Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

Iterative improvement of Bayesian parameter estimates for an engine model by means of experimental design

Sebastian Mosbach, Andreas Braumann, Peter L. W. Man,
Catharine A. Kastner, George P. E. Brownbridge, Markus Kraft [1]

released: 12 July 2011

[1] Department of Chemical Engineering
and Biotechnology
University of Cambridge
New Museums Site
Pembroke Street
Cambridge, CB2 3RA
United Kingdom
E-mail: mk306@cam.ac.uk

Preprint No. 104

Keywords: Model-based design of experiments, Bayesian parameter estimation, expensive models, internal combustion engines

Edited by
Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
New Museums Site
Pembroke Street
Cambridge CB2 3RA
United Kingdom

Fax: + 44 (0)1223 334796
E-Mail: c4e@cam.ac.uk
World Wide Web: http://como.cheng.cam.ac.uk/

Abstract

We implement an algorithm which estimates parameters of an internal combustion engine model using a Bayesian approach and employs an experimental design technique to iteratively suggest new experiments with the aim of decreasing the uncertainty in the parameter estimates. The primary focus here is the application of the methodology to a complex model whose computational expense limits the number of model evaluations to an extent which necessitates the use of surrogate models. In this work, we choose quadratic response surfaces as surrogates. The main goal of the considered engine model is to predict emissions formed by in-cylinder combustion during the closed-volume part of the engine cycle, employing detailed sub-models for the chemical kinetics of the fuel, turbulent mixing, and convective heat transfer. The model is applied here to an ultra-low emission Homogeneous Charge Compression Ignition (HCCI) engine fuelled with iso-octane. We find rapid convergence of the iterative algorithm in the considered case, as shown by a substantial reduction in parametric uncertainty in each iteration, using informative as well as non-informative priors.

Contents

1 Introduction

Computational models can be useful in reducing the development time and cost of new products in the chemical industry and many others. This is particularly relevant in areas where experimental measurements are expensive and time-consuming, as is the case with internal combustion engines. However, creating predictive yet computationally cheap models remains a significant challenge, and requires large amounts of research effort. Automating the model development process to the greatest possible extent is therefore desirable [25].

Any non-trivial model contains parameters which are often unknown or uncertain *a priori* and an obvious question to ask then is what values of the parameters lead to the best agreement between experiment and model. This problem has been referred to by a number of names, including inverse problem, calibration, and parameter estimation [2, 40]. A Bayesian approach to estimating parameters of a computer model was introduced in [23]. This has since been considered state of the art, and has been extended and applied in numerous contexts (see for example [4, 21, 22, 44]).

Parameter estimation is widely seen as an iterative procedure, alternating with experimental design [1, 4, 17]. The general aim of experimental design [16] is to determine the most useful set of experiments to perform based on available information in the form of previous experimental results or models. Experimental design has traditionally been used most in conjunction with statistical analysis to determine which process condition variables have a significant impact on the measured responses. It has become a standard technique for planning series of experiments in virtually all areas of research. Experimental design for the purposes of parameter estimation [13] attempts to tune model parameters for improved accuracy by finding process conditions for the next experiment such that the "information gain" is maximised, the goal being to minimise the variance in the parameters. In [17], applications of this approach across the field of chemical engineering are reviewed. In [35], an application in combustion chemical kinetics is presented. Advances in computing have caused an expansion in both the complexity of the models and the sophistication of the design techniques. In [11], an experimental design technique is described and demonstrated which integrates a method to select the parameters to be estimated for models with large numbers of parameters.

Complex, physics-based models in common use today are usually sufficiently computationally expensive so that the number of evaluations which can be performed in practice may be strongly limited. Whenever a model is too expensive to realistically be used in a parameter estimation or experimental design context, it is usually replaced by a much cheaper, often approximate model, called a surrogate. Surrogate models, in one form or another, have been in use across many fields in many applications – too numerous to review here exhaustively. In the area of combustion, the most common application is the fast integration of chemical kinetic equations (see [30] for a brief, incomplete overview). Common surrogates include quadratic response surfaces [12, 18, 38] and High-Dimensional Model Representation [33, 45]. In [30], cubic natural splines have been used to represent an internal combustion engine model as a whole, but applications outside chemical kinetics remain scarce.

It has been recognised that expensive models are particularly relevant to applications in practice [36]. Gaussian process models are a popular choice as surrogates in such cases [36, 44]. For example, in [21, 22], a Bayesian methodology employing Gaussian process models is used to estimate parameters of a complex, highly multivariate model for an imploding cylinder whilst accounting for *a priori* knowledge about the discrepancy between the model and the system. In [43], two Bayesian optimisation algorithms for expensive functions, again using Gaussian process models, are presented. Although works like these exist, direct optimisation attempts of expensive, black-box models are rare, according to a recent review [39], partly due to the intrinsic difficulty of the problem. In [24], it is shown that there may be a way to circumvent the problem in cases where the model possesses fidelity parameters such as a mesh spacing and there exist strong correlations between simulation results with low and high fidelity.

Despite these efforts in various fields, the authors are not aware of any use of expensive models for the purposes of parameter estimation or experimental design in the field of combustion in general, and in engine research in particular. While experimental design is a standard technique in the engine calibration area (see for example [10]) and is a well-established part of the engine development process (see for example [34]), only data-driven polynomial models have been used in this context and, to the authors' knowledge, no attempt has been made to include complex, expensive models.

The purpose of this paper is to apply an iterative method for reducing the uncertainty in a model by means of experimental design to a complex internal combustion engine model. This model employs detailed chemical kinetics, and takes into account turbulent mixing and convective heat transfer. Its computational expense limits the number of evaluations to 10^1-10^3 per day on a single computing core, depending on the case simulated. Since parameter estimation and experimental design techniques tend to require much larger numbers of evaluations, the use of surrogate models is inevitable in this application. We choose quadratic response surfaces as surrogates in this work. The main focus here is to test the Bayesian parameter estimation and the experimental design technique and to demonstrate convergence of the iterative algorithm for the considered internal combustion engine model. For this reason, we restrict ourselves in the first instance to generating "experimental" data by evaluating the model and introducing artificial errors.

The paper is structured as follows. In section 2 we briefly define our methodology, which includes more detailed accounts of the parameter estimation and experimental design techniques, and our choice of surrogate model. In section 3, we introduce the internal combustion engine and the engine model the methodology is applied to. We furthermore present the results for the iteratively improved parameter estimates. Finally, conclusions are drawn and recommendations for future extensions are made in section 4.

2 Methodology

In this section, we define the iterative algorithm and then give details on the parameter estimation and experimental design techniques chosen in this study. Subsequently, the surrogates used are introduced.

2.1 Definitions

We begin by introducing some terminology and notation used throughout. We are concerned with some system or apparatus, such as an internal combustion engine test cell, on which experimental measurements are carried out. Any experiment is characterised by a number of settings prescribed by the experimenter, *i.e.* a vector[1] of

$$\textit{process conditions}: \qquad \xi = (\xi_1, \ldots, \xi_M)^\top$$

with M components. For the engine example, these could include engine speed, load, inlet temperature, *etc*. The results of the measurements can be summarised into a vector of

$$\textit{experimental responses}: \qquad \eta^{\text{exp}} = (\eta_1^{\text{exp}}, \ldots, \eta_L^{\text{exp}})^\top$$

with L components. Examples include peak in-cylinder pressure, emissions, *etc*.

We then consider a model of the system, which takes as inputs the same process conditions as the system itself, and produces as output a vector of

$$\textit{model responses}: \qquad \eta = (\eta_1, \ldots, \eta_L)^\top,$$

which, like the experimental response vector, has L components as well. In this work, we shall not be concerned with multiple candidate models and their discrimination (see for example [17]). In general, a model depends in addition to the process conditions also on a vector of

$$\textit{model parameters}: \qquad \theta = (\theta_1, \ldots, \theta_P)^\top$$

with P components, for example heat transfer coefficients, chemical kinetic rate constants, *etc*. The values of these are unknown *a priori* and need to be determined by parameter estimation.

When we consider sequences of experiments or model evaluations at a sequence of process conditions we denote this by superscript indices in parentheses: The n^{th} experiment, performed at the process conditions $\xi^{(n)} = \left(\xi_1^{(n)}, \ldots, \xi_M^{(n)}\right)^\top$, yields responses $\eta^{\text{exp},(n)} = \left(\eta_1^{\text{exp},(n)}, \ldots, \eta_L^{\text{exp},(n)}\right)^\top$, and the n^{th} model evaluation, also performed at the process conditions $\xi^{(n)}$, yields responses $\eta^{(n)} := \eta(\xi^{(n)}, \theta) = \left(\eta_1^{(n)}, \ldots, \eta_L^{(n)}\right)^\top$.

2.2 General algorithm

Once the system and the model to be considered have been defined, given some experimental data, one can apply parameter estimation techniques to obtain model parameter values which lead to "the best" agreement between experiment and model. It is then natural to use experimental design in order to determine the process conditions at which the next experiment should be carried out such that the uncertainty in the parameter estimates is maximally reduced. This is precisely the algorithm we are adopting here:

1. Design a new experiment, *i.e.* determine a point in process condition space at which the new experiment is to be carried out.

[1] We assume all vectors to be column vectors.

2. Perform the experiment at the chosen point in process condition space.

3. Estimate the model parameters taking into account all experimental data obtained so far.

4. Repeat, *i.e.* go to step 1, until the model parameter estimates satisfy chosen requirements.

Algorithms similar to this have been used in a variety of applications outside of combustion (see for example [1, 4, 17]).

The main aim of this work is to establish convergence of the algorithm for a complex, computationally expensive model for internal combustion engines and thereby to test the chosen parameter estimation and experimental design techniques. Here, we choose the considered system to be the model itself, with some artificially introduced random errors. That is, we generate "experimental" data by evaluating the model using a chosen "true" value of the model parameters θ and adding randomly distributed noise to the model responses (an approach also taken for example in [21, 22]).

2.3 Bayesian parameter estimation

In this section, we explain how model parameters can be estimated with a Bayesian methodology similar to the construction given in [3] for multi-response experimental data. This also extends our previous work on parameter estimation and uncertainty propagation in the fields of granulation [6–8, 27] and combustion [37].

2.3.1 Preliminaries

In the following we set out the Bayesian framework used for the parameter estimation. The main idea of the Bayesian approach is to treat all quantities as random variables and systematically apply probability theory in order to derive conclusions. At any given point in time, our knowledge or belief about the values of the model parameters θ can be represented by a probability density $p(\theta)$, called the *prior* distribution. If new evidence is presented in the form of experimental data, given as a probability function $p(\eta^{\text{exp}}|\theta)$, then the knowledge about the model parameters can be updated, resulting in a *posterior* distribution $p(\theta|\eta^{\text{exp}})$ of the unknown parameters. This is the essence of Bayes' Theorem:

$$p(\theta|\eta^{\text{exp}}) \propto p(\eta^{\text{exp}}|\theta)p(\theta) \tag{1}$$

For ease of reference, we can also state the theorem as:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Thus, in order to make estimates of our model parameters using the posterior distribution, we need to construct the likelihood and the prior distribution.

2.3.2 Likelihood

The likelihood is essentially the distribution of the experimental data. At this point, we need to make an assumption about how the experimental responses are distributed. A common assumption, which we are going to adopt here as well, is that the experimental response is equal to the model response plus a Gaussian error:

$$\eta^{\mathrm{exp},(n)} = \eta\big(\xi^{(n)}, \theta\big) + \varepsilon^{(n)} \qquad \text{with} \qquad \varepsilon^{(n)} \sim \mathcal{N}_L(0, \Sigma), \tag{2}$$

where $\varepsilon^{(n)}$ is the vector of the measurement errors which are normally distributed with zero mean and covariance matrix Σ. It should be pointed out that (2) implicitly assumes the model to be noise-free, or in other words deterministic. For simplicity, the covariance matrix Σ of the experimental errors is assumed here to be independent of the process condition ξ, *i.e.* the system is assumed to be homoscedastic, but the exposition which follows can readily be generalised to the heteroscedastic case. The errors of individual experiments are independent, but the L components of the error vector for any one experiment may be correlated. So, written explicitly, equation (2) implies that the error $\varepsilon^{(n)}$ is distributed according to

$$p\big(\varepsilon^{(n)}|\Sigma\big) = \frac{1}{(2\pi)^{L/2}(\det\Sigma)^{1/2}} \exp\Big(-\tfrac{1}{2}{\varepsilon^{(n)}}^{\top}\Sigma^{-1}\varepsilon^{(n)}\Big).$$

Hence, as $\eta^{\mathrm{exp},(n)} \sim \mathcal{N}_L\big(\eta(\xi^{(n)}, \theta), \Sigma\big)$, the probability of observing a particular response $\eta^{\mathrm{exp},(n)}$ in the n^{th} experiment is given by

$$\begin{aligned} &p\big(\eta^{\mathrm{exp},(n)}\big|\theta, \Sigma\big) \\ &\quad= (2\pi)^{-L/2}(\det\Sigma)^{-1/2} \exp\Big\{-\tfrac{1}{2}\big[\eta^{\mathrm{exp},(n)} - \eta(\xi^{(n)}, \theta)\big]^{\top}\Sigma^{-1}\big[\eta^{\mathrm{exp},(n)} - \eta(\xi^{(n)}, \theta)\big]\Big\} \\ &\quad= (2\pi)^{-L/2}(\det\Sigma)^{-1/2} \exp\Big(-\tfrac{1}{2}{\varepsilon^{(n)}}^{\top}\Sigma^{-1}\varepsilon^{(n)}\Big). \end{aligned} \tag{3}$$

Making use of the assumption of independent experiments, the likelihood becomes

$$\begin{aligned} p\big(\eta^{\mathrm{exp},(1)}, \ldots, \eta^{\mathrm{exp},(N)}\big|\theta, \Sigma\big) &= \prod_{n=1}^{N} p\big(\eta^{\mathrm{exp},(n)}\big|\theta, \Sigma\big) \\ &= (2\pi)^{-NL/2}(\det\Sigma)^{-N/2} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}{\varepsilon^{(n)}}^{\top}\Sigma^{-1}\varepsilon^{(n)}\right). \end{aligned}$$

If we further introduce the $L \times L$ positive definite matrix

$$S(\theta) := \sum_{n=1}^{N} \varepsilon^{(n)}{\varepsilon^{(n)}}^{\top},$$

then the likelihood simplifies to

$$p\big(\eta^{\mathrm{exp},(1)}, \ldots, \eta^{\mathrm{exp},(N)}\big|\theta, \Sigma\big) = (2\pi)^{-NL/2}(\det\Sigma)^{-N/2} \exp\big\{-\tfrac{1}{2}\mathrm{tr}\big[\Sigma^{-1}S(\theta)\big]\big\}. \tag{4}$$

2.3.3 Prior distributions

The choice of prior distributions is a non-trivial, much-debated subject in the literature, and we shall make no attempt at an exhaustive treatment here. Noting that we have two objects, θ and Σ, for which we need to state prior beliefs, we begin by assuming their independence:

$$p(\theta, \Sigma) = p(\theta)p(\Sigma) \tag{5}$$

For our prior of θ, we consider a constant (uniform) distribution over a hypercube $\mathcal{C}$ which is defined as the region in P-dimensional space such that $\theta_j \in [-1, 1]$ for all $j = 1, \ldots, P$, which gives a prior probability density for θ as

$$p(\theta) = \frac{1}{|\mathcal{C}|}\mathbb{1}_{\{\theta\in\mathcal{C}\}} \tag{6}$$

where $|\cdot|$ denotes the size/volume of a set and $\mathbb{1}_{\{\cdot\}}$ is the indicator function.

For the prior of Σ, one can distinguish two scenarios: If Σ is known *a priori*, this case is referred to as *informative*. For example, one may want to use the estimator $\hat{\Sigma}$ for the covariance matrix, also called empirical covariance matrix, which has been obtained from previously collected experimental data. In this case, we can choose as an informative prior

$$p(\Sigma) = \delta(\Sigma - \hat{\Sigma}). \tag{7}$$

If Σ is unknown – the case referred to as *non-informative* – we choose the Inverse-Wishart non-informative prior

$$p(\Sigma) \propto (\det \Sigma)^{-\alpha-(L+1)/2} \exp\left[-\tfrac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Psi\right)\right], \tag{8}$$

where $\alpha > 0$ and $\Psi \in \mathbb{R}^{L\times L}$ positive definite are arbitrary parameters. The Inverse-Wishart non-informative prior is the multivariate generalisation of the widely used Inverse-Gamma non-informative prior. These priors are termed proper, meaning that they are genuine probability densities which integrate to unity, in contrast to improper priors which integrate to infinity [5].

2.3.4 Posterior distributions

Using Bayes' Theorem, we can now compute the posterior densities (up to constant factors):

$$\begin{aligned}
&p\left(\theta, \Sigma \middle| \eta^{\text{exp},(1)}, \ldots, \eta^{\text{exp},(N)}\right) \\
&\quad \overset{\text{Eqn. (1)}}{\propto} p\left(\eta^{\text{exp},(1)}, \ldots, \eta^{\text{exp},(N)} \middle| \theta, \Sigma\right) p(\theta, \Sigma) \\
&\quad \overset{\text{Eqn. (5)}}{\propto} p\left(\eta^{\text{exp},(1)}, \ldots, \eta^{\text{exp},(N)} \middle| \theta, \Sigma\right) p(\theta)p(\Sigma) \\
&\quad \overset{\text{Eqn. (6)}}{\propto} p\left(\eta^{\text{exp},(1)}, \ldots, \eta^{\text{exp},(N)} \middle| \theta, \Sigma\right) p(\Sigma) \cdot \mathbb{1}_{\{\theta\in\mathcal{C}\}} \\
&\quad \overset{\text{Eqn. (4)}}{\propto} (\det \Sigma)^{-N/2} \exp\left\{-\tfrac{1}{2}\mathrm{tr}\left[\Sigma^{-1}S(\theta)\right]\right\} p(\Sigma) \cdot \mathbb{1}_{\{\theta\in\mathcal{C}\}}
\end{aligned}$$

In the informative case, inserting Eqn. (7), we obtain

$$p\big(\theta, \Sigma \big| \eta^{\text{exp},(1)}, \dots, \eta^{\text{exp},(N)}\big) \\ \propto (\det \Sigma)^{-N/2} \exp\Big\{ -\tfrac{1}{2}\text{tr}\big[\Sigma^{-1} S(\theta)\big]\Big\} \delta(\Sigma - \hat{\Sigma}) \cdot \mathbb{1}_{\{\theta \in \mathcal{C}\}}, \tag{9}$$

whereas in the non-informative case, inserting Eqn. (8), we obtain

$$p\big(\theta, \Sigma \big| \eta^{\text{exp},(1)}, \dots, \eta^{\text{exp},(N)}\big) \\ \propto (\det \Sigma)^{-\alpha-(N+L+1)/2} \exp\Big\{ -\tfrac{1}{2}\text{tr}\big[\Sigma^{-1}\big(S(\theta) + \Psi\big)\big]\Big\} \cdot \mathbb{1}_{\{\theta \in \mathcal{C}\}}. \tag{10}$$

As we are interested in the marginal posterior density for θ, we integrate over all positive definite matrices Σ. In the informative case, using Eqn. (9), the posterior is obtained as

$$p\big(\theta \big| \eta^{\text{exp},(1)}, \dots, \eta^{\text{exp},(N)}\big) = \int_{\Sigma \text{ pos. def.}} p\big(\theta, \Sigma \big| \eta^{\text{exp},(1)}, \dots, \eta^{\text{exp},(N)}\big) \mathrm{d}\Sigma \\ \propto \exp\Big\{ -\tfrac{1}{2}\text{tr}\big[\hat{\Sigma}^{-1} S(\theta)\big]\Big\} \cdot \mathbb{1}_{\{\theta \in \mathcal{C}\}}. \tag{11}$$

and in the non-informative case, using Eqn. (10), we obtain [5]

$$p\big(\theta \big| \eta^{\text{exp},(1)}, \dots, \eta^{\text{exp},(N)}\big) = \int_{\Sigma \text{ pos. def.}} p\big(\theta, \Sigma \big| \eta^{\text{exp},(1)}, \dots, \eta^{\text{exp},(N)}\big) \mathrm{d}\Sigma \\ \propto \big[\det\big(S(\theta) + \Psi\big)\big]^{-\alpha - N/2} \cdot \mathbb{1}_{\{\theta \in \mathcal{C}\}}. \tag{12}$$

We have determined the expressions for the posteriors (11) and (12) only up to constant positive factors. While these normalisation factors can in principle be found, they turn out not to be needed. The posterior densities can be used to derive credibility regions (preferable to means and variance) of the unknown parameters as well as marginal distributions.

2.4 Experimental design

Experimental design can be regarded as an attempt to answer the question which future experiments are most informative in order to reduce the estimated parametric uncertainty in the model. In other words, experimental design provides a suggestion at which process conditions ξ the next experiment should be performed in order to maximally reduce the uncertainty in the unknown model parameters θ.

The experimental design technique we choose in this work is based on maximising the Fisher information matrix [2, 16, 20] in some sense. The Fisher information matrix M is defined as

$$M(\xi^{(1)}, \dots, \xi^{(N)}; \theta) \\ := \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log \prod_{i=1}^{N} p\big(\eta^{\text{exp},(i)} \big| \theta, \Sigma\big)\right)\left(\frac{\partial}{\partial\theta} \log \prod_{j=1}^{N} p\big(\eta^{\text{exp},(j)} \big| \theta, \Sigma\big)\right)^{\top}\right] \\ = \mathbb{E}\left[\left(\sum_{i=1}^{N} \frac{\partial}{\partial\theta} \log p\big(\eta^{\text{exp},(i)} \big| \theta, \Sigma\big)\right)\left(\sum_{j=1}^{N} \frac{\partial}{\partial\theta} \log p\big(\eta^{\text{exp},(j)} \big| \theta, \Sigma\big)\right)^{\top}\right].$$

The Fisher matrix is based on a set of N experiments here, and we take the view that the first $N-1$ of these have already been performed, at the process conditions $\xi^{(1)}, \ldots, \xi^{(N-1)}$, and the N^{th} experiment is the "next" one, for which we wish to determine the as yet unknown process conditions $\xi^{(N)}$. The derivatives of $\log p\big(\eta^{\text{exp},(n)}\big|\theta, \Sigma\big)$ with respect to θ are obtained from Eqn. (3) as

$$\begin{aligned}\frac{\partial \log p\big(\eta^{\text{exp},(i)}\big|\theta, \Sigma\big)}{\partial \theta} &= -\frac{1}{2}\left(\Big[\frac{\partial \eta(\xi^{(n)}, \theta)}{\partial \theta}\Big]^\top \Sigma^{-1}\varepsilon^{(n)} + {\varepsilon^{(n)}}^\top \Sigma^{-1}\frac{\partial \eta(\xi^{(n)}, \theta)}{\partial \theta}\right) \\ &= \Big[\frac{\partial \eta(\xi^{(n)}, \theta)}{\partial \theta}\Big]^\top \Sigma^{-1}\varepsilon^{(n)},\end{aligned}$$

where in the last step we have made use of the fact that Σ^{-1}, just like Σ, is symmetric. We introduce the following notation for the derivatives of the model function with respect to the model parameters θ for any arbitrary ξ, *i.e.* the Jacobi matrix,

$$J(\xi, \theta) := \frac{\partial \eta(\xi, \theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial \eta_1(\xi,\theta)}{\partial \theta_1} & \cdots & \frac{\partial \eta_1(\xi,\theta)}{\partial \theta_P} \\ \vdots & \ddots & \vdots \\ \frac{\partial \eta_L(\xi,\theta)}{\partial \theta_1} & \cdots & \frac{\partial \eta_L(\xi,\theta)}{\partial \theta_P} \end{pmatrix}. \tag{13}$$

Furthermore, we use $J^{(n)}$ as shorthand for $J\big(\xi^{(n)}, \theta\big)$. We then obtain the following expression for the Fisher information matrix:

$$\begin{aligned} &M\big(\xi^{(1)}, \ldots, \xi^{(N-1)}, \xi^{(N)}; \theta\big) \\ &\quad= \mathbb{E}\left[\left(\sum_{i=1}^{N} {J^{(i)}}^\top \Sigma^{-1}\,\varepsilon^{(i)}\right) \times \left(\sum_{j=1}^{N} {J^{(j)}}^\top \Sigma^{-1}\,\varepsilon^{(j)}\right)^\top\right] \\ &\quad= \mathbb{E}\left[\sum_{i=1}^{N} {J^{(i)}}^\top \Sigma^{-1}\,\varepsilon^{(i)} \cdot \sum_{j=1}^{N} {\varepsilon^{(j)}}^\top \Sigma^{-1} J^{(j)}\right] \\ &\quad= \sum_{i=1}^{N}\sum_{j=1}^{N} {J^{(i)}}^\top \Sigma^{-1} \underbrace{\mathbb{E}\big[\varepsilon^{(i)}{\varepsilon^{(j)}}^\top\big]}_{=0 \text{ if } i\neq j} \Sigma^{-1} J^{(j)} \\ &\quad= \sum_{i=1}^{N} {J^{(i)}}^\top \Sigma^{-1} \underbrace{\mathbb{E}\big[\varepsilon^{(i)}{\varepsilon^{(i)}}^\top\big]}_{=\Sigma} \Sigma^{-1}\, J^{(i)} \\ &\quad= \sum_{i=1}^{N} {J^{(i)}}^\top \Sigma^{-1} J^{(i)} \\ &\quad= \sum_{i=1}^{N-1} {J^{(i)}}^\top \Sigma^{-1} J^{(i)} + {J^{(N)}}^\top \Sigma^{-1} J^{(N)}. \end{aligned} \tag{14}$$

In the last line, we have summarised all contributions from the previous experiments, performed at $\xi^{(1)}, \ldots, \xi^{(N-1)}$, into the first term, and separated out the second term which is the only one featuring the experimental responses at the unknown, sought-after process

conditions. We note that due to the way the responses enter the Fisher matrix (14) via (13), intuitively, one would expect the "information" to be "large" in regions of process condition space where the responses are most sensitive to the model parameters.

In order to determine the process conditions $\xi^{(N)}$ of the next experiment, we maximise the Fisher information matrix in some sense. As the only dependence of the Fisher matrix we are interested in is that on the new, sought-after process condition, we abbreviate $M(\xi) := M(\xi^{(1)}, \ldots, \xi^{(N-1)}; \xi, \theta)$. The meaning of "maximal" is not uniquely defined for a matrix, and thus, maximisation of the Fisher matrix requires choosing a scalar objective function Φ, for which matrix invariants are commonly used. We consider here the following four optimality criteria, *i.e.* objective functions (see for example [17] for a brief review):

$$\text{A:} \quad \Phi(\xi) = -\mathrm{tr}\big(M^{-1}(\xi)\big) \tag{15}$$

$$\text{D:} \quad \Phi(\xi) = \det\big(M(\xi)\big) \tag{16}$$

$$\text{E:} \quad \Phi(\xi) = \min_i \big\{\lambda_i\big(M(\xi)\big)\big\} \tag{17}$$

$$\text{T:} \quad \Phi(\xi) = \mathrm{tr}\big(M(\xi)\big) \tag{18}$$

In equation (17), $\lambda_i(\cdot)$ denotes the i^{th} eigenvalue of its argument. The process conditions ξ^* of the next experiment to be carried out are then obtained by

$$\xi^* := \operatorname*{argmax}_{\xi} \Phi(\xi). \tag{19}$$

This optimisation problem is usually subject to constraints, either for physical reasons (*e.g.* a pressure has to be positive) or for reasons imposed by the apparatus or costs. Here, we apply constraints of the form $\xi_{\text{lb}} \leq \xi \leq \xi_{\text{ub}}$, with ξ_{lb} and ξ_{ub} denoting lower and upper bounds respectively.

2.5 Surrogate models

The described methodology involves, both in the parameter estimation and in the experimental design parts, large numbers of model evaluations, potentially in excess of 10^5. As indicated in the introduction, however, in practically relevant applications, which are our aim here, model evaluations are typically expensive, so one may be restricted to a much smaller number of evaluations. In such applications, therefore, it is inevitable to replace the actual model by a surrogate. In the current study, second order (*i.e.* quadratic) response surfaces are chosen as surrogate models. These have found wide-spread use in numerous applications [9, 12, 18, 38]. A large number of surrogate models, also called emulators or meta-models, have been proposed in the literature in a wide variety of contexts (for example [30, 33, 45]), but it is beyond the scope of the present work to investigate in detail which type of surrogate is most suitable.

Specifically, the l^{th} response of the model evaluated at the n^{th} process condition $\xi^{(n)}$ is replaced by the second order polynomial in model parameter space

$$\eta_l^{(n)}(\theta) = \beta_{l,0}^{(n)} + \sum_{i=1}^{P} \beta_{l,i}^{(n)} \theta_i + \sum_{i=1}^{P} \sum_{j \geq i}^{P} \beta_{l,ij}^{(n)} \theta_i \theta_j, \tag{20}$$

where $\beta_{l,0}^{(n)}$, $\beta_{l,i}^{(n)}$, and $\beta_{l,ij}^{(n)}$ are the constant, linear, and quadratic coefficients of the response surface respectively.

We determine the coefficients of our surrogate models (20) by performing least-squares fits to model evaluations on a Central Composite Design (CCD) in model parameter space. A CCD includes all points of a full factorial design (2^P corner points of the hypercube), two points on every axis, located symmetrically, and the centre point [32].

Calculating the Fisher information matrix according to (14) involves evaluating the Jacobi matrix $J(\xi, \theta)$ (Eqn. 13) at various ξ. It is straightforward to evaluate all but the last term of the Fisher matrix, since we can calculate the model responses for the process conditions $\xi^{(1)}, \ldots, \xi^{(N-1)}$ from the response surface (20). For the unknown process conditions $\xi^{(N)}$, the last term in (14), however, this approach is not suitable, because during the optimisation process a large number of evaluations at different ξs not known *a priori* needs to be performed. In order to overcome this problem, we create another set of response surfaces based on all the points that have been evaluated using the full model for the construction of the response surfaces at the process conditions $\xi^{(1)}, \ldots, \xi^{(N-1)}$. This new set of response surfaces is now defined on the combined space of the model parameters $\theta_1, \ldots, \theta_P$ and process conditions $\xi_1, \ldots, \xi_M$. Therefore, denoting the combined vector as $\vartheta = (\theta_1, \ldots, \theta_P, \xi_1, \ldots, \xi_M) = (\vartheta_1, \ldots, \vartheta_{P+M})$, the new response surface of the l^{th} response ($l = 1, \ldots, L$) takes the form

$$\gamma_l(\theta, \xi) = \gamma_l(\vartheta) = \beta_{l,0} + \sum_{i=1}^{P+M} \beta_{l,i}\vartheta_i + \sum_{i=1}^{P+M} \sum_{j \geq i}^{P+M} \beta_{l,ij}\vartheta_i\vartheta_j. \tag{21}$$

2.6 Remarks on implementation

Creation of the response surfaces was implemented in C++ in a code we call Model Development Suite (MoDS). It is designed to work with any model, treated as a black box, given as an executable which performs input and output through files in various formats including XML and delimiter separated value files (comma, tab, space, *etc*.).

The Bayesian parameter estimation described in subsection 2.3 was implemented in R. In order to obtain samples from the posterior distributions (11) and (12) we employ the Metropolis-Hastings algorithm [19, 28]. This algorithm employs a continuous-space discrete-time Markov Chain which has a stationary distribution identical to the distribution to be sampled from. The collection of samples is then used to derive for example credibility regions (preferable to means and variance) of the unknown parameters as well as marginal densities, which can be plotted using kernel density estimation.

The experimental design part described in subsection 2.4 was implemented in Matlab. The optimisation was carried out using the `fmincon` function. Derivatives of the model, *i.e.* of the response surfaces (20) and (21), required for the calculation of the Fisher matrix (14) through (13), were determined analytically.

3 Application to an internal combustion engine model

In this section, we apply the methodology introduced above to a complex model for internal combustion engines.

3.1 The considered engine

The internal combustion engine considered here is a heavy-duty Diesel engine converted for Homogeneous Charge Compression Ignition (HCCI) operation – a mode attractive due to ultra-low emissions of NO_x and particulate matter. The engine is operated with iso-octane as a fuel at constant load in steady-state. Table 1 summarises the engine specification and operating condition.

Table 1: *Engine specification and operating condition.*

Operating mode	HCCI
Bore	137.2 mm
Stroke	171.5 mm
Compression ratio	16.0-16.5
Inlet temperature	85°C-95°C
Speed	1800 RPM
Fuel	iso-octane
Equivalence ratio	0.3

3.2 The considered engine model

3.2.1 Model description

The model we use in this work to simulate the engine described above is the Stochastic Reactor Model (SRM), which has been successfully employed in a number of earlier studies such as port fuel injected HCCI combustion [26], single early direct injection HCCI [42], and dual injection HCCI [29]. Its applications include soot formation in HCCI [31] and direct injection spark ignition (DISI) engines [14], and cycle-to-cycle variations in spark ignition (SI) engines [15]. In [30], a surrogate model based on cubic natural splines was developed for the SRM and applied to multi-cycle transient simulation and control problems.

The SRM was originally inspired by Probability Density Function (PDF) transport methods. It employs detailed chemical kinetics and possesses sub-models for turbulent mixing

and convective heat transfer, and a highly detailed population balance model for soot formation.

As chemical kinetic fuel model, we use here a reduced mechanism for Primary Reference Fuels (PRFs, mixtures of n-heptane and iso-octane) consisting of 38 species and 44 reactions [41].

Since it includes detailed chemistry, the SRM can qualitatively predict emission trends of CO, CO_2, NO_x, and unburnt hydrocarbons at modest computational cost of the order of one hour on a single-core processor per engine cycle. While this is computationally relatively cheap compared to other approaches, run-times are still too large for the numbers of evaluations required here. Therefore, the use of surrogates is inevitable in this case.

The SRM contains by design several random sub-models. For the present work, however, the model is operated in such a way that it does not make use of any of its stochastic components, and hence is free of noise, as required by assumption (2).

3.2.2 Choice of process conditions, model parameters, and responses

Table 2: *Variables and their bounds.*

Variable type	Variable name	Lower bound	Upper bound	Coding
Process conditions	$\xi_1 = \mathrm{CR}$	16.0	16.5	none
	$\xi_2 = \mathrm{T_{in}}$	85°C	95°C	none
Model parameters	$\theta_1 = A_{33}$	$3.0 \times 10^6 \frac{\mathrm{cm}^3}{\mathrm{mol \cdot s}}$	$1.2 \times 10^7 \frac{\mathrm{cm}^3}{\mathrm{mol \cdot s}}$	logarithmic
	$\theta_2 = A_{37}$	$1.09 \times 10^{17} \frac{\mathrm{cm}^3}{\mathrm{mol \cdot s}}$	$1.32 \times 10^{17} \frac{\mathrm{cm}^3}{\mathrm{mol \cdot s}}$	linear
Responses	$\eta_1 = p_{\mathrm{max}}$	n/a	n/a	none
	$\eta_2 = \mathrm{CA50}$	n/a	n/a	none

An internal combustion engine experiment may require large numbers of process conditions and typically yields large numbers of responses. Likewise, any (sufficiently detailed) engine model, such as the SRM, has large numbers of process conditions, responses, and model parameters, possibly as many as of order 10^4 if one includes the parameters in the chemical kinetic fuel model. Here, as a first step, we consider the SRM as a function which maps two process conditions and two model parameters to two responses:

$$\mathrm{SRM} : (\xi_1, \xi_2; \theta_1, \theta_2) \mapsto (\eta_1, \eta_2)$$

As process conditions, we choose compression ratio CR and inlet temperature $\mathrm{T_{in}}$, so we have $\xi = (\xi_1, \xi_2) = (\mathrm{CR}, \mathrm{T_{in}})$. As model parameters, we choose $\theta = (\theta_1, \theta_2) = (A_{33}, A_{37})$, where A_{33} and A_{37} denote the forward pre-exponential factors of the rate expressions for reaction 33 and 37 of the chemical kinetic fuel model respectively:

$$\mathrm{CO + OH} \rightleftharpoons \mathrm{CO_2 + H} \qquad \text{(Reaction 33)}$$
$$\mathrm{H_2O_2 + M} \rightleftharpoons \mathrm{2OH + M} \qquad \text{(Reaction 37)}$$

As responses, we choose the peak in-cylinder pressure $p_{\max}$ and the crank angle at 50% heat release CA50. So we have $\eta = (\eta_1, \eta_2) = (p_{\max}, \mathrm{CA50})$. In short, the model function we consider is

$$\mathrm{SRM} : (\mathrm{CR}, \mathrm{T_{in}}; A_{33}, A_{37}) \mapsto (p_{\max}, \mathrm{CA50}).$$

Table 2 lists the variables used together with their chosen bounds. The bounds were determined such that stable combustion is achieved over the entire range, with CA50, *i.e.* essentially ignition timing, varying between about 0 and 10 CAD aTDC.

The last column in table 2 describes the type of normalisation or mapping of variable ranges $\theta_{\mathrm{lb}} \leq \theta \leq \theta_{\mathrm{ub}}$ to the interval[2] $[-1, +1]$, often referred to as *coding*. Two types of mapping are usually applied: linear

$$\theta_i' = \frac{\theta_i - (\theta_{i,\mathrm{ub}} + \theta_{i,\mathrm{lb}})/2}{(\theta_{i,\mathrm{ub}} - \theta_{i,\mathrm{lb}})/2}$$

and logarithmic coding

$$\theta_i' = \frac{\log \theta_i - (\log \theta_{i,\mathrm{ub}} + \log \theta_{i,\mathrm{lb}})/2}{(\log \theta_{i,\mathrm{ub}} - \log \theta_{i,\mathrm{lb}})/2}.$$

The main reason for coding variables is to avoid numerical ill-conditioning in some algorithms for quantities with vastly different orders of magnitude. Logarithmic coding is preferable to linear coding for quantities which vary over an order of magnitude or more. In the following, coded variables are indicated by a prime.

3.2.3 "Experimental" data

As the main aim of the present work is to test the methodology under well-defined conditions, we generate "experimental" data by evaluating the model using some "true" values of the model parameters θ – an approach also pursued in [22] for example. We arbitrarily choose the true system to be given by $\theta = (0, 0)$. In line with assumption (2), the model is deterministic, *i.e.* it does not contain random components. Therefore, we artificially introduce error by adding normally distributed noise with covariance matrix

$$\Sigma = \begin{pmatrix} 0.04\ \mathrm{bar}^2 & -0.028\ \mathrm{bar} \cdot \mathrm{CAD} \\ -0.028\ \mathrm{bar} \cdot \mathrm{CAD} & 0.02\ \mathrm{CAD}^2 \end{pmatrix} \tag{22}$$

to the responses η. The distribution of the responses, generated using 10^3 samples, is shown in Figure 1.

3.3 Results

One of the key ingredients for performing parameter estimation and experimental design with an expensive model is the surrogate model. Figure 2 shows a typical example of a response surface (Eqn. 20) fitted to a central composite design in coded model parameter

[2] The unit interval $[0, 1]$ is also frequently used.

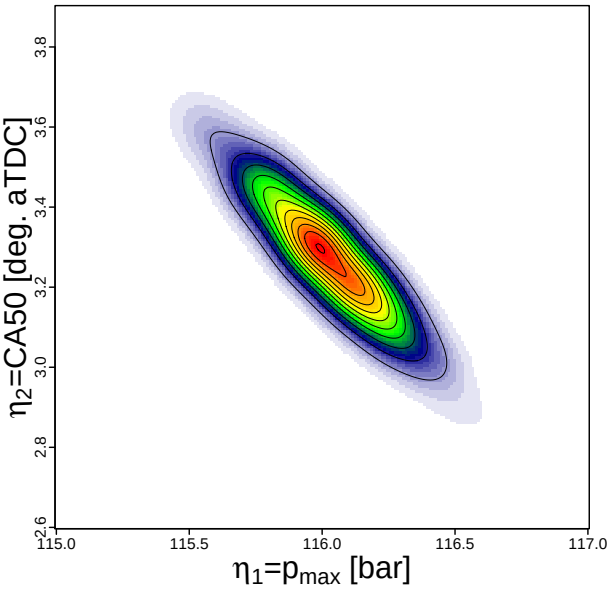


Figure 1: *Normally distributed "experimental" responses at* $\xi = (16.25, 90°C)$*, centred at "true" model response with covariance matrix* (22)*.*

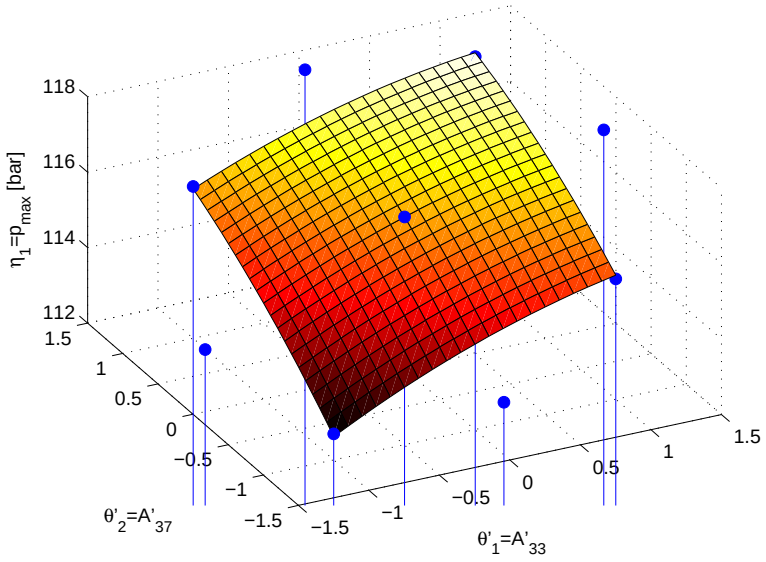


Figure 2: *Quadratic response surface in model parameter space for the first response* η_1 *(peak in-cylinder pressure) fitted to a Central Composite Design.*

space for the first response η_1, the peak pressure. The points represent the actual model evaluations. A CCD in a two-dimensional space consists of nine points: the four corner points of the hypercube, plus two points on both coordinate axes, located symmetrically (we choose $\pm\sqrt{2}$, resulting in a rotatable design), and the origin. In the particular application considered in this work, the response surfaces reproduce the actual model evaluations very well throughout. While some curvature is present as expected, the behaviour of the responses does not appear to deviate significantly from quadratic. Indeed that was the

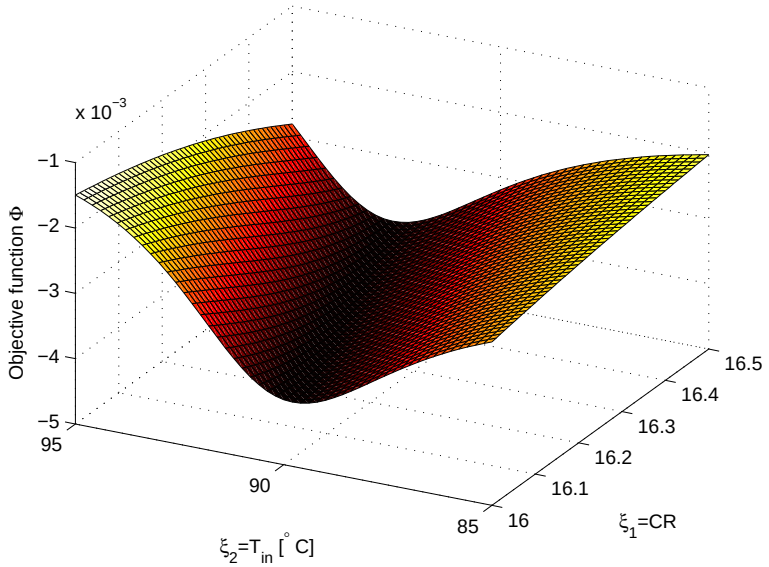


Figure 3: *Objective function surface for criterion A in iteration 2.*

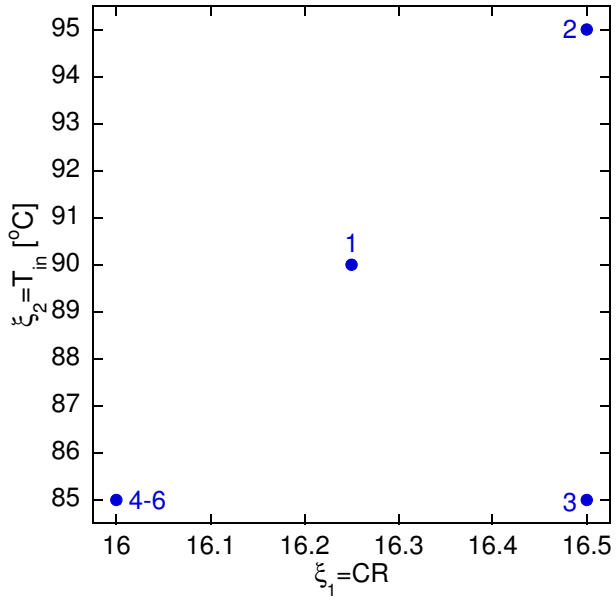
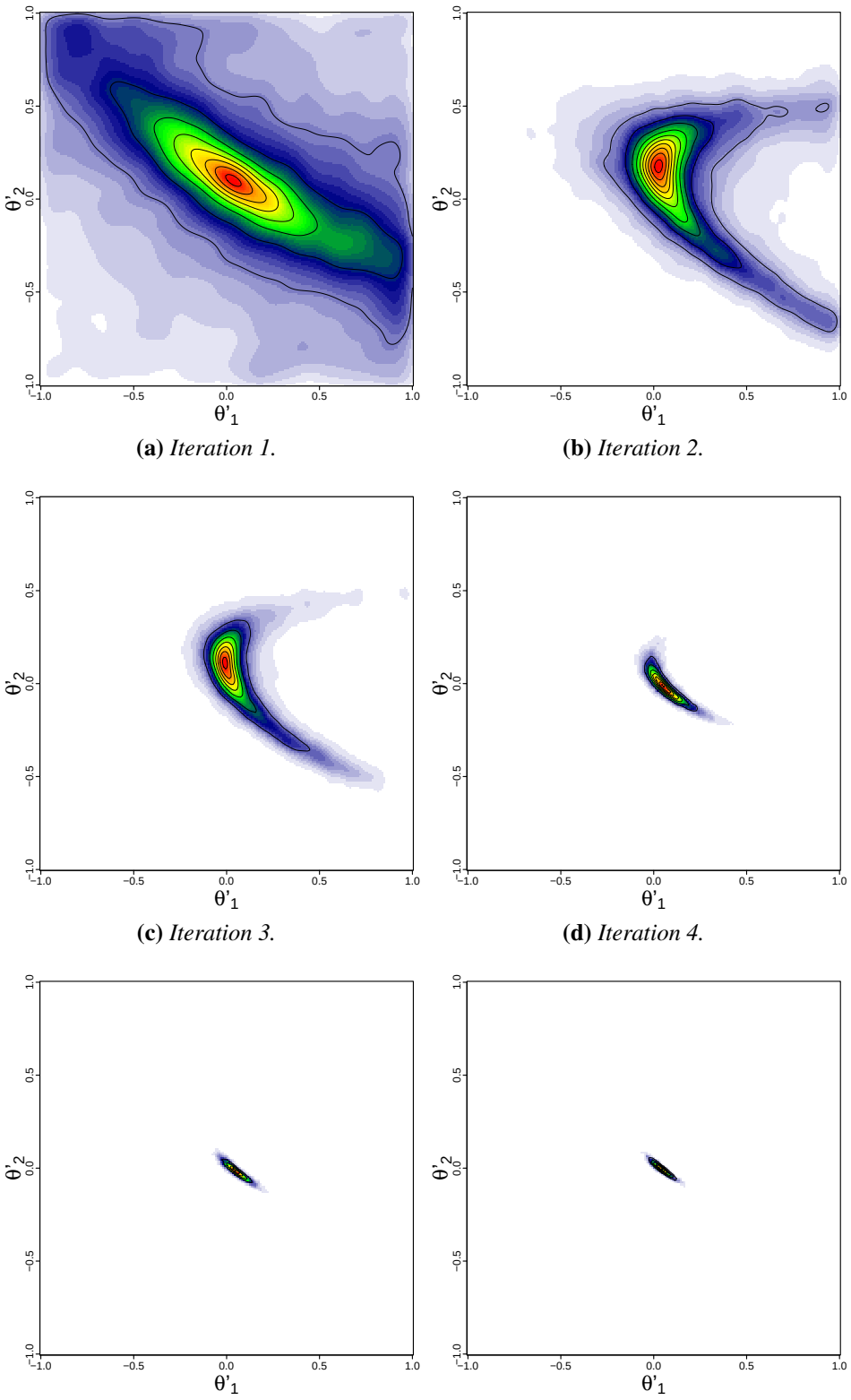


Figure 4: *Sequence of points in process condition space (compression ratio vs. inlet temperature). The first point was chosen arbitrarily, points 2 to 6 were suggested by the experimental design method using criteria D and T.*

main reason for choosing the hypercube in process condition space around and close to a stable engine operating point. At or near a critical operating point, say at the misfire boundary, one would expect strongly non-linear and curved surfaces. Quadratic surfaces may not be sufficient for such conditions.

We carried out six iterations of the algorithm described in subsection 2.2. As first "experimental" data point, we arbitrarily take the midpoint of the hypercube in process condition space, *i.e.* $\xi = (16.25, 90°\mathrm{C})$. We then perform the parameter estimation, and use the

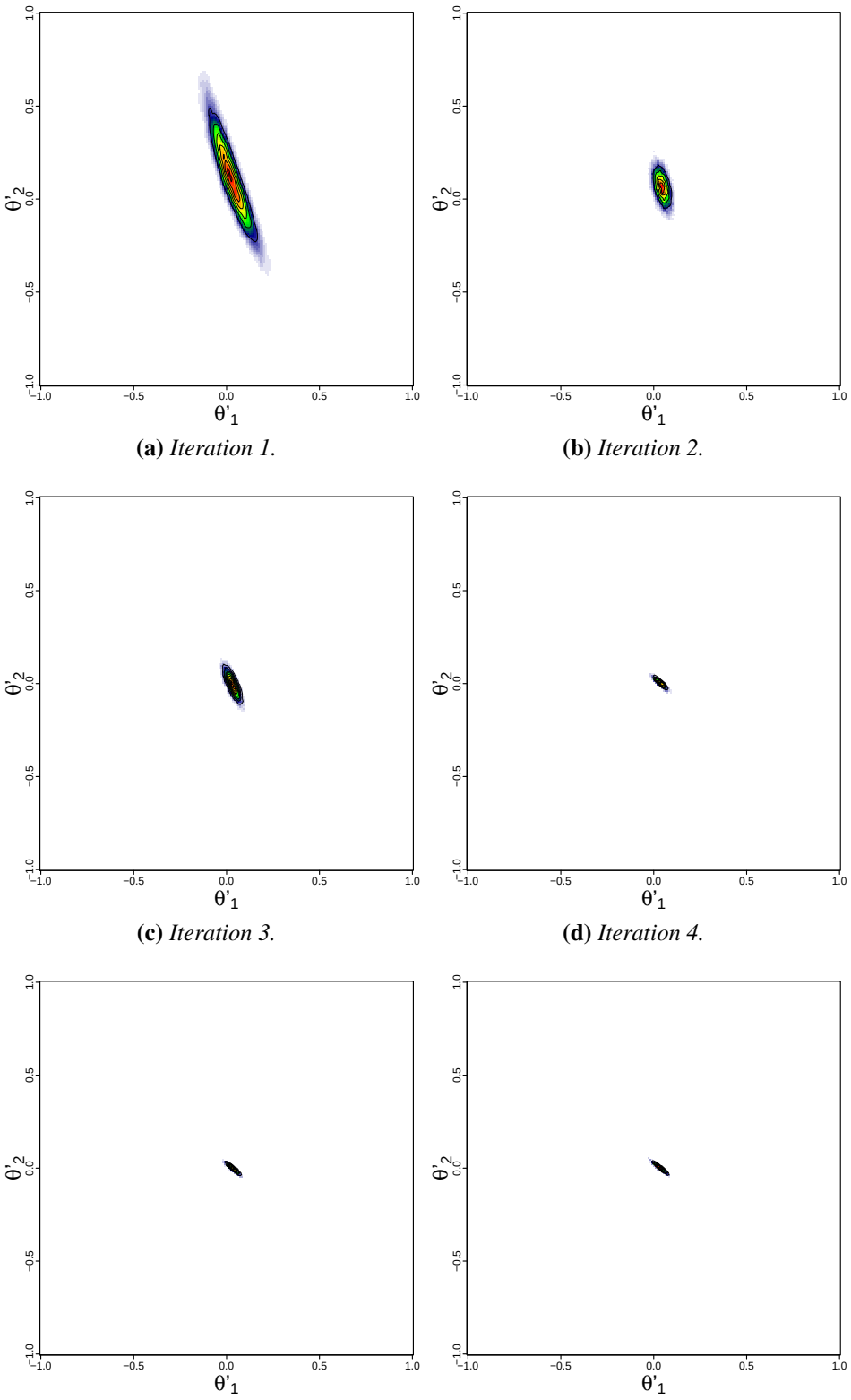


(a) *Iteration 1.* **(b)** *Iteration 2.*

(c) *Iteration 3.* **(d)** *Iteration 4.*

(e) *Iteration 5.* **(f)** *Iteration 6.*

Figure 5: *Evolution of the posterior densities of the coded model parameter θ' using the Inverse-Wishart non-informative prior through six iterations obtained with criteria D and T. The "true" values are $\theta' = (\theta'_1, \theta'_2) = (0, 0)$.*



(a) *Iteration 1.* **(b)** *Iteration 2.*

(c) *Iteration 3.* **(d)** *Iteration 4.*

(e) *Iteration 5.* **(f)** *Iteration 6.*

Figure 6: *Evolution of the posterior densities of the coded model parameter θ' using an informative prior through six iterations obtained with criteria D and T. The "true" values are $\theta' = (\theta'_1, \theta'_2) = (0, 0)$.*

experimental design technique to determine the next set of process conditions. The chosen experimental design methodology was found to frequently suggest corner points of the hypercube, which may not be surprising given the mathematical form of the Fisher matrix (14). Figure 3 shows a typical example of an objective function surface, here for the A-optimality criterion (15), which exhibits multiple local maxima located in different corner points of the hypercube. This implies that, when trying to identify the global maximum within the hypercube using local optimisation methods, care needs to be taken, as in the example shown, starting from the midpoint with a gradient-based method would lead to an incorrect corner point being identified as the maximum. The objective function surfaces do vary with the best estimate θ, of course, as the Fisher matrix (14) is a function of θ. However, even though the θs estimated using non-informative and informative priors differ, in none of the cases studied this made a difference as to which point was suggested for the next experiment.

In general, one would expect the different optimality criteria to suggest different points in process condition space, but in this case it was found that there is frequent agreement. Criteria D and T agreed in all iterations, and so did A and E. In the first iteration, all tested optimality criteria (A, D, E, and T) suggest $\xi^* = (16.5, 95°\mathrm{C})$ as the new experiment. We arbitrarily choose the D and T-optimality criteria, *i.e.* the determinant and trace of the Fisher information matrix. The sequence of points is shown in Figure 4.

Figures 5 and 6 show results of the Bayesian parameter estimation for the first six iterations, for the non-informative and the informative cases respectively. Recall that the "true" parameter values are $\theta' = (0, 0)$. Figure 5 shows the evolution of the posterior density of θ for the Inverse-Wishart non-informative prior. As parameters of this prior, we use $\alpha = \varepsilon$ and

$$\Psi = \begin{pmatrix} 2\varepsilon & 0 \\ 0 & 2\varepsilon \end{pmatrix},$$

with $\varepsilon = 10^{-3}$. All posterior density plots have been generated using 10^5 samples. Of these samples, the first 5×10^3 were discarded due to the fact that the Markov chain in the Metropolis-Hastings algorithm goes through an initial phase, known as *burn-in*, required to "forget" its initial state. It turns out that mean and standard deviation of the posterior densities in general do not yield very good values for the parameters and their uncertainties. Instead, the value of highest probability density together with the bounds of the high probability density region (at some given confidence level) should be used [4]. While this procedure leads to a good "best" estimate with error bars, one should bear in mind that the densities contain much more information, and in cases where the distributions are multimodal, it may not even make sense to pick a best estimate or an error bar. The algorithm is found to converge very rapidly in both the informative and the non-informative case, with the informative estimates significantly more accurate than the non-informative ones. This agrees with intuitive expectation, as the knowledge of the covariance matrix is made use of in the informative case.

4 Conclusions

We have implemented an algorithm which estimates model parameters using a Bayesian approach and employs an experimental design technique to iteratively suggest process conditions for new experiments with the aim of decreasing the uncertainty in the parameter estimates. The main focus here was the application to a complex internal combustion engine model whose computational expense limits the number of model evaluations to an extent which necessitates the use of a surrogate model. Quadratic response surfaces were chosen as surrogates in this work. We have found rapid convergence of the algorithm in the considered case, as shown by a substantial reduction in parametric uncertainty over six iterations, for both informative and non-informative priors used in the parameter estimation.

Regarding extensions of the present work, an important assumption has been that the experimental responses are equal to the model responses plus some Gaussian error (Eqn. 2), which has a number of implications. First of all, in practice, distributions of experimental errors may deviate from Gaussian which needs to be accounted for. Ignoring this may lead to systematically incorrect parameter estimates even with high credibility. Secondly, how much noise in the experimental data is tolerable? One would expect that more noise slows convergence, but at what point does the methodology break down? Thirdly, for stochastic models, ones which contain random components themselves, it needs to be investigated how much noise in the model is tolerable. And lastly, in practice, virtually all models have a systematic bias, *i.e.* they are unable to reproduce certain features of the data. If information on this is available *a priori*, it can be taken into account (see for example [22]), but this may not be possible in general. Furthermore, it is clear that if the model exhibits worse than quadratic non-linearities, more advanced surrogate models will be required. It is less clear, though, what type of surrogate is most suitable in a particular case. Several of these questions have been addressed individually at least to some extent in the literature, but await application to a relevant combustion problem such as an internal combustion engine as considered in this work.

Acknowledgements

Funding by the EPSRC, grant number EG/G028672/1, is gratefully acknowledged.

References

[1] S. P. Asprey and S. Macchietto. Statistical tools for optimal dynamic model building. *Computers and Chemical Engineering*, 24(2–7):1261–1267, 2000. doi:10.1016/S0098-1354(00)00328-8.

[2] Y. Bard. *Nonlinear Parameter Estimation*. Academic Press, New York, 1974.

[3] K. J. Beers. *Numerical Methods for Chemical Engineering*. Cambridge University Press, Cambridge, 1st edition, 2007.

[4] G. Blau, M. Lasinski, S. Orcun, S.-H. Hsu, J. Caruthers, N. Delgass, and V. Venkatasubramanian. High fidelity mathematical model building with experimental data: A Bayesian approach. *Computers and Chemical Engineering*, 32(4-5):971–989, 2008. doi:10.1016/j.compchemeng.2007.04.008.

[5] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, 1973.

[6] A. Braumann and M. Kraft. Incorporating experimental uncertainties into multivariate granulation modelling. *Chemical Engineering Science*, 65(3):1088–1100, 2010. doi:10.1016/j.ces.2009.09.063.

[7] A. Braumann, M. Kraft, and P. R. Mort. Parameter estimation in a multidimensional granulation model. *Powder Technology*, 197(3):196–210, 2010. doi:10.1016/j.powtec.2009.09.014.

[8] A. Braumann, P. L. W. Man, and M. Kraft. Statistical approximation of the inverse problem in multivariate population balance modeling. *Industrial and Engineering Chemistry Research*, 49(1):428–438, 2010. doi:10.1021/ie901230u.

[9] A. Braumann, P. L. W. Man, and M. Kraft. The inverse problem in granulation modelling – two different statistical approaches. *AIChE Journal*, in press, 2011. doi:10.1002/aic.12526.

[10] M. Castagné, Y. Bentolila, F. Chaudoye, A. Hallé, F. Nicolas, and D. Sinoquet. Comparison of engine calibration methods based on design of experiments (DoE). *Oil & Gas Science and Technology*, 63(4):563–582, 2008. doi:10.2516/ogst:2008029.

[11] Y. Chu and J. Hahn. Integrating parameter selection with experimental design under uncertainty for nonlinear dynamic systems. *AIChE Journal*, 54(9):2310–2320, 2008. doi:10.1002/aic.11562.

[12] S. G. Davis, A. B. Mhadeshwar, D. G. Vlachos, and H. Wang. A new approach to response surface development for detailed gas-phase and surface reaction kinetic model optimization. *International Journal of Chemical Kinetics*, 36(2):94–106, 2004. doi:10.1002/kin.10177.

[13] N. R. Draper and W. G. Hunter. Design of experiments for parameter estimation in multiresponse situations. *Biometrika*, 53(3–4):525–533, 1966. doi:10.1093/biomet/53.3-4.525.

[14] J. E. Etheridge, S. Mosbach, M. Kraft, H. Wu, and N. Collings. Modelling soot formation in a DISI engine. *Proceedings of the Combustion Institute*, 33(2):3159–3167, 2011. doi:10.1016/j.proci.2010.07.039.

[15] J. E. Etheridge, S. Mosbach, M. Kraft, H. Wu, and N. Collings. Modelling cycle to cycle variations in an SI engine with detailed chemical kinetics. *Combustion and Flame*, 158(1):179–188, 2011. doi:10.1016/j.combustflame.2010.08.006.

[16] R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

[17] G. Franceschini and S. Macchietto. Model-based design of experiments for parameter precision: State of the art. *Chemical Engineering Science*, 63(19):4846–4872, 2008. doi:10.1016/j.ces.2007.11.034.

[18] M. Frenklach, H. Wang, and M. J. Rabinowitz. Optimization and analysis of large chemical kinetic mechanisms using the solution mapping method — combustion of methane. *Progress in Energy and Combustion Science*, 18:47–73, 1992. doi:10.1016/0360-1285(92)90032-V.

[19] W. K. Hastings. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57(1):97–109, 1970. doi:10.1093/biomet/57.1.97.

[20] C. R. Hicks and K. V. Turner. *Fundamental Concepts in the Design of Experiments*. Oxford University Press, Oxford, 1999.

[21] D. Higdon, M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004. doi:10.1137/S1064827503426693.

[22] D. Higdon, J. Gattiker, B. Williams, and M. Rightley. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103 (482):570–583, 2008. doi:10.1198/016214507000000888.

[23] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society B*, 63(3):425–464, 2001. Stable URL: http://www.jstor.org/stable/2680584.

[24] P.-S. Koutsourelakis. Accurate uncertainty quantification using inaccurate computational models. *SIAM Journal on Scientific Computing*, 31(5):3274–3300, 2009. doi:10.1137/080733565.

[25] M. Kraft and S. Mosbach. The future of computational modelling in reaction engineering. *Philosophical Transactions of the Royal Society A*, 368(1924):3633–3644, 2010. doi:10.1098/rsta.2010.0124.

[26] M. Kraft, P. Maigaard, F. Mauss, M. Christensen, and B. Johansson. Investigation of combustion emissions in an HCCI engine – measurements and a new computational model. *Proceedings of the Combustion Institute*, 28(1):1195–1201, 2002. doi:10.1016/S0082-0784(00)80330-6.

[27] P. L. W. Man, A. Braumann, and M. Kraft. Resolving conflicting parameter estimates in multivariate population balance models. *Chemical Engineering Science*, 65(13):4038–4045, 2010. doi:10.1016/j.ces.2010.03.042.

[28] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21(6):1087–1091, 1953. doi:10.1063/1.1699114.

[29] S. Mosbach, H. Su, M. Kraft, A. Bhave, F. Mauss, Z. Wang, and J.-X. Wang. Dual injection HCCI engine simulation using a stochastic reactor model. *International Journal of Engine Research*, 8(1):41–50, 2007. doi:10.1243/14680874JER01806.

[30] S. Mosbach, A. M. Aldawood, and M. Kraft. Real-time evaluation of a detailed chemistry HCCI engine model using a tabulation technique. *Combustion Science and Technology*, 180(7):1263–1277, 2008. doi:10.1080/00102200802049414.

[31] S. Mosbach, M. S. Celnik, A. Raj, M. Kraft, H. R. Zhang, S. Kubo, and K.-O. Kim. Towards a detailed soot model for internal combustion engines. *Combustion and Flame*, 156(6):1156–1165, 2009. doi:10.1016/j.combustflame.2009.01.003.

[32] R. H. Myers and D. C. Montgomery. *Response Surface Methodology : Process and Product Optimization Using Designed Experiments*. John Wiley & Sons, New York, 2nd edition, 2002.

[33] H. Rabitz and Ö. F. Alış. General foundations of high-dimensional model representations. *Journal of Mathematical Chemistry*, 25(2-3):197–233, 1999. doi:10.1023/A:1019188517934.

[34] K. Röpke, editor. *Design of Experiments (DoE) in Engine Development, V: Innovative Development Methods for Vehicle Engines*. Expert Verlag, Renningen, 2011.

[35] T. Russi, A. Packard, R. Feeley, and M. Frenklach. Sensitivity analysis of uncertainty in model prediction. *Journal of Physical Chemistry A*, 112(12):2579–2588, 2008. doi:10.1021/jp076861c.

[36] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989. Stable URL: http://www.jstor.org/stable/2245858.

[37] M. Sander, R. I. A. Patterson, A. Braumann, A. Raj, and M. Kraft. Developing the PAH-PP soot particle model using process informatics and uncertainty propagation. *Proceedings of the Combustion Institute*, 33(1):675–683, 2011. doi:10.1016/j.proci.2010.06.156.

[38] P. Seiler, M. Frenklach, A. Packard, and R. Feeley. Numerical approaches for collaborative data processing. *Optimization and Engineering*, 7(4):459–478, 2006. doi:10.1007/s11081-006-0350-4.

[39] S. Shan and G. G. Wang. Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. *Structural and Multidisciplinary Optimization*, 41(2):219–241, 2009. doi:10.1007/s00158-009-0420-2.

[40] D. A. Sheen, X. You, H. Wang, and T. Løvås. Spectral uncertainty quantification, propagation and optimization of a detailed kinetic model for ethylene combustion. *Proceedings of the Combustion Institute*, 32(1):535–542, 2009. doi:10.1016/j.proci.2008.05.042.

[41] A. J. Smallbone, A. R. Coble, A. Bhave, S. Mosbach, M. Kraft, N. M. Morgan, and G. Kalghatgi. Simulating PM emissions and combustion stability in gasoline/diesel fuelled engines. SAE Paper No. 2011-01-1184, 2011.

[42] H. Su, A. Vikhansky, S. Mosbach, M. Kraft, A. Bhave, K.-O. Kim, T. Kobayashi, and F. Mauss. A computational study of an HCCI engine with direct injection during gas exchange. *Combustion and Flame*, 147(1-2):118–132, 2006. doi:10.1016/j.combustflame.2006.07.005.

[43] J. Villemonteix, E. Vazquez, and E. Walter. Bayesian optimization for parameter identification on a small simulation budget. 15th IFAC Symposium on System Identification, SYSID 2009, Saint-Malo, France, 2009. Available at: http://hal-supelec.archives-ouvertes.fr/hal-00368152/en.

[44] R. D. Wilkinson. Bayesian calibration of expensive multivariate computer experiments. In L. Biegler et al., editors, *Large-Scale Inverse Problems and Quantification of Uncertainty*, chapter 10. John Wiley & Sons, New York, 2010.

[45] T. Ziehn, K. J. Hughes, J. F. Griffiths, R. Porter, and A. S. Tomlin. A global sensitivity study of cyclohexane oxidation under low temperature fuel-rich conditions using HDMR methods. *Combustion Theory and Modelling*, 13(4):589–605, 2009. doi:10.1080/13647830902878398.