldsymbol{\theta}^p$. Also, $\hat{\zeta}(\mathbf{x})$ converges at an optimal rate to $\eta(\mathbf{x},\boldsymbol{\theta})$. This framework was extended to stochastic physical systems by Tuo and Wu (2015).

Defining the 'true' $\boldsymbol{\theta}^p$ as given in (1.5), Plumlee (2017) showed that the discrepancy function $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$ is orthogonal to the gradient of the simulator $\eta(\mathbf{x},\boldsymbol{\theta})$, and suggested the use of a Gaussian process prior on $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$ that respects this orthogonality property. This can be achieved through the use of a Gaussian process prior on $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$ with an appropriate covariance function, giving an approach known as Bayesian L_2-calibration. Orthogonality of $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$ and $\frac{\partial\eta(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ can be demonstrated as follows.

Taking the first derivative with respect to $\boldsymbol{\theta}$ for Equation (1.5) at $\boldsymbol{\theta}^p$ gives:

$$\frac{\partial}{\partial\boldsymbol{\theta}} \int_{\mathcal{X}} [\zeta(\mathbf{x}) - \eta(\mathbf{x},\boldsymbol{\theta})]^2 d\mathbf{x}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^p} = -\int_{\mathcal{X}} 2[\zeta(\mathbf{x}) - \eta(\mathbf{x},\boldsymbol{\theta}^p)] \frac{\partial\eta(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^p} d\mathbf{x} = 0,$$

and as $\delta_{\boldsymbol{\theta}^p}(\mathbf{x}) = \zeta(\mathbf{x}) - \eta(\mathbf{x},\boldsymbol{\theta}^p)$, this implies,

$$-\int_{\mathcal{X}} \delta_{\boldsymbol{\theta}^p}(\mathbf{x})\ \frac{\partial\eta(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^p} d\mathbf{x} = 0, \tag{1.6}$$

establishing the orthogonality result.

Gu and Wang (2018) criticised the above orthogonality condition as (1.6) can hold for any local minimum or even maximum. This orthogonality condition is necessary for minimising the L_2 norm but is not sufficient. The L_2 norm used for calibration may have multiple turning points, i.e. multiple points where the derivative is zero. They proposed an alternative prior for the discrepancy function, known as a scaled Gaussian process, which does not impose orthogonality. Instead, a prior distribution is placed on the L_2 norm of $\delta_{\boldsymbol{\theta}^p}(\cdot)$ that penalises large discrepancy functions.

In this thesis, we construct Bayesian optimal designs under the original Kennedy-O'Hagan framework, without imposing orthogonality through the prior for $\delta_{\boldsymbol{\theta}^p}(\cdot)$ (see Section 6.3). However, we briefly return to the topic of Bayesian L_2 calibration in Section 7.2.2.

Calibration can also be performed using a frequentist approach. Loeppky et al. (2006) introduced a likelihood alternative to the Bayesian methodology for estimating unknown parameters. This approach involved finding the values of the unknown calibration parameters that maximise the likelihood of the simulator and physical training data. Joseph and Melkote (2009) considered a parametric form of the discrepancy function. Wong et al. (2017) formulated a nonparametric model for the discrepancy function and used the L_2 norm (1.5) to define the 'true' calibration parameters, and estimate these by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i - \eta(\mathbf{x}_i, \boldsymbol{\theta}) \right\}^2 .$$

They estimated $\delta_{\boldsymbol{\theta}^p}(\cdot)$ by applying a nonparametric regression method to $\{\mathbf{x}_i, y_i - \eta(\mathbf{x}_i, \hat{\boldsymbol{\theta}})\}$, $i = 1, \ldots, n$, to deal with identifiability issues.

As an alternative to Kennedy-O'Hagan calibration, history matching has been used with computer simulators to obtain parameter sets that may plausibly contain $\boldsymbol{\theta}^p$ (Craig et al., 1997; Vernon et al., 2010). The basic idea of history matching is to use the mean and variance of an emulator of the simulator to calculate the 'implausibility' of an input combination $\boldsymbol{\theta}$, using the standardised difference between the emulator mean and the physical data (Wilkinson, 2014; Oakley and Youngman, 2017). Implausible regions of the parameter space are then ruled out, and the simulator is re-run where 'implausibility' is low. The emulator is then updated in the reduced parameter space, and new 'implausibility' measures are calculated. History matching does not assume a complete probability model for (1.1); in particular, no prior distribution is assumed for the discrepancy function or the parameters. As we shall see later in this thesis, a full probability model is needed for decision-theoretic design of experiments, leading us to focus on the Kennedy-O'Hagan calibration framework.

1.2 Design of experiments

In calibration, data from two types of experiments inform the estimation of the statistical model: the computer experiment, in which $\eta(\mathbf{x}_j^c, \boldsymbol{\theta}_j^c)$ is evaluated, and the physical experiment, in which the physical observation y_i is assumed to be observed at fixed and unknown values $\boldsymbol{\theta}^p$. This leads to two design problems: choice of the set of values of $\mathbf{x}_1^c, \ldots, \mathbf{x}_m^c$ and $\boldsymbol{\theta}_1^c, \ldots, \boldsymbol{\theta}_m^c$ at which to evaluate the simulator, and choice of conditions, $\mathbf{x}_1^p, \ldots, \mathbf{x}_n^p$ under which to observe the physical process.

Experimental design involves the specification of all aspects of an experiment. The choice of a design is often considered as an optimisation problem. Optimal experiment designs are the 'best' designs under a specific criterion, tailored to the experimental goals. Experimental design has been widely studied in theory and in practice, see, for example, Atkinson et al. (2007). The decisions that must be made when designing an experiment include which treatments, that is, combinations of values of the controllable variables, to run, the choice of sample size, specification of the experimental units to be studied and the choice of ranges or levels for each variable. In this thesis we address the design of the physical experiment and choose which combination of values of the controllable variables at which to observe the physical process.

1.2.1 Physical experiment

We define a design for a physical experiment as a set $\xi = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ of n points with each $\mathbf{x}_i$ chosen from a design space $\mathcal{X} \subset \mathbb{R}^{q_1}$. A n-size optimal design $\xi^\star$ is defined by comparison with the set Ξ of all possible designs of size n with respect to a specific criterion. The objective function, φ, reflects the aim of the experiment, which is to be maximised or minimised, and is used as the criterion for a design to be optimal in the set Ξ.

In the frequentist approach to design (Kiefer and Wolfowitz, 1959), many criteria have φ formulated as a function of the expected Fisher information matrix $I(\boldsymbol{\psi};\xi)$, for parameters $\boldsymbol{\psi}$ under a statistical model with likelihood function $\pi_l(\mathbf{y}|\boldsymbol{\psi},\xi)$.

Definition 1.1. The expected *Fisher information matrix* of the parameter vector $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{q_2})^{\mathrm{T}}$ is defined as the covariance matrix of the score function, i.e. the variance of the gradient of the log-likelihood function, $\log \pi_l(\mathbf{y}|\boldsymbol{\psi},\xi)$, with respect to $\boldsymbol{\psi}$. Assuming mild regularity conditions we have:

$$I(\boldsymbol{\psi};\xi) = \mathrm{Var}\left(\frac{d}{d\boldsymbol{\psi}} \log \pi_l(\mathbf{y}|\boldsymbol{\psi},\xi)\right) = -\mathbb{E}\left(\frac{d^2}{d\boldsymbol{\psi} d\boldsymbol{\psi}^{\mathrm{T}}} \log \pi_l(\mathbf{y}|\boldsymbol{\psi},\xi)\right), \tag{1.7}$$

where, for a function $f(\boldsymbol{\psi})$, $\frac{df}{d\boldsymbol{\psi}} = \left(\frac{\partial f}{\partial \psi_1}, \ldots, \frac{\partial f}{\partial \psi_{q_2}}\right)^{\mathrm{T}}$, and $\frac{d^2 f}{d\boldsymbol{\psi} d\boldsymbol{\psi}^{\mathrm{T}}}$ is the matrix with ijth element $\frac{\partial^2 f}{\partial \psi_i \partial \psi_j}$.

Common criteria include A-optimality, under which a design minimises $\mathrm{tr}[I(\boldsymbol{\psi};\xi)^{-1}]$ with respect to ξ, and D-optimality under which a design maximises $\log|I(\boldsymbol{\psi};\xi)|$ with respect to ξ given $\boldsymbol{\psi}$ (Fedorov, 1972; Pukelsheim and Torsney, 1991; Atkinson et al., 2007).

For nonlinear models, i.e. models that are nonlinear in the unknown parameters, or when interest is in estimating nonlinear functions of the model parameters in a linear model, the information matrix $I(\boldsymbol{\psi};\xi)$ may depend on the values of these parameters. This creates a problem for classical optimal design, as we require knowledge of the values of the model parameters prior to designing an optimal experiment to estimate them. There are three common approaches to address this problem.

- Locally optimal designs, using a point prior guess for $\boldsymbol{\psi}$.
- Formulation of criteria that optimise some summary of a classical objective function with respect to the prior information, e.g. an average or minimax crite-

rion (e.g. Pronzato and Walter, 1985). Optimising the expectation of classical objective functions with respect to a prior distribution is often referred to as "Pseudo-Bayesian" and has an asymptotic Bayesian justification via a normal approximation to the posterior distribution (Chaloner and Verdinelli, 1995).

- Fully Bayesian design (see Section 4.1) optimising a function of the posterior density.

Often, at least some information is available prior to an experiment and hence Bayesian methods can be very useful. In the Bayesian approach to designing experiments (e.g. Chaloner, 1984; Müller, 1999; Cook et al., 2008; Müller et al., 2006; Huan and Marzouk, 2013; Ryan et al., 2016) the available prior information about the parameters is exploited. As described by Chaloner and Verdinelli (1995), the design problem is formulated as a decision-theoretic problem (more in Section 4.1).

1.2.2 Computer experiments

We define a design for a computer experiment as a set $\xi^c = [(\mathbf{x}_1^c, \boldsymbol{\theta}_1^c), \ldots, (\mathbf{x}_m^c, \boldsymbol{\theta}_m^c)]$ of m choices of input combinations at which to collect simulator evaluations to build an emulator, chosen from a design space $\mathcal{X} \times \boldsymbol{\Theta}$. An optimal design $\xi^{c\star}$ of size m is defined by comparison with the set Ξ_c of all possible designs of size m with respect to a specific criterion. The design of computer experiments has been well-studied in the past 30 years and there is substantial literature, see for example Sacks et al. (1989), Santner et al. (2003) and Burstyn and Steinberg (2006), that indicates its rapid development.

For the design of computer experiments two classes of designs have been considered: the model-based and model-free approaches (see Pronzato and Müller (2012), for an overview). The model-based approach explicitly accounts for the statistical emulator and is separated into designs for estimation (see Section 4.1) and designs for prediction, e.g. selecting the design points to minimise prediction error. The model-free approach does not make use of any assumptions about the statistical emulator (for example, the Gaussian process prior) that will be used to approximate the simulator. The most popular model-free approach is the use of space-filling designs (see Lin and Tang (2015) for a recent review).

Space-filling properties are usually defined via summaries of Euclidean distances between design points (and sometimes other points in the design space). Common methods of finding space-filling designs include:

- optimising summaries of the distances between design points, i.e. geometric criteria (Johnson et al., 1990)

- sampling methods including simple random sampling, stratified random sampling and Latin Hypercube sampling (McKay et al., 1979)

- optimising statistical measures of uniformity (Fang et al., 2000).

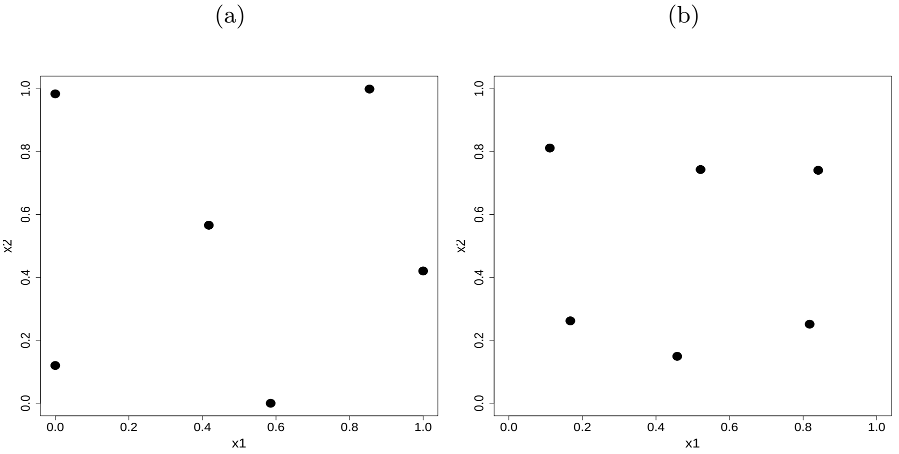


Figure 1.1: Example of (a) maximin and (b) minimax designs for two variables and $m = 6$ points

Johnson et al. (1990) developed designs based on geometric criteria. The two main categories are: maximin- and minimax-distance designs. First we define $d\left[(\mathbf{x}^c, \boldsymbol{\theta}^c), (\mathbf{x}^{c\prime}, \boldsymbol{\theta}^{c\prime})\right]$ to be a distance function, e.g. the Euclidean distance.

Definition 1.2. The Euclidean distance is the straight line distance between two vectors $\mathbf{w} = [w_1, \ldots, w_q]^{\mathrm{T}}, \mathbf{w}' = [w'_1, \ldots, w'_q]^{\mathrm{T}} \in \mathbb{R}^q$ given by:

$$d(\mathbf{w}, \mathbf{w}') = \sqrt{\sum_{h=1}^{q} (w_h - w'_h)^2}.$$

Maximin-distance criterion: a maximin-distance design $\xi^{c\star}$ maximises

$$\varphi_{Mm}(\xi^c) = \min_{j \neq j' \in \{1,\ldots,m\}} d\left[(\mathbf{x}^c_j, \boldsymbol{\theta}^c_j), (\mathbf{x}^c_{j'}, \boldsymbol{\theta}^c_{j'})\right].$$

Minimax-distance criterion: a minimax-distance design $\xi^{c\star}$ minimises

$$\varphi_{mM}(\xi^c) = \max_{(\mathbf{x}^c, \boldsymbol{\theta}^c) \in \mathcal{X} \times \boldsymbol{\Theta};\, j=1,\ldots,m} \min d\left[(\mathbf{x}^c, \boldsymbol{\theta}^c), (\mathbf{x}^c_j, \boldsymbol{\theta}^c_j)\right].$$

A maximin-distance design maximises the minimum distance between any two points in the design. Hence, the points are *spread* throughout the region and no pairs of points in the design are ‘too close’. A minimax-distance design minimises the biggest distance from all the points in the region to their nearest point in the design. Hence, the points *cover* the design region. In general, minimax-distance designs are not widely used because they are computationally difficult to generate. See Figure 1.1 for examples of maximin and minimax designs on $[0, 1]^2$ for two variables and $m = 6$.

Another approach to select the design points in a computer experiment utilises sampling methods such as: simple random sampling, stratified random sampling and Latin

Hypercube sampling. In simple random sampling the m points of the design are selected from the design region $\mathcal{X}$ at random, typically with respect to a uniform distribution. However, in high dimensions this method results in poor coverage of the design space and clustering of points. Stratified random sampling was proposed to overcome this problem. The design region is partitioned into m equally sized strata and one point is randomly selected from each stratum. Stratified sampling benefits from coverage of the whole experimental region. However, it is difficult for this method to cover a high-dimensional space. Latin Hypercube Designs (LHD) were proposed to overcome this problem.

When the output is influenced by only a few input variables, the design points should be evenly spaced across the projections onto these significant inputs. Latin Hypercube designs (McKay et al., 1979) were introduced to satisfy exactly this need; to allow for the projection of the design points into any single dimension to be equally spaced. The design region is divided into cells with equal size and then m cells are randomly selected satisfying the contraint that the projections of the selected cells on to each dimension do not overlap.

McKay et al. (1979) compared the above methods of sampling, and concluded that Latin Hypercube designs gave more accurate, i.e. lower variance, estimates of the mean and variance of the probability distribution of the output. Latin Hypercube designs are computationally inexpensive, easy to generate and have good projection properties. Because of these reasons, they have become the most popular sampling method for computer experiments. Different extensions of Latin Hypercube designs have also been proposed.

Several authors considered combining aspects of these three methods of designing computer experiments. For example, a geometric criterion can be used to find a maximin Latin Hypercube design, see Morris and Mitchell (1995) and Santner et al. (2003, Chapter 5). In addition, projection properties in higher dimensions have also been considered (e.g. Tang, 1993; Joseph et al., 2015).

1.2.3 Combining physical and computer experiments

In the physical experiment the observations are the result of a designed experiment on the physical process. There is little literature on this design problem, although Ranjan et al. (2011) and Williams et al. (2011) suggested designing and running these experiments in batches, updating the posterior distributions between batches and taking into account this updated information when designing the next batch. Leatherman et al. (2017) compared Mean Squared Error optimal designs for the combined physical and computer experiments using a particle swarm optimisation algorithm at a grid of inputs to find the starting design and a gradient-based quasi-Newton algorithm to search for the optimal design. More details about these methods and their limitations can be found in Chapter 6.

1.3 Aim and objectives

The aim of this thesis is to develop methodology for Bayesian optimal design of experiments for situations where the simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$ does not provide an adequate description of the mean of a system or a process, or the simulator might be expensive to evaluate, precluding its direct use in inference.

This work differs from previous research in the area as it is the first to find fully Bayesian optimal designs for a calibration model using the expected Shannon information gain utility function. Previous literature has addressed the design of follow-up experiments or locally-optimal designs. The Shannon information gain utility is combined with the approximate coordinate exchange (ACE) algorithm (Overstall and Woods, 2017) to construct optimal designs; we propose and use new methods, called ALIS and LIS, for approximating the evidence $\pi_e(\mathbf{y}|\xi)$, on which Shannon information gain depends, which reduce bias that might lead to overestimation of the information gain from a design.

Specific objectives of the thesis are to:

1. review the area of decision-theoretic design, especially the numerical approximation of the expected utility;
2. develop a new methodology for approximating the evidence in the evaluation of the expected Shannon information gain;
3. apply this methodology to approximate the expected utility for nonlinear models and combine this methodology with an optimisation algorithm to find Bayesian optimal designs for these models;
4. perform the first thorough comparison of several existing methods of approximating the expected Shannon information gain with the new proposed methods in terms of accuracy, precision and computational cost;
5. develop methodology to find Bayesian optimal designs for the physical process to enable collection of informative data that enable efficient estimation of the parameters $\boldsymbol{\theta}^p$ in the calibration model;
6. apply this methodology in cases where the simulator: (i) does not provide an accurate description of the mean; or (ii) is expensive to run or does not have an analytical form.

1.4 Thesis Organisation

In Chapter 2 we introduce a number of key concepts for Gaussian process models and their role in calibration. We review the Bayesian approach, which will be used throughout the thesis, and present fundamental results for the Gaussian process models. The Gaussian process allows inference about an observation at a new point, $\tilde{\mathbf{x}}$, via the posterior predictive distribution (Rasmussen and Williams, 2006). However, in the most general case, this distribution is not available in closed form. Hence, we apply Markov chain Monte Carlo (MCMC) methods in order to evaluate numerically intractable integrals.

Chapter 3 illustrates the impact of choice of design on calibration through some simple examples. First we introduce the Michaelis-Menten model, the estimation of which we initially treat as a simple Bayesian nonlinear regression problem; in later examples, this model is used as a known (i.e. computationally inexpensive) simulator in a calibration problem. We divide the latter into two cases; a known simulator with known calibration parameters and a known simulator with unknown calibration parameters.

In Chapter 4 we describe and apply the decision-theoretic approach to develop Bayesian optimal designs using the expected Shannon information gain utility function (Shannon, 1948) and illustrate the evaluation of the expected Shannon information gain on a simple example. A naïve nested Monte Carlo scheme is the simplest approach for approximating the expected utility, however in some cases it fails to give an accurate approximation. For this reason several authors have proposed more sophisticated methods that both reduce the computational burden and bias. Here, we introduce additional new methods for approximating the expected utility. In Chapter 5 we compare and assess through examples the different methods, including our new proposed methods, for approximating the expected Shannon information gain. We then describe the approximate coordinate exchange (ACE) algorithm for finding designs that maximise the expected utility. We combine these methods with the ACE algorithm to find Bayesian optimal designs.

Often the simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$ may not provide an adequate description of the mean, may be computationally expensive to run, or both. We address these problems in Chapter 6 following the Kennedy-O'Hagan calibration framework. For the first problem, we find optimal designs for the calibration model (1.1) assuming a Gaussian process prior on the unknown discrepancy function $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$. For the second problem, we assume a Gaussian process prior on the computationally expensive simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$ and find optimal designs to collect the physical experimental data to combine with simulator runs, in order to compute the posterior distribution for the unknown calibration parameters $\boldsymbol{\theta}^p$. We find Bayesian optimal designs by combining our new methods for approximating the expected utility with the ACE algorithm.

Chapter 7 concludes this thesis by summarising the research contributions and suggests future research directions.

Chapter 2

Gaussian Processes

In this chapter Gaussian process models are described in detail and the related concepts and methods used in this thesis are introduced. We begin by defining a Gaussian process and discussing its properties. After a brief introduction to Bayesian inference, we describe the Bayesian approach to Gaussian process modelling. Using results from the literature we give formulations and derivations of the prior, posterior and predictive distributions which are used in the following chapters.

2.1 Introduction to the Gaussian process model

Given data of the form $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq_1})^{\mathrm{T}} \in \mathcal{X}$ and y_i denotes a response measured at a specific point $\mathbf{x}_i$, we assume that

$$y_i = g(\mathbf{x}_i) + \varepsilon_i, \tag{2.1}$$

where ε_i represents the measurement error with $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ independently, and the deterministic function $g(\mathbf{x})$ is unspecified.

O'Hagan (1978) used Gaussian processes to model the behaviour of an unknown mathematical function. We adopt a nonparametric Bayesian approach by assuming the Gaussian process prior

$$g(\mathbf{x}) \sim \mathrm{GP}\left[\mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}, \sigma^2\kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi})\right], \tag{2.2}$$

where $\mathbf{f}(\mathbf{x}) = [f_0(\mathbf{x}), f_1(\mathbf{x}), \ldots, f_{k-1}(\mathbf{x})]^{\mathrm{T}}$ is a k-vector of known regression functions, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{k-1})^{\mathrm{T}} \in \mathcal{B}$ is also a k-vector of unknown trend parameters, $0 \leq \kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi}) \leq 1$ is the correlation function, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{q_1})^{\mathrm{T}} \in \Phi = (0, \infty)^{q_1}$ is the vector of (positive) correlation parameters and $\sigma^2 > 0$ is the constant variance. By nonparametric regression here we mean that the complexity of the approximating model for $g(\mathbf{x})$ increases with n.

The defining property of the prior (2.2) is that any finite collection of function evaluations $\mathbf{g} = [g(\mathbf{x}_1), \ldots, g(\mathbf{x}_n)]^{\mathrm{T}}$ has a multivariate normal distribution,

$$\mathbf{g} \sim \mathrm{N}\left[\mathbf{F}\boldsymbol{\beta}, \sigma^2 \mathbf{K}(\boldsymbol{\phi})\right], \tag{2.3}$$

where $\mathbf{F} = [\mathbf{f}(\mathbf{x}_1)\ \mathbf{f}(\mathbf{x}_2) \ldots \mathbf{f}(\mathbf{x}_n)]^{\mathrm{T}}$ is the $n \times k$ model matrix and $\mathbf{K}(\boldsymbol{\phi})$ is the correlation matrix with ijth entry $\mathbf{K}(\boldsymbol{\phi})_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\phi})$, $i, j = 1, \ldots, n$.

The set of allowable correlation functions is limited by the fact that $\mathbf{K}(\boldsymbol{\phi})$ must be positive-definite and symmetric for any choices of $[\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\boldsymbol{\phi}$, and also because $\kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi}) = 1$ when $\mathbf{x} = \mathbf{x}'$. A correlation function is separable when it can be written as a product of one-dimensional correlation functions,

$$\kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi}) = \prod_{r=1}^{q_1} \omega(x_r, x'_r; \phi_r), \tag{2.4}$$

where $\mathbf{x} = (x_1, \ldots, x_{q_1})^{\mathrm{T}}$ and $\mathbf{x}' = (x'_1, \ldots, x'_{q_1})^{\mathrm{T}}$.

One important family of correlation functions is the *powered exponential*, see for example Diggle et al. (1998), which has the form,

$$\omega(x, x'; \phi) = \exp\left[-\phi \mid\mid x - x' \mid\mid^{\nu}\right], \tag{2.5}$$

where $0 < \nu \leq 2$ is the decay parameter, $\phi > 0$ is the smoothness parameter and $|| \cdot ||$ denotes the Euclidean norm. We treat ν as fixed and known.

Another widely used family of correlation functions is the *Matèrn* (Matèrn, 1960), given by:

$$\omega(x, x'; \phi) = \frac{1}{2^{\nu-1}\Gamma(\nu)} (2\sqrt{\nu} \mid\mid x - x' \mid\mid \phi)^{\nu} K_{\nu}(2\sqrt{\nu} \mid\mid x - x' \mid\mid \phi),$$

where $\nu > 0$ is the decay parameter, $\phi > 0$ is the smoothness parameter and K_ν is the modified Bessel function of order ν.

In this thesis we will adopt the *squared exponential* correlation function,

$$\omega(x, x'; \phi) = \exp\left[-\phi \mid\mid x - x' \mid\mid^{2}\right]. \tag{2.6}$$

The squared exponential correlation function is a special case of the powered exponential correlation functions with $\nu = 2$. This correlation function is stationary[1] and a decreasing function of the Euclidean distance between x and x'. It also corresponds to the prior assumption that the model is very smooth in the sense that is infinitely differentiable.

[1] A stationary correlation function is a function of $x - x'$, and it is invariant to translations of the input space.

Combining Equations (2.1) and (2.3) we have

$$\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \sigma_\varepsilon^2 \sim N(\mathbf{F}\boldsymbol{\beta}, \sigma^2\mathbf{K}(\boldsymbol{\phi}) + \sigma_\varepsilon^2\mathbf{I}_n),$$

where $\mathbf{y} = (y_1, \dots, y_n)^{\mathrm{T}}$ is the n-vector of responses and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

It is useful to reparameterise the model to make computation easier. Here we adopt the reparameterisation

$$\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N\left(\mathbf{F}\boldsymbol{\beta}, \sigma^2[\mathbf{K}(\boldsymbol{\phi}) + \tau^2\mathbf{I}_n]\right), \tag{2.7}$$

where $\tau^2 = \sigma_\varepsilon^2/\sigma^2$ is the ratio of the noise to the process variation, known as the nugget (Diggle and Ribeiro, 2007). From now on we will use the reparameterised covariance matrix $\sigma^2[\mathbf{K}(\boldsymbol{\phi}) + \tau^2\mathbf{I}_n] = \sigma^2\boldsymbol{\Sigma}$.

Gaussian process modelling is a Bayesian alternative for classical nonparametric methods such as splines and local polynomial regression (Rasmussen and Williams, 2006, Chapter 6; Gramacy and Lee, 2008). In this thesis we follow a fully Bayesian approach and the Gaussian process provides a natural description of our prior beliefs about the function and allows us to update these beliefs using the data. In Gaussian process modelling, conjugate prior distributions are available for some parameters to simplify the calculations required for obtaining predictions.

In the next section we point out the relevance of Gaussian processes to calibration. We discuss inference for the Gaussian process model in Section 2.3.

2.2 The role of Gaussian processes in calibration

When computer simulations are time consuming and very computationally expensive to run, there is a need to find a computationally cheaper metamodel or *emulator*, that can replace the simulator to some degree. The emulator provides fast prediction of the outputs at untested input points, together with a measure of uncertainty about these predictions. A very popular emulator, is the Gaussian process, introduced to the field of computer experiments by Sacks et al. (1989). The Gaussian process emulator is a flexible and adaptive non-parametric smoother/interpolator[2], and can be used to gain knowledge into the simulator over the entire design region. Tasks such as validation and calibration, sensitivity and uncertainty analysis thus become feasible for expensive simulators (see Santner et al., 2003 and Fang et al., 2006).

We consider the calibration model (1.1). As mentioned in Section 1.1, in this model we have two groups of inputs but we also have two groups of responses. First we have the outputs after running the simulator for inputs $(\mathbf{x}_j^c, \boldsymbol{\theta}_j^c)$, $j = 1, \dots, m$, and second the

[2] A function f is an interpolator of the data (x_i, y_i) if $f(x_i) = y_i$, $i = 1, \dots, n$, and may be appropriate if $\sigma_\varepsilon^2 = 0$, i.e. $\varepsilon_i = 0$ in (2.1). A smoother is an estimate of the regression function g in (2.1) that does not need to pass through the data points.

observations of the physical process with inputs $\mathbf{x}_i^p$, $i = 1, \ldots, n$. The motivation for the notation distinguishing of $\mathbf{x}^p$ and $\mathbf{x}^c$ comes from the fact that we may not use the same values of the controllable variables that were used when observing the physical system and when running the simulator.

Let $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ be the vector of n responses from the physical experiment and $\mathbf{z} = [\eta(\mathbf{x}_1^c, \boldsymbol{\theta}_1^c), \ldots, \eta(\mathbf{x}_m^c, \boldsymbol{\theta}_m^c)]^{\mathrm{T}}$ be the outcomes of m-runs of the simulator. For simplicity we will assume that the regression parameter is known and fixed at $\rho = 1$. We represent prior uncertainty about both the simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$ and discrepancy $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$ by Gaussian processes,

$$\eta(\mathbf{x}, \boldsymbol{\theta}) \sim \mathrm{GP}\left(\mathbf{f}_\eta^{\mathrm{T}}(\mathbf{x}, \boldsymbol{\theta})\boldsymbol{\beta}_\eta,\ \sigma_\eta^2 \kappa_\eta[(\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}'); \boldsymbol{\phi}_\eta]\right), \tag{2.8}$$

and also,

$$\delta_{\boldsymbol{\theta}^p}(\mathbf{x}) \sim \mathrm{GP}\left(\mathbf{f}_\delta^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}_\delta, \sigma_\delta^2 \kappa_\delta(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi}_\delta)\right). \tag{2.9}$$

Above $\mathbf{f}_\eta(\mathbf{x}, \boldsymbol{\theta}) = (f_0^\eta(\mathbf{x}, \boldsymbol{\theta}), \ldots, f_{k_\eta - 1}^\eta(\mathbf{x}, \boldsymbol{\theta}))^{\mathrm{T}}$ and $\mathbf{f}_\delta(\mathbf{x}) = (f_0^\delta(\mathbf{x}), \ldots, f_{k_\delta - 1}^\delta(\mathbf{x}))^{\mathrm{T}}$ are the k_η- and k_δ-vectors of known regression functions, respectively, of the Gaussian process prior for the simulator and discrepancy. In addition, $\boldsymbol{\beta}_\eta = (\beta_0^\eta, \beta_1^\eta, \ldots, \beta_{k_\eta - 1}^\eta)^{\mathrm{T}}$ and $\boldsymbol{\beta}_\delta = (\beta_0^\delta, \beta_1^\delta, \ldots, \beta_{k_\delta - 1}^\delta)^{\mathrm{T}}$ are the corresponding parameter vectors that contain the unknown trend parameters of the Gaussian process prior for the simulator and discrepancy respectively. Further, $\kappa_\eta[(\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}'); \boldsymbol{\phi}_\eta]$ and $\kappa_\delta(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi}_\delta)$ are the correlation functions of the Gaussian process prior for the simulator and discrepancy respectively, with vector of correlation parameters $\boldsymbol{\phi}_\eta$ and $\boldsymbol{\phi}_\delta$. Finally, σ_η^2 and σ_δ^2 are the prior emulator variance and the prior variance of the Gaussian process for the discrepancy, respectively.

We define the $(n + m)$-vector $\mathbf{v} = \left[\mathbf{y}^{\mathrm{T}}\ \mathbf{z}^{\mathrm{T}}\right]^{\mathrm{T}}$ to contain the physical data and the outcomes of the m-runs of the simulator. The distribution of the combined vector of responses, $\mathbf{v}$, is:

$$\mathbf{v} \mid \boldsymbol{\theta}^p, \boldsymbol{\beta}_\eta, \boldsymbol{\beta}_\delta, \sigma_\eta^2, \sigma_\delta^2, \sigma_\varepsilon^2, \boldsymbol{\phi}_\eta, \boldsymbol{\phi}_\delta \sim N(\boldsymbol{\mu}_{\mathbf{v}}, \boldsymbol{\Sigma}_{\mathbf{v}}). \tag{2.10}$$

Above, the prior conditional expectation of $\mathbf{v}$ is

$$\boldsymbol{\mu}_{\mathbf{v}} = \begin{bmatrix} \mathbf{F}_\eta^p \boldsymbol{\beta}_\eta + \mathbf{F}_\delta^p \boldsymbol{\beta}_\delta \\ \mathbf{F}_\eta^c \boldsymbol{\beta}_\eta \end{bmatrix}, \tag{2.11}$$

where $\mathbf{F}_\eta^p = [\mathbf{f}_\eta(\mathbf{x}_1^p, \boldsymbol{\theta}^p)\ \mathbf{f}_\eta(\mathbf{x}_2^p, \boldsymbol{\theta}^p) \ldots\ \mathbf{f}_\eta(\mathbf{x}_n^p, \boldsymbol{\theta}^p)]^{\mathrm{T}}$, $\mathbf{F}_\delta^p = [\mathbf{f}_\delta(\mathbf{x}_1^p)\ \mathbf{f}_\delta(\mathbf{x}_2^p) \ldots\ \mathbf{f}_\delta(\mathbf{x}_n^p)]^{\mathrm{T}}$ and $\mathbf{F}_\eta^c = [\mathbf{f}_\eta(\mathbf{x}_1^c, \boldsymbol{\theta}_1^c)\ \mathbf{f}_\eta(\mathbf{x}_2^c, \boldsymbol{\theta}_2^c) \ldots\ \mathbf{f}_\eta(\mathbf{x}_m^c, \boldsymbol{\theta}_m^c)]^{\mathrm{T}}$.

In addition the prior conditional covariance matrix for $\mathbf{v}$ in (2.10) is the $(n+m) \times (n+m)$ matrix:

$$\boldsymbol{\Sigma}_{\mathbf{v}} = \sigma_\eta^2 \boldsymbol{\Sigma}_\eta + \begin{bmatrix} \sigma_\varepsilon^2 \mathbf{I}_n + \sigma_\delta^2 \boldsymbol{\Sigma}_\delta & \mathbf{0}_{n \times m} \\ \mathbf{0}_{m \times n} & \mathbf{0}_{m \times m} \end{bmatrix}. \tag{2.12}$$

Here, $\boldsymbol{\Sigma}_\delta$ is an $n \times n$ correlation matrix with ii'th entry $\kappa_\delta(\mathbf{x}_i^p, \mathbf{x}_{i'}^p; \boldsymbol{\phi}_\delta)$, and

$$\boldsymbol{\Sigma}_\eta = \begin{bmatrix} \boldsymbol{\Sigma}_\eta^{pp} & \boldsymbol{\Sigma}_\eta^{pc} \\ \boldsymbol{\Sigma}_\eta^{cp} & \boldsymbol{\Sigma}_\eta^{cc} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_\eta^{pp}$ is the $n\times n$ correlation matrix with ii'th entry $\kappa_\eta[(\mathbf{x}_i^p, \boldsymbol{\theta}^p), (\mathbf{x}_{i'}^p, \boldsymbol{\theta}^p); \boldsymbol{\phi}_\eta]$, $\boldsymbol{\Sigma}_\eta^{pc}$ is the $n \times m$ correlation matrix with ijth entry $\kappa_\eta[(\mathbf{x}_i^p, \boldsymbol{\theta}^p), (\mathbf{x}_j^c, \boldsymbol{\theta}_j^c); \boldsymbol{\phi}_\eta]$ and $\boldsymbol{\Sigma}_\eta^{pc} = (\boldsymbol{\Sigma}_\eta^{cp})^{\mathrm{T}}$. The $m \times m$ correlation matrix $\boldsymbol{\Sigma}_\eta^{cc}$ has jj'th entry $\kappa_\eta[(\mathbf{x}_j^c, \boldsymbol{\theta}_j^c), (\mathbf{x}_{j'}^c, \boldsymbol{\theta}_{j'}^c); \boldsymbol{\phi}_\eta]$.

The likelihood function for $\mathbf{v}$ is then given by:

$$\pi_l(\mathbf{v} \mid \boldsymbol{\theta}^p, \boldsymbol{\beta}_\eta, \boldsymbol{\beta}_\delta, \sigma_\eta^2, \sigma_\delta^2, \sigma_\varepsilon^2, \boldsymbol{\phi}_\eta, \boldsymbol{\phi}_\delta) \propto |\boldsymbol{\Sigma}_\mathbf{v}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{v} - \boldsymbol{\mu}_\mathbf{v})^{\mathrm{T}} \boldsymbol{\Sigma}_\mathbf{v}^{-1} (\mathbf{v} - \boldsymbol{\mu}_\mathbf{v})\right\}.$$

The model formulation is completed by specifying prior distributions for the parameters $\boldsymbol{\beta}_\eta$, $\boldsymbol{\beta}_\delta$, σ_η^2, σ_δ^2, σ_ε^2, $\boldsymbol{\phi}_\eta$ and $\boldsymbol{\phi}_\delta$. Prior distributions are also required for the unknown calibration parameters $\boldsymbol{\theta}^p$. Details of our choices of prior distributions for the Gaussian process parameters are given in the examples in Chapter 6.

In the next section we discuss inference for the Gaussian process model. We return to the calibration problem in Chapter 6.

2.3 Gaussian process inference

Now we return to the Gaussian process model and present fundamental results for Bayesian inference and prediction.

2.3.1 Conditional prediction with known hyperparameters

In order to make predictive inference about the response, $\tilde{y}$ at a new point $\tilde{\mathbf{x}} \in \mathcal{X}$, we need to derive the predictive distribution for the random variable $\tilde{y}$ using model (2.7).

Following Banerjee et al. (2004, Chapter 2), the joint prior distribution of $\mathbf{y}$ and $\tilde{y}$, conditional on all unknown model parameters $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2$, is given by,

$$\left.\begin{pmatrix} \tilde{y} \\ \mathbf{y} \end{pmatrix} \right| \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \quad \sim \quad N\left(\begin{pmatrix} \mathbf{f}^{\mathrm{T}}(\tilde{\mathbf{x}})\boldsymbol{\beta} \\ \mathbf{F}\boldsymbol{\beta} \end{pmatrix}, \sigma^2 \begin{pmatrix} (1+\tau^2) & \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \\ \mathbf{k}(\tilde{\mathbf{x}}) & \boldsymbol{\Sigma} \end{pmatrix}\right), \tag{2.13}$$

where $\mathbf{k}(\tilde{\mathbf{x}}) = [\kappa(\tilde{\mathbf{x}}, \mathbf{x}_1; \boldsymbol{\phi}), \ldots, \kappa(\tilde{\mathbf{x}}, \mathbf{x}_n; \boldsymbol{\phi})]^{\mathrm{T}}$ is the n-vector of correlations between the response at each of the existing input points $\mathbf{x}_i$ and the response at $\tilde{\mathbf{x}}$.

Standard results for multivariate normal distributions can be used to derive the following conditional posterior distribution

$$\tilde{y} \mid \mathbf{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N\left(\mu(\tilde{\mathbf{x}}), s^2(\tilde{\mathbf{x}})\right),$$

with

$$\mu(\tilde{\mathbf{x}}) = \mathbb{E}(\tilde{y} \mid \mathbf{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2) = \mathbf{f}(\tilde{\mathbf{x}})^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left[\mathbf{y} - \mathbf{F}\boldsymbol{\beta}\right], \tag{2.14}$$

$$s^2(\tilde{\mathbf{x}}) = \mathrm{var}(\tilde{y} \mid \mathbf{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2) = \sigma^2[(1+\tau^2) - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{k}(\tilde{\mathbf{x}})]. \tag{2.15}$$

Equation (2.14) is the prior mean plus a weighted linear combination of the residuals from a linear model with regressors $\mathbf{F}$ and coefficients $\boldsymbol{\beta}$. The weight assigned to a residual decreases with the distance between the corresponding $\mathbf{x}_i$ and the point $\tilde{\mathbf{x}}$, at which we are predicting. The smaller these weights, the closer the prediction is to the conditional prior mean. Equation (2.15) is the prior variance minus a quadratic form in the correlations between the response at each of the existing input points and the response at $\tilde{\mathbf{x}}$. The further $\tilde{\mathbf{x}}$ is from the points at which we have observed, the smaller this quadratic form will be. Summarising, the closer $\tilde{\mathbf{x}}$ is to the points we have observed, the more we potentially update the prior mean and reduce the posterior variance. The further from the points we have seen, the closer we stay to the prior mean and variance. The parameter vector $\boldsymbol{\phi}$ controls the strength of correlation and hence clearly influences how much the conditional prediction changes due to observing the data, $\mathbf{y}$.

2.3.2 Bayesian inference

We define $\boldsymbol{\psi} = (\boldsymbol{\beta}^{\mathrm{T}}, \sigma^2, \boldsymbol{\phi}^{\mathrm{T}}, \tau^2)^{\mathrm{T}}$ as the vector of unknown hyperparameters which belongs to the parameter space $\Psi = \mathbb{R}^k \times (0,\infty) \times (0,\infty)^{q_1} \times (0,\infty)$. To obtain an unconditional posterior predictive distribution, we integrate out these hyperparameters with respect to their posterior density, $\pi_a(\boldsymbol{\psi} \mid \mathbf{y})$:

$$\pi(\tilde{y} \mid \mathbf{y}) = \int_{\Psi} \pi_a(\tilde{y} \mid \mathbf{y}, \boldsymbol{\psi})\pi_a(\boldsymbol{\psi} \mid \mathbf{y})d\boldsymbol{\psi}.$$

We obtain $\pi_a(\boldsymbol{\psi} \mid \mathbf{y})$ using Bayes' theorem to update our prior beliefs for the unknown parameters $\boldsymbol{\psi}$ after observing data $\mathbf{y}$. The posterior density for $\boldsymbol{\psi}$ satisfies

$$\pi_a(\boldsymbol{\psi} \mid \mathbf{y}) \propto \pi_l(\mathbf{y} \mid \boldsymbol{\psi})\pi_b(\boldsymbol{\psi}), \tag{2.16}$$

where $\pi_b(\boldsymbol{\psi})$ is the prior density and $\pi_l(\mathbf{y} \mid \boldsymbol{\psi})$ the likelihood of the parameters given the data. The density $\pi_b(\boldsymbol{\psi})$ contains all prior information we have about the unknown parameters $\boldsymbol{\psi}$. The posterior density $\pi_a(\boldsymbol{\psi} \mid \mathbf{y})$, the density of the parameters after taking into account the observed data, provides inference about the unknown parameters.

A family of prior distributions is *conjugate* to a particular likelihood if the posterior distribution is in the same family as the prior distribution (Raiffa and Schlaifer, 1961). We use conjugate prior distributions for the Gaussian process variance, σ^2, and trend parameter, $\boldsymbol{\beta}$, conditional on the vector of correlation parameters $\boldsymbol{\phi}$ and the nugget τ^2.

For further background on Bayesian inference, see O'Hagan and Forster (2004).

2.3.3 Prior specification

The model specification requires assignment of prior distributions to the unknown parameters $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2$. A common approach is to assume that the trend parameters, $\boldsymbol{\beta}$, and the Gaussian process variance, σ^2, are independent of the correlation parameters, $\boldsymbol{\phi}$, and the nugget, τ^2. In addition, $\boldsymbol{\phi}$ and τ^2 are also independent, $\pi_b(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2) = \pi_b(\boldsymbol{\beta} \mid \sigma^2)\pi_b(\sigma^2)\pi_b(\boldsymbol{\phi})\pi_b(\tau^2)$.

For the trend parameter $\boldsymbol{\beta}$ and variance σ^2, we can consider (conditionally) conjugate prior distributions and assign $\boldsymbol{\beta} \mid \sigma^2$ a normal distribution and σ^2 an inverse-gamma distribution,

$$\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\mathbf{R}) \text{ and } \sigma^2 \sim IG(a, b).$$

Therefore,

$$\pi_b(\boldsymbol{\beta} \mid \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{k}{2}} \mid \mathbf{R} \mid^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\},$$

and

$$\pi_b(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\},$$

where $\boldsymbol{\beta}_0$ is the k-vector of known prior means, $\mathbf{R}$ is a known symmetric, positive definite $k \times k$ matrix and $a, b > 0$ are known hyperparameters.

The joint density for $\boldsymbol{\beta} \mid \sigma^2$ and σ^2 is given by:

$$\begin{aligned}\pi_b(\boldsymbol{\beta}, \sigma^2) &= \pi_b(\boldsymbol{\beta} \mid \sigma^2)\pi_b(\sigma^2) \\ &\propto (\sigma^2)^{-(\frac{2a+k}{2}+1)} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + b\right]\right\}, \qquad (2.17)\end{aligned}$$

which corresponds to a normal-inverse-gamma prior distribution,

$$(\boldsymbol{\beta}, \sigma^2) \sim NIG\left(\boldsymbol{\beta}_0, \mathbf{R}, a, b\right).$$

In later chapters, we assume exponential prior distributions for the vector of correlation parameters $\boldsymbol{\phi}$ and the nugget τ^2 to guarantee positive values.

In order to use a Gaussian process model to make predictions, the posterior distributions of the parameters and the posterior predictive distribution are required. In the following two sections we derive these distributions.

2.3.4 Posterior predictive distribution with known correlation parameters and nugget

In this section, we assume fixed and known correlation parameters, $\boldsymbol{\phi}$, and nugget, τ^2, and we derive Bayesian inference results for the Gaussian process model. We allow for uncertainty only in the trend parameter, $\boldsymbol{\beta}$, and variance, σ^2. Hence the posterior distribution for $\boldsymbol{\beta}$ and σ^2, and the posterior predictive distribution are available in analytical form.

The likelihood function for the model (2.7) is

$$\pi_l(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{y}-\mathbf{F}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta})\right]\right\}. \qquad (2.18)$$

Using Bayes' Theorem (2.16) and the joint prior density (2.17) we can calculate the unnormalised posterior density:

$$\begin{aligned}
\pi_a(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \boldsymbol{\phi}, \tau^2) &\propto \pi_l(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2)\pi_b(\boldsymbol{\beta}, \sigma^2) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta})\right\} \\
&\quad \times \left(\frac{1}{\sigma^2}\right)^{\frac{2a+k}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[b + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right]\right\} \\
&= \left(\frac{1}{\sigma^2}\right)^{\frac{k+2a_\star}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_\star)^{\mathrm{T}}\boldsymbol{\Sigma}_\star^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_\star) + b_\star\right]\right\},
\end{aligned} \qquad (2.19)$$

where

$$\begin{aligned}
\boldsymbol{\beta}_\star &= \left(\mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{R}^{-1}\right)^{-1}\left(\mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y} + \mathbf{R}^{-1}\boldsymbol{\beta}_0\right) \\
\boldsymbol{\Sigma}_\star &= \left(\mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{R}^{-1}\right)^{-1} \\
a_\star &= a + \frac{n}{2} \\
b_\star &= b + \frac{1}{2}\left[(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)^{\mathrm{T}}\left(\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}\right)^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)\right].
\end{aligned} \qquad (2.20)$$

The posterior density given by (2.19) corresponds to a normal-inverse-gamma distribution $\mathrm{NIG}(\boldsymbol{\beta}_\star, \boldsymbol{\Sigma}_\star, a_\star, b_\star)$.

The expression for $b_\star$ is given by the use of the Sherman-Woodbury-Morrison identity (Harville, 2008) to establish that:

$$\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\beta}_0 + \mathbf{y}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y} - \boldsymbol{\beta}_\star^{\mathrm{T}}\boldsymbol{\Sigma}_\star\boldsymbol{\beta}_\star = \left[(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)^{\mathrm{T}}(\boldsymbol{\Sigma}+\mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}})^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)\right].$$

The marginal posterior distribution for $\boldsymbol{\beta}$ has density,

$$\begin{aligned}
\pi(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\phi}, \tau^2) &= \int_0^\infty \pi_a(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \boldsymbol{\phi}, \tau^2) d\sigma^2 \\
&\propto \int_0^\infty (\sigma^2)^{-\frac{2a_\star+k}{2}-1} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_\star)^{\mathrm{T}}\boldsymbol{\Sigma}_\star^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_\star) + b_\star\right]\right\} d\sigma^2 \\
&= \left[1 + \frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_\star)^{\mathrm{T}}\boldsymbol{\Sigma}_\star^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_\star)}{2b_\star}\right]^{-\frac{2a_\star+k}{2}}.
\end{aligned} \qquad (2.21)$$

Hence, $\boldsymbol{\beta}$ follows a multivariate t-distribution,

$$\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\phi}, \tau^2 \sim t_{2a_\star}\left(k, \boldsymbol{\beta}_\star, \frac{b_\star}{a_\star}\boldsymbol{\Sigma}_\star\right), \tag{2.22}$$

with $2a_\star$ degrees of freedom, mean $\boldsymbol{\beta}_\star$ and variance $\frac{b_\star}{a_\star-1}\boldsymbol{\Sigma}_\star$.

The marginal posterior for σ^2 is an inverse-gamma distribution,

$$\sigma^2 \mid \mathbf{y}, \boldsymbol{\phi}, \tau^2 \sim IG\left(a_\star, b_\star\right), \tag{2.23}$$

following directly from the fact that the joint posterior density of $\boldsymbol{\beta} \mid \sigma^2$ and σ^2 conditional on $\mathbf{y}$, $\boldsymbol{\phi}$ and τ^2 is a normal-inverse-gamma distribution.

To obtain the posterior predictive distribution for $\tilde{y}$ at a new point $\tilde{\mathbf{x}}$ we use the conditional posterior predictive distribution $\tilde{y} \mid \mathbf{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N(\mu(\tilde{\mathbf{x}}), s^2(\tilde{\mathbf{x}}))$ where $\mu(\tilde{\mathbf{x}})$ and $s^2(\tilde{\mathbf{x}})$ are given by Equations (2.14) and (2.15) respectively and the conditional posterior distribution, $\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N(\boldsymbol{\beta}_\star, \sigma^2\boldsymbol{\Sigma}_\star)$, from (2.19).

We can rewrite Equation (2.14) as

$$\mu(\tilde{\mathbf{x}}) = \mathbf{a}_{\tilde{y}} + \mathbf{b}_{\tilde{y}}^{\mathrm{T}}\boldsymbol{\beta},$$

where $\mathbf{a}_{\tilde{y}} = \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y}$ and $\mathbf{b}_{\tilde{y}}^{\mathrm{T}} = \mathbf{f}(\tilde{\mathbf{x}})^{\mathrm{T}} - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F}$. Hence,

$$\mu(\tilde{\mathbf{x}}) \mid \mathbf{y}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N\left[\mathbf{a}_{\tilde{y}} + \mathbf{b}_{\tilde{y}}^{\mathrm{T}}\boldsymbol{\beta}_\star,\ \mathbf{b}_{\tilde{y}}^{\mathrm{T}}(\sigma^2\boldsymbol{\Sigma}_\star)\mathbf{b}_{\tilde{y}}\right].$$

Also, from the conditional posterior predictive distribution we have

$$\tilde{y} - \mu(\tilde{\mathbf{x}}) \mid \mathbf{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N\left[0, s^2(\tilde{\mathbf{x}})\right], \tag{2.24}$$

where the variance $s^2(\tilde{\mathbf{x}})$ is given by (2.15).

The right-hand side of (2.24) does not depend on $\mu(\tilde{\mathbf{x}})$, and hence does not depend on $\boldsymbol{\beta}$, and therefore $\tilde{y} - \mu(\tilde{\mathbf{x}})$ is statistically independent of $\mu(\tilde{\mathbf{x}})$ given $\mathbf{y}$, σ^2, $\boldsymbol{\phi}$ and τ^2. Hence, given $\mathbf{y}$, σ^2, $\boldsymbol{\phi}$ and τ^2,

$$\tilde{y} = [\tilde{y} - \mu(\tilde{\mathbf{x}})] + \mu(\tilde{\mathbf{x}}),$$

is a sum of two independent normal random variables. Thus,

$$\tilde{y} \mid \mathbf{y}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N(\tilde{\mu}_\star, \tilde{\sigma}_\star^2),$$

with:

$$
\begin{aligned}
\tilde{\mu}_\star &= \mathbf{a}_{\tilde{y}} + \mathbf{b}_{\tilde{y}}^{\mathrm{T}} \boldsymbol{\beta}_\star \\
&= \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{y} + [\mathbf{f}(\tilde{\mathbf{x}})^{\mathrm{T}} - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F}] \boldsymbol{\beta}_\star \\
&= \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{y} + [\mathbf{f}(\tilde{\mathbf{x}})^{\mathrm{T}} - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F}] (\mathbf{F}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F} + \mathbf{R}^{-1})^{-1} (\mathbf{F}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{y} + \mathbf{R}^{-1} \boldsymbol{\beta}_0) \\
&= (\mathbf{f}^{\mathrm{T}}(\tilde{\mathbf{x}}) - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F}) (\mathbf{F}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F} + \mathbf{R}^{-1})^{-1} \mathbf{R}^{-1} \boldsymbol{\beta}_0 \\
&\quad + [\mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} + (\mathbf{f}^{\mathrm{T}}(\tilde{\mathbf{x}}) - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F}) (\mathbf{F}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F} + \mathbf{R}^{-1})^{-1} \mathbf{F}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}] \mathbf{y}, \qquad (2.25)
\end{aligned}
$$

$$
\begin{aligned}
\tilde{\sigma}_\star^2 &= s^2(\tilde{\mathbf{x}}) + \mathbf{b}_{\tilde{y}}^{\mathrm{T}} (\sigma^2 \boldsymbol{\Sigma}_\star) \mathbf{b}_{\tilde{y}} \\
&= \sigma^2 [(1+\tau^2) - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{k}(\tilde{\mathbf{x}})] + [\mathbf{f}(\tilde{\mathbf{x}})^{\mathrm{T}} - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F}] [\sigma^2 \boldsymbol{\Sigma}_\star] [\mathbf{f}(\tilde{\mathbf{x}})^{\mathrm{T}} - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F}]^{\mathrm{T}} \\
&= \sigma^2 \left\{ (1+\tau^2) - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{k}(\tilde{\mathbf{x}}) \right. \\
&\qquad \left. + [\mathbf{f}(\tilde{\mathbf{x}})^{\mathrm{T}} - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F}] [\mathbf{F}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F} + \mathbf{R}^{-1}]^{-1} [\mathbf{f}^{\mathrm{T}}(\tilde{\mathbf{x}}) - \mathbf{k}(\tilde{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{F}]^{\mathrm{T}} \right\} \\
&= \sigma^2 \tilde{\Sigma}_\star. \qquad (2.26)
\end{aligned}
$$

The components in the above expression of $\tilde{\sigma}_\star^2$ are interpreted as: (a) the variability without taking into account any information provided from the data, (b) the decrease in variability resulting from conditioning on the data, and (c) the rise in variability as a result of the posterior uncertainty of the estimation of $\boldsymbol{\beta}$.

The posterior predictive density is obtained by integrating out the unknown σ^2 with respect to its posterior distribution,

$$
\begin{aligned}
\pi(\tilde{y} \mid \mathbf{y}, \boldsymbol{\phi}, \tau^2) &= \int_0^\infty \pi(\tilde{y} \mid \mathbf{y}, \sigma^2, \boldsymbol{\phi}, \tau^2) \pi(\sigma^2 \mid \mathbf{y}, \boldsymbol{\phi}, \tau^2) d\sigma^2 \\
&\propto \left[1 + \frac{(\tilde{y} - \tilde{\mu}_\star)^2}{2 b_\star \tilde{\Sigma}_\star} \right]^{-(\frac{2a_\star + 1}{2})}, \qquad (2.27)
\end{aligned}
$$

where $\tilde{\mu}_\star$ and $\tilde{\Sigma}_\star$ are given by Equations (2.25) and (2.26) respectively. Equation (2.27) indicates that the posterior predictive density for the output $\tilde{y}$ at a new point $\tilde{\mathbf{x}}$ is a univariate (scalar) t-distribution,

$$
\tilde{y} \mid \mathbf{y}, \boldsymbol{\phi}, \tau^2 \sim t_{2a_\star} \left(1, \tilde{\mu}_\star, \frac{b_\star}{a_\star} \tilde{\Sigma}_\star \right), \qquad (2.28)
$$

with $2a_\star$ degrees of freedom, mean $\tilde{\mu}_\star$ and variance $\frac{b_\star}{a_\star - 1} \tilde{\Sigma}_\star$.

2.3.5 Unconditional inference

Unnormalised posterior density for the parameters

In practice, the values of the correlation parameters $\boldsymbol{\phi}$ will usually be unknown. Hence, we need to allow for uncertainty in $\boldsymbol{\phi}$. We now consider two cases. First, $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\phi}$

are unknown, conditional on τ^2 and then $\boldsymbol{\beta}$, σ^2, $\boldsymbol{\phi}$ and τ^2 are unknown. In both cases, the posterior distribution of the parameters or the posterior predictive distribution cannot be derived analytically.

Case 1: $\boldsymbol{\phi}$ unknown, τ^2 known

As in Section 2.3.3, we assign a normal-inverse-gamma prior distribution to $(\boldsymbol{\beta}, \sigma^2)$. We also consider an independent proper prior density for $\boldsymbol{\phi}$, giving the joint prior density,

$$\pi_b(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}) = \pi_b(\boldsymbol{\beta}, \sigma^2)\pi_b(\boldsymbol{\phi}).$$

By Bayes' theorem, the marginal posterior density for $\boldsymbol{\phi}$ satisfies:

$$\pi(\boldsymbol{\phi} \mid \mathbf{y}, \tau^2) \propto \pi(\mathbf{y} \mid \boldsymbol{\phi}, \tau^2)\pi_b(\boldsymbol{\phi}).$$

Thus we require the marginal likelihood of the data, $\pi(\mathbf{y} \mid \boldsymbol{\phi}, \tau^2)$. This likelihood takes the form of a $t_{2a}\left[n, \mathbf{F}\boldsymbol{\beta}_0, \frac{b}{a}\left[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}\right]\right]$ distribution, obtained from,

$$\begin{aligned}
\pi(\mathbf{y} \mid \boldsymbol{\phi}, \tau^2) &= \int_0^\infty \pi(\mathbf{y} \mid \sigma^2, \boldsymbol{\phi}, \tau^2)\pi_b(\sigma^2) d\sigma^2 \\
&= \frac{(2\pi)^{-\frac{n}{2}} b^a \Gamma(a_\star)}{|\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}|^{\frac{1}{2}}\Gamma(a)} \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{a_\star + 1} \\
&\qquad \times \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^{\mathrm{T}}\left[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}\right]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0) + b\right]\right\} d\sigma^2 \\
&= \frac{(2\pi)^{-\frac{n}{2}} b^a \Gamma(a_\star)}{|\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}|^{\frac{1}{2}}\Gamma(a)} \left[b + \frac{(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^{\mathrm{T}}\left[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}\right]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)}{2}\right]^{-a_\star} \\
&\propto \frac{|\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}|^{-\frac{1}{2}}}{\left[1 + \frac{1}{2a}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^{\mathrm{T}}\left[\frac{b}{a}(\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}})\right]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)\right]^{\frac{2a+n}{2}}}.
\end{aligned}$$

Hence,

$$\pi(\boldsymbol{\phi} \mid \mathbf{y}, \tau^2) \propto \frac{|\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}|^{-\frac{1}{2}}}{[b + \frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^{\mathrm{T}}[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)]^{a_\star}}\pi_b(\boldsymbol{\phi}). \tag{2.29}$$

In the derivation of (2.29) we have used the fact that

$$\mathbf{y} \mid \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N\left[\mathbf{F}\boldsymbol{\beta}_0, \sigma^2\left(\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}\right)\right].$$

This can be seen as follows. First note that $\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N(\mathbf{F}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma})$ and $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\mathbf{R})$. Hence,

$$\mathbf{y} - \mathbf{F}\boldsymbol{\beta} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N(\mathbf{0}_n, \sigma^2\boldsymbol{\Sigma}).$$

The right hand side does not depend on $\boldsymbol{\beta}$ and so $\mathbf{y} - \mathbf{F}\boldsymbol{\beta}$ is statistically independent of $\boldsymbol{\beta}$ given σ^2, $\boldsymbol{\phi}$ and τ^2. Moreover as $\mathbf{y} - \mathbf{F}\boldsymbol{\beta}$ is independent of $\boldsymbol{\beta}$, it is also independent of $\mathbf{F}\boldsymbol{\beta}$, which has conditional distribution $N(\mathbf{F}\boldsymbol{\beta}_0, \sigma^2\mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}})$. Hence, given σ^2, $\boldsymbol{\phi}$ and

τ^2,

$$\mathbf{y} = (\mathbf{y} - \mathbf{F}\boldsymbol{\beta}) + \mathbf{F}\boldsymbol{\beta}$$

is a sum of two independent multivariate normal random variables, and so is a multivariate normal with the claimed mean and variance.

Case 2: $\boldsymbol{\phi}$ and τ^2 unknown

We assign a normal-inverse-gamma prior distribution to $(\boldsymbol{\beta}, \sigma^2)$. We consider proper prior densities for $\boldsymbol{\phi}$ and τ^2, which we assume independent of $\boldsymbol{\beta}$ and σ^2:

$$\pi_b(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2) = \pi_b(\boldsymbol{\beta}, \sigma^2)\pi_b(\boldsymbol{\phi})\pi_b(\tau^2).$$

By Bayes' theorem, the marginal posterior density for $\boldsymbol{\phi}$ and τ^2 satisfies:

$$\pi(\boldsymbol{\phi}, \tau^2 \mid \mathbf{y}) \propto \pi(\mathbf{y} \mid \boldsymbol{\phi}, \tau^2)\pi_b(\boldsymbol{\phi}, \tau^2).$$

Similar to the previous case we have that,

$$\pi(\boldsymbol{\phi}, \tau^2 \mid \mathbf{y}) \propto \frac{|\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}|^{-\frac{1}{2}}}{[b + \frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^{\mathrm{T}}[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}\mathbf{F}^{\mathrm{T}}]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)]^{a_\star}}\pi_b(\boldsymbol{\phi})\pi_b(\tau^2). \tag{2.30}$$

In both cases the marginal posterior densities (2.29) and (2.30) are not standard distributions and hence the posterior predictive distribution $\pi(\tilde{y} \mid \mathbf{y})$ does not have an analytical form. In the Bayesian framework, prediction is based on the predictive distribution given by:

$$\begin{aligned}
\text{Case 1: } \pi(\tilde{y} \mid \mathbf{y}) &= \int_{\Phi} \pi(\tilde{y} \mid \mathbf{y}, \boldsymbol{\phi}, \tau^2)\pi(\boldsymbol{\phi} \mid \mathbf{y}, \tau^2)d\boldsymbol{\phi}, \\
\text{Case 2: } \pi(\tilde{y} \mid \mathbf{y}) &= \int_0^\infty \int_{\Phi} \pi(\tilde{y} \mid \mathbf{y}, \boldsymbol{\phi}, \tau^2)\pi(\boldsymbol{\phi}, \tau^2 \mid \mathbf{y})d\boldsymbol{\phi}\, d\tau^2.
\end{aligned}$$

Another problem, which is similar to the above, occurs for the marginal posterior density $\pi(\boldsymbol{\beta} \mid \mathbf{y})$, which cannot be expressed analytically:

$$\begin{aligned}
\pi(\boldsymbol{\beta} \mid \mathbf{y}) &= \int_0^\infty \int_{\Phi} \int_0^\infty \pi_a(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \mid \mathbf{y})d\sigma^2\, d\boldsymbol{\phi}\, d\tau^2 \\
&= \int_0^\infty \int_{\Phi} \pi(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\phi}, \tau^2)\pi(\boldsymbol{\phi}, \tau^2 \mid \mathbf{y})d\boldsymbol{\phi}\, d\tau^2.
\end{aligned}$$

These integrals are unavailable in closed form and hence numerical evaluation is required. We employ sampling techniques based on Markov chain Monte Carlo methods, which are overviewed in the next section. Alternatively, $\boldsymbol{\phi}$ can be replaced by a point estimate e.g. the posterior mode or maximum likelihood estimate (MLE). In Section 6.4 we use the MLE plug-in approach (see for example Bayarri et al., 2007b).

2.4 Markov chain Monte Carlo

When performing Bayesian inference, the aim is to compute the joint posterior distribution for a set of random variables. However, this is not always easy because it often requires the calculation of integrals that are unavailable in closed form. In cases like this, we can use sampling techniques based on Markov chain Monte Carlo (MCMC) methods, see for example Gelman et al. (2013, Chapter 11). MCMC methods construct a Markov chain[3] $\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots$, with steady state distribution equal to the posterior density, $\pi_a(\boldsymbol{\psi} \mid \mathbf{y})$, of interest. The empirical distribution of the first $\tilde{M}$ values, $\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(\tilde{M})}$, will then converge to $\pi_a(\boldsymbol{\psi} \mid \mathbf{y})$ as $\tilde{M} \to \infty$. A good approximation to $\pi_a(\boldsymbol{\psi} \mid \mathbf{y})$ is obtained by running the chain for large finite $\tilde{M}$. There are general procedures for constructing Markov chains to match any $\pi_a(\boldsymbol{\psi} \mid \mathbf{y})$.

The chain is initialised with starting values, $\boldsymbol{\psi}^{(0)}$. The Markov property specifies that the distribution of $\boldsymbol{\psi}^{(i+1)}$ given all previous draws, $\boldsymbol{\psi}^{(i+1)} \mid \boldsymbol{\psi}^{(i)}, \boldsymbol{\psi}^{(i-1)}, \ldots$, depends only on the most recent value drawn $\boldsymbol{\psi}^{(i)}$.

2.4.1 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm (Metropolis et al., 1953 and Hastings, 1970) enables sampling from an essentially arbitrary target distribution. It proceeds as follows.

The first step is to initialise the chain with starting values $\boldsymbol{\psi}^{(0)}$ for the random variables. Let the current state of the chain be $\boldsymbol{\psi}^{(i)}$. The main loop of the algorithm consists of three components: generate a sample from a proposal density q, compute the acceptance probability, α, and accept or reject the candidate sample with probability α, or $1 - \alpha$, respectively (see Algorithm 1). In practice, we must allow some burn-in[4] time to let the empirical distribution of the chain become close enough to the target distribution.

Note that because the posterior density appears in both the numerator and the denominator of the acceptance ratio α as given in Algorithm 1, we only require an expression for the unnormalised density.

The Metropolis-Hastings algorithm is a general approach for sampling from a target density, in our case $\pi_a(\boldsymbol{\psi} \mid \mathbf{y})$. However, it requires the specification of a proposal density, which must be chosen carefully. The acceptance rates, which depend on the proposal distribution, must be continuously monitored for low and high values. The efficiency depends crucially on the scaling of the proposal density. If the variance of the proposal is too small, the acceptance rate will be high but the Markov chain will

[3] We construct a sequence of random variables $\{\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots\}$, such that at each time $t \geq 0$, the next state $\boldsymbol{\psi}^{(t+1)}$ is sampled from a distribution $f(\boldsymbol{\psi}^{(t+1)} | \boldsymbol{\psi}^{(t)})$ which depends only on the current state of the chain $\boldsymbol{\psi}^{(t)}$. This sequence is called *Markov chain*, and $f(\cdot \mid \cdot)$ is called the *transition kernel* of the chain (Gilks et al., 1996, Chapter 1).

[4] Burn-in is the procedure of throwing away some iterations at the beginning of an MCMC run. Inference is based on the assumption that the distribution of the simulated values $\boldsymbol{\psi}^{(i)}$, for large enough i, are close to the target distribution, $\pi_a(\boldsymbol{\psi} \mid \mathbf{y})$.

Algorithm 1: Metropolis-Hastings algorithm

```
Choose initial values for the chain, $\boldsymbol{\psi}^{(0)}$;
for $i = 1, 2, \dots, \tilde{M}$ do
    Propose: $\boldsymbol{\psi}^* \sim q(\boldsymbol{\psi}^* \mid \boldsymbol{\psi}^{(i-1)})$
    Acceptance probability: $\alpha(\boldsymbol{\psi}^* \mid \boldsymbol{\psi}^{(i-1)}) = \min\left\{1, \frac{q(\boldsymbol{\psi}^{(i-1)}|\boldsymbol{\psi}^*)\pi_a(\boldsymbol{\psi}^*|\mathbf{y})}{q(\boldsymbol{\psi}^*|\boldsymbol{\psi}^{(i-1)})\pi_a(\boldsymbol{\psi}^{(i-1)}|\mathbf{y})}\right\}$
    Sample $u \sim$ Uniform$(0, 1)$
    if $u < \alpha$ then
        Accept the proposal: $\boldsymbol{\psi}^{(i)} \leftarrow \boldsymbol{\psi}^*$
    else
        Reject the proposal: $\boldsymbol{\psi}^{(i)} \leftarrow \boldsymbol{\psi}^{(i-1)}$
```

converge slowly since all its increments will be small. Conversely, if the variance is too large, the Metropolis-Hastings algorithm will reject too high a proportion of its proposed moves and the chain will become 'stuck' at particular values of $\boldsymbol{\psi}$. A high acceptance rate does not necessarily indicate that the algorithm is behaving satisfactorily. Also, a low acceptance rate does not mean that the chain explores the entire support of the target distribution. Roberts et al. (1997) recommended for random walk algorithms (Gilks et al., 1996) the use of distributions with acceptance rates close to $\frac{1}{4}$ for models of high dimension and equal to $\frac{1}{2}$ for the models of dimension 1 or 2.

As in many MCMC methods, the draws are regarded as a sample from the target distribution only after the chain has passed the burn-in time and the effect of the fixed starting value has become so small that it can be ignored. The convergence occurs under mild regularity conditions such as irreducibility[5] and aperiodicity.[6]

2.5 Summary

In this chapter we introduced a number of key concepts for Gaussian process models and their role in calibration. We also reviewed the Bayesian approach which will be used throughout this thesis. For conjugate prior distributions for the trend and variance parameters, we define the posterior and predictive distributions, when the correlation parameters and the nugget are either known or unknown. Last, we introduced MCMC methods that we will employ in the next chapter in order to evaluate numerically intractable integrals.

[5] Given suitable technical conditions, for each x there exists a positive integer n such that $P^n(x, A) > 0$, where $P^n(x, A) = P[X_n \in A \mid X_0 = x]$ and A is any measurable set. In this case we say the chain is irreducible.

[6] If the transition kernel (see Footnote 3) has density $f(\cdot \mid \cdot)$, a sufficient condition for *aperiodicity* is that $f(\cdot \mid x)$ is positive in a neighbourhood of x, since the chain can then remain in this neighbourhood for an arbitrary number of times before visiting another set A.

Chapter 3

The impact of choice of design

In this chapter we illustrate, through examples, the impact of the choice of design in calibration. In the first section we introduce the Michaelis-Menten model, the estimation of which we treat as a Bayesian calibration problem. We start by discussing the choice of prior distributions and write down the posterior distribution for the model parameters. We then assess the differences in posterior inference from a small number of designs.

In the second example of this chapter, we assume the simplest case of the calibration model (1.1) where there is no random error, with the simulator being the Michaelis-Menten model with fixed calibration parameters θ_1^p and θ_2^p. We find the posterior distribution for the discrepancy function, $\delta_{\boldsymbol{\theta}^p}(x)$, given data from a simulated physical process, with $\delta_{\boldsymbol{\theta}^p}(x)$ being a sinusoidal function. Lastly, we consider different designs in order to see how the choice of the design affects uncertainty in the predictions.

Taking a step further towards the analysis of the calibration model (1.1), in the third example we assume that the simulator is again known but with unknown calibration parameters $\boldsymbol{\theta}^p$. Again the simulator is the Michaelis-Menten model. We also assume a Gaussian process prior for the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$ and implement Bayesian inference assuming prior distributions for the unknown calibration and correlation parameters. We then illustrate the impact of choice of the design by finding the posterior predictive distributions for different designs.

3.1 Michaelis-Menten model

The Michaelis-Menten model (Michaelis and Menten, 1913) is a nonlinear model that has been used in many applications (Bates and Watts, 1988, Chapters 2 and 3), for example in modelling enzyme kinetic data (Cornish-Bowden, 1995). The Michaelis-Menten model is also used in compartmental models to model the rate of flow from one compartment to another.

We will consider the Michaelis-Menten equation of the form:

$$\eta(x, \boldsymbol{\theta}) = \frac{\theta_1 x}{\theta_2 + x}, \quad x \in [0, \infty), \quad \theta_1, \theta_2 > 0, \tag{3.1}$$

with θ_2 known as the Michaelis-Menten constant. In the enzyme kinetic context, $\eta(x, \boldsymbol{\theta})$ is the reaction velocity, θ_1 is the maximum velocity of the reaction, x is the concentration of a substrate, and θ_2 is the half-saturation constant; that is, the value of x where $\eta(x, \boldsymbol{\theta})$ achieves half its maximum value (Lòpez-Fidalgo and Wong, 2002).

There are several methods proposed in the literature in order to estimate the parameters $\boldsymbol{\theta}$ of the Michaelis-Menten model. The majority of these methods are based on nonlinear least squares or on transformations of (3.1) to a linear relationship and application of linear regression techniques (Bliss and James, 1966; Glick et al., 1979; Currie, 1982). Raaijmaakers (1987) gave arguments supporting the use of maximum likelihood for the estimation of these parameters in the Michaelis-Menten model.

Designing experiments for the Michaelis-Menten equation has also been studied in the literature (see for example, Duggleby and Clarke, 1991; Boer et al., 2000). To overcome the dependence on the values of the unknown model parameters of locally optimal designs, Song and Wong (1998) proposed Bayesian optimal designs. They constructed Bayesian D-optimal designs (see Section 4.1.1) when the variance of the response depends on the independent variable. A Bayesian approach is applied to find an optimal design, by taking into account the prior information about the variance structure, and solve the problem of this dependence being only partially known. Dette and Biedermann (2003) found maximin D-optimal designs; that is designs that maximise the minimum of the D-efficiencies[1] over a certain interval for the nonlinear parameter.

In this thesis, we will concentrate on estimating the unknown parameters by taking a Bayesian decision-theoretic approach (see Section 4.1). Initially, we assume the statistical model:

$$y_i = \eta(x_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \dots n. \tag{3.2}$$

Here, y_i is the response measured at a specific point x_i, and ε_i is the random observational error, where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ independently.

The likelihood function is given by:

$$\pi_l(\mathbf{y} \mid \boldsymbol{\theta}, \sigma_\varepsilon^2) = (2\pi\sigma_\varepsilon^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{n} \left[\left(y_i - \frac{\theta_1 x_i}{\theta_2 + x_i} \right)^2 \right] \right\}, \tag{3.3}$$

with $\mathbf{y} = (y_1, \dots, y_n)^{\mathrm{T}}$ the vector of responses.

[1] The D-efficiency of a design ξ is

$$\mathrm{eff}(\boldsymbol{\theta}; \xi) = \left(\frac{\mid I(\boldsymbol{\theta}; \xi) \mid}{\mid I(\boldsymbol{\theta}; \xi^\star) \mid} \right)^{\frac{1}{p_\theta}},$$

where $I(\boldsymbol{\theta}; \xi)$ is the Fisher information matrix, $\xi^\star$ is the D-optimal design and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{p_\theta})^{\mathrm{T}} \in \boldsymbol{\Theta}$.

3.1.1 Example

We assume the statistical model (3.2), where $\eta(x, \boldsymbol{\theta})$ is the Michaelis-Menten model (3.1). The aim of this section is to demonstrate the impact of choice of design for the Michaelis-Menten equation. We generate six different designs, each with seven points, and estimate the unknown parameters θ_1 and θ_2 for each design. To generate the data we assume $\theta_1 = 0.15$ and $\theta_2 = 50$ (values taken from Berthouex and Brown, 2002, Chapter 35).

Prior specification

We assume uniform prior distributions for the unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)^{\mathrm{T}}$,

$$\theta_1 \sim \mathrm{Unif}[a_1, b_1] \text{ and } \theta_2 \sim \mathrm{Unif}[a_2, b_2],\ b_1 > a_1 > 0, \text{ and, } b_2 > a_2 > 0,$$

with $a_1 = a_2 = 0$ and $b_1 = b_2 = 200$. These priors where chosen to have support that includes point estimates from a literature data set [2].

We assume an inverse-gamma prior distribution for the error variance, σ_ε^2:

$$\sigma_\varepsilon^2 \sim \mathrm{IG}(a, b),\ a, b > 0,$$

with

$$a = \frac{n}{2} \text{ and } b = \frac{n}{2} S_0^2,$$

where S_0^2 is the mean of the squared residuals from a nonlinear least squares fit of the Michaelis-Menten equation to literature data (see Footnote 2) and n is the number of design points. For the given data we have $n = 7$ and $S_0^2 = 1.24 \times 10^{-4}$.

The joint prior density is given by,

$$\pi_b(\boldsymbol{\theta}, \sigma_\varepsilon^2) \propto \frac{I(a_1 < \theta_1 < b_1)}{b_1 - a_1} \frac{I(a_2 < \theta_2 < b_2)}{b_2 - a_2} (\sigma_\varepsilon^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma_\varepsilon^2}\right\}. \tag{3.4}$$

Posterior

The likelihood function for the model (3.2) is given by (3.3). Using Bayes' Theorem (2.16), and the joint prior density (3.4), the posterior density is proportional to:

$$\begin{aligned}\pi_a(\boldsymbol{\theta}, \sigma_\varepsilon^2 \mid \mathbf{y}) &\propto \pi_l(\mathbf{y} \mid \boldsymbol{\theta}, \sigma_\varepsilon^2)\pi_b(\boldsymbol{\theta}, \sigma_\varepsilon^2) \\ &\propto \frac{I(a_1 < \theta_1 < b_1)}{b_1 - a_1} \frac{I(a_2 < \theta_2 < b_2)}{b_2 - a_2} \left(\frac{1}{\sigma_\varepsilon^2}\right)^{a+1+\frac{n}{2}}\end{aligned}$$

[2] The data set was taken from Berthouex and Brown (2002, Chapter 35),

$$(x_1, \ldots, x_n)^{\mathrm{T}} = (28, 55, 83, 110, 138, 225, 375)^{\mathrm{T}}$$
$$\mathbf{y} = (0.053, 0.06, 0.112, 0.105, 0.099, 0.122, 0.125)^{\mathrm{T}}.$$

$$\times \exp\left\{-\frac{1}{\sigma_\varepsilon^2}\left[\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2+b\right]\right\}. \tag{3.5}$$

The marginal posterior distribution for $\boldsymbol{\theta}$ has density:

$$\begin{aligned}\pi_M(\boldsymbol{\theta}\mid\mathbf{y}) &= \int_0^\infty \pi_a(\boldsymbol{\theta},\sigma_\varepsilon^2\mid\mathbf{y})d\sigma_\varepsilon^2\\ &= \frac{I(a_1<\theta_1<b_1)I(a_2<\theta_2<b_2)}{(b_1-a_1)(b_2-a_2)}\\ &\quad\times\int_0^\infty\left(\frac{1}{\sigma_\varepsilon^2}\right)^{\frac{2a+n}{2}+1}\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left[\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2+2b\right]\right\}d\sigma_\varepsilon^2\\ &\propto \frac{I(a_1<\theta_1<b_1)I(a_2<\theta_2<b_2)}{(b_1-a_1)(b_2-a_2)}\left[1+\frac{\sum_{i=1}^{n}\left(y_i-\frac{\theta_1x_i}{\theta_2+x_i}\right)^2}{2b}\right]^{-(a+\frac{n}{2})}.\end{aligned}\tag{3.6}$$

MCMC implementation

In order to estimate the unknown parameters $\boldsymbol{\theta}$ of the Michaelis-Menten model, we will use the Metropolis-Hastings algorithm (Algorithm 1 in Section 2.4). We use the estimated parameters from the nonlinear least squares fit of the Michaelis-Menten equation to the literature data (see Footnote 2) to initiate the chain. That is, $(\theta_1^{(0)},\theta_2^{(0)})^{\mathrm{T}} = (0.153, 53.665)^{\mathrm{T}}$.

At each MCMC step we propose values for $\boldsymbol{\theta}$ from a Normal distribution,

$$\boldsymbol{\theta}^* \sim N(\boldsymbol{\theta}^{(i-1)}, c^2\mathbf{C}), \tag{3.7}$$

centred at the current iteration with

$$\mathbf{C} = \begin{bmatrix} 0.0009 & 0.8333\\ 0.8333 & 843.2894\end{bmatrix},$$

the covariance matrix for the estimators of the two parameters θ_1 and θ_2 from the nonlinear least squares fit of the Michaelis-Menten equation, scaled by a value c^2 (Laine, 2008). Following Gelman et al. (2013, Chapter 12) the most efficient proposal distribution has scale,

$$c \approx \frac{2.4}{\sqrt{p_\theta}},$$

where p_θ is the number of parameters. Efficiency is defined in terms of the effective sample size,

$$n_{\text{eff}} = \frac{n_{it}}{1+2\sum_{t=1}^{\infty}\rho_t},$$

(Gelman et al., 2013, page 286) where ρ_t is the autocorrelation of the sequence $\boldsymbol{\theta}$ at lag t and n_{it} is the number of iterations after we have discarded some samples as a

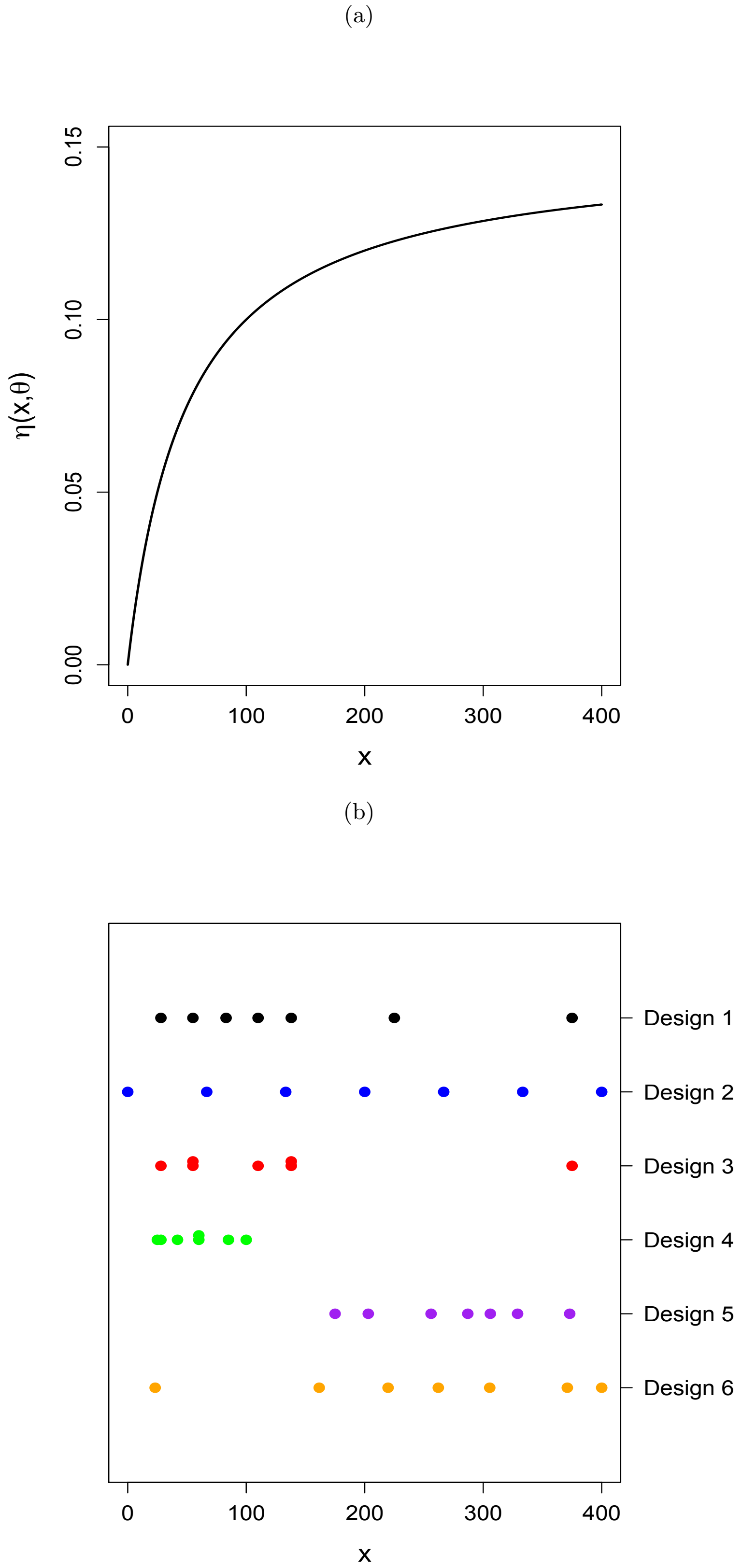


Figure 3.1: (a) The true Michaelis-Menten function; (b) The six different designs;

burn-in. The quantity n_{eff} gives the number of independent samples from that posterior distribution that would yield the same Monte Carlo error as the autocorrelated Markov chain.

Figure 3.1 (a) shows the true mean response we have assumed to generate the data for each of the six designs given in Figure 3.1 (b).

- Design 1 (taken from Berthouex and Brown, 2002, Chapter 35) is an ad-hoc design, where most of the points are concentrated where the true expected response is changing most quickly and some points are at the stationary part of the true expected response.

- Design 2 is a space-filling design (see Section 1.2.2).

- Design 3 consists of five of the seven points of Design 1, with two of the points repeated twice.

- Design 4 has all the points concentrated at the first half of the design space where the true expected response is changing most quickly.

- Design 5 has all the points concentrated at the part of the design space within the true expected response is stationary.

- Design 6 is again an ad-hoc design but now most of the points are concentrated at the stationary part of the true expected response, and fewer points where the true expected response is changing fastest.

For each of these designs we simulated a response vector $\mathbf{y}$ using the parameter values $(\theta_1, \theta_2)^{\mathrm{T}} = (0.15, 50)^{\mathrm{T}}$ and $\sigma_\varepsilon^2 = 0.05$. The posterior distribution for the simulated data was then calculated using the Metropolis-Hastings algorithm with proposal distribution (3.7). In each case, a chain length of $\tilde{M} = 20,000$ was used with the first $5,000$ iterations discarded as burn-in.

Figure 3.2 and Figure 3.3 are trace-plots which show the values each parameter took during the runtime of the chain. Inspecting these plots we notice that for most designs the chains converge to distributions centred on the true values of $\boldsymbol{\theta}$ ($\theta_1 = 0.15$, $\theta_2 = 50$).

Figure 3.4 shows the approximate posterior densities of θ_1 and θ_2 for each of the designs. All the posterior densities are centred at the true values (or very close to the true values) of θ_1 and θ_2. Design 4, as expected from Figures 3.2 and 3.3, has higher posterior variance for estimating θ_1 compared to the other five designs. Design 4 and Design 5 have higher posterior variance for estimating θ_2 compared to the other four designs (the posterior standard deviations for each of the designs can be found in Table 3.1). Hence we can conclude that the choice of design is important, and that Designs 4 and 5 are not very good designs for estimating the unknown parameters for this model.

Design	**Posterior st. dev.**	
	θ_1	θ_2
Design 1 ●	0.018	22.316
Design 2 ●	0.013	22.335
Design 3 ●	0.013	14.373
Design 4 ●	0.049	38.461
Design 5 ●	0.021	44.509
Design 6 ●	0.012	23.496

Table 3.1: Posterior standard deviation of θ_1 and θ_2

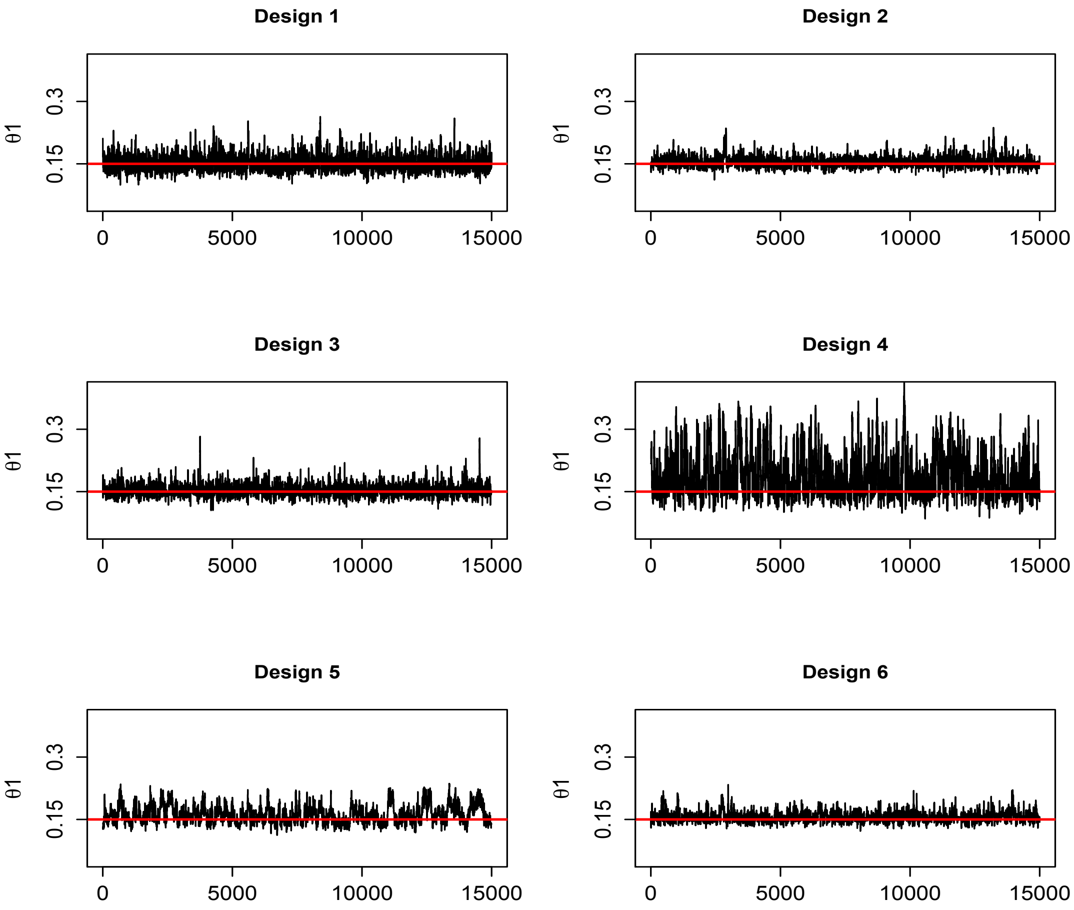


Figure 3.2: Trace plots for MCMC samples of θ_1 for the six designs and the Michaelis-Menten model and the true value of $\theta_1 = 0.15$ (red line)

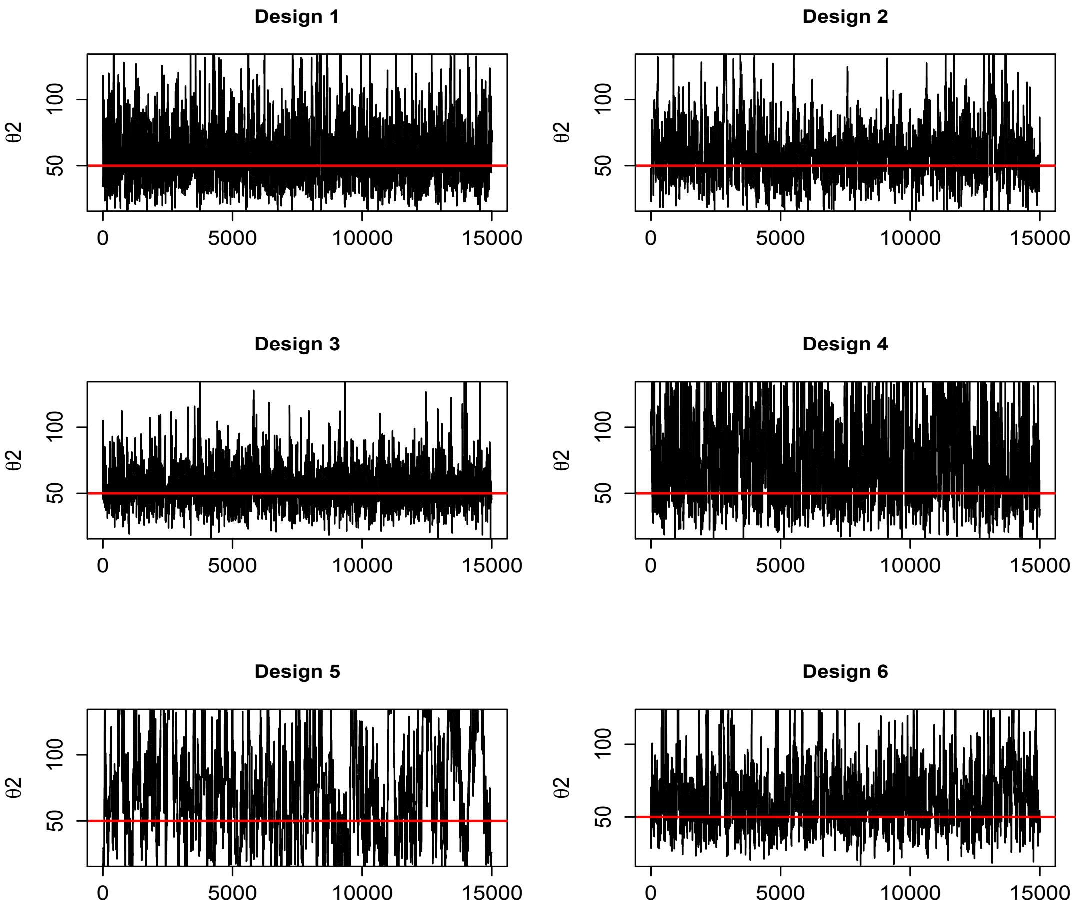
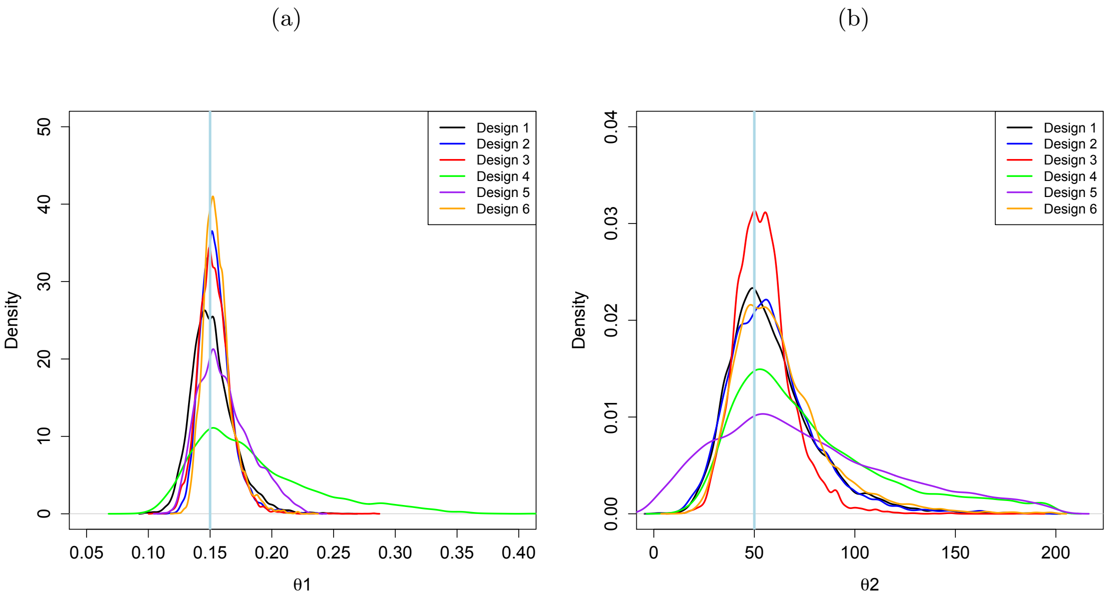


Figure 3.3: Trace plots for MCMC samples of θ_2 for the six designs and the Michaelis-Menten model and the true value of $\theta_2 = 50$ (red line)


(a)

(b)

Figure 3.4: (a) Approximate posterior densities for θ_1 for each design; (b) Approximate posterior densities for θ_2 for each design; the vertical lines in each plot are the true values of the unknown parameters ($\theta_1 = 0.15$, $\theta_2 = 50$)

3.2 Calibration model with a known simulator function

We present two simple examples to illustrate Bayesian inference for the Gaussian process within a calibration problem. We find the posterior distribution for $\delta_{\boldsymbol{\theta}^p}(\cdot)$ from model (1.1) given data from a simulated physical process.

3.2.1 Example 1: known simulator parameters θ^p and $\sigma_\varepsilon^2 = 0$

In this section we assume the calibration model (1.1) with simulator, $\eta(x, \boldsymbol{\theta})$, the Michaelis-Menten model described in Section 3.1. We assume the simplest case where the 'true' parameters $\boldsymbol{\theta}^p$ of the simulator are known and fixed at $\theta_1^p = 15$ and $\theta_2^p = 50$ (see Figure 3.5). We also assume that there is no random error, i.e. $\sigma_\varepsilon^2 = 0$. For simplicity we fix the regression parameter at $\rho = 1$. We can rewrite (1.1) as,

$$y_i - \eta(x_i, \boldsymbol{\theta}^p) = \delta_{\boldsymbol{\theta}^p}(x_i), \quad i = 1, \ldots, n.$$

We assume a Gaussian process prior on $\delta_{\boldsymbol{\theta}^p}(x_i)$ such that:

$$\boldsymbol{\delta}_{\boldsymbol{\theta}^p} \sim N\left[\mathbf{0}_n, \sigma^2 \mathbf{K}(\phi)\right],$$

where $\boldsymbol{\delta}_{\boldsymbol{\theta}^p} = [\delta_{\boldsymbol{\theta}^p}(x_1), \ldots, \delta_{\boldsymbol{\theta}^p}(x_n)]^{\mathrm{T}}$ and $\mathbf{K}(\phi)$ is the correlation matrix with ijth entry $\mathbf{K}(\phi)_{ij} = \kappa(x_i, x_j; \phi)$, $i, j = 1, \ldots, n$. We assume a conjugate prior distribution for the Gaussian process variance with $\sigma^2 \sim \mathrm{IG}(3, 2)$. We choose $a = 3$ to ensure finite prior variance. See Appendix C.1.1 for samples from the prior distribution on $\delta_{\boldsymbol{\theta}^p}(\cdot)$.

We also assume the squared exponential correlation function given in (2.6),

$$\kappa(x, x'; \phi) = \exp\left[-\phi\left(x - x'\right)^2\right].$$

Lastly, for the correlation parameter ϕ, we assume the prior distribution $\phi \sim \mathrm{Exp}(\lambda_\phi)$, with rate $\lambda_\phi = 200$, which ensures $\phi > 0$. See Appendix C.1.1 for further discussion on the choice of prior. In the Metropolis-Hastings algorithm, the proposal distribution for ϕ will be a sliding window proposal (Gramacy and Lee, 2008; Yang and Rodríguez, 2013) of the form:

$$\phi^* \mid \phi^{(i-1)} \sim \mathrm{Unif}\left[\frac{1}{\lambda_0}\phi_{i-1}, \lambda_0 \phi_{i-1}\right], \; \lambda_0 > 0. \tag{3.8}$$

Such a proposal distribution forms a window around the current value $\phi^{(i-1)}$. The window width is controlled by λ_0, which is a tuning parameter. This parameter λ_0 is held fixed throughout the sampling; usually a sensitivity analysis must be implemented in order to tune λ_0 to obtain reasonable convergence.

When generating the data, we assume that the discrepancy function has the form $\delta_{\boldsymbol{\theta}^p}(x) = \nu_1 \sin(\nu_2 x)$. We divide this example into two cases. In the first case we assume

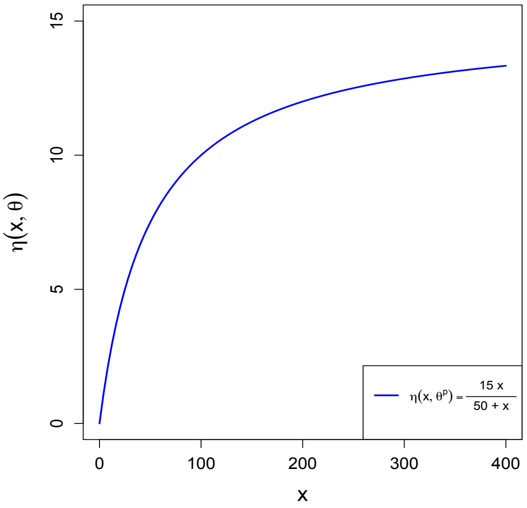


Figure 3.5: The Michaelis-Menten equation $\eta(x, \boldsymbol{\theta}^p) = \frac{15x}{50+x}$.

$\delta_{\boldsymbol{\theta}^p}(x) = 0.5\sin(0.05x)$ and in the second case we assume $\delta_{\boldsymbol{\theta}^p}(x) = 0.5\sin(0.1x)$.

Case 1: $\nu_1 = 0.5, \nu_2 = 0.05$

The assumed physical process, $\zeta(x)$, in this case has the form:

$$\zeta(x) = \frac{15x}{50+x} + 0.5\sin(0.05x).$$

In Figure 3.6 the blue solid line is the Michaelis-Menten equation and the green solid line is the physical process $\zeta(x)$. The simulated data are represented as black bullets.

To approximate the posterior distribution of ϕ, a Metropolis-Hastings algorithm is used to draw a sample $\{\phi_i\}_{i=1}^{\tilde{M}}$, $\tilde{M} = 150,000$, using the uniform proposal distribution (3.8). Convergence is assessed via diagnostic plots. For each value of ϕ from the chain, we calculate the predictive mean and variance of $\zeta(x)$ given by Equations (2.25) and (2.26) respectively, and generate samples from a t-distribution with this mean and variance. Then we sample realisations and calculate summaries of these. We find the median of the realisations and the 97.5% and 2.5% quantiles around the median to obtain the 95% credible interval for $\zeta(x)$.

In Figure 3.7 the red line is the posterior median of the Gaussian process model, the blue lines are the 95% probability bounds (see also Figure 3.10), the green line is the true model and the three black lines are three realisations from the Gaussian process posterior for $\zeta(x)$. Uncertainty is pinched to zero at the design points but as we move away from the points uncertainty increases. Also, we notice that the posterior median of the Gaussian process model over-smooths the true function.

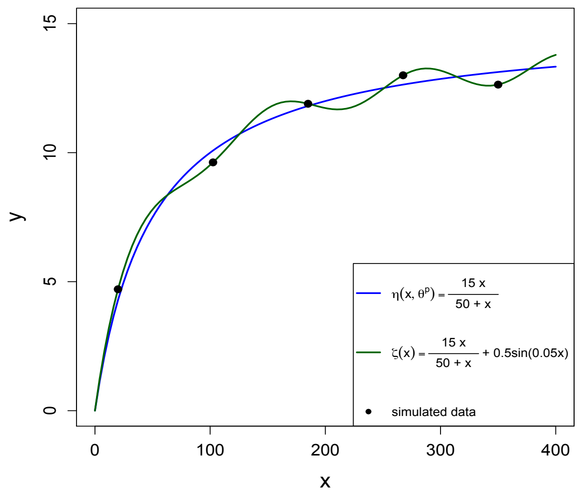


Figure 3.6: The Michaelis-Menten equation $\eta(x, \boldsymbol{\theta}^p) = \frac{15x}{50+x}$ (blue line), the assumed physical process, $\zeta(x) = \frac{15x}{50+x} + 0.5\sin(0.05x)$ (green line) and the simulated data (black bullets)

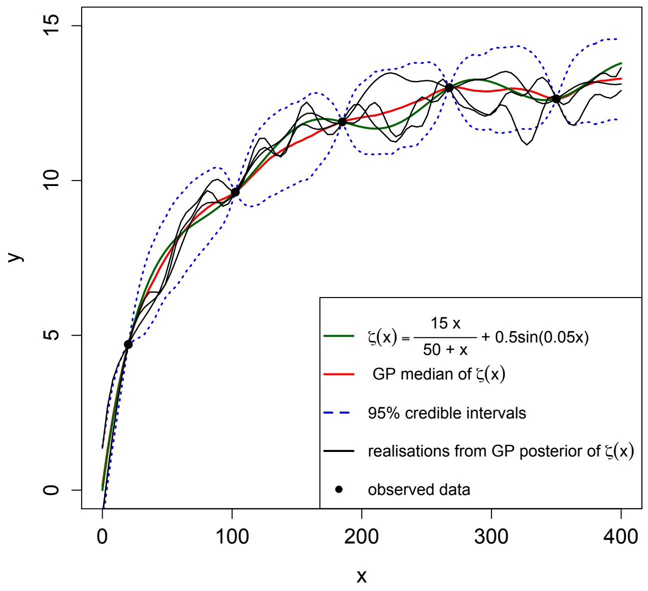


Figure 3.7: Posterior median for $\zeta(x)$ (red line); pointwise 95% credible intervals (blue lines); true model (green line); three realisations from the GP model (black lines)

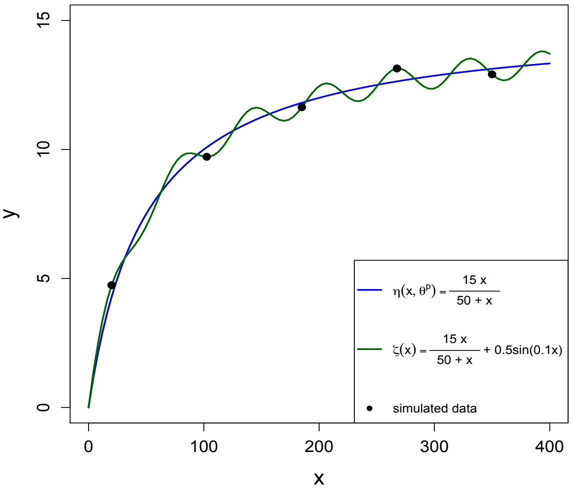


Figure 3.8: The Michaelis-Menten equation $\eta(x, \boldsymbol{\theta}^p) = \frac{15x}{50+x}$ (blue line), the true model $\zeta(x) = \frac{15x}{50+x} + 0.5\sin(0.1x)$ (green line) and the simulated data (black bullets)

Case 2: $\nu_1 = 0.5, \nu_2 = 0.1$

The assumed physical process, $\zeta(x)$, in this case has the form:

$$\zeta(x) = \frac{15x}{50+x} + 0.5\sin(0.1x).$$

In Figure 3.8 the blue solid line is the Michaelis-Menten equation and the green solid line is the assumed physical process, $\zeta(x)$. The simulated data are represented as black bullets. The difference here is that we have a more complex function than before, which is more 'wiggly' due to the different frequency of the sinusoidal discrepancy term.

As in the previous case, in Figure 3.9 the red line is the posterior median of the Gaussian process model, the blue lines are the pointwise 95% credible intervals and the green line is the true model (see also Figure 3.10). The three black lines are realisations from the Gaussian process posterior for $\zeta(x)$. Again, uncertainty is pinched to zero at the design points and increases as we move away from the points. As before, the posterior median of the Gaussian process model over-smooths the true function. However, in both cases the credible intervals reflect this lack of knowledge.

In Figures 3.10 (a) and (b) the red line is the posterior median of the Gaussian process model, for Case 1 and Case 2, respectively. In both cases, the curve is a smooth line and passes through the five data points. The blue lines are the pointwise 95% credible intervals and make it clear that we have relatively little information away from these five points. The uncertainty is pinched to zero at the five design points because we set the error variance $\sigma_\varepsilon^2 = 0$ in the statistical model. The green line is the true model. Figure 3.10 (b) shows that the red line over-smooths the true function. Due to the small number of data points, the credible intervals do not clearly show the higher frequency

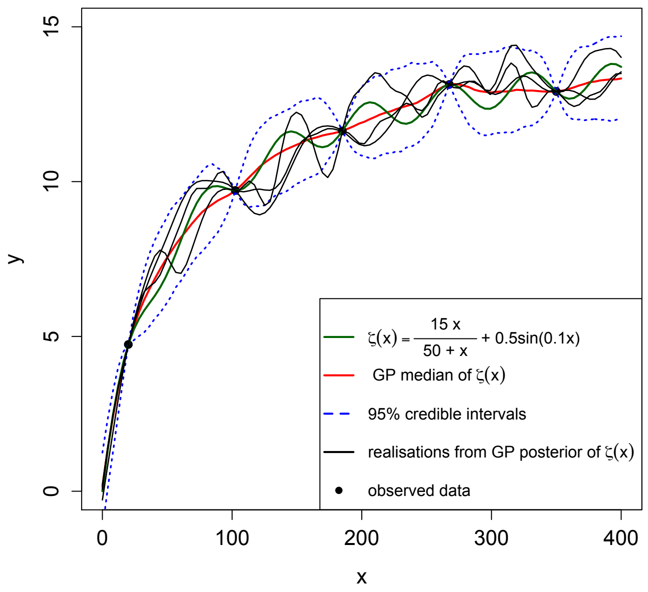
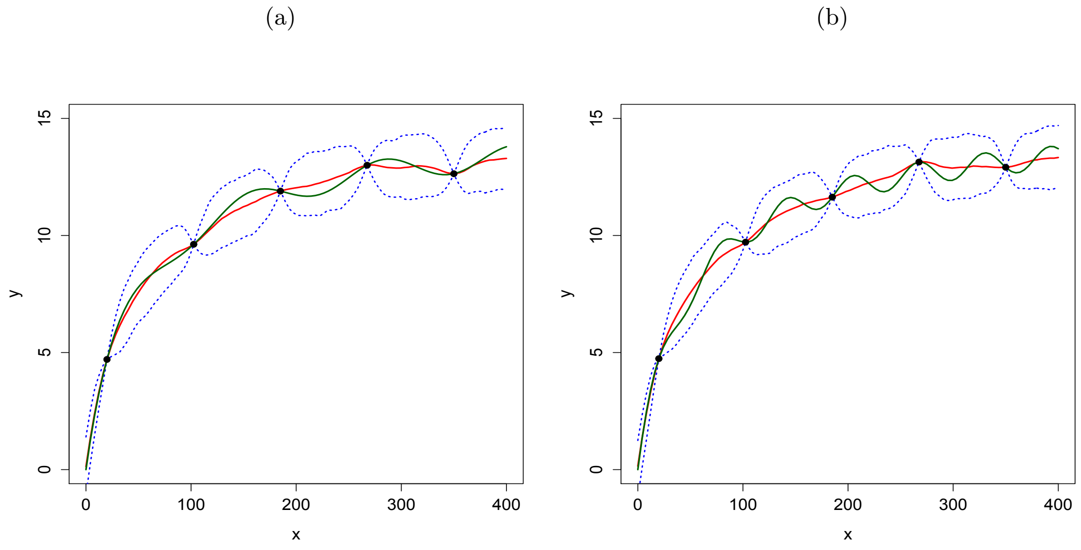


Figure 3.9: Posterior median for $\zeta(x)$ (red line); pointwise 95% credible intervals (blue lines); true model (green line); three realisations from the GP model (black lines)



(a)

(b)

Figure 3.10: Posterior median for $\zeta(x)$ (red line); pointwise 95% credible intervals (blue lines); true model (green line) for (a) Case 1: $\zeta(x) = \frac{15x}{50+x} + 0.5\sin(0.05x)$; (b) Case 2: $\zeta(x) = \frac{15x}{50+x} + 0.5\sin(0.1x)$

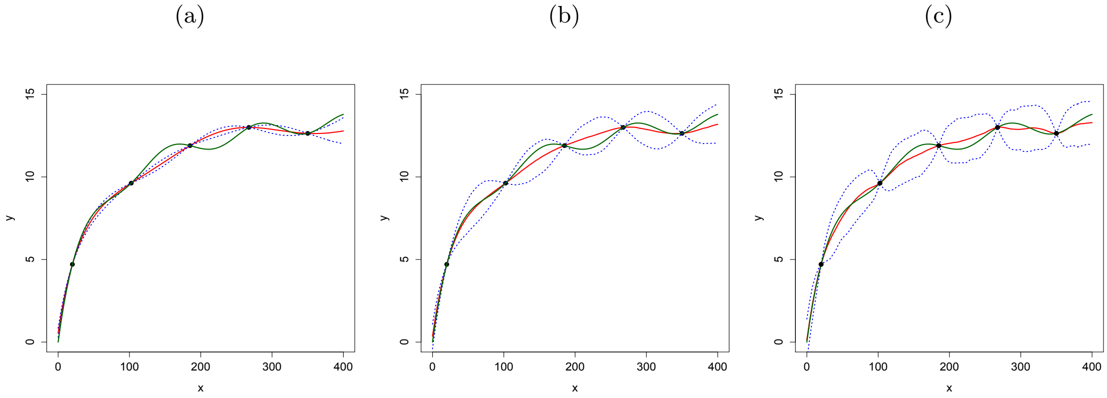


Figure 3.11: Posterior median for $\zeta(x)$ (red line), pointwise 95% credible intervals (blue line) and the true model (green line) for Case 1: $\zeta(x) = \frac{15x}{50+x} + 0.5\sin(0.05x)$ and, (a) a fixed value of the correlation parameter, $\phi = 10^{-4}$; (b) Plug-in approach using MLE for the correlation parameter (c) the correlation parameter is estimated using MCMC

of the discrepancy function in Case 2.

The correlation parameter

The correlation parameter ϕ influences how the posterior prediction changes due to knowing the response at the design points. In the previous section we estimated the values of the correlation parameter by using MCMC methods (Section 2.4) and more specifically the Metropolis-Hastings algorithm. Here we show how the results are changing by estimating ϕ using MCMC, using a fixed arbitrary value and using the maximum likelihood estimate (MLE).

In Figure 3.11 (a) we show how the Gaussian process model approximates the true model $\zeta(x) = \frac{15x}{50+x} + 0.5\sin(0.05x)$ when using a fixed value for the correlation parameter, $\phi = 10^{-4}$. Although uncertainty is negligible we can notice that there are regions where the Gaussian process median approximates the true model very poorly and the GP model is overconfident in an incorrect prediction. This happens because we have chosen an inappropriate value of ϕ that assumes that the correlation between data points decays too slowly with the difference in x.

The use of the MLE or estimating the correlation parameter ϕ using MCMC, as shown in Figures 3.11 (b) and (c), increases the uncertainty. However, Figures 3.11 (b) and (c) show a more realistic representation of the true function, while in Figure 3.11 (a) the prediction intervals consistently fail to include the true function.

Design comparison

As we mentioned in Section 1.2 the choice of the design points is an important part of the calibration problem. We illustrate an example in which we change the design

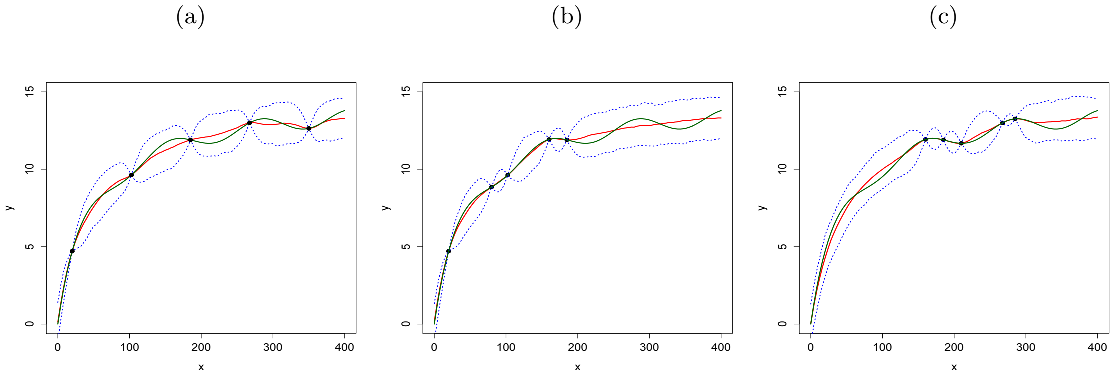


Figure 3.12: Posterior median for $\zeta(x)$ (red line), pointwise 95% credible intervals (blue lines), true model (green line) and the simulated data (black bullets) as we move some design points from the initial equally spaced design, for Case 1: $\zeta(x) = \frac{15x}{50+x} + 0.5\sin(0.05x)$

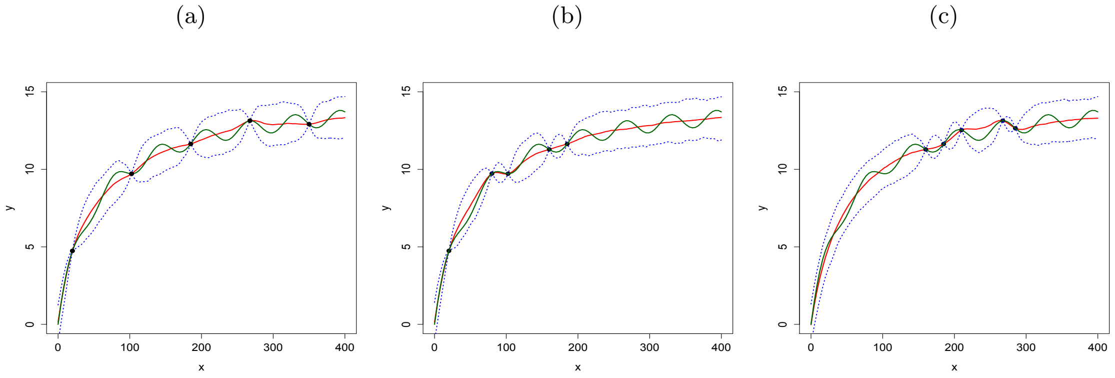


Figure 3.13: Posterior median for $\zeta(x)$ (red line), pointwise 95% credible intervals (blue lines), true model (green line) and the simulated data (black bullets) as we move some design points from the initial equally spaced design, for Case 2: $\zeta(x) = \frac{15x}{50+x} + 0.5\sin(0.1x)$

points from the initial design used in Case 1 and Case 2.

For the discrepancy functions given in Cases 1 and 2, we illustrate how the predictive uncertainty changes with the design. Again in these figures, the red line is the median of the Gaussian process model, the blue dashed lines are pointwise 95% credible intervals, the green line is the true model and our chosen design points are shown as black bullets. The three plots in each of Figures 3.12 and 3.13 show a sequence in which the design points are changed.

We move some design points, from the initial equally spaced design given in Figures 3.6 and 3.8, closer together in order to learn how quickly the curve is changing and hence learn more about the correlation parameter ϕ. As the space between the points decreases our uncertainty about $\zeta(x)$ between these points decreases as well, whereas as the space between the points increases our uncertainty about $\zeta(x)$ between these points increases.

We notice in Figure 3.12 (b) and (c) for Case 1 that the median of the Gaussian process model adapts to shape of the true model quickly in regions with a high density of design points, and the uncertainty is reduced. In this case, the true function does not change very quickly and the median of the Gaussian process can learn the shape of the true function. Similarly, in Figure 3.13 (b) and (c) for Case 2 we notice that the uncertainty is decreased in regions with high density of design points, however the median of the Gaussian process is still different from the true model. In Case 2 we have a more complex function which is more 'wiggly' due to the different frequency of the sinusoidal discrepancy term.

3.2.2 Example 2: unknown simulator parameters θ^p and $\sigma_\varepsilon^2 \neq 0$

In this section we again assume the calibration model (1.1) with simulator, $\eta(x, \boldsymbol{\theta})$, the Michaelis-Menten equation described in Section 3.1. However, now we assume that the parameters of the simulator, $\boldsymbol{\theta}^p$, are unknown and hence prior distributions on these parameters are required. We assume a Gaussian process prior for the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$, and we give an example where we find posterior predictive distributions to illustrate the calibration problem and demonstrate the impact of choice of design.

The regression parameter is known and fixed at $\rho = 1$. We have:

$$y_i = \eta(x_i, \boldsymbol{\theta}^p) + \delta_{\boldsymbol{\theta}^p}(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

with $\eta(x_i, \boldsymbol{\theta}) = \frac{\theta_1 x_i}{\theta_2 + x_i}$ and assume a Gaussian process prior on $\delta_{\boldsymbol{\theta}^p}(\cdot)$, similar to the previous example, such that:

$$\boldsymbol{\delta}_{\boldsymbol{\theta}^p} \sim N\left[\mathbf{0}_n, \sigma^2 \mathbf{K}(\phi)\right],$$

where $\boldsymbol{\delta}_{\boldsymbol{\theta}^p} = [\delta_{\boldsymbol{\theta}^p}(x_1), \ldots, \delta_{\boldsymbol{\theta}^p}(x_n)]^{\mathrm{T}}$ and $\mathbf{K}(\phi)$ is the correlation matrix with ijth entry $\mathbf{K}(\phi)_{ij} = \kappa(x_i, x_j; \phi)$, $i, j = 1, \ldots, n$. The random error is normally distributed with zero mean and variance σ_ε^2, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

We define $\boldsymbol{\eta} = [\eta(x_1, \boldsymbol{\theta}^p), \ldots, \eta(x_n, \boldsymbol{\theta}^p)]^{\mathrm{T}}$ and $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$. Hence,

$$\mathbf{y} \sim N\left(\boldsymbol{\eta}, \sigma^2 \mathbf{K}(\phi) + \sigma_\varepsilon^2 \mathbf{I}_n\right).$$

The reparameterisation described in Section 2.3.3 gives that

$$\mathbf{y} \sim N\left(\boldsymbol{\eta}, \sigma^2 \boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\Sigma} = \mathbf{K}(\phi) + \tau^2 \mathbf{I}_n$. More on the calibration model can be found in Chapter 6.

Prior specification

First, we specify the prior distributions for the unknown parameters as follows:

$$\theta_1^p \sim \text{Unif}[a_1, b_1], \quad \theta_2^p \sim \text{Unif}[a_2, b_2], \quad \sigma^2 \sim \text{IG}(a, b),$$
$$\phi \sim \exp(\lambda_\phi), \quad \tau^2 \sim \exp(\lambda_{\tau^2}),$$

with $b_1 > a_1 > 0$, $b_2 > a_2 > 0$, $a, b > 0$, $\lambda_\phi > 0$ and $\lambda_{\tau^2} > 0$. The joint prior density is given by:

$$\begin{aligned}\pi_b(\theta_1^p, \theta_2^p, \sigma^2, \phi, \tau^2) &= \pi_b(\theta_1^p)\pi_b(\theta_2^p)\pi_b(\sigma^2)\pi_b(\phi)\pi_b(\tau^2) \\ &\propto \frac{I(a_1 < \theta_1 < b_1)}{b_1 - a_1}\frac{I(a_2 < \theta_2 < b_2)}{b_2 - a_2} \\ &\qquad \times \left(\frac{1}{\sigma^2}\right)^{(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\}\pi_b(\phi)\pi_b(\tau^2). \end{aligned} \tag{3.9}$$

The posterior density results from applying Bayes' Theorem (2.16):

$$\pi_a(\theta_1^p, \theta_2^p, \sigma^2, \phi, \tau^2 \mid \mathbf{y}) \propto \pi_l(\mathbf{y} \mid \theta_1^p, \theta_2^p, \sigma^2, \phi, \tau^2)\pi_b(\theta_1^p, \theta_2^p, \sigma^2, \phi, \tau^2),$$

where

$$\pi_l(\mathbf{y} \mid \theta_1^p, \theta_2^p, \sigma^2, \phi, \tau^2) = \frac{|\mathbf{\Sigma}|^{-\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left([\mathbf{y} - \boldsymbol{\eta}]^{\mathrm{T}}\,\mathbf{\Sigma}^{-1}\,[\mathbf{y} - \boldsymbol{\eta}]\right)\right\},$$

is the likelihood function (see Section 2.3.4).

We can derive the posterior distribution, the conditional marginal distributions and the conditional predictive distributions as shown analytically in Sections 2.3.4 and 2.3.5. However, most integrals do not have an analytical solution and, as a result, we employ sampling techniques based on Markov chain Monte Carlo methods (see Section 2.4).

Example

To illustrate the calibration problem and demonstrate the impact of choice of design, we find posterior predictive distributions for the model (1.1) where the simulator is known with unknown parameters.

To simulate data, we assume $\theta_1^p = 15$, $\theta_2^p = 50$, $\delta_{\boldsymbol{\theta}^p}(x) = 0.5\sin(0.1x)$ and $\sigma_\varepsilon^2 = 0.05$. We assume the GP prior model (2.9) for $\delta_{\boldsymbol{\theta}^p}(\cdot)$ with the squared exponential correlation function $\kappa(x, x'; \phi) = \exp[-\phi(x - x')^2]$. We assume a priori $\theta_1^p \sim \text{Unif}[8, 24]$ and $\theta_2^p \sim \text{Unif}[20, 85]$. Figure C.3 (a) shows that we get a reasonable range of different shapes of the expected response of $\eta(x, \boldsymbol{\theta})$. In addition we assume $\sigma^2 \sim IG(3, 2)$, $\phi \sim \text{Exp}(200)$, $\tau^2 \sim \text{Exp}(15)$, which gives small noise-to-signal ratio. See also Appendix C.1.2 for samples from the prior distribution of $\delta_{\boldsymbol{\theta}^p}(x)$. A Metropolis-Hastings algorithm is used to draw a dependent sample $(\phi_i, \tau_i^2, \boldsymbol{\theta}_i^p)$, $i = 1, \ldots, 150,000$ using proposal distributions for ϕ and τ^2 of the form (3.8) and a normal proposal distribution for $\boldsymbol{\theta}^p$.

We examine four different designs with different sizes, shown in Figure 3.14 (b) and

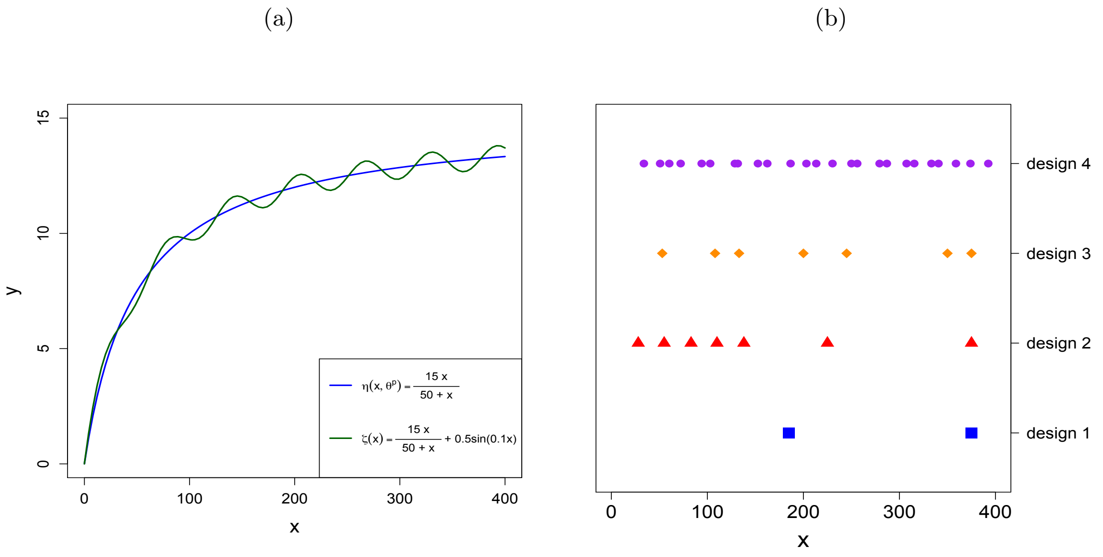


Figure 3.14: (a) True functions $\eta(x, \boldsymbol{\theta}^p)$ and $\zeta(x)$; (b) The four designs outlined in Table 3.2.

outlined in Table 3.2.

- Design 1 is a two-point maximin D-optimal design. The design is found by maximising the minimum D-efficiency over the parameter space $[20, 85]$ (see Dette and Biedermann, 2003).

- Design 2 is an ad hoc design with seven points. Most of the points of this design are concentrated where the true model is changing fastest and we also have some points at the stationary part of the model.

- Design 3 is a random Latin Hypercube design (McKay et al., 1979) with seven points (see Section 1.2.2).

- Design 4 is a random Latin Hypercube design (McKay et al., 1979) with 25 points.

We fit the calibration model to simulated data from each design and approximate the resulting posterior distributions for the unknown parameters $\boldsymbol{\theta}^p$ (Figures 3.15 (a) and 3.15 (b)), the discrepancy $\delta_{\boldsymbol{\theta}^p}(x)$ (Figures 3.16 (b), 3.17 (b), 3.18 (b), and 3.19 (b)) and reality $\zeta(x)$ (Figures 3.16 (a), 3.17 (a), 3.18 (a), 3.19 (a)). There are clear differences between the designs.

Table 3.2 holds the values of the posterior standard deviations and root mean squared errors for the two parameters θ_1^p, θ_2^p and the reality $\zeta(x)$ averaged across the design space, for each of the four designs.

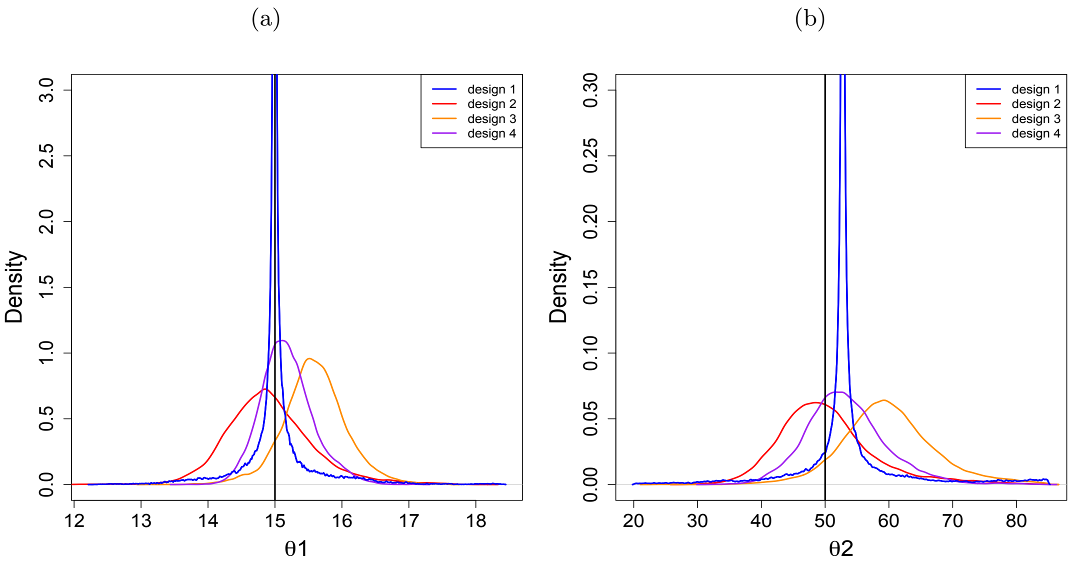


Figure 3.15: (a) Approximate posterior density of θ_1^p for each design; (b) Approximate posterior density of θ_2^p for each design

Design	**Posterior st. dev.**			**RMSE**		
	θ_1^p	θ_2^p	$\zeta(x)$	θ_1^p	θ_2^p	$\zeta(x)$
Design 1 ■ 2 point maximin D-optimal design	0.4482	7.5818	0.1367	0.4484	8.2457	0.3955
Design 2 ▲ 7 point ad hoc design	0.6675	7.7606	0.2074	0.6716	7.7678	0.3629
Design 3 ◆ 7 point Latin Hypercube design	0.5657	8.2990	0.1542	0.8158	13.4297	0.3998
Design 4 ● 25 point Latin Hypercube design	0.3961	6.3918	0.0557	0.4345	7.1979	0.1292

Table 3.2: Posterior standard deviation and root mean squared errors of θ_1^p, θ_2^p and $\zeta(x)$ averaged across the design space

Design 1 results in poor estimation of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$ as shown in Figure 3.16 (b), and overconfidence in the predictions as can be seen in Figure 3.16 (a). As noted by Brynjarsdóttir and O'Hagan (2014), an analysis that does not account for model discrepancy may lead to biased and over-confident parameter estimators and predictions. This is the case here, since the design does not take into account the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$. The posterior standard deviations are larger than might be anticipated from Figure 3.15 due to the long tails of the distribution and the fact that the estimate of θ_2^p is biased.

Design 2 performs reasonably (see Figure 3.17 and Table 3.2). Uncertainty is small

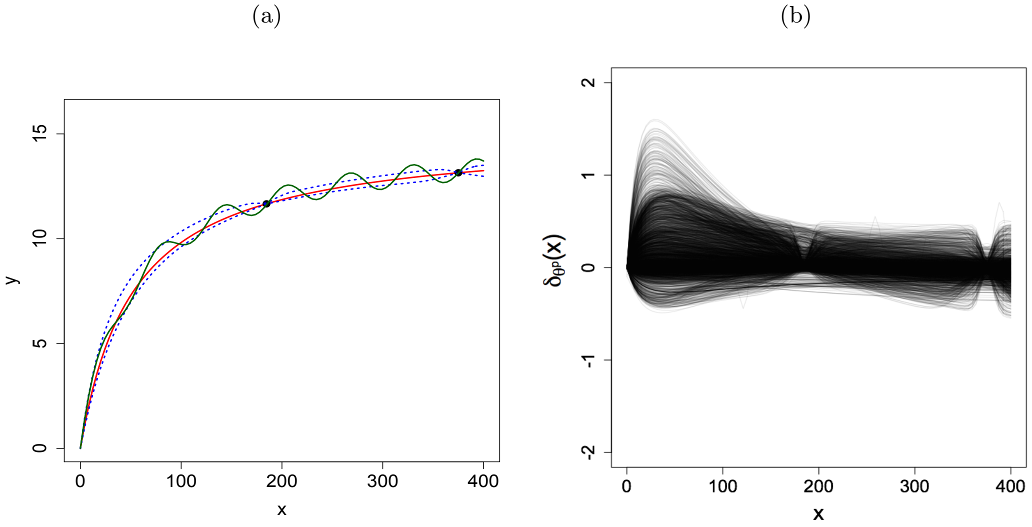


Figure 3.16: Design 1: (a) The true model (green line) the posterior median of $\zeta(x)$ (red line) and 95% credible intervals; (b) Samples from the posterior distribution of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$

for small values of x, where the majority of design points are placed, and the model is able to capture the sinusoidal form of the discrepancy function. As we move to the stationary part of the model, where we also have fewer points, the uncertainty increases and is only pinched to zero at the design points and we have a poor estimation of the discrepancy function.

Design 3, which makes no use of $\eta(x, \boldsymbol{\theta})$, has the worst performance (see Table 3.2). The bias in the posterior distribution for $\boldsymbol{\theta}^p$ may arise from non-identifiability. By this we mean the difficulty of identifying the discrepancy function, $\delta_{\boldsymbol{\theta}^p}(\cdot)$, that corresponds to the 'true' values of the simulator parameters, $\boldsymbol{\theta}^p$, since for any value of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ we can find a different discrepancy function, $\delta_{\boldsymbol{\theta}}(\cdot)$ (see Section 1.1 and Section 7.2.2).

Lastly, Design 4, with more points, most accurately captures the high-frequency discrepancy (see Figure 3.19). However, the posterior distribution for $\boldsymbol{\theta}^p$ is biased compared to the parameter values assumed in the simulation, also a consequence of non-identifiability (see Table 3.2).

3.3 Summary

In this chapter we gave some examples that demonstrated the importance of the choice of design. Existing optimal designs and space-filling designs result in poor estimation of the discrepancy function and this motivates us to find a methodology for finding Bayesian optimal designs that will be more suitable for estimation of the unknown parameters. We take a fully Bayesian approach by using a utility function and an optimisation algorithm in order to find designs for nonlinear models such as the Michaelis-

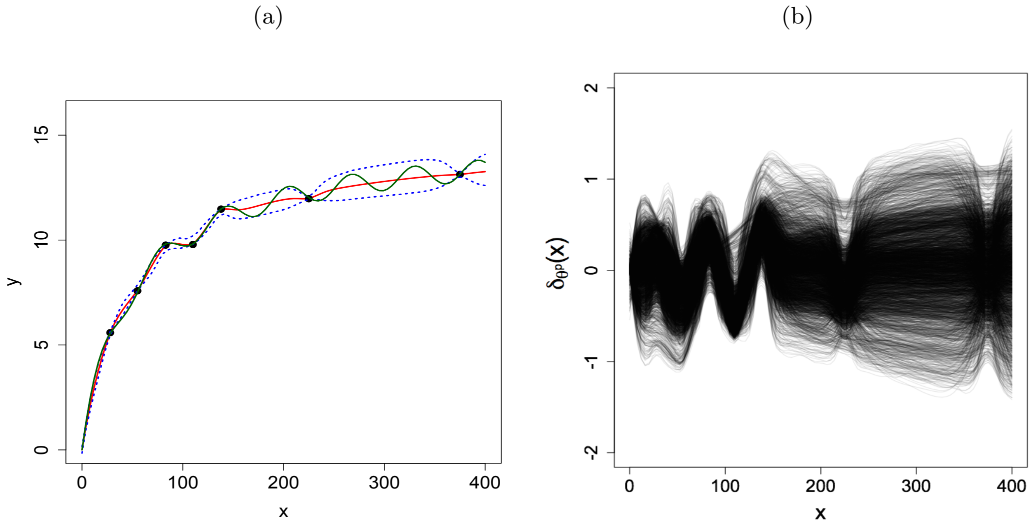


Figure 3.17: Design 2: (a) The true model (green line) the posterior median of $\zeta(x)$ (red line) and 95% credible intervals; (b) Samples from the posterior distribution of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$

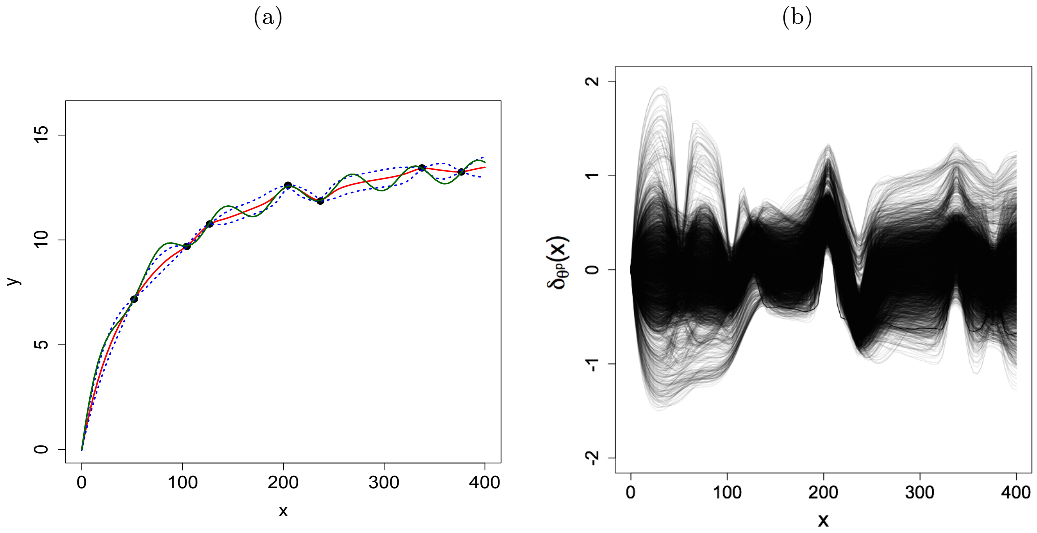


Figure 3.18: Design 3: (a) The true model (green line) the posterior median of $\zeta(x)$ (red line) and 95% credible intervals; (b) Samples from the posterior distribution of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$

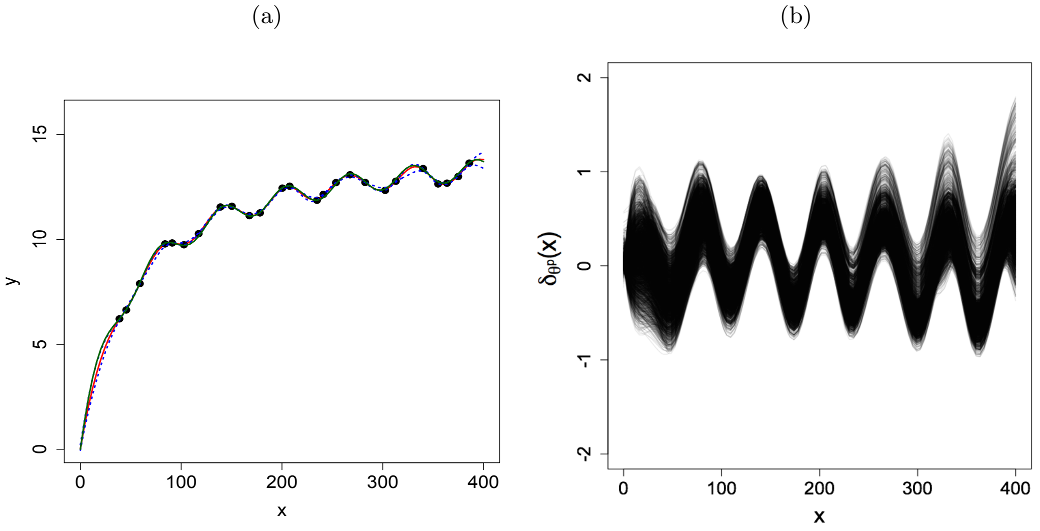


Figure 3.19: Design 4: (a) The true model (green line) the posterior median of $\zeta(x)$ (red line) and 95% credible intervals; (b) Samples from the posterior distribution of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$

Menten model (see Chapter 5) and the calibration model for known or unknown simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$ (see Chapter 6).

Chapter 4

Methods for approximating the expected Shannon information gain in Bayesian optimal design

The objective of this chapter is to describe the decision-theoretic approach to develop Bayesian optimal designs. We introduce Bayesian optimal designs that maximise the expected Shannon information gain utility and illustrate expected utility evaluation for the simple linear model. In general, Bayesian design is easy in principle and hard in practice. For many nonlinear models, the expected utility will be intractable, and involve high-dimensional integrals with respect to $\mathbf{y}$, necessitating numerical approximation. Naïve nested Monte Carlo is the most straightforward approximation method, however in some cases it fails to give an accurate estimate of the expected utility. For this reason a number of methods have been proposed to reduce the computational burden and reduce bias. Motivated by the simple linear example, we consider several alternative numerical methods for estimating the expected Shannon information gain. We illustrate some of the existing improved methods and propose two further new methods, called Laplace importance sampling (LIS) and approximate Laplace importance sampling (ALIS).

4.1 Decision-theoretic Bayesian designs

Decision theory (e.g. Berger, 1985, Chapter 1) addresses the problem of choosing an action, a, from a set, A, of possible actions under uncertainty about a parameter, $\boldsymbol{\psi} \in \Psi$. The uncertainty about $\boldsymbol{\psi}$ is typically represented by a probability distribution with density $\pi(\boldsymbol{\psi})$. The theory proposes that a should be chosen to maximise the expectation, with respect to $\boldsymbol{\psi}$, of a utility function, $u(a, \boldsymbol{\psi})$, or equivalently to minimise the expectation of a loss function. Bayesian experimental design can be viewed as a decision problem where the utility function is chosen to reflect the aims of the experiment, for example parameter inference or prediction.

Assume that given the design decision $\xi \in \Xi$ and parameter values $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{q_2})^{\mathrm{T}} \in \Psi$, we will observe data $\mathbf{y} \in \mathcal{Y}$ arising from the probability density function $\pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)$. Also, we assume the parameters $\boldsymbol{\psi}$ have prior density $\pi_b(\boldsymbol{\psi})$ and that we have a utility function $u(\xi, \boldsymbol{\psi}, \mathbf{y})$ quantifying performance in relation to the aims of the experiment. A Bayesian optimal design, $\xi^\star \in \Xi$, maximises the expected utility $U(\xi) = \mathbb{E}[u(\xi, \boldsymbol{\psi}, \mathbf{y})]$, where the expectation is with respect to the future data $\mathbf{y}$ and model parameters $\boldsymbol{\psi}$. That is,

$$\xi^\star \in \arg\max_{\xi \in \Xi} U(\xi),$$

where

$$\begin{aligned} U(\xi) &= \mathbb{E}[u(\xi, \boldsymbol{\psi}, \mathbf{y})] \\ &= \int_{\Psi} \int_{\mathcal{Y}} u(\xi, \boldsymbol{\psi}, \mathbf{y}) \pi(\mathbf{y}, \boldsymbol{\psi}|\xi) d\mathbf{y} d\boldsymbol{\psi} \\ &= \int_{\Psi} \int_{\mathcal{Y}} u(\xi, \boldsymbol{\psi}, \mathbf{y}) \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi) \pi_b(\boldsymbol{\psi}) d\mathbf{y} d\boldsymbol{\psi}. \end{aligned} \tag{4.1}$$

With a few exceptions, for most models, prior distributions and utility functions the integral (4.1) does not have a closed-form solution. Thus, to find Bayesian designs, (4.1) must be approximated either analytically, traditionally using asymptotic results (Chaloner and Verdinelli, 1995), or alternatively using numerical methods. Since many experiments have a small number of runs, asymptotic approximations may be inappropriate. Recently, progress has been made using Monte Carlo approaches to approximate the utility (Ryan et al., 2015), which may be more accurate for experiments with few runs. However, this raises several challenges, as we discuss in the next few sections.

For reviews of Bayesian design of experiments and related computational methods, see Chaloner and Verdinelli (1995), Ryan et al. (2015) and Woods et al. (2017).

4.1.1 Utility functions

Bernardo (1979) discussed the choice of a utility function when the goal of the experiment is inference, i.e. selection of a probability distribution that describes uncertainty about the parameter $\boldsymbol{\psi}$. He strongly advocated the utility function

$$u(\xi, \boldsymbol{\psi}, \mathbf{y}) = \log \pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi) - \log \pi_b(\boldsymbol{\psi}). \tag{4.2}$$

He argued that, in order to encourage the scientist to be honest, the utility should be maximised at (and only at) the posterior distribution, i.e. it should be a proper scoring rule[1]. It was shown that (4.2) is the the unique local[2] proper scoring rule. Also when the purpose of the experiment is inference about $\boldsymbol{\psi}$ it is common to use a utility function which can be thought of as the reduction in the surprisal[3] about the true parameter value using the posterior rather than the prior distribution.

In this case, the expected utility can be shown to be equal to the expected gain in Shannon information, or equivalently the expected Kullback-Leibler divergence from posterior density $\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)$ to the prior density $\pi_b(\boldsymbol{\psi})$ (Shannon, 1948; Lindley, 1956). A simple way to show that the expected utility is equal to the expected Kullback-Leibler divergence from posterior to prior density is given below.

The expected utility is given by:

$$\mathbb{E}[u(\xi,\boldsymbol{\psi},\mathbf{y})] = \int_{\Psi}\int_{\mathcal{Y}} \log\frac{\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)}{\pi_b(\boldsymbol{\psi})}\pi(\mathbf{y},\boldsymbol{\psi}|\xi)d\mathbf{y}d\boldsymbol{\psi}. \tag{4.3}$$

The Kullback-Leibler divergence from posterior to the prior density is:

$$d_{KL}[\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)||\pi_b(\boldsymbol{\psi})] = \int_{\Psi}\log\frac{\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)}{\pi_b(\boldsymbol{\psi})}\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)d\boldsymbol{\psi}.$$

Hence, the expected Kullback-Leibler divergence is given by:

$$\mathbb{E}\left\{d_{KL}\left[\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)||\pi_b(\boldsymbol{\psi})\right]\right\} = \int_{\mathcal{Y}}\int_{\Psi}\log\frac{\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)}{\pi_b(\boldsymbol{\psi})}\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)d\boldsymbol{\psi}\pi_e(\mathbf{y}|\xi)d\mathbf{y}.$$

Using Fubini's theorem and assuming mild regularity conditions we have,

$$\begin{aligned}\mathbb{E}\left\{d_{KL}\left[\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)||\pi_b(\boldsymbol{\psi})\right]\right\} &= \int_{\Psi}\int_{\mathcal{Y}}\log\frac{\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)}{\pi_b(\boldsymbol{\psi})}\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)\pi_e(\mathbf{y}|\xi)d\mathbf{y}d\boldsymbol{\psi}\\ &= \int_{\Psi}\int_{\mathcal{Y}}\log\frac{\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)}{\pi_b(\boldsymbol{\psi})}\pi(\mathbf{y},\boldsymbol{\psi}|\xi)d\mathbf{y}d\boldsymbol{\psi},\end{aligned}$$

which is the expected utility as given in Equation (4.3).

In common with other authors in this thesis we work with an alternative expression for (4.2) which can be derived using Bayes' rule (2.16):

$$\frac{\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)}{\pi_b(\boldsymbol{\psi})} = \frac{\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)\pi_e(\mathbf{y}|\xi)}{\pi_b(\boldsymbol{\psi})\pi_e(\mathbf{y}|\xi)} = \frac{\pi_l(\mathbf{y}|\boldsymbol{\psi},\xi)}{\pi_e(\mathbf{y}|\xi)}.$$

Hence we can replace $\log\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi) - \log\pi_b(\boldsymbol{\psi})$ with $\log\pi_l(\mathbf{y}|\boldsymbol{\psi},\xi) - \log\pi_e(\mathbf{y}|\xi)$. It

[1] A real function u is a proper scoring rule if for each density $g(\cdot)$,

$$\sup_f \int_{\Psi} u(f(\cdot),\boldsymbol{\psi})g(\boldsymbol{\psi})d\boldsymbol{\psi} = \int_{\Psi} u(g(\cdot),\boldsymbol{\psi})g(\boldsymbol{\psi})d\boldsymbol{\psi},$$

and the supremum is only attained at $g(\cdot)$ (see Definition 2 in Bernardo, 1979).

[2] Let u be the real function that describes the utility $u(f(\cdot),\boldsymbol{\psi})$ obtained by the scientist if the density function $f(\cdot)$ is reported as the final conclusion after an experiment has been performed and $\boldsymbol{\psi}$ is the unknown parameter. The function u is a local utility function if $u(f(\cdot),\boldsymbol{\psi}) = u(f(\boldsymbol{\psi}),\boldsymbol{\psi})$ for all values of $\boldsymbol{\psi}\in\Psi$ (see Definition 3 in Bernardo, 1979).

[3] The surprisal given a density f is $-\log f(\boldsymbol{\psi})$ (Baldi and Itti, 2010).

follows that

$$u(\xi, \boldsymbol{\psi}, \mathbf{y}) = \log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi) - \log \pi_e(\mathbf{y}|\xi), \tag{4.4}$$

where

$$\pi_e(\mathbf{y}|\xi) = \int_{\Psi} \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)\pi_b(\boldsymbol{\psi})d\boldsymbol{\psi},$$

is commonly called the evidence, a quantity of importance in model selection, e.g. Friel and Wyse (2012).

This leads to the following form of the expected utility:

$$\begin{aligned} U(\xi) &= \mathbb{E}[u(\xi, \boldsymbol{\psi}, \mathbf{y})] \\ &= \int_{\Psi}\int_{\mathcal{Y}} [\log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi) - \log \pi_e(\mathbf{y}|\xi)]\pi(\mathbf{y}, \boldsymbol{\psi}|\xi)d\mathbf{y}d\boldsymbol{\psi} \\ &= \int_{\Psi}\int_{\mathcal{Y}} \log \frac{\pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)}{\pi_e(\mathbf{y}|\xi)}\pi(\mathbf{y}, \boldsymbol{\psi}|\xi)d\mathbf{y}d\boldsymbol{\psi}. \end{aligned} \tag{4.5}$$

Many authors have used the following approximation to $U(\xi)$ justified via an asymptotic approximation to the posterior distribution of $\boldsymbol{\psi}$:

$$\varphi(\xi) = \mathbb{E}[\log |I(\boldsymbol{\psi}; \xi)|] = \int_{\Psi} \log |I(\boldsymbol{\psi}; \xi)|\pi_b(\boldsymbol{\psi})d\boldsymbol{\psi}, \tag{4.6}$$

where $I(\boldsymbol{\psi}; \xi)$ denotes the expected Fisher information, given in (1.7), for parameters $\boldsymbol{\psi}$ under the design ξ. Designs that maximise $\varphi(\xi)$ are sometimes referred to as (pseudo-) Bayesian $D-$optimal designs. This expression also results from taking the expectation of the utility function,

$$u(\xi, \boldsymbol{\psi}, \mathbf{y}) = \log |I(\boldsymbol{\psi}; \xi)|,$$

which does not depend on $\mathbf{y}$. The integral (4.6) can be approximated via Monte Carlo integration, via sampling from the prior distribution for $\boldsymbol{\psi}$, or numerical quadrature (for the latter, see Woods et al., 2006; Gotwalt et al., 2009).

In some cases, the goal of the experiment may be prediction rather than parameter inference. In this case an expected utility that quantifies the uncertainty of the posterior predictive distribution will be adopted. Suppose that, given the responses $\mathbf{y}$ obtained from design ξ, interest lies in predicting the response $\tilde{y}$ at one new design point $\tilde{\mathbf{x}}$. Then an appropriate (expected) utility is the expected Shannon information gain between the prior and the posterior predictive distribution. The prior predictive density (the marginal density $\pi(\tilde{y})$) does not depend on the design, and so maximisation of the expected gain in Shannon information for $\tilde{y}$ is equivalent to maximisation of

$$U(\xi) = \int_{\mathcal{Y}}\int_{\mathcal{Y}_p} \log \pi(\tilde{y}|\mathbf{y}, \xi)\pi(\tilde{y}, \mathbf{y}|\xi)d\tilde{y}d\mathbf{y}, \tag{4.7}$$

see San Martini and Spezzaferri (1984) and Verdinelli et al. (1993) where

$$\pi(\tilde{y}|\mathbf{y},\xi) = \int_{\Psi} \pi(\tilde{y}|\boldsymbol{\psi},\mathbf{y})\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)d\boldsymbol{\psi}$$

is the posterior predictive density. If convenient, we can rewrite $\pi(\tilde{y},\mathbf{y}|\xi)$ using Bayes' theroem (2.16).

Another common utility function is the Negative Squared Error Loss, given by

$$u(\xi,\boldsymbol{\psi},\mathbf{y}) = -\sum_{w=1}^{q_2}[\psi_w - \mathbb{E}(\psi_w|\mathbf{y},\xi)]^2, \tag{4.8}$$

where q_2 is the number of components of $\boldsymbol{\psi}$. Minimising the expected negative squared error loss is equivalent to maximising the expectation of the average posterior variance of $\boldsymbol{\psi}$ with respect to the marginal distribution of $\mathbf{y}$ (e.g. Overstall et al., 2018),

$$U(\xi) = \mathbb{E}\left\{\mathbb{E}\left[u(\xi,\boldsymbol{\psi},\mathbf{y})|\mathbf{y}\right]\right\} = \mathbb{E}[-\mathrm{tr}\{\mathrm{var}(\boldsymbol{\psi}|\mathbf{y},\xi)\}]. \tag{4.9}$$

As before, (4.9) may be approximated using an asymptotic normal approximation to the posterior distribution of $\boldsymbol{\psi}$, as follows:

$$\varphi(\xi) = -\mathbb{E}[\mathrm{tr}\{I(\boldsymbol{\psi};\xi)^{-1}\}] = -\int_{\Psi} \mathrm{tr}\{I(\boldsymbol{\psi};\xi)^{-1}\}\pi_b(\boldsymbol{\psi})d\boldsymbol{\psi}.$$

Designs that maximise $\varphi(\xi)$ are referred to as (pseudo-) Bayesian $A-$optimal designs.

4.1.2 Monte Carlo approximation of the expected utility

In this section, we focus on numerical evaluation of the expected Shannon information gain utility $U(\xi)$, given in (4.5), using Monte Carlo integration methods. An obvious way to approximate $U(\xi)$ is via

$$\tilde{U}(\xi) = \frac{1}{k_1}\sum_{h=1}^{k_1}\left[\log \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_h,\xi) - \log \tilde{\pi}_e^h\right], \tag{4.10}$$

where $(\boldsymbol{\psi}_h,\mathbf{y}_h)$, for $h = 1,\ldots,k_1$, are independent samples from the joint prior density $\pi(\boldsymbol{\psi},\mathbf{y}|\xi)$, and $\tilde{\pi}_e^h$ is an estimate of the evidence $\pi_e(\mathbf{y}_h|\xi)$.

There are several existing methods for estimating the evidence in (4.10), which vary in accuracy and computational expense. The simplest is 'naïve Monte Carlo', discussed in Section 4.1.3, which gives biased results. Other existing methods are Laplace approximations, discussed in Sections 4.2.1 and 4.2.2, and nested importance sampling discussed in Section 4.2.3. A novel method is discussed in Section 4.3, and shown in Chapter 5 to be more efficient than existing methods.

To find an optimal design, we wish to maximise $\tilde{U}(\xi)$. The problem with all Monte

Carlo approximations of $U(\xi)$ is that $\mathbf{y}_h$ depends on ξ, for $h = 1, \dots, k_1$. This means that every evaluation of $\tilde{U}(\xi)$ for a new ξ requires a new sample to be generated. This will be computationally expensive and, perhaps more importantly, $U(\xi)$ will not be a smooth function, which means that conventional optimisation algorithms cannot be applied. For low-dimensional problems (one variable and a small number of design points), Müller and Parmigiani (1996) performed stochastic optimisation by fitting curves to the Monte Carlo samples, effectively conducting a noisy computer experiment to construct a statistical emulator for the approximation $\tilde{U}(\xi)$. However, for problems with a large number of design variables this approach is computationally very expensive. For high-dimensional design spaces, typically a very large number of function evaluations is required to build an accurate emulator. Thus it is desirable to reduce the dimensionality of the problem. Overstall and Woods (2017) achieved this by using a coordinate exchange algorithm (Meyer and Nachtsheim, 1995) to break up the optimisation in to a sequence of one-dimensional problems. The need to emulate high-dimensional functions is therefore eliminated, resulting in an effective and computationally efficient design selection methodology. For further details of their algorithm, see Section 5.2.

4.1.3 Naïve Monte Carlo and its bias

In (4.10), the simplest way to approximate the evidence, $\pi_e(\mathbf{y}_h|\xi)$, is via

$$\tilde{\pi}_e^h = \frac{1}{k_2} \sum_{k=1}^{k_2} \pi_l(\mathbf{y}_h|\tilde{\boldsymbol{\psi}}_{hk}, \xi),$$

where $\tilde{\boldsymbol{\psi}}_{hk}$ is another sample from the prior density $\pi_b(\boldsymbol{\psi})$, for $h = 1, \dots, k_1$, and $k = 1, \dots k_2$. We refer to this approximation as the *naïve Monte Carlo* (nMC) method, outlined in Algorithm 2.

Algorithm 2: The naïve Monte Carlo method

```
Generate a sample $\boldsymbol{\psi}_h$, $h = 1, \dots, k_1$, from $\pi_b(\boldsymbol{\psi})$;
for $h = 1, \dots, k_1$ do
    Generate a response $\mathbf{y}_h$ from $\pi_l(\mathbf{y}|\boldsymbol{\psi}_h, \xi)$;
    Generate a sample $\{\tilde{\boldsymbol{\psi}}_{hk}\}_{k=1}^{k_2}$ from $\pi_b(\boldsymbol{\psi})$;
    for $k = 1, \dots, k_2$ do
        Calculate $\tilde{u}_{hk} = \pi_l(\mathbf{y}_h|\tilde{\boldsymbol{\psi}}_{hk}, \xi)$;
    Estimate the evidence $\pi_e(\mathbf{y}_h|\xi)$ via $\tilde{\pi}_e^h = \frac{1}{k_2}\sum_{k=1}^{k_2} \tilde{u}_{hk}$;
    Calculate $\tilde{u}_h = \log \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_h, \xi) - \log \tilde{\pi}_e^h$;
Estimate the expected Shannon information gain utility by $\tilde{U}(\xi) = \frac{1}{k_1}\sum_{h=1}^{k_1} \tilde{u}_h$;
```

Ryan (2003) has shown that the naïve Monte Carlo integration method yields a biased estimator $\tilde{U}(\xi)$ of $U(\xi)$. Asymptotically, the bias is

$$\mathbb{E}[\tilde{U}(\xi) - U(\xi)] \approx \frac{C(\xi)}{k_2},$$

where

$$C(\xi) = \frac{1}{2}\mathbb{E}\left\{ \frac{1}{[\pi_e(\mathbf{y}|\xi)]^2} \text{var}\left[\frac{\pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)}{\pi_e(\mathbf{y}|\xi)} \middle| \mathbf{y} \right] \right\}.$$

Hence k_2 controls the bias. Ryan (2003) also showed that k_1 controls the variance. Large values of k_1 and k_2 make the approximation problem computationally expensive, and hence one must consider a trade-off between k_1 and k_2. Increasing k_1 results in reduction of variance and increasing k_2 results in reduction of the positive bias. If the function $C(\xi)$ is approximately constant over ξ, the bias will be roughly constant in ξ for fixed k_2, and thus of no consequence when comparing designs. For fixed computational effort it will be therefore best to choose a fixed k_2 and choose k_1 to be larger than k_2.

A severe practical problem that occurs if moderate inner loop sample sizes, k_2, are used is that the evidence $\pi_e(\mathbf{y}|\xi)$ can often be estimated as zero, leading to a numerical estimate of infinity for the expected utility. This occurs when the posterior distribution is much more concentrated than the prior distribution since then the inner loop sample consists largely of values that are far from the region of highest posterior density and hence have zero likelihood.

Huan and Marzouk (2013) overcome the numerical issues with zero evidence by using the same sample of parameter values in the inner loop as in the outer loop. The approximation to the expected utility (4.10) is now given by

$$\tilde{U}(\xi) = \frac{1}{k_1}\sum_{h=1}^{k_1}\left[\log \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_h, \xi) - \frac{1}{k_1}\sum_{k=1}^{k_1} \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_k, \xi) \right], \qquad (4.11)$$

where $\{\boldsymbol{\psi}_k\}_{k=1}^{k_1}$ is a sample from the prior density, $\pi_b(\boldsymbol{\psi})$. Note that the same sample is used in the outer and inner summation. We refer to this approach as the '*reuse*' method. It usually gives finite estimates of the expected utility gain because, for each $\mathbf{y}_h$, the inner loop sample now contains the value $\boldsymbol{\psi}_h$ which is used to generate the response and which usually has nonneglible posterior density. The biases of naïve Monte Carlo and 'reuse' estimators are asymptotically of the same order, although 'reuse' estimators offer substantial gains in computational efficiency because finite estimates of the expected utility can be obtained with much smaller values of k_2. However for finite inner loop sample sizes, this method can result in large negative bias (see examples in Sections 5.1.2 and 5.1.3).

In the next section we illustrate through the simple linear example how naïve Monte Carlo approximation of the expected Shannon information gain results in positive bias and overestimation of the information gain for a given design.

Linear Regression example

For the naïve Monte Carlo (nMC) method, we now illustrate how the bias and variance change with the sizes of the inner and outer loop samples. We do so using an illustrative example in which the expected Shannon information gain is available analytically. We assume the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.12}$$

where $\mathbf{X}$ is the $n \times 2$ model matrix $\mathbf{X} = [\mathbf{1}_n \quad (x_1, \dots, x_n)^{\mathrm{T}}]$ with x_i the value of an explanatory variable for the ith run, $i = 1, \dots, n$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}} \in \mathcal{B} \subset \mathbb{R}^2$ contains the unknown regression parameters, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ is the observation error, and σ_ε^2 is the known constant variance.

The conjugate prior distribution for $\boldsymbol{\beta}$ is a multivariate normal, $N(\boldsymbol{\beta}_0, \sigma_\varepsilon^2 \mathbf{R})$, for which:

$$\pi_b(\boldsymbol{\beta}) = (2\pi\sigma_\varepsilon^2)^{-\frac{p}{2}} |\mathbf{R}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}} \mathbf{R}^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\},$$

where p is the number of unknown regression parameters in $\boldsymbol{\beta}$.

The likelihood function is given by:

$$\pi_l(\mathbf{y}|\boldsymbol{\beta}, \sigma_\varepsilon^2, \xi) = (2\pi\sigma_\varepsilon^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}.$$

Using Bayes' theorem (2.16) we obtain the posterior density,

$$\begin{aligned}
\pi_a(\boldsymbol{\beta}|\mathbf{y}, \sigma_\varepsilon^2, \xi) &\propto \pi_l(\mathbf{y}|\boldsymbol{\beta}, \sigma_\varepsilon^2, \xi)\pi_b(\boldsymbol{\beta}) \\
&= (2\pi\sigma_\varepsilon^2)^{-(\frac{n+p}{2})} |\mathbf{R}|^{-\frac{1}{2}} \exp\Bigg\{ -\frac{1}{2\sigma_\varepsilon^2} \Big[(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}} \mathbf{R}^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
&\qquad\qquad + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \Big] \Bigg\} \\
&\propto \frac{1}{(2\pi\sigma_\varepsilon^2)^{p/2} |\mathbf{S}^*|^{1/2}} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \Big[(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \mathbf{S}^{*-1} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \Big] \right\},
\end{aligned}$$

with

$$\begin{aligned}
\boldsymbol{\beta}^* &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{-1})^{-1} (\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{R}^{-1}\boldsymbol{\beta}_0) \\
\mathbf{S}^* &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{-1})^{-1}.
\end{aligned}$$

Hence $\boldsymbol{\beta} \mid \mathbf{y}, \sigma_\varepsilon^2$ is normal with mean $\boldsymbol{\beta}^*$ and variance $\sigma_\varepsilon^2 \mathbf{S}^*$.

The expected Shannon information gain is:

$$U(\xi) = \int_{\mathcal{B}} \int_{\mathcal{Y}} \log[\pi_a(\boldsymbol{\beta}|\mathbf{y}, \sigma_\varepsilon^2, \xi) - \log \pi_b(\boldsymbol{\beta})] \pi(\boldsymbol{\beta}, \mathbf{y}|\xi) d\mathbf{y} d\boldsymbol{\beta}$$

$$
\begin{aligned}
&= \int_{\mathcal{B}}\int_{\mathcal{Y}} \Big[-\frac{p}{2}\log 2\pi\sigma_\varepsilon^2 - \frac{1}{2}\log|\mathbf{S}^*| - \frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{S}^{*-1}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \\
&\qquad - \Big\{ -\frac{p}{2}\log 2\pi\sigma_\varepsilon^2 - \frac{1}{2}\log|\mathbf{R}| - \frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \Big\} \Big] \pi(\boldsymbol{\beta},\mathbf{y}|\xi)d\mathbf{y}d\boldsymbol{\beta} \\
&= \int_{\mathcal{B}}\int_{\mathcal{Y}} \Big[-\frac{p}{2}\log 2\pi\sigma_\varepsilon^2 - \frac{1}{2}\log|\mathbf{S}^*| \\
&\qquad\qquad - \frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{S}^{*-1}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \Big] \pi_l(\mathbf{y}|\boldsymbol{\beta},\sigma_\varepsilon^2,\xi)\pi_b(\boldsymbol{\beta})d\mathbf{y}d\boldsymbol{\beta} \\
&\quad + \int_{\mathcal{B}}\int_{\mathcal{Y}} \Big[\frac{p}{2}\log 2\pi\sigma_\varepsilon^2 + \frac{1}{2}\log|\mathbf{R}| \\
&\qquad\qquad + \frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \Big] \pi_l(\mathbf{y}|\boldsymbol{\beta},\sigma_\varepsilon^2,\xi)\pi_b(\boldsymbol{\beta})d\mathbf{y}d\boldsymbol{\beta} \\
&= \mathrm{SIG}_1 + \mathrm{SIG}_2.
\end{aligned}
$$

For the second integral we have

$$
\begin{aligned}
\mathrm{SIG}_2 &= \int_{\mathcal{B}}\int_{\mathcal{Y}} \Big[\frac{p}{2}\log 2\pi\sigma_\varepsilon^2 + \frac{1}{2}\log|\mathbf{R}| \qquad (4.13) \\
&\qquad + \frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \Big] \pi_l(\mathbf{y}|\boldsymbol{\beta},\sigma_\varepsilon^2,\xi)\pi_b(\boldsymbol{\beta})d\mathbf{y}d\boldsymbol{\beta} \\
&= \frac{p}{2}\log 2\pi\sigma_\varepsilon^2 + \frac{1}{2}\log|\mathbf{R}| + \frac{1}{2\sigma_\varepsilon^2}\int_{\mathcal{B}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\pi_b(\boldsymbol{\beta})d\boldsymbol{\beta} \\
&= \frac{p}{2}\log 2\pi\sigma_\varepsilon^2 + \frac{1}{2}\log|\mathbf{R}| + \frac{1}{2\sigma_\varepsilon^2}\mathrm{tr}(\sigma_\varepsilon^2\mathbf{R}^{-1}\mathbf{R}) \\
&= \frac{p}{2}\log 2\pi\sigma_\varepsilon^2 + \frac{1}{2}\log|\mathbf{R}| + \frac{p}{2}. \qquad (4.14)
\end{aligned}
$$

By a similar argument, the first integral is

$$
\mathrm{SIG}_1 = \frac{1}{2}\log\left|\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{R}^{-1}\right| - \frac{p}{2}\log(2\pi\sigma_\varepsilon^2) - \frac{p}{2}, \qquad (4.15)
$$

agreeing with known results from Chaloner and Verdinelli (1995).

Combining Equations (4.14) and (4.15) we get:

$$
\begin{aligned}
U(\xi) &= \frac{1}{2}\log\left|\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{R}^{-1}\right| - \frac{p}{2}\log(2\pi\sigma_\varepsilon^2) - \frac{p}{2} + \left(\frac{p}{2}\log 2\pi\sigma_\varepsilon^2 + \frac{1}{2}\log|\mathbf{R}| + \frac{p}{2}\right) \\
&= \frac{1}{2}\log\left|\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{R}^{-1}\right| + \frac{1}{2}\log|\mathbf{R}|. \qquad (4.16)
\end{aligned}
$$

We now show how the bias of the naïve Monte Carlo (nMC) method changes for different values of k_1 and k_2 using the exact value of the expected utility as shown in (4.16).

We assume $\sigma_\varepsilon^2 = \frac{1}{3}$ and $\boldsymbol{\beta} \sim N(\mathbf{0}_p, \sigma_\varepsilon^2\mathbf{I}_p)$. We use the design presented in Figure 4.1 for all the results that we are going to illustrate for this linear regression example. This design is an expected Shannon information gain optimal design found for the linear model using the ACE algorithm (Section 5.2) and nMC (Section 4.1.3). Figure 4.2

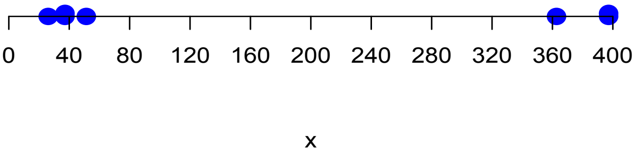


Figure 4.1: Expected Shannon information gain optimal design with $n = 7$, found for the linear model using ACE (Section 5.2) and nMC (Section 4.1.3); two of the points are repeated twice

shows the distribution of 100 estimates of the expected utility obtained using nMC for different combinations of k_1 and k_2. The true value of the Shannon information gain, obtained using (4.16), is shown by the red horizontal line.

From Figure 4.2, we notice considerable differences for the different pairs of k_1 and k_2. As the number of samples in the inner loop, k_2, increases, the bias decreases as anticipated from the asymptotic theory (Ryan, 2003). As the number of samples in the outer loop, k_1, increases, the variance decreases. A very large inner loop size ($k_2 = 100,000$) moves the estimates of the expected utility much closer to the true value but also increases the computational expense of the approximating method.

4.2 Existing improved methods for approximating the expected utility

The approximation of the expected Shannon information gain utility function,

$$U(\xi) = \int_{\Psi} \int_{\mathcal{Y}} \log \frac{\pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)}{\pi_e(\mathbf{y}|\xi)} \pi(\mathbf{y}, \boldsymbol{\psi}|\xi) d\mathbf{y} d\boldsymbol{\psi},$$

requires the solution of intractable integrals and for this reason numerical approximation methods are used as described in Section 4.1.2. An obvious way to approximate the expected utility is via (4.10); that is to take an independent sample of $(\boldsymbol{\psi}_h, \mathbf{y}_h)$, $h = 1, \ldots, k_1$, from the joint prior density $\pi(\mathbf{y}, \boldsymbol{\psi}|\xi)$ and approximate the evidence, $\pi_e(\mathbf{y}|\xi)$, using another sample from a known distribution.

There are several existing methods for estimating the evidence in (4.10), which vary in accuracy and computational expense. A summary of these methods can be found in Table 5.1 of Chapter 5.

The most straightforward approach, naïve Monte Carlo (nMC), approximates the evidence using another sample from the prior density $\pi_b(\boldsymbol{\psi})$ (see Section 4.1.3). However, the positive bias of this method overestimates the information gain from an experiment. Also, it requires large values of k_1 and k_2 to obtain sufficient precision and accuracy

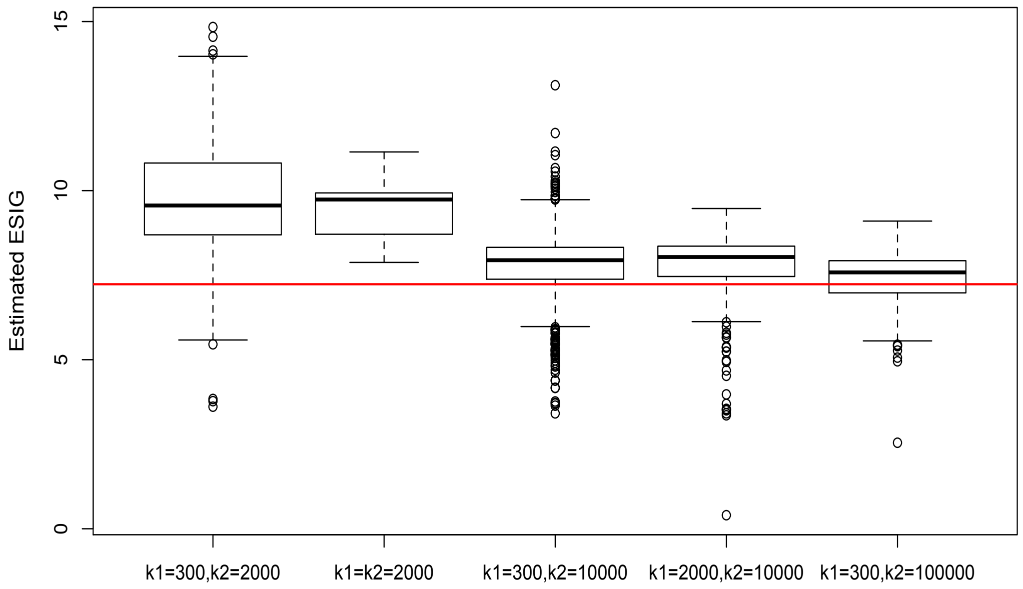


Figure 4.2: Estimated expected Shannon information gain for the linear model (4.12) using nMC and different combinations of k_1 and k_2, and the true value of the Shannon information gain obtained using (4.16) (red line)

of the approximation. As each design assessment requires $k_1(1 + k_2)$ likelihood evaluations, this leads to computationally expensive optimisation when searching for an optimal design, for even moderate k_1 and k_2. For diffuse prior distributions and informative experiments, a problem of zero approximation to the evidence can also occur (Section 4.1.3). These challenges have led to the development of new improved methods that will give better approximations to the evidence, $\pi_e(\mathbf{y}|\xi)$, in order to reduce the positive bias.

Long et al. (2013) aimed to reduce the computational expense of sampling techniques, i.e. naïve Monte Carlo approximation, by employing the Laplace approximation. This approximation uses a second-order Taylor series expansion of the log-posterior density, $\log \pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi)$, about the posterior mode $\hat{\boldsymbol{\psi}}$, leading to a Gaussian approximation of the posterior distribution of $\boldsymbol{\psi}$. This fundamental asymptotic method was first introduced by Pierre Simon Laplace (Stigler, 1986) under the assumption that a sufficient number of observations, n, is available.

In Sections 4.2.1 and 4.2.2 we consider two different methods of approximating the expected Shannon information gain based on Laplace approximations. The $(k_1 \times k_2)$ inner likelihood evaluations are replaced with k_1 optimisations, each of which takes only a few iterations of a quasi-Newton algorithm. The first of these, which we call Laplace Approximation I (LA1), coincides with the approximation proposed by Overstall et al. (2018), and follows from the expression of the utility function as the difference between the log-likelihood and log-evidence (see Section 4.1.1); an estimate for the evidence is found based on a Laplace approximation and is used within (4.10) to approximate the expected Shannon information gain. The second approximation, which we call Laplace Approximation II (LA2), coincides with that derived by Long et al. (2013), and follows from the alternative expression of the utility function, discussed in Section 4.1.1, as the difference of the log-posterior and log-prior densities.

Importance sampling is a Monte Carlo integration method commonly used for approximating a target integral of interest. In Section 4.2.3 we describe an alternative approximation of the expected Shannon information gain that uses importance sampling to estimate the evidence, introduced by Feng (2015).

In Section 4.3 we propose two further new approximation methods to estimate the evidence in the expected Shannon information gain, called Laplace importance sampling (LIS) and approximate Laplace importance sampling (ALIS). These methods combine features of importance sampling and Laplace approximations.

4.2.1 Approximating the evidence - Laplace Approximation I

In this section an approximation to the evidence, $\pi_e(\mathbf{y}|\xi)$, required to estimate the expected utility via (4.10), is found using a Laplace approximation.

Recall the utility function is given by

$$u(\xi, \boldsymbol{\psi}, \mathbf{y}) = \log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi) - \log \pi_e(\mathbf{y}|\xi),$$

with the expected Shannon information gain given by Equation (4.5).

We can express the evidence as:

$$\pi_e(\mathbf{y}|\xi) = \int_\Psi \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)\pi_b(\boldsymbol{\psi})d\boldsymbol{\psi} = \int_\Psi \exp\left[\log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi)\right] d\boldsymbol{\psi}, \tag{4.17}$$

where $\pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi) = \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)\pi_b(\boldsymbol{\psi})$ is the unnormalised posterior density.

A second order Taylor series expansion of $\log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi)$ about the posterior mode $\hat{\boldsymbol{\psi}}$ gives:

$$\begin{aligned}\log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi) \approx \log \pi_u(\hat{\boldsymbol{\psi}}|\mathbf{y}, \xi) + \frac{\partial \log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi)}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \\ - \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}),\end{aligned}$$

where $\mathbf{H}(\hat{\boldsymbol{\psi}})$ is the negative Hessian of the log-unnormalised posterior density,

$$\mathbf{H}(\hat{\boldsymbol{\psi}}) = -\frac{\partial^2 \log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi)}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^{\mathrm{T}}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}.$$

By definition the first derivative of the log-unnormalised posterior density is zero at the posterior mode, and so the second term of the Taylor expansion vanishes, giving:

$$\log \pi_u(\boldsymbol{\psi}|\mathbf{y},\xi) \approx \log \pi_u(\hat{\boldsymbol{\psi}}|\mathbf{y},\xi) - \frac{1}{2}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}}). \tag{4.18}$$

Exponentiating the above gives:

$$\pi_u(\boldsymbol{\psi}|\mathbf{y},\xi) \approx \pi_u(\hat{\boldsymbol{\psi}}|\mathbf{y},\xi)\exp\left[-\frac{1}{2}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})\right],$$

and integrating this expression results in:

$$\begin{aligned}\pi_e(\mathbf{y}|\xi) &= \int_\Psi \pi_u(\boldsymbol{\psi}|\mathbf{y},\xi)d\boldsymbol{\psi}\\ &\approx \int_\Psi \pi_u(\hat{\boldsymbol{\psi}}|\mathbf{y},\xi)\exp\left[-\frac{1}{2}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})\right]d\boldsymbol{\psi}\\ &= \frac{\pi_u(\hat{\boldsymbol{\psi}}|\mathbf{y},\xi)(2\pi)^{q_2/2}}{\left|\mathbf{H}(\hat{\boldsymbol{\psi}})\right|^{1/2}},\end{aligned}$$

where q_2 is the number of unknown parameters $\boldsymbol{\psi}$.

Hence in (4.10), the Laplace approximation to the evidence is

$$\pi_e(\mathbf{y}|\xi) \approx \frac{\pi_u(\hat{\boldsymbol{\psi}}|\mathbf{y},\xi)(2\pi)^{q_2/2}}{\left|\mathbf{H}(\hat{\boldsymbol{\psi}})\right|^{1/2}}.$$

The approximation of the expected Shannon information gain (4.5) is given by:

$$\begin{aligned}U(\xi) \approx \int_\Psi\int_\mathcal{Y} \Big[&\log \pi_l(\mathbf{y}|\boldsymbol{\psi},\xi) - \log \pi_u(\hat{\boldsymbol{\psi}}|\mathbf{y},\xi)\\ &-\frac{1}{2}\log\left[(2\pi)^{q_2}\left|\mathbf{H}(\hat{\boldsymbol{\psi}})^{-1}\right|\right]\Big]\pi(\mathbf{y},\boldsymbol{\psi}|\xi)d\mathbf{y}d\boldsymbol{\psi}.\end{aligned} \tag{4.19}$$

Hence the approximation (4.10) becomes:

$$\begin{aligned}\tilde{U}(\xi) &= \frac{1}{k_1}\sum_{h=1}^{k_1}\left[\log \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_h,\xi) - \log \tilde{\pi}_e^h\right]\\ &= \frac{1}{k_1}\sum_{h=1}^{k_1}\left[\log \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_h,\xi) - \log \pi_u(\hat{\boldsymbol{\psi}}_h|\mathbf{y}_h,\xi) - \frac{1}{2}\log\left[(2\pi)^{q_2}\left|\mathbf{H}(\hat{\boldsymbol{\psi}}_h)^{-1}\right|\right]\right],\end{aligned}$$

where $\hat{\boldsymbol{\psi}}_h$ is obtained using a quasi-Newton algorithm (see Section 4.3).

We refer to this approximation as Laplace Approximation I or *LA1*.

Overstall et al. (2018) showed through some examples that such a normal-based approximation together with the ACE algorithm (Section 5.2) is able to find efficient Bayesian optimal designs.

4.2.2 Laplace approximation II

Laplace approximation II, or *LA2*, is a different approximation to the expected Shannon information gain that does not directly use an approximation to the evidence. If we take the alternative expression of the utility given in Section 4.1.1,

$$u(\xi, \boldsymbol{\psi}, \mathbf{y}) = \log \pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi) - \log \pi_b(\boldsymbol{\psi}),$$

it is possible to obtain the following approximation to (4.3):

$$\begin{aligned} U(\xi) \approx \int_{\mathcal{Y}} \Bigg[-\frac{1}{2}\log(2\pi)^{q_2}|\mathbf{H}(\hat{\boldsymbol{\psi}})^{-1}| - \frac{q_2}{2} - \log \pi_b(\hat{\boldsymbol{\psi}}) \\ -\frac{1}{2}\mathrm{tr}\left[\mathbf{Q}(\hat{\boldsymbol{\psi}})\mathbf{H}(\hat{\boldsymbol{\psi}})^{-1}\right]\Bigg] \pi_e(\mathbf{y}|\xi)d\mathbf{y}. \end{aligned} \tag{4.20}$$

To get the above result, we take a second order Taylor series expansion of the log-posterior density, $\log \pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi)$, about the posterior mode, $\hat{\boldsymbol{\psi}}$:

$$\begin{aligned} \log \pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi) &\approx \log \pi_a(\hat{\boldsymbol{\psi}}|\mathbf{y}, \xi) + \frac{\partial \log \pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi)}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \\ &\qquad - \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \\ &= \log \pi_a(\hat{\boldsymbol{\psi}}|\mathbf{y}, \xi) - \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}), \end{aligned} \tag{4.21}$$

as $\frac{\partial \log \pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} = 0$ by definition and,

$$\frac{\partial^2 \log \pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi)}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^{\mathrm{T}}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} = \frac{\partial^2 \log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi)}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^{\mathrm{T}}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} = -\mathbf{H}(\hat{\boldsymbol{\psi}}),$$

which holds from Bayes' Theorem (2.16) and by the fact that the evidence, $\pi_e(\mathbf{y}|\xi)$, does not depend on the unknown parameters $\boldsymbol{\psi}$.

In the previous section we showed that the evidence,

$$\pi_e(\mathbf{y}|\xi) \approx \frac{\pi_u(\hat{\boldsymbol{\psi}}|\mathbf{y}, \xi)(2\pi)^{q_2/2}}{\left|\mathbf{H}(\hat{\boldsymbol{\psi}})\right|^{1/2}} = \frac{\pi_l(\mathbf{y}|\hat{\boldsymbol{\psi}}, \xi)\pi_b(\hat{\boldsymbol{\psi}})(2\pi)^{q_2/2}}{\left|\mathbf{H}(\hat{\boldsymbol{\psi}})\right|^{1/2}}.$$

Again using Bayes' Theorem, the posterior density is given by

$$\pi_a(\hat{\boldsymbol{\psi}}|\mathbf{y}, \xi) = \frac{\pi_l(\mathbf{y}|\hat{\boldsymbol{\psi}}, \xi)\pi_b(\hat{\boldsymbol{\psi}})}{\pi_e(\mathbf{y}|\xi)},$$

and combining these two results we get:

$$\pi_a(\hat{\boldsymbol{\psi}}|\mathbf{y}, \xi) \approx \frac{\left|\mathbf{H}(\hat{\boldsymbol{\psi}})\right|^{1/2}}{(2\pi)^{q_2/2}}.$$

Plugging this result back into Equation (4.21) we get a normal approximation to the posterior density with mean $\hat{\boldsymbol{\psi}}$ and variance $\mathbf{H}(\hat{\boldsymbol{\psi}})^{-1}$:

$$\log \pi_a(\boldsymbol{\psi}|\mathbf{y},\xi) \approx -\log\left[(2\pi)^{q_2/2}\left|\mathbf{H}(\hat{\boldsymbol{\psi}})\right|^{-1/2}\right] - \frac{1}{2}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}}). \tag{4.22}$$

Hence the approximation of the expected Shannon information gain (4.3) becomes

$$\begin{aligned}
U(\xi) &\approx \int_\Psi\int_{\mathcal{Y}}\left[-\frac{1}{2}\log(2\pi)^{q_2}|\mathbf{H}(\hat{\boldsymbol{\psi}})^{-1}| - \frac{1}{2}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})\right.\\
&\qquad\left. - \log\pi_b(\boldsymbol{\psi})\right]\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)\pi_e(\mathbf{y}|\xi)d\mathbf{y}d\boldsymbol{\psi}\\
&= \int_{\mathcal{Y}}\left[-\frac{1}{2}\log(2\pi)^{q_2}|\mathbf{H}(\hat{\boldsymbol{\psi}})^{-1}| - \underbrace{\int_\Psi \frac{1}{2}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)d\boldsymbol{\psi}}_{I_1}\right.\\
&\qquad\left. - \underbrace{\int_\Psi \log\pi_b(\boldsymbol{\psi})\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)d\boldsymbol{\psi}}_{I_2}\right]\pi_e(\mathbf{y}|\xi)d\mathbf{y}.
\end{aligned}$$

For the first integral, I_1, we use the normal approximation to the posterior density (4.22), and the known formula of the expectation of a quadratic form[4]:

$$\begin{aligned}
I_1 &= \int_\Psi \frac{1}{2}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)d\boldsymbol{\psi}\\
&\approx \frac{1}{2}\mathrm{tr}[\mathbf{H}(\hat{\boldsymbol{\psi}})\mathbf{H}(\hat{\boldsymbol{\psi}})^{-1}]\\
&= \frac{1}{2}\mathrm{tr}[\mathbf{I}_{q_2}]\\
&= \frac{q_2}{2}.
\end{aligned}$$

In order to approximate the second integral, I_2, we take a second-order Taylor series expansion of the log-likelihood, $\log\pi_l(\mathbf{y}|\boldsymbol{\psi},\xi)$, about the posterior mode $\hat{\boldsymbol{\psi}}$:

$$\begin{aligned}
\log\pi_l(\mathbf{y}|\boldsymbol{\psi},\xi) &\approx \log\pi_l(\mathbf{y}|\hat{\boldsymbol{\psi}},\xi) + \frac{\partial\log\pi_l(\mathbf{y}|\boldsymbol{\psi},\xi)}{\partial\boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})\\
&\qquad + \frac{1}{2}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})^{\mathrm{T}}\left[-\mathbf{H}(\hat{\boldsymbol{\psi}})-\mathbf{Q}(\hat{\boldsymbol{\psi}})\right](\boldsymbol{\psi}-\hat{\boldsymbol{\psi}}),
\end{aligned} \tag{4.23}$$

where

$$\mathbf{Q}(\hat{\boldsymbol{\psi}}) = \frac{\partial^2\log\pi_b(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathrm{T}}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}.$$

[4] The expectation of a quadratic form is

$$\mathbb{E}[\mathbf{c}^{\mathrm{T}}\Lambda\mathbf{c}] = \mathrm{tr}[\Lambda\Sigma_c] + \mu_c^{\mathrm{T}}\Lambda\mu_c,$$

where $\mathbf{c}$ is a vector of n_c random variables, Λ is a n_c-dimensional symmetric matrix, μ_c is the expected value of $\mathbf{c}$ and Σ_c is the variance-covariance matrix of $\mathbf{c}$ (Mathai and Provost, 1992, Chapter 3).

and

$$-\mathbf{H}(\hat{\boldsymbol{\psi}}) - \mathbf{Q}(\hat{\boldsymbol{\psi}}) = \frac{\partial^2 \log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathrm{T}}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}.$$

This results from

$$\begin{aligned}
\mathbf{H}(\hat{\boldsymbol{\psi}}) &= -\frac{\partial^2 \log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathrm{T}}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\
&= -\frac{\partial^2 \log[\pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)\pi_b(\boldsymbol{\psi})]}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathrm{T}}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\
&= \left[-\frac{\partial^2 \log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathrm{T}}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} - \frac{\partial^2 \log \pi_b(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathrm{T}}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}\right].
\end{aligned}$$

Note that, to approximate I_2, (4.23) only needs to be an accurate approximation in a small neighbourhood around $\hat{\boldsymbol{\psi}}$, because if the sample size n is large then the posterior distribution will be concentrated around $\hat{\boldsymbol{\psi}}$.

Using Equation (4.18) and (4.23), the log-prior density, $\log \pi_b(\boldsymbol{\psi})$, is given by:

$$\begin{aligned}
\log \pi_b(\boldsymbol{\psi}) &= \log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi) - \log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi) \\
&\approx \log \pi_u(\hat{\boldsymbol{\psi}}|\mathbf{y}, \xi) - \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) - \log \pi_l(\mathbf{y}|\hat{\boldsymbol{\psi}}, \xi) \\
&\quad - \frac{\partial \log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)}{\partial\boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) - \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}}\left[-\mathbf{H}(\hat{\boldsymbol{\psi}}) - \mathbf{Q}(\hat{\boldsymbol{\psi}})\right](\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \\
&= \log \pi_b(\hat{\boldsymbol{\psi}}) + \frac{\partial \log \pi_b(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) + \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{Q}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}),
\end{aligned}$$

from

$$\begin{aligned}
\frac{\partial \log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)}{\partial\boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) &= \frac{\partial \log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi)}{\partial\boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \\
&\quad - \frac{\partial \log \pi_b(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}),
\end{aligned}$$

where

$$\frac{\partial \log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi)}{\partial\boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) = 0,$$

by definition.

Then we approximate the second integral I_2, with respect to $\boldsymbol{\psi}$ as:

$$\begin{aligned}
I_2 &= \int_\Psi \log \pi_b(\boldsymbol{\psi})\pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi)d\boldsymbol{\psi} \\
&\approx \underbrace{\int_\Psi \log \pi_b(\hat{\boldsymbol{\psi}})\pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi)d\boldsymbol{\psi}}_{I_3} \\
&\qquad + \underbrace{\int_\Psi \frac{\partial \log \pi_b(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})\pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi)d\boldsymbol{\psi}}_{I_4}
\end{aligned}$$

$$\underbrace{+\int_{\Psi}\frac{1}{2}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{Q}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}})\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)d\boldsymbol{\psi}\,.}_{I_5}$$

We know that

$$I_3 = \mathbb{E}_{\boldsymbol{\psi}|\mathbf{y}}\left[\log \pi_b(\hat{\boldsymbol{\psi}})\right] = \log \pi_b(\hat{\boldsymbol{\psi}}),$$

and

$$I_5 \approx \frac{1}{2}\mathrm{tr}[\mathbf{H}(\hat{\boldsymbol{\psi}})^{-1}\mathbf{Q}(\hat{\boldsymbol{\psi}})].$$

Lastly we have that

$$\begin{aligned} I_4 &= \int_{\Psi}\frac{\partial \log \pi_b(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}\ \boldsymbol{\psi}\ \pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)d\boldsymbol{\psi} - \int_{\Psi}\frac{\partial \log \pi_b(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}\ \hat{\boldsymbol{\psi}}\ \pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)d\boldsymbol{\psi} \\ &= \frac{\partial \log \pi_b(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}\ \mathbb{E}_{\boldsymbol{\psi}|\mathbf{y}}[\boldsymbol{\psi}] - \frac{\partial \log \pi_b(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}\ \hat{\boldsymbol{\psi}} \\ &\approx 0, \end{aligned} \tag{4.24}$$

following from the assumption $\boldsymbol{\psi}|\mathbf{y} \sim N[\hat{\boldsymbol{\psi}}, \mathbf{H}(\hat{\boldsymbol{\psi}})^{-1}]$, approximately.

Hence, we have proved that the approximation of the expected Shannon information gain is given by Equation (4.20).

Similar to (4.10), for the Laplace Approximation II we have,

$$\begin{aligned} \tilde{U}(\xi) = \frac{1}{k_1}\sum_{h=1}^{k_1}\Bigg[-\frac{1}{2}\log\left[(2\pi)^{q_2}\left|\mathbf{H}(\hat{\boldsymbol{\psi}}_h)^{-1}\right|\right] - \frac{q_2}{2} \\ - \log \pi_b(\hat{\boldsymbol{\psi}}_h) - \frac{1}{2}\mathrm{tr}\left[\mathbf{Q}(\hat{\boldsymbol{\psi}}_h)\mathbf{H}(\hat{\boldsymbol{\psi}}_h)^{-1}\right]\Bigg], \end{aligned}$$

where $\hat{\boldsymbol{\psi}}_h$ is obtained by maximising $\pi_u(\boldsymbol{\psi}|\mathbf{y}_h,\xi)$ with respect to $\boldsymbol{\psi}$ using a quasi-Newton algorithm (see Section 4.3), where $\mathbf{y}_h \sim \pi_e(\mathbf{y}|\xi)$.

A connection between Laplace Approximation I and Laplace Approximation II can be found in Appendix A.

In the next section we consider importance sampling as an alternative approach for approximating the evidence in (4.10).

4.2.3 Approximating the evidence - Importance sampling

The aim is to find an improved way of estimating the evidence $\pi_e(\mathbf{y}|\xi) = \mathbb{E}[\pi_l(\mathbf{y}|\boldsymbol{\psi},\xi)] = \int_{\Psi}\pi_l(\mathbf{y}|\boldsymbol{\psi},\xi)\pi_b(\boldsymbol{\psi})d\boldsymbol{\psi}$, needed in order to estimate the expected utility via (4.10), that will give reduced bias and will be computationally inexpensive.

Another approach for approximating the evidence is *importance sampling*. Importance sampling is a Monte Carlo integration method commonly used for approximating a target integral of interest. Suppose that we wish to approximate

$$I = \mathbb{E}[f(\mathbf{x})] = \int_{\mathbb{R}^{q_1}} f(\mathbf{x})h(\mathbf{x})d\mathbf{x}, \tag{4.25}$$

where $h(\mathbf{x})$ is a probability density function on $\mathbb{R}^{q_1}$ with support[5] Q, so that $h(\mathbf{x}) = 0$ when $\mathbf{x} \notin Q$. Suppose moreover that we have available an *importance density*, $q(\mathbf{x})$, such that $q(\mathbf{x}) > 0$ for all $\mathbf{x} \in Q$ with $f(\mathbf{x})h(\mathbf{x}) > 0$. Then, if $\mathbf{x}_1, \dots, \mathbf{x}_M$ is an independent sample from q,

$$\hat{I} = \frac{1}{M}\sum_{i=1}^{M} \frac{f(\mathbf{x}_i)h(\mathbf{x}_i)}{q(\mathbf{x}_i)} \tag{4.26}$$

is an unbiased estimator of I. The adjustment factor $w_i = \frac{h(\mathbf{x}_i)}{q(\mathbf{x}_i)}$ is called the importance ratio or weight.

Moreover, the variance of $\hat{I}$ is finite provided

$$\int_Q \frac{f(\mathbf{x})^2 h(\mathbf{x})^2}{q(\mathbf{x})} d\mathbf{x} = \mathbb{E}_q\left[\frac{f(\mathbf{x})^2 h(\mathbf{x})^2}{q(\mathbf{x})^2}\right]$$

is finite, which roughly means that q must have heavy enough tails given functions f and h. The optimal importance density is $q(\mathbf{x}) \propto f(\mathbf{x})h(\mathbf{x})$, which makes the variance of $\hat{I}$ zero (Geweke, 1989).

As described in Section 4.1.2 the Monte Carlo approximation of the expected Shannon information gain is given by

$$\tilde{U}(\xi) = \frac{1}{k_1}\sum_{h=1}^{k_1}\left[\log \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_h, \xi) - \log \tilde{\pi}_e^h\right],$$

where $\tilde{\pi}_e^h$ is an estimate of the evidence

$$\pi_e(\mathbf{y}_h|\xi) = \int_\Psi \pi_l(\mathbf{y}_h|\boldsymbol{\psi}, \xi)\pi_b(\boldsymbol{\psi})d\boldsymbol{\psi},$$

which is of the form (4.25) with $f = \pi_l(\mathbf{y}_h|\boldsymbol{\psi}, \xi)$ and $h = \pi_b(\boldsymbol{\psi})$. Hence we can estimate the evidence by taking an independent sample $\tilde{\boldsymbol{\psi}}_{h1}, \dots, \tilde{\boldsymbol{\psi}}_{hk_2}$, from an importance density $q_{\boldsymbol{\psi}}^h(\boldsymbol{\psi})$ and evaluating the importance sampling estimator,

$$\pi_e(\mathbf{y}_h|\xi) \approx \tilde{\pi}_e^h = \frac{1}{k_2}\sum_{k=1}^{k_2} w_{hk}\pi_l(\mathbf{y}_h|\tilde{\boldsymbol{\psi}}_{hk}, \xi). \tag{4.27}$$

Above,

$$w_{hk} = \frac{\pi_b(\tilde{\boldsymbol{\psi}}_{hk})}{q_{\boldsymbol{\psi}}^h(\tilde{\boldsymbol{\psi}}_{hk})},$$

[5] The support of a real-valued function g is the subset of the domain containing those elements which are not mapped to zero, $\mathrm{supp}(g) = \{\mathbf{x} \in \mathcal{X} | g(\mathbf{x}) \neq 0\}$.

and we assume that the likelihood function $\pi_l(\mathbf{y}_h|\tilde{\boldsymbol{\psi}}_{hk},\xi)$ can be evaluated for each $\mathbf{y}_h \sim \pi_l(\mathbf{y}|\boldsymbol{\psi}_h,\xi)$ from the outer sample.

To see how to choose a good importance density, note that if we could sample from the posterior density $\pi_a(\boldsymbol{\psi}|\mathbf{y},\xi)$, with weights $w_{hk} = \pi_b(\tilde{\boldsymbol{\psi}}_{hk})/\pi_a(\tilde{\boldsymbol{\psi}}_{hk}|\mathbf{y}_h,\xi)$, then the approximation (4.27) would be exactly the evidence. Thus, a good importance distribution should be similar to the posterior distribution, i.e. $q_{\boldsymbol{\psi}}(\boldsymbol{\psi})$ should be approximately proportional to $\pi_l(\mathbf{y}|\boldsymbol{\psi},\xi)\pi_b(\boldsymbol{\psi})$.

A number of different approximations to the posterior distribution have been used to form the importance distribution.

(i) Nested importance sampling

Feng (2015) uses a normal approximation to the posterior as an importance distribution,

$$q^h_{\boldsymbol{\psi}}(\boldsymbol{\psi}) \sim N\left(\hat{\boldsymbol{\mu}}^h, \hat{\boldsymbol{\Sigma}}^h\right).$$

The posterior mean $\boldsymbol{\mu}^h$ and posterior covariance $\boldsymbol{\Sigma}^h$ for $\mathbf{y}_h$, are estimated using self-normalised importance sampling from the prior density $\pi_b(\boldsymbol{\psi})$ (Owen, 2013, Chapter 9). Here,

$$\begin{aligned}\boldsymbol{\mu}^h &= \mathbb{E}[\boldsymbol{\psi}|\mathbf{y}_h,\xi]\\ &= \int_\Psi \boldsymbol{\psi}\ \pi_a(\boldsymbol{\psi}|\mathbf{y}_h,\xi)d\boldsymbol{\psi}\\ &= \int_\Psi \boldsymbol{\psi}\frac{\pi_l(\mathbf{y}_h|\boldsymbol{\psi},\xi)}{\pi_e(\mathbf{y}_h|\xi)}\pi_b(\boldsymbol{\psi})d\boldsymbol{\psi}. \end{aligned}\tag{4.28}$$

The evidence $\pi_e(\mathbf{y}_h|\xi)$ is approximated by

$$\begin{aligned}\pi_e(\mathbf{y}_h|\xi) &= \int_\Psi \pi_l(\mathbf{y}_h|\boldsymbol{\psi},\xi)\pi_b(\boldsymbol{\psi})d\boldsymbol{\psi}\\ &\approx \frac{1}{k_1}\sum_{i=1}^{k_1}\pi_l(\mathbf{y}_h|\boldsymbol{\psi}_i,\xi),\end{aligned}\tag{4.29}$$

with $\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_{k_1}$ a sample from the prior density $\pi_b(\boldsymbol{\psi})$. Then the approximate evidence is used with Equation (4.28) to get the estimate of the posterior mean,

$$\hat{\boldsymbol{\mu}}^h = \sum_{i=1}^{k_1}\boldsymbol{\psi}_i\frac{\pi_l(\mathbf{y}_h|\boldsymbol{\psi}_i,\xi)}{\sum_{i=1}^{k_1}\pi_l(\mathbf{y}_h|\boldsymbol{\psi}_i,\xi)}.\tag{4.30}$$

The posterior covariance is given by:

$$\boldsymbol{\Sigma}^h = \mathrm{Var}[\boldsymbol{\psi}|\mathbf{y}_h,\xi]$$

$$= \int_{\Psi} (\boldsymbol{\psi} - \boldsymbol{\mu}^h)(\boldsymbol{\psi} - \boldsymbol{\mu}^h)^{\mathrm{T}} \frac{\pi_l(\mathbf{y}_h|\boldsymbol{\psi}, \xi)}{\pi_e(\mathbf{y}_h|\xi)} \pi_b(\boldsymbol{\psi}) d\boldsymbol{\psi}.$$

Estimate (4.29) of the evidence and the estimate of the posterior mean, $\hat{\boldsymbol{\mu}}^h$, can be used to obtain an estimate of the posterior variance:

$$\hat{\boldsymbol{\Sigma}}^h = \sum_{i=1}^{k_1} (\boldsymbol{\psi}_i - \hat{\boldsymbol{\mu}}^h)(\boldsymbol{\psi}_i - \hat{\boldsymbol{\mu}}^h)^{\mathrm{T}} \frac{\pi_l(\mathbf{y}_h|\boldsymbol{\psi}_i, \xi)}{\sum_{i=1}^{k_1} \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_i, \xi)}. \tag{4.31}$$

The mean (4.30) and variance (4.31) are used to form a multivariate normal importance distribution or, if fatter tails are needed, they are used to form a multivariate t importance distribution to approximate the evidence. This distribution is used to perform the sampling required for approximation (4.27), and approximate the expected Shannon information gain via (4.10). We will refer to this method as *nested importance sampling* (nIS).

Feng (2015) demonstrated that this nested importance sampling scheme is not very robust especially for small inner sample size, k_2. For this reason a minimum effective sample size was introduced, which is used as a cutoff for reverting to using the original naïve Monte Carlo approach of sampling from the prior distribution, as described in Section 4.1.2.

The effective sample size (ESS) is used as a diagnostic to show when the weights,

$$w_{hk} = \frac{\pi_l(\mathbf{y}_h|\tilde{\boldsymbol{\psi}}_{hk}, \xi)}{q^h_{\boldsymbol{\psi}}(\tilde{\boldsymbol{\psi}}_{hk})},$$

are problematic. We assume that $w_{hk} > 0$ (if $w_{hk} = 0$ for all $k = 1, \ldots, k_2$ then the importance sampling has failed). The effective sample size compares the variance of $\hat{I}$ under the importance distribution to the variance that would be obtained if the prior distribution were used as the importance distribution. Different derivations can be used to find a useful expression of ESS (Owen, 2013, Chapter 9). A popular formula is given by

$$\mathrm{ESS} = \frac{1}{\sum_{h=1}^{k_1} (\bar{w}_h)^2},$$

where

$$\bar{w}_h = \frac{\pi_l(\mathbf{y}_h|\boldsymbol{\psi}_h, \xi)}{\sum_{i=1}^{k_1} \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_i, \xi)}, \quad h = 1, \ldots k_1,$$

are the normalised weights.

4.2.4 Other methods for approximating the evidence

Various other approaches have been proposed to approximate the evidence. Newton and Raftery (1994) proposed the use of the harmonic mean estimator. Chib (1995) proposed that the posterior can be estimated by a Monte Carlo average based on draws from the Gibbs sampler. DiCiccio et al. (1997) investigated theoretical and empirical properties of Laplace approximation, Bartlett's adjustment, importance sampling and bridge sampling for estimating the evidence. Gelman and Meng (1998) investigated path sampling for estimating normalising constants. However, all these methods require MCMC samples from the posterior in order to estimate the evidence. Thus it would be computationally expensive to use them in our design utility approximations, since an MCMC chain would need to be run for each iteration of the outer loop (Ryan, 2003).

A possible way to extend the importance sampling approach might be to use annealed importance sampling (Neal, 2001) which adaptively defines an importance sampling distribution to approximate the posterior. However, this method requires a temperature cooling scheme which will be difficult to choose within Monte Carlo loops and optimisation schemes. Skilling (2006) proposed nested sampling for the approximation of the evidence which again is computationally expensive, and likely to be too burdensome for repeated use at each iteration of the outer loop (Friel and Wyse, 2012). Power posteriors were explored by Friel and Pettitt (2008) for estimating the evidence. Similar to the annealed importance sampling method the power posterior approach also requires a temperature scheme to be chosen, which will be difficult within Monte Carlo loops that are repeated many times in the search for an optimal design.

Along similar lines to a Laplace approximation we could consider other deterministic approximations such as Variational Bayes methods (Parise and Welling, 2007) or an integrated nested Laplace approximation (INLA) (Rue et al., 2009). The potential application of these methods within design optimisation problems is an area for future research.

4.3 Laplace importance sampling for approximating the expected utility

We now approximate the evidence, $\pi_e(\mathbf{y}|\xi)$, by importance sampling using a Laplace approximation to the posterior distribution as the importance distribution. This approximation to the evidence is then used to approximate the expected Shannon information gain via (4.10). Compared to naïve Monte Carlo, sampling from an approximation to the posterior distribution is much less likely to result in a zero estimate of the evidence.

Recall that the basic idea of the Laplace approximation is to approximate the log-unnormalised posterior density $\log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi)$, with a quadratic Taylor series expansion around the posterior mode $\hat{\boldsymbol{\psi}}$,

$$\log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi) \approx \log \pi_u(\hat{\boldsymbol{\psi}}|\mathbf{y}, \xi) - \frac{1}{2}\left[(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}}\mathbf{H}(\hat{\boldsymbol{\psi}})(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})\right],$$

where as before,

$$\mathbf{H}(\hat{\boldsymbol{\psi}}) = -\frac{\partial^2 \log \pi_u(\boldsymbol{\psi}|\mathbf{y}, \xi)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathrm{T}}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}.$$

This implies that if $\mathbf{H}(\hat{\boldsymbol{\psi}})$ is positive-definite, then $\exp\left[\log \pi_u(\boldsymbol{\psi}|\mathbf{y},\xi)\right]$ is approximately proportional to the density of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma} = \mathbf{H}(\hat{\boldsymbol{\psi}})^{-1}$.

Kuk (1999) first proposed to use the Laplace approximation to form an importance distribution in the context of estimating the likelihood function of generalised linear mixed models. We implement this idea within a nested Monte Carlo scheme for approximating the expected Shannon information gain utility and explore how the bias and variance change with the inner and outer loop Monte Carlo sample sizes.

We describe two methods, Laplace Importance Sampling (LIS) and Approximate Laplace Importance Sampling (ALIS), outlined in Algorithm 3, which both approximate the evidence using (4.27) with an importance density of the form:

$$q^h_{\boldsymbol{\psi}}(\boldsymbol{\psi}) \sim N(\hat{\boldsymbol{\mu}}^h, \hat{\boldsymbol{\Sigma}}^h).$$

If the posterior density has fatter tails it is possible that a t importance distribution might give a better approximation thus we will include a t distribution in our comparisons as well. Oh and Berger (1993) suggested to use a low number, ν, of degrees of freedom for the t distribution. In the examples presented in this thesis we use $\nu = 5$ which seems to have adequate performance in our numerical examples given in Chapters 5 and 6. The choice of the degrees of freedom in this design problems is an area of future research.

In both cases, $\hat{\boldsymbol{\Sigma}}^h$ is obtained via

$$\hat{\boldsymbol{\Sigma}}^h = \mathbf{H}(\hat{\boldsymbol{\mu}}^h)^{-1},$$

where $\mathbf{H}(\hat{\boldsymbol{\mu}}^h)$ is the negative Hessian of the log-unnormalised posterior density evaluated at the mean $\hat{\boldsymbol{\mu}}^h$ of the importance density.

In LIS, $\hat{\boldsymbol{\mu}}^h$ is obtained via,

$$\hat{\boldsymbol{\mu}}^h = \hat{\boldsymbol{\psi}}_h \in \arg\max_{\boldsymbol{\psi}} \pi_u(\boldsymbol{\psi}|\mathbf{y}_h, \xi). \tag{4.32}$$

The posterior mode, $\hat{\boldsymbol{\psi}}_h$, is found using a quasi-Newton algorithm[6], the Broyden-Fletcher-Goldfarb-Shanno algorithm (Bonnans et al., 2006, Chapter 4); our implementation is from Press et al. (2007, Chapter 10). This algorithm typically converges in a few iterations, using as initial values the parameter values, $\boldsymbol{\psi}_h$, known to have generated the hypothetical data, $\mathbf{y}_h$. We do not use Fisher scoring because in most of our examples the expected Fisher information matrix is difficult to calculate and this matrix can suffer from problems with numerical ill-conditioning, see also Section 4.3.3.

Intuitively, LIS seems to be the natural way to construct a good importance sampling

[6] Quasi-Newton algorithms essentially employ the Newton-Raphson method with an estimated Hessian matrix which is guaranteed to be positive-definite.

distribution. However, for some models it might be the case that the posterior mode $\hat{\boldsymbol{\psi}}_h$ does not change significantly from the initial values $\boldsymbol{\psi}_h$ we have sampled from the prior distribution. Hence the additional computational expense of finding the posterior mode may be unnecessary. For this reason, we also introduce ALIS, a simpler version of LIS.

In ALIS, $\hat{\boldsymbol{\mu}}^h$ is obtained via

$$\hat{\boldsymbol{\mu}}^h = \begin{cases} \boldsymbol{\psi}_h, & \text{if } \mathbf{H}(\boldsymbol{\psi}_h) \text{ is positive-definite} \\ \hat{\boldsymbol{\psi}}_h & \text{otherwise}\,. \end{cases} \tag{4.33}$$

As shown in Equation (4.33), if $\mathbf{H}(\boldsymbol{\psi}_h)$ is positive-definite, the mean of the importance sampling distribution is the true parameter vector $\boldsymbol{\psi}_h$ sampled from the prior distribution, that is known to have generated the hypothetical data $\mathbf{y}_h$. This will reduce the computational expense because we only need to proceed to the optimisation for the few occasions when $\mathbf{H}(\boldsymbol{\psi}_h)$ is not positive-definite.

Algorithm 3: ALIS/LIS Algorithm

```
Generate a sample $\boldsymbol{\psi}_h$, $h = 1, \dots, k_1$, from $\pi_b(\boldsymbol{\psi})$;
for $h = 1, \dots, k_1$ do
    Generate a response $\mathbf{y}_h$ from $\pi_l(\mathbf{y}|\boldsymbol{\psi}_h, \xi)$;
    Calculate $\hat{\boldsymbol{\mu}}^h$ and $\hat{\boldsymbol{\Sigma}}^h$ using Algorithm 4 or Algorithm 5;
    Generate a sample $\{\tilde{\boldsymbol{\psi}}_{hk}\}_{k=1}^{k_2}$, from the importance density $q^h_{\boldsymbol{\psi}}(\boldsymbol{\psi})$ with mean $\hat{\boldsymbol{\mu}}^h$
    and variance $\hat{\boldsymbol{\Sigma}}^h$;
    for $k = 1, \dots, k_2$ do
        Calculate $\tilde{u}_{hk} = \frac{\pi_l(\mathbf{y}_h|\tilde{\boldsymbol{\psi}}_{hk},\xi)\pi_b(\tilde{\boldsymbol{\psi}}_{hk})}{q^h_{\boldsymbol{\psi}}(\tilde{\boldsymbol{\psi}}_{hk})}$;
    Estimate the evidence $\pi_e(\mathbf{y}_h|\xi)$ via $\tilde{\pi}^h_e = \frac{1}{k_2}\sum_{k=1}^{k_2} \tilde{u}_{hk}$;
    Calculate $\tilde{u}_h = \log \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_h, \xi) - \log \tilde{\pi}^h_e$;
Estimate the expected Shannon information gain utility $\tilde{U}(\xi) = \frac{1}{k_1}\sum_{h=1}^{k_1} \tilde{u}_h$;
```

Algorithm 4: LIS step

```
Calculate the posterior mode, $\hat{\boldsymbol{\psi}}_h$ of $\pi_u(\boldsymbol{\psi}|\mathbf{y}_h, \xi)$;
Set $\hat{\boldsymbol{\mu}}^h = \hat{\boldsymbol{\psi}}_h$ and $\hat{\boldsymbol{\Sigma}}^h = \mathbf{H}(\hat{\boldsymbol{\mu}}^h)^{-1}$
```

Ryan et al. (2015) first suggested implementing LIS in the Bayesian design framework for a particular Pharmacokinetics example, and using different utility functions to that employed in this thesis. LIS was then used within an MCMC algorithm in order to search for near-optimal designs for the particular PK study.

Recently, Beck et al. (2018) examined the performance of LIS under statistical models with fixed error variance and designs consisting of a single replicated design point. They provided some theoretical error analysis for these examples, and limited comparisons

Algorithm 5: ALIS step

```
Calculate $\mathbf{H}(\boldsymbol{\psi}_h)$;
if $\mathbf{H}(\boldsymbol{\psi}_h)$ positive-definite then
    Set $\hat{\boldsymbol{\mu}}^h = \boldsymbol{\psi}_h$ and $\hat{\boldsymbol{\Sigma}}^h = \mathbf{H}(\hat{\boldsymbol{\mu}}^h)^{-1}$;
else
    Calculate the posterior mode, $\hat{\boldsymbol{\psi}}_h$ of $\pi_u(\boldsymbol{\psi}|\mathbf{y}_h, \xi)$;
    Set $\hat{\boldsymbol{\mu}}^h = \hat{\boldsymbol{\psi}}_h$ and $\hat{\boldsymbol{\Sigma}}^h = \mathbf{H}(\hat{\boldsymbol{\mu}}^h)^{-1}$;
```

to other methods. Optimal values of the inner and outer sample sizes are presented to achieve given error tolerances in the estimate of the expected Shannon information gain for minimum computational resource. In this thesis, in the numerical comparisons in Chapter 5, we take the opposite approach of assuming a fixed computational budget for both samples. Interestingly, Beck et al. (2018) dismissed ALIS due to discrepancy between $\boldsymbol{\psi}_h$ and $\hat{\boldsymbol{\psi}}_h$; in Chapter 5, we find the effectiveness of ALIS is very much dependent on the example under study.

4.3.1 ALIS/LIS for nuisance parameters

We will now study the case where the model contains nuisance parameters. Any parameter, e.g. the variance components, which is not of immediate interest is called a nuisance parameter; such parameters must still taken into account when studying the parameters which are of interest.

We partition the parameter vector as $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ are the p_θ parameters of interest and $\boldsymbol{\gamma} \in \Gamma$ are the p_γ nuisance parameters. The expected utility now takes the form:

$$\begin{aligned}
U(\xi) &= \mathbb{E}[u(\xi, \boldsymbol{\theta}, \mathbf{y})] \\
&= \int_{\boldsymbol{\Theta}} \int_{\mathcal{Y}} [\log \pi_M(\mathbf{y}|\boldsymbol{\theta}, \xi) - \log \pi_e(\mathbf{y}|\xi)] \pi(\mathbf{y}, \boldsymbol{\theta}|\xi) d\mathbf{y} d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\Theta}} \int_{\mathcal{Y}} \log \frac{\pi_M(\mathbf{y}|\boldsymbol{\theta}, \xi)}{\pi_e(\mathbf{y}|\xi)} \pi(\mathbf{y}, \boldsymbol{\theta}|\xi) d\mathbf{y} d\boldsymbol{\theta},
\end{aligned}$$

where

$$\pi_M(\mathbf{y}|\boldsymbol{\theta}, \xi) = \int_\Gamma \pi(\mathbf{y}, \boldsymbol{\gamma}|\boldsymbol{\theta}, \xi) d\boldsymbol{\gamma} = \int_\Gamma \pi_l(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \xi) \pi_b(\boldsymbol{\gamma}|\boldsymbol{\theta}) d\boldsymbol{\gamma},$$

is the marginal distribution of the data, $\mathbf{y}$, after integrating out the nuisance parameters, $\boldsymbol{\gamma}$.

Hence the approximation (4.10) of the expected Shannon information gain, becomes

$$\tilde{U}(\xi) \approx \frac{1}{k_1} \sum_{h=1}^{k_1} \left[\log \tilde{\pi}_M^h - \log \tilde{\pi}_e^h \right],$$

with, as before,

$$\pi_e(\mathbf{y}_h|\xi) \approx \tilde{\pi}_e^h = \frac{1}{k_2}\sum_{k=1}^{k_2} \pi_l(\mathbf{y}_h|\tilde{\boldsymbol{\psi}}_{hk},\xi)\frac{\pi_b(\tilde{\boldsymbol{\psi}}_{hk})}{q_{\boldsymbol{\psi}}^h(\tilde{\boldsymbol{\psi}}_{hk})},$$

$\{\tilde{\boldsymbol{\psi}}_{hk}\}_{k=1}^{k_2} = \{\tilde{\boldsymbol{\theta}}_{hk}, \tilde{\gamma}_{hk}\}_{k=1}^{k_2}$, and now a second importance sampling approximation is used to estimate the likelihood marginal to the nuisance parameters $\boldsymbol{\gamma}$,

$$\pi_M(\mathbf{y}_h|\boldsymbol{\theta}_h,\xi) \approx \tilde{\pi}_M^h = \frac{1}{k_3}\sum_{s=1}^{k_3} \pi_l(\mathbf{y}_h|\boldsymbol{\theta}_h,\tilde{\tilde{\gamma}}_{hs},\xi)\frac{\pi_b(\tilde{\tilde{\gamma}}_{hs}|\boldsymbol{\theta}_h)}{q_{\gamma|\theta}^h(\tilde{\tilde{\gamma}}_{hs})},$$

$\{\tilde{\tilde{\gamma}}_{hs}\}_{s=1}^{k_3}$. Here $q_{\gamma|\theta}^h$ is the importance density and $\pi_b(\boldsymbol{\gamma}|\boldsymbol{\theta})$ is the prior density of the nuisance parameters $\boldsymbol{\gamma}$ given the parameters of interest $\boldsymbol{\theta}$.

We choose $q_{\gamma|\theta}^h$ to approximate the conditional posterior density $\pi_a(\boldsymbol{\gamma}|\mathbf{y},\boldsymbol{\theta})$ via a multivariate normal approximation to the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$,

$$\left.\begin{pmatrix}\boldsymbol{\theta}_h\\ \gamma_h\end{pmatrix}\right|\mathbf{y}_h,\xi \quad \sim N\left(\hat{\boldsymbol{\mu}}^h,\hat{\boldsymbol{\Sigma}}^h\right), \tag{4.34}$$

where $\hat{\boldsymbol{\mu}}^h = \left[\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^h \;\; \hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^h\right]^{\mathrm{T}}$ is defined by (4.32) for LIS, and (4.33) for ALIS, and

$$\hat{\boldsymbol{\Sigma}}^h = \left[\mathbf{H}(\hat{\boldsymbol{\mu}}^h)\right]^{-1} = \begin{bmatrix}\mathbf{H}_{\boldsymbol{\theta\theta}}^h & \mathbf{H}_{\boldsymbol{\theta\gamma}}^h\\ (\mathbf{H}_{\boldsymbol{\theta\gamma}}^h)^{\mathrm{T}} & \mathbf{H}_{\boldsymbol{\gamma\gamma}}^h\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{H}_h^{\boldsymbol{\theta\theta}} & \mathbf{H}_h^{\boldsymbol{\theta\gamma}}\\ (\mathbf{H}_h^{\boldsymbol{\theta\gamma}})^{\mathrm{T}} & \mathbf{H}_h^{\boldsymbol{\gamma\gamma}}\end{bmatrix}.$$

Here

$$\mathbf{H}_{\boldsymbol{\theta\theta},ij}^h = -\left[\frac{\partial^2\log\pi_u(\hat{\boldsymbol{\mu}}^h|\mathbf{y}_h,\xi)}{\partial\theta_i\partial\theta_j}\right],\; \mathbf{H}_{\boldsymbol{\theta\gamma},ij'}^h = -\left[\frac{\partial^2\log\pi_u(\hat{\boldsymbol{\mu}}^h|\mathbf{y}_h,\xi)}{\partial\theta_i\partial\gamma_{j'}}\right],$$

$$\mathbf{H}_{\boldsymbol{\gamma\gamma},i'j'}^h = -\left[\frac{\partial^2\log\pi_u(\hat{\boldsymbol{\mu}}^h|\mathbf{y}_h,\xi)}{\partial\gamma_{i'}\partial\gamma_{j'}}\right],$$

where $i,j = 1,\ldots,p_\theta$ and $i',j' = 1,\ldots,p_\gamma$. There are partition formulas to obtain $\mathbf{H}^{\boldsymbol{\theta\theta}}$ from $\mathbf{H}_{\boldsymbol{\theta\theta}}$, $\mathbf{H}_{\boldsymbol{\theta\gamma}}$ and $\mathbf{H}_{\boldsymbol{\gamma\gamma}}$, see, for example, Graybill (1983, Chapter 8).

It follows from standard results on multivariate normal distributions (Banerjee et al., 2004, Chapter 2) that if (4.34) holds then:

$$\gamma_h \mid \mathbf{y}_h,\boldsymbol{\theta}_h,\xi \sim N\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^h + \mathbf{H}_h^{\boldsymbol{\gamma\theta}}\left(\mathbf{H}_h^{\boldsymbol{\theta\theta}}\right)^{-1}(\boldsymbol{\theta}_h - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^h),\; \mathbf{H}_h^{\boldsymbol{\gamma\gamma}} - \mathbf{H}_h^{\boldsymbol{\gamma\theta}}\left(\mathbf{H}_h^{\boldsymbol{\theta\theta}}\right)^{-1}\mathbf{H}_h^{\boldsymbol{\theta\gamma}}\right),$$

with $\mathbf{H}_h^{\boldsymbol{\gamma\theta}} = (\mathbf{H}_h^{\boldsymbol{\theta\gamma}})^{\mathrm{T}}$.

We use this approximate conditional posterior as the importance distribution to approximate the marginal likelihood to integrate out the nuisance parameters $\boldsymbol{\gamma}$, and approximate expected Shannon information gain with ALIS and LIS as outlined in Algorithm 6.

Algorithm 6: ALIS/LIS Algorithm for nuisance parameters

```
Generate a sample $\boldsymbol{\psi}_h = (\boldsymbol{\theta}_h, \boldsymbol{\gamma}_h)^{\mathrm{T}}$, $h = 1, \ldots, k_1$, from $\pi_b(\boldsymbol{\psi})$;
for $h = 1, \ldots, k_1$ do
    Generate a response $\mathbf{y}_h$ from $\pi_l(\mathbf{y}|\boldsymbol{\psi}_h, \xi)$;
    Calculate $\hat{\boldsymbol{\mu}}^h$ and $\hat{\boldsymbol{\Sigma}}^h$ using Algorithm 4 or Algorithm 5;
    Generate a sample $\{\tilde{\boldsymbol{\psi}}_{hk}\}_{k=1}^{k_2} = \{\tilde{\boldsymbol{\theta}}_{hk}, \tilde{\boldsymbol{\gamma}}_{hk}\}_{k=1}^{k_2}$, from the importance density $q_{\boldsymbol{\psi}}^h(\boldsymbol{\psi})$
    with mean $\hat{\boldsymbol{\mu}}^h$ and variance $\hat{\boldsymbol{\Sigma}}^h$;
    for $k = 1, \ldots, k_2$ do
        Calculate $\tilde{u}_{hk} = \frac{\pi_l(\mathbf{y}_h|\tilde{\boldsymbol{\psi}}_{hk}, \xi)\pi_b(\tilde{\boldsymbol{\psi}}_{hk})}{q_{\boldsymbol{\psi}}^h(\tilde{\boldsymbol{\psi}}_{hk})}$;
    Estimate the evidence $\pi_e(\mathbf{y}_h|\xi)$ by $\tilde{\pi}_e^h = \frac{1}{k_2}\sum_{k=1}^{k_2} \tilde{u}_{hk}$;
    Calculate $\hat{\boldsymbol{\mu}}^h_{\gamma|\theta} = \hat{\boldsymbol{\mu}}^h_{\boldsymbol{\gamma}} + \mathbf{H}^{\gamma\boldsymbol{\theta}}_h \left(\mathbf{H}^{\boldsymbol{\theta\theta}}_h\right)^{-1} (\boldsymbol{\theta}_h - \hat{\boldsymbol{\mu}}^h_{\boldsymbol{\theta}})$, $\hat{\boldsymbol{\Sigma}}^h_{\gamma|\theta} = \mathbf{H}^{\gamma\gamma}_h - \mathbf{H}^{\gamma\boldsymbol{\theta}}_h \left(\mathbf{H}^{\boldsymbol{\theta\theta}}_h\right)^{-1} \mathbf{H}^{\boldsymbol{\theta}\gamma}_h$;
    Generate a sample $\{\tilde{\tilde{\boldsymbol{\gamma}}}_{hs}\}_{s=1}^{k_3}$, from the importance density $q^h_{\gamma|\theta}(\boldsymbol{\gamma})$ with mean $\hat{\boldsymbol{\mu}}^h_{\gamma|\theta}$
    and variance $\hat{\boldsymbol{\Sigma}}^h_{\gamma|\theta}$;
    for $s = 1, \ldots, k_3$ do
        Calculate $\tilde{u}_{hs} = \frac{\pi_l(\mathbf{y}_h|\boldsymbol{\theta}_h, \tilde{\tilde{\boldsymbol{\gamma}}}_{hs}, \xi)\pi_b(\tilde{\tilde{\boldsymbol{\gamma}}}_{hs})}{q^h_{\gamma|\theta}(\tilde{\tilde{\boldsymbol{\gamma}}}_{hs})}$;
    Estimate the marginal likelihood $\pi_M(\mathbf{y}_h|\boldsymbol{\theta}_h, \xi)$ by $\tilde{\pi}_M^h = \frac{1}{k_3}\sum_{s=1}^{k_3} \tilde{u}_{hs}$;
    Calculate $\tilde{u}_h = \log \tilde{\pi}_M^h - \log \tilde{\pi}_e^h$;
Estimate the expected Shannon information gain utility by $\tilde{U}(\xi) = \frac{1}{k_1}\sum_{h=1}^{k_1} \tilde{u}_h$;
```

4.3.2 ALIS/LIS for transformed parameters

Often we need to construct an importance distribution that guarantees that the parameters of interest satisfy some constraints, e.g. are always positive. We do this by constructing a normal approximation to the posterior distribution of a transformed version of the parameter, $\boldsymbol{\psi}' = T(\boldsymbol{\psi})$, e.g. $T = \log$ to ensure positivity. Another reason for transforming the parameters is to put them on a scale where the normal approximation to the posterior distribution is more accurate, e.g. when the posterior distribution is log-normal then the transformation $\boldsymbol{\psi}' = (\log \psi_1, \ldots, \log \psi_{q_2})^{\mathrm{T}}$ will make the approximation exact.

Let π_l, $\pi_l^{\psi'}$ denote the likelihood in $\boldsymbol{\psi}$ and $\boldsymbol{\psi}'$ parameterisations, respectively, and similarly let π_b, $\pi_b^{\psi'}$ denote the prior densities for $\boldsymbol{\psi}$ and $\boldsymbol{\psi}'$ respectively. Then the unnormalised posterior denisty of $\boldsymbol{\psi}'$ is given by,

$$\begin{aligned}\pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y}, \xi) &= \pi_l^{\psi'}(\mathbf{y}|\boldsymbol{\psi}', \xi)\pi_b^{\psi'}(\boldsymbol{\psi}') \\ &= \pi_l(\mathbf{y}|T^{-1}(\boldsymbol{\psi}'), \xi)\pi_b(T^{-1}(\boldsymbol{\psi}')) \left|\det \mathcal{J}\left[T^{-1}(\boldsymbol{\psi}')\right]\right|, \qquad (4.35)\end{aligned}$$

where $\mathcal{J}\left[T^{-1}(\boldsymbol{\psi}')\right]$ is the Jacobian matrix[7] of T^{-1}. It is necessary to calculate the

[7] The Jacobian matrix for the transformation $(x, y) \to (z, u)$ is:

$$\mathcal{J} = \begin{bmatrix} \frac{dz}{dx} & \frac{dz}{dy} \\ \frac{du}{dx} & \frac{du}{dy} \end{bmatrix}.$$

negative Hessian of the log-unnormalised posterior density (4.35) with respect to $\boldsymbol{\psi}'$, $\mathbf{H}_{\boldsymbol{\psi}'}(\boldsymbol{\psi}')$, i.e. find the derivatives of $\log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y},\xi)$ with respect to $\boldsymbol{\psi}'$. For $T(\psi_i) = \log \psi_i$, $i = 1, \ldots, q_2$,

$$\frac{\partial T^{-1}(\psi_i')}{\partial \psi_j'} = \frac{\partial \exp \psi_i'}{\partial \psi_j'} = \begin{cases} \exp \psi_i', & i = j \\ 0, & i \neq j\,, \end{cases}$$

$$\log \left| \frac{\partial T^{-1}(\psi_i')}{\partial \psi_i'} \right| = \log \exp \psi_i' = \psi_i' \qquad \Leftrightarrow \qquad \frac{\partial \log \left| \frac{\partial T^{-1}(\psi_i')}{\partial \psi_i'} \right|}{\partial \psi_i'} = 1,$$

and hence

$$\begin{aligned} \frac{\partial \log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y},\xi)}{\partial \psi_i'} &= \frac{\partial \log \pi_l(\mathbf{y}|T^{-1}(\boldsymbol{\psi}'),\xi)}{\partial \psi_i'} + \frac{\partial \log \pi_b(T^{-1}(\boldsymbol{\psi}'))}{\partial \psi_i'} + 1, \\ \frac{\partial^2 \log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y},\xi)}{\partial \psi_i' \partial \psi_j'} &= \frac{\partial^2 \log \pi_l(\mathbf{y}|T^{-1}(\boldsymbol{\psi}'),\xi)}{\partial \psi_i' \partial \psi_j'} + \frac{\partial^2 \log \pi_b(T^{-1}(\boldsymbol{\psi}'))}{\partial \psi_i' \partial \psi_j'}, \end{aligned} \tag{4.36}$$

where $j = 1, \ldots, q_2$.

To estimate the evidence, $\pi_e(\mathbf{y}|\xi)$, in the approximate expected Shannon information gain (4.10) by importance sampling, as shown in Equation (4.27), we sample $\{\tilde{\boldsymbol{\psi}}'_{hk}\}_{k=1}^{k_2}$ from the importance density of the transformed parameters $\boldsymbol{\psi}'$,

$$q_{\boldsymbol{\psi}'}^h(\boldsymbol{\psi}') \propto N\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\psi}'}^h, \boldsymbol{\Sigma}_{\boldsymbol{\psi}'}^h\right),$$

where $\hat{\boldsymbol{\mu}}_{\boldsymbol{\psi}'}^h$ is defined by

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\psi}'}^h = \hat{\boldsymbol{\psi}}_h' \in \arg\max_{\boldsymbol{\psi}'} \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y}_h,\xi), \tag{4.37}$$

for LIS (Algorithm 8), and

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\psi}'}^h = \begin{cases} \boldsymbol{\psi}_h', & \text{if } \mathbf{H}_{\boldsymbol{\psi}'}(\boldsymbol{\psi}_h') \text{ is positive-definite} \\ \hat{\boldsymbol{\psi}}_h' & \text{otherwise}\,, \end{cases} \tag{4.38}$$

for ALIS (Algorithm 9). The variance of the importance density is

$$\boldsymbol{\Sigma}_{\boldsymbol{\psi}'}^h = \left[\mathbf{H}_{\boldsymbol{\psi}'}(\hat{\boldsymbol{\mu}}_{\boldsymbol{\psi}'}^h)\right]^{-1}. \tag{4.39}$$

It is necessary to work out the implied importance density for the untransformed parameters $\boldsymbol{\psi}$. It follows directly from the form of $q_{\boldsymbol{\psi}'}^h$ that

$$q_{\boldsymbol{\psi}}^h(\boldsymbol{\psi}) = q_{\boldsymbol{\psi}'}^h\left(T(\boldsymbol{\psi})\right) \left|\det \mathcal{J}\left[T(\boldsymbol{\psi})\right]\right|,$$

where $\mathcal{J}\left[T(\boldsymbol{\psi})\right]$ is the Jacobian matrix of T, the transformation from $\boldsymbol{\psi}$ to $\boldsymbol{\psi}'$.

The approximation of the expected Shannon information gain for the transformed parameters $\boldsymbol{\psi}'$ with ALIS and LIS is outlined in Algorithm 7.

Algorithm 7: ALIS/LIS Algorithm for transformed parameters

```
Generate a sample $\boldsymbol{\psi}_h$, $h = 1, \ldots, k_1$, from $\pi_b(\boldsymbol{\psi})$;
Calculate the transformed sample $\{\boldsymbol{\psi}'_h\}_{h=1}^{k_1} = \{T(\boldsymbol{\psi}_h)\}_{h=1}^{k_1}$;
for $h = 1, \ldots, k_1$ do
    Generate a response $\mathbf{y}_h$ from $\pi_l(\mathbf{y}|\boldsymbol{\psi}_h, \xi)$;
    Calculate $\hat{\boldsymbol{\mu}}^h_{\boldsymbol{\psi}'}$ and $\hat{\boldsymbol{\Sigma}}^h_{\boldsymbol{\psi}'}$ using Algorithm 8 or Algorithm 9;
    Generate a sample $\{\tilde{\boldsymbol{\psi}}'_{hk}\}_{k=1}^{k_2}$, from the importance density $q^h_{\boldsymbol{\psi}'}(\boldsymbol{\psi}')$ with mean
    $\hat{\boldsymbol{\mu}}^h_{\boldsymbol{\psi}'}$ and variance $\hat{\boldsymbol{\Sigma}}^h_{\boldsymbol{\psi}'}$, and calculate $\tilde{\boldsymbol{\psi}}_{hk} = T^{-1}(\tilde{\boldsymbol{\psi}}'_{hk})$;
    for $k = 1, \ldots, k_2$ do
        Calculate $\tilde{u}_{hk} = \frac{\pi_l(\mathbf{y}_h|\tilde{\boldsymbol{\psi}}_{hk},\xi)\pi_b\left(\tilde{\boldsymbol{\psi}}_{hk}\right)}{q^h_{\boldsymbol{\psi}'}\left(T(\tilde{\boldsymbol{\psi}}_{hk})\right)\left|\det \mathcal{J}\left[T(\tilde{\boldsymbol{\psi}}_{hk})\right]\right|}$;
    Estimate the evidence $\pi_e(\mathbf{y}_h|\xi)$ via $\tilde{\pi}^h_e = \frac{1}{k_2}\sum_{k=1}^{k_2} \tilde{u}_{hk}$;
    Calculate $\tilde{u}_h = \log \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_h, \xi) - \log \tilde{\pi}^h_e$;
Estimate the expected Shannon information gain utility by $\tilde{U}(\xi) = \frac{1}{k_1}\sum_{h=1}^{k_1} \tilde{u}_h$;
```

Algorithm 8: LIS step for transformed parameters

```
Calculate the posterior mode, $\hat{\boldsymbol{\psi}}'_h$ of $\pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y}_h, \xi)$;
Set $\hat{\boldsymbol{\mu}}^h_{\boldsymbol{\psi}'} = \hat{\boldsymbol{\psi}}'_h$ and $\hat{\boldsymbol{\Sigma}}^h_{\boldsymbol{\psi}'} = \mathbf{H}_{\boldsymbol{\psi}'}(\hat{\boldsymbol{\mu}}^h_{\boldsymbol{\psi}'})^{-1}$
```

Algorithm 9: ALIS step for transformed parameters

```
Calculate $\mathbf{H}_{\boldsymbol{\psi}'}(\boldsymbol{\psi}'_h)$;
if $\mathbf{H}_{\boldsymbol{\psi}'}(\boldsymbol{\psi}'_h)$ positive-definite then
    Set $\hat{\boldsymbol{\mu}}^h_{\boldsymbol{\psi}'} = \boldsymbol{\psi}'_h$ and $\hat{\boldsymbol{\Sigma}}^h_{\boldsymbol{\psi}'} = \mathbf{H}_{\boldsymbol{\psi}'}(\hat{\boldsymbol{\mu}}^h_{\boldsymbol{\psi}'})^{-1}$;
else
    Calculate the posterior mode, $\hat{\boldsymbol{\psi}}'_h$ of $\pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y}_h, \xi)$;
    Set $\hat{\boldsymbol{\mu}}^h_{\boldsymbol{\psi}'} = \hat{\boldsymbol{\psi}}'_h$ and $\hat{\boldsymbol{\Sigma}}^h_{\boldsymbol{\psi}'} = \mathbf{H}_{\boldsymbol{\psi}'}(\hat{\boldsymbol{\mu}}^h_{\boldsymbol{\psi}'})^{-1}$;
```

4.3.3 Different methods for constructing the importance sampling distribution

In order to arrive at the formulation of ALIS and LIS described in Section 4.3, we first attempted to construct the importance sampling distribution in a number of different ways.

First, we tried to use $\hat{\boldsymbol{\mu}}^h = \boldsymbol{\psi}_h$ as the mean of the importance distribution for all $h = 1, \ldots, k_1$, and the inverse of the observed Fisher information matrix as the covariance matrix. However the observed Fisher information matrix is often not positive-definite and hence is not always invertible. Hence, we attempted to regularise the observed

Fisher information matrix by adding the Hessian of the log-prior density which will always be positive-definite. Even so, the negative Hessian of the log-unnormalised posterior density is often not positive-definite at $\boldsymbol{\psi}_h$ but will be at $\hat{\boldsymbol{\psi}}_h$. Instead we could have used the expected Fisher information matrix, which is by definition positive-definite. However, often it is non-trivial to obtain the expected Fisher information matrix for some nonlinear models, especially when σ_ε^2 is integrated out of the likelihood function. Also, although the expected Fisher information matrix is positive-definite in theory, it is often numerically close to singular.

Second, we tried to identify an approximate posterior mean by performing a very small number (one or two) of Newton-Raphson steps, using $\boldsymbol{\psi}_h$ as the starting values. However, again the negative Hessian of the log-unnormalised posterior density was often not positive-definite. In the final form of ALIS, we thus decided to selectively use a quasi-Newton algorithm, initialized at $\boldsymbol{\psi}_h$, to obtain the posterior mode in cases where the negative Hessian of the log-unnormalised posterior density was numerically indefinite or singular.

In the next chapter we perform the first thorough comparison of the different methods introduced in the previous sections (Section 4.1.3, Section 4.2.1, Section 4.2.2 and Section 4.2.3) and the new proposed methods, ALIS and LIS (Section 4.3), in terms of their relative performance and computational cost in the context of expected Shannon information gain estimation.

4.4 Summary

In this chapter we described several existing methods for numerical estimation of the expected Shannon information gain utility, and provided details for two unexplored methods, ALIS and LIS. We described the general approach to fully Bayesian designs of experiments, dealing with several important challenges including estimation of the utility function, and design optimisation. We illustrated through the simple linear example how the most simple approach (Naïve Monte Carlo) used to approximate the expected Shannon information gain results in positive bias and overestimation of the information gain for a given design. In the next chapter, we show that ALIS and LIS give an efficient compromise between accuracy and computational cost of estimation of the expected utility.

Chapter 5

Assessments of Shannon information gain approximations in Bayesian design

In this chapter we perform the first thorough comparison of the different methods introduced in Chapter 4 in terms of their relative performance and computational cost in the context of expected Shannon information gain estimation. An optimisation algorithm, the approximate coordinate exchange (ACE) algorithm, is then described to optimise the expected utility. We combine ACE with the different methods in Chapter 4 to find Bayesian optimal designs for nonlinear models.

5.1 Introduction

We aim to approximate the expected Shannon information gain using

$$\tilde{U}(\xi) = \frac{1}{k_1} \sum_{h=1}^{k_1} \left[\log \pi_l(\mathbf{y}_h | \boldsymbol{\psi}_h, \xi) - \log \tilde{\pi}_e^h \right],$$

with a variety of approximations $\tilde{\pi}_e^h$ to the evidence, $\pi_e(\mathbf{y}|\xi)$; see Table 5.1.

Using expression (4.3) we can also apply the Laplace approximation II (LA2), described in Section 4.2.2, where an approximation to the expected Shannon information gain takes the form

$$\tilde{U}(\xi) = \frac{1}{k_1} \sum_{h=1}^{k_1} \left[-\frac{1}{2} \log \left[(2\pi)^{q_2} \left| \mathbf{H}(\hat{\boldsymbol{\psi}}_h)^{-1} \right| \right] - \frac{q_2}{2} \right.$$
$$\left. - \log \pi_b(\hat{\boldsymbol{\psi}}_h) - \frac{1}{2} \mathrm{tr} \left[\mathbf{Q}(\hat{\boldsymbol{\psi}}_h) \mathbf{H}(\hat{\boldsymbol{\psi}}_h)^{-1} \right] \right],$$

which does not directly use an approximation to the evidence.

Method	Approximate Evidence
Naïve Monte Carlo (nMC)	$\tilde{\pi}^h_e = \frac{1}{k_2}\sum_{k=1}^{k_2} \pi_l(\mathbf{y}_h\|\tilde{\boldsymbol{\psi}}_{hk},\xi), \quad \tilde{\boldsymbol{\psi}}_{hk} \sim \pi_b(\boldsymbol{\psi})$
Importance Sampling $q^h_{\boldsymbol{\psi}}$ is a $N\left(\hat{\boldsymbol{\mu}}^h,\hat{\boldsymbol{\Sigma}}^h\right)$ or $t_\nu\left(q_2,\hat{\boldsymbol{\mu}}^h,\frac{\nu-2}{\nu}\hat{\boldsymbol{\Sigma}}^h\right)$ density	$\tilde{\pi}^h_e = \frac{1}{k_2}\sum_{k=1}^{k_2} \frac{\pi_b(\tilde{\boldsymbol{\psi}}_{hk})}{q^h_{\boldsymbol{\psi}}(\tilde{\boldsymbol{\psi}}_{hk})}\pi_l(\mathbf{y}_h\|\tilde{\boldsymbol{\psi}}_{hk},\xi), \quad \tilde{\boldsymbol{\psi}}_{hk} \sim q^h_{\boldsymbol{\psi}}(\boldsymbol{\psi})$ For nuisance parameters or transformed parameters see Sections 4.3.1 and 4.3.2, respectively
Nested Importance sampling (nIS)	$\hat{\boldsymbol{\mu}}^h$ and $\hat{\boldsymbol{\Sigma}}^h$ defined via Equation (4.30) and Equation (4.31), respectively
Laplace Importance sampling (LIS)	$\hat{\boldsymbol{\mu}}^h = \hat{\boldsymbol{\psi}}_h$ $\hat{\boldsymbol{\Sigma}}^h = \mathbf{H}(\hat{\boldsymbol{\psi}}_h)^{-1}$ where $\mathbf{H}(\hat{\boldsymbol{\psi}}_h)$ is the negative Hessian of $\log \pi_u(\boldsymbol{\psi}\|\mathbf{y},\xi)$ evaluated at $\hat{\boldsymbol{\psi}}_h \in \arg\max_{\boldsymbol{\psi}} \pi_u(\boldsymbol{\psi}\|\mathbf{y}_h,\xi)$
Approximate Laplace Importance sampling (ALIS)	$\hat{\boldsymbol{\mu}}^h = \begin{cases} \boldsymbol{\psi}_h, & \mathbf{H}(\boldsymbol{\psi}_h) \text{ positive-definite} \\ \hat{\boldsymbol{\psi}}_h & \text{otherwise} \end{cases}$ $\hat{\boldsymbol{\Sigma}}^h = \mathbf{H}(\hat{\boldsymbol{\mu}}^h)^{-1}$
Laplace Approximation I (LA1)	$\tilde{\pi}^h_e = \log \pi_u(\hat{\boldsymbol{\psi}}_h\|\mathbf{y}_h,\xi) + \frac{1}{2}\log\left[(2\pi)^{q_2}\left\|\mathbf{H}(\hat{\boldsymbol{\psi}}_h)^{-1}\right\|\right]$

Table 5.1: Methods described in Chapter 4 for approximating the evidence in the expected Shannon information gain

Firstly we show the advantage of ALIS and LIS over nMC through the simple linear regression example, and then we compare all the methods for three nonlinear regression models. In the examples in Section 5.1.2 and Section 5.1.3 we also assess the 'reuse' method (4.11) (Huan and Marzouk, 2013) for approximating the expected utility.

5.1.1 Linear Regression example (continued)

Continuing the example from Section 4.1.3, we will now compare nMC with the new methods for approximating the expected utility, ALIS and LIS, proposed in Section 4.3. We acknowledge that this example is unusually favourable to LIS as here the true posterior is a normal distribution, and so the approximate posterior will be exactly

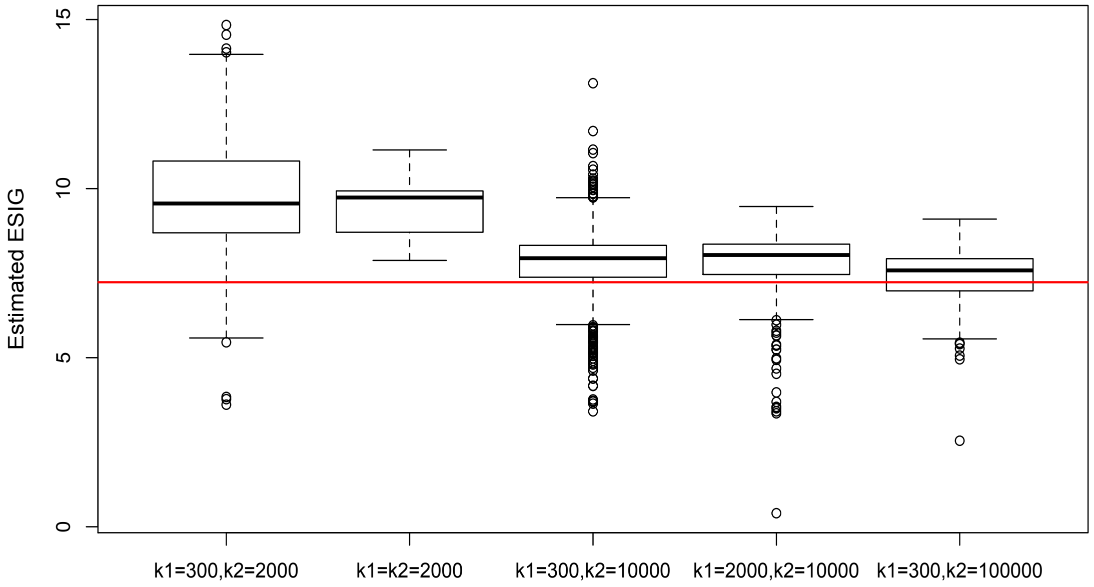


Figure 5.1: Estimated expected Shannon information gain for the linear model (4.12) using nMC and different combinations of k_1 and k_2, and the true value of the Shannon information gain obtained using (4.16) (red line)

equal to the true posterior distribution. Hence the approximation $\tilde{\pi}_e^h$ to the evidence will be exact.

Figures 5.1, 5.2 and 5.3, show the distribution of 100 estimates of the expected Shannon information gain utility approximated using nMC, ALIS and LIS, respectively, for different pairs of inner and outer loop sizes. The scale in these three figures is chosen to be comparable. It is clear that the estimate of the expected utility in Figures 5.2 and 5.3, is much closer to the true value even for very small values of the inner and outer loop sample sizes. For this reason we do not include results for $k_1 = 300, k_2 = 100000$. In Figure 5.1, results for $k_1 = k_2 = 300$ are omitted due to occurrence of the zero evidence problem (see Section 4.1.3).

Figure 5.4 shows the same results as Figure 5.2 for ALIS but with a smaller y-axis scale than before in order to better illustrate any differences between the different pairs of k_1 and k_2. We notice that for small inner loop sample size ($k_2 = 300$) the bias is nonzero, but substantially smaller than when using nMC. For large values of the inner loop sample size ($k_2 = 2000, 10000$) the bias is negligible. The variance of the approximation of the expected utility decreases as k_1 increases. We will discuss the interaction between sample size and computational expense further in the following examples.

Figure 5.5 shows the same results as Figure 5.3 for LIS but again on a smaller y-axis

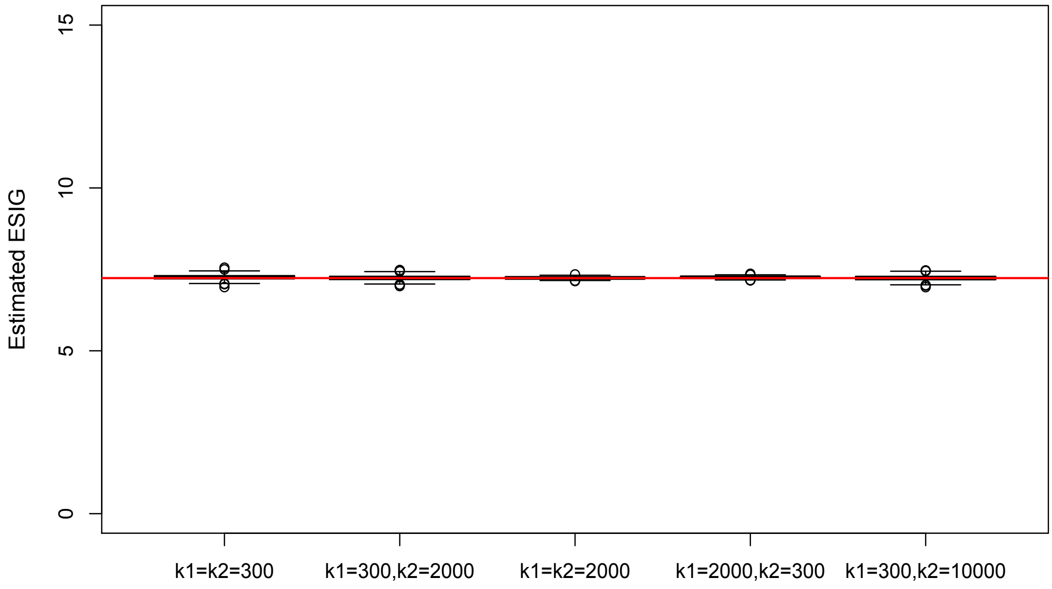


Figure 5.2: Estimated ESIG for the linear model (4.12) using ALIS and different combinations of k_1 and k_2, and the true value of the Shannon information gain obtained using (4.16) (red line)

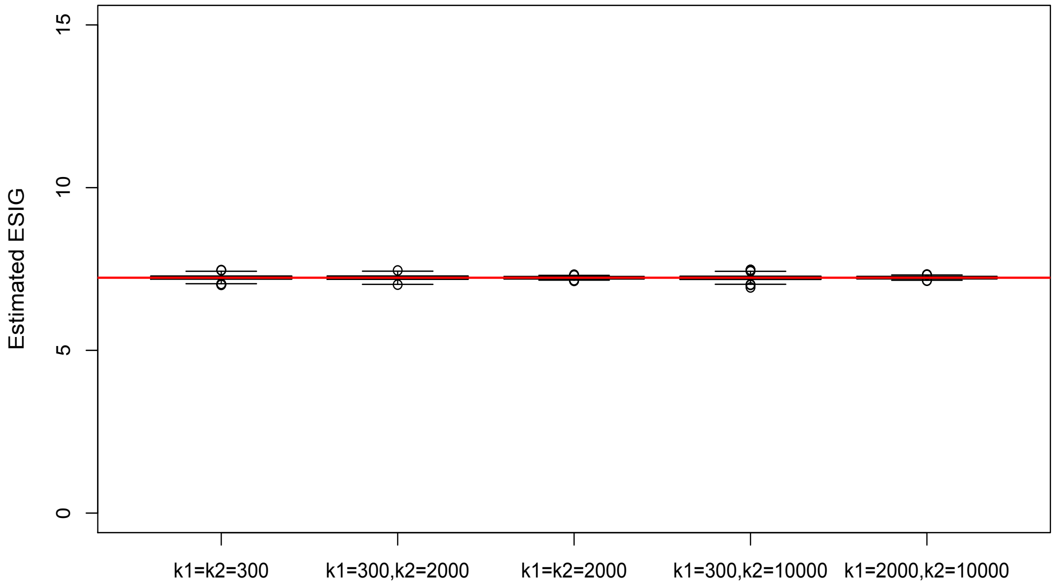


Figure 5.3: Estimated ESIG for the linear model (4.12) using LIS and different combinations of k_1 and k_2, and the true value of the Shannon information gain obtained using (4.16) (red line)

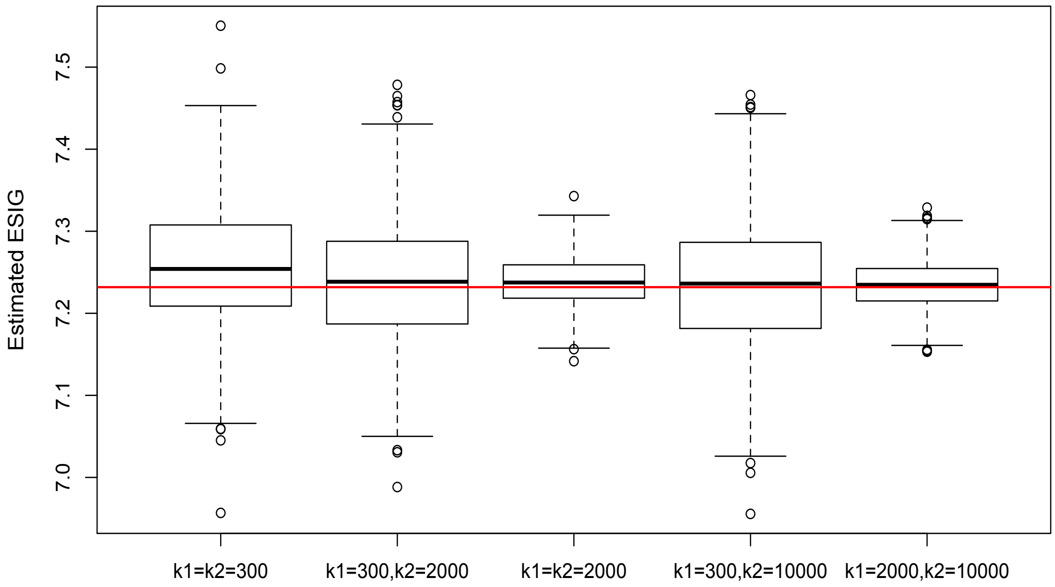


Figure 5.4: Estimated ESIG for the linear model (4.12) using ALIS and different combinations of k_1 and k_2, and the true value of the Shannon information gain obtained using (4.16) (red line) on a smaller y-axis scale

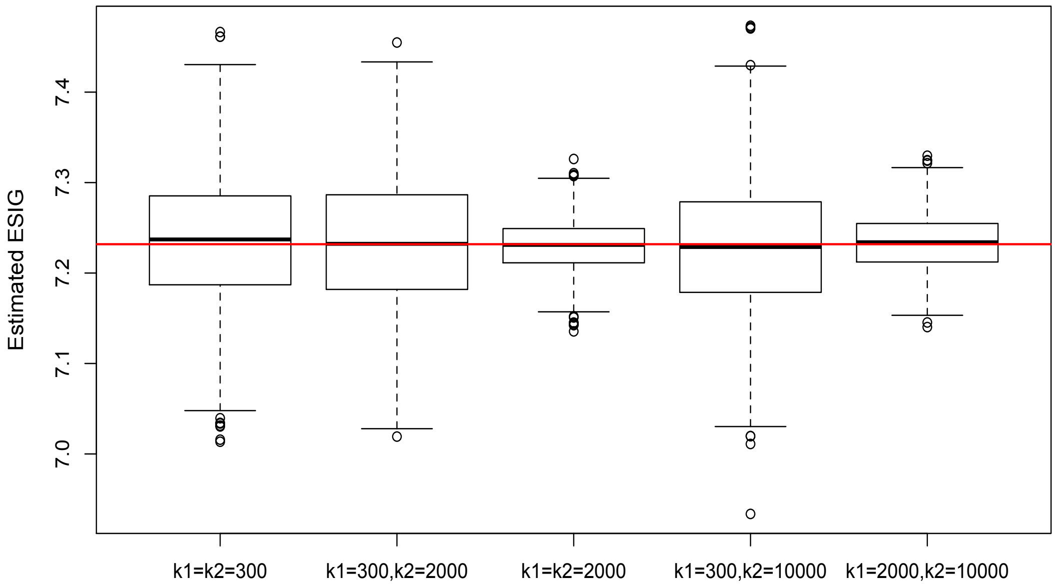


Figure 5.5: Estimated ESIG for the linear model (4.12) using LIS and different combinations of k_1 and k_2, and the true value of the Shannon information gain obtained using (4.16) (red line) on a smaller y-axis scale

scale. We notice that the bias has decreased compared to ALIS and the approximation of the expected utility estimates are centred closer to the true value for the different pairs of k_1 and k_2. The only thing that changes in this figure is the variance, which is controlled by the size of k_1.

In this example, where the posterior is actually a normal distribution, for LIS the approximate posterior will be exactly equal to the true posterior distribution, and so the approximation of the evidence will be exact. This will result in the method being exactly unbiased. However, in this example, using LIS (centering on the mode and using the negative Hessian of the log-unnormalised posterior density) rather than ALIS (centering on the true parameters that generated the data) has not resulted in a substantial improvement, supporting the use of ALIS with more complex examples.

5.1.2 Michaelis-Menten model

The first nonlinear example in this section is the Michaelis-Menten model,

$$y_i = \frac{\theta_1 x_i}{\theta_2 + x_i} + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and θ_1, θ_2, $\sigma^2 > 0$ are unknown parameters (see Section 3.1).

We assume a conjugate inverse-gamma prior distribution, $\sigma_\varepsilon^2 \sim \mathrm{IG}(a, b)$, where $a = 3$ and $b = 2$ are known hyperparameters. We also assume independent log-normal prior distributions, $\theta_1 \sim \log N(\mu_1, \sigma_1^2)$ and $\theta_2 \sim \log N(\mu_2, \sigma_2^2)$, where $\mu_1 = 4.38$, $\sigma_1 = 0.07$, $\mu_2 = 1.19$ and $\sigma_2 = 0.84$. These prior distributions result in $\mathbb{E}[\theta_1] = 80$ and $\mathbb{E}[\theta_2] = 5$ and imply that the 10% and 90% quantiles of noise-to-signal ratio (σ_ε divided by the maximum expected response $\eta(400, \boldsymbol{\theta})$) are 0.009 and 0.02 [1], respectively. We chose a more diffuse prior distribution for θ_2 as this is the parameter that has a greater influence on the shape of the response. See Appendix C.1.3 for examples of the shape of the expected response of the Michaelis-Menten model for different values of θ_1 and θ_2 sampled from these prior distributions.

We integrate out σ_ε^2 to obtain the marginal likelihood with respect to $\boldsymbol{\theta} = (\theta_1, \theta_2)^{\mathrm{T}}$, which is available in closed form:

$$\begin{aligned}\pi_M(\mathbf{y}|\boldsymbol{\theta}, \xi) &= \int_0^\infty \pi_l(\mathbf{y}|\boldsymbol{\theta}, \sigma_\varepsilon^2, \xi)\pi_b(\sigma_\varepsilon^2) d\sigma_\varepsilon^2 \\ &= \int_0^\infty (2\pi\sigma_\varepsilon^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^n \left[\left(y_i - \frac{\theta_1 x_i}{\theta_2 + x_i}\right)^2\right]\right\} \\ &\qquad\qquad \times (\sigma_\varepsilon^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma_\varepsilon^2}\right\} d\sigma_\varepsilon^2\end{aligned}$$

[1] We choose a small noise-to-signal ratio as this is the case which is most interesting in a computational point of view, where existing methods for approximating the evidence that use samples from the prior distribution fail to give a very good estimate of the evidence.

$$\propto \left[1 + \frac{\sum_{i=1}^{n}\left(y_i - \frac{\theta_1 x_i}{\theta_2 + x_i}\right)^2}{2b}\right]^{-(a+\frac{n}{2})}.$$

The log-unnormalised posterior density is then given by:

$$\begin{aligned}\log \pi_u(\boldsymbol{\theta}|\mathbf{y}, \xi) = &-\left(a + \frac{n}{2}\right)\log\left[2b + \sum_{i=1}^{n}\left(y_i - \frac{\theta_1 x_i}{\theta_2 + x_i}\right)^2\right] + \text{constant} \\ &- \log[\theta_1\sigma_1\sqrt{2\pi}] - \frac{(\log\theta_1 - \mu_1)^2}{2\sigma_1^2} \\ &- \log[\theta_2\sigma_2\sqrt{2\pi}] - \frac{(\log\theta_2 - \mu_2)^2}{2\sigma_2^2}. \end{aligned} \tag{5.1}$$

In order to ensure positive values for θ_1 and θ_2 we use the transformation $\boldsymbol{\theta}' = (\log\theta_1, \log\theta_2)^{\mathrm{T}}$ in LIS and ALIS (Section 4.3). Hence, we take a normal approximation to the posterior distribution of $\boldsymbol{\theta}'$ as described in Section 4.3.2. To calculate the negative Hessian of the log-unnormalised posterior density, $\mathbf{H}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}')$, we first have to find the derivatives of the log-unnormalised posterior density $\log\pi_u^{\theta'}(\boldsymbol{\theta}'|\mathbf{y}, \xi)$ with respect to $\boldsymbol{\theta}'$ using Equations (4.36). The derivatives can be found in Appendix B.2.1.

We compare approximations of the expected Shannon information gain for space-filling designs with $n = 5, 10, 20$ points, given in Figure 5.6. We use both normal and t importance distributions for the importance sampling methods. The ESIG using nMC, ALIS, LIS and nIS is estimated for two combinations of inner and outer sample sizes: (i) $k_1 = 2000, k_2 = 10000$, and (ii) $k_1 = k_2 = 300$. For LA1 and LA2 (single loop methods) we use: (i) $k_1 = 2000$, (ii) $k_1 = 300$, and for the 'reuse' method: (i) $k_1 = 2000$, (ii) $k_1 = 300$.

Figure 5.7 shows the distribution of 100 estimates of the ESIG obtained using the different methods, and different combinations of inner and outer sample sizes, for the $n = 5$ space-filling design, ξ_5. We treat as the 'true' ESIG the nMC approximation with $k_1 = k_2 = 1,000,000$ (red line), as these sample sizes should lead to negligible bias. We notice that ALIS and LIS have small bias and variance even for small k_1 and k_2 compared to all other methods for the same sample sizes. Increasing k_1 and k_2 reduces the variance and bias of nMC (see Section 4.1.2), and nMC,2000 has ESIG similar to the importance sampling based methods (ALIS, LIS, nIS). Similarly, increasing k_1 and k_2 reduces the variance and bias of the 'reuse' method. Increasing k_1 and k_2 also makes a big improvement to nIS, because small k_1 and k_2 leads to small effective sample size and hence in most iterations of nIS, samples from the prior are used rather than samples from the approximate posterior distribution (see Section 4.2.3). For ALIS and LIS, increasing (k_1, k_2) from $(300, 300)$ to $(2000, 10000)$ has little effect on the mean of the distribution, perhaps because the bias is already small even for $(k_1, k_2) = (300, 300)$. However, the variance is reduced. In this particular example, changing from a normal importance distribution to a t importance distribution sightly improves ALIS but makes

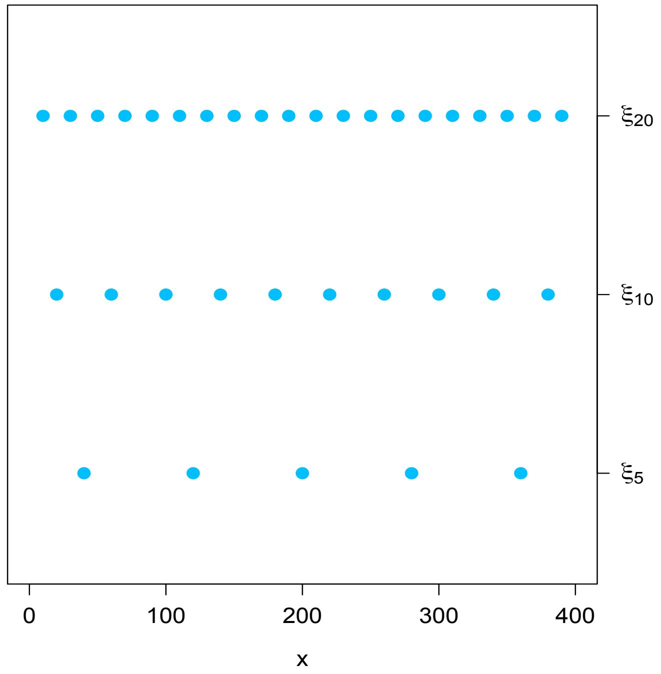


Figure 5.6: The space-filling designs, $\xi_5, \xi_{10}, \xi_{20}$ used with the Michaelis-Menten example with $n = 5, 10, 20$, respectively

little difference for LIS. LA1 and LA2 have less bias than nMC, nIS, and 'reuse' with $k_1 = k_2 = 300$, but for larger k_1 and k_2, use of LA1 and LA2 result in more bias than these three methods.

Figure 5.8 shows the results for applying the methods to the 5-run space filling design, ξ_5, in terms of relative root mean squared error (rRMSE) with respect to a nMC approximation with $k_1 = k_2 = 1,000,000$. The figure plots rRMSE against computational log-time. The nMC,1000000 approximation is treated as the 'true' ESIG because it should lead to negligible bias and variance. Four clusters can be distinguished in Figure 5.8: a cluster of nMC and nIS for small k_1 and k_2; a cluster of LA1, LA2 and 'reuse' for small k_1 and k_2; a cluster of ALIS and LIS for small k_1 and k_2; and a cluster of all methods for large k_1 and k_2. The least computationally expensive methods are LA1,300 and LA2,300 but these methods result in higher rRMSE compared with most other methods. As expected, nMC,300, nIS,300 and nIS,t,300 give the highest rRMSE (nIS for small k_1 and k_2 leads to small effective sample size and hence the prior distribution is often being used as the importance distribution rather than an approximate posterior distribution, see Section 4.2.3). ALIS and LIS with $k_1 = k_2 = 300$ and both normal and t importance distributions have lower rRMSE than other methods with similar computational expense. Increasing k_1 and k_2 for all methods has decreased the rRMSE but increased the computational expense.

Next we present results for the designs with more runs ($n = 10,\ 20$) which also support the results and insights from above.

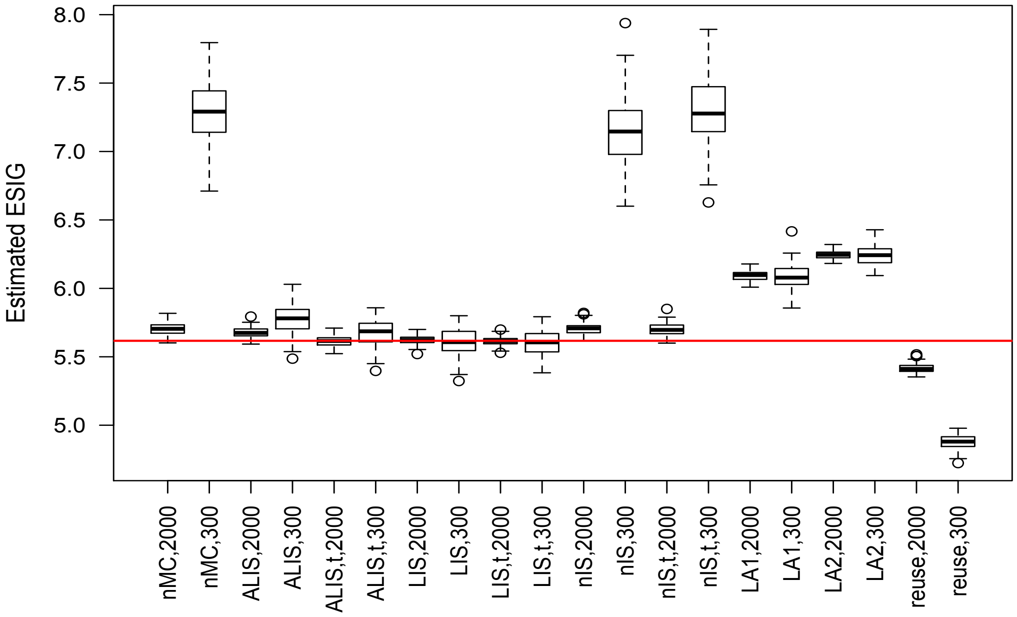


Figure 5.7: Estimated expected Shannon information gain for the parameters $\boldsymbol{\theta}$ of the Michaelis-Menten model for all methods (see Table 5.1, Section 4.2.2 and (4.11)) for the $n = 5$ space-filling design, ξ_5, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$ (the notation nMC,2000 denotes estimation of the ESIG using naïve Monte Carlo with $k_1 = 2000$ and $k_2 = 10000$, nMC,300 is the ESIG evaluated 100 times using nMC with $k_1 = k_2 = 300$, etc)

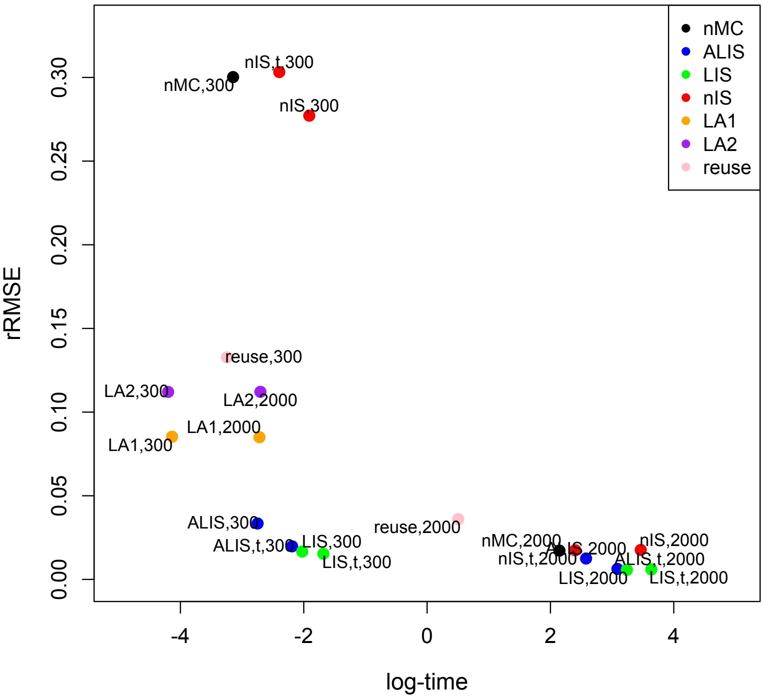


Figure 5.8: The rRMSE against log-time for the $n = 5$ space-filling design, ξ_5, for the Michaelis-Menten example

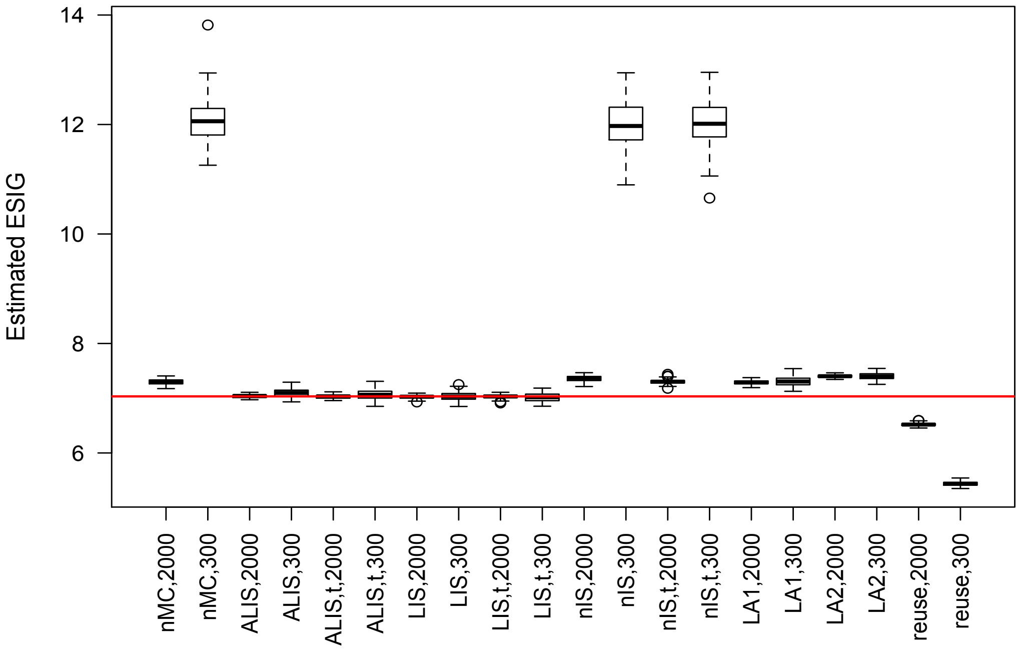


Figure 5.9: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the Michaelis-Menten model for all methods for the $n = 10$ space-filling design, ξ_{10}, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$

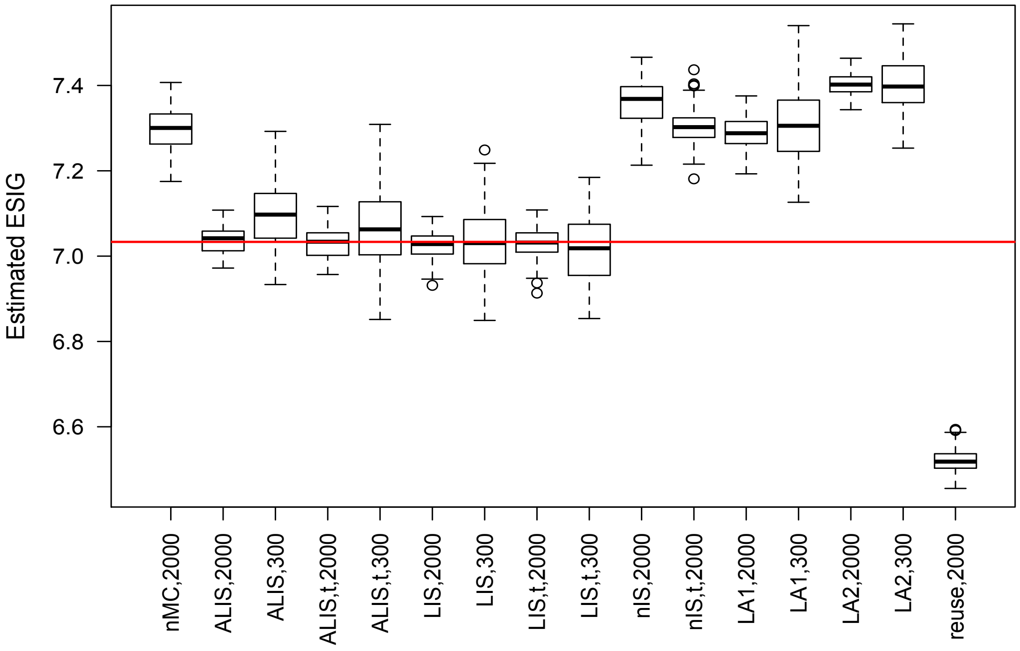


Figure 5.10: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the Michaelis-Menten model for all methods for the $n = 10$ space-filling design, ξ_{10}, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$ (nMC,300, nIS,300, nIS,t,300 and reuse,300 are omitted because these methods exhibit large bias)

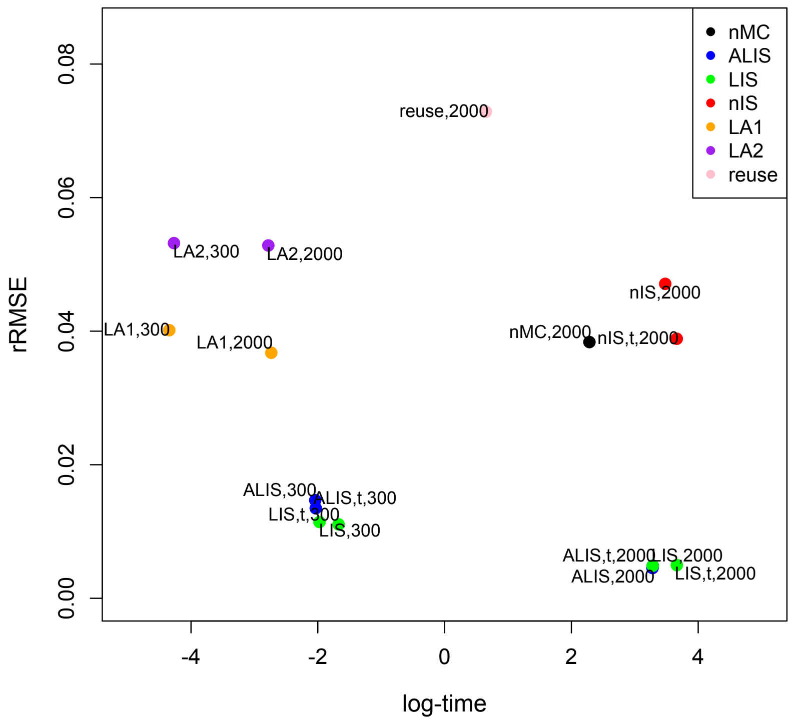


Figure 5.11: The rRMSE against log-time for the $n = 10$ space-filling design, ξ_{10}, for the Michaelis-Menten example

Figures 5.9 and 5.10 show 100 estimates of the ESIG for each method for the 10-run space-filling design, ξ_{10}. In Figure 5.10, the results for nMC,300, nIS,300, nIS,t,300 and reuse,300 are omitted because these methods exhibit large bias. The results are similar to the results of the $n = 5$ space-filling design presented in Figure 5.7, but with higher information gains in general. For this 10-run design nMC and nIS result in greater bias than the 5-run design possibly due to the greater difference between the prior and the posterior distributions.

In Figure 5.11 we assess the $n = 10$ space-filling design, ξ_{10}, in terms of rRMSE against log-time. The same procedure was followed as before. The least computationally expensive methods are LA1,300 and LA2,300 however these have higher rRMSE than other methods. We omit nMC,300, nIS,300, nIS,t,300 and reuse,300 because these methods exhibit large bias. Compared to the Laplace approximation methods, ALIS,300 and LIS,300 with both normal and t importance distributions give much reduced rRMSE for a moderate increase in computational cost. ALIS,300 and LIS,300 are also substantially cheaper and more accurate than nMC,2000, nIS,2000 and reuse,2000. Only ALIS,2000 and LIS,2000 for both normal and t importance distributions are more accurate, but this comes at the price of a significant increase in computational expense which does not seem worthwhile for this example.

Figure 5.12 shows 100 estimates of the ESIG for each method for the $n = 20$ space-

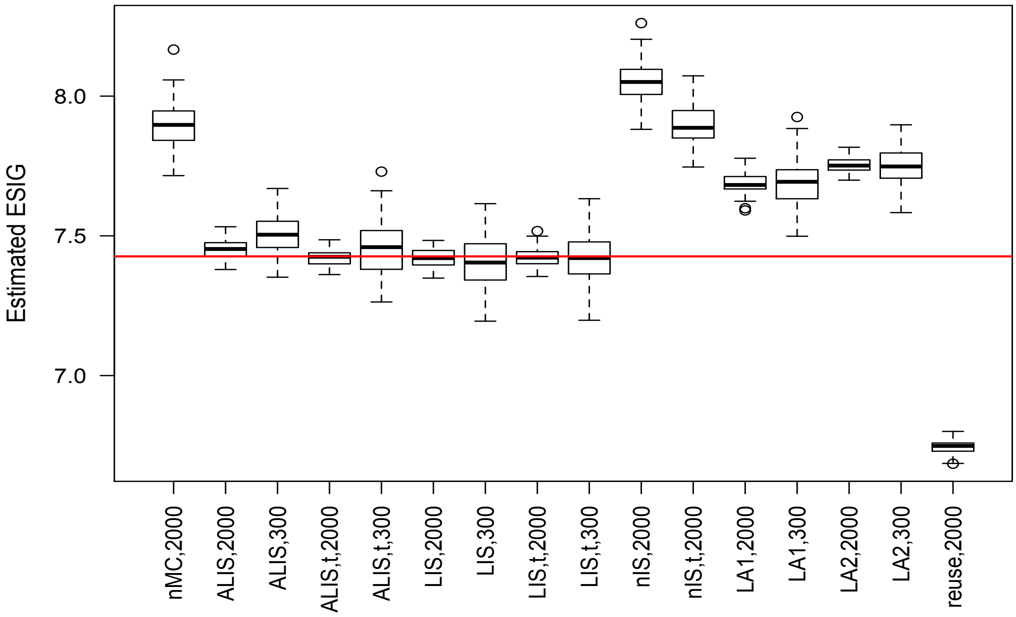


Figure 5.12: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the Michaelis-Menten model for all methods for the $n = 20$ space-filling design, ξ_{20}, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$ (nMC,300, nIS,300, nIS,t,300 and reuse,300 are omitted because these methods exhibit large bias)

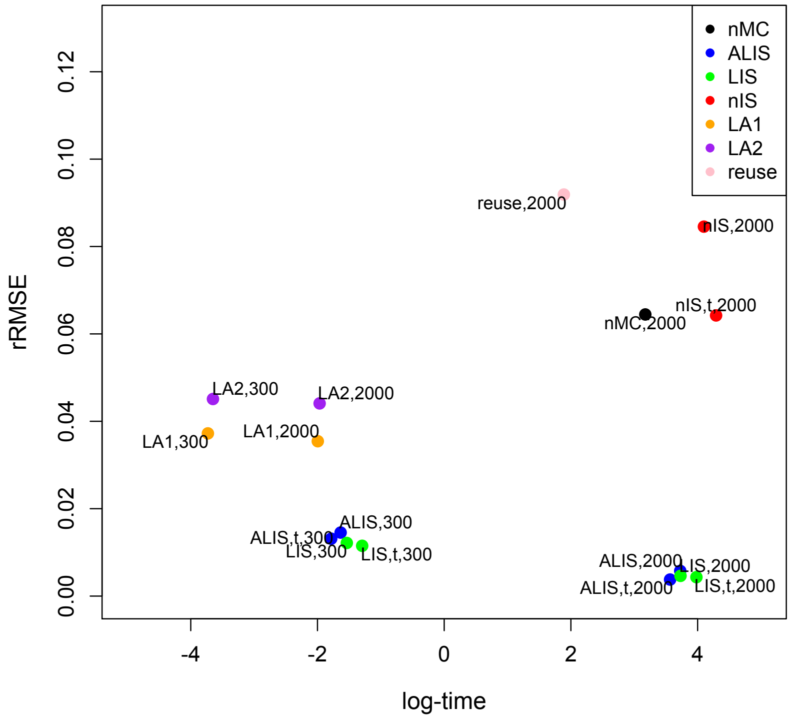


Figure 5.13: The rRMSE against log-time for the $n = 20$ space-filling design, ξ_{20}, for the Michaelis-Menten example

filling design, ξ_{20}. Similar results are obtained as in Figures 5.7 and 5.9 but with the information gain again being higher. We can notice that increasing the number of runs, n, of the design has increased the bias in the 'reuse' method. Also, increasing n has decreased the bias in LA1 and LA2, as we would expect, as the Laplace approximation relies on asymptotic results and hence as n increases the bias decreases.

In Figure 5.13 we assess the 20-run space-filling design, ξ_{20}, in terms of rRMSE against log-time. Similar results apply here as for the previous designs. LIS,300 and ALIS,300 give a highly accurate approximation at relatively low computational cost.

Figures 5.7-5.13 show that ALIS and LIS with small k_1 and k_2 have given results with less bias than other methods for low computational cost (small k_1 and k_2).

5.1.3 Biochemical Oxygen Demand (BOD) model

A data set on biochemical oxygen demand (BOD) was analysed by Bates and Watts (1988, Chapter 2) assuming the following model:

$$y_i = \theta_1(1 - \exp\{-\theta_2 x_i\}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{5.2}$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, y_i is BOD (mg/L) and x_i is time (in days).

We assume independent log-normal prior distributions, $\theta_1 \sim \log N(\mu_1, \sigma_1^2)$ and $\theta_2 \sim \log N(\mu_2, \sigma_2^2)$, with $\mu_1 = 3.38$, $\sigma_1 = 0.20$, $\mu_2 = 1.098$, $\sigma_2 = 1.12$. The prior means were chosen to match the means as given by DiCiccio et al. (1997) ($\mathbb{E}(\theta_1) = 30$ and $\mathbb{E}(\theta_2) = 3$). The prior variances were chosen to show differences between the methods (for smaller and bigger variances the results were similar for all methods). We also assume a non-informative prior distribution on σ_ε with $\pi_b(\sigma_\varepsilon) \propto \sigma_\varepsilon^{-1}$. See Appendix C.1.4 for examples of the shape of the expected response of the BOD model for different values of θ_1 and θ_2 sampled from these prior distributions.

The likelihood is given by:

$$\pi_l(\mathbf{y}|\boldsymbol{\theta}, \sigma_\varepsilon^2, \xi) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(\mathbf{y} - \boldsymbol{\eta})^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\eta})\right\},$$

where $\boldsymbol{\eta} = [\eta(x_1, \boldsymbol{\theta}) \ldots \eta(x_n, \boldsymbol{\theta})]^{\mathrm{T}}$ and $\eta(x_i, \boldsymbol{\theta}) = \theta_1(1 - \exp\{-\theta_2 x_i\})$.

We integrate out σ_ε^2 to obtain the marginal likelihood:

$$\begin{aligned}
\pi_M(\mathbf{y}|\boldsymbol{\theta}, \xi) &= \int_0^\infty \pi_l(\mathbf{y}|\boldsymbol{\theta}, \sigma_\varepsilon^2, \xi)\pi_b(\sigma_\varepsilon)d\sigma_\varepsilon \\
&\propto \int_0^\infty \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(\mathbf{y} - \boldsymbol{\eta})^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\eta})\right\} \frac{1}{\sigma_\varepsilon}\frac{1}{2\sigma_\varepsilon}d\sigma_\varepsilon^2 \\
&= \int_0^\infty \frac{1}{2}\frac{1}{(2\pi)^{n/2}}\frac{1}{(\sigma_\varepsilon^2)^{n/2+1}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(\mathbf{y} - \boldsymbol{\eta})^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\eta})\right\} d\sigma_\varepsilon^2 \\
&\propto \left[(\mathbf{y} - \boldsymbol{\eta})^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\eta})\right]^{-n/2}.
\end{aligned}$$

The above integral is evaluated by comparison with an inverse-gamma density and is finite provided that the residuals of the sum of squares, $(\mathbf{y}-\boldsymbol{\eta})^{\mathrm{T}}(\mathbf{y}-\boldsymbol{\eta})$, is non zero which is true with probability one.

The log-unnormalised posterior density for $\boldsymbol{\theta}=(\theta_1,\theta_2)^{\mathrm{T}}$ is then given by:

$$\begin{aligned}\log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi) = &-\frac{n}{2}\log\left[(\mathbf{y}-\boldsymbol{\eta})^{\mathrm{T}}(\mathbf{y}-\boldsymbol{\eta})\right] + \text{constant} \\ &- \log[\theta_1\sigma_1\sqrt{2\pi}] - \frac{(\log\theta_1-\mu_1)^2}{2\sigma_1^2} - \log[\theta_2\sigma_2\sqrt{2\pi}] - \frac{(\log\theta_2-\mu_2)^2}{2\sigma_2^2}.\end{aligned} \tag{5.3}$$

Similarly to the previous example (Section 5.1.2), we aim to construct an importance distribution that guarantees positive values of all parameters $\boldsymbol{\theta}$. Hence, we take a normal approximation to the distribution of $\boldsymbol{\theta}'=(\log\theta_1,\log\theta_2)^{\mathrm{T}}$ as described in Section 4.3.2. In order to calculate the negative Hessian of the log-unnormalised posterior, $\mathbf{H}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}')$, in ALIS and LIS (Section 4.3), we first have to find the derivatives of the log-unnormalised posterior density $\log\pi_u^{\theta'}(\boldsymbol{\theta}'|\mathbf{y},\xi)$ with respect to $\boldsymbol{\theta}'$ using Equations (4.36). These derivatives can be found in Appendix B.2.2.

We compare approximations of the ESIG for the design given in Bates and Watts (1988, Appendix 1) ($n=6$) and for space-filling designs with $n=10,20$ as shown in Figure 5.14, for all methods similar to the previous example. We use both normal and t importance distributions for the importance sampling methods.

As in Section 5.1.2, Figure 5.15 shows the distribution of 100 estimates of the ESIG obtained using the different methods, and different combinations of inner and outer sample sizes. For each choice of method and sample size, 100 Monte Carlo estimates were calculated for the $n=6$ design, ξ_6. Again, we treat as the ‘true’ ESIG the nMC approximation with $k_1=k_2=1{,}000{,}000$ (red line) because the very large Monte Carlo sample size should lead to negligible bias and variance. We notice that ALIS and LIS for even small k_1 and k_2 have less bias and variance compared to the other methods for the same sample sizes. Increasing k_1 and k_2 reduces the variance and bias of nMC and also makes a big improvement to nIS and ‘reuse’. For ALIS and LIS, increasing (k_1,k_2) from $(300,300)$ to $(2000,10000)$ has little effect on the mean of the distribution, perhaps because the bias is already small even for $(k_1,k_2)=(300,300)$. However, the variance is reduced. Changing from a normal to a t importance distribution slightly improves ALIS, LIS and nIS. LA1 and LA2 have less bias than nMC,300, nIS,300, nIS,t,300 and reuse,300, but the bias is larger than for the other methods for larger k_1 and k_2. In this example we can see greater differences in the ESIG between LA1 and LA2 and the other methods; for this nonlinear example, the Laplace approximation methods overestimate the information gain for this particular design.

Figure 5.16 shows the results for the BOD example from applying all the methods to the $n=6$ run design, ξ_6, in terms of relative root mean squared error (rRMSE) with

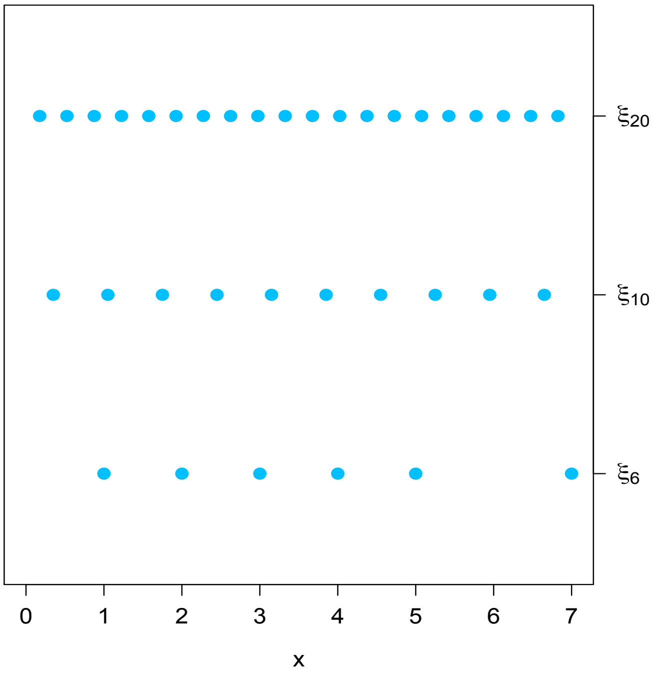


Figure 5.14: BOD example: The $n = 6$ design, ξ_6, from Bates and Watts (1988, Appendix 1), and $n = 10, 20$ space-filling designs, ξ_{10} and ξ_{20}, respectively

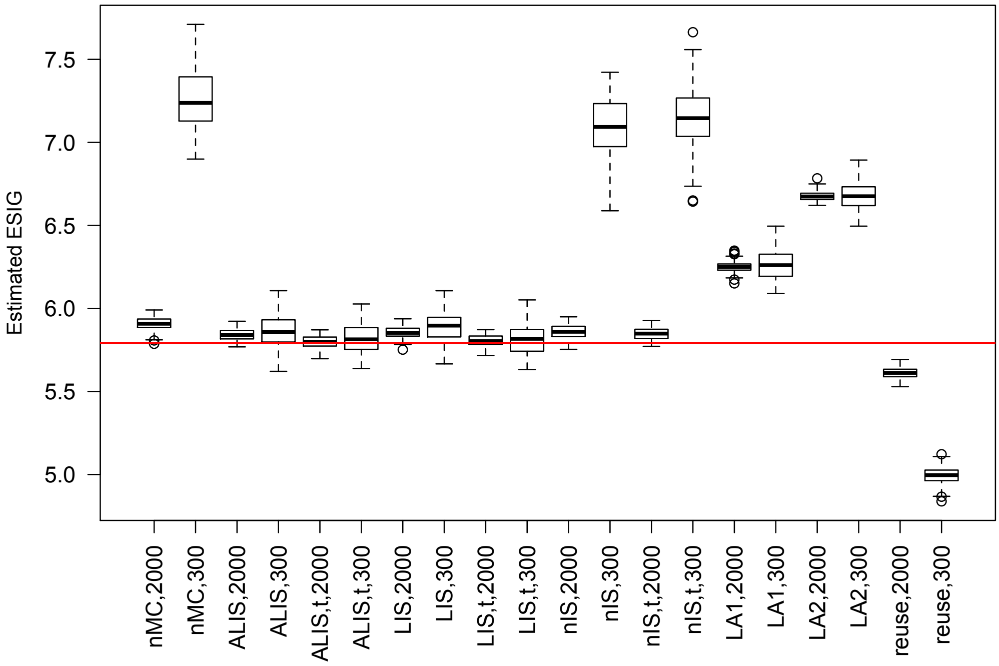


Figure 5.15: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the BOD model for all methods (see Table 5.1, Section 4.2.2 and (4.11)) for the $n = 6$ design, ξ_6, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$ (for notation see Figure 5.7)

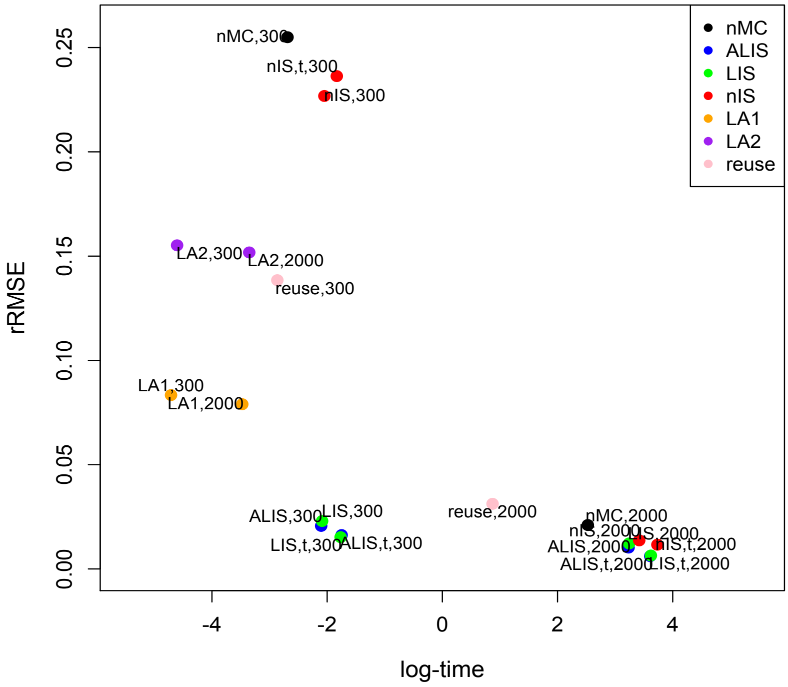


Figure 5.16: The rRMSE against log-time for the $n = 6$ design, ξ_6, for the BOD example

respect to a nMC approximation with $k_1 = k_2 = 1,000,000$ against computational log-time, similar to the example in Section 5.1.2. As before, the methods are separated into four clusters. The least computationally expensive methods are LA1 and LA2 but these have higher rRMSE than other methods. Also, reuse,300 is computationally cheap but has higher rRMSE than other methods. ALIS,300 and LIS,300 for both normal and t importance distributions have small rRMSE and are computationally cheap. As expected nMC,300, nIS,300 and nIS,t,300 have the highest rRMSE. Increasing k_1 and k_2 for all methods decreases the rRMSE but increases the computational expense.

Figure 5.17 and 5.18 show the distribution of 100 estimates of the ESIG for the 10-run space-filling design, ξ_{10}, with the only difference that nMC,300, nIS,300, nIS,t,300 and reuse,300 are omitted from Figure 5.18 because these methods exhibit large bias. Similar comments apply here as in Figure 5.15. The difference between LA1 and LA2 and the other methods has decreased for this design as the Laplace approximation relies on asymptotic results and hence as n increases the bias decreases.

In Figure 5.19 we assess the 10-run space-filling design, ξ_{10}, in terms of rRMSE against log-time. Similar results apply as for the previous design and the previous example. We can notice here a bigger difference in rRMSE between LA1 and LA2 (LA1 results in lower rRMSE than LA2).

Figure 5.20 shows the distribution of 100 estimates of the ESIG for each method for the 20-run space-filling design, ξ_{20}, following the same procedure as for the previous example and the previous designs. We can notice similar results as in Figure 5.15 and

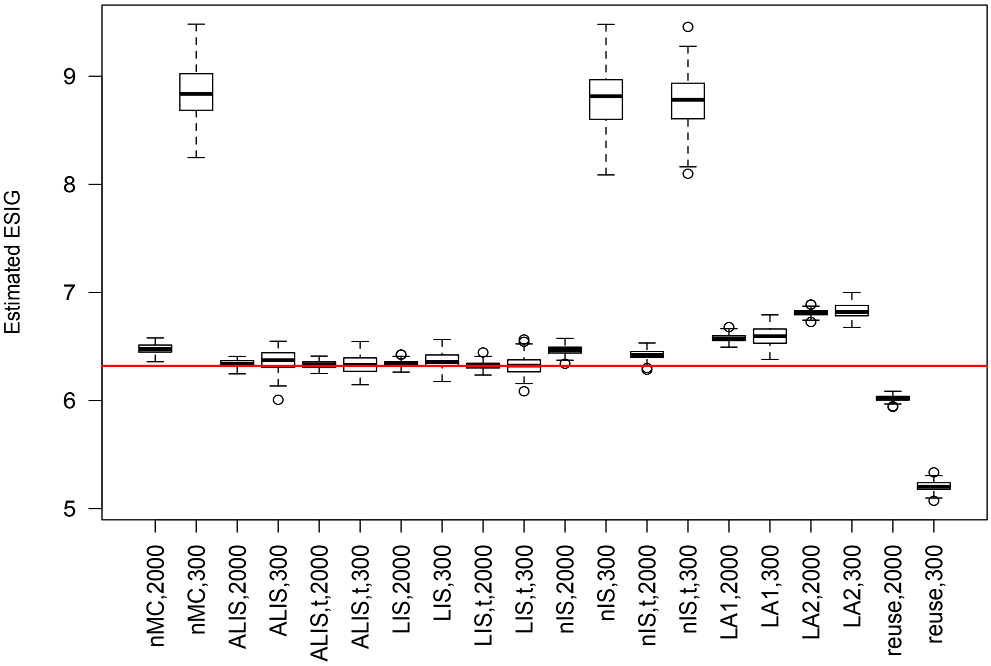


Figure 5.17: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the BOD model for all methods for the $n = 10$ space-filling design, ξ_{10}, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$

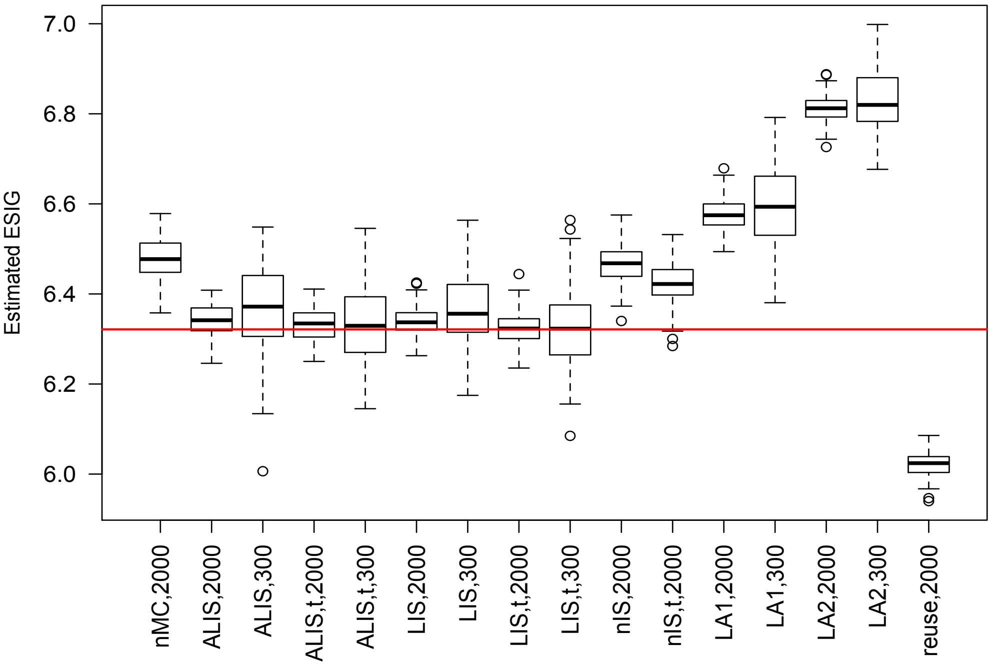


Figure 5.18: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the BOD model for all methods for the $n = 10$ space-filling design, ξ_{10}, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$ (nMC,300, nIS,300, nIS,t,300 and reuse,300 are omitted because these methods exhibit large bias)

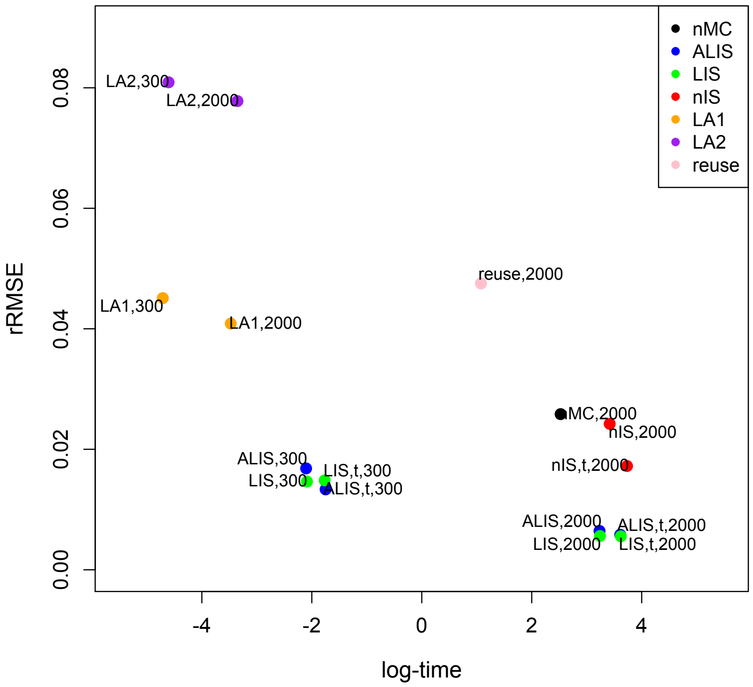


Figure 5.19: The rRMSE against log-time for the $n = 10$ space-filling design, ξ_{10}, for the BOD example (nMC,300, nIS,300, nIS,t,300 and reuse,300 are omitted because these methods exhibit large bias)

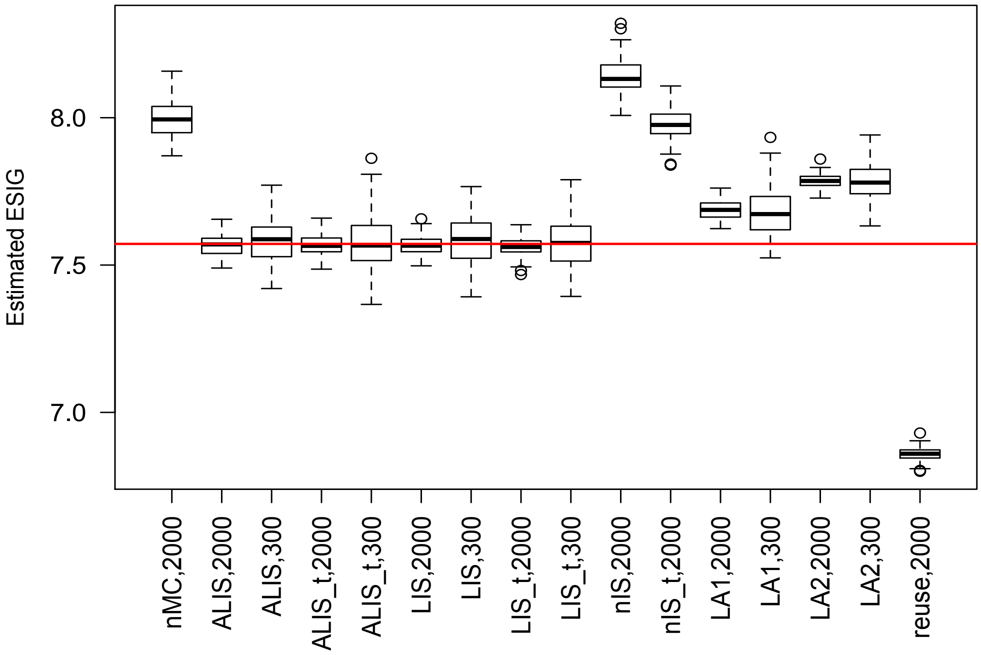


Figure 5.20: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the BOD model for all methods for the $n = 20$ space-filling design, ξ_{20}, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$ (nMC,300, nIS,300, nIS,t,300 and reuse,300 are omitted because these methods exhibit large bias)

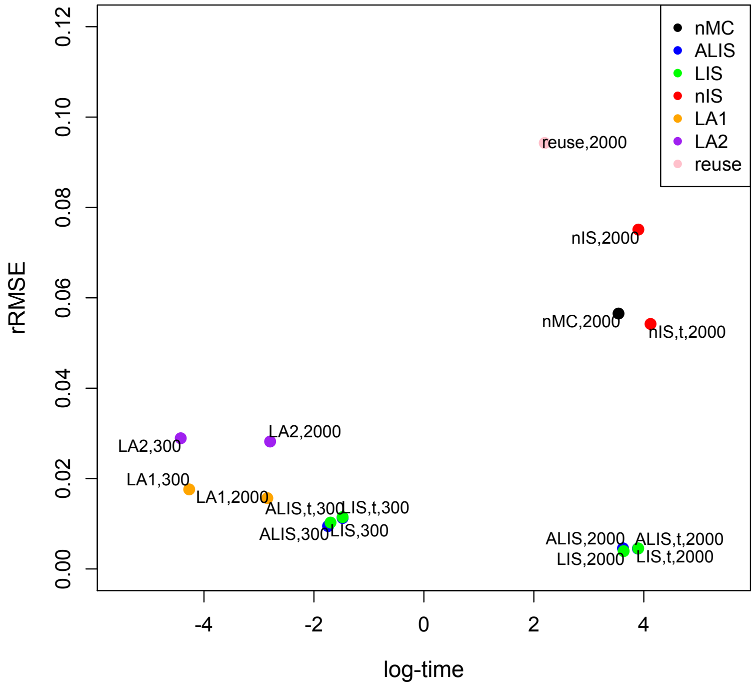


Figure 5.21: The rRMSE against log-time for the $n = 20$ space-filling design, ξ_{20}, for the BOD example (nMC,300, nIS,300, nIS,t,300 and reuse,300 are omitted because these methods exhibit large bias)

Figure 5.18. Increasing the number of runs of the design, n, has decreased the bias in LA1 and LA2, which is now less than the bias from nMC,2000, nIS,2000 and nIS,t,2000.

In Figure 5.21 we assess the 20-run space-filling design, ξ_{20}, in terms of rRMSE against log-time for all the methods. The same comments apply as for the previous designs (ξ_6 and ξ_{10}) and the previous example. LIS,300 and ALIS,300 give a highly accurate approximation at relatively low computational cost.

As in the previous example, Figures 5.15-5.21 show that ALIS and LIS with small k_1 and k_2 have given results with smaller bias than other methods with low computational cost (small k_1 and k_2). The bias of the 'reuse' method remained large compared to the other methods as we increased n. Also, for this example we notice that the difference between LA1 and LA2 and the other methods decreases as n increases; the Laplace approximation relies on asymptotic results and hence as n increases the bias decreases. In addition, for this example the difference between the two Laplace approximations is also higher (LA1 results in less bias than LA2) possibly due to the additional assumption required in LA2 (see Section 4.2.2).

5.1.4 Lubricant model

The last nonlinear example in this chapter is a 10-dimensional model. Following Bates and Watts (1988, Chapter 3), the kinematic viscosity of a lubricant is given as a function of temperature (°C), x_1, and pressure (atm), x_2. The model is given by:

$$y_i = \frac{\theta_1}{\theta_2 + x_{1i}} + \theta_3 x_{2i} + \theta_4 x_{2i}^2 + \theta_5 x_{2i}^3 + (\theta_6 + \theta_7 x_{2i}^2) x_{2i} \exp\left\{ -\frac{x_{1i}}{\theta_8 + \theta_9 x_{2i}^2} \right\} + \varepsilon_i, \quad (5.4)$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and we define $\theta_{10} = \log \sigma_\varepsilon$. We assume independent normal prior distributions, $\theta_j \sim N(\mu_j, \sigma_j^2)$, $j = 1, \ldots, 10$, with mean and standard deviation equal to the maximum likelihood estimates and their standard errors from data available from Bates and Watts (1988, Chapter 3) (see also DiCiccio et al., 1997), which can be found in Table 5.2.

The likelihood is

$$\pi_l(\mathbf{y}|\boldsymbol{\theta}, \xi) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} (\mathbf{y} - \boldsymbol{\eta})^{\mathrm{T}} (\mathbf{y} - \boldsymbol{\eta}) \right\},$$

where $\boldsymbol{\eta} = [\eta(x_{11}, x_{21}, \boldsymbol{\theta}) \ldots \eta(x_{1n}, x_{2n}, \boldsymbol{\theta})]^{\mathrm{T}}$ with

$$\eta(x_{1i}, x_{2i}, \boldsymbol{\theta}) = \frac{\theta_1}{\theta_2 + x_{1i}} + \theta_3 x_{2i} + \theta_4 x_{2i}^2 + \theta_5 x_{2i}^3 + (\theta_6 + \theta_7 x_{2i}^2) x_{2i} \exp\left\{ -\frac{x_{1i}}{\theta_8 + \theta_9 x_{2i}^2} \right\},$$

and the log-unnormalised posterior density is:

$$\begin{aligned} \log \pi_u(\boldsymbol{\theta}|\mathbf{y}, \xi) = &-\frac{n}{2} \log[2\pi\sigma_\varepsilon^2] - \frac{1}{2\sigma_\varepsilon^2} \left[(\mathbf{y} - \boldsymbol{\eta})^{\mathrm{T}} (\mathbf{y} - \boldsymbol{\eta}) \right] \\ &- \sum_{j=1}^{10} \left\{ \frac{1}{2} \log[2\pi\sigma_j^2] + \frac{1}{2\sigma_j^2} (\theta_j - \mu_j)^2 \right\}. \quad (5.5) \end{aligned}$$

The derivatives of this density can be found in Appendix B.2.3. The different methods listed in Table 5.1 are used to approximate the evidence, and hence the expected Shannon information gain. In addition we also employ LA2.

We compare approximations of the expected Shannon information gain for the design given in Bates and Watts (1988, Appendix 1) with $n = 53$ and for a subset of the original design with $n = 20$, chosen at random with stratification to include five design points at each level of temperature. These designs are shown in Figure 5.22. The ESIG using nMC and nIS is estimated for $k_1 = 2000$, $k_2 = 10000$; using ALIS and LIS for $k_1 = k_2 = 300$; and using LA1 and LA2 (single loop methods) $k_1 = 300$. We choose these combinations of inner and outer sample sizes based on the results from previous examples. For this example the ‘reuse’ method is omitted because as shown in the previous examples it performs poorly.

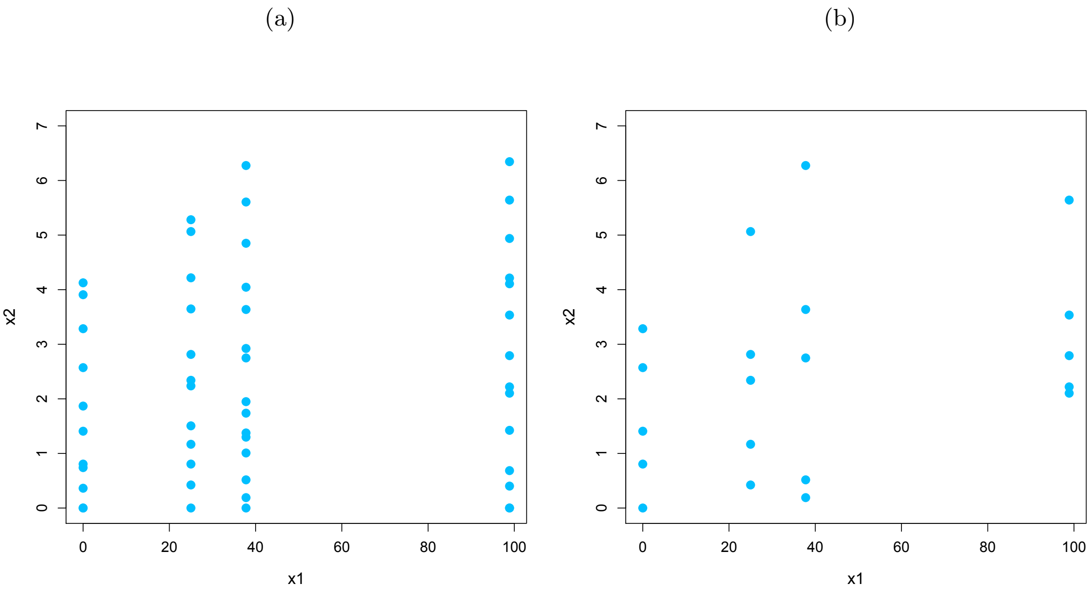


Figure 5.22: (a) The $n = 53$ design, ξ_{53}, as given by Bates and Watts (1988, Appendix 1) for the lubricant model; (b) A sub-design, ξ_{20}, with $n = 20$ chosen from the 53-run design given by Bates and Watts (1988, Appendix 1) for the lubricant model

Parameter	**Mean**	**St. dev.**
θ_1	1054.54	24.63
θ_2	206.55	5.29
θ_3	1.46	0.04
θ_4	-0.26	0.01
θ_5	0.02	0.002
θ_6	0.40	0.03
θ_7	0.04	0.001
θ_8	57.40	2.37
θ_9	-0.48	0.075
$\theta_{10} = \log \sigma_\varepsilon$	-1.50	0.10

Table 5.2: The prior means and standard deviations of the unknown parameters θ_j, $j = 1, \ldots, 10$ of the lubricant model (5.4)

Figure 5.23 shows boxplots of the Monte Carlo distribution of estimates of the ESIG obtained using the different methods for the $n = 53$ design, ξ_{53}. For each choice of method 100 Monte Carlo estimates were calculated. The nMC method has resulted in ESIG much higher than all the other methods. We also notice that the estimated ESIG using LA2 is sometimes also very large with the distribution for this approximation having a long right tail.

In order to reduce the estimated ESIG of nMC for this example we increase k_2. Figure 5.24 shows boxplots of the Monte Carlo distribution of 100 estimates of the ESIG obtained using nMC for two different combinations of k_1 and k_2: (i) $k_1 = 2000$, $k_2 =$

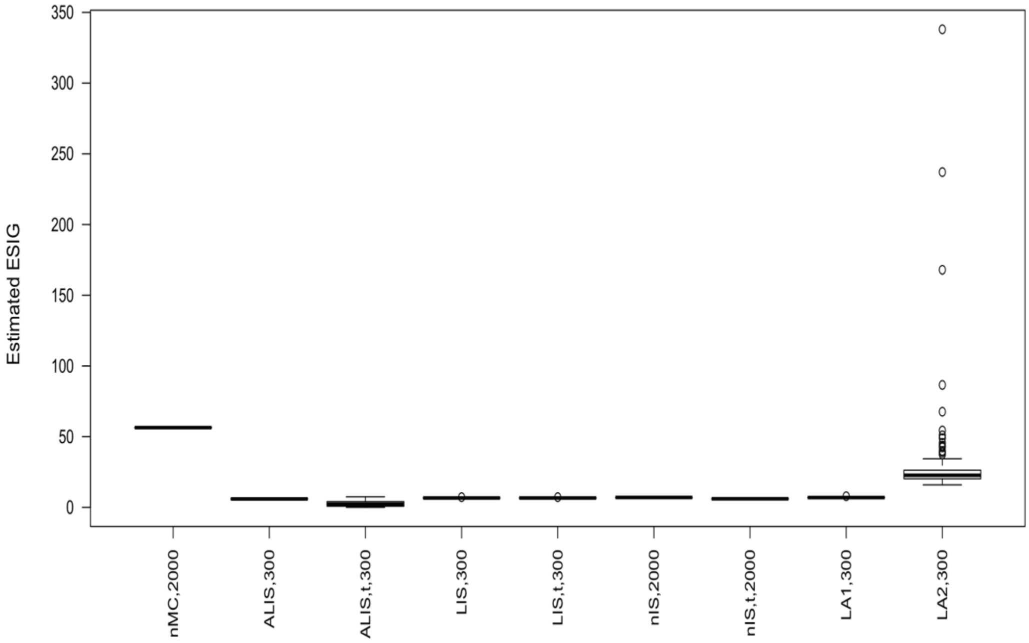


Figure 5.23: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the lubricant model for all methods (see Table 5.1 and Section 4.2.2) for the $n = 53$ design, ξ_{53} (for notation see Figure 5.7)

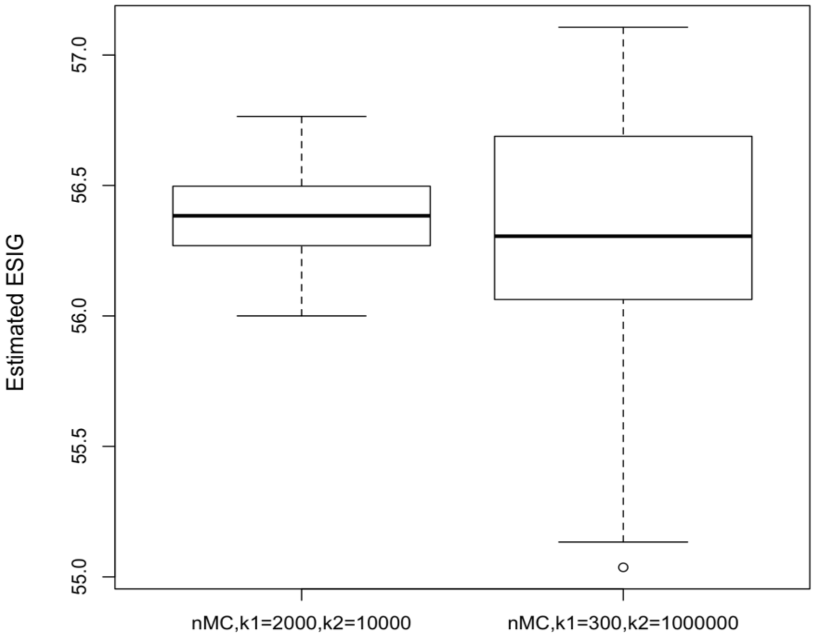


Figure 5.24: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the lubricant model found using nMC for different combinations of k_1 and k_2, for the $n = 53$ design, ξ_{53}

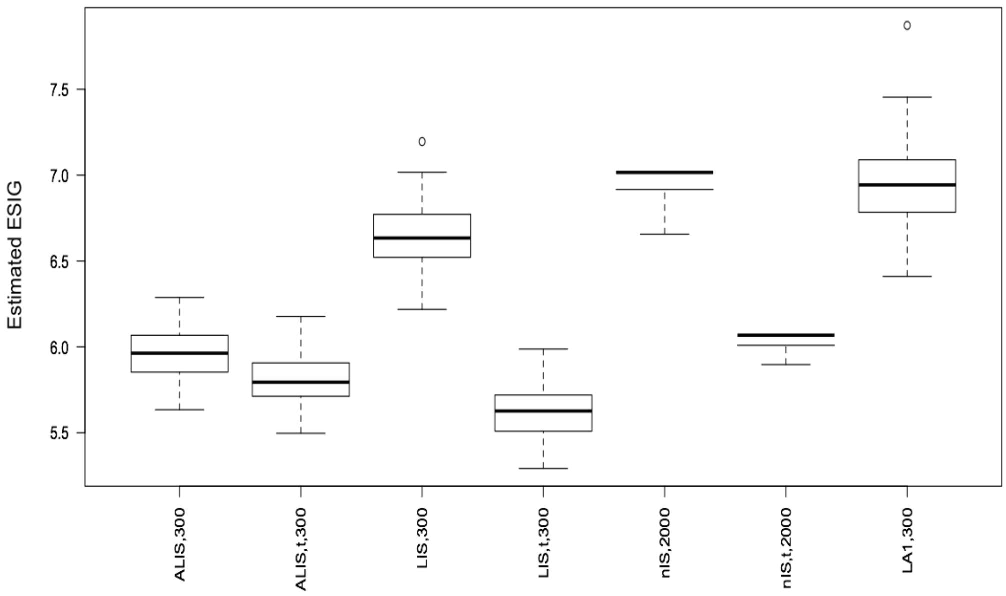


Figure 5.25: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the lubricant model for all methods for the $n = 53$ design, ξ_{53} (nMC and LA2 are omitted)

10000; (ii) $k_1 = 300$, $k_2 = 1000000$. Increasing k_2 does not change the estimated ESIG. For this 10-parameter nonlinear example, it appears that nMC is not a good method for approximating the expected utility as it seems to overestimate the information gain even for very large values of k_2.

In order to display differences between the methods, in Figure 5.25 nMC and LA2 are omitted. ALIS and LIS give lower estimates of the ESIG than other methods. For all the importance sampling methods (ALIS, LIS and nIS), changing from a normal to a t importance distribution has reduced the estimated ESIG. The nIS method has sometimes resulted in infinite ESIG estimates because the evidence, $\pi_e(\mathbf{y}|\xi)$, were approximated as zero and hence the approximate expected utility is

$$\begin{aligned}
\tilde{U}(\xi) &= \frac{1}{k_1}\sum_{h=1}^{k_1}\left[\log \pi_l(\mathbf{y}_h|\boldsymbol{\theta}_h, \xi) - \log \tilde{\pi}_e^h\right] \\
&= \frac{1}{k_1}\sum_{h=1}^{k_1}[\log \pi_l(\mathbf{y}_h|\boldsymbol{\theta}_h, \xi) - \log(0)] \\
&= \frac{1}{k_1}\sum_{h=1}^{k_1}[\log \pi_l(\mathbf{y}_h|\boldsymbol{\theta}_h, \xi) - (-\infty)] \\
&= \infty,
\end{aligned}$$

(see also Section 4.1.3). This zero evidence phenomenon can occur when most of the sampled $\boldsymbol{\theta}$ in the inner loop are a long way from the region of high likelihood or high posterior density, as can happen with nMC or importance sampling with a poorly chosen importance density.

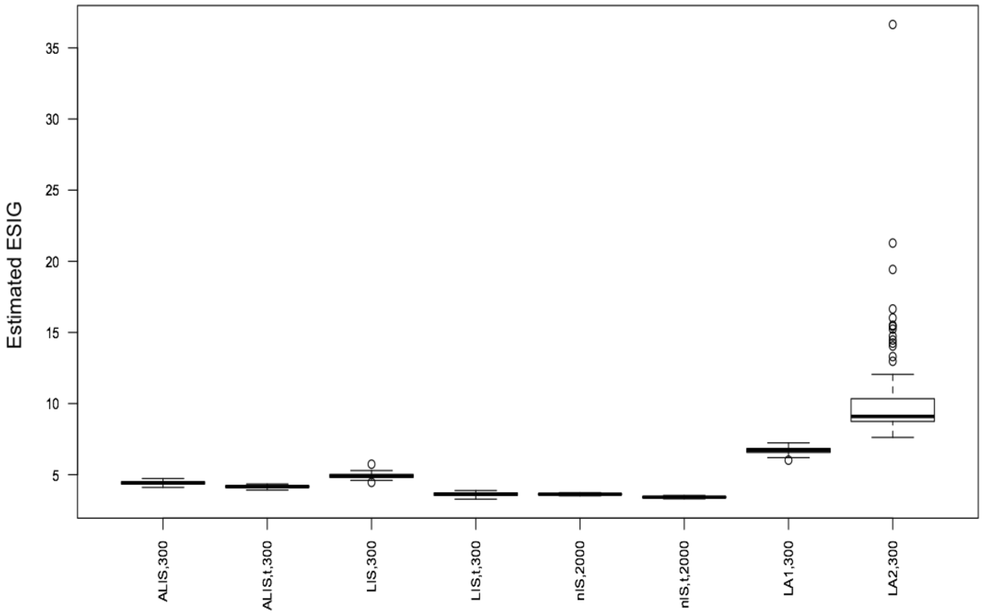


Figure 5.26: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the lubricant model for all methods for the $n = 20$ design, ξ_{20} (nMC is omitted)

Figure 5.26 shows the same results as Figure 5.25 but for the $n = 20$ design. The nMC method is omitted as it exhibits large ESIG. Similar comments apply here as for the previous design.

In order to see any differences between the methods, LA2 is also omitted in Figure 5.27 because of the long right tail. Similar to the 53-run design LIS,300 and ALIS,300 result in approximations with lower ESIG. For this design, nIS,2000 and nIS,t,2000 also result in approximations with lower ESIG, similar to that of ALIS and LIS. The better performance of nIS for the 20-run design may be due to the prior distribution being a better approximation of the posterior distribution in this case. In contrast, for the 53-run design, there may be a greater difference between the prior and posterior distributions. This may lead to a small effective sample size in the estimation of the posterior mean and covariance, and so the importance distribution will usually revert to the prior distribution (see Section 4.2.3), giving performance comparable to nMC.

For this example a plot to compare the relative root mean squared error (rMSE) against computational time is not included because we cannot find a ‘true’ value of the ESIG for comparison.

For this model we have seen that for the importance sampling methods (ALIS, LIS and nIS) changing from a normal to a t distribution results in lower estimates of the ESIG. This might be a consequence of the posterior distribution having fatter tails than the prior distribution and hence a t distribution is more appropriate. Also, the estimated ESIG with LA1 and LA2 for this model is different than the other methods. Possibly the asymptotic normal approximation to the posterior density underpinning

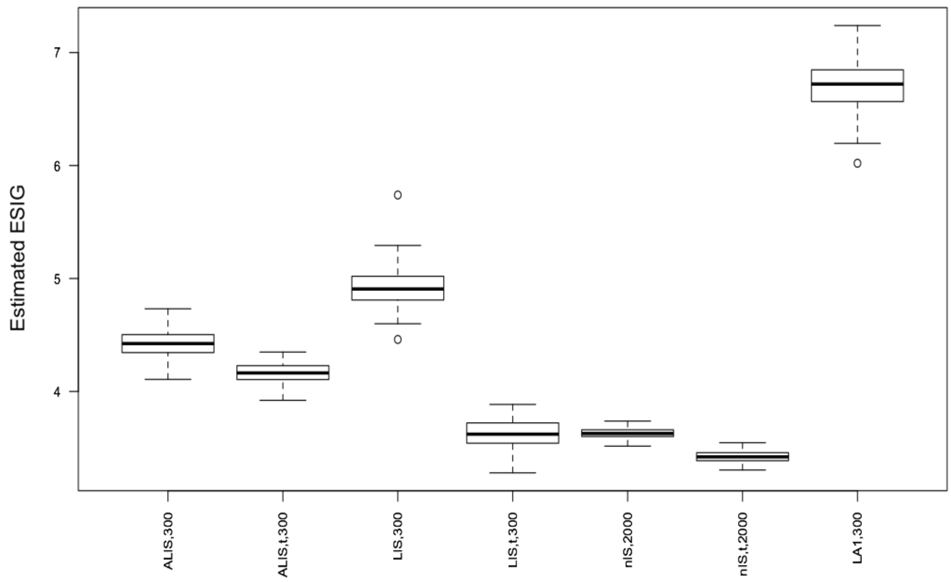


Figure 5.27: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the lubricant model for all methods for the $n = 20$ design, ξ_{20} (LA2 is omitted)

the Laplace approximation is inaccurate due to the large number of parameters relative to the sample size.

5.2 Optimisation of the utility function

A fundamental problem is how to optimise the objective function, and hence find optimal designs, in a computationally efficient manner. The most common approach used to search numerically for an optimal exact design is to use an exchange algorithm. This fall into two main classes: point exchange algorithms and coordinate exchange algorithms. Point exchange algorithms (Fedorov, 1972; Johnson and Nachtsheim, 1983) involve systematically exchanging design points with points from a candidate set in order to improve the value of the objective function. These approaches may be computationally expensive for problems with many continuous factors due to the need to use a very large candidate set. Coordinate exchange algorithms (Meyer and Nachtsheim, 1995) instead change one element, or 'coordinate', of the design at a time, without the need for a candidate set. A 'coordinate' is the value taken by an individual variable in a single run. These algorithms apply when the objective function can be evaluated exactly and deterministically, as is usually the case with frequentist design. However, in Bayesian design typically a Monte Carlo approximation to the objective function, i.e. the expected utility, is used, and so the standard exchange algorithms are not applicable.

For low-dimensional Bayesian design problems (one variable and a small number of design points), Müller and Parmigiani (1996) and more recently Weaver et al. (2016), performed stochastic optimisation for expected utility maximisation. This is performed by conducting a noisy computer experiment to construct a statistical emulator for the Monte Carlo approximation $\tilde{U}(\xi)$ for a small number of designs and smooth the resulting values of the approximation to the utility. However application of this idea to high-dimensional design problems suffers from the curse of dimensionality.

A more recent development that is also compatible with noisy evaluations of the objective function is the Approximate Coordinate Exchange (ACE) algorithm (Overstall and Woods, 2017). For each coordinate, a noisy Monte Carlo estimate of the expected utility is made for a small number of potential changes to the coordinate. These noisy evaluations of the expected utility are then smoothed using a Gaussian process emulator. The fitted emulator is smooth, and so can be optimised directly unlike the noisy evaluations themselves. The advantage of embedding the emulation step within a coordinate exchange algorithm is that it is only necessary to emulate one-dimensional functions, thereby eliminating the computational expense that occurs when using Gaussian processes in high dimensions (Rasmussen and Williams, 2006), e.g. through the need to use a large space-filling design.

The ACE algorithm

The algorithm is divided into two phases. Phase I of the algorithm is application of a coordinate exchange algorithm by constructing a sequence of one-dimensional emulators; see Algorithm 10. This phase tends to produce designs with clusters of similar design points. Phase II checks if the points in each cluster can be reduced by using a point exchange algorithm, using the optimal design from Phase I as a candidate list.

Algorithm 10: The ACE algorithm (Overstall and Woods, 2017)

```
Start with a randomly chosen initial design $\xi$;
repeat
    for $i = 1, \dots, n$ do
        for $j = 1, \dots, q_1$ do
            Generate a $1d$ space-filling design $\mathcal{d}_j = [x_j^1, \dots, x_j^Q] \in \mathcal{X}_j^Q$;
            for $k = 1, \dots, Q$ do
                Evaluate $\tilde{U}_k = \tilde{U}(\xi_{ij}(x_j^k))$;
            Construct a $1d$ Gaussian process emulator $\hat{U}(x)$ from data $\{x_j^k, \tilde{U}_k\}$;
            Set $x_{ij} = \arg\max_{x \in \mathcal{X}_j} \hat{U}(x)$ with probability $\tilde{p}$ obtained from Algorithm 11;
until convergence;
```

Above we use the notation $\xi_{ij}(x) = [\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{w}_{ij}(x), \mathbf{x}_{i+1}, \dots, \mathbf{x}_n]$, where $\mathbf{w}_{ij}(x) = (x_{i1}, \dots, x_{ij-1}, x, x_{ij+1} \dots, x_{iq_1})^{\mathrm{T}}$.

In Algorithm 10 the notation $\xi_{ij}(x) = [\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{w}_{ij}(x), \mathbf{x}_{i+1}, \dots, \mathbf{x}_n]$, where $\mathbf{w}_{ij}(x) = (x_{i1}, \dots, x_{ij-1}, x, x_{ij+1} \dots, x_{iq_1})^{\mathrm{T}}$, defines a new design with a new proposed ijth coordinate x. In addition $\tilde{U}_k = \tilde{U}(\xi_{ij}(x_j^k))$, $k = 1, \ldots, Q$, is the evaluation of the expected utility for the new design $\xi_{ij}(x_j^k)$ with the new proposed coordinate x_j^k, with the proposals x_j^k coming from a one-dimensional space-filling design $\mathcal{D}_j$. The emulator is given by the posterior mean of a Gaussian process,

$$\hat{U}(x) = \hat{\mu}_{ij} + \hat{\sigma}_{ij}\mathbf{a}^{\mathrm{T}}(x, \mathcal{D}_j)A(\mathcal{D}_j)^{-1}\mathbf{z}_{ij},$$

with $\hat{\mu}_{ij} = \sum_{k=1}^{Q} \tilde{U}_k/Q$, $\hat{\sigma}_{ij}^2 = \sum_{k=1}^{Q}(\tilde{U}_k - \hat{\mu}_{ij})^2/(Q-1)$ and $\mathbf{z}_{ij}$ is a vector having kth entry $(\tilde{U}_k - \hat{\mu}_{ij})/\hat{\sigma}_{ij}$. Under the common assumption of a squared exponential correlation structure, the vector $\mathbf{a}$ is the Q vector of correlations between variable x and each coordinate x_j^k and A is the $Q \times Q$ correlation matrix between all the coordinates x_j^k (see Woods et al., 2017).

In Algorithm 10, the proposed change to the coordinate is accepted with probability resulting from an independent check on the difference in expected utilities between the current and proposed designs. See Algorithm 11, which essentially describes obtaining the posterior probability that the proposed design has larger expected utility than the current design, using a separate Monte Carlo sample from the joint distribution of $\boldsymbol{\psi}$ and $\mathbf{y}$.

Algorithm 11: Accept/reject step of the ACE Algorithm 10

```
Given the current design $\xi$ and the new proposed coordinate $x$;
Let $\xi_{ij}(x)$ be the design formed by replacing the $ij$th coordinate of $\xi$ with $x$;
for $s = 1, \ldots, B$ do
    Sample $\tilde{\boldsymbol{\psi}}_s$ from $\pi_b(\boldsymbol{\psi})$;
    Sample $\mathbf{y}_1 \sim \pi_l(\mathbf{y}|\tilde{\boldsymbol{\psi}}_s, \xi_{ij}(x))$ and $\mathbf{y}_2 \sim \pi_l(\mathbf{y}|\tilde{\boldsymbol{\psi}}_s, \xi)$;
    Set $U_{1s} = \tilde{u}(\xi_{ij}(x), \tilde{\boldsymbol{\psi}}_s, \mathbf{y}_1)$ and $U_{2s} = \tilde{u}(\xi, \tilde{\boldsymbol{\psi}}_s, \mathbf{y}_2)$;
Assume $U_1 \sim N(b_1 + b_2, a)$ and $U_2 \sim N(b_1, a)$;
Calculate the posterior probability, $\tilde{p}$, that $b_2 > 0$ using “data” $U_1$ and $U_2$;
```

In Algorithm 11 the notation $\tilde{u}(\xi, \boldsymbol{\psi}, \mathbf{y})$ defines the estimate of $u(\xi, \boldsymbol{\psi}, \mathbf{y})$ obtained by estimating the evidence using a specified approximation.

Convergence of the algorithm is assessed graphically from trace plots of the approximate expected utility against the iteration number. Also, to avoid local optima, the algorithm is run P times with each run starting from a different, randomly chosen, initial design, ξ. See Overstall and Woods (2017) for more details. The ACE algorithm is implemented in the R package `acebayes` (Overstall et al., 2017).

As we find optimal designs numerically, all our designs may only be near-optimal.

5.3 Bayesian optimal designs using the approximate coordinate exchange algorithm

In this section we combine the expected utility approximation methods presented in Table 5.1, Section 4.2.2 and (4.11) with ACE (see Section 5.2) to find Bayesian optimal designs for the Michaelis-Menten model (Section 5.1.2), the BOD model (Section 5.1.3) and the Lubricant model (Section 5.1.4).

When using ACE, for the emulator building step we used $k_1 = k_2 = 2000$ and $k_1 = 2000$ for the single loop methods, for the Michaelis-Menten model and the BOD model. For the lubricant model we used $k_1 = k_2 = 300$ for ALIS and LIS and $k_1 = 300$ for LA1 and LA2. For nMC and nIS a larger Monte Carlo sample size was needed because the evaluation of the expected utility fails for small sample sizes; we used $k_1 = 2000$, $k_2 = 10000$. In Algorithm 11 of ACE we set $B = k_1 = k_2 = 10000$. Post-hoc the ESIG is approximated for each design found using ALIS with $k_1 = k_2 = 300$ which as shown in the previous section is computationally efficient and results in less bias than the other methods.

Michaelis-Menten model

We employ the ACE algorithm to find expected Shannon information gain optimal designs with $n = 5, 10, 20$. For each design, 10 random starts of the ACE algorithm are used, each starting from a different random Latin hypercube design.

Figures 5.28, 5.30 and 5.32 show the optimal designs produced using the different methods for $n = 5, 10, 20$ runs, respectively. Figure 5.28 shows some small differences in the design points found for each method, with the main pattern being that all methods tend to position some design points at the start of the region, where the expected response is changing more quickly, and some at the end of the region, where the expected response is more stable. The designs produced using LA2 and the 'reuse' method have more differences with the designs found using all the other methods; the former results in a design where some points are kept in the middle and no points are placed at the end of the design region, and the latter, results in a design where some points are also kept in the middle of the design region. Similar patterns can also be noticed in Figures 5.30 and 5.32 for $n = 10, 20$, with the addition of points at the start of the design region and some points in the middle of the design region. LA1 produces an optimal 20-run design different from the other methods, where there are no points at the end of the design region (Figure 5.32).

Figures 5.29, 5.31 and 5.33 give the estimated ESIG for the optimal designs for the different methods, approximated using ALIS with $k_1 = k_2 = 300$. All the optimal designs have higher ESIG than the space-filling design, and all optimal designs have similar ESIG. For the $n = 20$ optimal designs, LA2 and the 'reuse' have produced

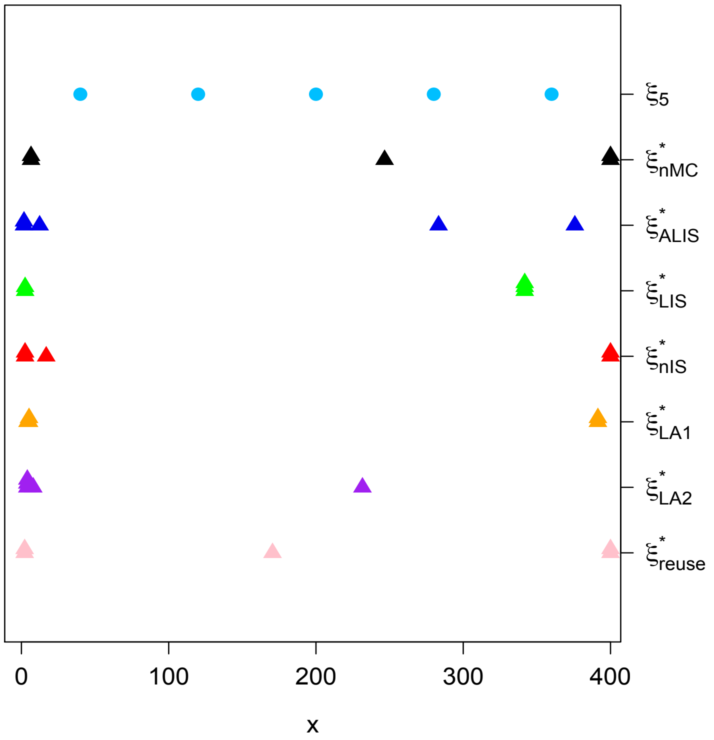


Figure 5.28: Expected Shannon information gain optimal designs with $n = 5$ for the Michaelis-Menten model and the 5-run space-filling design, ξ_5

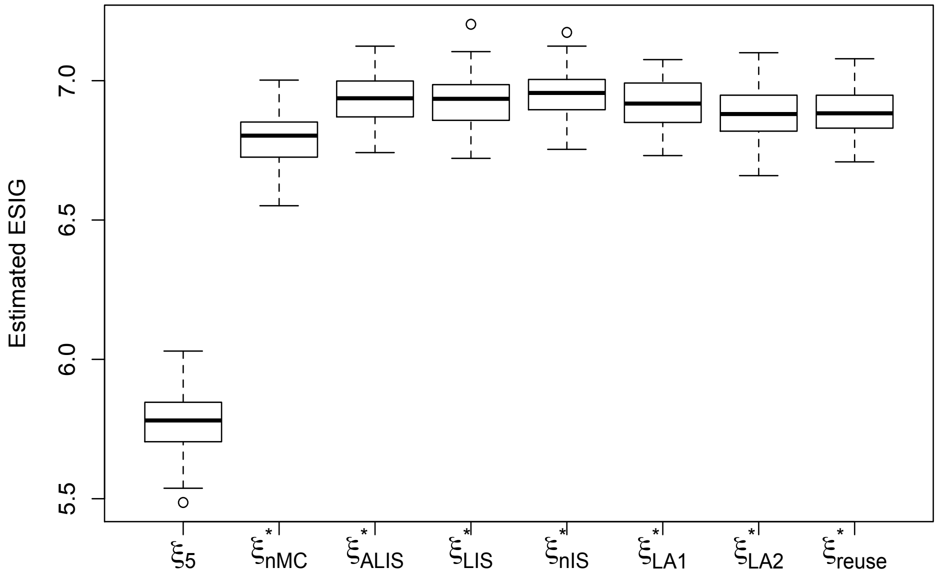


Figure 5.29: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the Michaelis-Menten model found using ALIS for the $n = 5$ optimal designs and the space-filling design, ξ_5

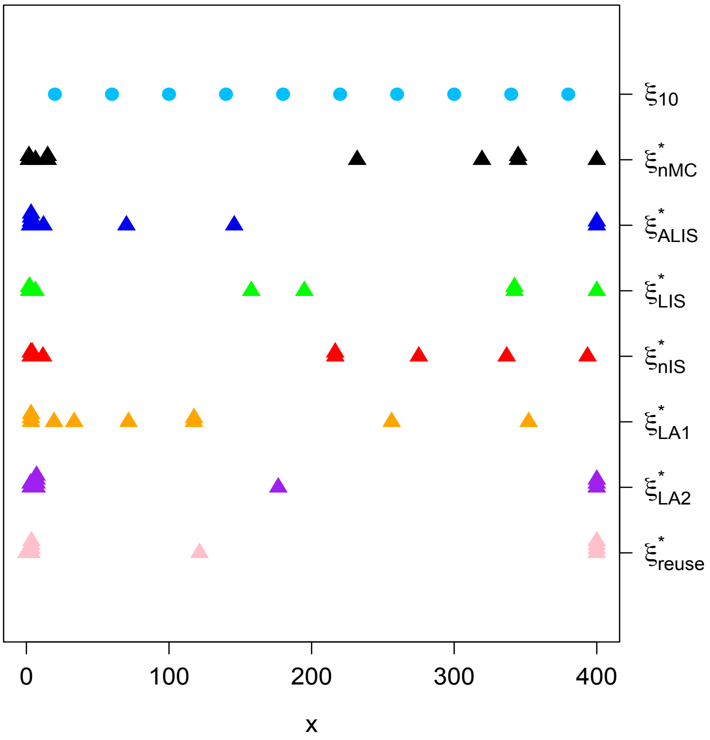


Figure 5.30: ESIG optimal designs with $n = 10$ for the Michaelis-Menten model and the 10-run space-filling design, ξ_{10}

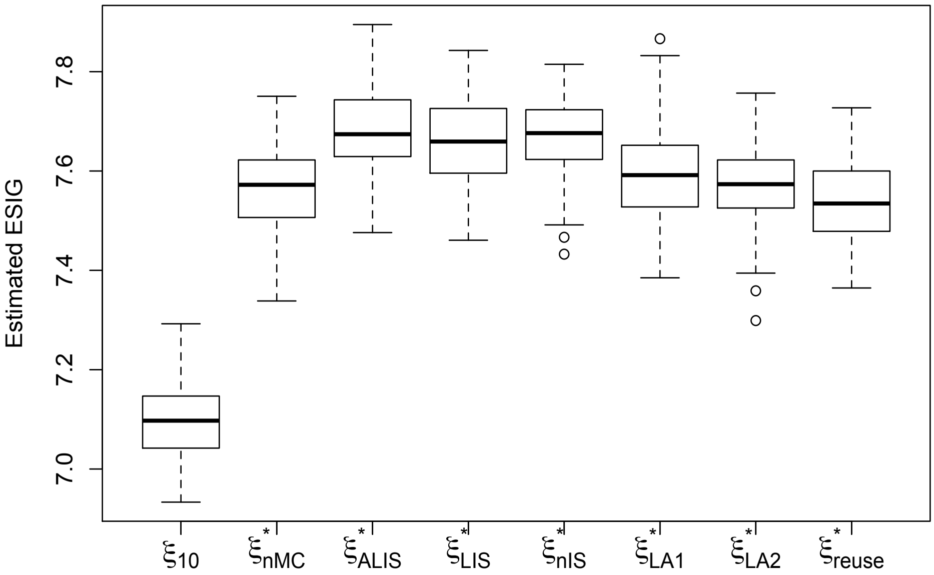


Figure 5.31: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the Michaelis-Menten model found using ALIS for the $n = 10$ optimal designs and the space-filling design, ξ_{10}

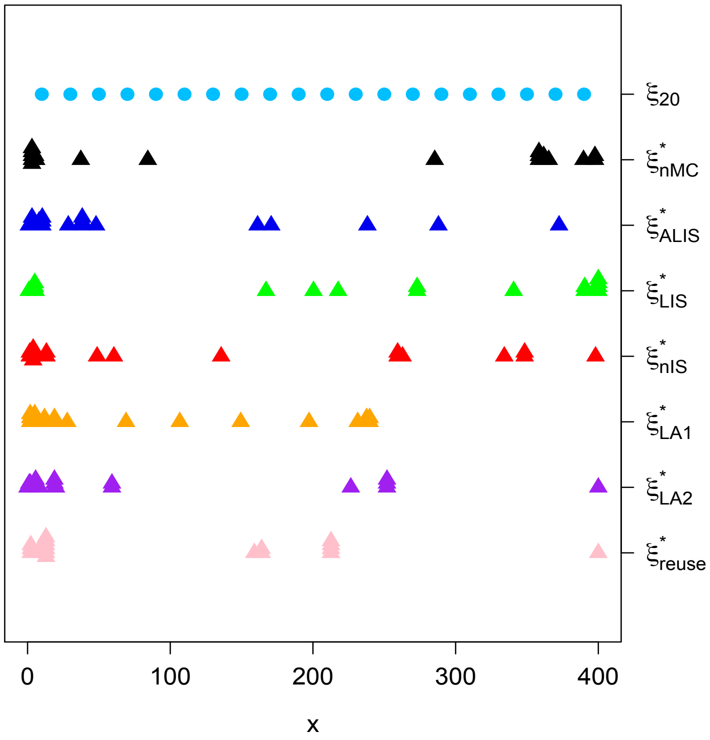


Figure 5.32: ESIG optimal designs with $n = 20$ for the Michaelis-Menten model and the 20-run space-filling design, ξ_{20}

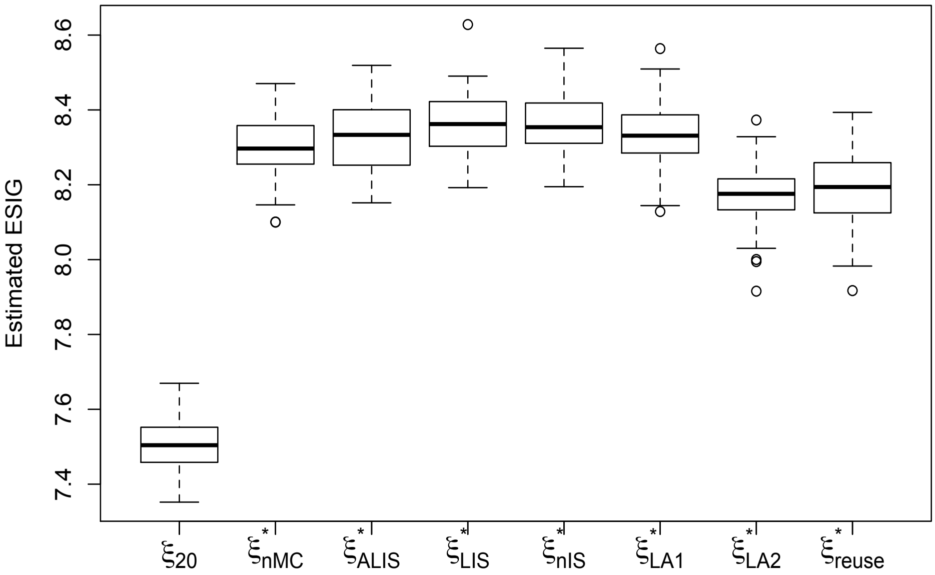


Figure 5.33: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the Michaelis-Menten model found using ALIS for the $n = 20$ optimal designs and the space-filling design, ξ_{20}

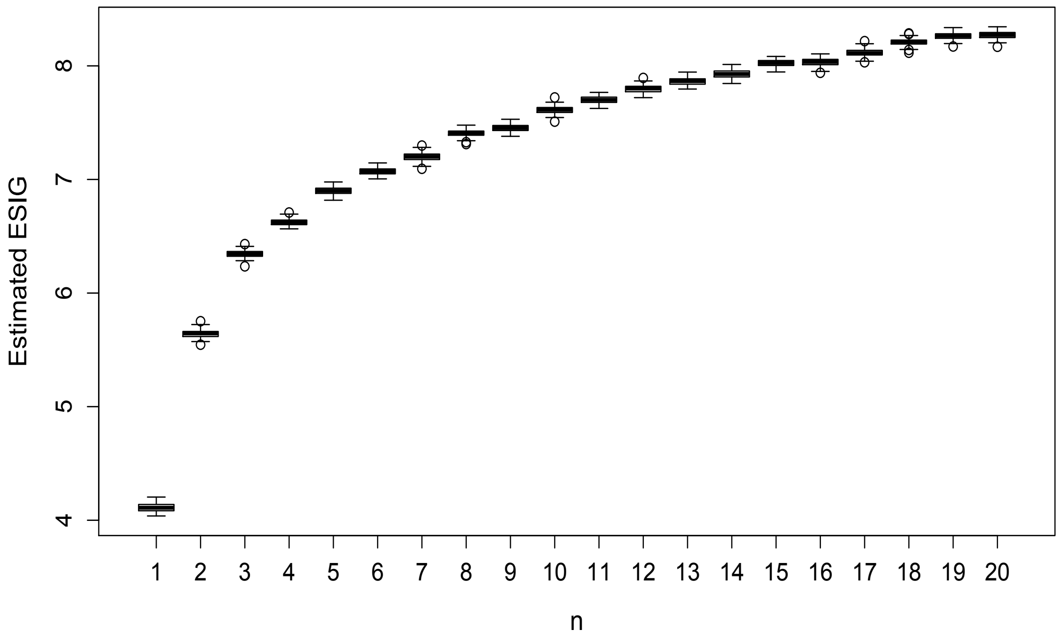


Figure 5.34: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the Michaelis-Menten model for optimal designs with $n = 1, \ldots, 20$, found using ACE and ALIS with $k_1 = k_2 = 2000$

designs where the ESIG is lower than from the designs produced from all the other methods; however the difference is small.

Note that as we increase the number of runs, n, of the design the ESIG does not change very quickly. To see that clearly, we find optimal designs using ALIS and $k_1 = k_2 = 300$ with $n = 1, 2, \ldots, 20$ runs, and then estimate the ESIG 100 times using ALIS and $k_1 = k_2 = 2000$ for each design. Figure 5.34 shows the increasing relationship between the runs of the design, n, and ESIG. The rate of increase decreases with n.

Biochemical Oxygen Demand (BOD) model

In this section we find optimal designs using the ACE algorithm for the BOD model. Again, 10 random starting designs are used in ACE, each starting from a different random LHS design. The same procedure was followed to find the optimal designs as for the Michaelis-Menten model. For this model LA2 is omitted as the optimisation failed to converge due to infinite objective function values.

Figures 5.35, 5.37 and 5.39 show the optimal designs produced using the different methods for $n = 6, 10, 20$ runs, respectively. Figure 5.35 shows some small differences in the design points found for each method, with the main pattern similar to that of the Michaelis-Menten model optimal designs; all methods tend to position some design points at the start of the region, where the expected response is changing more quickly,

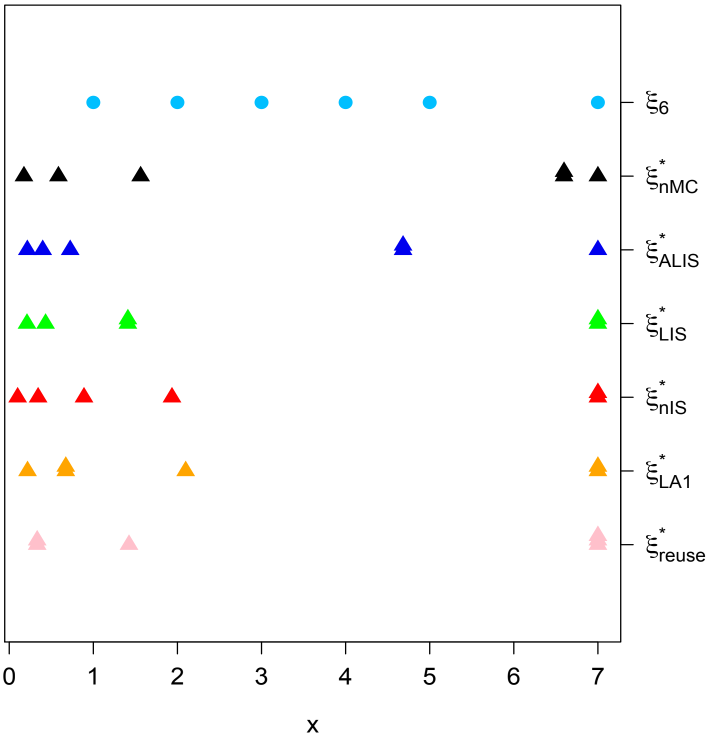


Figure 5.35: ESIG optimal designs with $n = 6$ for the BOD model and the 6-run design, ξ_6, given by Bates and Watts (1988)

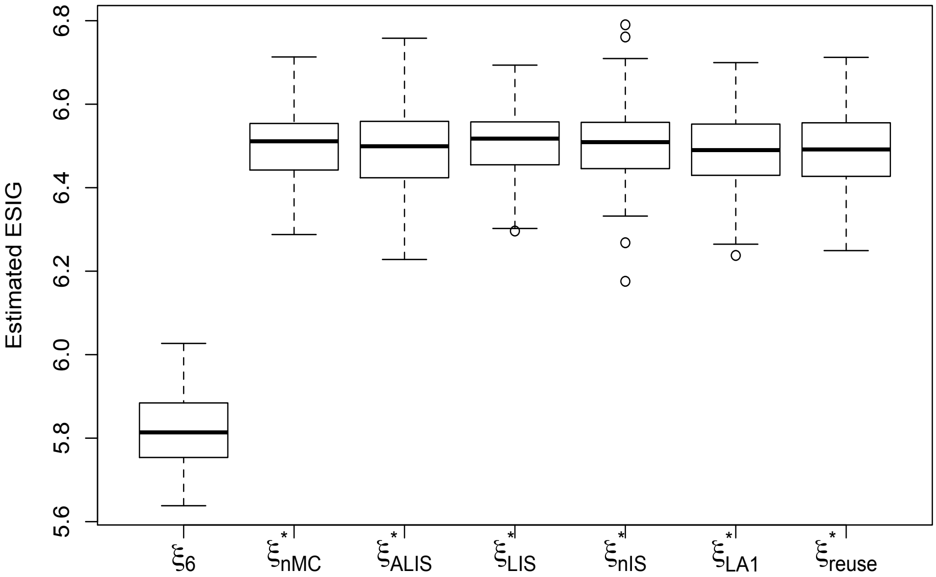


Figure 5.36: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the BOD model for the $n = 6$ optimal designs and and the 6-run design, ξ_6, given by Bates and Watts (1988)

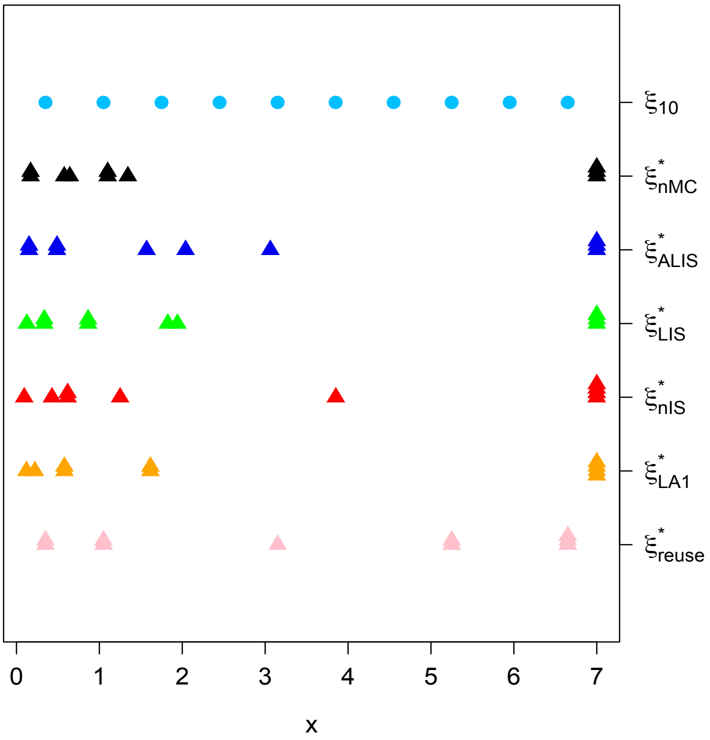


Figure 5.37: ESIG optimal designs with $n = 10$ for the BOD model and the space-filling design, ξ_{10}

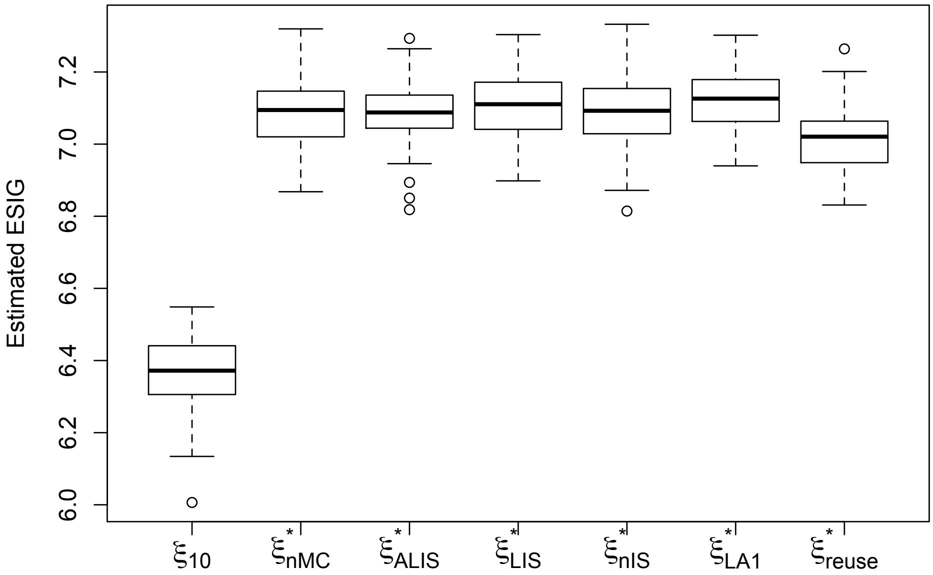


Figure 5.38: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the BOD model for the $n = 10$ optimal designs and the space-filling design, ξ_{10}

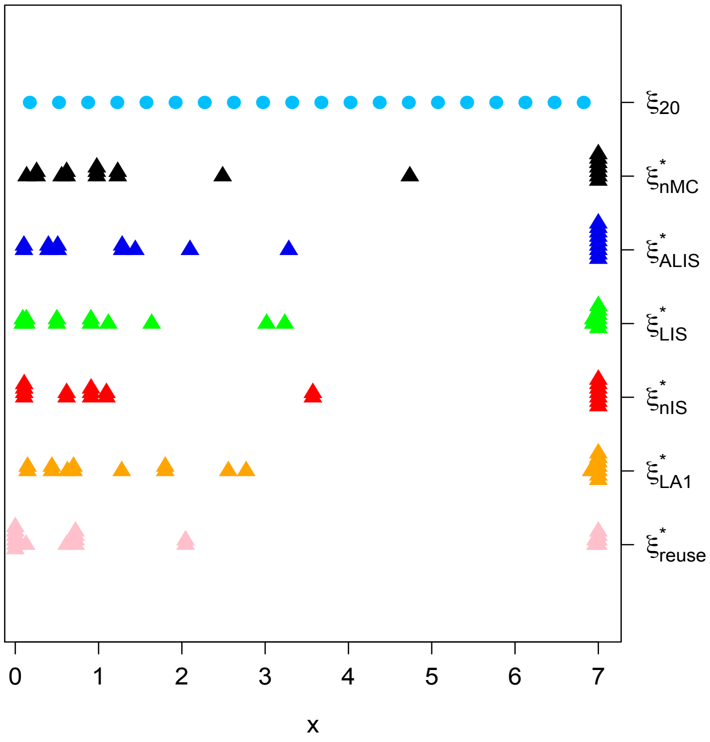


Figure 5.39: ESIG optimal designs with $n = 20$ for the BOD model and the space-filling design, ξ_{20}

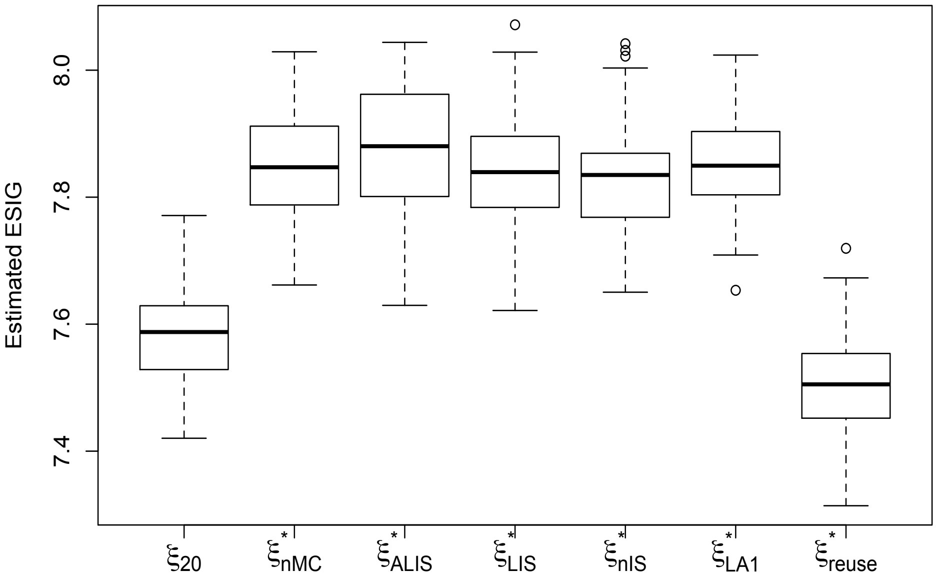


Figure 5.40: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the BOD model for the $n = 20$ optimal designs and the space-filling design, ξ_{20}

and some at the end of the region, where the expected response is more stable. Similar patterns can also be noticed in Figures 5.37 and 5.39 for $n = 10, 20$, with the addition of more repeated points at the start and the end of the design region and some points in the middle of the design region. In Figure 5.37 we notice that the 'reuse' method produces a design where the points cover the design region.

Figures 5.36, 5.38 and 5.40 give the estimated ESIG for the optimal designs for the different methods, approximated using ALIS with $k_1 = k_2 = 300$. For each of the methods 100 Monte Carlo estimates were calculated. In Figures 5.36 and 5.38, all the optimal designs have higher ESIG than the 6-run design given by Bates and Watts (1988) and the space-filling design with $n = 10$, respectively, and all optimal designs have similar ESIG. In Figure 5.40 we notice that the optimal design found with the 'reuse' method has similar ESIG to the space-filling and lower ESIG than the optimal designs found with all the other methods, which have higher ESIG than the space-filling design.

For this model, perhaps the poorer performance of the optimal designs obtained from the 'reuse' method is due to the Monte Carlo size used to estimate the ESIG being insufficient.

Lubricant model

In this section we find optimal designs using the ACE algorithm for the lubricant model. Again, 10 random starting designs are used in ACE. The same procedure was followed to find the optimal designs as for the Michaelis-Menten model and the BOD model. For this model we omitted the 'reuse' method due to the poor performance of the optimal designs found for the previous examples.

Figure 5.41 shows the 53-run design as given in Bates and Watts (1988, Appendix 1).

Figure 5.42 shows the optimal designs with $n = 53$ found using ACE with the different methods. We notice that the optimal design produced by LA1 is almost a "one-factor-at-a-time" design (Czitrom, 1999) with variation in x_2 almost only occurring at the lowest value of x_1. Similar, but less extreme, patterns are observed in the designs found using ALIS and LIS. The optimal designs found using nMC, nIS and LA2 vary the values of x_1 and x_2 in the design region more uniformly compared to the other methods.

Figure 5.43 shows boxplots of the Monte Carlo distribution of 100 estimates of the ESIG obtained using ALIS with $k_1 = k_2 = 300$ for the optimal designs found with the different methods. All optimal designs have higher ESIG than the 53-run design from Bates and Watts (1988, Appendix 1). However, the optimal designs found with ALIS,

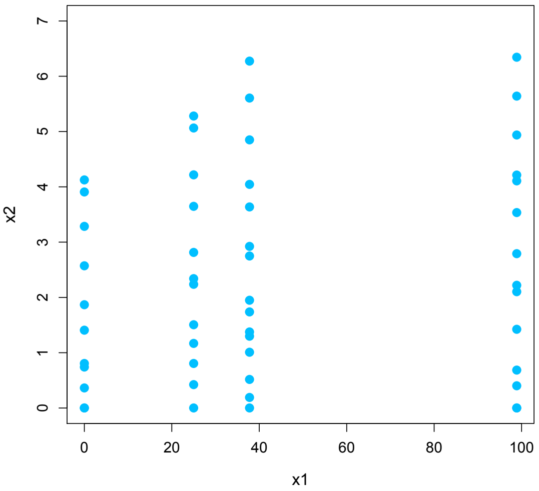


Figure 5.41: The $n = 53$ design, ξ_{53}, as given in Bates and Watts (1988, Appendix 1) for the lubricant model

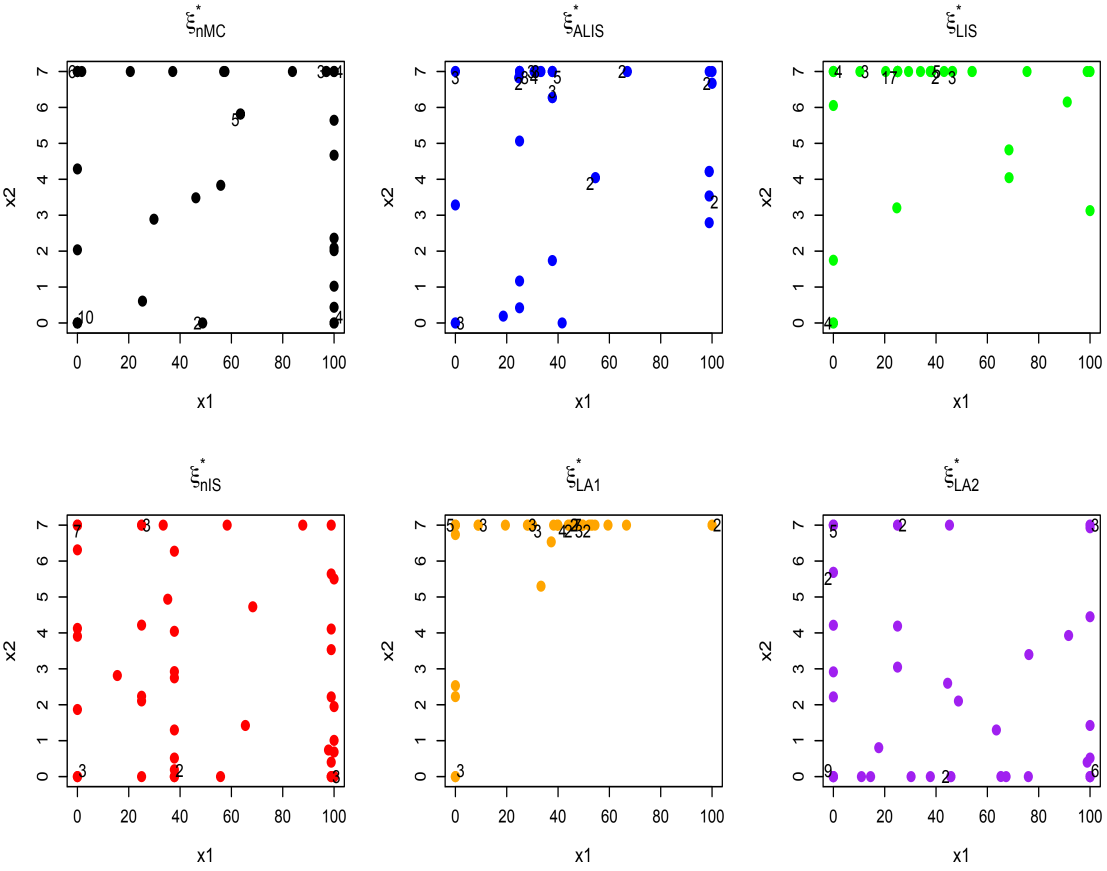


Figure 5.42: ESIG optimal designs with $n = 53$ for the lubricant model (the numbers on some points show how many times the point is repeated)

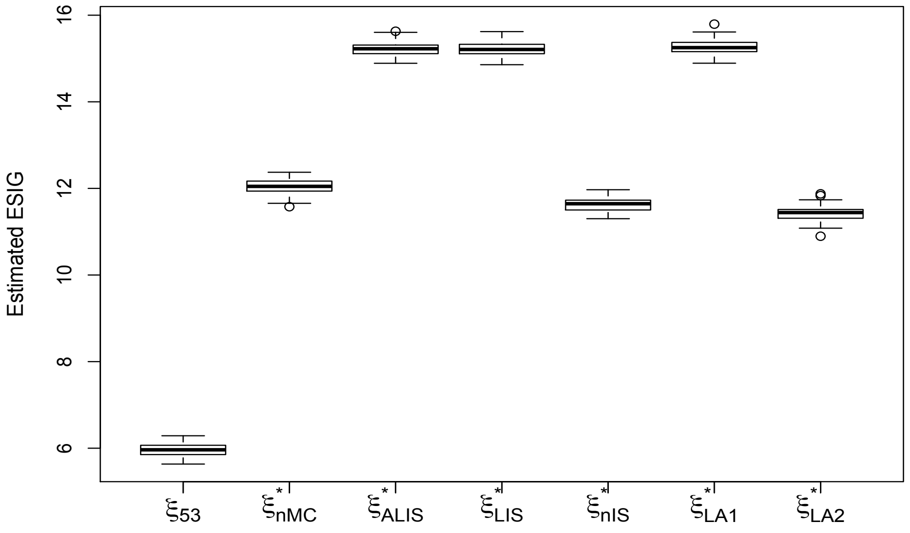


Figure 5.43: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the lubricant model for the 53-run optimal designs shown in Figure 5.42 and the design ξ_{53} shown in Figure 5.41

Parameter	**Mean**	**St. dev.**
θ_3	0.0	5.0
θ_4	0.0	5.0
θ_5	0.0	5.0

Table 5.3: The means and standard deviations of the unknown parameters θ_3, θ_4 and θ_5 of the lubricant model (5.4) used to find $n = 20$ optimal designs

LIS and LA1, which are also more similar to each other, have higher ESIG than the optimal designs found with the other methods.

To investigate the sensitivity of the designs to the choice of the prior distribution, for the same model we change the means and standard deviations of the prior distributions on θ_3, θ_4 and θ_5 and keep the prior distributions of all other parameters fixed (as given in Table 5.2), and find optimal designs using ACE with $n = 20$. The new prior means and standard deviations for θ_3, θ_4 and θ_5 are given in Table 5.3. We follow the same procedure as before.

Figure 5.45 shows the ESIG optimal designs found using ACE with the different methods. The designs found using ALIS and LIS are similar with most points for x_2 take the highest value. LA1 and LA2 have produced designs with fewer distinct values of x_1. For these prior distributions we could not find optimal designs using nMC and nIS due to the zero evidence problem (see Section 4.1.3).

In Figure 5.46 we present the ESIG estimated 100 times and approximated using ALIS with $k_1 = k_2 = 300$, for the four different optimal designs and the 20-run design, ξ_{20}

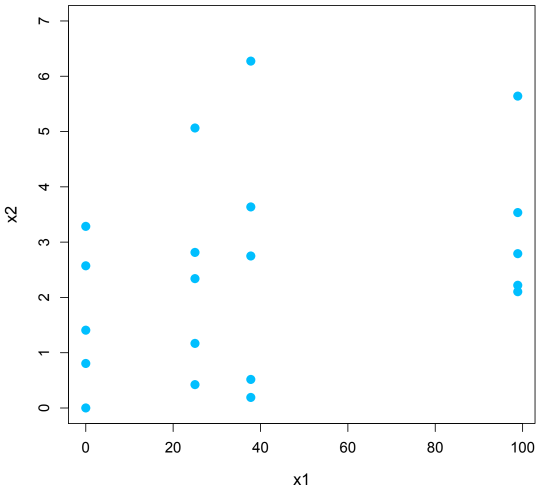


Figure 5.44: A sub-design with $n = 20$, ξ_{20}, chosen from the 53-run design given in Bates and Watts (1988, Appendix 1) for the lubricant model

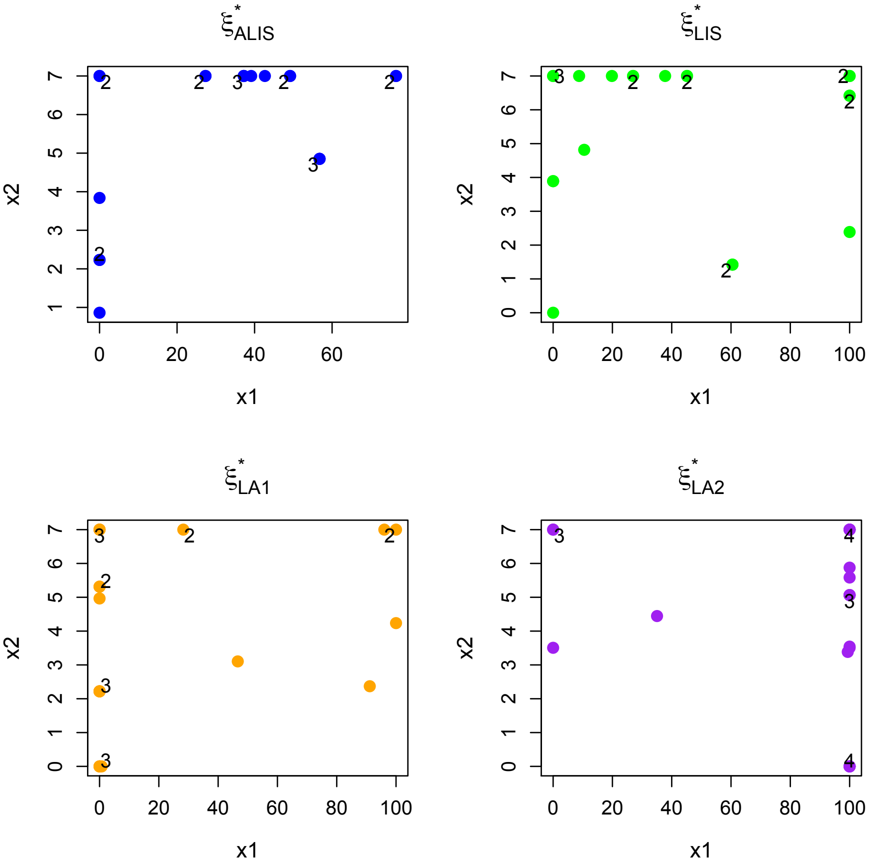


Figure 5.45: ESIG optimal designs with $n = 20$ for the lubricant model (the numbers on some points show how many times the point is repeated)

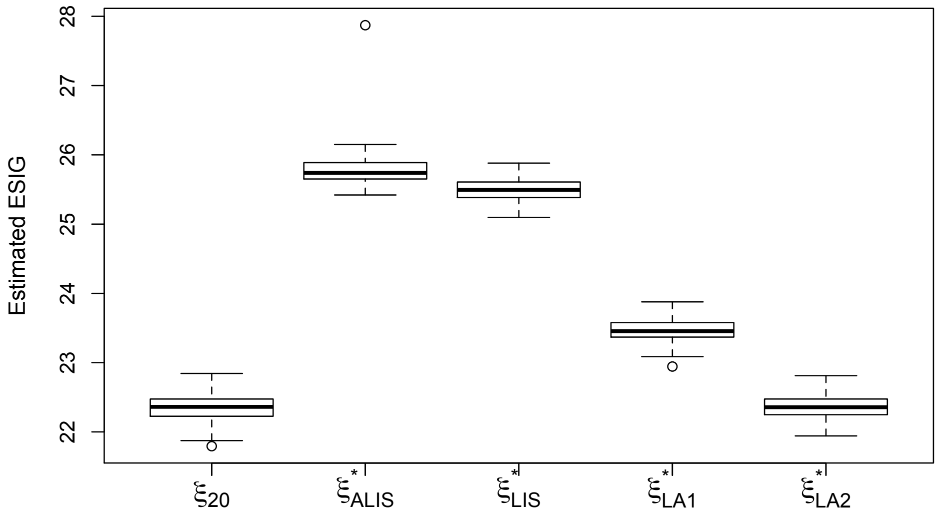


Figure 5.46: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the lubricant model for the 20-run optimal designs shown in Figure 5.45 and the design ξ_{20} shown in Figure 5.44

(a subset of the design given by Bates and Watts, 1988, Appendix 1, see Section 5.1.4 and Figure 5.44). All optimal designs, except the optimal design found using LA2 have higher ESIG than ξ_{20}. In this higher dimensional example, the optimal designs found with ALIS and LIS have higher ESIG than the optimal designs found with the other methods. It is likely that the poorer performance of the designs from LA1 and LA2 occurs due to the larger experiment sizes being required to produce accurate asymptotic results.

5.4 Summary

In this chapter we have showed through nonlinear examples that the new proposed methods of approximating the evidence and hence approximating the expected Shannon information gain, ALIS and LIS, with a moderate Monte Carlo sample size provide a good balance between bias and computational expense. We have also illustrated that ALIS and LIS perform better than existing improved methods. Lastly, we found Bayesian optimal designs by combining the methods introduced in Chapter 4 with the ACE algorithm and showed that for complex models ALIS and LIS have produced designs that have higher expected Shannon information gain than the designs produced with the other methods.

Chapter 6

Bayesian optimal designs for a calibration model

In this chapter we focus on finding fully Bayesian optimal designs for the calibration model (1.1), with a particular focus on designs for the physical experiment. We describe existing methods in the literature for finding optimal designs for the physical experiment in the calibration problem. We use the Kennedy-O'Hagan calibration framework to address design under the two key problems, present in many systems or processes, of model discrepancy and computationally expensive models. We assume that only one of the two problems holds at a time, and Gaussian process priors are used to model unknown functions. ALIS and LIS are used to approximate the expected Shannon information gain and they are combined with the ACE algorithm to find Bayesian optimal designs.

6.1 Statistical calibration

The Kennedy-O'Hagan calibration framework addresses two key problems that are present in many systems or processes:

- the function $\eta(\mathbf{x}, \boldsymbol{\theta})$ may not provide an adequate description of the mean;
- the model may be expensive to evaluate, precluding direct use of the model in inference.

Both of these problems can be addressed using Gaussian processes. This can be done simultaneously (see Section 2.2), but for the purposes of this chapter we assume that only one of these statements holds.

We consider the following statistical model for the physical observation y_i:

$$y_i = \zeta(\mathbf{x}_i) + \varepsilon_i = \eta(\mathbf{x}_i, \boldsymbol{\theta}^p) + \delta_{\boldsymbol{\theta}^p}(\mathbf{x}_i) + \varepsilon_i, \ i = 1, \ldots, n. \tag{6.1}$$

The discrepancy function, $\delta_{\boldsymbol{\theta}^p}(\cdot)$, encodes the difference between the simulator evaluated at the 'true' $\boldsymbol{\theta}^p$, $\eta(\mathbf{x}_i, \boldsymbol{\theta}^p)$, and the mean, $\zeta(\mathbf{x}_i)$, of the physical process. We assume ε_i is the random error and $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ independently.

We divide the calibration problem into the following sub-problems:

SP1. known simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$, with $\delta_{\boldsymbol{\theta}^p}(\mathbf{x}) = 0$ (nonlinear design, see Chapter 5);

SP2. known simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$, with discrepancy;

SP3. unknown simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$, and $\delta_{\boldsymbol{\theta}^p}(\mathbf{x}) = 0$;

SP4. unknown simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$, with discrepancy.

In Section 6.2, we review existing methods in the literature for finding optimal designs for the physical experiment to estimate the parameter $\boldsymbol{\theta}^p = (\theta_1^p, \ldots, \theta_{p_\theta}^p)^{\mathrm{T}}$ in the calibration model (6.1), or predict the mean of the physical system.

In Sections 6.3 and 6.4 we develop novel methodology for optimal design of the physical experiment in Sub-problem 2 (SP2) and Sub-problem 3 (SP3), respectively. Sub-problem 4 (SP4) is left as future work, with further discussion in Section 7.2.

6.2 Experimental designs for simulator calibration

We define a design for a physical experiment as a set $\xi = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ of n points from a design space $\mathcal{X} \subset \mathbb{R}^{q_1}$. A n-size optimal design $\xi^\star$ is defined by comparison with the set Ξ of all possible designs of size n with respect to a specific criterion. We define a design for a computer experiment as a set $\xi^c = [(\mathbf{x}_1^c, \boldsymbol{\theta}_1^c), \ldots, (\mathbf{x}_m^c, \boldsymbol{\theta}_m^c)]$ of m set of choices of input combinations at which to run the simulator in order to collect simulator evaluations to build an emulator, from a design space $\mathcal{X} \times \boldsymbol{\Theta}$. An optimal design $\xi^{c\star}$ of size m is defined by comparison with the set Ξ_c of all possible designs of size m with respect to a specific criterion.

In their important paper, Kennedy and O'Hagan (2001) suggested a sequential design for the simulator, to ensure joint coverage of the calibration input space and the control input space. Also, they recommended that points should be 'close' to physical observations in order to infer the discrepancy function. However, they did not suggest an optimality criterion nor methods for finding optimal designs.

For the simpler problem of Gaussian process interpolation in computer experiments, the most popular design criteria to assess the quality of prediction are functions of the Mean Square Prediction Error (MSPE), see Sacks et al. (1989) and Santner et al. (2003, Chapter 6). The optimal designs are found by minimising the Maximum Mean Square Prediction Error (MMSPE) or the Integrated Mean Square Prediction Error (IMSPE). IMSPE is more commonly used and averages the mean square prediction error over the design space (Hardin and Sloane, 1993). A quasi-Newton algorithm was proposed by Sacks et al. (1989) to find IMPSE-optimal designs.

A number of authors have sought to extend IMSPE criteria to the calibration problem (Ranjan et al., 2011; Williams et al., 2011; Leatherman et al., 2017). Given the combined vector of responses $\mathbf{v} = [\mathbf{y}^{\mathrm{T}}\ \mathbf{z}^{\mathrm{T}}]^{\mathrm{T}}$ and the full vector of model parameters $\boldsymbol{\psi} = [(\boldsymbol{\theta}^p)^{\mathrm{T}}, \boldsymbol{\beta}_\eta^{\mathrm{T}}, \boldsymbol{\beta}_\delta^{\mathrm{T}}, \sigma_\eta^2, \sigma_\delta^2, \sigma_\varepsilon^2, \boldsymbol{\phi}_\eta^{\mathrm{T}}, \boldsymbol{\phi}_\delta^{\mathrm{T}}]^{\mathrm{T}}$ (see Section 2.2), an IMPSE-optimal combined design $[\xi^\star, \xi^{c\star}]$ minimises the objective function

$$\varphi(\xi, \xi^c|\boldsymbol{\psi}) = \int_{\mathcal{X}} \mathbb{E}\left[\left(\mathbb{E}[\zeta(\mathbf{x})|\mathbf{v}, \boldsymbol{\psi}] - \zeta(\mathbf{x})\right)^2 |\boldsymbol{\psi} \right] d\mathbf{x}.$$

Note that this is essentially a local optimality criterion, as it conditions on particular values of the parameters. This seems undesirable, as it means that, for example, the performance of the design for estimating $\boldsymbol{\theta}^p$ is not considered. Our fully Bayesian approach considers the amount of information gained about $\boldsymbol{\theta}^p$.

Ranjan et al. (2011) discussed the design of follow-up experiments for calibration (the selection of new trials that improve the predictive ability of the calibration model). Designs are considered for both the physical experiment and the computer experiment. Two ideas are used to reduce the computational expense of design construction by reducing the dimension of the optimisation problem: replication (forced replicates of field observations leads to a simple estimation procedure for σ_ε^2) and alignment of physical trials and computer trials. The main focus of the paper was the prediction of the physical process at unobserved trial locations, that is, to select new trials that improve the predictive ability of the calibration model (6.1). They constructed IMSPE-optimal follow-up designs for the calibration setting using posterior point estimates of the calibration parameters $\boldsymbol{\theta}^p$ and found that adding physical points gives greater reduction in IMPSE than adding simulator points.

Williams et al. (2011), similar to Ranjan et al. (2011), focussed on batch sequential design optimisation using standard space-filling designs as the initial physical and simulator designs in order to achieve an accurate prediction of the discrepancy (IMPSE for $\delta_{\boldsymbol{\theta}^p}(\cdot)$ to minimise the integrated posterior variance of the discrepancy function conditional on the calibration parameters $\boldsymbol{\theta}^p$). Batch sequential criteria were developed to add new simulation runs for calibration of computer models based on maximising the expected improvement, and MSE-based and distance-based criteria to achieve accurate predictions of quantities of interest. The proposed sequential design criteria are influenced by the existing literature on computer experiments with extensions to allow design augmentation in batches.

Leatherman et al. (2017) focussed on predicting the mean of the physical system based on physical observations and simulator runs, and constructed local IMPSE-optimal designs (local to the parameters). The designs depend on the assumed values for the parameters $\boldsymbol{\theta}^p$ which are unknown prior to experimentation; however a simulation study was performed to examine the prediction accuracy of a range of local IMPSE-optimal designs in order to find out if there is a choice of parameter values that allows accurate empirical predictions for a range of "test-bed" response surfaces. A class of designs was constructed using particle swarm optimisation (Kennedy and Eberhart, 1995) to find an initial design which was then refined using a gradient-based quasi-Newton algorithm to find the optimal designs under IMPSE. These designs were also compared with space-filling designs. They concluded that there is no optimal design that predicts better than all other designs for all "test-beds" and all design sizes.

Huan and Marzouk (2013) proposed an algorithmic approach for optimal Bayesian designs for simulators with zero discrepancy and polynomial chaos emulation. Calibration in such situations is typically a simple statistically identifiable problem. A utility function (Shannon information gain) is used, reflecting expected information gain. A mathematical approximation to the computationally expensive simulator and the naïve Monte Carlo integration method are used to evaluate the expected information gain. Stochastic approximation algorithms are then used to make optimisation feasible.

In this thesis we take a fully Bayesian approach to find optimal designs for the calibration model (6.1). Prior information about unknown parameters and models is represented by prior distributions, and the aim of the experiment is described in the decision-theoretic framework by the utility function. Our goal is to estimate unknown calibration parameters $\boldsymbol{\theta}^p$. Similarly to the previous chapter, the designs found maximise an approximation to the expected Shannon information gain,

$$U(\xi) = \int_{\Psi}\int_{\mathcal{Y}} \log \frac{\pi_l(\mathbf{y}|\boldsymbol{\psi},\xi)}{\pi_e(\mathbf{y}|\xi)} \pi(\mathbf{y},\boldsymbol{\psi}|\xi) d\mathbf{y} d\boldsymbol{\psi}.$$

We approximate the evidence, $\pi_e(\mathbf{y}|\xi)$, in the expected Shannon information gain using the ALIS and LIS approximations described in Section 4.3. The expected Shannon information gain is maximised using the approximate coordinate exchange (ACE) algorithm (Overstall and Woods, 2017) described in Section 5.2.

In the next section we consider SP2 from Section 6.1 and use a Gaussian process prior to model the unknown discrepancy function $\delta_{\boldsymbol{\theta}^p}(\cdot)$.

6.3 Known simulator with discrepancy

Inadequacy of the simulator $\eta(\mathbf{x},\boldsymbol{\theta})$ can be addressed by adopting the extended model given in Equation (6.1),

$$y_i = \zeta(\mathbf{x}_i) + \varepsilon_i = \eta(\mathbf{x}_i,\boldsymbol{\theta}^p) + \delta_{\boldsymbol{\theta}^p}(\mathbf{x}_i) + \varepsilon_i, \quad i = 1,\ldots,n,$$

with $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$. As briefly described in Section 3.2.2, we assume that $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$, the discrepancy between the simulator $\eta(\mathbf{x},\boldsymbol{\theta}^p)$ and the mean $\zeta(\mathbf{x})$ of the physical process, is an unknown function about which we have limited insight and whose form is unknown.

For this reason we assume a Gaussian process prior

$$\delta_{\boldsymbol{\theta}^p}(\mathbf{x}) \sim \mathrm{GP}\left[0, \sigma^2 \kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi})\right].$$

This prior has constant zero mean and covariance $\sigma^2\kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi})$, where σ^2 is the constant variance, $\kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi})$ is the correlation function and $\boldsymbol{\phi}$ is the vector of correlation parameters. For more details on Gaussian processes see Chapter 2.

We assume that the simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$ is a known function but the calibration parameters $\boldsymbol{\theta}^p$ are unknown. Hence, we have that the physical observation y_i comes from a normal distribution with mean $\eta(\mathbf{x}_i, \boldsymbol{\theta}^p)$ and variance $\sigma^2 + \sigma_\varepsilon^2$, and

$$\mathbf{y} \mid \boldsymbol{\theta}^p, \sigma^2, \boldsymbol{\phi}, \sigma_\varepsilon^2 \sim N\left[\boldsymbol{\eta}, \sigma^2\mathbf{K}(\boldsymbol{\phi}) + \sigma_\varepsilon^2\mathbf{I}_n\right],$$

where $\boldsymbol{\eta} = [\eta(\mathbf{x}_1, \boldsymbol{\theta}^p), \ldots, \eta(\mathbf{x}_n, \boldsymbol{\theta}^p)]^{\mathrm{T}}$ is the mean vector, $\mathbf{K}(\boldsymbol{\phi})$ is the $n \times n$ correlation matrix and $\mathbf{I}_n$ is the $n \times n$ identity matrix. We use the reparameterisation $\tau^2 = \sigma_\varepsilon^2/\sigma^2$ described in Section 2.1 to obtain

$$\mathbf{y} \mid \boldsymbol{\theta}^p, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N\left[\boldsymbol{\eta}, \sigma^2\boldsymbol{\Sigma}\right],$$

where $\boldsymbol{\Sigma} = \mathbf{K}(\boldsymbol{\phi}) + \tau^2\mathbf{I}_n$.

The likelihood function is given by

$$\pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}[\mathbf{y} - \boldsymbol{\eta}]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}[\mathbf{y} - \boldsymbol{\eta}]\right\},$$

where $\boldsymbol{\psi} = [(\boldsymbol{\theta}^p)^{\mathrm{T}}, \sigma^2, \boldsymbol{\phi}^{\mathrm{T}}, \tau^2]^{\mathrm{T}}$. Hence the log-likelihood function is:

$$\log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2\sigma^2}[\mathbf{y} - \boldsymbol{\eta}]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}[\mathbf{y} - \boldsymbol{\eta}]. \tag{6.2}$$

The model specification requires prior distributions on the unknown parameters $\boldsymbol{\psi}$.

We aim to approximate the expected Shannon information gain using

$$\tilde{U}(\xi) = \frac{1}{k_1}\sum_{h=1}^{k_1}\left[\log \pi_l(\mathbf{y}_h|\boldsymbol{\psi}_h, \xi) - \log \tilde{\pi}_e^h\right],$$

with $(\boldsymbol{\psi}_h, \mathbf{y}_h) \sim \pi(\boldsymbol{\psi}, \mathbf{y}|\xi)$, $h = 1, \ldots, k_1$, and the approximation $\tilde{\pi}_e^h$ to the evidence, $\pi_e(\mathbf{y}_h|\xi)$, found using ALIS and LIS (see Section 4.3). Technically, in order to use ALIS and LIS to approximate the evidence we have to calculate $\mathbf{H}(\boldsymbol{\psi})$, the negative Hessian of the log-unnormalised posterior density. However, we shall instead obtain a similar approximation by using

$$\mathbf{H}(\boldsymbol{\psi}) = I(\boldsymbol{\psi}; \xi) - \mathbf{Q}(\boldsymbol{\psi}),$$

with $I(\boldsymbol{\psi}; \xi)$ the expected Fisher information matrix, and $\mathbf{Q}(\boldsymbol{\psi})$ the Hessian of the log-prior density. The reason for this choice is that for multivariate normal data, there is a simple expression for the expected Fisher information matrix (see Lemma 6.1).

Lemma 6.1. Assume $\mathbf{y} = [y_1, \dots, y_n]^{\mathrm{T}}$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}(\boldsymbol{\psi})$, covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\psi})$, and $\boldsymbol{\psi} = [\psi_1, \psi_2, \dots, \psi_{q_2}]^{\mathrm{T}}$ is the q_2-dimensional vector of parameters. Then the ijth element, $I_{i,j}$, for $0 \leq i, j \leq q_2$ of the expected Fisher information matrix is given by:

$$I_{i,j} = \frac{\partial \boldsymbol{\mu}(\boldsymbol{\psi})}{\partial \psi_i}^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\psi})^{-1} \frac{\partial \boldsymbol{\mu}(\boldsymbol{\psi})}{\partial \psi_j} + \frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\psi})^{-1} \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\psi})}{\partial \psi_i} \boldsymbol{\Sigma}(\boldsymbol{\psi})^{-1} \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\psi})}{\partial \psi_j} \right), \qquad (6.3)$$

see Porat and Friedlander (1986).

In order to ensure positive values for σ^2, $\boldsymbol{\phi}$ and τ^2, we construct the importance distribution by taking a normal approximation to the posterior of the transformed parameters,

$$\boldsymbol{\psi}' = \left[(\boldsymbol{\theta}^p)^{\mathrm{T}}, \log \sigma^2, \log \phi_1, \dots, \log \phi_{q_1 + p_\theta}, \log \tau^2 \right]^{\mathrm{T}},$$

as described in Section 4.3.2.

6.3.1 Example: Michaelis-Menten simulator and $\delta_{\theta^p}(x) \neq 0$

In this section we compare the performance of four different designs for estimating the unknown calibration parameters in terms of the expected Shannon information gain utility given in Equation (4.5). The expected utility is approximated using LIS[1] as described in Section 4.3. We then find Bayesian optimal designs using ACE (Section 5.2).

We assume the statistical model given in Equation (6.1), where

$$\eta(x_i, \boldsymbol{\theta}) = \frac{\theta_1 x_i}{\theta_2 + x_i}, \quad i = 1, \dots, n.$$

Thus the simulator is the Michaelis-Menten model and $\boldsymbol{\psi} = [\theta_1^p, \theta_2^p, \sigma^2, \phi, \tau^2]^{\mathrm{T}}$. An example of the Michaelis-Menten model, $\eta(x, \boldsymbol{\theta})$, and reality, $\zeta(x)$, is shown in Figure 6.1. The correlation parameter ϕ here is a scalar as there is a single control variable x. We assume prior distributions for the unknown parameters, $\theta_1^p \sim \log N(\mu_1, \sigma_1^2)$, $\theta_2^p \sim \log N(\mu_2, \sigma_2^2)$, $\sigma^2 \sim \mathrm{IG}(a, b)$, $\phi \sim \mathrm{Exp}(\lambda_\phi)$ and $\tau^2 \sim \mathrm{Exp}(\lambda_{\tau^2})$. We also assume a Gaussian process prior for the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x) \sim \mathrm{GP}\left[0, \sigma^2 \kappa(x, x'; \phi)\right]$ with the squared exponential correlation function (2.6). The log-likelihood function is given by Equation (6.2).

The log-prior density for $\boldsymbol{\psi}$ is given by:

$$\log \pi_b(\boldsymbol{\psi}) = \log \pi_b(\theta_1^p) + \log \pi_b(\theta_2^p) + \log \pi_b(\sigma^2) + \log \pi_b(\phi) + \log \pi_b(\tau^2)$$

[1] Previous examples in Chapter 5 have shown that LIS is more stable than ALIS for this type of example and we will use LIS as the default in this section.

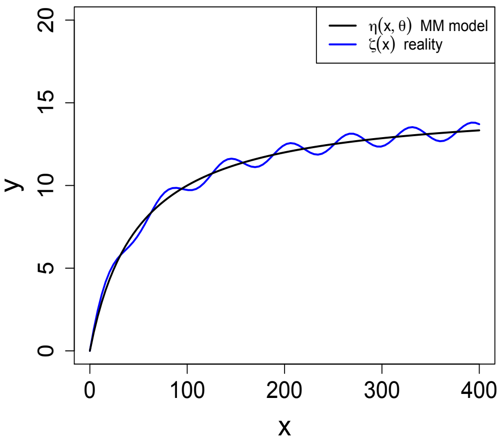


Figure 6.1: The Michaelis-Menten model for $\theta_1^p = 15$ and $\theta_2^p = 50$ (black line) and an example of reality $\zeta(x)$ where we assumed a sinusoidal function for $\delta_{\boldsymbol{\theta}^p}(x)$ (blue line)

$$= -\log\left[\theta_1^p\sigma_1(2\pi)^{1/2}\right] - \frac{(\log\theta_1^p - \mu_1)^2}{2\sigma_1^2} - \log\left[\theta_2^p\sigma_2(2\pi)^{1/2}\right] - \frac{(\log\theta_2^p - \mu_2)^2}{2\sigma_2^2}$$
$$+ \log\left[\frac{b^a}{\Gamma(a)}\right] - (a+1)\log\sigma^2 - \frac{b}{\sigma^2} + \log\lambda_\phi - \lambda_\phi\phi + \log\lambda_{\tau^2} - \lambda_{\tau^2}\tau^2,$$

and the log-unnormalised posterior density, $\log\pi_u(\boldsymbol{\psi}|\mathbf{y},\xi) = \log\pi_l(\mathbf{y}|\boldsymbol{\psi},\xi) + \log\pi_b(\boldsymbol{\psi})$.

For this example, we aim to construct an importance distribution that guarantees positive values of all parameters $\boldsymbol{\psi}$. Hence, we take a normal approximation to the distribution of $\boldsymbol{\psi}' = [\log\theta_1^p, \log\theta_2^p, \log\sigma^2, \log\phi, \log\tau^2]^{\mathrm{T}}$ as described in Section 4.3.2. In order to calculate the negative Hessian of the log-unnormalised posterior density, $\mathbf{H}_{\boldsymbol{\psi}'}(\boldsymbol{\psi}')$, in ALIS and LIS (Section 4.3), we first have to find the derivatives of the log-unnormalised posterior density $\log\pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y},\xi)$ with respect to $\boldsymbol{\psi}'$ using Equations (4.36). These derivatives can be found in Appendix B.2.4.

For the hyperparameters we assume $\mu_1 = 4.38$, $\sigma_1 = 0.07$, $\mu_2 = 1.19$, $\sigma_2 = 0.84$, to match the hyperparameters used for the nonlinear Michaelis-Menten model in Section 5.1.2, $a = 3$, $b = 2$ and $\lambda_\phi = 200$, $\lambda_{\tau^2} = 50$ to guarantee small values for the correlation parameter ϕ and the nugget τ^2. See Appendix C.1.5 for examples of the shape of the expected response of $\eta(x,\boldsymbol{\theta})$, and samples from the prior distribution of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$. The choice of prior distribution of τ^2 implies that the 10% and 90% quantiles of the noise-to-signal ratio (σ_ε divided by the maximum expected response, $\eta(400,\boldsymbol{\theta})$) are 0.0005 and 0.003, respectively. The prior distributions of σ^2 and τ^2 imply that the 10% and 90% quantiles of the error variance σ_ε^2, are 0.003 and 0.112, respectively.

We compare the performance of the four different designs shown in Figure 6.2. Each design consists of ten points.

- Design 1, $\xi_D^\star$, is a two-point maximin $D-$optimal design as described in Section

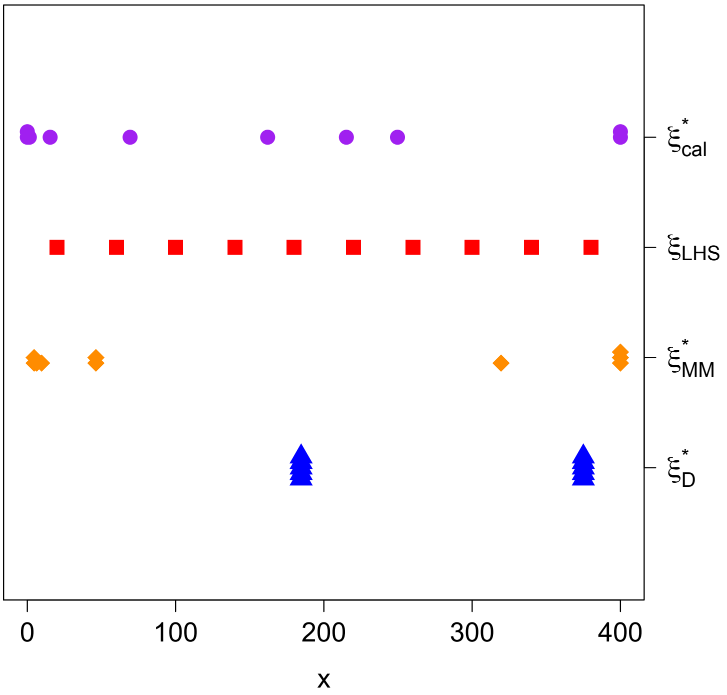


Figure 6.2: The four designs compared for the calibration model with known simulator, the Michaelis-Menten model and $\delta_{\boldsymbol{\theta}^p}(x) \neq 0$

3.2.2 (Dette and Biedermann, 2003). Each point of the design is repeated five times. This design is based on the assumption that the Michaelis-Menten model is correct.

- Design 2, $\xi^\star_{MM}$, is a ten-point optimal design for the Michaelis-Menten model found be combining LIS with ACE and assuming $\delta_{\boldsymbol{\theta}^p}(x) = 0$, i.e. that the model is correct (see Section 5.1.2). This design features five support points, where the first point is replicated twice, the third point is replicated twice, the sixth point is repeated three times and the fifth point is replicated twice. We notice that most points of this design are concentrated at the part of the design space where the simulator is changing most quickly and also there are some points at the stationary part of the curve.

- Design 3, ξ_{LHS}, is a space-filling design, specifically a maximin Latin Hypercube design with ten points (Santner et al., 2003, Chapter 5) and does not take into account the model.

- Design 4, $\xi^\star_{cal}$, is a ten point optimal design for the calibration model assuming the Michaelis-Menten model with discrepancy, found again by combining LIS with ACE. This design features eight support points, where the first and the eighth point are replicated twice. This design appears to be a compromise between Design 2 and Design 3, which we would expect since a design with a greater

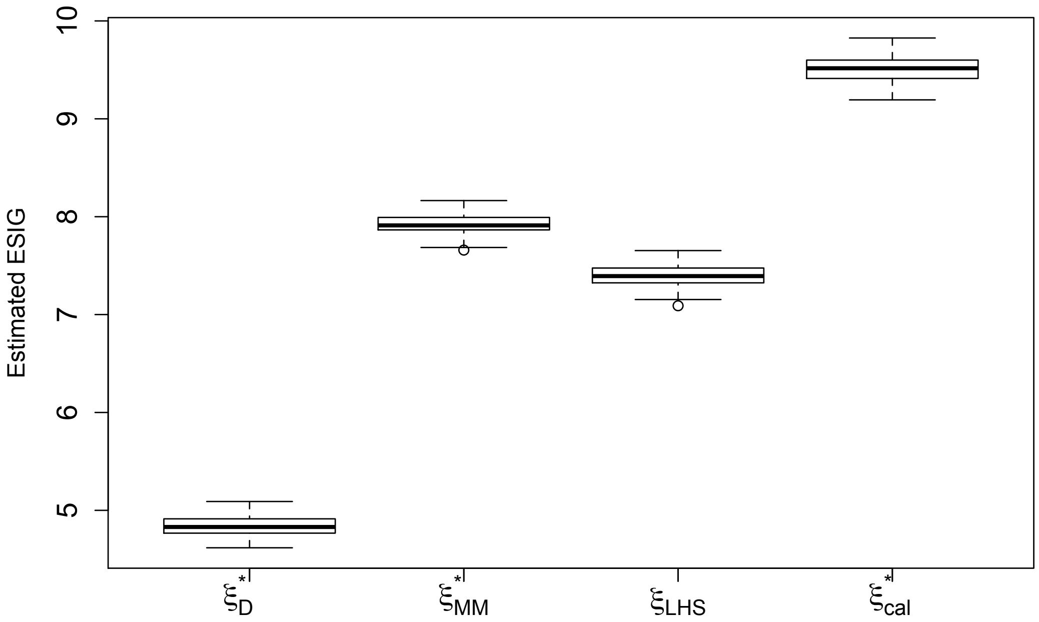


Figure 6.3: Estimated expected Shannon information gain for the parameters $\boldsymbol{\psi}$ of the Michaelis-Menten calibration model (estimated using LIS with $k_1 = k_2 = 300$ for each of the four designs shown in Figure 6.2)

spread of points will be able to better capture the discrepancy function.

Figure 6.3 shows, for each design, 100 independent estimates of the expected Shannon information gain obtained using LIS with $k_1 = k_2 = 300$ which as shown in Chapter 5 is computationally efficient. Design 1, $\xi^{\star}_D$ (maximin D-optimal design), appears to have the worst performance, i.e. the lowest expected Shannon information gain. Design 2, $\xi^{\star}_{MM}$ (Bayesian nonlinear regression design), and Design 3, ξ_{LHS} (maximin LHS design), have a similar performance with $\xi^{\star}_{MM}$ performing a little better. The design with the best performance is Design 4, $\xi^{\star}_{cal}$ (Bayesian optimal calibration design). Hence, we are able to conclude that designs tailored to the calibration problem can perform better than either existing optimal designs or space-filling designs.

In Appendix C.2, for each of the designs given in Figure 6.2 we give further comparisons using LIS and comparisons using nMC and ALIS for two combinations of k_1 and k_2.

In Figure 6.4 we present Bayesian optimal designs for the Michaelis-Menten calibration model as we change the number of points ($n = 5, 10, 20$) and keep the prior information fixed. These designs were found by combining LIS with the ACE algorithm. Most of the new points are added where the simulator is changing most quickly and tend to be spread over the design region in order to capture the form of the discrepancy function.

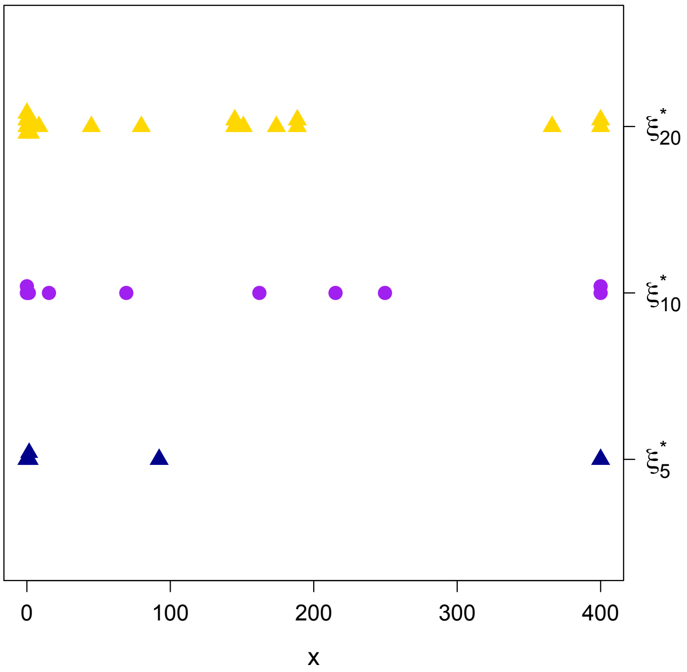


Figure 6.4: Bayesian optimal designs for the Michaelis-Menten calibration model where $\xi_5^\star$ is the optimal design with $n = 5$, $\xi_{10}^\star$ is the optimal design with $n = 10$ (also given in Figure 6.2 as $\xi_{cal}^\star$) and $\xi_{20}^\star$ is the optimal design with $n = 20$

Optimal design	a	b	**Mean**	**Variance**
$\xi_1^\star$	20	0.5	0.03	3.8×10^{-5}
$\xi_2^\star$	15	1	0.07	4×10^{-4}
$\xi_3^\star$	12	1.5	0.13	0.01
$\xi_4^\star$	10	2	0.22	0.06
$\xi_5^\star$	7	3	0.5	0.05
$\xi_6^\star$	5	4	1.0	0.33
$\xi_{cal}^\star$	3	2	1.0	1.0
$\xi_7^\star$	3	4	2.0	4.0
$\xi_8^\star$	3	7	3.5	12.25

Table 6.1: The values of the hyperparameters a and b of the inverse-gamma prior distribution of the Gaussian process variance σ^2 and the implied mean and variance for each combination of a and b used to obtain the Bayesian optimal designs presented in Figure 6.5

Next we find Bayesian optimal designs as we change the prior of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$. In particular we change the values for the hyperparameters a and b of the inverse-gamma distribution for the Gaussian process variance σ^2.

In Figure 6.5 we present ten Bayesian optimal designs found using LIS and ACE for

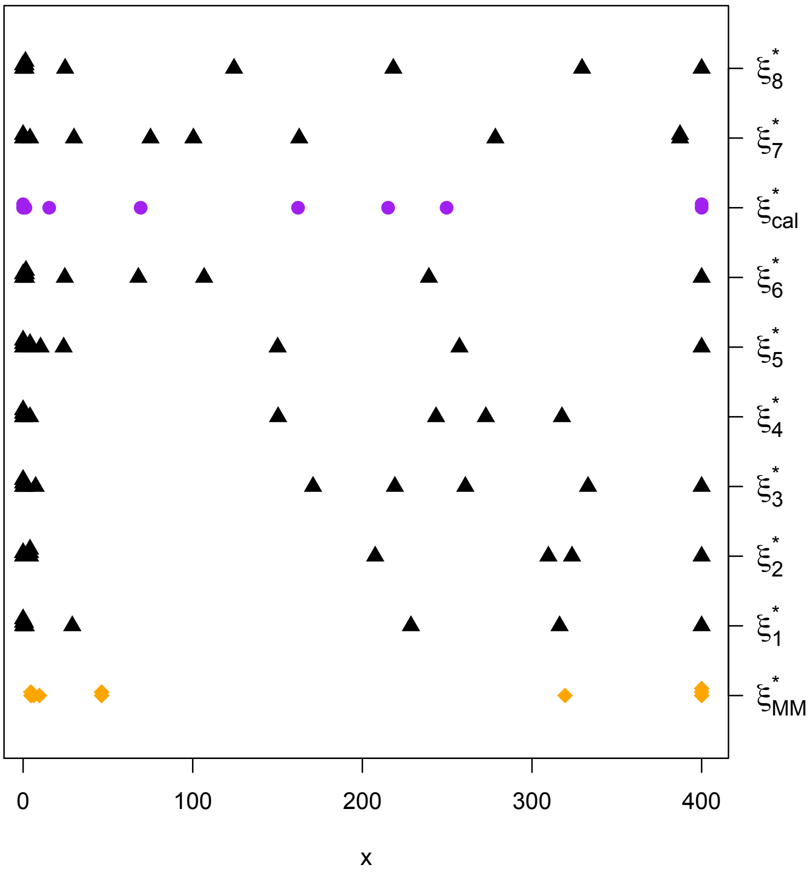


Figure 6.5: Bayesian optimal designs for the parameters $\boldsymbol{\psi}$ of the Michaelis-Menten calibration model found using LIS and $k_1 = k_2 = 300$ for different priors on the discrepancy for each design; the design with points plotted using orange diamonds is the Bayesian optimal design for the nonlinear Michaelis-Menten model (i.e. $\delta_{\boldsymbol{\theta}^p}(x) = 0$) and the design with points plotted using purple bullets is the Bayesian optimal design for the Michaelis-Menten calibration model, as given in Figure 6.2

each combination of values of a and b as given in Table 6.1. In this figure we have also included the optimal design for the nonlinear Michaelis-Menten model and the optimal design for the Michaelis-Menten calibration model for the original prior distribution ($a = 3$, $b = 2$). Small prior mean and variance for σ^2 implies small variance for the Gaussian process model and the Bayesian optimal designs obtained are similar to the Bayesian optimal design for the Michaelis-Menten model ($\delta_{\boldsymbol{\theta}^p}(x) = 0$). Large prior mean and variance for σ^2 implies large variance for the Gaussian process model and the Bayesian optimal designs obtained are similar to designs that are equally spaced across the design region (space-filling designs).

Lastly, we find Bayesian optimal designs for the calibration model assuming that $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^{\mathrm{T}}$ are nuisance parameters (for more information on nuisance parameters see Section 4.3.1). Hence the designs are tailored to estimate the parameters of interest $\boldsymbol{\theta}^p = (\theta_1^p, \theta_2^p)^{\mathrm{T}}$.

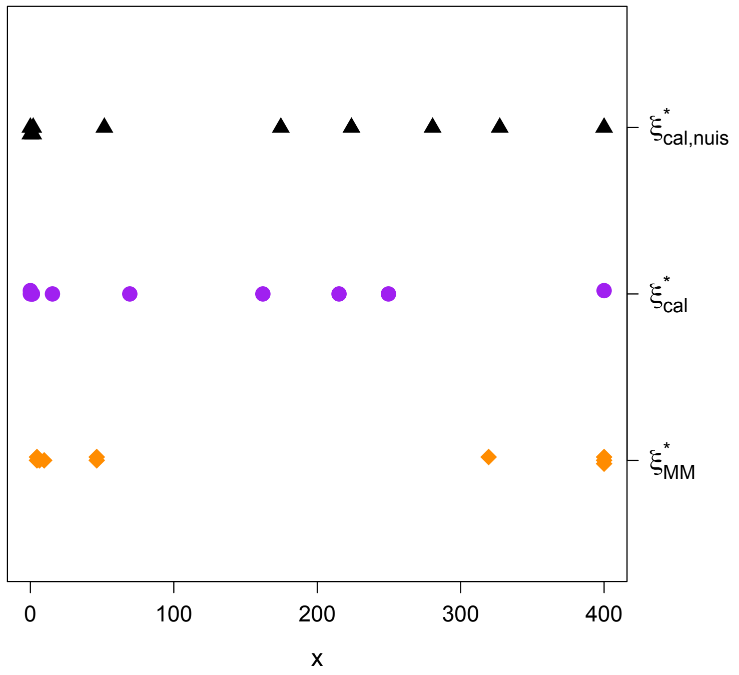


Figure 6.6: Bayesian optimal designs for (i) the Michaelis-Menten model, $\xi^\star_{MM}$ (orange); (ii) the calibration model, $\xi^\star_{cal}$ (purple, also given in Figure 6.4) (iii) the calibration model when $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^{\mathrm{T}}$ are treated as nuisance parameters, $\xi^\star_{cal,nuis}$ (black)

In Figure 6.6 we compare Bayesian optimal designs found using LIS and ACE for: (i) the Michaelis-Menten model with no discrepancy, $\xi^\star_{MM}$; (ii) the calibration model, $\xi^\star_{cal}$; (iii) the calibration model where σ^2, ϕ and τ^2 are nuisance parameters, $\xi^\star_{cal,nuis}$. The first two designs are also presented in Figure 6.2. Treating $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^{\mathrm{T}}$ as nuisance parameters results in an optimal design with points that are more spread over the design region. As we would expect, this is more similar to the optimal design for the calibration model instead of the Michaelis-Menten model with no discrepancy.

Figure 6.7 shows boxplots for each design presented in Figure 6.6, corresponding to the distribution of 100 approximations of the ESIG for the full parameter vector $\boldsymbol{\psi}$ for (a) the optimal design obtained for the Michaelis-Menten model with no discrepancy, $\xi^\star_{MM}$; (b) the optimal design obtained by treating $\boldsymbol{\psi} = (\theta^p_1, \theta^p_2, \sigma^2, \phi, \tau^2)^{\mathrm{T}}$ as interest parameters, $\xi^\star_{cal}$; and (c) the optimal design obtained by treating $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^{\mathrm{T}}$ as nuisance parameters, $\xi^\star_{cal,nuis}$. To perform the calculation, LIS was used with $k_1 = k_2 = 300$. Designs $\xi^\star_{cal}$ and $\xi^\star_{cal,nuis}$ have very similar performance for estimating $\boldsymbol{\psi}$. This means that the optimal design found treating $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^{\mathrm{T}}$ as nuisance parameters also performs well under the calibration model treating all parameters as interest parameters.

Figure 6.8, similarly to Figure 6.7, shows boxplots for each design presented in Figure 6.6, corresponding to the distribution of 100 estimates of the ESIG for a known (correct) Michaelis-Menten model. To estimate the expected utility, LIS was used with $k_1 = k_2 =$

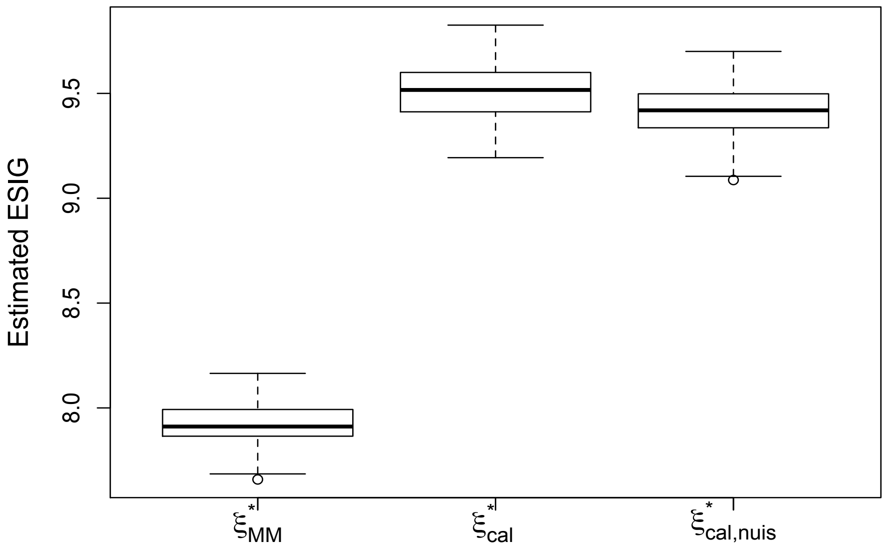


Figure 6.7: Estimated ESIG for the parameters $\boldsymbol{\psi}$ of the Michaelis-Menten calibration model for the optimal designs obtained for the Michaelis-Menten model with no discrepancy, $\xi^\star_{MM}$, the calibration model by treating the full vector $\boldsymbol{\psi}$ as parameters of interest, $\xi^\star_{cal}$, and the calibration model when treating $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^\mathrm{T}$ as nuisance parameters, $\xi^\star_{cal,nuis}$

300. The Bayesian optimal design for the Michaelis-Menten model with no discrepancy, $\xi^\star_{MM}$, has the best performance. The Bayesian optimal design for the calibration model when treating $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^\mathrm{T}$ as nuisance parameters, $\xi^\star_{cal,nuis}$, has similar performance to the optimal design for the Michaelis-Menten model, as it is an optimal design suitable for estimating the parameters of interest. The design with the worst performance is the Bayesian optimal design for the calibration model with the full vector $\boldsymbol{\psi}$ treated as parameters of interest, $\xi^\star_{cal}$.

Lastly, in Figure 6.9 we present boxplots for each design presented in Figure 6.6, corresponding to the distribution of 100 estimates of the ESIG for the calibration model where $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^\mathrm{T}$ are treated as nuisance parameters. To estimate the expected utility, LIS was used with $k_1 = k_2 = k_3 = 300$ (see Section 4.3.1). The results here are similar to the results presented in Figure 6.7.

The performance of Bayesian optimal designs for the calibration model is affected little whether $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^\mathrm{T}$ are treated as interest parameters or nuisance parameters when finding or assessing the design. However, the optimal design for estimating $\boldsymbol{\psi}$ in the calibration model is less effective than the other two designs if the Michaelis-Menten model is correct.

In this chapter we found designs for a calibration model without addressing the identifiability issue discussed in detail in Section 1.1. To find designs for closely related

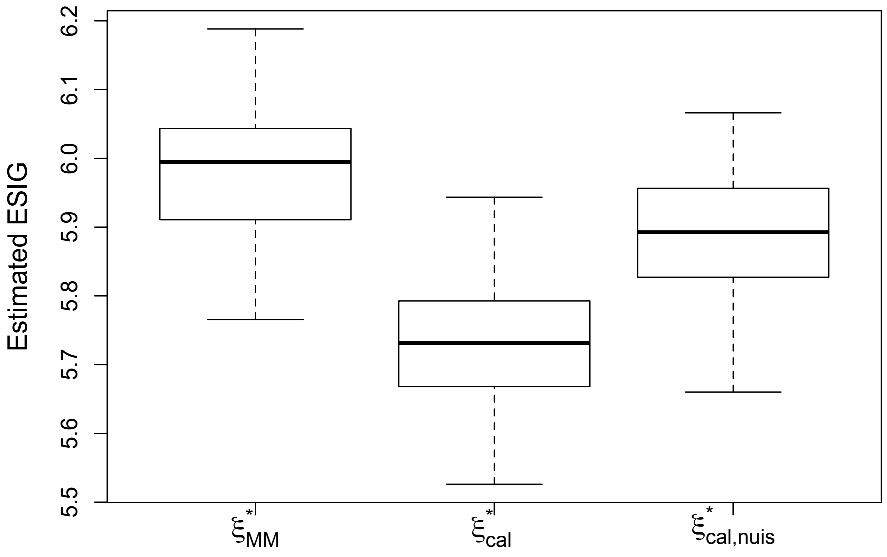


Figure 6.8: Estimated ESIG for the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)^{\mathrm{T}}$ of the Michaelis-Menten model with no discrepancy, for the optimal designs obtained for the Michaelis-Menten model with no discrepancy, $\xi^{\star}_{MM}$, the calibration model by treating the full vector $\boldsymbol{\psi}$ as parameters of interest, $\xi^{\star}_{cal}$, and the calibration model when treating $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^{\mathrm{T}}$ as nuisance parameters, $\xi^{\star}_{cal,nuis}$

identifiable formulation such as L_2-calibration different priors on the discrepancy function can be used, see Plumlee (2017).

In the next section we assume that $\eta(\mathbf{x}, \boldsymbol{\theta})$ is a computationally expensive or unknown simulator and the discrepancy function $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$ is zero (SP3 of Section 6.1).

6.4 Computationally expensive or unknown simulator

In this section, we consider SP3 from Section 6.1. Namely, we assume that the simulator, $\eta(\mathbf{x}, \boldsymbol{\theta})$, is computationally expensive to run, and hence its value is unknown except at a small number of input combinations $(\mathbf{x}^c_1, \boldsymbol{\theta}^c_1), \ldots, (\mathbf{x}^c_m, \boldsymbol{\theta}^c_m)$, where simulator evaluations $\eta(\mathbf{x}^c_j, \boldsymbol{\theta}^c_j)$ have been collected in a computer experiment. Uncertainty about the simulator output at untried input combinations is modelled by placing a Gaussian process prior on $\eta(\mathbf{x}, \boldsymbol{\theta})$, and conditioning on the computer experiment data. Also, for the purposes of this section we assume that the discrepancy between the simulator and reality, $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$, is zero. We approximate the expected Shannon information gain using ALIS and LIS approximations and find Bayesian optimal designs for the physical experiment using ACE.

Due to its computational expense, the simulator cannot be used directly in inference

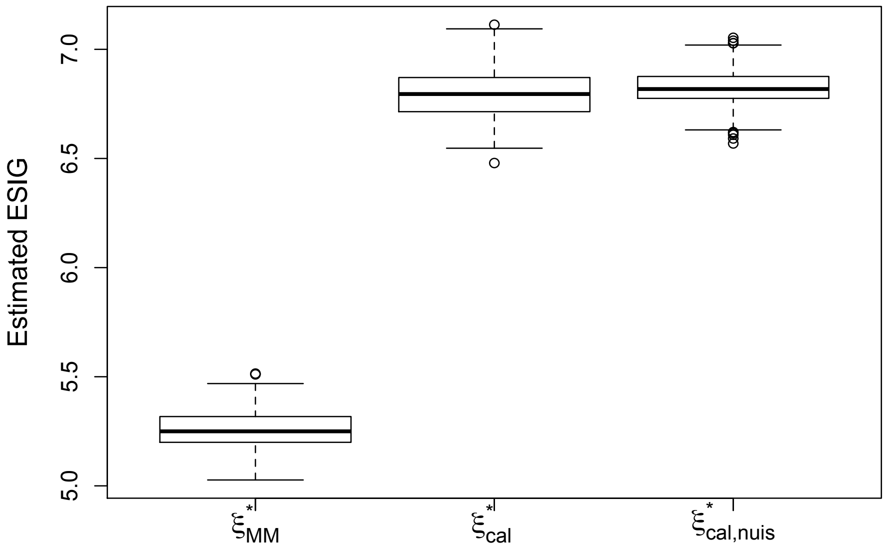


Figure 6.9: Estimated ESIG for the parameters $\boldsymbol{\theta}^p = (\theta_1^p, \theta_2^p)^{\mathrm{T}}$ of the calibration model, when treating $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^{\mathrm{T}}$ as nuisance parameters, for three different designs; the designs are the optimal design obtained under the Michaelis-Menten model with no discrepancy, $\xi^\star_{MM}$, the calibration model by treating the full vector $\boldsymbol{\psi}$ as parameters of interest, $\xi^\star_{cal}$, and the calibration model when treating $\boldsymbol{\gamma} = (\sigma^2, \phi, \tau^2)^{\mathrm{T}}$ as nuisance parameters, $\xi^\star_{cal,nuis}$

or when constructing designs. In this case, the calibration model (6.1) takes the form

$$y_i = \eta(\mathbf{x}_i^p, \boldsymbol{\theta}^p) + \varepsilon_i, \quad i = 1, \dots, n, \tag{6.4}$$

where the random error $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ independently.

Let $\mathbf{y} = [y_1, \dots, y_n]^{\mathrm{T}}$ be the vector of n observations from the physical experiment and $\mathbf{z} = [\eta(\mathbf{x}_1^c, \boldsymbol{\theta}_1^c), \dots, \eta(\mathbf{x}_m^c, \boldsymbol{\theta}_m^c)]^{\mathrm{T}}$ the vector of simulator evaluations from the m-run computer experiment.

We represent prior uncertainty about the simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$ by a Gaussian process,

$$\eta(\mathbf{x}, \boldsymbol{\theta}) \sim \mathrm{GP}\left(\mathbf{f}^{\mathrm{T}}(\mathbf{x}, \boldsymbol{\theta})\boldsymbol{\beta},\ \sigma^2\kappa[(\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}'); \boldsymbol{\phi}]\right), \tag{6.5}$$

where $\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) = (f_0(\mathbf{x}, \boldsymbol{\theta}), \dots, f_{k_\eta - 1}(\mathbf{x}, \boldsymbol{\theta}))^{\mathrm{T}}$ is the k_η-vector of known regression functions and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{k_\eta - 1})^{\mathrm{T}}$ is the corresponding k_η-parameter vector that contains the unknown regression parameters for the emulator of the simulator. In addition, $\kappa[(\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}'); \boldsymbol{\phi}]$ is the correlation function, with vector of correlation parameters $\boldsymbol{\phi}$, and σ^2 is the prior variance (see also Section 2.2).

Following the results presented in Chapter 2 we have that,

$$\mathbf{z} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi} \sim N(\mathbf{F}^c\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma}_{cc}), \tag{6.6}$$

where $\mathbf{F}^c = [\mathbf{f}(\mathbf{x}_1^c, \boldsymbol{\theta}_1^c)\ \mathbf{f}(\mathbf{x}_2^c, \boldsymbol{\theta}_2^c) \ldots \mathbf{f}(\mathbf{x}_m^c, \boldsymbol{\theta}_m^c)]^{\mathrm{T}}$ is the $m \times k_\eta$ model matrix of the computer experiment and $\boldsymbol{\Sigma}_{cc}$ is defined through the correlation function with jj'th entry $\boldsymbol{\Sigma}_{cc,jj'} = \kappa[(\mathbf{x}_j^c, \boldsymbol{\theta}_j^c), (\mathbf{x}_{j'}^c, \boldsymbol{\theta}_{j'}^c); \boldsymbol{\phi}]$, $j, j' = 1, \ldots, m$.

The prior joint density of $\boldsymbol{\beta} \mid \sigma^2$ and σ^2 corresponds to a normal-inverse-gamma distribution (see Section 2.3.3),

$$(\boldsymbol{\beta}, \sigma^2) \sim NIG\left(\boldsymbol{\beta}_0, \mathbf{R}, a, b\right).$$

The conditional posterior density $\pi_a(\boldsymbol{\beta}, \sigma^2 | \mathbf{z}, \boldsymbol{\phi})$ (see Section 2.3.4) also corresponds to a normal-inverse-gamma distribution $\mathrm{NIG}(\boldsymbol{\beta}_\star, \boldsymbol{\Sigma}_\star, a_\star, b_\star)$ with

$$\begin{aligned}
\boldsymbol{\beta}_\star &= (\mathbf{F}^{c\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c + \mathbf{R}^{-1})^{-1} (\mathbf{F}^{c\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{z} + \mathbf{R}^{-1} \boldsymbol{\beta}_0) \\
\boldsymbol{\Sigma}_\star &= (\mathbf{F}^{c\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c + \mathbf{R}^{-1})^{-1} \\
a_\star &= a + \frac{m}{2} \\
b_\star &= b + \frac{1}{2} \left[(\mathbf{z} - \mathbf{F}^c \boldsymbol{\beta}_0)^{\mathrm{T}} (\boldsymbol{\Sigma}_{cc} + \mathbf{F}^c \mathbf{R} \mathbf{F}^{c\mathrm{T}})^{-1} (\mathbf{z} - \mathbf{F}^c \boldsymbol{\beta}_0) \right], \qquad (6.7)
\end{aligned}$$

as shown in (2.20).

The conditional posterior distribution

$$\boldsymbol{\beta} \mid \mathbf{z}, \sigma^2, \boldsymbol{\phi} \sim N(\boldsymbol{\beta}_\star, \sigma^2 \boldsymbol{\Sigma}_\star), \qquad (6.8)$$

follows from Equation (2.19), and the marginal posterior for σ^2 is an inverse-gamma distribution

$$\sigma^2 \mid \mathbf{z}, \boldsymbol{\phi} \sim IG\left(a_\star, b_\star\right), \qquad (6.9)$$

see Equation (2.23). Both $\boldsymbol{\beta} | \mathbf{z}, \sigma^2, \boldsymbol{\phi}$ and $\sigma^2 | \mathbf{z}, \boldsymbol{\phi}$ are conditionally independent of $\boldsymbol{\theta}^p$ given $\mathbf{z}$.

6.4.1 Conditional prediction with known hyperparameters

The joint prior distribution of physical data, $\mathbf{y}$, and simulator evaluations, $\mathbf{z}$, conditional on all unknown model parameters $\boldsymbol{\theta}^p, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2$, is given by:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} \Bigg| \boldsymbol{\theta}^p, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \quad \sim \quad N\left(\begin{pmatrix} \mathbf{F}^p \boldsymbol{\beta} \\ \mathbf{F}^c \boldsymbol{\beta} \end{pmatrix}, \sigma^2 \begin{pmatrix} \boldsymbol{\Sigma}_{pp} + \tau^2 \mathbf{I}_n & \boldsymbol{\Sigma}_{cp}^{\mathrm{T}} \\ \boldsymbol{\Sigma}_{cp} & \boldsymbol{\Sigma}_{cc} \end{pmatrix} \right), \qquad (6.10)$$

where $\tau^2 = \sigma_\varepsilon^2 / \sigma^2$, $\mathbf{F}^p = [\mathbf{f}(\mathbf{x}_1^p, \boldsymbol{\theta}^p)\ \mathbf{f}(\mathbf{x}_2^p, \boldsymbol{\theta}^p) \ldots\ \mathbf{f}(\mathbf{x}_n^p, \boldsymbol{\theta}^p)]^{\mathrm{T}}$ is the $n \times k_\eta$ model matrix for the physical experiment and $\boldsymbol{\Sigma}_{pp}$, $\boldsymbol{\Sigma}_{cp}$ are defined through the correlation function with entries given by $\boldsymbol{\Sigma}_{pp,ii'} = \kappa[(\mathbf{x}_i^p, \boldsymbol{\theta}^p), (\mathbf{x}_{i'}^p, \boldsymbol{\theta}^p); \boldsymbol{\phi}]$ and $\boldsymbol{\Sigma}_{cp,ji} = \kappa[(\mathbf{x}_j^c, \boldsymbol{\theta}_j^c), (\mathbf{x}_i^p, \boldsymbol{\theta}^p); \boldsymbol{\phi}]$, where $i, i' = 1, \ldots, n$ and $j = 1, \ldots, m$.

Standard results on multivariate normal distributions can be used to derive the following conditional posterior distribution

$$\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}^p, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right),$$

with

$$\boldsymbol{\mu}_y = \mathbb{E}(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}^p, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2) = \mathbf{F}^p\boldsymbol{\beta} + \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\left[\mathbf{z} - \mathbf{F}^c\boldsymbol{\beta}\right], \tag{6.11}$$

and

$$\boldsymbol{\Sigma}_y = \mathrm{var}(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}^p, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2) = \sigma^2[\boldsymbol{\Sigma}_{pp} + \tau^2\mathbf{I}_n - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\boldsymbol{\Sigma}_{cp}]. \tag{6.12}$$

Hence the likelihood for the physical data $\mathbf{y}$, conditional on the simulator evaluations $\mathbf{z}$, is

$$\pi_l(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}^p, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_y|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_y)^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)\right\}.$$

We obtain the marginal distribution of the physical data conditional on simulator evaluations, by using the fact that $\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}^p, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ and $\boldsymbol{\beta} \mid \mathbf{z}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N(\boldsymbol{\beta}_\star, \sigma^2\boldsymbol{\Sigma}_\star)$. It follows that

$$\mathbf{y} - \boldsymbol{\mu}_y \mid \mathbf{z}, \boldsymbol{\theta}^p, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N(\mathbf{0}_n, \sigma^2\boldsymbol{\Sigma}_y^\star), \tag{6.13}$$

where $\boldsymbol{\Sigma}_y = \sigma^2\boldsymbol{\Sigma}_y^\star$. The right hand side of (6.13) does not depend on $\boldsymbol{\beta}$, and so $\mathbf{y} - \boldsymbol{\mu}_y$ is statistically independent of $\boldsymbol{\beta}$ given $\mathbf{z}$, $\boldsymbol{\theta}^p$, σ^2, $\boldsymbol{\phi}$ and τ^2. Moreover $\mathbf{y} - \boldsymbol{\mu}_y$ is also conditionally independent of $\boldsymbol{\mu}_y$, which is a linear transformation of $\boldsymbol{\beta}$ (Equation (6.11)). The vector $\boldsymbol{\mu}_y|\mathbf{z}, \boldsymbol{\theta}^p, \sigma^2, \boldsymbol{\phi}, \tau^2$ also follows a multivariate normal distribution with mean and variance given by:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\mu}_y|\mathbf{z}, \boldsymbol{\theta}^p, \sigma^2, \boldsymbol{\phi}, \tau^2] &= \mathbb{E}(\mathbf{F}^p\boldsymbol{\beta} + \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\left[\mathbf{z} - \mathbf{F}^c\boldsymbol{\beta}\right]) \\
&= \mathbb{E}[\mathbf{F}^p\boldsymbol{\beta}] + \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbb{E}\left[\mathbf{z} - \mathbf{F}^c\boldsymbol{\beta}\right] \\
&= \mathbf{F}^p\boldsymbol{\beta}_\star + \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\left[\mathbf{z} - \mathbf{F}^c\boldsymbol{\beta}_\star\right] \\
&= \mathbf{F}^p(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c + \mathbf{R}^{-1})^{-1}(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{z} + \mathbf{R}^{-1}\boldsymbol{\beta}_0) \\
&+ \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\left[\mathbf{z} - \mathbf{F}^c(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c + \mathbf{R}^{-1})^{-1}(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{z} + \mathbf{R}^{-1}\boldsymbol{\beta}_0)\right],
\end{aligned} \tag{6.14}$$

and

$$\begin{aligned}
\mathrm{var}[\boldsymbol{\mu}_y|\mathbf{z}, \boldsymbol{\theta}^p, \sigma^2, \boldsymbol{\phi}, \tau^2] &= \mathrm{var}[\mathbf{F}^p\boldsymbol{\beta} + \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}(\mathbf{z} - \mathbf{F}^c\boldsymbol{\beta})] \\
&= \mathrm{var}[(\mathbf{F}^p - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c)\boldsymbol{\beta} + \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{z}] \\
&= \sigma^2(\mathbf{F}^p - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c)\boldsymbol{\Sigma}_\star(\mathbf{F}^p - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c)^{\mathrm{T}} \\
&= \sigma^2(\mathbf{F}^p - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c)(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c + \mathbf{R}^{-1})^{-1}(\mathbf{F}^p - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c)^{\mathrm{T}}.
\end{aligned}$$

Hence, given $\mathbf{z}$, $\boldsymbol{\theta}^p$, σ^2, $\boldsymbol{\phi}$ and τ^2,

$$\mathbf{y} = (\mathbf{y} - \boldsymbol{\mu}_y) + \boldsymbol{\mu}_y,$$

is a sum of two independent multivariate normal random variables. Thus,

$$\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}^p, \sigma^2, \boldsymbol{\phi}, \tau^2 \sim N[\tilde{\boldsymbol{\mu}}_y, \sigma^2 \tilde{\boldsymbol{\Sigma}}_y],$$

where $\tilde{\boldsymbol{\mu}}_y = \mathbb{E}[\boldsymbol{\mu}_y | \mathbf{z}, \boldsymbol{\theta}^p, \sigma^2, \boldsymbol{\phi}, \tau^2]$ given by (6.14) and $\tilde{\boldsymbol{\Sigma}}_y$ given by:

$$\tilde{\boldsymbol{\Sigma}}_y = \{\boldsymbol{\Sigma}_y^\star + (\mathbf{F}^p - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c)(\mathbf{F}^{c\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c + \mathbf{R}^{-1})^{-1} (\mathbf{F}^p - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c)^{\mathrm{T}}\}. \tag{6.15}$$

We can then integrate out σ^2 with respect to its marginal posterior distribution, $\sigma^2 \mid \mathbf{z}, \boldsymbol{\phi}, \tau^2 \sim \mathrm{IG}(a_\star, b_\star)$, given in (6.9), to obtain:

$$\begin{aligned}
\pi(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}^p, \boldsymbol{\phi}, \tau^2) &= \int_0^\infty \pi(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}^p, \sigma^2, \boldsymbol{\phi}, \tau^2) \pi(\sigma^2 \mid \mathbf{z}, \boldsymbol{\phi}, \tau^2) d\sigma^2 \\
&= \int_0^\infty \left[\frac{1}{(\sigma^2)^{n/2} \sqrt{(2\pi)^n |\tilde{\boldsymbol{\Sigma}}_y|}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_y^{-1} (\mathbf{y} - \tilde{\boldsymbol{\mu}}_y) \right\} \right. \\
&\qquad \left. \times \frac{(b_\star)^{a_\star}}{\Gamma(a_\star)} (\sigma^2)^{-(a_\star+1)} \exp\left\{ -\frac{b_\star}{\sigma^2} \right\} \right] d\sigma^2 \\
&= \frac{(b_\star)^{a_\star}}{\Gamma(a_\star) \sqrt{(2\pi)^n |\tilde{\boldsymbol{\Sigma}}_y|}} \int_0^\infty (\sigma^2)^{-(a_\star+1+n/2)} \\
&\qquad \times \exp\left\{ -\frac{1}{\sigma^2} \left[\frac{1}{2} (\mathbf{y} - \tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_y^{-1} (\mathbf{y} - \tilde{\boldsymbol{\mu}}_y) + b_\star \right] \right\} d\sigma^2 \\
&= \frac{(b_\star)^{a_\star} \Gamma\left(a_\star + \frac{n}{2}\right)}{\Gamma(a_\star) \sqrt{(2\pi)^n |\tilde{\boldsymbol{\Sigma}}_y|}} \left[b_\star + \frac{(\mathbf{y} - \tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_y^{-1} (\mathbf{y} - \tilde{\boldsymbol{\mu}}_y)}{2} \right]^{-(a_\star+n/2)} \\
&= \frac{(b_\star)^{-n/2} \Gamma\left(a_\star + \frac{n}{2}\right)}{\Gamma(a_\star) \sqrt{(2\pi)^n |\tilde{\boldsymbol{\Sigma}}_y|}} \left[1 + \frac{(\mathbf{y} - \tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_y^{-1} (\mathbf{y} - \tilde{\boldsymbol{\mu}}_y)}{2b_\star} \right]^{-(a_\star+n/2)}.
\end{aligned} \tag{6.16}$$

Equation (6.16) indicates that the predictive distribution of the physical data $\mathbf{y}$ is a multivariate t-distribution,

$$\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}^p, \boldsymbol{\phi}, \tau^2 \sim t_{2a_\star} \left(n, \tilde{\boldsymbol{\mu}}_y, \frac{b_\star}{a_\star} \tilde{\boldsymbol{\Sigma}}_y \right),$$

with $2a_\star$ degrees of freedom, mean $\tilde{\boldsymbol{\mu}}_y$ given by (6.14), and variance $\frac{b_\star}{a_\star - 1} \tilde{\boldsymbol{\Sigma}}_y$, where $\tilde{\boldsymbol{\Sigma}}_y$ is given by (6.15).

Hence we have obtained the marginal posterior predictive distribution of the physical data $\mathbf{y}$ given simulator runs $\mathbf{z}$, calibration parameters $\boldsymbol{\theta}^p$, correlation parameters $\boldsymbol{\phi}$ and the nugget τ^2. This distribution can be used when designing physical experiments.

6.4.2 Approximation of the expected utility

We are interested in finding designs that maximise an approximation to the expected Shannon information gain

$$U(\xi) = \int_{\Psi}\int_{\mathcal{Y}} \log \frac{\pi_l(\mathbf{y}|\mathbf{z}, \boldsymbol{\psi}, \xi)}{\pi_e(\mathbf{y}|\mathbf{z}, \xi)} \pi(\mathbf{y}, \boldsymbol{\psi}|\mathbf{z}, \xi) d\mathbf{y} d\boldsymbol{\psi}.$$

We approximate the expected Shannon information gain using the ALIS and LIS approximations described in Section 4.3, and find optimal designs using the approximate coordinate exchange (ACE) algorithm (see Section 5.2).

The negative Hessian of the log-unnormalised posterior density, $\mathbf{H}(\boldsymbol{\psi})$, must be calculated in order to use ALIS and LIS, where $\boldsymbol{\psi} = [(\boldsymbol{\theta}^p)^{\mathrm{T}}, \tau^2]^{\mathrm{T}}$ and we assume that the vector of correlation parameters, $\boldsymbol{\phi}$, for the emulator is held fixed at the maximum likelihood estimates from the computer experiment. The log-unnormalised posterior density is given by:

$$\begin{aligned}\log \pi_u(\boldsymbol{\theta}^p, \tau^2|\mathbf{y}, \mathbf{z}, \xi) &= \log[\pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}^p, \tau^2, \xi)\pi_b(\boldsymbol{\theta}^p)\pi_b(\tau^2)] \\ &= \log \pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}^p, \tau^2, \xi) + \log \pi_b(\boldsymbol{\theta}^p) + \log \pi_b(\tau^2),\end{aligned}$$

where, from (6.16),

$$\begin{aligned}\log \pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}^p, \tau^2, \xi) = &-\frac{1}{2}\log|\tilde{\boldsymbol{\Sigma}}_y| - \left(a_\star + \frac{n}{2}\right)\log\left[2b_\star + (\mathbf{y} - \tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_y^{-1}(\mathbf{y} - \tilde{\boldsymbol{\mu}}_y)\right] \\ &+ \text{constant}.\end{aligned} \tag{6.17}$$

In order to guarantee positivity of sampled parameters we construct the importance distributions of ALIS and LIS by making a normal approximation to the posterior distribution of a transformed parameter vector $\boldsymbol{\psi}'$, with $\psi'_h = T_h(\psi_h)$, i.e. ψ'_h depends only on the hth component of $\boldsymbol{\psi}'$ and not the other components, as described in Section 4.3.2. For the first example in Section 6.4.3 we use the transformation $\boldsymbol{\psi}' = (\theta^p, \log\tau^2)^{\mathrm{T}}$ and for the second example in Section 6.4.4 we use the transformation $\boldsymbol{\psi}' = (\log\theta^p_1, \log\theta^p_2, \log\tau^2)^{\mathrm{T}}$.

We now find the first and second derivatives of (6.17) required to construct the ALIS and LIS importance densities, which we use in the examples presented in Sections 6.4.3 and 6.4.4.

We have $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{q_2})^{\mathrm{T}}$ and $\boldsymbol{\psi}' = (\psi'_1, \ldots, \psi'_{q_2})^{\mathrm{T}}$. For $h = 1, \ldots, q_2$, the first derivatives are given by:

$$\begin{aligned}\frac{\partial \log \pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\psi}, \xi)}{\partial \psi'_h} &= \frac{\partial \log \pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\psi}, \xi)}{\partial \psi_h}\frac{\partial \psi_h}{\partial \psi'_h} \\ \frac{\partial \log \pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\psi}, \xi)}{\partial \psi_h} &= -\frac{1}{2}\frac{\partial \log|\tilde{\boldsymbol{\Sigma}}_y|}{\partial \psi_h} + a_\star\frac{\partial \log[2b_\star]}{\partial \psi_h}\end{aligned}$$

$$\begin{aligned}
&-\left(a_\star + \frac{n}{2}\right)\left[\frac{2\frac{\partial b_\star}{\partial \psi_h}}{2b_\star + (\mathbf{y}-\tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_y^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y)}\right] \\
&-\left(a_\star + \frac{n}{2}\right)\left[\frac{-2\left(\frac{\partial \tilde{\boldsymbol{\mu}}_y}{\partial \psi_h}\right)^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_y^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y) + (\mathbf{y}-\tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y^{-1}}{\partial \psi_h}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y)}{2b_\star + (\mathbf{y}-\tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_y^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y)}\right],
\end{aligned}$$

where

$$\frac{\partial \log|\tilde{\boldsymbol{\Sigma}}_y|}{\partial \psi_h} = \mathrm{tr}\left[\tilde{\boldsymbol{\Sigma}}_y^{-1}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \psi_h}\right].$$

We first find the derivatives with respect to the true calibration parameters $\boldsymbol{\theta}^p$. For $k = 1, \dots, p_\theta$, we have:

$$\begin{aligned}
\frac{\partial \tilde{\boldsymbol{\mu}}_y}{\partial \theta_k^p} &= \frac{\partial \mathbf{F}^p}{\partial \theta_k^p}(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c + \mathbf{R}^{-1})^{-1}(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{z} + \mathbf{R}^{-1}\boldsymbol{\beta}_0) \\
&\quad + \frac{\partial \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_k^p}\boldsymbol{\Sigma}_{cc}^{-1}\left[\mathbf{z} - \mathbf{F}^c(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c + \mathbf{R}^{-1})^{-1}(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{z} + \mathbf{R}^{-1}\boldsymbol{\beta}_0)\right], \\
\frac{\partial \tilde{\boldsymbol{\Sigma}}_y^{-1}}{\partial \theta_k^p} &= -\tilde{\boldsymbol{\Sigma}}_y^{-1}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta_k^p}\tilde{\boldsymbol{\Sigma}}_y^{-1}, \\
\frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta_k^p} &= \frac{\partial \boldsymbol{\Sigma}_{pp}}{\partial \theta_k^p} - \frac{\partial \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_k^p}\boldsymbol{\Sigma}_{cc}^{-1}\boldsymbol{\Sigma}_{cp} - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\frac{\partial \boldsymbol{\Sigma}_{cp}}{\partial \theta_k^p} \\
&\quad + \left(\frac{\partial \mathbf{F}^p}{\partial \theta_k^p} - \frac{\partial \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_k^p}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c\right)(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c + \mathbf{R}^{-1})^{-1}(\mathbf{F}^p - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c)^{\mathrm{T}} \\
&\quad + (\mathbf{F}^p - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c)(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c + \mathbf{R}^{-1})^{-1}\left(\frac{\partial \mathbf{F}^p}{\partial \theta_k^p} - \frac{\partial \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_k^p}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c\right)^{\mathrm{T}}, \\
\frac{\partial \log[2b_\star]}{\partial \theta_k^p} &= 0, \\
\frac{\partial b_\star}{\partial \theta_k^p} &= 0, \\
\frac{\partial \boldsymbol{\Sigma}_{pp,ii'}}{\partial \theta_k^p} &= 2\phi_{\theta_k}\boldsymbol{\Sigma}_{pp,ii'}(\theta_k^p - \theta_k^p) = 0, \quad i, i' = 1, \dots, n, \\
\frac{\partial \boldsymbol{\Sigma}_{cp,ji}}{\partial \theta_k^p} &= 2\phi_{\theta_k}\boldsymbol{\Sigma}_{cp,ji}(\theta_{jk}^c - \theta_k^p), \quad i = 1, \dots, n, \quad j = 1, \dots, m.
\end{aligned}$$

Then we find the derivative of the variance $\tilde{\boldsymbol{\Sigma}}_y$ with respect to the nugget τ^2. The mean $\tilde{\boldsymbol{\mu}}_y$ does not depend on the nugget. We have:

$$\begin{aligned}
\frac{\partial \tilde{\boldsymbol{\Sigma}}_y^{-1}}{\partial \tau^2} &= -\tilde{\boldsymbol{\Sigma}}_y^{-1}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \tau^2}\tilde{\boldsymbol{\Sigma}}_y^{-1}, \\
\frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \tau^2} &= \mathbf{I}_n.
\end{aligned}$$

For $h, s = 1, \dots, q_2$, the second derivatives are given by:

$$\frac{\partial^2 \log \pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\psi}, \xi)}{\partial \psi_h' \partial \psi_s'} = \frac{\partial}{\partial \psi_s'}\left[\frac{\partial \log \pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\psi}, \xi)}{\partial \psi_h'}\right]$$

$$\begin{aligned}
&= \frac{\partial \psi_s}{\partial \psi'_s}\frac{\partial}{\partial \psi_s}\left[\frac{\partial \log \pi(\mathbf{y}|\mathbf{z},\boldsymbol{\psi},\xi)}{\partial \psi'_h}\right] \\
&= \frac{\partial \psi_s}{\partial \psi'_s}\frac{\partial}{\partial \psi_s}\left[\frac{\partial \psi_h}{\partial \psi'_h}\frac{\partial \log \pi(\mathbf{y}|\mathbf{z},\boldsymbol{\psi},\xi)}{\partial \psi_h}\right] \\
&= \frac{\partial \psi_s}{\partial \psi'_s}\left[\frac{\partial^2 \psi_h}{\partial \psi_s \partial \psi'_h}\frac{\partial \log \pi(\mathbf{y}|\mathbf{z},\boldsymbol{\psi},\xi)}{\partial \psi_h} + \frac{\partial \psi_h}{\partial \psi'_h}\frac{\partial^2 \log \pi(\mathbf{y}|\mathbf{z},\boldsymbol{\psi},\xi)}{\partial \psi_s \partial \psi_h}\right] \\
&= \frac{\partial \psi_s}{\partial \psi'_s}\left[\frac{\partial \psi_h}{\partial \psi'_h}\frac{\partial^2 \log \pi(\mathbf{y}|\mathbf{z},\boldsymbol{\psi},\xi)}{\partial \psi_s \partial \psi_h}\right].
\end{aligned}$$

Here,

$$\frac{\partial^2 \log \pi(\mathbf{y}|\mathbf{z},\boldsymbol{\psi},\xi)}{\partial \psi_s \partial \psi_h} = -\frac{1}{2}\frac{\partial^2 \log|\tilde{\boldsymbol{\Sigma}}_y|}{\partial \psi_s \partial \psi_h} + a_\star \frac{\partial^2 \log[2b_\star]}{\partial \psi_s \partial \psi_h} - \left(a_\star + \frac{n}{2}\right)\left(\frac{\frac{\partial A}{\partial \psi_s}B - A\frac{\partial B}{\partial \psi_s}}{B^2}\right),$$

where

$$\begin{aligned}
A &= 2\frac{\partial b_\star}{\partial \psi_h} - 2\left(\frac{\partial \tilde{\boldsymbol{\mu}}_y}{\partial \psi_h}\right)^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_y^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y) + (\mathbf{y}-\tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y^{-1}}{\partial \psi_h}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y) \\
\frac{\partial A}{\partial \psi_s} &= 2\frac{\partial^2 b_\star}{\partial \psi_s \partial \psi_h} - 2\left(\frac{\partial^2 \tilde{\boldsymbol{\mu}}_y}{\partial \psi_s \partial \psi_h}\right)^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_y^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y) - 2\left(\frac{\partial \tilde{\boldsymbol{\mu}}_y}{\partial \psi_h}\right)^{\mathrm{T}}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y^{-1}}{\partial \psi_s}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y) \\
&\quad + 2\left(\frac{\partial \tilde{\boldsymbol{\mu}}_y}{\partial \psi_h}\right)^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_y^{-1}\left(\frac{\partial \tilde{\boldsymbol{\mu}}_y}{\partial \psi_s}\right) - 2\left(\frac{\partial \tilde{\boldsymbol{\mu}}_y}{\partial \psi_s}\right)^{\mathrm{T}}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y^{-1}}{\partial \psi_h}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y) \\
&\quad + (\mathbf{y}-\tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}}\frac{\partial^2 \tilde{\boldsymbol{\Sigma}}_y^{-1}}{\partial \psi_s \partial \psi_h}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y) \\
B &= 2b_\star + (\mathbf{y}-\tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_y^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y) \\
\frac{\partial B}{\partial \psi_s} &= 2\frac{\partial b_\star}{\partial \psi_s} - 2\left(\frac{\partial \tilde{\boldsymbol{\mu}}_y}{\partial \psi_s}\right)^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_y^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y) + (\mathbf{y}-\tilde{\boldsymbol{\mu}}_y)^{\mathrm{T}}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y^{-1}}{\partial \psi_s}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_y),
\end{aligned}$$

and

$$\frac{\partial^2 \log|\tilde{\boldsymbol{\Sigma}}_y|}{\partial \psi_s \partial \psi_h} = \operatorname{tr}\left[-\tilde{\boldsymbol{\Sigma}}_y^{-1}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \psi_s}\tilde{\boldsymbol{\Sigma}}_y^{-1}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \psi_h} + \tilde{\boldsymbol{\Sigma}}_y^{-1}\frac{\partial^2 \tilde{\boldsymbol{\Sigma}}_y}{\partial \psi_s \partial \psi_h}\right]$$

We first find the second derivatives with respect to the true calibration parameters $\boldsymbol{\theta}^p$. For $k, r = 1, \ldots, p_\theta$ we have:

$$\begin{aligned}
\frac{\partial^2 \tilde{\boldsymbol{\mu}}_y}{\partial \theta^p_k \partial \theta^p_r} &= \frac{\partial^2 \mathbf{F}^p}{\partial \theta^p_k \partial \theta^p_r}(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c + \mathbf{R}^{-1})^{-1}(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{z} + \mathbf{R}^{-1}\boldsymbol{\beta}_0) \\
&\quad + \frac{\partial^2 \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta^p_k \partial \theta^p_r}\boldsymbol{\Sigma}_{cc}^{-1}\left[\mathbf{z} - \mathbf{F}^c(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{F}^c + \mathbf{R}^{-1})^{-1}(\mathbf{F}^{c\mathrm{T}}\boldsymbol{\Sigma}_{cc}^{-1}\mathbf{z} + \mathbf{R}^{-1}\boldsymbol{\beta}_0)\right], \\
\frac{\partial^2 \tilde{\boldsymbol{\Sigma}}_y^{-1}}{\partial \theta^p_k \partial \theta^p_r} &= \tilde{\boldsymbol{\Sigma}}_y^{-1}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta^p_k}\tilde{\boldsymbol{\Sigma}}_y^{-1}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta^p_r}\tilde{\boldsymbol{\Sigma}}_y^{-1} - \tilde{\boldsymbol{\Sigma}}_y^{-1}\frac{\partial^2 \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta^p_k \partial \theta^p_r}\tilde{\boldsymbol{\Sigma}}_y^{-1} + \tilde{\boldsymbol{\Sigma}}_y^{-1}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta^p_r}\tilde{\boldsymbol{\Sigma}}_y^{-1}\frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta^p_k}\tilde{\boldsymbol{\Sigma}}_y^{-1},
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta_k^p \partial \theta_r^p} &= \frac{\partial^2 \boldsymbol{\Sigma}_{pp}}{\partial \theta_k^p \partial \theta_r^p} - \frac{\partial^2 \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_k^p \partial \theta_r^p} \boldsymbol{\Sigma}_{cc}^{-1} \boldsymbol{\Sigma}_{cp} - \frac{\partial \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_k^p} \boldsymbol{\Sigma}_{cc}^{-1} \frac{\partial \boldsymbol{\Sigma}_{cp}}{\partial \theta_r^p} \\
&\quad - \frac{\partial \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_r^p} \boldsymbol{\Sigma}_{cc}^{-1} \frac{\partial \boldsymbol{\Sigma}_{cp}}{\partial \theta_k^p} - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \frac{\partial^2 \boldsymbol{\Sigma}_{cp}}{\partial \theta_r^p \partial \theta_k^p} \\
&\quad + \left(\frac{\partial^2 \mathbf{F}^p}{\partial \theta_k^p \partial \theta_r^p} - \frac{\partial^2 \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_k^p \partial \theta_r^p} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c \right) (\mathbf{F}^{c\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c + \mathbf{R}^{-1})^{-1} (\mathbf{F}^p - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c)^{\mathrm{T}} \\
&\quad + \left(\frac{\partial \mathbf{F}^p}{\partial \theta_k^p} - \frac{\partial \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_k^p} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c \right) (\mathbf{F}^{c\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c + \mathbf{R}^{-1})^{-1} \left(\frac{\partial \mathbf{F}^p}{\partial \theta_r^p} - \frac{\partial \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_r^p} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c \right)^{\mathrm{T}} \\
&\quad + \left(\frac{\partial \mathbf{F}^p}{\partial \theta_r^p} - \frac{\partial \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_r^p} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c \right) (\mathbf{F}^{c\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c + \mathbf{R}^{-1})^{-1} \left(\frac{\partial \mathbf{F}^p}{\partial \theta_k^p} - \frac{\partial \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_k^p} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c \right)^{\mathrm{T}} \\
&\quad + (\mathbf{F}^p - \boldsymbol{\Sigma}_{cp}^{\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c) (\mathbf{F}^{c\mathrm{T}} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c + \mathbf{R}^{-1})^{-1} \left(\frac{\partial^2 \mathbf{F}^p}{\partial \theta_k^p \partial \theta_r^p} - \frac{\partial^2 \boldsymbol{\Sigma}_{cp}^{\mathrm{T}}}{\partial \theta_k^p \partial \theta_r^p} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{F}^c \right)^{\mathrm{T}}, \\
\frac{\partial^2 \boldsymbol{\Sigma}_{pp,ii'}}{\partial \theta_k^p \partial \theta_r^p} &= 0, \\
\frac{\partial^2 \boldsymbol{\Sigma}_{cp,ji}}{\partial \theta_k^p \partial \theta_r^p} &= 4 \phi_{\theta_k} \phi_{\theta_r} \boldsymbol{\Sigma}_{cp,ji} (\theta_{jk}^c - \theta_k^p)(\theta_{jr}^c - \theta_r^p) - 2 \phi_{\theta_k} \boldsymbol{\Sigma}_{cp,ji} \frac{\partial \theta_k^p}{\partial \theta_r^p} \\
&= 4 \phi_{\theta_k} \phi_{\theta_r} \boldsymbol{\Sigma}_{cp,ji} (\theta_{jk}^c - \theta_k^p)(\theta_{jr}^c - \theta_r^p) - 2 \phi_{\theta_k} \boldsymbol{\Sigma}_{cp,ji} \delta_{kr},
\end{aligned}
$$

where

$$
\delta_{kr} = \begin{cases} 1, & \text{if } k = r \\ 0, & \text{otherwise}, \end{cases} \tag{6.18}
$$

is the Kronecker delta and $i, i' = 1, \ldots, n$ and $j = 1, \ldots, m$.

The second derivatives of the variance $\tilde{\boldsymbol{\Sigma}}_y$ involving the nugget τ^2 is,

$$
\begin{aligned}
\frac{\partial^2 \tilde{\boldsymbol{\Sigma}}_y^{-1}}{\partial \theta_k^p \partial \tau^2} &= \tilde{\boldsymbol{\Sigma}}_y^{-1} \frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta_k^p} \tilde{\boldsymbol{\Sigma}}_y^{-1} \frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \tau^2} \tilde{\boldsymbol{\Sigma}}_y^{-1} - \tilde{\boldsymbol{\Sigma}}_y^{-1} \frac{\partial^2 \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta_k^p \partial \tau^2} \tilde{\boldsymbol{\Sigma}}_y^{-1} + \tilde{\boldsymbol{\Sigma}}_y^{-1} \frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \tau^2} \tilde{\boldsymbol{\Sigma}}_y^{-1} \frac{\partial \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta_k^p} \tilde{\boldsymbol{\Sigma}}_y^{-1}, \\
\frac{\partial^2 \tilde{\boldsymbol{\Sigma}}_y}{\partial \theta_k^p \partial \tau^2} &= \mathbf{0}_{n \times n}, \\
\frac{\partial^2 \tilde{\boldsymbol{\Sigma}}_y}{\partial [\tau^2]^2} &= \mathbf{0}_{n \times n}, \\
\frac{\partial^2 \tilde{\boldsymbol{\mu}}_y}{\partial \theta_k^p \partial \tau^2} &= \mathbf{0}_n.
\end{aligned}
$$

Finally, the second derivatives of $b_\star$ with respect to ψ_s:

$$
\frac{\partial f(b_\star)}{\partial \psi_s} = 0,
$$

for any function f and any $s = 1, \ldots, q_2$.

In this section we have presented the derivatives of $\log \pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}^p, \tau^2, \xi)$. The derivatives of $\log \pi_b(\boldsymbol{\psi})$ can be found in Appendix B.2.5 for the example in Section 6.4.3 and in Appendix B.2.6 for the example in Section 6.4.4.

In Section 6.4.3 we present an example for calibration problem SP3 of Section 6.1. We compare the ESIG for a fixed physical design as we increase the number of runs of the computer experiment, m. Lastly, we compare optimal designs for this model with optimal designs found for a corresponding nonlinear model.

6.4.3 Example 1: Unknown simulator and $\delta_{\theta^p}(x) = 0$ - Cantilever beam function

We estimate the expected Shannon information gain with ALIS and LIS approximations (Section 4.3) for the statistical model (6.4) and assume a Gaussian process prior for the simulator as given in Equation (6.5). We combine ALIS and LIS with the ACE algorithm (Section 5.2) and find Bayesian optimal designs.

For the purpose of this example, we generate simulator runs using the model of cantilever beam displacement (in inches) (Wu et al., 2001) given by:

$$\eta(\mathbf{x}, \theta) = \frac{4L^3}{\theta w t} \sqrt{\left(\frac{x_1}{t^2}\right)^2 + \left(\frac{x_2}{w^2}\right)^2}. \tag{6.19}$$

For the purposes of finding Bayesian optimal designs, $\eta(\mathbf{x}, \theta)$ is treated as an unknown expensive simulator.

Equation (6.19) is used to model a simple uniform cantilever beam with horizontal and vertical loads as shown in Figure 6.10. The beam length L is a constant with value $L = 100$ inches, w is the width of the cross-section with value $w = 4$ inches and t is thickness of the cross-section with value $t = 2$ inches. The controllable variables x_1 and x_2 are the vertical and horizontal load (Newtons), respectively. Both x_1 and x_2 take values in the range $[-2000, 2000]$. The negative values of the force here imply load in the opposite direction from the one given in Figure 6.10. The calibration parameter, θ^p, is Young's modulus of the beam material for which a normal prior distribution, $\theta^p \sim N[2.9 \times 10^7, (1.45 \times 10^6)^2]$, is assumed as given by Surjanovic and Bingham (2017). Figure 6.11 shows ellipsoidal contours of the simulator for a given value of the calibration parameter.

We also assume that during calibration, the correlation parameters $\boldsymbol{\phi} = (\phi_{x_1}, \phi_{x_2}, \phi_\theta)^{\mathrm{T}}$ for the emulator are held fixed at the maximum likelihood estimates from the computer experiment. In other words, the correlation parameters will not be updated with the physical experiment data. However, the prior distributions for the parameters θ^p and $\tau^2 = \sigma_\varepsilon^2/\sigma^2$ will be updated following the physical experiment. We maximise the expected Shannon information gain for θ^p and treat τ^2 as a nuisance parameter. We integrate out $\boldsymbol{\beta} \mid \mathbf{z}, \sigma^2, \boldsymbol{\phi}, \tau^2$ and $\sigma^2 \mid \mathbf{z}, \boldsymbol{\phi}, \tau^2$ using their marginal posterior distributions,

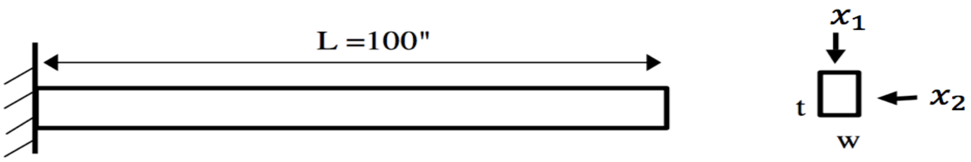


Figure 6.10: A beam under vertical and horizontal loads (taken from Wu et al., 2001)

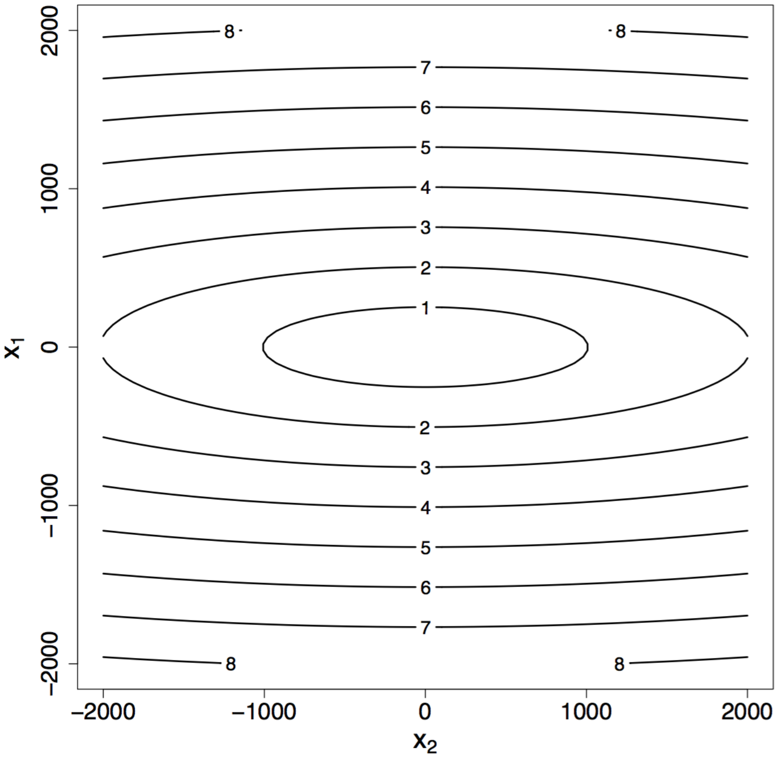


Figure 6.11: Contour plot of the cantilever beam function for $\theta^p = 3.15 \times 10^7$

as shown in Section 6.4.1, to obtain

$$\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}^p, \tau^2 \sim t_{2a_\star}\left[n, \tilde{\boldsymbol{\mu}}_y, \frac{b_\star}{a_\star}\tilde{\boldsymbol{\Sigma}}_y\right],$$

as given in Equation (6.16), where $\tilde{\boldsymbol{\mu}}_y$, $\tilde{\boldsymbol{\Sigma}}_y$ are given by Equations (6.14) and (6.15) respectively, and $a_\star$, $b_\star$ are given by Equation (6.7).

For the design of the computer experiment, $\xi^c = [(\mathbf{x}_1^c, \theta_1^c), \dots, (\mathbf{x}_m^c, \theta_m^c)]$, we use three-dimensional LHS designs scaled to the range $[-2000, 2000] \times [-2000, 2000] \times [2.5 \times 10^7, 3.2 \times 10^8]$ with the range of θ^p chosen as the 1% and 99% prior quantiles of the prior distribution. We transform the variables according to:

$$\theta^{p\prime} = \frac{\theta^p - 2.9 \times 10^7}{1.9 \times 10^6}, \quad x_1' = \frac{x_1}{1173.8} \text{ and } x_2' = \frac{x_2}{1173.8},$$

so each column of the transformed design matrix has zero mean and standard deviation one. Hence $\theta^{p\prime} \sim N(\mu_1, \sigma_1^2)$, is the prior distribution for the transformed calibration parameter $\theta^{p\prime}$ with $\mu_1 = 0$ and $\sigma_1 = 0.73$. The new range of the transformed simulator designs $\xi^{c\prime} = [(\mathbf{x}_1^{c\prime}, \theta_1^{c\prime}), \dots, (\mathbf{x}_m^{c\prime}, \theta_m^{c\prime})]$ is $[-1.7, 1.7] \times [-1.7, 1.7] \times [-1.65, 1.65]$.

We assume the product of one-dimensional correlation functions (2.4) where each one-dimensional correlation function is the squared exponential correlation function (2.6). We also assume prior distributions for the trend parameter $\boldsymbol{\beta}$, the variance of the Gaussian process σ^2 and the nugget τ^2 as given below:

$$\boldsymbol{\beta} \sim N(\mathbf{0}_1, \sigma^2\mathbf{I}_1), \quad \sigma^2 \sim \mathrm{IG}(3, 2), \quad \tau^2 \sim \mathrm{Exp}(20).$$

The choice of prior distribution of τ^2 implies that the 10% and 90% quantiles of the noise-to-signal ratio (σ_ε divided by the maximum expected response $\eta(2000, 2000, \theta)$) are 0.007 and 0.04, respectively. The prior distributions of σ^2 and τ^2 imply that the 10% and 90% quantiles of the error variance σ_ε^2 , are 0.003 and 0.112, respectively.

We assume that the regression trend function for the Gaussian process is $\mathbf{f}(\mathbf{x}) = 1$, so $\mathbf{F}^c = \mathbf{1}_m$ and $\mathbf{F}^p = \mathbf{1}_n$. We denote by ξ the physical design for which the Shannon information gain is estimated. The design ξ has points in the range $[-1.7, 1.7] \times [-1.7, 1.7]$.

In Appendix C.3 a detailed investigation is presented into the change in the Gaussian process fit to the cantilever beam function as the number of simulator runs, m, is changed ($m = 30, 60, 90$). To summarise the results, the GP models fit well for all values of m considered, with the fit of the posterior mean to the simulator response obviously improving as m increases. This improvement is most obvious near the edges of the design region. The posterior variance decreases as m increases.

We assume the expected Shannon information gain utility function (4.5). We approximate the evidence in the ESIG using ALIS and LIS approximations (Section 4.3) which are then combined with the ACE algorithm (Section 5.2) to find Bayesian optimal

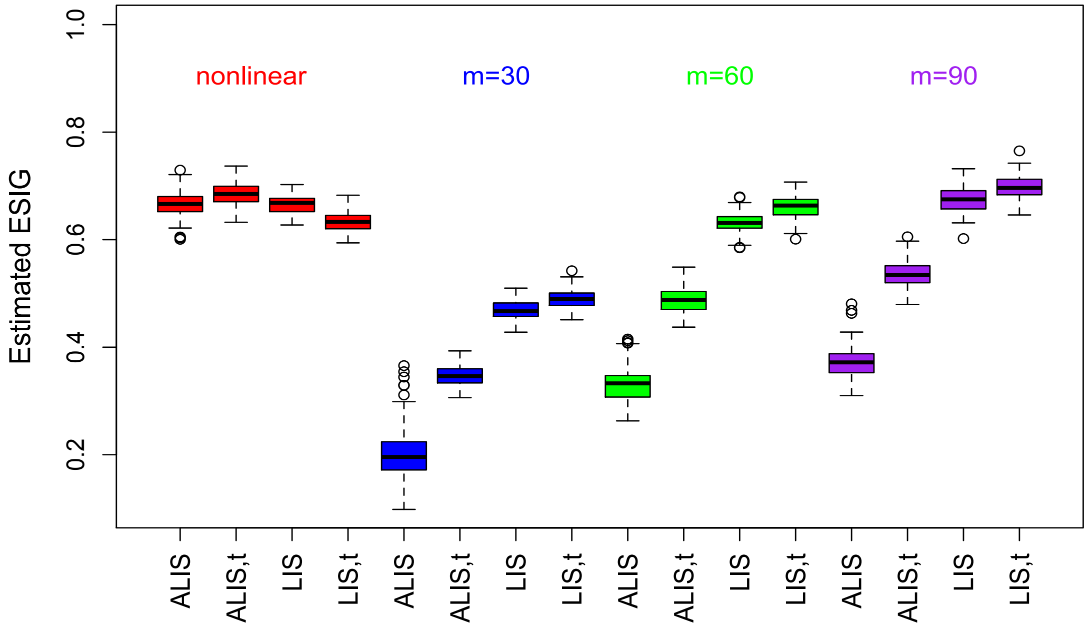


Figure 6.12: Estimated ESIG for the parameter θ of the nonlinear model (cantilever beam example) when treating σ_ε^2 as a nuisance parameter (red) and the parameter of interest θ^p of the calibration model when τ^2 is treated as a nuisance parameter, as we increase the number of simulator runs, m, and found using ALIS and LIS

designs.

The log-unnormalised marginal posterior density is given by:

$$\begin{aligned}\log \pi_u(\theta^{p\prime}, \tau^2 | \mathbf{y}, \mathbf{z}, \phi, \xi) &= \log \pi(\mathbf{y}|\mathbf{z}, \theta^{p\prime}, \boldsymbol{\phi}, \tau^2, \xi) + \log \pi_b(\theta^{p\prime}) + \log \pi_b(\tau^2) \\ &= \log \pi(\mathbf{y}|\mathbf{z}, \theta^{p\prime}, \boldsymbol{\phi}, \tau^2, \xi) - \frac{1}{2}\log\left(2\pi\sigma_1^2\right) - \frac{(\theta^{p\prime} - \mu_1)^2}{2\sigma_1^2} \\ &\quad + \log \lambda_{\tau^2} - \lambda_{\tau^2}\tau^2,\end{aligned}$$

and $\log \pi(\mathbf{y}|\mathbf{z}, \theta^{p\prime}, \boldsymbol{\phi}, \tau^2, \xi)$ is given by Equation (6.17).

For this example, we aim to construct an importance distribution that guarantees that positive values for the nugget, τ^2, will be sampled. Hence, we take a normal approximation to the posterior distribution of $\boldsymbol{\psi}' = (\theta^{p\prime}, \log \tau^2)^{\mathrm{T}}$ as described in Section 4.3.2. In order to calculate the negative Hessian of the log-unnormalised posterior density, $\mathbf{H}_{\boldsymbol{\psi}'}(\boldsymbol{\psi}')$, in ALIS and LIS (see Section 4.3), we first have to find the derivatives of the log-unnormalised posterior density $\log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y}, \mathbf{z}, \xi)$ with respect to $\boldsymbol{\psi}'$ using Equations (4.36). The derivatives of the log-marginal predictive density of the physical data, $\log \pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\psi}, \xi)$, can be found in Section 6.4.2 and the derivatives of the log-prior density, $\log \pi_b(\boldsymbol{\psi})$, can be found in Appendix B.2.5 which give the derivatives of $\log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y}, \mathbf{z}, \xi)$ using Equations (4.36). We also treat the nugget τ^2 as a nuisance parameter which we integrate out as described in Section 4.3.1.

In Figure 6.12, we compare the expected Shannon information gain for a space-filling physical design with $n = 10$ runs calculated (i) for the nonlinear model assuming the cantilever beam function (6.19) is known and can be evaluated, and the response has normally distributed errors (see Appendix B.2.5); and (ii) the calibration model assuming (6.19) can only be evaluated for a m-run computer experiment. The number of simulator runs is set to $m = 30, 60$ and 90. We present 100 estimates of the expected Shannon information gain for the two models and the different sizes of simulator runs, m. To perform the calculation we use ALIS and LIS with $k_1 = k_2 = 2000$ for the nonlinear model (σ_ε^2 is integrated-out analytically, see Appendix B.2.5) and $k_1 = k_2 = k_3 = 2000$ for the calibration model, and both normal and t importance distributions.

As the number of simulator runs is increased, the approximate ESIG also increases. For $m = 60$ and $m = 90$ the computer experiment produces a good approximation to the simulator. Hence, the approximate ESIG based on these computer experiments is roughly equal to the ESIG for θ under the nonlinear regression model.

Figure 6.12 shows differences between ALIS and LIS for the same number of simulator runs, m. These differences arise as ALIS sometimes "fails" because of poorly conditioned matrices $\mathbf{H}_{\psi'}(\psi')$ that are still just positive-definite. Hence ALIS does not enter the optimisation step (see Section 4.3). However, the ill-conditioning results in very large variances, and so very large values of some sampled parameters, for which the likelihood evaluation cannot be performed. The results in Figure 6.12 are conditional on the likelihood evaluation being possible, which results in the ALIS estimator having negative bias. For the results presented in the rest of this example we will use LIS.

Next we find Bayesian optimal designs by combining LIS with the ACE algorithm. Similar to the previous examples we use 10 different random starts in ACE. We present optimal designs for the physical experiment with $n = 10$ for both the nonlinear model when σ_ε^2 is treated as a nuisance parameter and the calibration model when τ^2 is treated as a nuisance parameter with different numbers of simulator runs. Finally, we approximate the ESIG of the optimal designs found with ACE for each model.

Figure 6.13 shows Bayesian optimal designs for the nonlinear cantilever beam function when σ_ε^2 is treated as a nuisance parameter, and the calibration problem when treating τ^2 as a nuisance parameter and for different number of simulator runs ($m = 30, 60, 90$). In Appendix B.2.5 we present more near-optimal designs obtained from different random starts of ACE. We notice that as we increase m, the optimal designs obtained for the calibration model become more similar to the optimal design for the nonlinear model. That is, for the physical design, there are many values of x_1 (vertical load) but x_2 (horizontal load) only takes the extreme values ±1.7. The calibration designs for $m = 60, 90$ also mainly contain values of x_2 close to the edges of the range, but the calibration design for $m = 30$ exhibits a greater spread of values of x_2, including values in the interior of the range. This difference is probably due to the smaller computer experiment producing a poorer approximation to the simulator.

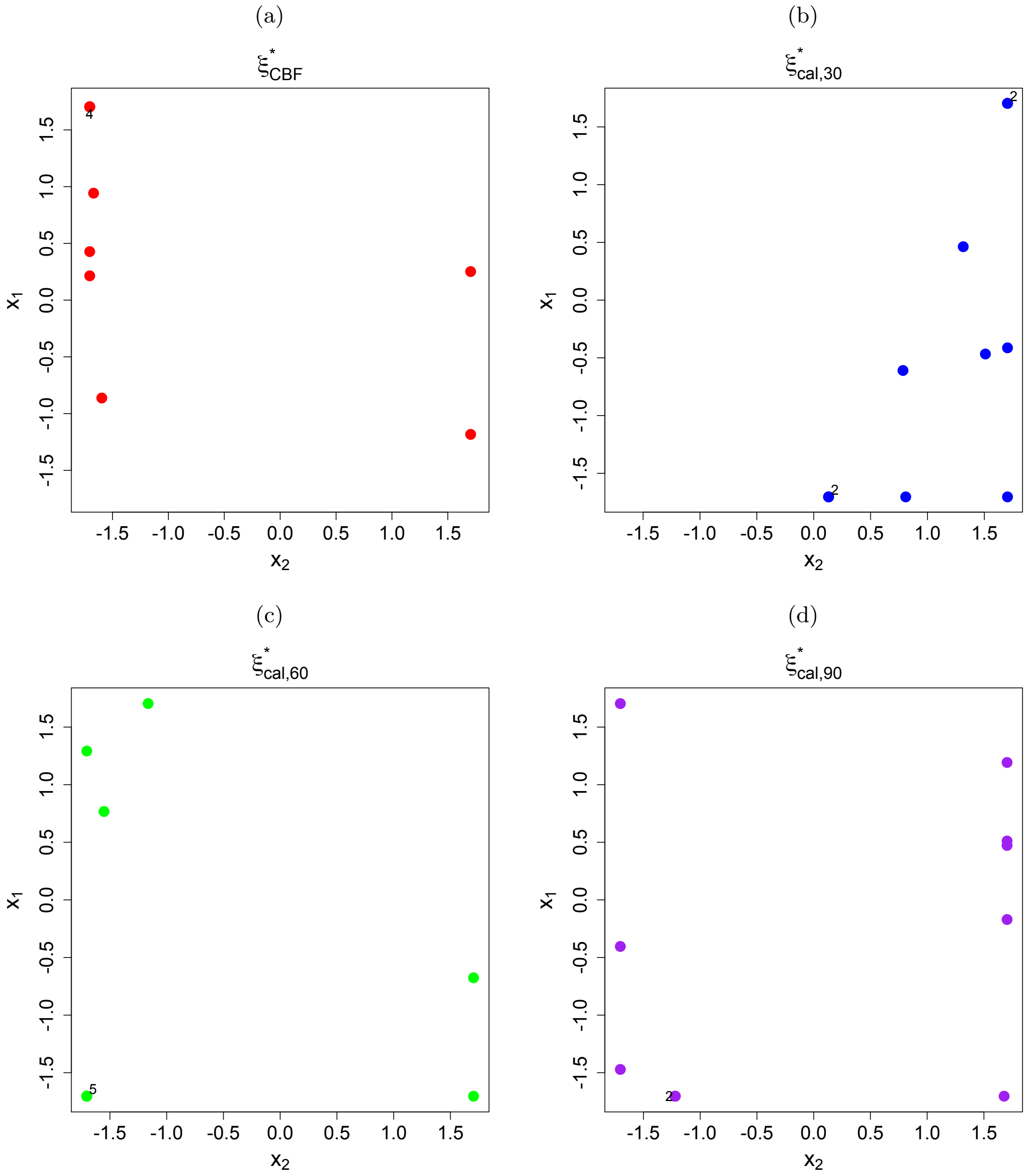


Figure 6.13: Cantilever beam example: (a) Bayesian optimal design for the nonlinear model when treating σ_ε^2 as a nuisance parameter ($\xi^\star_{CBF}$); Bayesian optimal design for the calibration model when treating τ^2 as a nuisance parameter and (b) $m = 30$ ($\xi^\star_{cal,30}$); (c) $m = 60$ ($\xi^\star_{cal,60}$); (d) $m = 90$ ($\xi^\star_{cal,90}$); the number on some points in each plot shows how many times the point is repeated

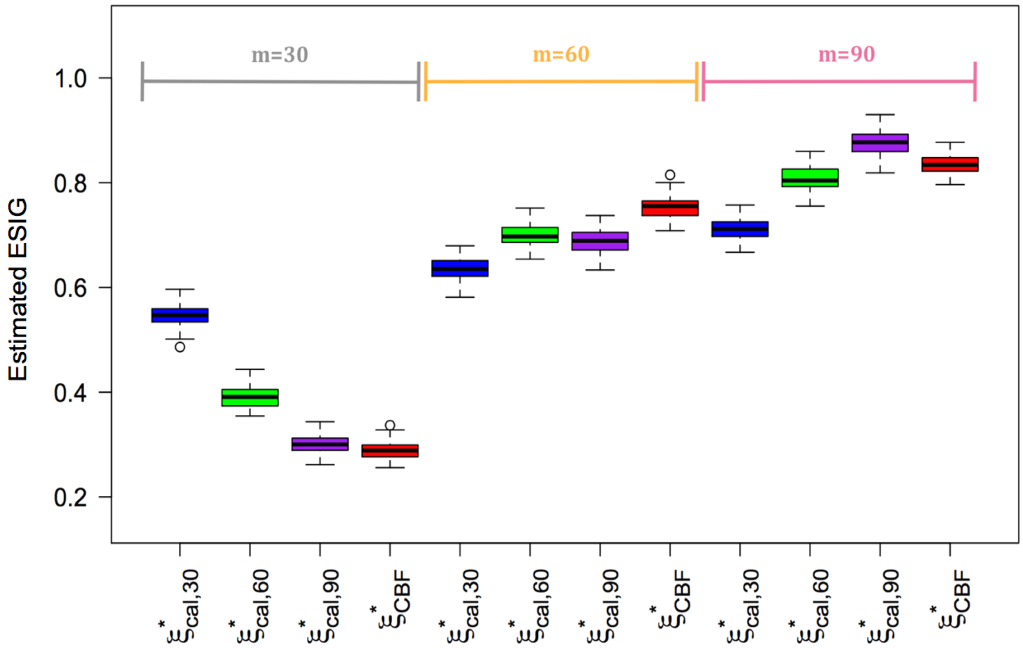


Figure 6.14: Cantilever beam example: estimated ESIG for the parameter of interest $\theta^{p\prime}$ for the calibration model when treating τ^2 as a nuisance parameter and as we increase the number of simulator runs, m, for the optimal designs shown in Figure 6.13, found using LIS with $k_1 = k_2 = k_3 = 2000$

Figure 6.14 shows boxplots for each design presented in Figure 6.13, corresponding to the distribution of 100 estimates of the ESIG for the calibration model when treating τ^2 as a nuisance parameter and as we increase the number of simulator runs ($m = 30, 60, 90$). To perform the calculation, LIS was used with $k_1 = k_2 = k_3 = 2000$. For the calibration model with $m = 30$ simulator runs (boxplots under the grey line) the optimal design found under this model, $\xi^\star_{cal,30}$, has the best performance. For the calibration model with $m = 60$ simulator runs (boxplots under the orange line) all optimal designs have similar performance with design $\xi^\star_{CBF}$ performing slightly better. Finally, for the calibration model with $m = 90$ simulator runs (boxplots under the pink line) the design for the calibration model $\xi^\star_{cal,90}$ has the best performance. This plot provides further evidence that the computer experiment with $m = 30$ has provided an approximation to the simulator that is quite different from that obtained from the larger computer experiments.

In Figure 6.15 we present boxplots for each design presented in Figure 6.13, corresponding to the distribution of 100 estimates of the ESIG for the nonlinear model when treating σ^2_ε as a nuisance parameter. Again to perform the calculation, LIS was used with $k_1 = k_2 = 2000$ (σ^2_ε is integrated-out analytically). The optimal designs found for the calibration model when treating τ^2 as nuisance parameter with $m = 60$ and $m = 90$ simulator runs have very similar performance to the optimal design found for the nonlinear model.

We have shown (Appendix C.3) that the Gaussian process posterior mean adapts very

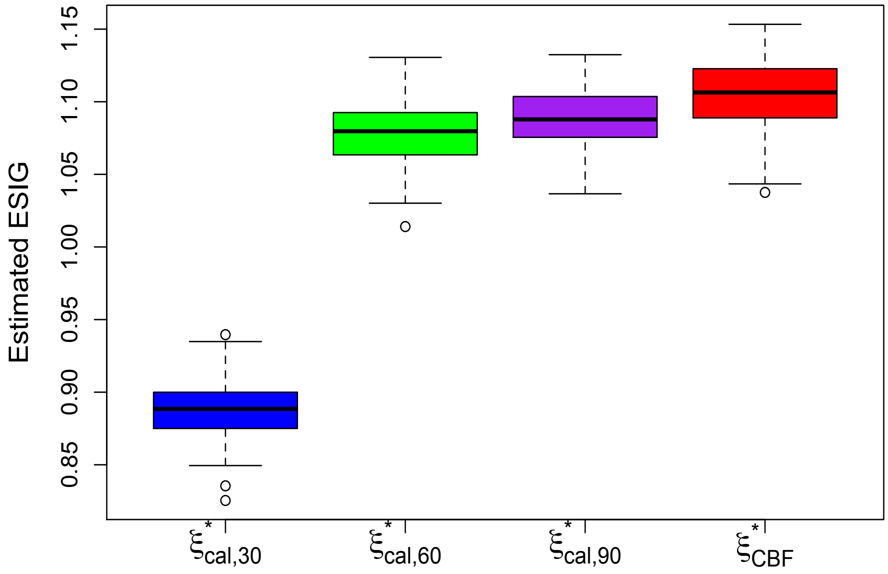


Figure 6.15: Cantilever beam example: estimated ESIG for the nonlinear model when treating σ^2_ε as a nuisance parameter for the optimal designs shown in Figure 6.13, found using LIS with $k_1 = k_2 = 2000$

quickly to the true model as we increase the number of simulator runs m and the Gaussian process posterior variance decreases. As a result, optimal designs found using the calibration model with large m are reasonably efficient for estimating the parameters of the nonlinear model. Similarly, optimal designs under the nonlinear model are reasonably efficient for estimating the parameters of the calibration model.

In this example, because the simulator has a simple equation we are able to compare optimal designs found for the calibration model with unknown simulator with optimal designs for a nonlinear model. However, in general this will not be the case. This example shows that efficient designs can be obtained even in the unknown simulator case.

6.4.4 Example 2: Unknown simulator and $\delta_{\theta^p}(x) = 0$ - Michaelis-Menten model

Similarly to the previous example, we estimate the expected Shannon information gain with ALIS and LIS approximations (Section 4.3) for the statistical model (6.4) and assume a Gaussian process prior for the simulator as given in Equation (6.5). We combine ALIS and LIS with the ACE algorithm (Section 5.2) to find Bayesian optimal designs.

For the purposes of this example, we generate simulator runs using the Michaelis-

Menten model (3.1),

$$\eta(x, \boldsymbol{\theta}) = \frac{\theta_1 x}{\theta_2 + x}.$$

Again $\eta(x, \boldsymbol{\theta})$ is treated as an unknown expensive simulator for the purposes of finding Bayesian optimal calibration designs. We assume that during calibration, the correlation parameters $\boldsymbol{\phi} = (\phi_x, \phi_{\theta_1}, \phi_{\theta_2})^{\mathrm{T}}$ are held fixed at the maximum likelihood estimates from the computer experiment. In other words, the correlation parameters will not be updated with the physical experiment data. However, the prior distributions for the parameters $\boldsymbol{\theta}^p = (\theta_1^p, \theta_2^p)^{\mathrm{T}}$ and $\tau^2 = \sigma_\varepsilon^2/\sigma^2$ will be updated following the physical experiment. We maximise the expected Shannon information gain for $\boldsymbol{\theta}^p$ and treat τ^2 as a nuisance parameter. Similarly to the previous example, we integrate out $\boldsymbol{\beta} \mid \mathbf{z}, \sigma^2, \boldsymbol{\phi}, \tau^2$ and $\sigma^2 \mid \mathbf{z}, \boldsymbol{\phi}, \tau^2$ using their marginal posterior distributions as shown in Section 6.4.1.

For the design of the computer experiment, $\xi^c = [(x_1^c, \boldsymbol{\theta}_1^c), \ldots, (x_m^c, \boldsymbol{\theta}_m^c)]$, we use three-dimensional LHS designs scaled to the range $[0, 400] \times [68.05, 94.29] \times [0.25, 45.59]$ with the range of θ_1^p and θ_2^p chosen as the 1% and 99% quantiles of their prior distributions. We denote by ξ the physical design for which the Shannon information gain is estimated. The physical design ξ has points in $[0, 400]$. We assume the product of one-dimensional correlation functions (2.4) where each one-dimensional correlation function is the squared exponential correlation function (2.6). We also assume the following prior distributions:

$$\theta_1^p \sim \log N(4.38, 0.07^2), \quad \theta_2^p \sim \log N(1.19, 0.84^2), \quad \boldsymbol{\beta} \sim N(\mathbf{0}_1, \sigma^2 \mathbf{I}_1),$$

$$\sigma^2 \sim \mathrm{IG}(3, 2), \quad \tau^2 \sim \mathrm{Exp}(50).$$

The choice of prior distribution of τ^2 implies that the 10% and 90% quantiles of the noise-to-signal ratio (σ_ε divided by the maximum expected response, $\eta(400, \boldsymbol{\theta})$) are 0.0004 and 0.003, respectively. The prior distributions of σ^2 and τ^2 imply that the 10% and 90% quantiles of the error variance σ_ε^2, are 0.001 and 0.047, respectively.

As in the previous example (Section 6.4.3) we assume that the regression trend function for the Gaussian process is $\mathbf{f}(\mathbf{x}) = 1$, so $\mathbf{F}^c = \mathbf{1}_m$ and $\mathbf{F}^p = \mathbf{1}_n$.

We maximise the expected Shannon information gain utility function (4.5) for the parameters $\boldsymbol{\psi} = [(\boldsymbol{\theta}^p)^{\mathrm{T}}, \tau^2]^{\mathrm{T}}$, and approximate the evidence in this utility using ALIS and LIS approximations (see Section 4.3). The log-unnormalised posterior density is given by:

$$\begin{aligned}
\log \pi_u(\boldsymbol{\theta}^p, \tau^2 | \mathbf{y}, \mathbf{z}, \xi) &= \log \pi(\mathbf{y} | \mathbf{z}, \boldsymbol{\theta}^p, \boldsymbol{\phi}, \tau^2, \xi) + \log \pi_b(\boldsymbol{\theta}^p) + \log \pi_b(\tau^2) \\
&= \log \pi(\mathbf{y} | \mathbf{z}, \boldsymbol{\theta}^p, \boldsymbol{\phi}, \tau^2, \xi) - \log \left[\theta_1^p \sigma_1 (2\pi)^{1/2}\right] - \frac{(\log \theta_1^p - \mu_1)^2}{2\sigma_1^2} \\
&\quad - \log \left[\theta_2^p \sigma_2 (2\pi)^{1/2}\right] - \frac{(\log \theta_2^p - \mu_2)^2}{2\sigma_2^2} + \log \lambda_{\tau^2} - \lambda_{\tau^2} \tau^2,
\end{aligned}$$

where $\log \pi(\mathbf{y} | \mathbf{z}, \boldsymbol{\theta}^p, \boldsymbol{\phi}, \tau^2, \xi)$ is given by Equation (6.17).

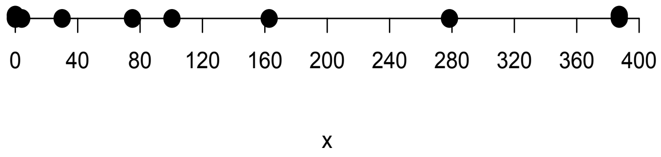


Figure 6.16: ESIG optimal design with $n = 10$, found for the Michaelis-Menten calibration model with $\delta_{\boldsymbol{\theta}^p}(x) \neq 0$, using LIS with $k_1 = k_2 = 300$; two of the points are repeated twice (also given in Figure 6.2 as $\xi^{\star}_{cal}$)

For this example, we aim to construct an importance distribution that guarantees positive values of all parameters $\boldsymbol{\psi}$. Hence, we take a normal approximation to the distribution of $\boldsymbol{\psi}' = [\log \theta^p_1, \log \theta^p_2, \log \tau^2]^{\mathrm{T}}$, as described in Section 4.3.2. In order to calculate the negative Hessian of the log-unnormalised posterior density, $\mathbf{H}_{\boldsymbol{\psi}'}(\boldsymbol{\psi}')$, in ALIS and LIS (Section 4.3), we first have to find the derivatives of the log-unnormalised posterior density $\log \pi^{\psi'}_u(\boldsymbol{\psi}'|\mathbf{y}, \mathbf{z}, \xi)$ with respect to $\boldsymbol{\psi}'$ using Equations (4.36). The derivatives of $\log \pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\psi}, \xi)$ and $\log \pi_b(\boldsymbol{\psi})$ can be found in Section 6.4.2 and Appendix B.2.6 respectively, which give the derivatives of $\log \pi^{\psi'}_u(\boldsymbol{\psi}'|\mathbf{y}, \mathbf{z}, \xi)$ using Equations (4.36).

In Figure 6.17 we compare the distribution of 100 independent estimates of the ESIG for the design shown in Figure 6.16 with $n = 10$ runs, as we increase the number of simulator runs, $m = 30, 50, 60, 90, 150$. We treat τ^2 as a nuisance parameter (see Section 4.3.1). To perform the calculation, ALIS and LIS were used with $k_1 = k_2 = k_3 = 300$ and both normal and t importance distributions.

As we increase the number of simulator runs, m, we expect that the ESIG will also increase as we become more certain about the simulator, so that it eventually becomes essentially known. However, Figure 6.17 shows exactly the opposite. This happens because the marginal posterior distribution of the Gaussian process variance, σ^2, from the computer experiment, has increasing mean and variance ($a_\star$ and $b_\star$ depend on the simulator design, and both increase with m), the prior distribution on τ^2 is fixed, and hence the implied prior distribution on the error variance, $\sigma^2_\varepsilon = \sigma^2 \times \tau^2$ also has increasing mean and variance (see Figure 6.18). This increasing "size" of the error variance leads to lower ESIG. This is a consequence of the model not being stationary in x and $\boldsymbol{\theta}^p$. These features result in a more diffuse distribution for $\mathbf{y}$ for larger m and hence sampling of more extreme values.

Figure 6.19 shows a sample from the prior distributions of θ^p_1 and θ^p_2 and the simulator designs as we increase m. As we increase the number of simulator runs, these designs cover a slightly wider region; however points sampled from the extremes of the prior distributions will never be included in the design. This causes problems for Gaussian process predictions for the unknown simulator $\eta(x, \boldsymbol{\theta})$ near these points.

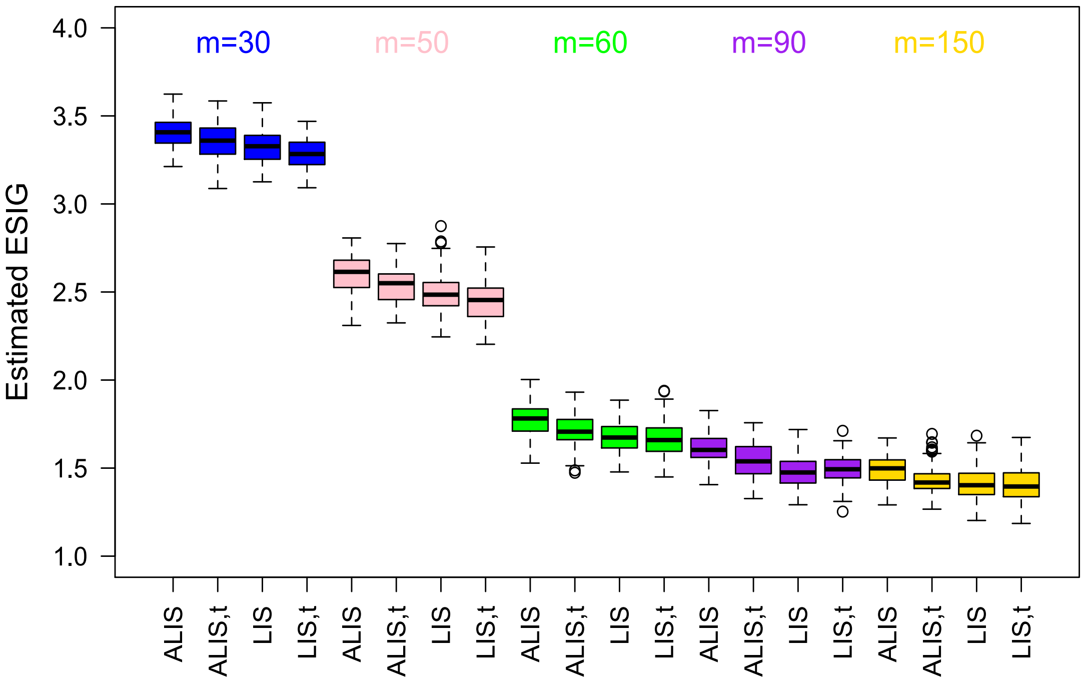


Figure 6.17: Estimated ESIG for the parameters of interest $\boldsymbol{\theta}^p$ of the Michaelis-Menten calibration model when treating τ^2 as a nuisance parameter, using ALIS and LIS with $k_1 = k_2 = k_3 = 300$, as we increase the number of simulator runs, m, for a fixed physical design and fixed priors on the unknown parameters

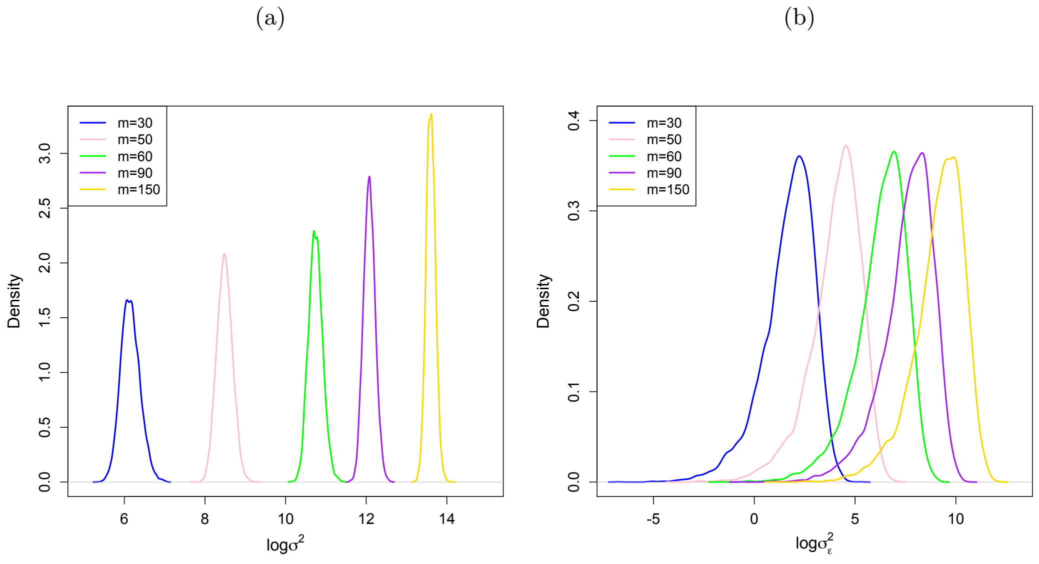


Figure 6.18: (a) The prior density of the log Gaussian process variance, $\log \sigma^2$, and (b) the implied prior distribution of the log error variance, $\log \sigma_\varepsilon^2$ as we increase the number of simulator runs, m, in the computer experiment for the Michaelis-Menten calibration example, and keep the prior of τ^2 fixed

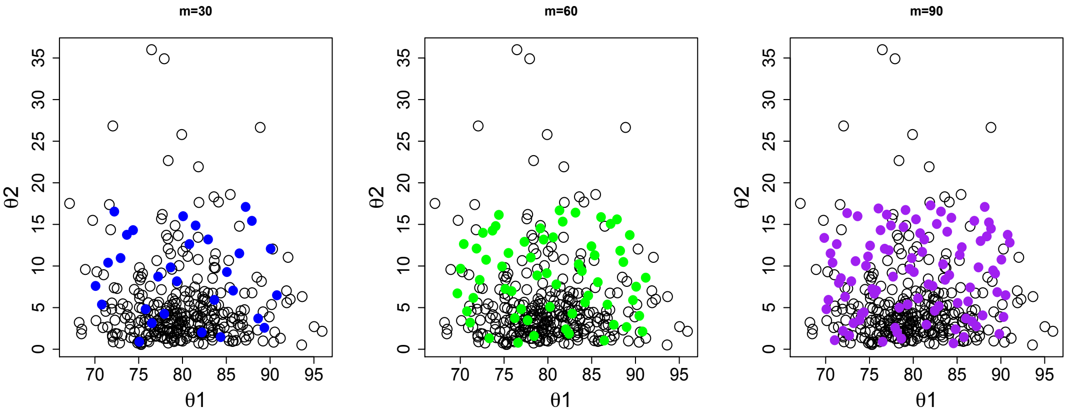


Figure 6.19: A sample from the prior distributions of θ_1^p and θ_2^p and the simulator designs with: $m = 30$ (blue); $m = 60$ (green); $m = 90$ (purple) for the Michaelis-Menten calibration model

In Figures 6.20, 6.21 and 6.22 we show predictions from the Gaussian process fit of simulator runs to the simulator outputs (generated from the Michaelis-Menten model) as we increase the number of simulator runs, $m = 30, 60, 90$. The red line is the mean of the Gaussian process, the blue dashed lines are 95% probability bounds and the black line is the true Michaelis-Menten model for a given $\boldsymbol{\theta}$, as shown for each plot. We present four plots for each value of m (plot (a): $\theta_1 = 90$, $\theta_2 = 12$; plot (b): $\theta_1 = 90$, $\theta_2 = 8.5$; plot (c): $\theta_1 = 90$, $\theta_2 = 30$; plot (d): $\theta_1 = 90$, $\theta_2 = 0.9$). For plots (c) and (d), θ_2 samples extreme points from the prior distribution. As we increase m, the uncertainty in plots (a) and (b) decreases and the mean of the Gaussian process adapts to the shape of the true function. However, for plot (c), the uncertainty is large for $m = 30$, improves a little for $m = 60$, and increases again for $m = 90$. These results are consistent with the increases in σ^2 seen in Figure 6.18. Lastly, for plot (d) as we increase m the variance of the Gaussian process decreases, but the mean does not converge to the true Michaelis-Menten model, which has quite a different shape compared to the other three plots.

To overcome these issues we match the implied distribution of the error variance of the calibration model to the prior distribution for the error variance for a nonlinear regression model based on the Michaelis-Menten equation. We treat τ^2 of the calibration model and σ_ε^2 of the nonlinear Michaelis-Menten model as nuisance parameters (see Section 4.3.1). Then we match the implied distribution of the error variance, $\sigma_\varepsilon^2 = \sigma^2 \times \tau^2$, of the calibration model, with the error variance, σ_ε^2, of the nonlinear Michaelis-Menten model by changing the prior distribution on τ^2. The values of the hyperparameters of the inverse-gamma distribution of the error variance, σ_ε^2, for the nonlinear Michaelis-Menten model are $a = 2.9$ and $b = 16.9$ and we change the prior distribution of τ^2 to match the implied distribution of σ_ε^2 for the calibration model with the error variance of the nonlinear Michaelis-Menten model: (i) for $m = 30$, $\lambda_{\tau^2} = 50$, (ii) for $m = 60$, $\lambda_{\tau^2} = 3000$ and (iii) for $m = 90$, $\lambda_{\tau^2} = 6000$. We use the same prior

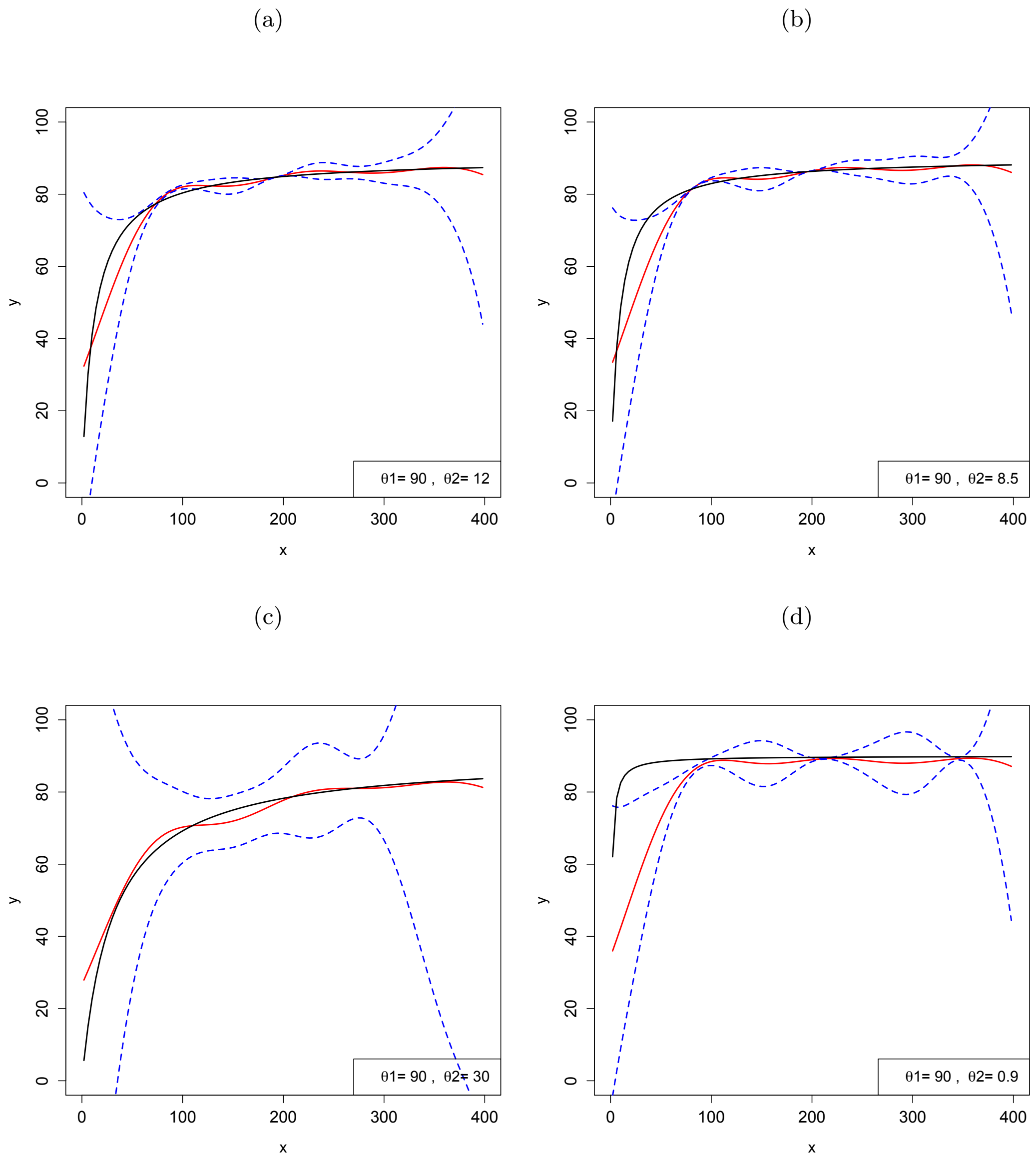


Figure 6.20: Posterior predictive mean for the Gaussian process fit with $m = 30$ simulator runs (red line); 95% probability bounds (blue lines); the true Michaelis-Menten function for a given $\boldsymbol{\theta}$ (black line)

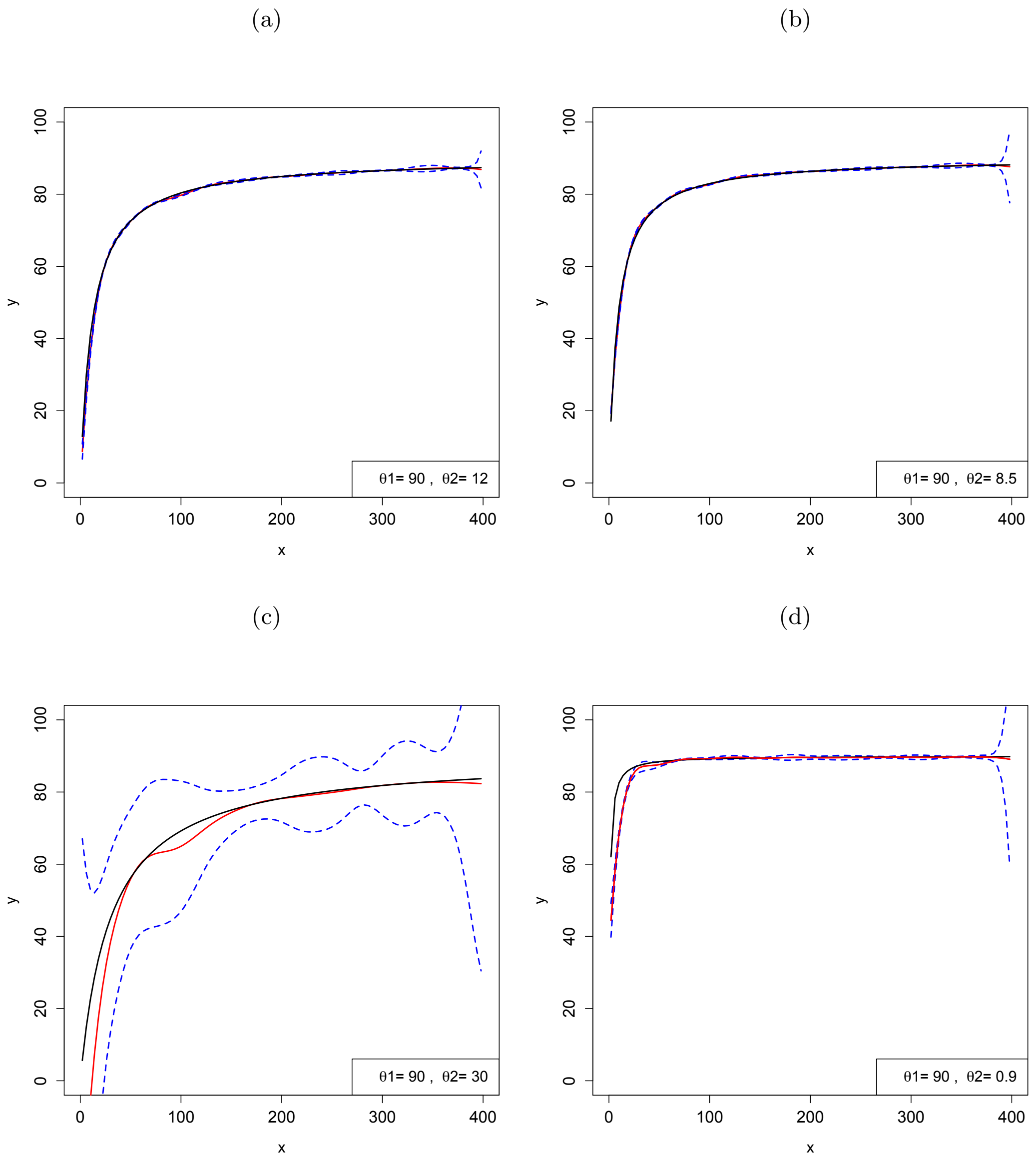


Figure 6.21: Posterior predictive mean for the Gaussian process fit with $m = 60$ simulator runs (red line); 95% probability bounds (blue lines); the true Michaelis-Menten function for a given $\boldsymbol{\theta}$ (black line)

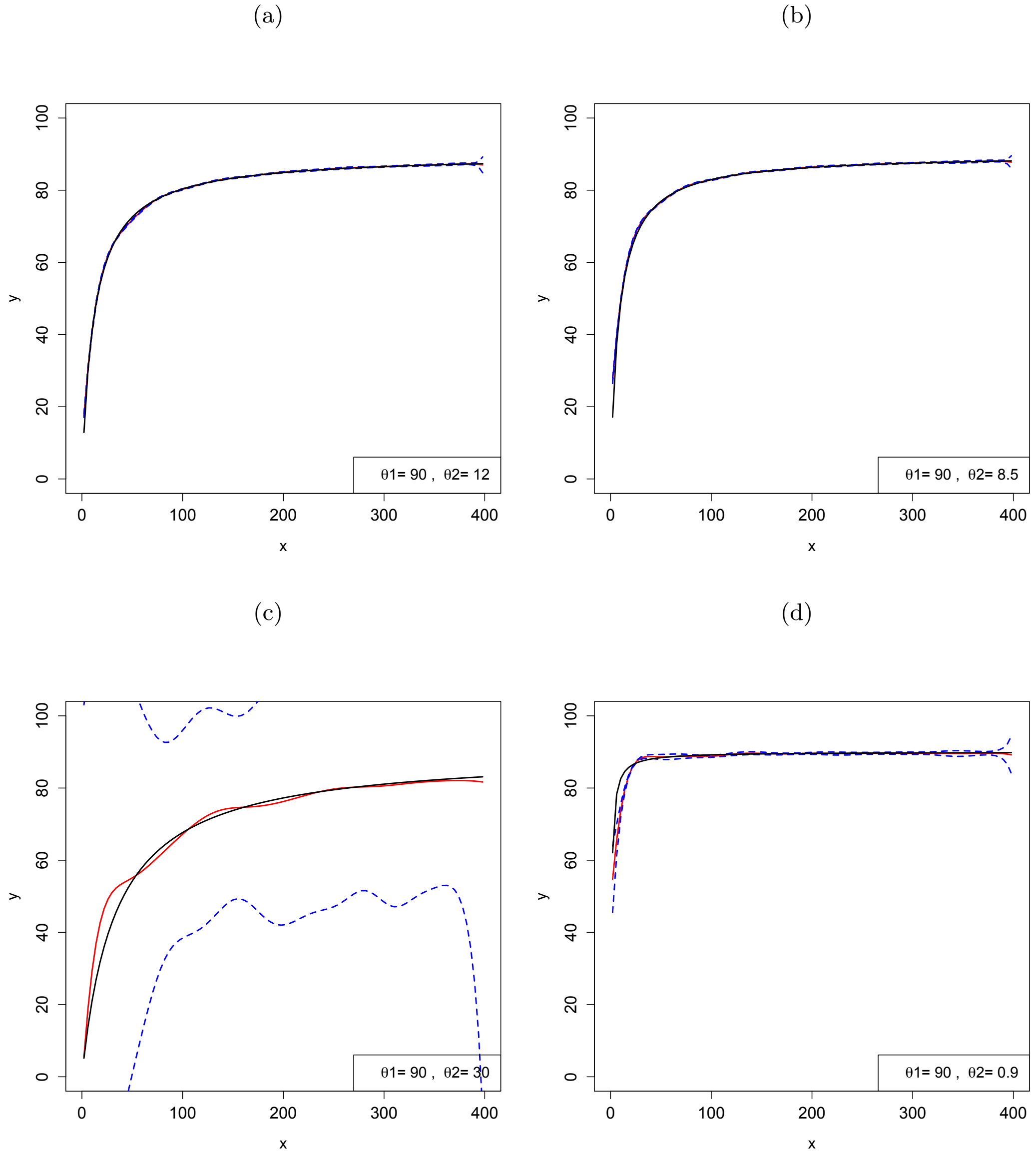


Figure 6.22: Posterior predictive mean for the Gaussian process fit with $m = 90$ simulator runs (red line); 95% probability bounds (blue lines); the true Michaelis-Menten function for a given $\boldsymbol{\theta}$ (black line)

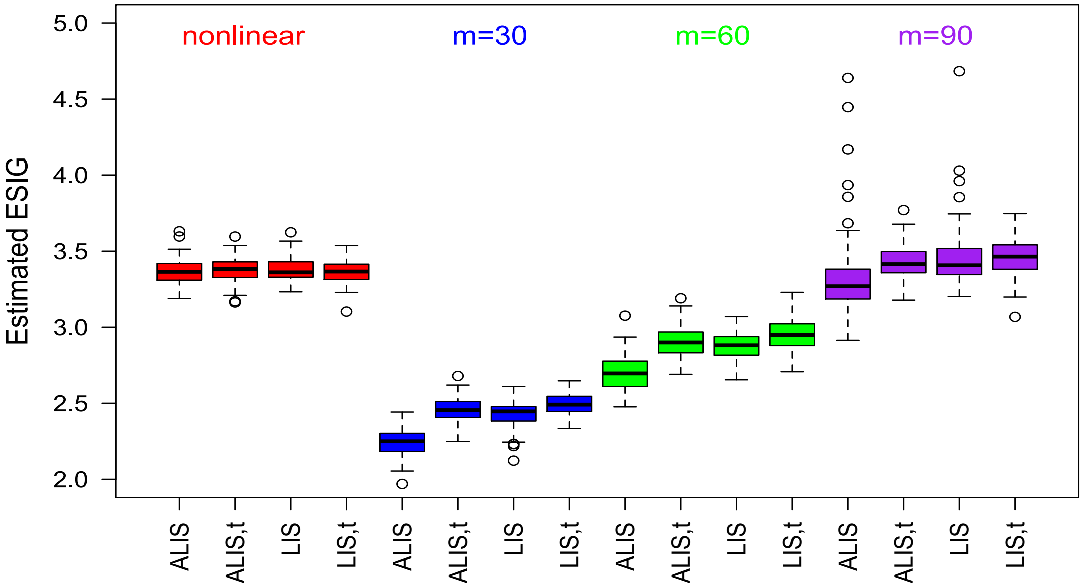


Figure 6.23: Estimated ESIG for the parameters of interest $\boldsymbol{\theta}^p$ of the calibration model (6.4) for the Michaelis-Menten example, as we increase the number of simulator runs, m, and change the prior of the nuisance parameter τ^2 (which is treated as a nuisance parameter), to keep the implied prior distribution of σ_ε^2 fixed, and also for the parameters $\boldsymbol{\theta}$ of the nonlinear Michaelis-Menten model where σ_ε^2 is a nuisance parameter

distribution on $\boldsymbol{\theta}^p$ (calibration model) and $\boldsymbol{\theta}$ (nonlinear Michaelis-Menten model).

In Figure 6.23 we present 100 estimates of the ESIG for the nonlinear Michaelis-Menten model where σ_ε^2 is treated as a nuisance parameter, and the calibration model where τ^2 is treated as a nuisance parameter, as we increase the number of simulator runs ($m = 30, 60, 90$). The mean and variance of the implied distribution of σ_ε^2 for the calibration model, is held approximately fixed by changing the mean and variance of τ^2. The ESIG for the parameters of interest $\boldsymbol{\theta}^p$ for the calibration model increases as m increases. For $m = 90$ the emulator is a better approximation of the simulator for the range of values of $\boldsymbol{\theta}^p$ that appear in the importance sample, and the ESIG is similar to the ESIG for the parameters $\boldsymbol{\theta}$ of the nonlinear Michaelis-Menten model for the design shown in Figure 6.16.

Next we find Bayesian optimal designs, as described before, using LIS and $k_1 = k_2 = 300$, and ACE for 10 random starts for the nonlinear Michaelis-Menten model where σ_ε^2 is a nuisance parameter (σ_ε^2 is integrated-out analytically, see Section 5.1.2), and LIS with $k_1 = k_2 = k_3 = 300$ for the calibration model where τ^2 is a nuisance parameter and $m = 30$. It was only possible to find optimal designs for the case $m = 30$ due to numerical issues we discuss in Section 6.5, caused by the Gaussian process fit failing as shown in Figures 6.20, 6.21 and 6.22.

The left hand panel in Figure 6.24 shows four Bayesian near-optimal designs, denoted

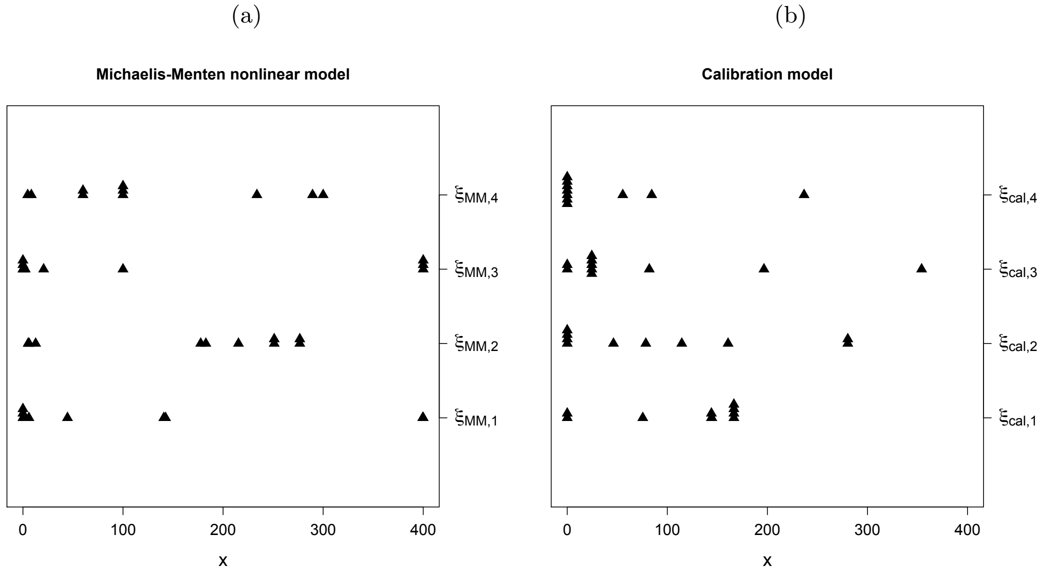


Figure 6.24: Bayesian near-optimal designs for (a) the Michaelis-Menten nonlinear regression model when σ_ε^2 is treated as a nuisance parameter; (b) the Michaelis-Menten calibration model when τ^2 is treated as a nuisance parameter and $m = 30$

$\xi_{MM,1}$, $\xi_{MM,2}$, $\xi_{MM,3}$ and $\xi_{MM,4}$, for the nonlinear Michaelis-Menten model found from four of the random starts of ACE, with prior hyperparameters $\mu_1 = 4.38$, $\sigma_1 = 0.07$, $\mu_2 = 1.19$, $\sigma_2 = 0.84$, $a = 2.915$ and $b = 16.92$, and σ_ε^2 treated as a nuisance parameter. The right hand panel in Figure 6.24 shows four Bayesian near-optimal designs, denoted $\xi_{cal,1}$, $\xi_{cal,2}$, $\xi_{cal,3}$ and $\xi_{cal,4}$, for the calibration model with $m = 30$, again obtained from four of the random starts of ACE, with hyperparameters $\mu_1 = 4.38$, $\sigma_1 = 0.07$, $\mu_2 = 1.19$, $\sigma_2 = 0.84$, $a = 3$, $b = 2$ and $\lambda_{\tau^2} = 50$, and τ^2 treated as a nuisance parameter. Designs $\xi_{MM,1}$ and $\xi_{MM,3}$ are similar to optimal designs for the Michaelis-Menten from previous examples, having most points where the function is changing fastest and some points where function is stable. Designs $\xi_{MM,2}$ and $\xi_{MM,4}$ have most points in the region where the function is changing fastest, however they do not have any points at the end of the design region as has been seen before. This may be a result of the new prior distribution for σ_ε^2, as the variance is now larger. For the calibration model, all designs follow this pattern with no points near the end of the design region. A potential explanation for this might be that at this edge of the design region we are very uncertain about the output of simulator for most parameter values (see Figures 6.20, 6.21 and 6.22).

In Figure 6.25 we present 100 estimates of the ESIG for the parameters $\boldsymbol{\theta}$ of the nonlinear Michaelis-Menten model when σ_ε^2 is treated as a nuisance parameter (σ_ε^2 is integrated-out analytically), found using LIS and $k_1 = k_2 = 300$ for the designs presented in Figure 6.24 (a). All designs have similar performance with $\xi_{MM,2}$ performing slightly better. We denote this design by $\xi_{MM}^\star$.

Figure 6.26 shows 100 estimates of the ESIG found using LIS and $k_1 = k_2 = k_3 = 300$

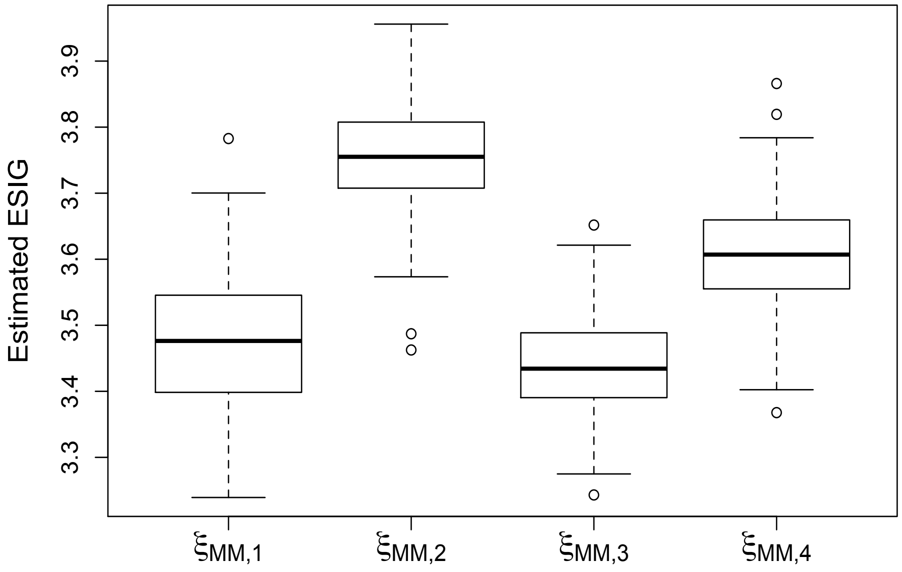


Figure 6.25: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the nonlinear Michaelis-Menten model when σ_ε^2 is treated as a nuisance parameter for the designs from Figure 6.24 (a)

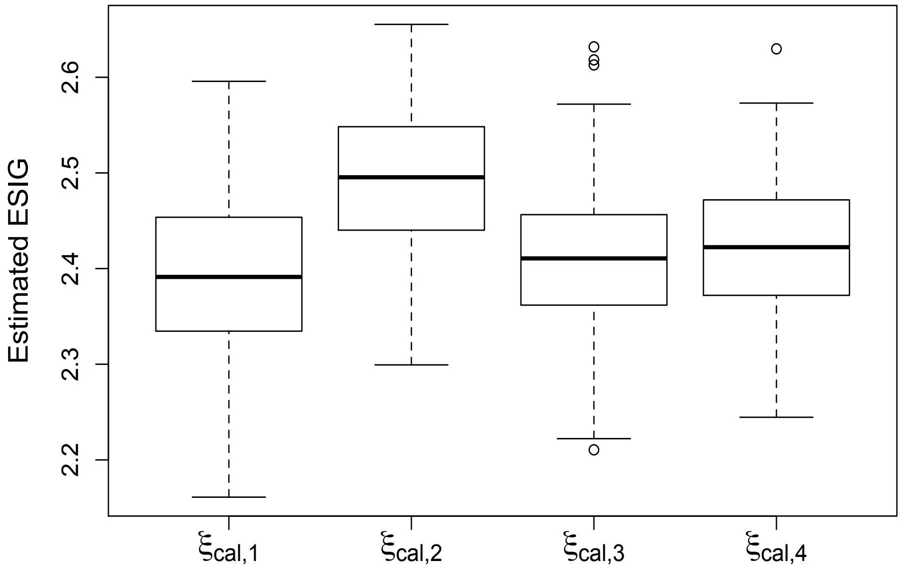


Figure 6.26: Estimated ESIG for the parameters $\boldsymbol{\theta}^p$ of the Michaelis-Menten calibration model where τ^2 is treated as a nuisance parameter and $m = 30$ for the designs from Figure 6.24 (b)

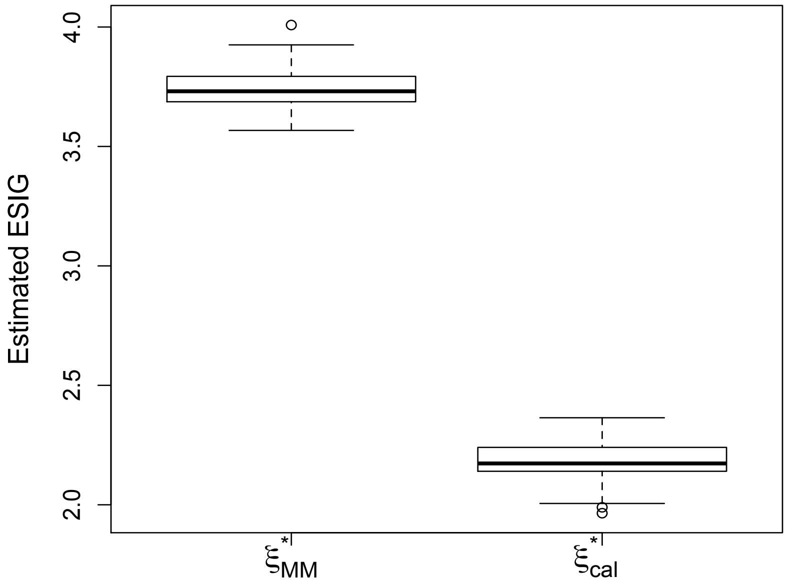


Figure 6.27: Estimated ESIG for the parameters $\boldsymbol{\theta}$ of the nonlinear Michaelis-Menten model when σ^2_ε is treated as a nuisance parameter, for the designs $\xi^\star_{MM}$ and $\xi^\star_{cal}$

for the calibration model when τ^2 is treated as a nuisance parameter and with $m = 30$, for the designs presented in Figure 6.24 (b). Again, all designs have similar performance with design $\xi_{cal,2}$ performing slightly better. We denote this design by $\xi^\star_{cal}$.

Figure 6.27 shows 100 estimates of the ESIG from LIS with $k_1 = k_2 = 300$ for the parameters $\boldsymbol{\theta}$ of the nonlinear Michaelis-Menten model, where σ^2_ε is a nuisance parameter, for $\xi^\star_{MM}$ and $\xi^\star_{cal}$. Figure 6.28 shows 100 estimates of the ESIG from LIS with $k_1 = k_2 = k_3 = 300$ for the calibration parameters $\boldsymbol{\theta}^p$, where τ^2 is a nuisance parameter, and $m = 30$ for $\xi^\star_{MM}$ and $\xi^\star_{cal}$. The Bayesian optimal design for the calibration model is not optimal under the Michaelis-Menten model as shown in Figure 6.27. However, the optimal design found under the Michaelis-Menten model is reasonably efficient for estimating the parameters of the calibration model.

In this example, again the simulator has a simple equation and hence we are able to compare optimal designs found for the calibration model with unknown simulator with optimal designs for a nonlinear model. This example shows that optimal designs obtained in the unknown simulator case can be inefficient under the nonlinear Michaelis-Menten model. In this example, this seems to be probably due to the relatively poor fit of the Gaussian process emulator.

For this example we are not able to construct optimal designs for ALIS and LIS for larger m because the prior distribution of the nugget τ^2 results in very small sampled values and the negative Hessian matrix $\mathbf{H}(\boldsymbol{\psi})$ becomes ill-conditioned (non-invertible).

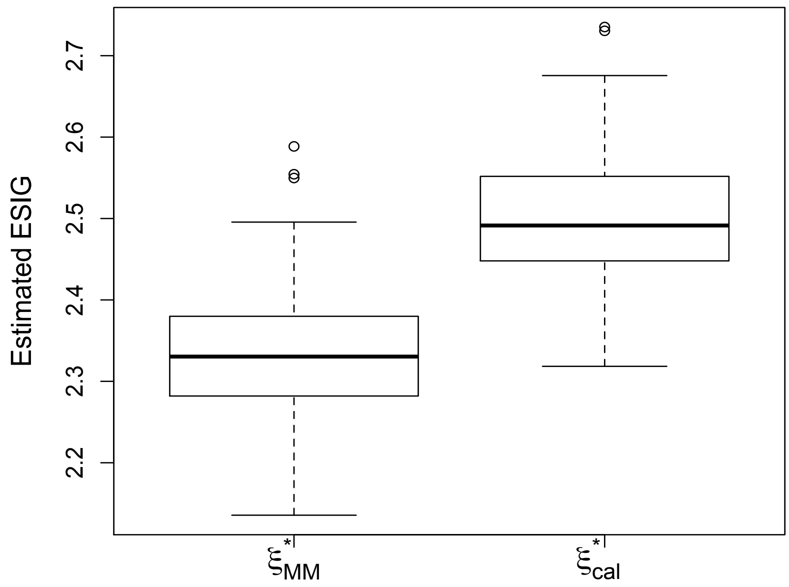


Figure 6.28: Estimated ESIG for the parameters of interest $\boldsymbol{\theta}^p$ of the Michaelis-Menten calibration model when treating τ^2 as a nuisance parameter and $m = 30$, for the designs $\xi^\star_{MM}$ and $\xi^\star_{cal}$

Numerical problems caused by these issues are discussed in the next section.

6.5 Numerical issues

As seen in Example 6.4.3, for calibration with an unknown simulator, ALIS can fail due to the Hessian matrix $\mathbf{H}(\boldsymbol{\psi})$ being ill-conditioned. These matrices result in very large parameter variances in the ALIS importance distribution. As a consequence, we obtain very large values of some sampled parameters, for which the likelihood evaluation cannot be performed. Conditional on the likelihood evaluation being possible the ALIS estimator seems to result in negative bias in the estimate of the expected Shannon information gain. Hence we would recommend using LIS rather than ALIS for this type of problem.

In Example 6.4.4 the Gaussian process posterior mean did not adapt to the true model very quickly and the Gaussian process variance, σ^2, increased with the number of simulator runs, m. A potential explanation for this is that a stationary Gaussian process is a poor approximation to the model. To deal with this issue we altered the prior distribution on τ^2 for different values of m in order to keep the implied prior distribution on σ^2_ε fairly constant. To achieve this for large m, the prior mean of τ^2 had to be made very small. This results in numerical issues in ALIS and LIS as we are unable to invert the variance covariance matrix $\tilde{\boldsymbol{\Sigma}}_y$ and the likelihood function is not defined. Hence, the common parameterisation $\tau^2 = \sigma_\varepsilon^2/\sigma^2$ appears to perform poorly in cases like this.

6.6 Summary

We have developed the necessary methods to address two key problems within the Kennedy-O'Hagan calibration framework, namely Bayesian design when: (i) the function $\eta(\mathbf{x}, \boldsymbol{\theta})$ does not provide an accurate description of the mean; and (ii) the model may be expensive to evaluate or unknown precluding direct use of the model in inference. We showed how ALIS and LIS can be used within these very general settings to approximate the expected Shannon information gain. We have shown that designs tailored to the calibration problem perform better than either existing optimal designs or space-filling designs. For each of these two problems we found Bayesian optimal designs using the ACE algorithm and compared them with Bayesian optimal designs for nonlinear models. Last, we have shown that optimal designs for the calibration model with $\delta_{\boldsymbol{\theta}^p}(\mathbf{x}) = 0$ perform as well as optimal designs obtained when we know the model if suitable experiments can be performed.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis we have discussed the problem of Bayesian design for nonlinear models, particularly physical models within the Kennedy-O'Hagan framework. The objective of this research was to develop methodology for Bayesian optimal designs for the physical experiment to be combined with limited simulator runs to perform inference for the unknown parameters. We sought Bayesian optimal designs that maximise the expected Shannon information gain when the aim of the experiment was to estimate unknown parameters (Chapter 5 and Chapter 6). Throughout the thesis, we have discussed the challenges of approximating this expected utility which, in many cases, is intractable and involves high-dimensional integrals. We also discussed how existing methods, in some cases, fail to give an accurate approximation of the expected utility.

We have developed, assessed and compared new methods for approximating the expected Shannon information gain, namely Laplace importance sampling (LIS) and approximate Laplace importance sampling (ALIS); see Chapter 4. We firstly applied these methods in the search for Bayesian design for nonlinear models, and showed that their use provides better approximations than existing methods; they provide a good balance between bias and computational expense. Combined with an optimisation algorithm, we have also showed that these new methods can produce designs that have better performance than the designs produced with the other methods (Chapter 5).

We also developed the necessary methods to address two key design problems within the Kennedy-O'Hagan calibration framework, namely Bayesian design when: (i) the function $\eta(\mathbf{x}, \boldsymbol{\theta})$ does not provide an accurate description of the mean; and (ii) the model may be expensive to evaluate or unknown, precluding its direct use in inference. In Chapter 6 we showed how ALIS and LIS can be used to approximate the expected Shannon information gain in these two cases, and hence facilitate the search for optimal designs.

The methods in this thesis help to overcome the computational complexity of Bayesian optimal design, and address the reliance of previous methods on either knowing the functional form of the simulator, as in traditional nonlinear model design, or assuming the calibration parameters are known, as in most of the existing optimal design methods for calibration.

One limitation of this research is that the derivatives of the model are required. However, this is not a huge problem when a Gaussian process prior is used to model the simulator. The research in this thesis is tailored to optimal designs maximising the expected Shannon information gain. However ALIS and LIS provide a better approximation to the posterior distribution (Section 4.3) that should be beneficial in the construction of distributions of interest when using a different utility function.

Lastly, it is well-known that when the discrepancy function is present, the calibration parameters are not uniquely identifiable without the use of informative prior distributions. To allow unique estimation of both the calibration parameters and the discrepancy function, the Gaussian process prior of the discrepancy function must be formulated appropriately and satisfy some constraints; this topic is discussed below in Section 7.2.2.

7.2 Future work

7.2.1 ALIS and LIS

ALIS and LIS could be extended to approximate different utility functions, for example the expected Shannon information gain between prior and posterior predictive distributions when the aim of the experiment is prediction, or the negative square error loss for either parameters or predictions. For these utility functions, different distributions must be approximated; however a better approximation to the posterior distribution, as given in Section 4.3, will aid in the construction of approximations to the distributions of interest. For example, to approximate the negative squared error loss utility function given by

$$u(\xi, \boldsymbol{\psi}, \mathbf{y}) = -\sum_{w=1}^{q_2} [\psi_w - \mathbb{E}(\psi_w|\mathbf{y}, \xi)]^2,$$

$\mathbb{E}(\psi_w|\mathbf{y}, \xi)$ can be approximated via ALIS and LIS as

$$\tilde{\mathbb{E}}^h(\psi_w|\mathbf{y}, \xi) = \frac{\sum_{k=1}^{k_2} \tilde{\psi}_{kw} \frac{\pi_l(\mathbf{y}_h|\tilde{\boldsymbol{\psi}}_{hk}, \xi)\pi_b(\tilde{\boldsymbol{\psi}}_{hk})}{q^h_{\boldsymbol{\psi}}(\tilde{\boldsymbol{\psi}}_{hk})}}{\sum_{k=1}^{k_2} \frac{\pi_l(\mathbf{y}_h|\tilde{\boldsymbol{\psi}}_{hk}, \xi)\pi_b(\tilde{\boldsymbol{\psi}}_{hk})}{q^h_{\boldsymbol{\psi}}(\tilde{\boldsymbol{\psi}}_{hk})}}$$

where $\tilde{\boldsymbol{\psi}}_{hk}$, $h = 1, \ldots, k_1$, $k = 1, \ldots, k_2$, is a sample from the importance distribution (see also Section 4.3 for notation). Note that using an approximation to the posterior is not necessary the optimal choice of importance density for estimating the posterior mean.

As discussed in Chapter 5, the ALIS and LIS approximations are combined with the ACE algorithm to obtain Bayesian optimal designs. This combination involves a trade-off between accuracy of the expected Shannon information gain and computational expense of the optimisation of this utility. Namely, should the computational budget be spent on more precise and accurate approximations to the expected utility, or on performing more random starts of the ACE algorithm? Future work could investigate this trade-off and provide recommendations for specific classes of problem. A further refinement to the computational methodology could be to vary the values of the outer Monte Carlo sample size in the ALIS and LIS approximations as we progress through the iterations in ACE, with larger sample sizes for later iterations. A better estimate of the expected utility is more important for later iterations of ACE, where smaller improvements in the expected utility are anticipated.

7.2.2 Design for calibration

Our methodology for finding fully Bayesian optimal deigns for calibration could be extended to experiments where the simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$ is both computationally expensive (with no closed form) and subject to non-zero discrepancy. Gaussian process priors for both the simulator and the unknown discrepancy function must be assumed. For $\eta(\mathbf{x}, \boldsymbol{\theta})$, we have

$$\eta(\mathbf{x}, \boldsymbol{\theta}) \sim \mathrm{GP}\left(\mathbf{f}_\eta^{\mathrm{T}}(\mathbf{x}, \boldsymbol{\theta})\boldsymbol{\beta}_\eta,\ \sigma_\eta^2 \kappa_\eta[(\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}'); \boldsymbol{\phi}_\eta]\right),$$

and for $\delta_{\boldsymbol{\theta}^p}(\mathbf{x})$,

$$\delta_{\boldsymbol{\theta}^p}(\mathbf{x}) \sim \mathrm{GP}\left(\mathbf{f}_\delta^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}_\delta, \sigma_\delta^2 \kappa_\delta(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi}_\delta)\right),$$

as described in more detail in Section 2.2. The distribution of the combined $(n+m)$-vector of responses $\mathbf{v} = [\mathbf{y}^{\mathrm{T}}\ \ \mathbf{z}^{\mathrm{T}}]^{\mathrm{T}}$ from the physical and the computer experiment is:

$$\mathbf{v} \mid \boldsymbol{\psi} \sim N(\boldsymbol{\mu}_{\mathbf{v}}, \boldsymbol{\Sigma}_{\mathbf{v}}),$$

where $\boldsymbol{\psi} = [(\boldsymbol{\theta}^p)^{\mathrm{T}}, \boldsymbol{\beta}_\eta^{\mathrm{T}}, \boldsymbol{\beta}_\delta^{\mathrm{T}}, \sigma_\eta^2, \sigma_\delta^2, \sigma_\varepsilon^2, \boldsymbol{\phi}_\eta^{\mathrm{T}}, \boldsymbol{\phi}_\delta^{\mathrm{T}}]^{\mathrm{T}}$, with

$$\boldsymbol{\mu}_{\mathbf{v}} = \mathbb{E}[\mathbf{v}] = \begin{bmatrix} \mathbf{F}_\eta^p \boldsymbol{\beta}_\eta + \mathbf{F}_\delta^p \boldsymbol{\beta}_\delta \\ \mathbf{F}_\eta^c \boldsymbol{\beta}_\eta \end{bmatrix},$$

and

$$\boldsymbol{\Sigma}_{\mathbf{v}} = \mathrm{cov}[\mathbf{v}] = \sigma_\eta^2 \boldsymbol{\Sigma}_\eta + \begin{bmatrix} \sigma_\varepsilon^2 \mathbf{I}_n + \sigma_\delta^2 \boldsymbol{\Sigma}_\delta & \mathbf{0}_{n\times m} \\ \mathbf{0}_{m\times n} & \mathbf{0}_{m\times m} \end{bmatrix},$$

where

$$\boldsymbol{\Sigma}_\eta = \begin{bmatrix} \boldsymbol{\Sigma}_\eta^{pp} & \boldsymbol{\Sigma}_\eta^{pc} \\ \boldsymbol{\Sigma}_\eta^{cp} & \boldsymbol{\Sigma}_\eta^{cc} \end{bmatrix}.$$

The joint distribution of $\mathbf{y}$ and $\mathbf{z}$, conditional on all unknown model parameters $\boldsymbol{\psi}$ is:

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{z}\end{pmatrix}\Bigg|\, \boldsymbol{\psi} \;\sim\; N\left(\begin{pmatrix}\mathbf{F}^p_\eta\boldsymbol{\beta}_\eta + \mathbf{F}^p_\delta\boldsymbol{\beta}_\delta\\ \mathbf{F}^c_\eta\boldsymbol{\beta}_\eta\end{pmatrix},\; \begin{pmatrix}\sigma^2_\eta\boldsymbol{\Sigma}^{pp}_\eta + \sigma^2_\varepsilon\mathbf{I}_n + \sigma^2_\delta\boldsymbol{\Sigma}_\delta & \sigma^2_\eta\boldsymbol{\Sigma}^{cp\mathrm{T}}_\eta\\ \sigma^2_\eta\boldsymbol{\Sigma}^{cp}_\eta & \sigma^2_\eta\boldsymbol{\Sigma}^{cc}_\eta\end{pmatrix}\right),$$

The correlation matrices $\boldsymbol{\Sigma}^{pp}_\eta$, $\boldsymbol{\Sigma}^{cp}_\eta$, $\boldsymbol{\Sigma}^{cc}_\eta$ and $\boldsymbol{\Sigma}_\delta$ are defined through the correlation functions with entries given by:

$$\begin{aligned}
\boldsymbol{\Sigma}^{pp}_{\eta,ii'} &= \kappa_\eta[(\mathbf{x}^p_i,\boldsymbol{\theta}^p),(\mathbf{x}^p_{i'},\boldsymbol{\theta}^p);\boldsymbol{\phi}_\eta],\\
\boldsymbol{\Sigma}^{cp}_{\eta,ji} &= \kappa_\eta[(\mathbf{x}^c_j,\boldsymbol{\theta}^c_j),(\mathbf{x}^p_i,\boldsymbol{\theta}^p);\boldsymbol{\phi}_\eta],\\
\boldsymbol{\Sigma}^{cc}_{\eta,jj'} &= \kappa_\eta[(\mathbf{x}^c_j,\boldsymbol{\theta}^c_j),(\mathbf{x}^c_{j'},\boldsymbol{\theta}^c_{j'});\boldsymbol{\phi}_\eta],\\
\boldsymbol{\Sigma}_{\delta,ii'} &= \kappa_\delta[(\mathbf{x}^p_i,\mathbf{x}^p_{i'});\boldsymbol{\phi}_\delta],
\end{aligned}$$

where $i, i' = 1,\dots,n$, and $j, j' = 1,\dots,m$.

Standard results for multivariate normal distributions can be used to derive the following conditional distribution

$$\mathbf{y}\mid\mathbf{z},\boldsymbol{\psi}\sim N\left(\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y\right),$$

with

$$\begin{aligned}
\boldsymbol{\mu}_y &= \mathbb{E}(\mathbf{y}\mid\mathbf{z},\boldsymbol{\psi}) = \mathbf{F}^p_\eta\boldsymbol{\beta}_\eta + \mathbf{F}^p_\delta\boldsymbol{\beta}_\delta + \boldsymbol{\Sigma}^{cp\mathrm{T}}_\eta{\boldsymbol{\Sigma}^{cc}_\eta}^{-1}\left[\mathbf{z}-\mathbf{F}^c_\eta\boldsymbol{\beta}_\eta\right],\\
\boldsymbol{\Sigma}_y &= \mathrm{var}(\mathbf{y}\mid\mathbf{z},\boldsymbol{\psi}) = \sigma^2_\eta\boldsymbol{\Sigma}^{pp}_\eta + \sigma^2_\varepsilon\mathbf{I}_n + \sigma^2_\delta\boldsymbol{\Sigma}_\delta - \boldsymbol{\Sigma}^{cp\mathrm{T}}_\eta{\boldsymbol{\Sigma}^{cc}_\eta}^{-1}\boldsymbol{\Sigma}^{cp}_\eta.
\end{aligned}$$

As before, the model specification requires prior distributions for the unknown parameters $\boldsymbol{\psi} = [(\boldsymbol{\theta}^p)^\mathrm{T},\boldsymbol{\beta}^\mathrm{T}_\eta,\boldsymbol{\beta}^\mathrm{T}_\delta,\sigma^2_\eta,\sigma^2_\delta,\sigma^2_\varepsilon,\boldsymbol{\phi}^\mathrm{T}_\eta,\boldsymbol{\phi}^\mathrm{T}_\delta]^\mathrm{T}$.

In order to use the ALIS and LIS approximations to the expected Shannon information gain, the negative Hessian of the log-unnormalised posterior density, $\mathbf{H}(\boldsymbol{\psi})$, is required. The log-unnormalised posterior density is given by:

$$\begin{aligned}
\log\pi_u(\boldsymbol{\psi}|\mathbf{y},\mathbf{z},\xi) = {}& \log[\pi_l(\mathbf{y}|\mathbf{z},\boldsymbol{\psi},\xi)] + \log[\pi_b(\boldsymbol{\theta}^p)] + \log[\pi_b(\boldsymbol{\beta}_\eta)] + \log[\pi_b(\boldsymbol{\beta}_\delta)]\\
&+ \log[\pi_b(\sigma^2_\eta)] + \log[\pi_b(\sigma^2_\delta)] + \log[\pi_b(\sigma^2_\varepsilon)] + \log[\pi_b(\boldsymbol{\phi}_\eta)] + \log[\pi_b(\boldsymbol{\phi}_\delta)].
\end{aligned}$$

Once the derivatives of this expression have been obtained the procedure described in Sections 4.3 and 5.2 can be followed to estimate the expected Shannon information gain for a given design and find Bayesian optimal designs using the ACE algorithm as in Chapters 5 and 6.

It is well-known that when a discrepancy function is present, the calibration parameters are not identifiable (see Section 1.1). To resolve this identifiability problem, Plumlee (2017) suggested Bayesian L_2-calibration, which involves the use of a Gaussian process prior with a correlation function that incorporates the constraint that $\delta_{\boldsymbol{\theta}^p}(\cdot)$ is orthogonal to the gradient of the simulator $\eta(\mathbf{x},\boldsymbol{\theta})$. Specifically, when the simulator $\eta(\mathbf{x},\boldsymbol{\theta})$

is known, the correlation function of the discrepancy function $\delta_{\boldsymbol{\theta}^P}(\mathbf{x})$ is chosen as:

$$\tilde{\kappa}_\delta(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi}_\delta) = \kappa_\delta(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi}_\delta) - h_\theta(\mathbf{x})^{\mathrm{T}} H_\theta^{-1} h_\theta(\mathbf{x}'),$$

where $\kappa_\delta(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi}_\delta)$ is an arbitrary choice of correlation function (e.g. the squared exponential),

$$h_\theta(\mathbf{x}) = \int_{\mathcal{X}} \frac{\partial \eta(\mathbf{t}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \kappa_\delta(\mathbf{x}, \mathbf{t}; \boldsymbol{\phi}_\delta) d\mathbf{t},$$

and

$$H_\theta = \int_{\mathcal{X}} \int_{\mathcal{X}} \frac{\partial \eta(\mathbf{t}', \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left[\frac{\partial \eta(\mathbf{t}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]^{\mathrm{T}} \kappa_\delta(\mathbf{t}', \mathbf{t}; \boldsymbol{\phi}_\delta) d\mathbf{t}' d\mathbf{t}.$$

A harder problem is when the simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$ is unknown or expensive to evaluate. For this problem the derivatives and integrals must be approximated numerically, see Plumlee (2017) for more details. A very interesting extension of our results would be to apply ALIS and LIS as part of a methodology to find optimal designs under these L_2-calibration prior distributions.

Bibliography

Arendt, P. D., Apley, D. W. and Chen, W. (2016) A preposterior analysis to predict identifiability in the experimental calibration of computer models. *Institute of Industrial Engineers Transactions*, **48**, 75–88.

Atkinson, A. C., Donev, A. N. and Tobias, R. (2007) *Optimum Experimental Designs, with SAS.* Oxford: Oxford University Press, 2nd edn.

Baldi, P. and Itti, L. (2010) Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks*, **23**, 649–666.

Banerjee, S., Carlin, B. and Gelfand, A. (2004) *Hierarchical Modeling and Analysis for Spatial Data.* Boca Raton: Chapman and Hall/CRC, 2nd edn.

Bates, D. M. and Watts, D. G. (1988) *Nonlinear Regression Analysis and its Applications.* Wiley, New York.

Bayarri, M. J., Berger, J. O., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R. J., Paulo, R., Sacks, J. and Walsh, D. (2007a) Computer model validation with functional output. *The Annals of Statistics*, **35**, 1874–1906.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C. H. and Tu, J. (2007b) A framework for validation of computer models. *Technometrics*, **49**, 138–154.

Beck, J., Dia, B. M., Espath, L. F. R., Long, Q. and Tempone, R. (2018) Fast Bayesian experimental design: Laplace-based importance sampling for the expected information gain. *Computer Methods in Applied Mechanics and Engineering*, **334**, 523–553.

Berger, J. (1985) *Statistical Decision Theory and Bayesian Analysis.* New York: Springer, 2nd edn.

Bernardo, J. M. (1979) Expected information as expected utility. *The Annals of Statistics*, **7**, 686–690.

Berthouex, P. M. and Brown, L. C. (2002) *Statistics for Environmental Engineers.* Boca Raton: CRC Press, 2nd edn.

Bliss, C. I. and James, A. T. (1966) Fitting the rectangular hyperbola. *Biometrics*, **22**, 573–602.

Boer, E. P. J., Rasch, D. and Hendrix, E. M. T. (2000) Locally optimal designs in non-linear regression: a case study for the Michaelis-Menten function. In *Advances in Stochastic Simulation Methods. Statistics for Industry and Technology* (eds. N. Balakrishnan, V. B. Melas and S. Ermakov), 177–188. Birkhäuser, Boston.

Bonnans, J. F., Gilbert, J. C., Lemaréchal, C. and Sagastizábal, C. A. (2006) *Numerical Optimization: Theoretical and Practical Aspects.* New York: Springer, 2nd edn.

Brynjarsdóttir, J. and O'Hagan, A. (2014) Learning about physical parameters: the importance of model discrepancy. *Inverse Problems*, **30**, 114007.

Burstyn, D. and Steinberg, D. M. (2006) Comparison of designs for computer experiments. *Journal of Statistical Planning and Inference*, **136**, 1103–1119.

Chaloner, K. (1984) Optimal Bayesian experimental design for linear models. *Annals of Statistics*, **12**, 283–300.

Chaloner, K. and Verdinelli, I. (1995) Bayesian experimental design: a review. *Statistical Science*, **10**, 273–304.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Cook, A., Gibson, G. and Gilligan, C. (2008) Optimal observation times in experimental epidemic processes. *Biometrics*, **64**, 860–868.

Cornish-Bowden, A. (1995) *Analysis of Enzyme Kinetic Data.* Oxford: Oxford University Press.

Craig, P. S., Goldstein, M., Seheult, A. H. and Smith, J. A. (1997) Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments. In *Case Studies in Bayesian Statistics. Lecture Notes in Statistics* (eds. C. Gatsonis, J. S. Hodges, R. E. Kass, R. McCulloch, P. Rossi and N. D. Singpurwalla), vol. 121, 37–93. New York: Springer.

Currie, D. J. (1982) Estimating Michaelis-Menten parameters: bias, variance and experimental design. *Biometrics*, **38**, 907–919.

Czitrom, V. (1999) One-factor-at-a-time versus designed experiments. *The American Statistician*, **53**, 126–131.

Dancik, G. (2007) mlegp: an R package for Gaussian process modelling and sensitivity analysis. URL https://cran.r-project.org/web/packages/mlegp/mlegp.pdf.

Dette, H. and Biedermann, S. (2003) Robust and efficient designs for the Michaelis-Menten model. *Journal of the American Statistical Association*, **98**, 679–686.

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997) Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of American Statistical Association*, **92**, 903–915.

Diggle, P., Moyeed, R. and Tawn, J. (1998) Model-based geostatistics. *Journal of the Royal Statistical Society: Series C*, **47**, 299–350.

Diggle, P. and Ribeiro, J. (2007) *Model-based Geostatistics.* New York: Springer.

Duggleby, R. G. and Clarke, R. B. (1991) Experimental designs for estimating the parameters of the Michaelis-Menten equation from progress curves of enzyme-catalyzed reactions. *Biochimica et Biophysica Acta*, **1080**, 231–236.

Fang, K., Li, R. and Sudjianto, A. (2006) *Design and Modelling for Computer Experiments.* Boca Raton: Chapman and Hall/CRC.

Fang, K. T., Lin, D. K. J., Winker, P. and Zhang, Y. (2000) Uniform design: theory and application. *Technometrics*, **42**, 237–248.

Fedorov, V. (1972) *Theory of Optimal Experiments.* New York: Academic Press.

Feng, C. (2015) *Optimal Bayesian experimental design in the presence of model error.* Master's thesis, Center for Computational Engineering, Massachusetts Institute of Technology.

Forrester, A. I. J. (2010) Black-box calibration for complex-system simulation. *The Royal Society Interface*, **368**, 3567–3579.

Friel, N. and Pettitt, A. N. (2008) Marginal likelihood estimation via power posteriors. *Journal of Royal Statistical Society, Series B*, **70**, 589–607.

Friel, N. and Wyse, J. (2012) Estimating the evidence - a review. *Statistica Neerlandica*, **66**, 288–308.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian Data Analysis.* Boca Raton: Chapman and Hall/CRC Press, 3rd edn.

Gelman, A. and Meng, X. L. (1998) Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, **13**, 163–185.

Geweke, J. (1989) Bayesian inference in Econometric models using Monte Carlo integration. *Econometrica*, **57**, 1317–1339.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter), chap. 1. Boca Raton: CRC Press.

Glick, N., Landman, A. D. and Roufogalis, B. D. (1979) Correcting Lineweaver-Burk calculations of V and K_m. *Trends in Biochemical Sciences*, **4**, 82–83.

Goh, J., Bingham, D., Holloway, J. P., Grosskopf, M. J., Kuranz, C. C. and Rutter, E. (2013) Prediction and computer model calibration using outputs from multi-fidelity simulators. *Technometrics*, **55**, 501–512.

Gotwalt, C. M., Jones, B. A. and Steinberg, D. M. (2009) Fast computation of designs robust to parameter uncertainty for nonlinear settings. *Technometrics*, **51**, 88–105.

Gramacy, R. B., Bingham, D., Holloway, J. P., Grosskopf, M. J., Kuranz, C. C., Rutter, E., Trantham, M. and Drake, P. R. (2015) Calibrating a large computer experiment simulating radiative shock hydrodynamics. *The Annals of Applied Statistics*, **9**, 1141–1168.

Gramacy, R. B. and Lee, H. K. H. (2008) Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, **103**, 1119–1130.

Graybill, F. A. (1983) *Matrices with Applications in Statistics.* Belmont CA: Wadsworth, second edn.

Gu, M. and Wang, L. (2018) Scaled Gaussian stochastic process for computer model calibration and prediction. *arXiv:1707.08215.*

Han, G., Santner, T. J. and Rawlinson, J. J. (2009) Simultaneous determination of tuning and calibration parameters for computer experiments. *Technometrics*, **51**, 464–474.

Hardin, R. H. and Sloane, N. J. A. (1993) A new approach to the construction of optimal designs. *Journal of Statistical Planning and Inference*, **37**, 339–369.

Harville, D. (2008) *Matrix Algebra from a Statistician's Perspective.* New York: Springer.

Hastings, W. K. (1970) Monte Carlo sampling using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Higdon, D., Gattiker, J., Williams, B. and Rightley, M. (2008) Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, **103**, 570–583.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A. and Ryne, R. D. (2004) Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, **26**, 448–466.

Huan, X. and Marzouk, Y. (2013) Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, **232**, 288–317.

Johnson, M., Moore, L. M. and Ylvisaker, D. (1990) Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, **26**, 131–148.

Johnson, M. and Nachtsheim, C. (1983) Some guidelines for constructing exact D-optimal designs on convex designs spaces. *Technometrics*, **25**, 271–277.

Joseph, R. V., Gul, E. and Ba, S. (2015) Maximum projection designs for computer experiments. *Biometrika*, **102**, 371–380.

Joseph, V. R. and Melkote, S. N. (2009) Statistical adjustments to engineering models. *Journal of Quality Technology*, **41**, 362–375.

Kennedy, J. and Eberhart, R. (1995) Particle swarm optimization. In *Proceedings of International Conference on Neural Networks, 1995*, vol. 4, 1942–1948. IEEE.

Kennedy, M. and O'Hagan, A. (2001) Bayesian calibration for computer models (with discussion). *Journal of the Royal Statistical Society: Series B*, **63**, 425–464.

Kiefer, J. and Wolfowitz, J. (1959) Optimum designs in regression problems. *The Annals of Mathematical Statistics*, **30**, 271–294.

Kuk, A. Y. C. (1999) Laplace importance sampling for generalized linear mixed models. *Journal of Statistical Computation and Simulation*, **63**, 143–158.

Laine, M. (2008) *Adaptive MCMC Methods with Applications in Environmental and Models*. Ph.D. thesis, Finnish Meteorological Institute, Lappeenranta, Finland.

Leatherman, E. R., Dean, A. M. and Santner, T. J. (2017) Designing combined physical and computer experiments to maximize prediction accuracy. *Computational Statistics and Data Analysis*, **113**, 346–362.

Leatherman, E. R., Guo, H., Gilbert, S. L., Hutchinson, I. D., Maher, S. A. and Santner, T. J. (2014) Using a statistically calibrated biphasic finite element model of the human knee joint to identify robust designs for a meniscal substitute. *Journal of Biomechanical Engineering*, **136**, 0710071–0710078.

Lin, D. C. and Tang, B. (2015) Latin Hypercube and space-filling designs. In *Handbook of Design and Analysis of Experiments* (eds. A. Dean, M. Morris, J. Stufken and D. Bingham), chap. 17. Boca Raton: CRC Press.

Lindley, D. V. (1956) On the measure of information provided by an experiment. *The Annals of Statistics*, **27**, 986–1005.

Loeppky, J. L., Bingham, D. and Welch, W. J. (2006) Computer model calibration or tuning in practice. *Tech. Rep. 221*, Department of Statistics, The University of British Columbia, Vancouver.

Long, Q., Scavino, M., Tempone, R. and Wang, S. (2013) Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, **259**, 24–39.

Lòpez-Fidalgo, J. and Wong, W. K. (2002) Design issues for the Michaelis-Menten model. *Journal of Theoretical Biology*, **215**, 1–11.

Matèrn, B. (1960) Spatial variation, stochastic models and their application to some problems in forest surveys and other sampling investigation. *Medd. Statens Skogsforskningsinst*, **5**, 1–144.

Mathai, A. M. and Provost, S. B. (1992) *Quadratic Forms in Random Variables.* New York: Dekker.

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239–245.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Meyer, R. K. and Nachtsheim, C. J. (1995) The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, **37**, 60–69.

Michaelis, L. and Menten, M. L. (1913) Die kinetnik der invertinwirkung. *Biochem. Z.*, **49**, 334–336.

Morris, M. and Mitchell, T. (1995) Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, **43**, 381–402.

Müller, P. (1999) Simulation-based optimal design. *Bayesian Statistics*, **6**, 459–474.

Müller, P., Berry, D., Grieve, A. and Krams, M. (2006) A Bayesian decision-theoretic dose-finding trial. *Decision Analysis*, **3**, 197–207.

Müller, P. and Parmigiani, G. (1996) Optimal design via curve fitting of Monte Carlo experiments. *Journal of the American Statistical Association*, **90**, 1322–1330.

Murphy, J. M., Booth, B. B. B., Collins, M., Harris, G. R., Sexton, D. M. H. and Webb, M. J. (2007) A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philosophical Transactions of the Royal Society A. Mathematical, Physical and Engineering Scienses*, **365**, 1993–2028.

Neal, R. M. (2001) Annealed importance sampling. *Statistics and Computing*, **11**, 125–139.

Newton, M. A. and Raftery, A. E. (1994) Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of Royal Statistical Society, Series B*, **56**, 3–48.

Norrisa, D., Leesmana, G., Sinkoa, P. and Grassa, G. (2000) Development of predictive pharmacokinetic simulation models for drug discovery. *Journal of Controlled Release*, **65**, 55–62.

Oakley, J. E. and Youngman, B. D. (2017) Calibration of stochastic computer simulators using likelihood emulation. *Technometrics*, **59**, 80–92.

Oh, M. S. and Berger, J. O. (1993) Integration of multimodal functions by Monte Carlo importance sampling. *Journal of American Statistical Association*, **88**, 450–456.

O'Hagan, A. (1978) Curve fitting and optimal design for prediction (with discussion). *Journal of the Royal Statistical Society: Series B*, **40**, 1–42.

O'Hagan, A. and Forster, J. (2004) *Bayesian Inference*, vol. 2B of *Kendall's Advanced Theory of Statistics*. Oxford University Press: Arnold, 2nd edn.

Overstall, A. M., McGree, J. M. and Drovandi, C. C. (2018) An approach for finding fully Bayesian optimal designs using normal-based approximations to loss functions. *Statistics and Computing*, **28**, 343–358.

Overstall, A. M. and Woods, D. C. (2017) Bayesian design of experiments using approximate coordinate exchange. *Technometrics*, **59**, 458–470.

Overstall, A. M., Woods, D. C. and Adamou, M. (2017) acebayes: An R package for Bayesian optimal design of experiments via approximate coordinate exchange. *arXiv:1705.08096*.

Owen, A. B. (2013) *Monte Carlo Theory, Methods and Examples*. URL https://statweb.stanford.edu/~owen/mc/.

Parise, S. and Welling, M. (2007) Bayesian model scoring in markov random fields. In *Advances in neural information processing systems* (eds. B. Schölkopf, J. Platt and T. Hoffman), vol. 19, 1073–1080. MIT Press.

Plumlee, M. (2017) Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, **112**, 1274–1285.

Porat, B. and Friedlander, B. (1986) Computation of the exact information matrix of Gaussian time series with stationary random components. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **34**, 118–130.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007) *Numerical Recipes: The Art of Scientific Computing*. New York, USA: Cambridge University Press, 3rd edn.

Pronzato, L. and Müller, W. G. (2012) Design of computer experiments: space filling and beyond. *Statistics and Computing*, **22**, 681–701.

Pronzato, L. and Walter, E. (1985) Robust experimental design via stochastic approximation. *Mathematical Biosciences*, **75**, 103–120.

Pukelsheim, F. and Torsney, B. (1991) Optimal weights for experimental designs on linearly independent support points. *The Annals of Statistics*, **19**, 1614–1625.

Raaijmaakers, J. G. W. (1987) Statistical analysis of the Michaelis-Menten equation. *Biometrics*, **43**, 793–803.

Raiffa, H. and Schlaifer, R. (1961) *Applied Statistical Decision Theory.* New York: Wiley.

Ranjan, P., Lu, W., Bingham, D., Reese, S., Williams, B. J., Chou, C. C., Doss, F., Grosskopf, M. and Holloway, J. P. (2011) Follow-up experimental design of computer models and physical processes. *Journal of Statistical Theory and Practice*, **5**, 119–136.

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

Reese, C., Wilson, A., Hamada, M., Martz, H. and Ryan, K. J. (2004) Integrated analysis of computer and physical experiments. *Technometrics*, **46**, 153–164.

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithm. *The Annals of Applied Probability*, **7**, 110–120.

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 319–392.

Ryan, E. G., Drovandi, C. C., McGree, J. M. and Pettitt, A. N. (2016) A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, **84**, 128–154.

Ryan, E. G., Drovandi, C. C. and Pettitt, A. N. (2015) Fully Bayesian experimental design for pharmacokinetic studies. *Entropy*, **17**, 1063–1089.

Ryan, K. J. (2003) Estimating expected information gains for experimental designs with application to the random Fatigue-Limit model. *Journal of Computational and Graphical Statistics*, **12**, 585–603.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989) Design and analysis of computer experiments. *Statistical Science*, **4**, 409–423.

San Martini, A. and Spezzaferri, F. (1984) A predictive model selection criterion. *Journal of the Royal Statistical Society: Series B*, **46**, 296–303.

Santner, T., Williams, B. J. and Notz, W. I. (2003) *The Design and Analysis of Computer Experiments.* New York: Springer.

Shannon, C. E. (1948) A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

Skilling, J. (2006) Nested sampling for general Bayesian computation. *Bayesian Analysis*, **1**, 833–859.

Song, D. and Wong, W. K. (1998) Optimal two-point designs for the Michaelis-Menten model with heteroscedastic errors. *Communications in Statistics-Theory and Methods*, **27**, 1503–1516.

Stigler, S. M. (1986) Laplace's 1774 memoir on inverse probability. *Statistical Science*, **1**, 359–363.

Storlie, C. B., William, L. A., Ryan, E. M., Gattiker, J. R. and Higdon, D. M. (2015) Calibration of computational models with categorical parameters and correlated outputs via Bayesian smoothing spline ANOVA. *Journal of the American Statistical Association*, **110**, 68–82.

Surjanovic, S. and Bingham, D. (2017) Virtual library of simulation experiments: test functions and datasets. URL https://www.sfu.ca/~ssurjano/canti.html.

Tang, B. (1993) Orthogonal array-based Latin hypercubes. *Journal of American Statistical Association*, **88**, 1392–1397.

Tuo, R. and Wu, J. C. F. (2015) Efficient calibration for imperfect computer models. *The Annals of Statistics*, **43**, 2331–2352.

— (2016) A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, **4**, 767–795.

Verdinelli, I., Polson, N. and Singpurwalla, N. (1993) Shannon information and Bayesian design for prediction in the accelerated life testing. In *Reliability and Decision Making* (eds. R. E. Barlow, C. A. Clarotti and F. Spizzichino), 247–256. London: Chapman and Hall/CRC.

Vernon, I., Goldstein, M. and Bower, R. G. (2010) Galaxy formation: a Bayesian uncertainty analysis. *Bayesian Analysis*, **5**, 619–670.

Wang, S., Chen, W. and Tsui, K. L. (2009) Bayesian validation of computer models. *Technometrics*, **51**, 439–451.

Weaver, B. P., Williams, B. J., Anderson-Cook, C. M. and Higdon, D. M. (2016) Computational enhancements to Bayesian design of experiments using Gaussian processes. *Bayesian Analysis*, **11**, 191–213.

Wilkinson, R. D. (2010) Bayesian calibration of expensive multivariate computer experiments. In *Large-Scale Inverse Problems and Quantification of Uncertainty* (eds.

L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, Y. Marzouk, L. Tenorio, B. B. Waanders and K. Willcox), 195–215. Chichester: Wiley.

— (2014) Accelerating ABC methods using Gaussian processes. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (eds. S. Kaski and J. Corander), vol. 33, 1015–1023. PMLR.

Williams, B., Higdon, D., Gattiker, J., Moore, L., McKay, M. and Keller-McNulty, S. (2006) Combining experimental data and computer simulations, with an application to flyer plate experiments. *Bayesian Analysis*, **1**, 765–792.

Williams, B., Loeppky, J., Moore, L. and Macklem, M. (2011) Batch sequential design to achieve predictive maturity with calibrated computer models. *Reliability Engineering and System Safety*, **96**, 1208–1219.

Wong, R. K. W., Storlie, C. B. and Lee, T. C. M. (2017) A frequentist approach to computer model calibration. *Journal of the Royal Statistical Society: Series B*, **79**, 635–648.

Woods, D. C., Lewis, S. M., Eccleston, J. A. and Russell, K. G. (2006) Designs for generalized linear models with several variables and model uncertainty. *Technometrics*, **48**, 284–292.

Woods, D. C., Overstall, A. M., Adamou, M. and Waite, T. W. (2017) Bayesian design of experiments for generalised linear models and dimensional analysis with industrial and scientific application. *Quality Engineering*, **29**, 91–103.

Wu, Y. T., Shin, Y., Sues, R. H. and Cesare, M. A. (2001) Safety-factor based approach for probability-based design optimization. In *Proc. 42nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, vol. 196, 199–342. Seattle, Washington.

Yang, Z. and Rodríguez, C. E. (2013) Searching for efficient Markov chain Monte Carlo proposal kernels. *Proceedings of the National Academy of Sciences*, **110**, 19307–19312.

Appendix A

A.1 Connection between Laplace approximation I (LA1) and Laplace approximation II (LA2)

In this section we illustrate a connection between Laplace Approximation I, discussed in Section 4.2.1, and Laplace Approximation II, discussed in Section 4.2.2. In particular, we show that LA2 can be derived from LA1 by using a Taylor approximation to the likelihood in addition to the normal approximation to the posterior density used in LA1.

Plugging the second order Taylor series expansion (4.23) of the log-likelihood, $\log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)$, about the posterior mode $\hat{\boldsymbol{\psi}}$, back into Equation (4.19) we get:

$$
\begin{aligned}
U(\xi) \approx \int_{\Psi}\int_{\mathcal{Y}} \Bigg[& \log \pi_l(\mathbf{y}|\hat{\boldsymbol{\psi}}, \xi) + \frac{\partial \log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \\
& + \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}} \left[-\mathbf{H}(\hat{\boldsymbol{\psi}}) - \mathbf{Q}(\hat{\boldsymbol{\psi}}) \right] (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \\
& - \log \pi_u(\hat{\boldsymbol{\psi}}|\mathbf{y}, \xi) - \frac{1}{2}\log\left[(2\pi)^{q_2} \left| \mathbf{H}(\hat{\boldsymbol{\psi}})^{-1} \right| \right] \Bigg] \pi(\mathbf{y}, \boldsymbol{\psi}|\xi) d\mathbf{y} d\boldsymbol{\psi} \\
= \int_{\Psi}\int_{\mathcal{Y}} \Bigg[& \log \pi_l(\mathbf{y}|\hat{\boldsymbol{\psi}}, \xi) + \frac{\partial \log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \\
& + \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}} \left[-\mathbf{H}(\hat{\boldsymbol{\psi}}) - \mathbf{Q}(\hat{\boldsymbol{\psi}}) \right] (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \\
& - [\log \pi_l(\mathbf{y}|\hat{\boldsymbol{\psi}}, \xi) + \log \pi_b(\hat{\boldsymbol{\psi}})] - \frac{1}{2}\log[(2\pi)^{q_2} |\mathbf{H}(\hat{\boldsymbol{\psi}})^{-1}|] \Bigg] \pi(\mathbf{y}, \boldsymbol{\psi}|\xi) d\mathbf{y} d\boldsymbol{\psi} \\
= \int_{\mathcal{Y}} \Bigg[& - \log \pi_b(\hat{\boldsymbol{\psi}}) - \frac{1}{2}\log\left[(2\pi)^{q_2} \left| \mathbf{H}(\hat{\boldsymbol{\psi}})^{-1} \right| \right] \\
& + \underbrace{\int_{\Psi} \frac{\partial \log \pi_l(\mathbf{y}|\boldsymbol{\psi}, \xi)}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi) d\boldsymbol{\psi}}_{I_6} \\
& + \underbrace{\int_{\boldsymbol{\Psi}} \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}} \left[-\mathbf{H}(\hat{\boldsymbol{\psi}}) - \mathbf{Q}(\hat{\boldsymbol{\psi}}) \right] (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \pi_a(\boldsymbol{\psi}|\mathbf{y}, \xi) d\boldsymbol{\psi}}_{I_7} \Bigg] \pi_e(\mathbf{y}|\xi) d\mathbf{y}.
\end{aligned}
$$

Firstly I_6 can be solved as:

$$\begin{aligned}
I_6 &= \int_\Psi \frac{\partial \log \pi_u(\boldsymbol{\psi}|\mathbf{y},\xi)}{\partial \boldsymbol{\psi}} \Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \pi_a(\mathbf{y}|\boldsymbol{\psi},\xi) d\boldsymbol{\psi} \\
&\quad - \underbrace{\int_\Psi \frac{\partial \log \pi_b(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \pi_a(\mathbf{y}|\boldsymbol{\psi},\xi) d\boldsymbol{\psi}}_{I_4} \\
&\approx 0,
\end{aligned}$$

because $\frac{\partial \log \pi_u(\boldsymbol{\psi}|\mathbf{y},\xi)}{\partial \boldsymbol{\psi}} \Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) = 0$ by definition, and $I_4 \approx 0$ as shown in (4.24).

Secondly we approximate I_7 as:

$$\begin{aligned}
I_7 &= -\int_\Psi \frac{1}{2} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}} \mathbf{H}(\hat{\boldsymbol{\psi}}) (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \pi_a(\mathbf{y}|\boldsymbol{\psi},\xi) d\boldsymbol{\psi} \\
&\quad - \int_\Psi \frac{1}{2} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^{\mathrm{T}} \mathbf{Q}(\hat{\boldsymbol{\psi}}) (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \pi_a(\mathbf{y}|\boldsymbol{\psi},\xi) d\boldsymbol{\psi} \\
&\approx -\frac{q_2}{2} - \frac{1}{2} \mathrm{tr} \left[\mathbf{Q}(\hat{\boldsymbol{\psi}}) \mathbf{H}(\hat{\boldsymbol{\psi}})^{-1} \right].
\end{aligned}$$

Hence Equation (4.19) becomes

$$U(\xi) \approx \int_{\mathcal{Y}} \left[-\frac{1}{2} \log(2\pi)^{q_2} \left| \mathbf{H}(\hat{\boldsymbol{\psi}})^{-1} \right| - \frac{q_2}{2} - \log \pi_b(\hat{\boldsymbol{\psi}}) - \frac{1}{2} \mathrm{tr} \left[\mathbf{Q}(\hat{\boldsymbol{\psi}}) \mathbf{H}(\hat{\boldsymbol{\psi}})^{-1} \right] \right] \pi_e(\mathbf{y}|\xi) d\mathbf{y},$$

which is identical to (4.20).

Both of these methods assume a normal approximation to the posterior density. The main difference between Laplace Approximation I and Laplace Approximation II is an additional second-order approximation to the log-likelihood (or equivalently to the log-prior density) assumed to hold over the region of highest posterior density. This requires that the posterior is quite highly concentrated around $\hat{\boldsymbol{\psi}}$, which will be the case for large n.

Appendix B

B.1 Fisher Information Matrix for the multivariate normal distribution

When $\mathbf{y} \sim N\left[\boldsymbol{\mu}(\boldsymbol{\psi}), \boldsymbol{\Sigma}(\boldsymbol{\psi})\right]$ follows a multivariate normal distribution, the Fisher information matrix for parameters $\boldsymbol{\psi}$ is given by Equation (6.3). When the mean and variance depend on parameter vectors $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, i.e. $\boldsymbol{\mu}(\boldsymbol{\psi}_1)$ and $\boldsymbol{\Sigma}(\boldsymbol{\psi}_2)$, then:

$$I(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2; \xi) = \operatorname{diag}\left[I(\boldsymbol{\psi}_1; \xi), I(\boldsymbol{\psi}_2; \xi)\right],$$

where

$$I(\boldsymbol{\psi}_1; \xi)_{i,j} = \frac{\partial \boldsymbol{\mu}(\boldsymbol{\psi}_1)^{\mathrm{T}}}{\partial \psi_{1_i}} \boldsymbol{\Sigma}(\boldsymbol{\psi}_2)^{-1} \frac{\partial \boldsymbol{\mu}(\boldsymbol{\psi}_1)}{\partial \psi_{1_j}}, \tag{B.1}$$

and

$$I(\boldsymbol{\psi}_2; \xi)_{i',j'} = \frac{1}{2} \operatorname{tr}\left[\boldsymbol{\Sigma}(\boldsymbol{\psi}_2)^{-1} \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\psi}_2)}{\partial \psi_{2i'}} \boldsymbol{\Sigma}(\boldsymbol{\psi}_2)^{-1} \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\psi}_2)}{\partial \psi_{2j'}}\right]. \tag{B.2}$$

B.2 Derivatives

In order to use LIS and ALIS to estimate the expected Shannon information gain, we need to calculate the first and second derivatives of the log-unnormalised posterior density of interest.

For most models we are interested in finding the derivatives with respect to $\boldsymbol{\psi}'$, a transformation of the original parameters $\boldsymbol{\psi}$. The most common transformation we use is $\boldsymbol{\psi}' = (\log \psi_1, \ldots, \log \psi_{q_2})^{\mathrm{T}}$ in order to ensure positive values for the parameters $\boldsymbol{\psi}$ when sampling from the importance density. We use the chain rule:

$$\frac{\partial \log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y}, \xi)}{\partial \boldsymbol{\psi}'} = \frac{\partial \log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y}, \xi)}{\partial \boldsymbol{\psi}} \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\psi}'}, \tag{B.3}$$

and

$$\frac{\partial^2 \log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y}, \xi)}{\partial \boldsymbol{\psi}'^{\mathrm{T}} \partial \boldsymbol{\psi}'} = \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\psi}'} \frac{\partial}{\partial \boldsymbol{\psi}} \left[\frac{\partial \log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y}, \xi)}{\partial \boldsymbol{\psi}} \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\psi}'}\right]$$

$$= \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\psi}'} \left[\frac{\partial^2 \log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y},\xi)}{\partial \boldsymbol{\psi}^{\mathrm{T}} \partial \boldsymbol{\psi}} \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\psi}'} + \frac{\partial \log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y},\xi)}{\partial \boldsymbol{\psi}} \frac{\partial^2 \boldsymbol{\psi}}{\partial \boldsymbol{\psi}^{\mathrm{T}} \partial \boldsymbol{\psi}'} \right]. \tag{B.4}$$

All the derivatives derived in this appendix have been checked numerically for a variety of different designs, data sets and parameter values.

B.2.1 Michaelis-Menten model

In this section we calculate the derivatives of the log-unnormalised posterior density, $\log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)$, of the Michaelis-Menten model given in Equation (5.1).

First derivatives:

$$\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1} = 2\left(a + \frac{n}{2}\right) \frac{\sum_{i=1}^n \left(\frac{x_i y_i}{\theta_2 + x_i} - \frac{\theta_1 x_i^2}{(\theta_2+x_i)^2}\right)}{2b + \sum_{i=1}^n \left(y_i - \frac{\theta_1 x_i}{\theta_2 + x_i}\right)^2} - \frac{1}{\theta_1} - \frac{\log\theta_1 - \mu_1}{\sigma_1^2 \theta_1},$$

$$\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_2} = -2\left(a + \frac{n}{2}\right) \frac{\sum_{i=1}^n \left(\frac{\theta_1 x_i y_i}{(\theta_2 + x_i)^2} - \frac{\theta_1^2 x_i^2}{(\theta_2+x_i)^3}\right)}{2b + \sum_{i=1}^n \left(y_i - \frac{\theta_1 x_i}{\theta_2 + x_i}\right)^2} - \frac{1}{\theta_2} - \frac{\log\theta_2 - \mu_2}{\sigma_2^2 \theta_2}.$$

Second derivatives:

$$\begin{aligned}\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1^2} = 2\left(a + \frac{n}{2}\right) &\left\{ \frac{\sum_{i=1}^n \left(-\frac{x_i^2}{(\theta_2+x_i)^2}\right)\left[2b + \sum_{i=1}^n \left(y_i - \frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right]}{\left[2b + \sum_{i=1}^n \left(y_i - \frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right]^2} \right. \\ &\left. - \frac{2\left[\sum_{i=1}^n \left(\frac{x_i y_i}{\theta_2+x_i} - \frac{\theta_1 x_i^2}{(\theta_2+x_i)^2}\right)\right]^2}{\left[2b + \sum_{i=1}^n \left(y_i - \frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right]^2} \right\} \\ &+ \frac{1}{\theta_1^2} - \frac{1 - \log\theta_1 + \mu_1}{\theta_1^2 \sigma_1^2},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1 \partial \theta_2} = 2\left(a + \frac{n}{2}\right) &\left\{ \frac{\sum_{i=1}^n \left(-\frac{x_i y_i}{(\theta_2+x_i)^2} + \frac{2\theta_1 x_i^2}{(\theta_2+x_i)^3}\right)[2b + \sum_{i=1}^n \left(y_i - \frac{\theta_1 x_i}{\theta_2+x_i}\right)^2]}{\left[2b + \sum_{i=1}^n \left(y_i - \frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right]^2} \right. \\ &\left. - \frac{2\sum_{i=1}^n \left(\frac{x_i y_i}{\theta_2+x_i} - \frac{\theta_1 x_i^2}{(\theta_2+x_i)^2}\right)\sum_{i=1}^n \left(\frac{\theta_1 x_i y_i}{(\theta_2+x_i)^2} - \frac{\theta_1^2 x_i^2}{(\theta_2+x_i)^3}\right)}{\left[2b + \sum_{i=1}^n \left(y_i - \frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right]^2} \right\},\end{aligned}$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_2^2} = -2\left(a+\frac{n}{2}\right)\left\{\frac{\sum_{i=1}^n \left(-\frac{2\theta_1 x_i y_i}{(\theta_2+x_i)^3}+\frac{3\theta_1^2 x_i^2}{(\theta_2+x_i)^4}\right)\left[2b+\sum_{i=1}^n\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right]}{\left[2b+\sum_{i=1}^n\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right]^2}\right.$$

$$\left.-\frac{2\left[\sum_{i=1}^n\left(\frac{\theta_1 x_i y_i}{(\theta_2+x_i)^2}-\frac{\theta_1^2 x_i^2}{(\theta_2+x_i)^3}\right)\right]^2}{\left[2b+\sum_{i=1}^n\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right]^2}\right\}+\frac{1}{\theta_2^2}-\frac{1-\log\theta_2+\mu_2}{\theta_2^2\sigma_2^2},$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1 \partial \theta_2} = \frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_2 \partial \theta_1}.$$

Using Equations (B.3), (B.4) and (4.36) we obtain the derivatives of the log-unnormalised posterior density $\log \pi_u^{\theta'}(\boldsymbol{\theta}'|\mathbf{y},\xi)$, with respect to $\boldsymbol{\theta}' = (\log\theta_1, \log\theta_2)^{\mathrm{T}}$.

As described in Section 4.3.2 we are required to derive the implied importance density for the untransformed parameters $\boldsymbol{\theta}$. We have that the importance density of the transformed parameters $\boldsymbol{\theta}'$, $q^h_{\boldsymbol{\theta}'}(\boldsymbol{\theta}')$, is a normal density with mean $\hat{\boldsymbol{\mu}}^h_{\boldsymbol{\theta}'}$, defined using Equations (4.37) and (4.38) for LIS and ALIS respectively, and variance $\hat{\boldsymbol{\Sigma}}^h_{\boldsymbol{\theta}'}$ defined using Equation (4.39). Hence,

$$q^h_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = q^h_{\boldsymbol{\theta}'}(T(\boldsymbol{\theta}))\left|\det \mathcal{J}[T(\boldsymbol{\theta})]\right|,$$

where

$$\mathcal{J}[T(\boldsymbol{\theta})] = \begin{bmatrix} \frac{1}{\theta_1} & 0 \\ 0 & \frac{1}{\theta_2} \end{bmatrix}, \tag{B.5}$$

is the Jacobian matrix of the transformation from $\boldsymbol{\theta}'$ to $\boldsymbol{\theta}$.

B.2.2 Biochemical Oxygen Demand (BOD) model

We will now calculate the derivatives of the log-unnormalised posterior density $\log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)$ for the BOD model given in Equation (5.3).

First derivatives:

$$\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1} = -\frac{n\sum_{i=1}^n\left[y_i-\theta_1(1-\exp\{-\theta_2 x_i\})\right]\left[-(1-\exp\{-\theta_2 x_i\})\right]}{\sum_{i=1}^n\left[y_i-\theta_1(1-\exp\{-\theta_2 x_i\})\right]^2} - \frac{1}{\theta_1}-\frac{\log\theta_1-\mu_1}{\sigma_1^2\theta_1},$$

$$\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_2} = -\frac{n\sum_{i=1}^n\left[y_i-\theta_1(1-\exp\{-\theta_2 x_i\})\right]\left[-\theta_1 x_i\exp\{-\theta_2 x_i\}\right]}{\sum_{i=1}^n\left[y_i-\theta_1(1-\exp\{-\theta_2 x_i\})\right]^2} - \frac{1}{\theta_2}-\frac{\log\theta_2-\mu_2}{\sigma_2^2\theta_2}.$$

Second derivatives:

$$\begin{aligned}
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_1^2} &= \frac{2n\left[\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]\,[-1+\exp\{-\theta_2x_i\}]\right]^2}{\left[\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]^2\right]^2} \\
&\quad -\frac{n\sum_{i=1}^n [-1+\exp\{-\theta_2x_i\}]^2}{\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]^2}+\frac{1}{\theta_1^2}-\frac{1-\log\theta_1+\mu_1}{\theta_1^2\sigma_1^2}, \\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2^2} &= \frac{2n\left[\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]\,[-\theta_1x_i\exp\{-\theta_2x_i\}]\right]^2}{\left[\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]^2\right]^2} \\
&\quad -\frac{n\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]\,\left[\theta_1x_i^2\exp\{-\theta_2x_i\}\right]}{\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]^2} \\
&\quad -\frac{n\left[\sum_{i=1}^n [-\theta_1x_i\exp\{-\theta_2x_i\}]\right]^2}{\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]^2}+\frac{1}{\theta_2^2}-\frac{1-\log\theta_2+\mu_2}{\theta_2^2\sigma_2^2}, \\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_1\partial\theta_2} &= \frac{2n\left[\sum_{i=1}^n [-\theta_1x_i\exp\{-\theta_2x_i\}]\,[y_i-\theta_1(1-\exp\{-\theta_2x_i\})]\right]^2}{\left[\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]^2\right]^2} \\
&\quad -\frac{n\sum_{i=1}^n [-\theta_1x_i\exp\{-\theta_2x_i\}]\,[-1+\exp\{-\theta_2x_i\}]}{\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]^2} \\
&\quad -\frac{n\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]\,[-x_i\exp\{-\theta_2x_i\}]}{\sum_{i=1}^n [y_i-\theta_1(1-\exp\{-\theta_2x_i\})]^2}, \\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_1\partial\theta_2} &= \frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2\partial\theta_1}.
\end{aligned}$$

Using Equations (B.3), (B.4) and (4.36) we obtain the derivatives of the log-unnormalised posterior density $\log\pi_u^{\theta'}(\boldsymbol{\theta}'|\mathbf{y},\xi)$, with respect to $\boldsymbol{\theta}'=(\log\theta_1,\log\theta_2)^{\mathrm{T}}$.

Again, we have to work out the implied importance density for the untransformed parameters $\boldsymbol{\theta}$, as described in Section 4.3.2 and as demonstrated in the previous example. The Jacobian matrix is given by Equation (B.5).

B.2.3 Lubricant model

We now calculate the derivatives of the log-unnormalised posterior density $\log\pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)$ for the lubricant model given in Equation (5.5).

For simplicity we denote $\eta_i=\eta(x_{1i},x_{2i},\boldsymbol{\theta})=\frac{\theta_1}{\theta_2+x_{1i}}+\theta_3x_{2i}+\theta_4x_{2i}^2+\theta_5x_{2i}^3+(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}$, and we assume $\theta_{10}=\log\sigma_\varepsilon^2$.

First derivatives:

$$\begin{aligned}
\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_1} &= \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^n(y_i-\eta_i)\left(\frac{1}{\theta_2+x_{2i}}\right)-\frac{1}{\sigma_1^2}(\theta_1-\mu_1), \\
\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^n(y_i-\eta_i)\left(\frac{\theta_1}{(\theta_2+x_{2i})^2}\right)-\frac{1}{\sigma_2^2}(\theta_2-\mu_2),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_3} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\,(-x_{2i}) - \frac{1}{\sigma_3^2}(\theta_3-\mu_3),\\
\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_4} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\,(-x_{2i}^2) - \frac{1}{\sigma_4^2}(\theta_4-\mu_4),\\
\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_5} &= -\frac{1}{\sigma_\varepsilon^2 2}\sum_{i=1}^{n}(y_i-\eta_i)\,(-x_{2i}^3) - \frac{1}{\sigma_5^2}(\theta_5-\mu_5),\\
\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_6} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\right) - \frac{1}{\sigma_6^2}(\theta_6-\mu_6),\\
\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_7} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\right) - \frac{1}{\sigma_7^2}(\theta_7-\mu_7),\\
\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_8} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-[\theta_6+\theta_7x_{2i}^2]x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\left\{\frac{x_{1i}}{(\theta_8+\theta_9x_{2i}^2)^2}\right\}\right)\\
&\qquad - \frac{1}{\sigma_8^2}(\theta_8-\mu_8),\\
\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_9} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-[\theta_6+\theta_7x_{2i}^2]x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\left\{\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9x_{2i}^2)^2}\right\}\right)\\
&\qquad - \frac{1}{\sigma_9^2}(\theta_9-\mu_9),\\
\frac{\partial \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_{10}} &= -n + \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i) - \frac{1}{\sigma_{10}^2}(\theta_{10}-\mu_{10}).
\end{aligned}$$

Second derivatives:

$$\begin{aligned}
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1^2} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left[-\frac{1}{\theta_2+x_{2i}}\right]^2 - \frac{1}{\sigma_1^2},\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1\partial\theta_2} &= \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left[\frac{\theta_1}{(\theta_2+x_{2i})^2}\right]\left[\frac{1}{\theta_2+x_{2i}}\right] + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\eta_i)\left[-\frac{1}{(\theta_2+x_{2i})^2}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1\partial\theta_3} &= \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i})\left[\frac{1}{\theta_2+x_{2i}}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1\partial\theta_4} &= \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i}^2)\left[\frac{1}{\theta_2+x_{2i}}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1\partial\theta_5} &= \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i}^3)\left[\frac{1}{\theta_2+x_{2i}}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1\partial\theta_6} &= \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\right)\left[\frac{1}{\theta_2+x_{2i}}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1\partial\theta_7} &= \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\right)\left[\frac{1}{\theta_2+x_{2i}}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_1\partial\theta_8} &= \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\frac{x_{1i}}{(\theta_8+\theta_9x_{2i}^2)^2}\right)\left[\frac{1}{\theta_2+x_{2i}}\right],
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_1\partial\theta_9} &= \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)\left[\frac{1}{\theta_2+x_{2i}}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_1\partial\theta_{10}} &= -\frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left[\frac{1}{\theta_2+x_{2i}}\right],
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2^2} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left[\frac{\theta_1}{(\theta_2+x_{1i})^2}\right]^2-\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\eta_i)\left[-\frac{2\theta_1}{(\theta_2+x_{2i})^2}\right]-\frac{1}{\sigma_2^2},\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2\partial\theta_3} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i})\left[\frac{\theta_1}{(\theta_2+x_{2i})^2}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2\partial\theta_4} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i}^2)\left[\frac{\theta_1}{(\theta_2+x_{2i})^2}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2\partial\theta_5} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i}^3)\left[\frac{\theta_1}{(\theta_2+x_{2i})^2}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2\partial\theta_6} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)\left[\frac{\theta_1}{(\theta_2+x_{2i})^2}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2\partial\theta_7} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)\left[\frac{\theta_1}{(\theta_2+x_{2i})^2}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2\partial\theta_8} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)\left[\frac{\theta_1}{(\theta_2+x_{2i})^2}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2\partial\theta_9} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)\left[\frac{\theta_1}{(\theta_2+x_{2i})^2}\right],\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_2\partial\theta_{10}} &= \frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left[\frac{\theta_1}{(\theta_2+x_{2i})^2}\right],
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_3^2} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i})^2-\frac{1}{\sigma_3^2},\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_3\partial\theta_4} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i}^2)(-x_{2i}),\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_3\partial\theta_5} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i}^3)(-x_{2i}),\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_3\partial\theta_6} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)(-x_{2i}),\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_3\partial\theta_7} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)(-x_{2i}),
\end{aligned}
$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_3\partial\theta_8} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)(-x_{2i}),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_3\partial\theta_9} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)(-x_{2i}),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_3\partial\theta_{10}} = \frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)(-x_{2i}),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_4^2} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i}^2)^2-\frac{1}{\sigma_4^2},$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_4\partial\theta_5} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i}^3)(-x_{2i}^2),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_4\partial\theta_6} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)(-x_{2i}^2),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_4\partial\theta_7} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)(-x_{2i}^2),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_4\partial\theta_8} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)(-x_{2i}^2),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_4\partial\theta_9} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)(-x_{2i}^2),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_4\partial\theta_{10}} = \frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)(-x_{2i}^2),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_5^2} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(-x_{2i}^3)^2-\frac{1}{\sigma_5^2},$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_5\partial\theta_6} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)(-x_{2i}^3),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_5\partial\theta_7} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)(-x_{2i}^3),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_5\partial\theta_8} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)(-x_{2i}^3),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_5\partial\theta_9} = -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)(-x_{2i}^3),$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_5\partial\theta_{10}} = \frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)(-x_{2i}^3),$$

$$
\begin{aligned}
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_6^2} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)^2 - \frac{1}{\sigma_6^2},\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_6 \partial \theta_7} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)\left\{-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right),\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_6 \partial \theta_8} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)\\
&\qquad\qquad \times\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)\\
&\quad -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}}{(\theta_8+\theta_9 x_{2i}^2)^2}\right),\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_6 \partial \theta_9} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)\\
&\qquad\qquad \times\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)\\
&\quad -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9 x_{2i}^2)^2}\right),\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_6 \partial \theta_{10}} &= \frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right),
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_7^2} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)^2 - \frac{1}{\sigma_7^2},\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_7 \partial \theta_8} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)\\
&\qquad\qquad \times\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)\\
&\quad -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}}{(\theta_8+\theta_9 x_{2i}^2)^2}\right),\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_7 \partial \theta_9} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7 x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9 x_{2i}^2)^2}\right)\\
&\qquad\qquad \times\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\right)\\
&\quad -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9 x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9 x_{2i}^2)^2}\right),
\end{aligned}
$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_7\partial\theta_{10}} = \frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-x_{2i}^3\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\right),$$

$$\begin{aligned}
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_8^2} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\frac{x_{1i}}{(\theta_8+\theta_9x_{2i}^2)^2}\right)^2,\\
&- \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\left[\frac{x_{1i}}{(\theta_8+\theta_9x_{2i}^2)^2}\right]^2\right)\\
&- \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\left[-\frac{2x_{1i}}{(\theta_8+\theta_9x_{2i}^2)^3}\right]\right)\\
&- \frac{1}{\sigma_8^2},\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_8\partial\theta_9} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9x_{2i}^2)^2}\right)^2\\
&- \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\right.\\
&\qquad\qquad\left.\times\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9x_{2i}^2)^2}\frac{x_{1i}}{(\theta_8+\theta_9x_{2i}^2)^2}\right)\\
&- \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\frac{-2x_{1i}x_{2i}^2}{(\theta_8+\theta_9x_{2i}^2)^3}\right),\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_8\partial\theta_{10}} &= -\frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\frac{x_{1i}}{(\theta_8+\theta_9x_{2i}^2)^2}\right),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_9^2} &= -\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9x_{2i}^2)^2}\right)^2\\
&- \frac{1}{\sigma_\varepsilon^2 2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\left[\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9x_{2i}^2)^2}\right]^2\right)\\
&- \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\left[-\frac{2x_{1i}x_{2i}^3}{(\theta_8+\theta_9x_{2i}^2)^3}\right]\right)\\
&- \frac{1}{\sigma_9^2},\\
\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial\theta_9\partial\theta_{10}} &= -\frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)\left(-(\theta_6+\theta_7x_{2i}^2)x_{2i}\exp\left\{-\frac{x_{1i}}{\theta_8+\theta_9x_{2i}^2}\right\}\frac{x_{1i}x_{2i}^2}{(\theta_8+\theta_9x_{2i}^2)^2}\right),
\end{aligned}$$

$$\frac{\partial^2 \log \pi_u(\boldsymbol{\theta}|\mathbf{y},\xi)}{\partial \theta_{10}^2} = \frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\eta_i)^2 - \frac{1}{\sigma_{10}^2}.$$

B.2.4 Calibration model: Michaelis-Menten simulator with $\delta_{\theta^p}(x) \neq 0$

We now calculate the derivatives of the log-unnormalised posterior density $\log \pi_u(\boldsymbol{\psi}|\mathbf{y},\xi)$ for the calibration model given in Section 6.3.1, where the simulator is the Michaelis-Menten model and there is non-zero discrepancy, i.e. $\delta_{\boldsymbol{\theta}^p}(x) \neq 0$.

The first derivatives of the log-unnormalised posterior density are:

$$\begin{aligned}
\frac{\partial \log \pi_u(\boldsymbol{\psi}|\mathbf{y},\xi)}{\partial \theta_1^p} &= \frac{1}{\sigma^2}\left(\frac{\partial \eta(\mathbf{x},\boldsymbol{\theta})}{\partial \theta_1^p}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}[\mathbf{y}-\eta(\mathbf{x},\boldsymbol{\theta})] - \frac{1}{\theta_1^p} - \frac{\log\theta_1^p - \mu_1}{\sigma_1^2\theta_1^p},\\
\frac{\partial \log \pi_u(\boldsymbol{\psi}|\mathbf{y},\xi)}{\partial \theta_2^p} &= \frac{1}{\sigma^2}\left(\frac{\partial \eta(\mathbf{x},\boldsymbol{\theta})}{\partial \theta_2^p}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}[\mathbf{y}-\eta(\mathbf{x},\boldsymbol{\theta})] - \frac{1}{\theta_2^p} - \frac{\log\theta_2^p - \mu_2}{\sigma_2^2\theta_2^p},\\
\frac{\partial \log \pi_u(\boldsymbol{\psi}|\mathbf{y},\xi)}{\partial \sigma^2} &= -\frac{n}{\sigma^2} + \frac{1}{2(\sigma^2)^2}[\mathbf{y}-\eta(\mathbf{x},\boldsymbol{\theta})]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}[\mathbf{y}-\eta(\mathbf{x},\boldsymbol{\theta})] - \frac{(a+1)\sigma^2+b}{(\sigma^2)^2},\\
\frac{\partial \log \pi_u(\boldsymbol{\psi}|\mathbf{y},\xi)}{\partial \phi} &= -\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi}\right] - \frac{1}{2\sigma^2}[\mathbf{y}-\eta(\mathbf{x},\boldsymbol{\theta})]^{\mathrm{T}}\left[-\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi}\boldsymbol{\Sigma}^{-1}\right][\mathbf{y}-\eta(\mathbf{x},\boldsymbol{\theta})]\\
&\quad - \lambda_\phi,\\
\frac{\partial \log \pi_u(\boldsymbol{\psi}|\mathbf{y},\xi)}{\partial \tau^2} &= -\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \tau^2}\right] - \frac{1}{2\sigma^2}[\mathbf{y}-\eta(\mathbf{x},\boldsymbol{\theta})]^{\mathrm{T}}\left[-\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \tau^2}\boldsymbol{\Sigma}^{-1}\right][\mathbf{y}-\eta(\mathbf{x},\boldsymbol{\theta})]\\
&\quad - \lambda_{\tau^2},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial \eta(\mathbf{x},\boldsymbol{\theta})}{\partial \theta_1^p} &= \left[\frac{x_1}{\theta_2^p + x_1},\ldots,\frac{x_n}{\theta_2^p + x_n}\right]^{\mathrm{T}},\\
\frac{\partial \eta(\mathbf{x},\boldsymbol{\theta})}{\partial \theta_2^p} &= \left[-\frac{\theta_1^p x_1}{(\theta_2^p + x_1)^2},\ldots,-\frac{\theta_1^p x_n}{(\theta_2^p + x_n)^2}\right]^{\mathrm{T}},\\
\frac{\partial \boldsymbol{\Sigma}}{\partial \phi}ij &= -(x_i-x_j)^2\exp[-\phi(x_i-x_j)^2] = -(x_i-x_j)^2\mathbf{K}(\phi)_{ij},\\
\frac{\partial \boldsymbol{\Sigma}}{\partial \tau^2} &= \mathbf{I}_n.
\end{aligned}$$

Using Equations (B.3), (B.4) and (4.36) we obtain the derivatives of the log-unnormalised posterior density $\log \pi_u^{\psi'}(\boldsymbol{\psi}'|\mathbf{y},\xi)$, with respect to $\boldsymbol{\psi}' = (\log\theta_1^p, \log\theta_2^p, \log\sigma^2, \log\phi, \log\tau^2)^{\mathrm{T}}$.

The second derivatives of the log-likelihood can now be easily calculated using Equation (6.3). In this case, the mean and variance depend on different vectors of parameters and hence we have the special case described in Section B.1. The information matrix for $\boldsymbol{\psi} = (\theta_1^p, \theta_2^p, \sigma^2, \phi, \tau^2)^{\mathrm{T}}$ has the form

$$I(\boldsymbol{\psi};\xi) = \begin{bmatrix} I_{11} & I_{12} & 0 & 0 & 0 \\ I_{21} & I_{22} & 0 & 0 & 0 \\ 0 & 0 & I_{33} & I_{34} & I_{35} \\ 0 & 0 & I_{43} & I_{44} & I_{45} \\ 0 & 0 & I_{53} & I_{54} & I_{55} \end{bmatrix}, \tag{B.6}$$

where

$$\begin{aligned}
I_{11} &= \frac{\partial \boldsymbol{\eta}}{\partial \theta_1^p}^{\mathrm{T}} \frac{\boldsymbol{\Sigma}^{-1}}{\sigma^2} \frac{\partial \boldsymbol{\eta}}{\partial \theta_1^p}, \\
I_{12} = I_{21} &= \frac{\partial \boldsymbol{\eta}}{\partial \theta_1^p}^{\mathrm{T}} \frac{\boldsymbol{\Sigma}^{-1}}{\sigma^2} \frac{\partial \boldsymbol{\eta}}{\partial \theta_2^p}, \\
I_{22} &= \frac{\partial \boldsymbol{\eta}}{\partial \theta_2^p}^{\mathrm{T}} \frac{\boldsymbol{\Sigma}^{-1}}{\sigma^2} \frac{\partial \boldsymbol{\eta}}{\partial \theta_2^p}, \\
I_{33} &= \frac{1}{2}\mathrm{tr}[\mathbf{I}_n], \\
I_{34} = I_{43} &= \frac{1}{2}\mathrm{tr}[\phi \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi}], \\
I_{35} = I_{53} &= \frac{1}{2}\mathrm{tr}[\tau^2 \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \tau^2}], \\
I_{44} &= \frac{1}{2}\mathrm{tr}[\phi^2 \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi}], \\
I_{45} = I_{54} &= \frac{1}{2}\mathrm{tr}[\boldsymbol{\Sigma}^{-1} \phi \frac{\partial \boldsymbol{\Sigma}}{\partial \phi} \boldsymbol{\Sigma}^{-1} \tau^2 \frac{\partial \boldsymbol{\Sigma}}{\partial \tau^2}], \\
I_{55} &= \frac{1}{2}\mathrm{tr}[(\tau^2)^2 \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \tau^2} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \tau^2}].
\end{aligned}$$

The second derivatives of the log-prior density are:

$$\begin{aligned}
\frac{\partial^2 \log \pi_b(\theta_1^p)}{\partial \theta_1^{p2}} &= -\frac{1}{\theta_1^p \sigma_1^2}, \\
\frac{\partial^2 \log \pi_b(\theta_2^p)}{\partial \theta_2^{p2}} &= -\frac{1}{\theta_2^p \sigma_2^2}, \\
\frac{\partial^2 \log \pi_b(\sigma^2)}{\partial [\sigma^2]^2} &= -\frac{b}{(\sigma^2)^2}, \\
\frac{\partial^2 \log \pi_b(\phi)}{\partial \phi^2} &= -\lambda_\phi, \\
\frac{\partial^2 \log \pi_b(\tau^2)}{\partial [\tau^2]^2} &= -\lambda_{\tau^2},
\end{aligned}$$

and, again, using (B.3), (B.4) and (4.36) we obtain the derivatives with respect to $\boldsymbol{\psi}'$.

As described in Section 4.3.2 and demonstrated in previous examples, we must derive the implied importance density for the untransformed parameters $\boldsymbol{\psi}$. We have that the importance density of the transformed parameters $\boldsymbol{\psi}'$, $q^h_{\boldsymbol{\psi}'}(\boldsymbol{\psi}')$, is a normal density with mean $\hat{\boldsymbol{\mu}}^h_{\boldsymbol{\psi}'}$, defined using Equations (4.37) and (4.38) for LIS and ALIS respectively, and variance $\boldsymbol{\Sigma}^h_{\boldsymbol{\psi}'}$ defined using Equation (4.39). Hence,

$$q^h_{\boldsymbol{\psi}}(\boldsymbol{\psi}) = q^h_{\boldsymbol{\psi}'}(T(\boldsymbol{\psi}))\left|\det \mathcal{J}[T(\boldsymbol{\psi})]\right|,$$

where

$$\mathcal{J}[T(\boldsymbol{\psi})] = \begin{bmatrix} \frac{1}{\theta^p_1} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\theta^p_2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sigma^2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\phi} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\tau^2} \end{bmatrix}, \tag{B.7}$$

is the Jacobian matrix.

B.2.5 Unknown simulator and $\delta_{\theta^p}(x) = 0$ - Cantilever Beam function

Results required to find Bayesian optimal designs for a known cantilever beam simulator

We first assume that the simulator $\eta(\mathbf{x}, \boldsymbol{\theta})$ is known and is given by the cantilever beam function. We have

$$y_i = \frac{4L^3}{\theta w t}\sqrt{\left(\frac{x_{1i}}{t^2}\right)^2 + \left(\frac{x_{2i}}{w^2}\right)^2} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$. We assume a normal distribution for the unknown parameter $\theta \sim N(\mu_1, \sigma^2_1)$ and a conjugate inverse-gamma prior distribution for $\sigma^2_\varepsilon \sim \mathrm{IG}(a, b)$.

The likelihood function is given by:

$$\pi_l(\mathbf{y}|\theta, \sigma^2_\varepsilon, \xi) = \frac{1}{(2\pi\sigma^2_\varepsilon)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2_\varepsilon}(\mathbf{y} - \boldsymbol{\eta})^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\eta})\right\},$$

where $\boldsymbol{\eta} = [\eta(\mathbf{x}_1, \theta) \ldots \eta(\mathbf{x}_n, \theta)]^{\mathrm{T}}$ and $\eta(\mathbf{x}_i, \theta) = \frac{4L^3}{\theta w t}\sqrt{\left(\frac{x_{1i}}{t^2}\right)^2 + \left(\frac{x_{2i}}{w^2}\right)^2}$.

We integrate out σ^2_ε to obtain the marginal likelihood:

$$\begin{aligned} \pi(\mathbf{y}|\theta, \xi) &= \int_0^\infty \pi_l(\mathbf{y}|\theta, \sigma^2_\varepsilon)\pi_b(\sigma^2_\varepsilon)d\sigma^2_\varepsilon \\ &= \int_0^\infty (2\pi\sigma^2_\varepsilon)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2_\varepsilon}\left[(\mathbf{y} - \boldsymbol{\eta})^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\eta})\right]\right\}(\sigma^2_\varepsilon)^{-(a+1)} \exp\{-b\sigma_\varepsilon^{-2}\}d\sigma^2_\varepsilon \\ &\propto \left[1 + \frac{(\mathbf{y} - \boldsymbol{\eta})^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\eta})}{2b}\right]^{-(a+\frac{n}{2})}. \end{aligned}$$

The log-unnormalised marginal posterior density is then given by:

$$\begin{aligned}\log \pi_u(\theta|\mathbf{y},\xi) = &-\left(a+\frac{n}{2}\right)\log\left[2b+(\mathbf{y}-\boldsymbol{\eta})^{\mathrm{T}}(\mathbf{y}-\boldsymbol{\eta})\right]+\left(a+\frac{n}{2}\right)\log[2b]\\ &-\frac{1}{2}\log[2\pi\sigma_1^2]-\frac{1}{2\sigma_1^2}(\theta-\mu_1)^2.\end{aligned}$$

We estimate the expected Shannon information gain using ALIS and LIS approximations which are then combined with the ACE algorithm to find Bayesian optimal designs.

We take a normal approximation to the posterior distribution of θ as described in Section 4.3.2. To calculate the negative Hessian, $\mathbf{H}(\theta)$, of the log-unnormalised posterior density required by the ALIS and LIS approximations (see Section 4.3), we first have to find the derivatives of the log-unnormalised posterior density $\log \pi_u(\theta|\mathbf{y},\xi)$ with respect to θ. The derivatives are given below.

For convenience we assume

$$A_i = \left(\frac{x_{1i}}{t^2}\right)^2 + \left(\frac{x_{2i}}{w^2}\right)^2.$$

$$\begin{aligned}\frac{\partial \log \pi_u(\theta|\mathbf{y},\xi)}{\partial\theta} &= -2\left(a+\frac{n}{2}\right)\frac{\sum_{i=1}^n\left(y_i\frac{4L^3}{\theta^2wt}\sqrt{A_i}-\frac{16L^6}{\theta^3w^2t^2}A_i\right)}{2b+\sum_{i=1}^n\left(y_i-\frac{4L^3}{\theta wt}\sqrt{A_i}\right)^2}-\frac{\theta-\mu_1}{\sigma_1^2},\\ \frac{\partial^2 \log \pi_u(\theta|\mathbf{y},\xi)}{\partial\theta^2} &= -2\left(a+\frac{n}{2}\right)\\ &\times\left\{\frac{\sum_{i=1}^n\left[-y_i\frac{8L^3}{\theta^3wt}\sqrt{A_i}+\frac{48L^6}{\theta^4w^2t^2}A_i\right]\left[2b+\sum_{i=1}^n\left(y_i-\frac{4L^3}{\theta wt}\sqrt{A_i}\right)^2\right]}{\left[2b+\sum_{i=1}^n\left(y_i-\frac{4L^3}{\theta wt}\sqrt{A_i}\right)^2\right]^2}\right.\\ &\left.-\frac{2\left[\sum_{i=1}^n\left(y_i\frac{4L^3}{\theta^2wt}\sqrt{A_i}-\frac{16L^6}{\theta^3w^2t^2}A_i\right)\right]\left[\sum_{i=1}^n\left(y_i\frac{4L^3}{\theta^2wt}\sqrt{A_i}-\frac{16L^6}{\theta^3w^2t^2}A_i\right)\right]}{\left[2b+\sum_{i=1}^n\left(y_i-\frac{4L^3}{\theta wt}\sqrt{A_i}\right)^2\right]^2}\right\}\\ &-\frac{1}{\sigma_1^2}.\end{aligned}$$

Results required to find Bayesian designs when the simulator is unknown

Now, we assume $\eta(\mathbf{x},\boldsymbol{\theta})$ is unknown and we assume the calibration model given in Equation (6.4) and a Gaussian process prior for the unknown simulator as described in the example in Section 6.4.3. We calculate the derivatives of the log-prior density, $\log \pi_b(\boldsymbol{\psi})$ with respect to the unknown parameters $\boldsymbol{\psi}'$, where $\boldsymbol{\psi}' = (\theta^p, \log\tau^2)^{\mathrm{T}}$. We have that:

$$\theta^p \sim \log N(\mu_1,\sigma_1^2),\ \ \tau^2 \sim \mathrm{Exp}(\lambda_{\tau^2}).$$

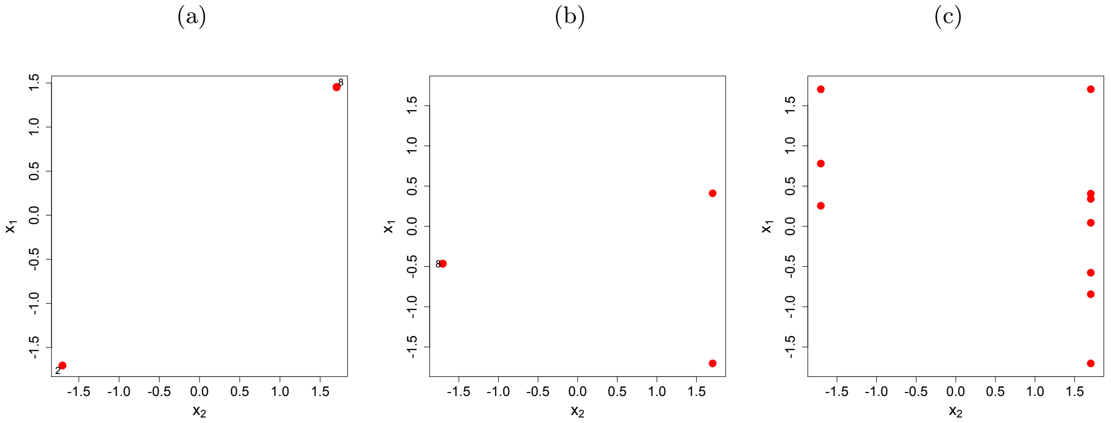


Figure B.1: Cantilever beam example: Bayesian near-optimal-ESIG designs from different random starts of ACE for the nonlinear model when σ_ε^2 is treated as a nuisance parameter (the numbers on some points show how many times the point is repeated)

The log-prior density is:

$$\begin{aligned}\log \pi_b(\boldsymbol{\psi}) &= \log \pi_b(\theta^p) + \log \pi_b(\tau^2)\\ &= \log\left[\theta^p \sigma_1 (2\pi)^{1/2}\right] + \frac{(\log \theta^p - \mu_1)^2}{2\sigma_1^2} + \log \lambda_{\tau^2} + \lambda_{\tau^2}\tau^2.\end{aligned}$$

First Derivatives:

$$\begin{aligned}\frac{\partial \log \pi_b(\theta_1^p)}{\partial \theta^p} &= -\frac{\theta^p - \mu_1}{\sigma_1^2},\\ \frac{\partial \log \pi_b(\tau^2)}{\partial \log \tau^2} &= -\lambda_{\tau^2}\tau^2.\end{aligned}$$

Second derivatives:

$$\begin{aligned}\frac{\partial^2 \log \pi_b(\theta_1^p)}{\partial [\theta^p]^2} &= -\frac{1}{\sigma_1^2},\\ \frac{\partial^2 \log \pi_b(\tau^2)}{\partial [\log \tau^2]^2} &= -\lambda_{\tau^2}\tau^2.\end{aligned}$$

In Section 6.4.3 we found Bayesian optimal designs by combining LIS with the ACE algorithm where we used 10 random starts. We presented optimal designs for the physical experiment with $n = 10$ for both the nonlinear model when σ_ε^2 is treated as a nuisance parameter and the calibration model when τ^2 is treated as a nuisance parameter with different numbers of simulator runs ($m = 30, 60, 90$). Here we present near-optimal designs from different random starts of ACE, see Figures B.1-B.4. For each model and size of computer experiment, we see a wide variety of different designs. However, they all have similar estimated ESIG, see Table B.1.

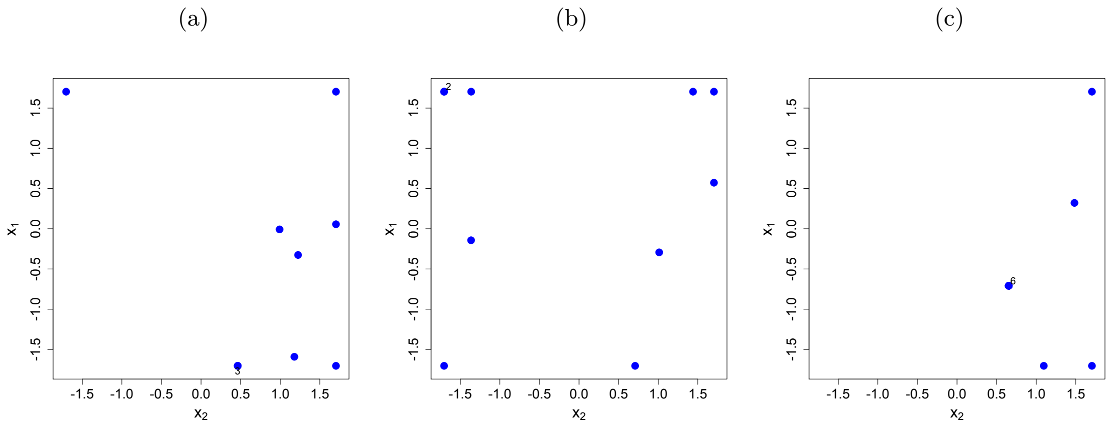


Figure B.2: Cantilever beam example: Bayesian near-optimal-ESIG designs from different random starts of ACE for the calibration model when τ^2 is treated as a nuisance parameter and $m = 30$

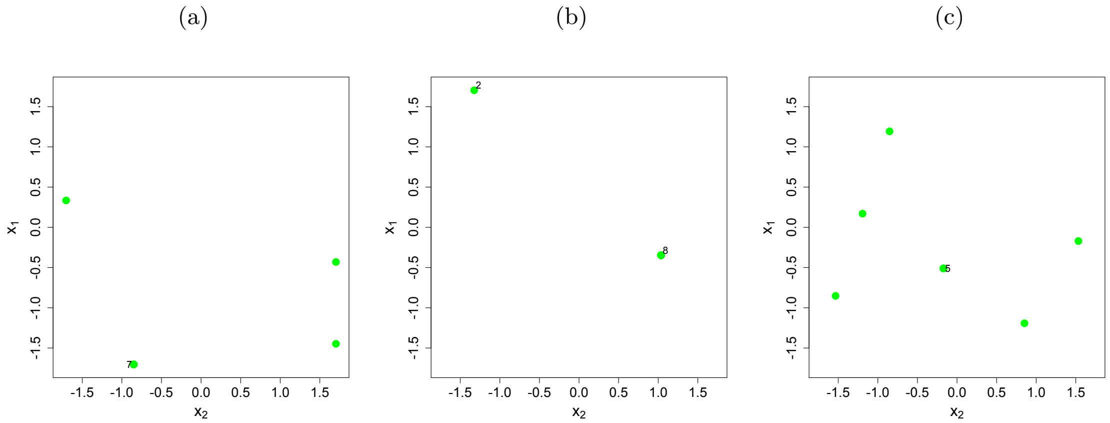


Figure B.3: Cantilever beam example: Bayesian near-optimal-ESIG designs from different random starts of ACE for the calibration model when τ^2 is treated as a nuisance parameter and $m = 60$

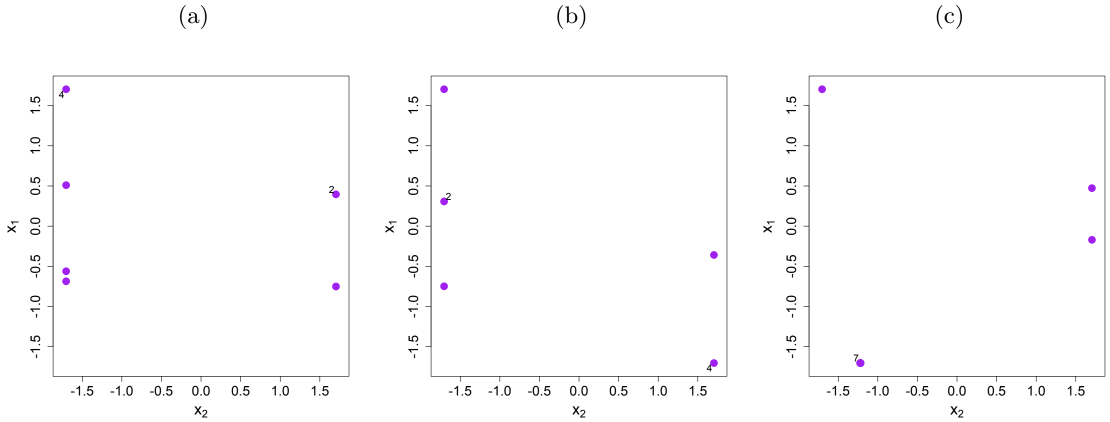


Figure B.4: Cantilever beam example: Bayesian near-optimal-ESIG designs from different random starts of ACE for the calibration model when τ^2 is treated as a nuisance parameter and $m = 90$

Design	Mean of 100 estimates of the ESIG			
	Nonlinear model	**Calibration model**, $m = 30$	**Calibration model**, $m = 60$	**Calibration model**, $m = 90$
$\xi_{CBF,1}$ Fig. B.1 (a)	1.087	0.282	0.747	0.831
$\xi_{CBF,2}$ Fig. B.1 (b)	1.082	0.271	0.739	0.829
$\xi_{CBF,3}$ Fig. B.1 (c)	1.079	0.279	0.699	0.818
$\xi_{cal,30,1}$ Fig. B.2 (a)	0.880	0.535	0.629	0.708
$\xi_{cal,30,2}$ Fig. B.2 (b)	0.873	0.529	0.622	0.699
$\xi_{cal,30,3}$ Fig. B.2 (c)	0.868	0.520	0.615	0.693
$\xi_{cal,60,1}$ Fig. B.3 (a)	1.074	0.385	0.692	0.801
$\xi_{cal,60,2}$ Fig. B.3 (b)	1.070	0.381	0.686	0.799
$\xi_{cal,60,3}$ Fig. B.3 (c)	1.063	0.376	0.679	0.783
$\xi_{cal,90,1}$ Fig. B.4 (a)	1.084	0.297	0.687	0.871
$\xi_{cal,90,2}$ Fig. B.4 (b)	1.078	0.292	0.681	0.868
$\xi_{cal,90,3}$ Fig. B.4 (c)	1.072	0.284	0.679	0.859

Table B.1: Mean of 100 estimates of the expected Shannon information gain for the designs given in Figures B.1-B.4; the ESIG was estimated under (i) the nonlinear model where σ_ε^2 is treated as a nuisance parameter; (ii) the calibration model where τ^2 is treated as a nuisance parameter and (a) $m = 30$; (b) $m = 60$; and (c) $m = 90$

B.2.6 Unknown simulator and $\delta_{\theta^p}(x) = 0$ - Michaelis-Menten model

Finally, we calculate the derivatives of the log-prior density, $\log \pi_b(\boldsymbol{\psi})$, with respect to the unknown parameters $\boldsymbol{\psi}'$, where $\boldsymbol{\psi}' = (\log \theta_1^p, \log \theta_2^p, \log \tau^2)^{\mathrm{T}}$, for the example given in Section 6.4.4. We have that:

$$\theta_1^p \sim \log N(\mu_1, \sigma_1^2), \quad \theta_2^p \sim \log N(\mu_2, \sigma_2^2), \quad \tau^2 \sim \mathrm{Exp}(\lambda_{\tau^2}).$$

The log-prior density is:

$$\log \pi_b(\boldsymbol{\psi}) = \log \pi_b(\theta_1^p) + \log \pi_b(\theta_2^p) + \log \pi_b(\tau^2)$$

$$= \log\left[\theta_1^p \sigma_1 (2\pi)^{1/2}\right] + \frac{(\log\theta_1^p - \mu_1)^2}{2\sigma_1^2} + \log\left[\theta_2^p \sigma_2 (2\pi)^{1/2}\right] + \frac{(\log\theta_2^p - \mu_2)^2}{2\sigma_2^2}$$
$$+ \log\lambda_{\tau^2} + \lambda_{\tau^2}\tau^2.$$

First Derivatives:

$$\begin{aligned}
\frac{\partial \log \pi_b(\theta_1^p)}{\partial \log \theta_1^p} &= -1 - \frac{\log\theta_1^p - \mu_1}{\sigma_1^2}, \\
\frac{\partial \log \pi_b(\theta_2^p)}{\partial \log \theta_2^p} &= -1 - \frac{\log\theta_2^p - \mu_2}{\sigma_2^2}, \\
\frac{\partial \log \pi_b(\tau^2)}{\partial \log \tau^2} &= -\lambda_{\tau^2}\tau^2.
\end{aligned}$$

Second derivatives:

$$\begin{aligned}
\frac{\partial^2 \log \pi_b(\theta_1^p)}{\partial [\log \theta_1^p]^2} &= -\frac{1}{\sigma_1^2}, \\
\frac{\partial^2 \log \pi_b(\theta_2^p)}{\partial [\log \theta_2^p]^2} &= -\frac{1}{\sigma_2^2}, \\
\frac{\partial^2 \log \pi_b(\tau^2)}{\partial [\log \tau^2]^2} &= -\lambda_{\tau^2}\tau^2.
\end{aligned}$$

Appendix C

C.1 The choice of hyperparameters for different examples

C.1.1 Example 3.2.1

For the example given in Section 3.2.1 we assume the calibration model (1.1) and we also assume a known simulator, the Michaelis-Menten model, with known and fixed parameters $\boldsymbol{\theta}^p = (15, 50)^{\mathrm{T}}$. Therefore

$$\eta(x, \boldsymbol{\theta}^p) = \frac{15x}{50 + x}.$$

We also assume a Gaussian process prior on the discrepancy function $\delta_{\boldsymbol{\theta}^p}(\cdot)$ such that

$$\boldsymbol{\delta}_{\boldsymbol{\theta}^p} \sim N\left[\mathbf{0}_n, \sigma^2 \mathbf{K}(\phi)\right],$$

where $\boldsymbol{\delta}_{\boldsymbol{\theta}^p} = [\delta_{\boldsymbol{\theta}^p}(x_1), \ldots, \delta_{\boldsymbol{\theta}^p}(x_n)]^{\mathrm{T}}$. We also assume $\sigma^2 \sim \mathrm{IG}(a, b)$, $\phi \sim \mathrm{Exp}(\lambda_\phi)$ with densities

$$\pi_b(\sigma^2) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp\left\{-\frac{b}{\sigma^2}\right\},$$
$$\pi_b(\phi) = \lambda_\phi \exp\{-\lambda_\phi \phi\},$$

with $a = 3$, $b = 2$, $\lambda_\phi = 200$ and $\sigma_\varepsilon^2 = 0$. This choice of prior distribution for the correlation parameter ϕ suggests that if two points x and x' in the range $[0, 400]$ are ‘close’ then the correlation function $\kappa(x, x'; \phi)$ is close to one and as the distance between the two points is increased the correlation function $\kappa(x, x'; \phi)$ decreases and tends to zero. See Figure C.1 for the density of $\kappa(x, x'; \phi)$ for different values of ϕ sampled from the prior distribution and for three fixed distances between two points. In Figure C.2 (b) we present samples from the prior distribution of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$ for this particular choice of hyperparameters.

This choice of hyperparameters results in a sensible prior distribution in relation to the “size” of the model (see Figure C.2 (a) for the expected response of the Michaelis-Menten model), for the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$. Samples from the posterior distribution of the discrepancy function are presented in Section 3.2.1.

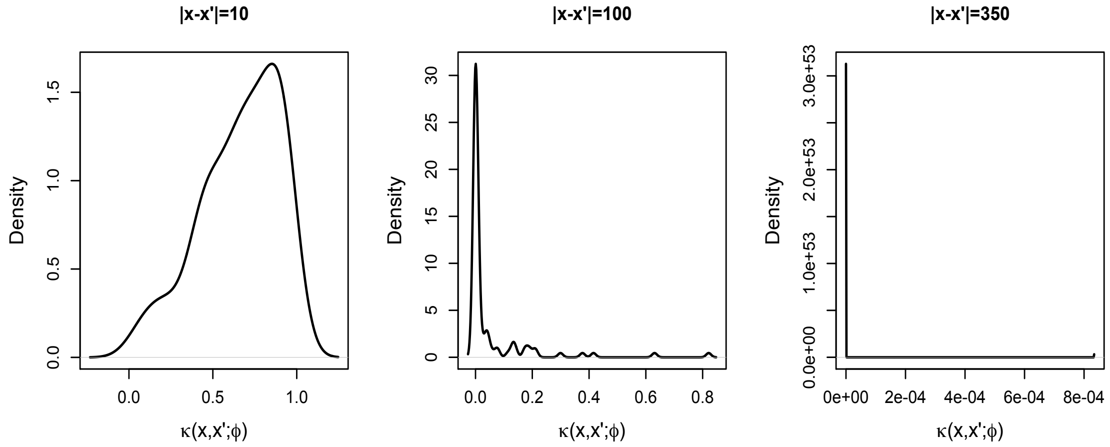


Figure C.1: The density of $\kappa(x, x'; \phi)$ for different values of ϕ sampled from the prior distribution and for three fixed distances between two points: (i) $|x - x'| = 10$; (ii) $|x - x'| = 100$; and (iii) $|x - x'| = 350$

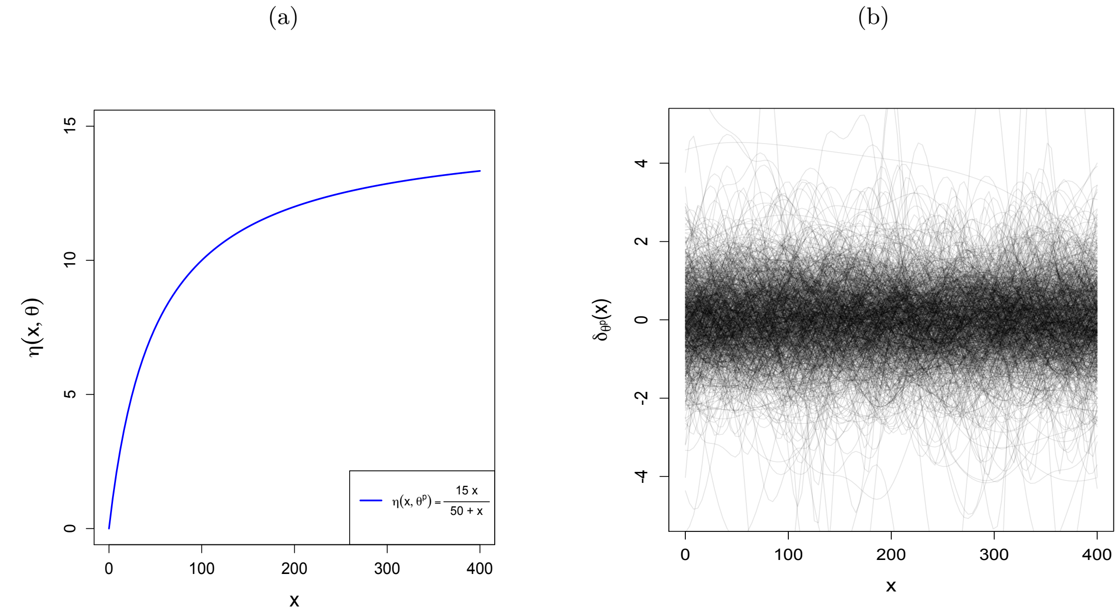


Figure C.2: Example 3.2.1: (a) The expected response of the Michaelis-Menten model, $\eta(x, \boldsymbol{\theta}^p) = \frac{15x}{50+x}$; (b) Samples from the prior distribution of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$

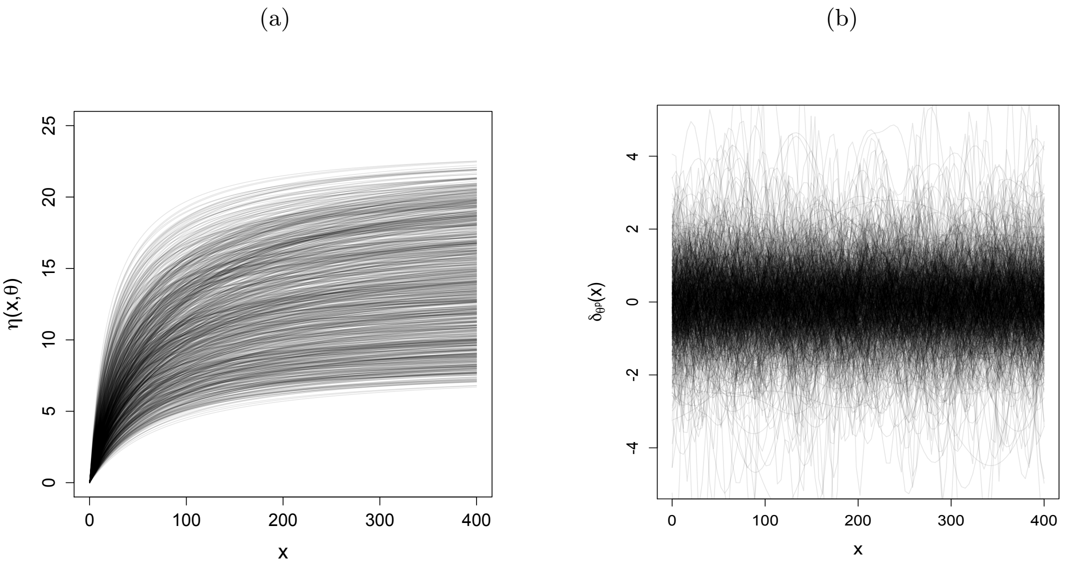


Figure C.3: Example 3.2.2: Examples of (a) the shape of the expected response of the Michaelis-Menten model for different values of θ_1^p and θ_2^p sampled from the prior distributions; (b) realisations from the prior distribution of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$ for the particular choice of hyperparameters

C.1.2 Example 3.2.2

For the example given in Section 3.2.2, we assume the calibration model (1.1) and a known simulator, the Michaelis-Menten model, with unknown parameters $\boldsymbol{\theta}^p$:

$$\eta(x, \boldsymbol{\theta}) = \frac{\theta_1 x}{\theta_2 + x}.$$

We also assume a Gaussian process prior on the discrepancy function $\delta_{\boldsymbol{\theta}^p}(\cdot)$ such that

$$\boldsymbol{\delta}_{\boldsymbol{\theta}^p} \sim N\left[\mathbf{0}_n, \sigma^2 \mathbf{K}(\phi)\right],$$

where $\boldsymbol{\delta}_{\boldsymbol{\theta}^p} = [\delta_{\boldsymbol{\theta}^p}(x_1), \ldots, \delta_{\boldsymbol{\theta}^p}(x_n)]^{\mathrm{T}}$ and $\theta_1^p \sim \mathrm{Unif}[8, 24]$, $\theta_2^p \sim \mathrm{Unif}[20, 85]$, $\sigma^2 \sim \mathrm{IG}(3, 2)$, $\phi \sim \mathrm{Exp}(200)$ and $\tau^2 \sim \mathrm{Exp}(15)$ (similar to Example C.1.1). In Figure C.1 we present the density of $\kappa(x, x'; \phi)$, for different values of ϕ sampled from the prior distribution and for three fixed distances between two points. In Figure C.3 we present (a) examples of the shape of the expected response of the Michaelis-Menten model for different values of θ_1^p and θ_2^p sampled from these prior distributions and (b) samples from the prior distribution of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$ for this particular choice of hyperparameters.

This choice of hyperparameters results in a sensible prior for the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$ in relation to the "size" of the model for different values of θ_1^p and θ_2^p. Samples from the posterior distribution of the discrepancy function are presented in Section 3.2.2.

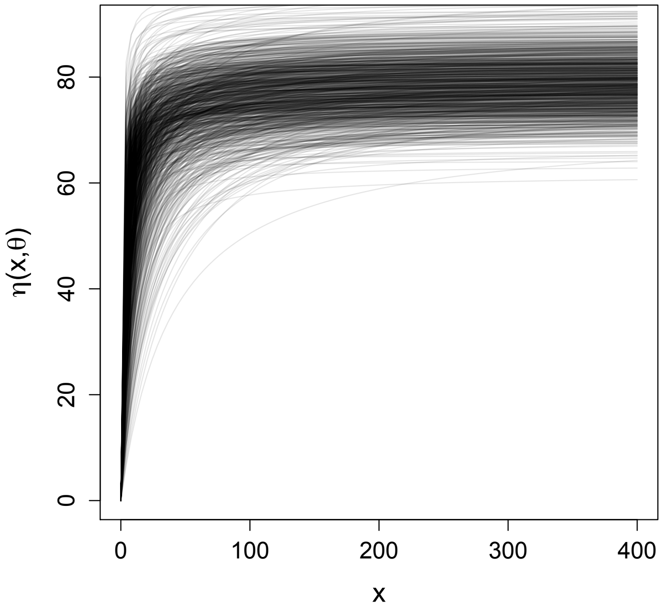


Figure C.4: Example 5.1.2: The expected response of the Michaelis-Menten model for different values of θ_1 and θ_2 sampled from their prior distributions

C.1.3 Example 5.1.2

For the Michaelis-Menten model given in Section 5.1.2 we assume log-normal prior distributions for the unknown parameters θ_1 and θ_2 with $\mu_1 = 4.38$, $\sigma_1 = 0.07$, $\mu_2 = 1.19$ and $\sigma_2 = 0.84$. In Figure C.4 we present examples of the shape of the expected response of the Michaelis-Menten model for different values of θ_1 and θ_2 sampled from these prior distributions.

Figure C.4 describes the variability of the simulator output due to the variability in the parameters. Also, the variability shown in this figure explains why the optimal designs obtained for the Michaelis-Menten model have most points where the curve is changing more quickly and also some points at the stable part of the curve.

C.1.4 Example 5.1.3

For the Biochemical Oxygen Demand (BOD) model given in Section 5.1.2 we assume log-normal prior distributions for the unknown parameters θ_1 and θ_2 with $\mu_1 = 3.38$, $\sigma_1 = 0.20$, $\mu_2 = 1.098$, $\sigma_2 = 1.12$. In Figure C.5 we show the shape of the expected response of the BOD model for different values of θ_1 and θ_2 sampled from these prior distributions.

Similarly to the previous example, Figure C.5 describes the variability of the simulator

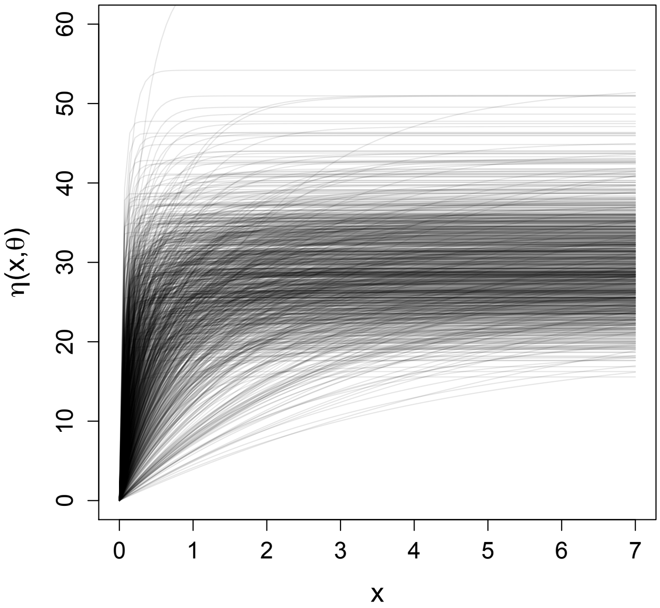


Figure C.5: Example 5.1.3: The expected response of the BOD model for different values of θ_1 and θ_2 sampled from their prior distributions

output due to the variability in the parameters. Again, the variability shown in this figure explains why the optimal designs obtained for the BOD model have most points where the curve is changing more quickly and also some points at the stable part of the curve.

C.1.5 Example 6.3.1

For the example given in Section 6.3.1 we assume the calibration model (6.1) with the simulator being the Michaelis-Menten model with unknown parameters $\boldsymbol{\theta}$, similar to Example C.1.2, such that:

$$\eta(x, \boldsymbol{\theta}) = \frac{\theta_1 x}{x + \theta_2}.$$

We assume independent log-normal prior distributions $\theta_1^p \sim \log N(4.38, 0.07^2)$ and $\theta_1^p \sim \log N(1.19, 0.84^2)$ (as in Example C.1.3). We also assume a Gaussian process prior for the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$

$$\boldsymbol{\delta}_{\boldsymbol{\theta}^p} \sim N\left[\mathbf{0}_n, \sigma^2 \mathbf{K}(\phi)\right],$$

where $\boldsymbol{\delta}_{\boldsymbol{\theta}^p} = [\delta_{\boldsymbol{\theta}^p}(x_1), \ldots, \delta_{\boldsymbol{\theta}^p}(x_n)]^{\mathrm{T}}$. We use $\sigma^2 \sim \mathrm{IG}(3, 2)$, $\phi \sim \mathrm{Exp}(200)$ and $\tau^2 \sim \mathrm{Exp}(20)$. In Figure C.1 we present the density of the correlation function $\kappa(x, x'; \phi)$, for different values of ϕ sampled from the prior distribution and for three fixed distances between two points.

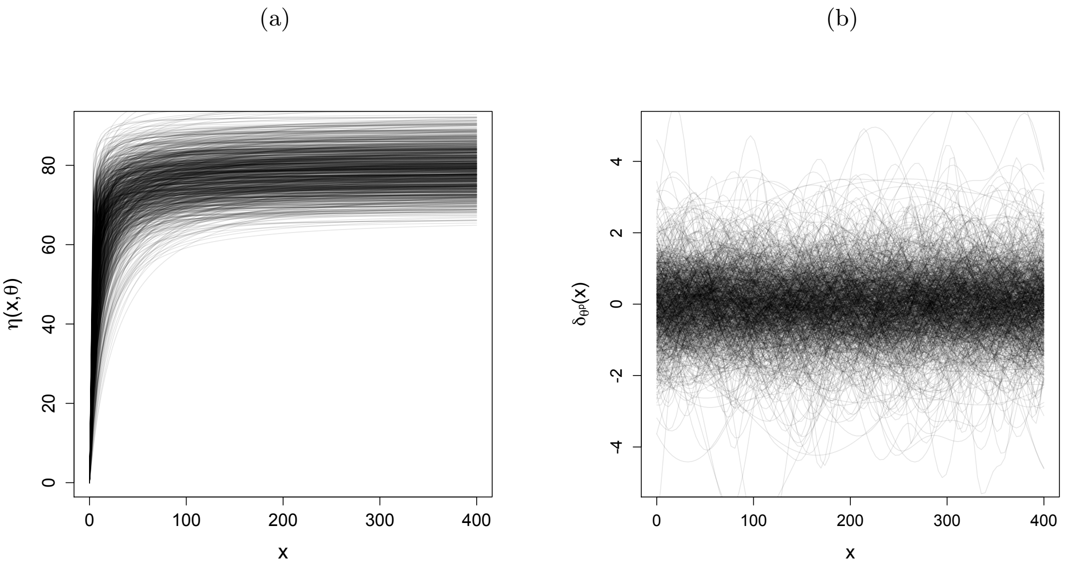


Figure C.6: Example 6.3.1: Examples of (a) the shape of the expected response of the Michaelis-Menten model for different values of θ_1^p and θ_2^p sampled from the prior distributions; (b) realisations from the prior distribution of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$ for the particular choice of hyperparameters

In Figure C.6 we present (a) examples of the shape of the expected response of the Michaelis-Menten model for different values of θ_1^p and θ_2^p sampled from these prior distributions, and (b) samples from the prior distribution of the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$ for this particular choice of hyperparameters. This choice of hyperparameters results in a sensible prior for the discrepancy function $\delta_{\boldsymbol{\theta}^p}(x)$ in relation to the "size" of the model for different values of θ_1^p and θ_2^p.

C.2 Example: Michaelis-Menten simulator and $\delta_{\theta^p}(x) \neq 0$

We assume the example given in Section 6.3.1. We compare estimates of the expected Shannon information gain found using nMC, ALIS and LIS for two combinations of k_1 and k_2; (i) $k_1 = k_2 = 300$ and (ii) $k_1 = 2000, k_2 = 10000$. Both normal and t importance distributions are used in ALIS and LIS. The expected Shannon information gain is approximated for the designs given in Figure 6.2. We treat as the 'true' ESIG the nMC approximation with $k_1 = k_2 = 1,000,000$ (red line) because should lead to negligible bias.

Figure C.7 shows the distribution of 100 estimates of the ESIG found using nMC, ALIS and LIS for the D-optimal design, $\xi_D^\star$. Increasing k_1 and k_2 reduces the variance and bias of nMC. For this design, the ESIG using ALIS and LIS also changes by increasing k_2 which controls the bias and by increasing k_1 which reduces the variance. Changing the importance distribution from a normal to a t also makes a difference.

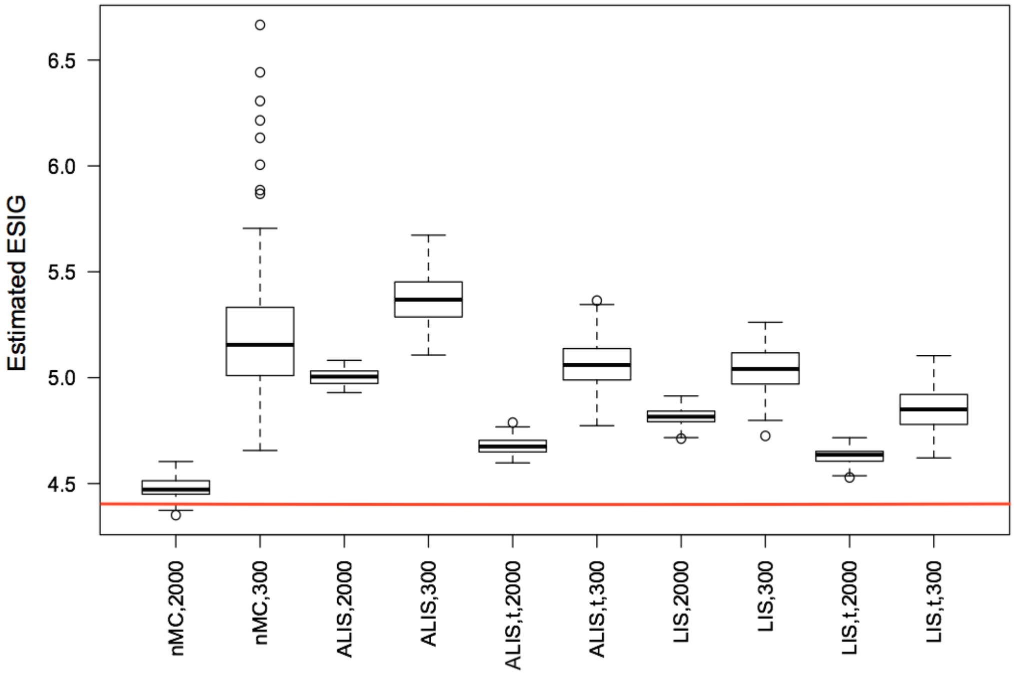


Figure C.7: Estimated ESIG for the parameters $\boldsymbol{\psi}$ of the calibration model found using nMC, ALIS and LIS for different combinations of k_1 and k_2, for the D-optimal design, $\xi_D^\star$, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$

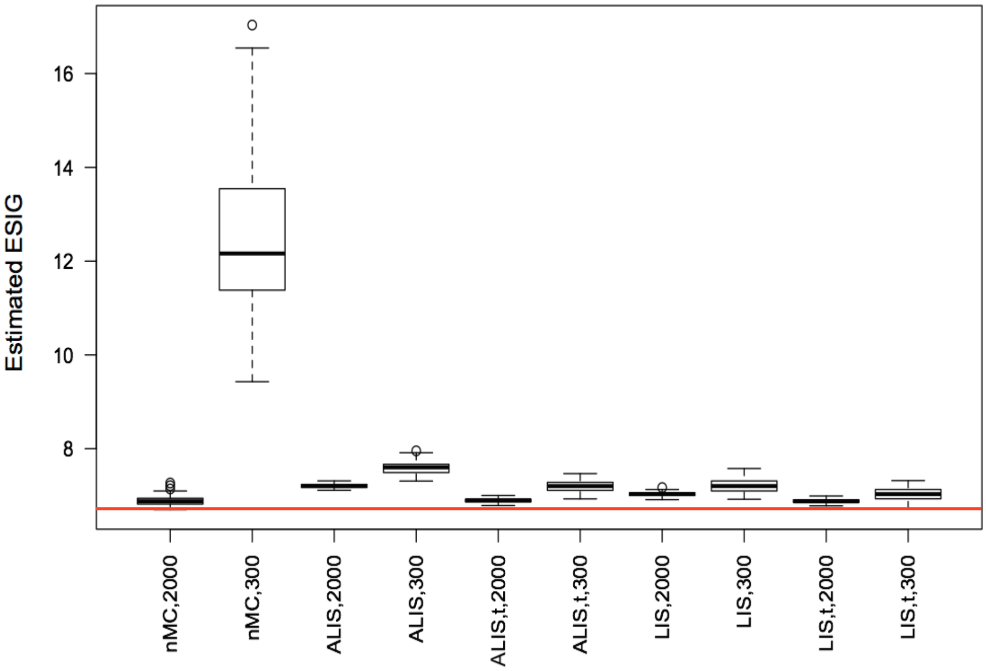


Figure C.8: Estimated ESIG for the parameters $\boldsymbol{\psi}$ of the calibration model found using nMC, ALIS and LIS for different combinations of k_1 and k_2, for the Bayesian optimal design of the Michaelis-Menten model found using ACE, $\xi_{MM}^\star$, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$

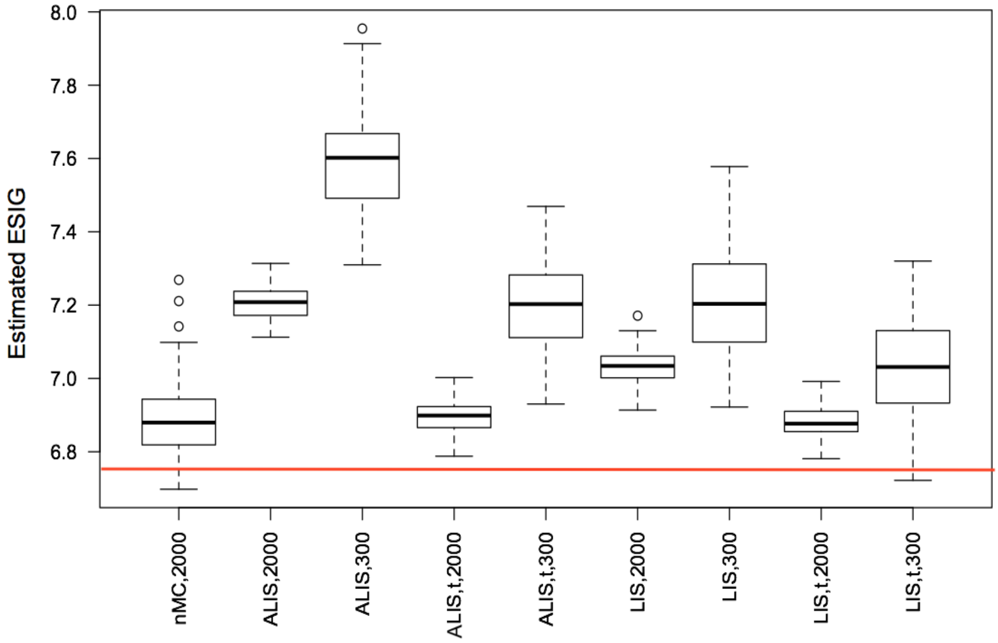


Figure C.9: Estimated ESIG for the parameters $\boldsymbol{\psi}$ of the calibration model found using nMC, ALIS and LIS for different combinations of k_1 and k_2, for the Bayesian optimal design for the Michaelis-Menten model found using ACE, $\xi^{\star}_{MM}$, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$ (nMC,300 is omitted because this method exhibits large bias)

Figures C.9 and C.8 show the distribution of 100 estimates of the ESIG found using nMC, ALIS and LIS for the Bayesian optimal design for the Michaelis-Menten model found using ACE, $\xi^{\star}_{MM}$. In the latter plot nMC,300 results have been omitted due to high positive bias. We can see similar patterns as discussed for Figure C.7.

In Figure C.10 we show the distribution of 100 estimates of the ESIG found using nMC, ALIS and LIS for the maximin LHS design, ξ_{LHS}. We can see similar patterns as in Figures C.7 and C.9. Again, we have not included nMC,300 results due to high bias. For this design, which is equally spaced in one dimension, the prior distribution of the correlation parameter, ϕ, is more similar to the posterior distribution than the approximation to the posterior distribution used in ALIS and LIS, and for this reason nMC appears to perform better than either ALIS or LIS.

In Figure C.11, similarly to the previous figures, we show the distribution of 100 estimates of the ESIG found using nMC, ALIS and LIS for the Bayesian optimal design for the calibration model found using ACE, $\xi^{\star}_{cal}$. Clearly in this figure we can see that ALIS with a t importance distribution gives approximations with the least bias.

Figure C.12 shows 100 estimates of the ESIG for two replicates of the 10-run Bayesian optimal design for the calibration model found using ACE. For this design, the advantage of ALIS and LIS over nMC is much clearer.

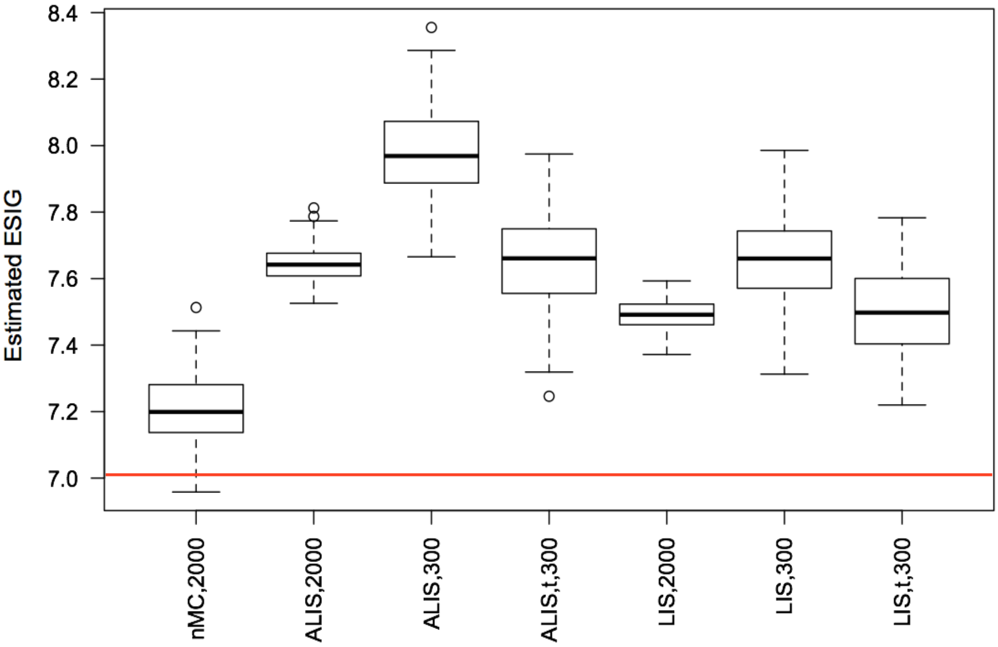


Figure C.10: Estimated ESIG for the parameters $\boldsymbol{\psi}$ of the calibration model found using nMC, ALIS and LIS for different combinations of k_1, k_2, for the maximin LHS design, ξ_{LHS}, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$

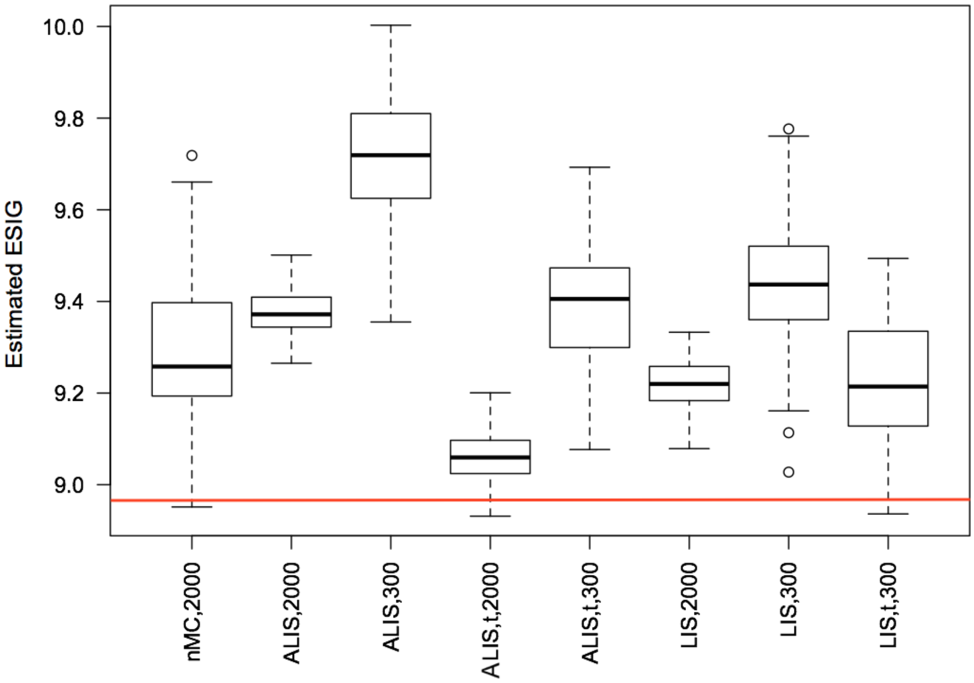


Figure C.11: Estimated ESIG for the parameters $\boldsymbol{\psi}$ of the calibration model found using nMC, ALIS and LIS for different combinations of k_1, k_2, for the Bayesian optimal design for the calibration model found using ACE, $\xi^{\star}_{cal}$, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$

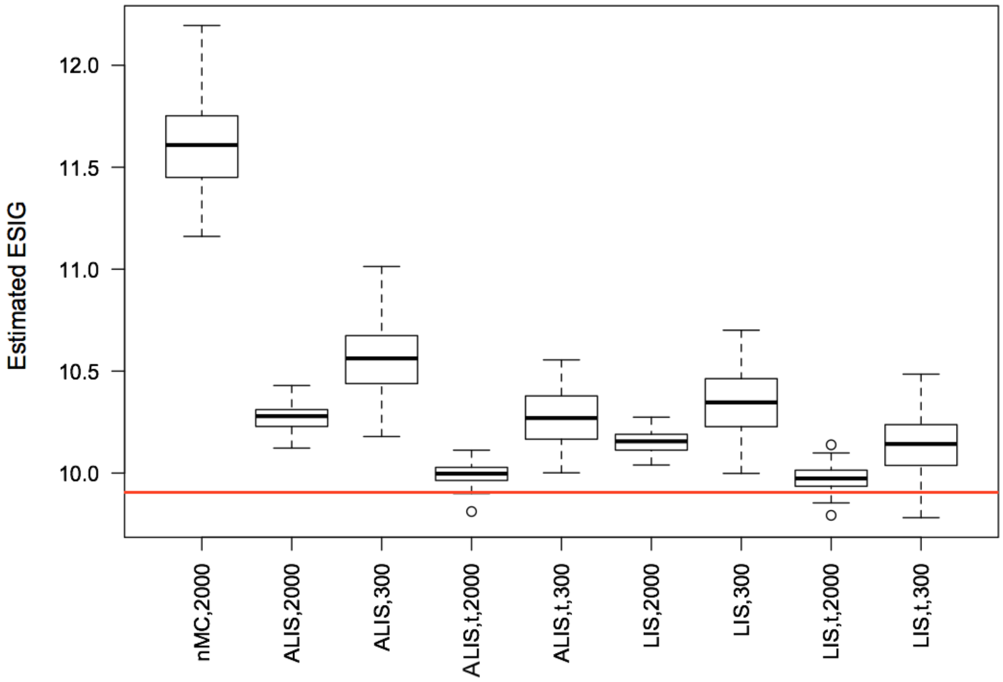


Figure C.12: Estimated ESIG for the parameters $\boldsymbol{\psi}$ of the calibration model found using nMC, ALIS and LIS for different combinations of k_1 and k_2, for two replicates of the 10-run Bayesian optimal design for the calibration model, $\xi^{\star}_{cal}$, and the 'true' ESIG (red line) obtained from nMC with $k_1 = k_2 = 1,000,000$

The examples where nMC is performing better than ALIS and LIS are when evaluating designs that probably do not give much information about the calibration parameters. In these examples the prior distribution is a better approximation to the posterior than an asymptotic Laplace approximation. For the designs that give information about the calibration parameters, nMC requires big sample sizes k_1 and k_2 in order to reduce the bias. For this particular example we chosen LIS to approximate the expected Shannon information gain as is a bit more accurate than ALIS, and empirically is not much more computationally expensive, at least for small Monte Carlo sample sizes.

C.3 Unknown simulator and $\delta_{\theta^p}(x) = 0$ - Cantilever Beam function

For the example presented in Section 6.4.3, we fit a Gaussian process with a constant mean to the simulator outputs, $\mathbf{z}$, obtained using a computer experiment observed under a design $\xi^c = [(\mathbf{x}^c_1, \theta^c_1), \ldots, (\mathbf{x}^c_m, \theta^c_m)]$. The prior distributions used are the same given in Section 6.4.3. We use `mlegp` package (Dancik, 2007), to model the effect of Young's modulus of beam material, θ, on the output of the cantilever beam function. We fit a GP to $m = 30, 60, 90$ simulator runs.

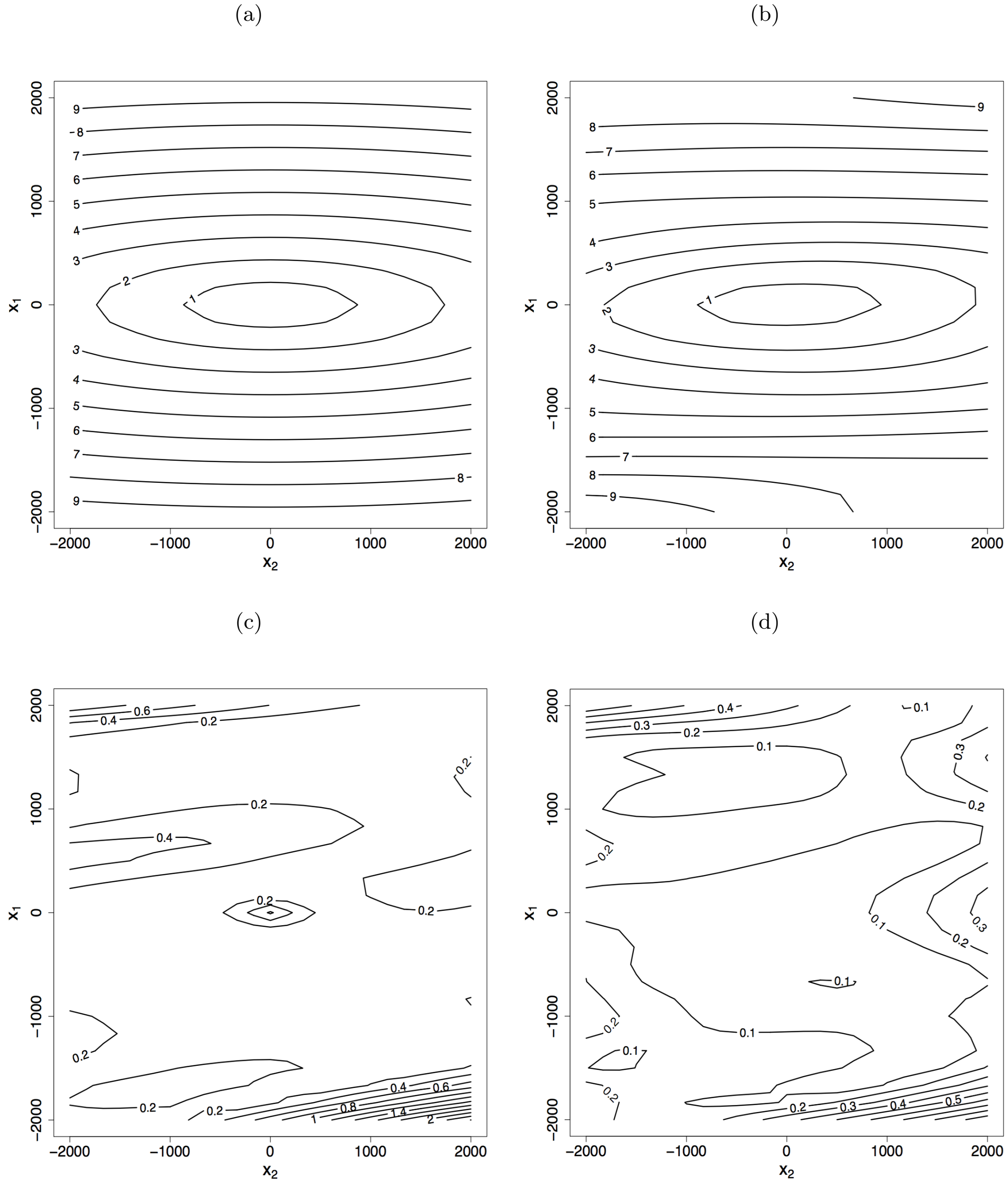


Figure C.13: Gaussian process fit for $m = 30$. (a) Contour plot of the cantilever beam function for a fixed $\theta = 2.71 \times 10^7$; (b) Posterior mean of the Gaussian process fit on the cantilever beam function; (c) The root squared difference between the response and the posterior mean; (d) Posterior standard deviation of the Gaussian process fit on the cantilever beam function

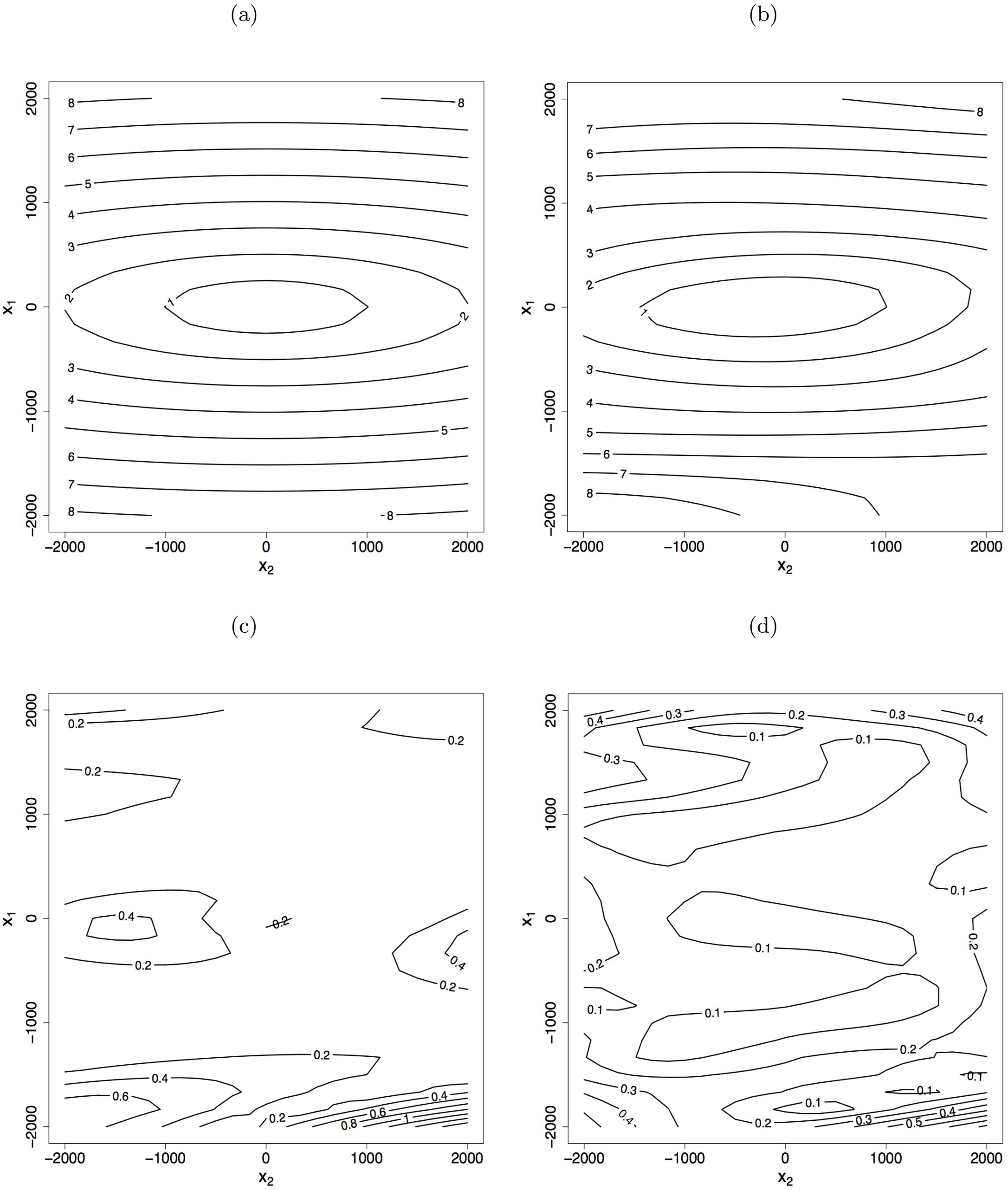


Figure C.14: Gaussian process fit for $m = 30$. (a) Contour plot of the cantilever beam function for a fixed $\theta = 3.15 \times 10^7$; (b) Posterior mean of the Gaussian process fit on the cantilever beam function; (c) The root squared difference between the response and the posterior mean; (d) Posterior standard deviation of the Gaussian process fit on the cantilever beam function

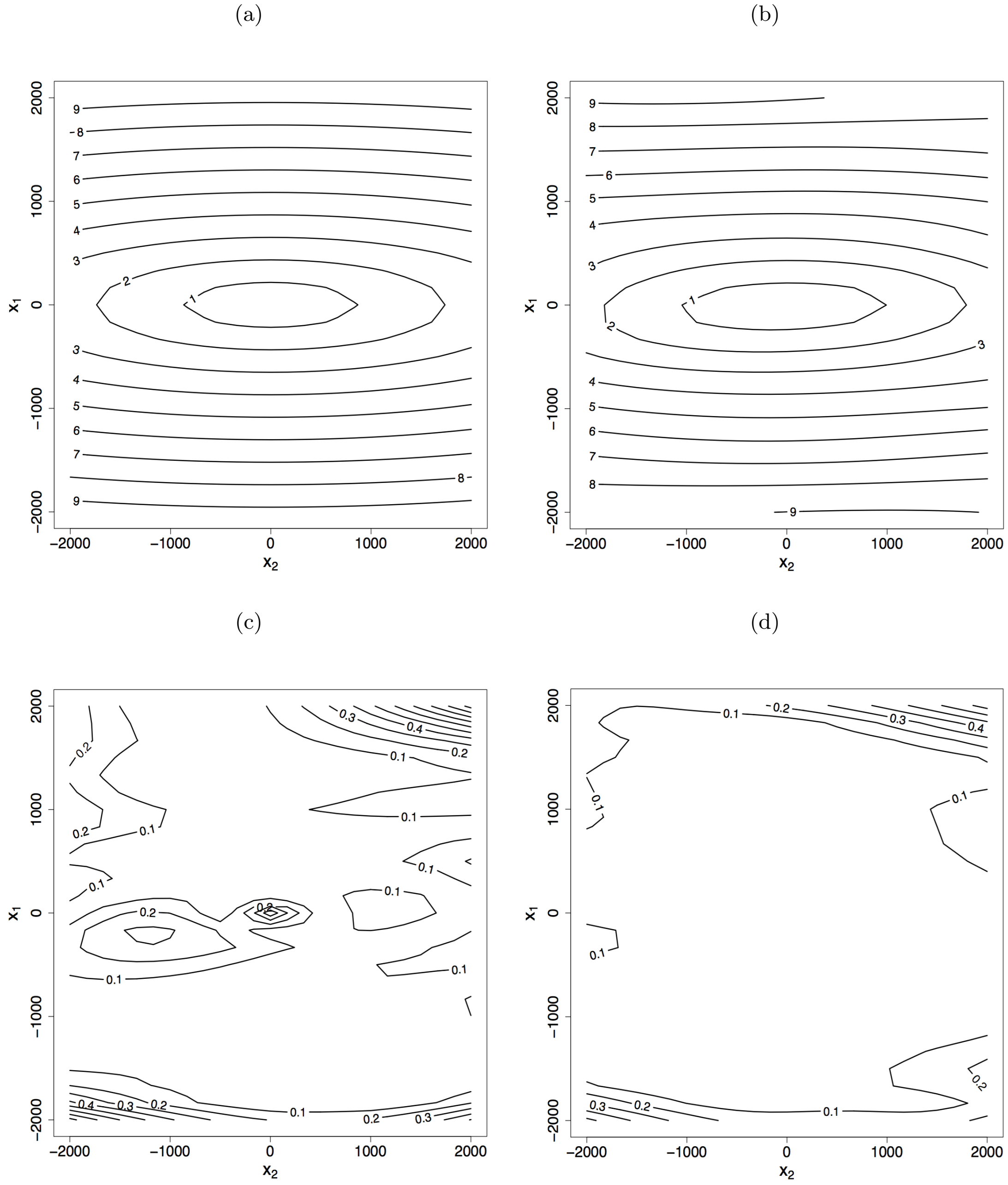


Figure C.15: Gaussian process fit for $m = 60$. (a) Contour plot of the cantilever beam function for a fixed $\theta = 2.71 \times 10^7$; (b) Posterior mean of the Gaussian process fit on the cantilever beam function; (c) The root squared difference between the response and the posterior mean; (d) Posterior standard deviation of the Gaussian process fit on the cantilever beam function

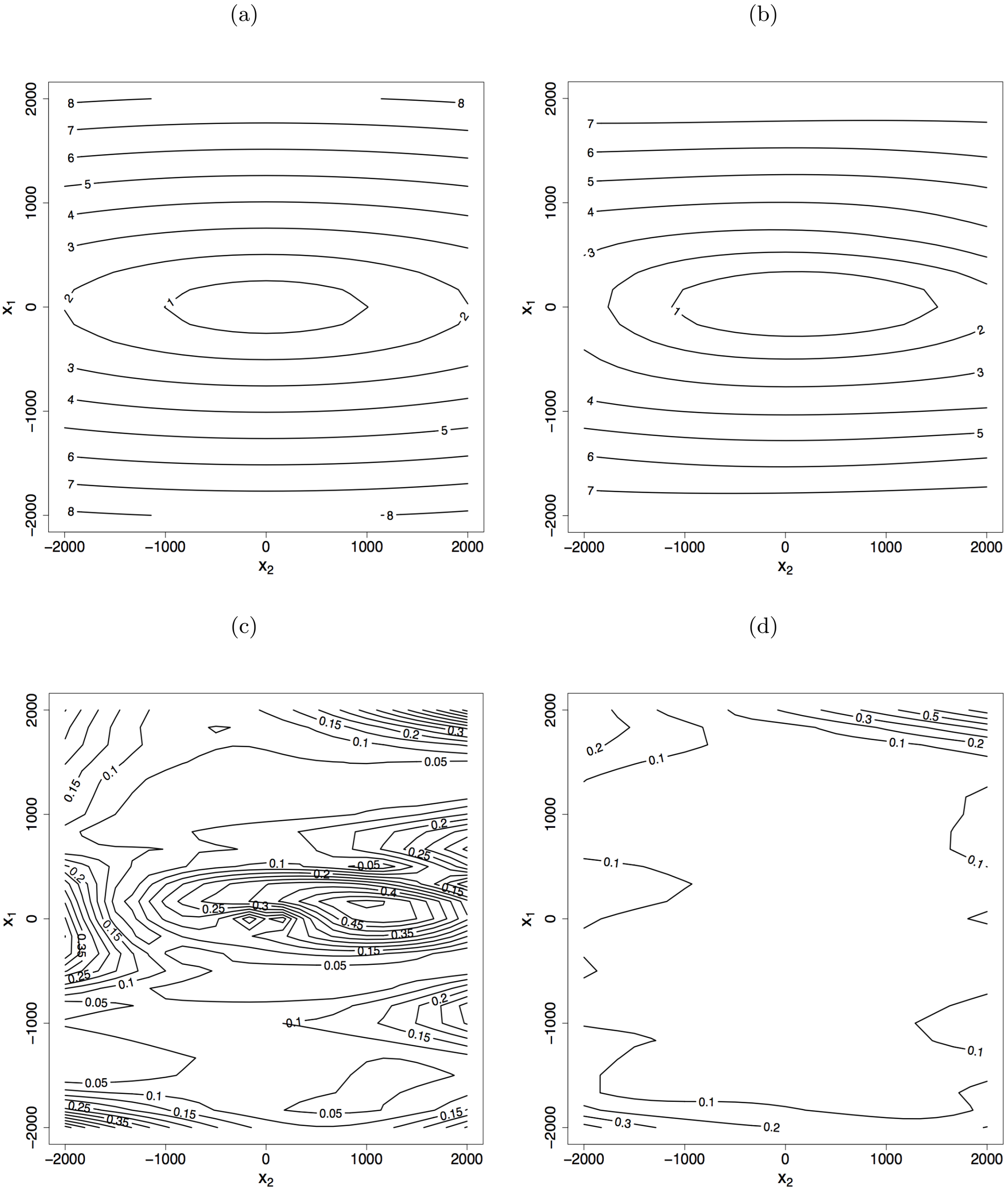


Figure C.16: Gaussian process fit for $m = 60$. (a) Contour plot of the cantilever beam function for a fixed $\theta = 3.15 \times 10^7$; (b) Posterior mean of the Gaussian process fit on the cantilever beam function; (c) The root squared difference between the response and the posterior mean; (d) Posterior standard deviation of the Gaussian process fit on the cantilever beam function

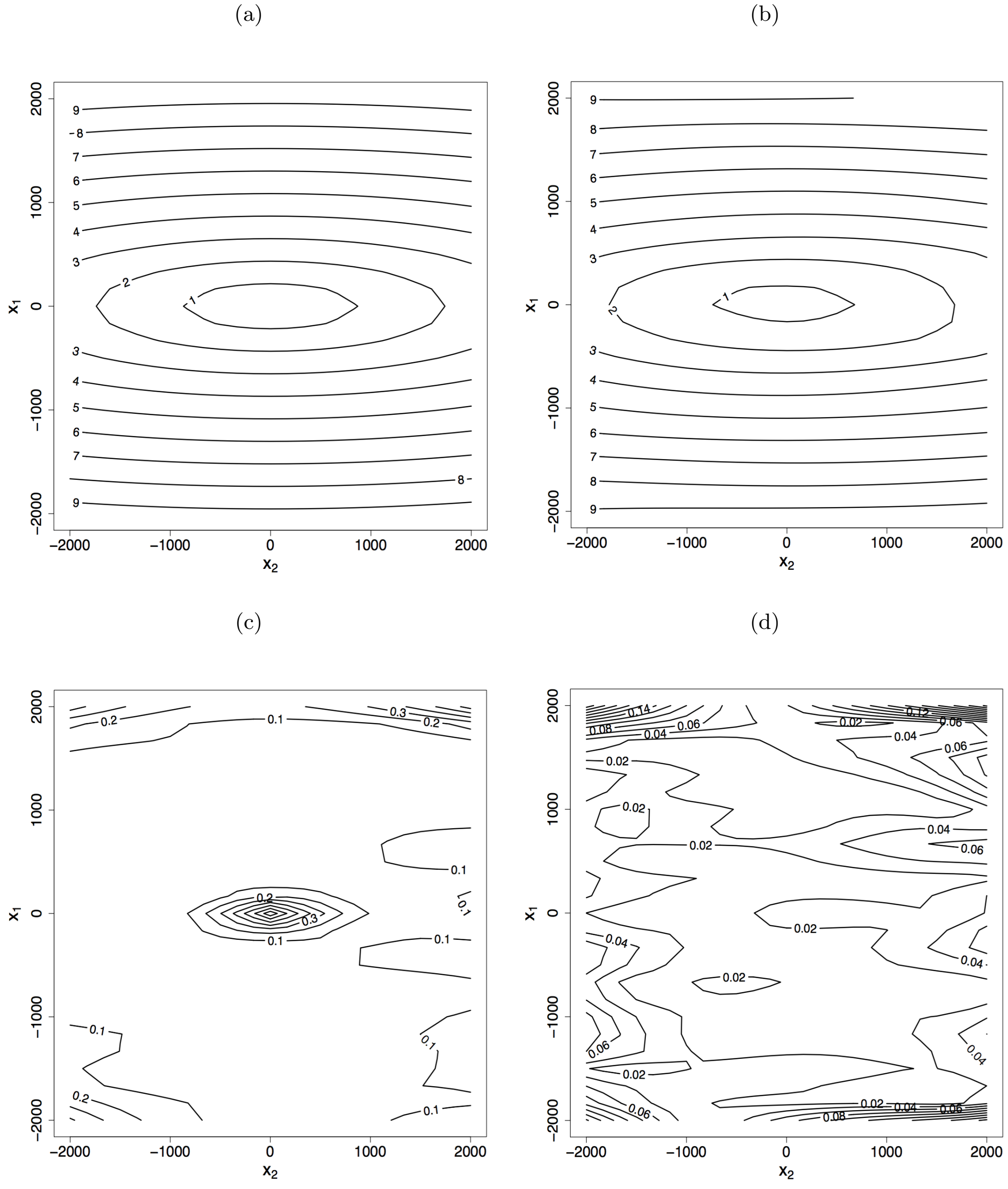
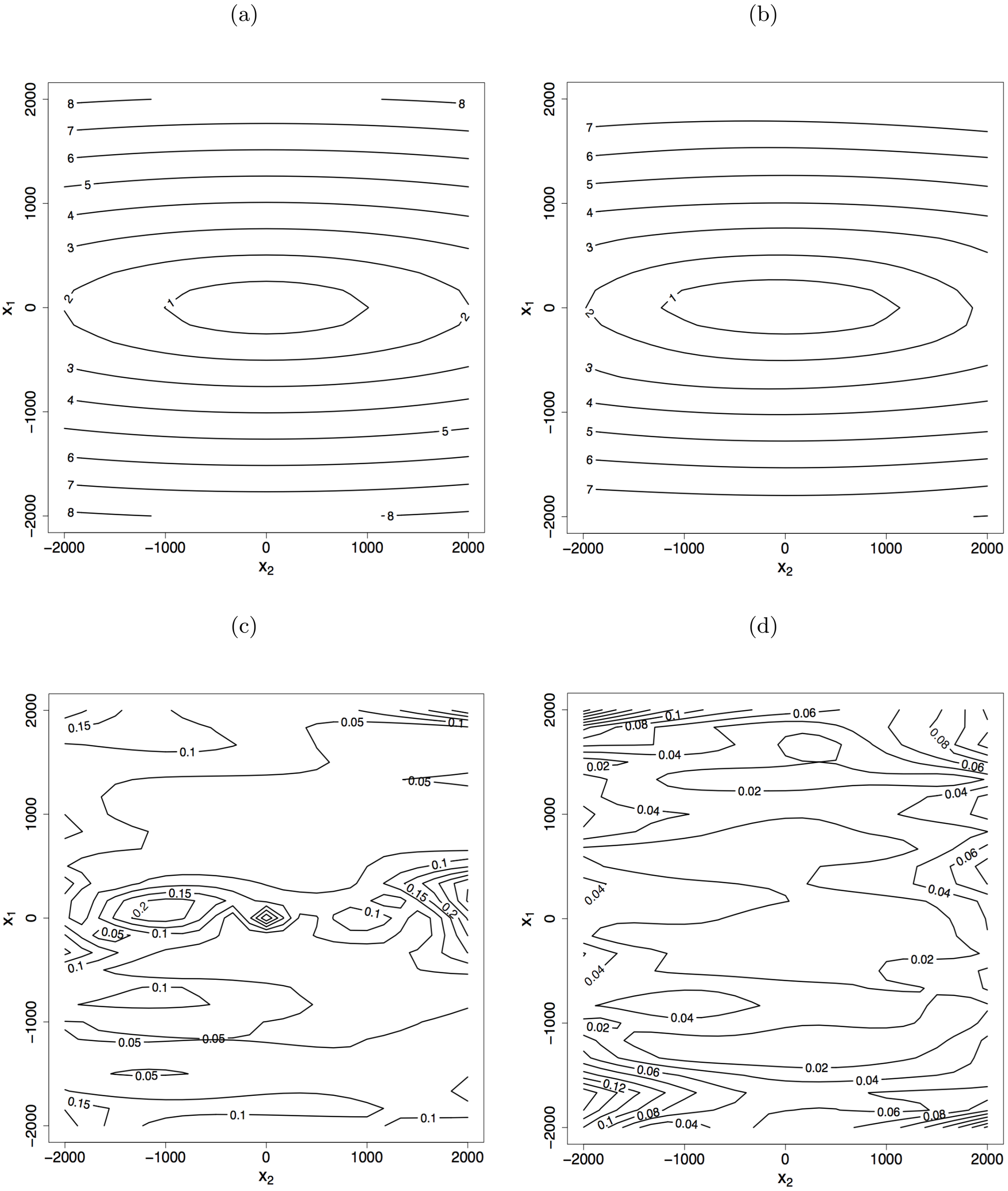


Figure C.17: Gaussian process fit for $m = 90$. (a) Contour plot of the cantilever beam function for a fixed $\theta = 2.71 \times 10^7$; (b) Posterior mean of the Gaussian process fit on the cantilever beam function; (c) The root squared difference between the response and the posterior mean; (d) Posterior standard deviation of the Gaussian process fit on the cantilever beam function



(a)

(b)

(c)

(d)

Figure C.18: Gaussian process fit for $m = 90$. (a) Contour plot of the cantilever beam function for a fixed $\theta = 3.15 \times 10^7$; (b) Posterior mean of the Gaussian process fit on the cantilever beam function; (c) The root squared difference between the response and the posterior mean; (d) Posterior standard deviation of the Gaussian process fit on the cantilever beam function

Figures C.13 - C.18 present results from these GP models from $m = 30$ (C.13 and C.14), 60 (C.15 and C.16), and 90 (C.17 and C.18) runs. For each figure, plot (a) presents the simulator output for a value of θ not in the computer experiment design, plot (b) gives the posterior predictive mean for this θ value from the GP model, plot (c) gives the root squared difference between the true simulator output and the posterior predictive mean, and plot (d) gives the posterior predictive standard deviation.

As we increase the number of simulator runs m, the posterior mean of the GP more closely resembles the true response; especially noticeable at the edge of the design space. The root squared difference between the true response and the posterior mean becomes smaller for larger m. The posterior standard deviation of the GP decreases as we increase m.

These figures demonstrate that as we increase the number of observations m we get a GP mean that adapts better to the true response and the GP standard deviation is smaller. For space-filling designs and Bayesian optimal designs the estimated ESIG tends to the value of the estimated ESIG for the nonlinear model as m increases.