Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

Bayesian parameter estimation for a jet-milling model using Metropolis-Hastings and Wang-Landau sampling

Catharine A. Kastner, Andreas Braumann, Peter L. W. Man,

Sebastian Mosbach, George P. E. Brownbridge, Jethro Akroyd,

Markus Kraft [1], Chrismono Himawan [2]

released: 8 September 2012

[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

[2] Particle Generation Control
& Engineering UK
Product Development
GlaxoSmithKline RD
Office: 5S090, Park Road
Ware Herts SG12 0DP
United Kingdom

Preprint No. 109

Keywords: MCMC, Metropolis-Hastings, Wang-Landau, parameter estimation, population balance equation

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:	mk306@cam.ac.uk
World Wide Web:	`http://como.cheng.cam.ac.uk/`

Abstract

Bayesian parameter estimates for a computationally expensive multi-response jet-milling model are computed using the Metropolis-Hastings and Wang-Landau Markov Chain Monte Carlo sampling algorithms. The model is accompanied by data obtained from 74 experiments at different process settings which is used to estimate the model parameters. The experimentally measured quantities are the 10th, 50th and 90th quantiles of the resulting particle size distributions. Parameter estimation is performed on a population balance jet-milling model comprised of three subprocesses: jet expansion, milling and classification. The model contains eight parameters requiring estimation and can compute the same quantities that are determined in the experiments. As the model is computationally expensive to solve, the sampling algorithms are applied to a surrogate model to establish algorithm specific parameters and to obtain model parameter estimates. The resulting parameter estimates are given with a discussion of their reliability and the observed behaviour of the two sampling algorithms. Comparison of the autocorrelation function between samples generated by the two algorithms shows that the Wang-Landau algorithm exhibits more rapid decay. Trace plots of the parameter samples from the two algorithms appear to be analogous and encourage the supposition that the Markov Chains have converged to the distribution of interest. One- and two-dimensional density plots indicate a unimodal distribution for all parameters, which suggests that the obtained estimates are unique. The two-dimensional density plots also suggest correlation between at least two of the model parameters. The realised distribution generated by both algorithms produced consistent results and demonstrated similar behaviour. For the application considered in this work, the Wang-Landau algorithm is found to exhibit superior performance with respect to the correlation and equivalent performance in all other respects.

Contents

1 Introduction

The use of models to emulate physical systems is widespread in both research and industry. The popularity of modelling stems from practical concerns, primarily the expense and difficulty involved in performing experiments [25]. A well-developed model can be employed instead of performing costly and time-consuming experimental work; however the integrity of the model is a fundamental concern. Any model sufficiently sophisticated to emulate complicated physical systems will contain parameters requiring estimation. The method utilised to perform such estimation is a non-trivial decision, as poor parameter estimates will undermine the usefulness of the model. Issues that must be considered are both the accuracy and the computational expense of obtaining the parameter estimates.

Although many methods have been proposed for parameter estimation, the current state of the art is often held to be the use of Bayesian theory in conjunction with Markov Chain Monte Carlo (MCMC). Bayesian theory is used to define uncertainty of model parameters as a distribution based on prior information and existing data. The basic tenet of Bayesian methods is that information about one aspect of a system can be used to make statements regarding other parts. Generally, there is uncertainty inherent in the experimental responses. This makes using Bayesian theory particularly appropriate as we can exploit properties of the experimental uncertainty in the derivation for the distribution of the model parameters. Establishing such a distribution is non-trivial and there is a multitude of suggested methods; a summary of which would exceed the scope of this work and will not be attempted.

Once this *posterior* distribution is constructed, a realisation of the distribution is needed to obtain the parameter estimates, for which we turn to sampling algorithms. The prominent concern in implementing any sampling algorithm is that the method move through the support of the target distribution rapidly, or shows ‘good mixing’ [19], as poor mixing will require more computations. A comprehensive list of sampling algorithms and theory is beyond the scope of this paper; however it would be remiss to fail to mention some of the more important developments.

The two fundamental and most widely used samplers are the Metropolis-Hastings algorithm [23], which is implemented in this paper, and the Gibbs sampler [12] which is a special case of single-component Metropolis-Hastings [21]. When using either of these algorithms, correlation between the parameters can cause slow mixing which may be improved by using reparameterisation techniques; a review of which is included in [19]. Another strategy to improve mixing is by modifying the distribution being sampled from by using importance sampling [17], simulated tempering [18] or simulated annealing [13]. Another approach is to implement self-correcting features to tune the algorithm as it iterates. An algorithm developed specifically to address directional adaption is presented in [20] with a general review of such methods presented in [3, 32]. The Wang-Landau algorithm [37], which is compared to Metropolis-Hastings in this paper, combines adaptive methodology with alternate distribution sampling, simulated tempering in this specific case. The methods available to quantify the quality and performance of the algorithms are extensive. In this paper we will make use of very basic illustrative plots, however a review of more sophisticated methods can be found in [14].

While there is extensive literature regarding the theoretical aspects of sampling techniques, as well as innumerable studies of various aspects of the implementation, the tendency is to focus on features of the implementation, such as the required sample size. In this paper the focus instead is directed towards a straight-forward practical application of and comparison between the performance of the algorithms. The overall methodology used in this paper has been previously applied for single response data in a particle granulation context in [8–11, 26, 36] and for a multireponse internal combustion engine model in [29].

The **purpose of this paper** is to demonstrate model parameter estimation for a population balance jet-milling model via the application of two MCMC algorithms, Metropolis-Hastings and Wang-Landau. These algorithms are applied to a multi-response computationally expensive model where quadratic response surfaces are used as a surrogate model. The entire parameter estimation process is shown, along with demonstrations of some basic tools to ascertain the validity of the results and to compare the performance of the two MCMC algorithms. The emphasis throughout is on general techniques which may be applied elsewhere.

The structure of this paper is as follows: Section 2 has a description of the experimental system. Section 3 describes the jet-milling model. Section 4 contains an overview of the Bayesian theory and the MCMC sampling algorithms used along with details of our specific application and the construction of the surrogate model. In section 5 we describe specific settings for the implementation used. In section 6 we show the results of the implementation and compare the results from the two algorithms. In section 7 we draw conclusions and discuss recommendations for future work.

2 Experimental system

Jet milling is used as a size reduction unit operation when the desired particle size is less than $10\,\mu$m [22]. This process produces a very finely ground product with a narrow size distribution without risk of contamination. The feed particles and a gas are injected into a mixing chamber where the inflow of the gas promotes breakage by causing particle to particle and particle to wall collisions. As it is not possible to classify particles during the milling process, an air classification system is typically added externally that sifts the end product into coarse and fine material. The coarse particles are recycled back into the system and the fine ones proceed to the next unit operation. A basic schematic illustrating a simple jet milling system, as portrayed in [22], can be seen in **Figure 1**.

The experimental results used in the forthcoming parameter estimation were acquired by GlaxoSmithKline (GSK) using 8-inch micronisers. A collection of 74 batches were performed using the process conditions described in **Table 1**. The end product was analysed using the Malvern wet method [1, 2]. The measurements made were x_{10}, x_{50}, and x_{90}, which are the 10th, 50th and 90th percentiles of the resulting particle size distribution (PSD).

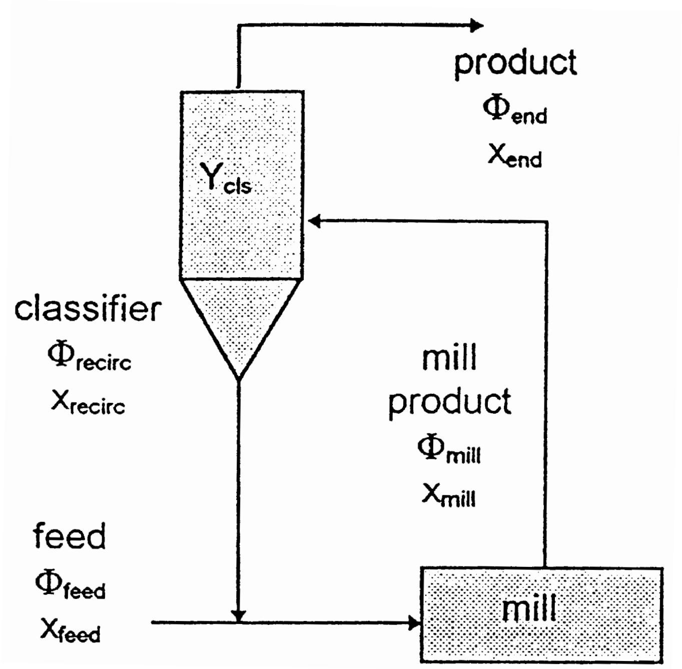


Figure 1: *General schematic of a jet milling system (taken from [22]).*

Table 1: *Process conditions used in experiments.*

Process conditions	Value	Units
Constants		
Feed gas temperature	293.15	K
Grinding gas temperature	293.15	K
Process conditions varied		
Feed particle size distribution (PSD) characteristics:		
x_{10}	100 – 300	μm
x_{50}	250 – 500	μm
x_{90}	550 – 800	μm
Grinding pressure	2 – 8.5	barg
Feed pressure	5 – 9.5	barg
Feed mass flow rate	2 – 7.5	kg/h

3 Model description

The model used in this work is a population balance model of a jet milling process, developed by GSK, where the entire process is divided into three subprocesses:

1. Jet expansion,
2. Milling (breakage),
3. Classification.

The overall process is modelled with a population balance model for an ideally mixed vessel with inflow and classified outflow that has eight parameters requiring estimation. The model inputs are the grinding pressure, the feed pressure, the feed mass flow rate and a Rosin Rammler characterisation of the initial PSD. The model outputs the resulting PSD as a characterisation of x_{10}, x_{50} and x_{90} as well as the specific energy input and the fraction of solid material in the total flow. The collection of parameters employed in the model, their sources and physical definitions are shown in **Table 2**.

The population balance equation is:

$$\frac{\mathrm{d}w}{\mathrm{d}t} = -(\overline{I} - \overline{B})\,\overline{S}\,\overline{w} + \dot{m}_{\text{feed}}\,\overline{X}_{w,\text{feed}} - \overline{P}_{\text{disc}}\,\overline{w}\ , \tag{1}$$

where w represents a discretised characterisation of the particle distribution.

The jet expansion subprocess, using the construction in [31], gives the mass flow of the gas as:

$$\dot{m}_{\text{gas}} = p_{\text{grind}} A_0 \sqrt{\frac{\varkappa\, M_{\text{w}}}{R\, T_{\text{gas}}} \left(\frac{2}{k+1}\right)^{\frac{k+1}{k-1}}}\ , \tag{2}$$

from which we can calculate the kinetic energy, as given in [27], as:

$$E_{\text{k}} = \frac{1}{2}\dot{m}_{\text{gas}}\, v_{\text{gas}}^{2} \tag{3}$$

and the specific energy, also from [27], as:

$$E_{\text{sp}} = \frac{E_{\text{k}}}{\dot{m}_{\text{feed}}}\ , \tag{4}$$

and the fraction of solid material in the total flow is computed as:

$$X_{\text{solid}} = \frac{\dot{m}_{\text{feed}}}{\dot{m}_{\text{feed}}\, \dot{m}_{\text{gas}}}\ . \tag{5}$$

Table 2: *Variables used in the model.*

Parameter	Description	Source of value
Population balance equation		
$\overline{I}$	Identity matrix	
$\overline{B}$	Breakage matrix	Model parameter
$\overline{S}$	Breakage frequency vector	Model parameter
$\dot{m}_{\text{feed}}$	Feed mass flow rate	Model parameter
$\overline{X}_{w,\text{feed}}$	Feed mass distribution	Process condition
$\overline{P}_{\text{disc}}$	Probability discharge	Model parameter
Jet expansion		
A_0	Nozzle area	Process condition
E_{k}	Kinetic energy	Model response
E_{sp}	Specific energy	Model response
M_{w}	Molecular mass	Process condition
p_{grind}	Grinding pressure	Process condition
R	Universal gas constant	Process condition
T_{gas}	Grinding gas temperature	Process condition
v_{gas}	Gas velocity	Process condition
X_{solid}	Solid mass fraction	Model response
$\varkappa$	Heat capacity ratio	Process condition
Breakage		
k_{r}	Feed PSD parameter	Process condition
m	Feed PSD parameter	Process condition
$\dot{m}_{\text{gas}}$	Gas mass flow rate	Model parameter
τ	Empirical parameter	*Parameter estimation*
β	Empirical parameter	*Parameter estimation*
λ	Empirical parameter	*Parameter estimation*
Air classification		
k_{disc}	Discharge consant	*Parameter estimation*
M_{holdup}	Mass of solid in milling chamber	Model parameter
x_{50d}	Cut size	*Parameter estimation*
k_{x50d}	Holdup constant (on cut size)	*Parameter estimation*
$k_{\sigma\text{50d}}$	Holdup constant (on spread)	*Parameter estimation*
σ_{50d}	Spread	*Parameter estimation*

The breakage subprocess is described by the breakage matrix $\overline{B}$, where the ith and jth element of $\overline{B}$ is:

$$B_{ij} = \frac{S_{i-1} - S_i}{S_j}, \tag{6}$$

which uses the elements of the breakage frequency vector $\overline{S}$ [6]. The ith element of $\overline{S}$ is defined as:

$$S_i = (m_{\text{gas}}\tau)^{\beta} \left(\frac{x_i}{x_{max}} \right)^{\lambda}, \tag{7}$$

where x_i is the size class under consideration and x_{max} is the uppermost size of the discretisation of the particle distribution [22].

For the air classification submodel, we use the construction in [22] as:

$$P_{\text{disc},i} = k_{\text{disc}}\, \dot{m}_{\text{gas}} \left[1 - 0.5 \left(1 + \text{erf} \left(\frac{\ln x_i - \ln x_{\text{d}}}{\sqrt{2}\, \ln\, \sigma_{\text{d}}} \right) \right) \right], \tag{8}$$

where

$$x_{\text{d}} = x_{\text{50d}}[k_{\text{x50d}}(1 + M^2_{\text{holdup}})] \tag{9}$$

and

$$\sigma_{\text{d}} = \sigma_{\text{50d}}[k_{\sigma\text{50d}}(1 + M^2_{\text{holdup}})]\,. \tag{10}$$

This model has been implemented in Matlab and a single evaluation takes approximately 5 seconds CPU time. Since use of sampling algorithms typically requires making a large number of model evaluations, using the model directly is prohibitively slow. A commonly used method in this situation is to use a surrogate model. In addition, past experience with this model has provided recommended parameter range constraints to define the model parameter space which are set out in **Table 3**.

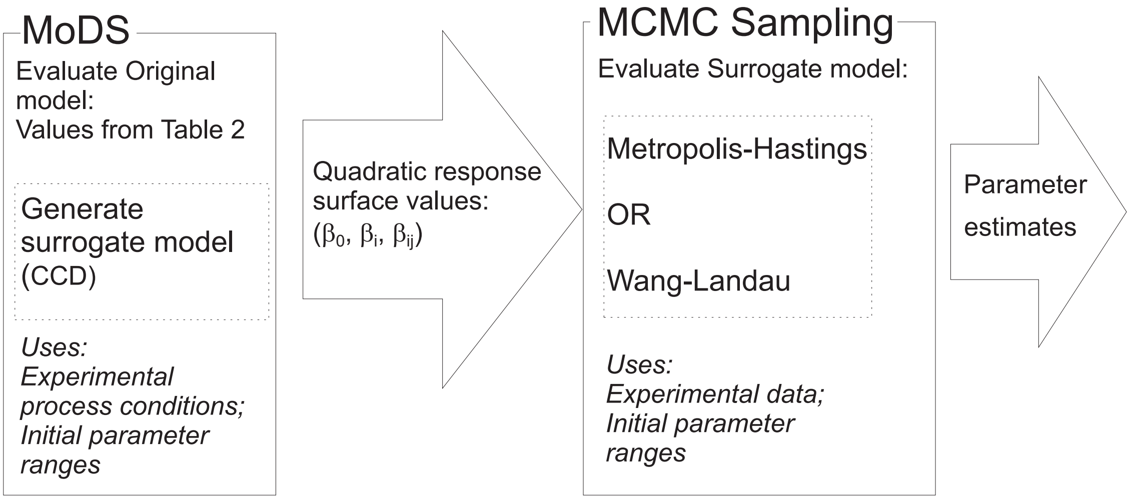
4 Parameter estimation methodology

In this section we briefly discuss the construction of the surrogate model and review the Bayesian theory employed along with the implementation of the MCMC algorithms. Both the Metropolis-Hastings and the Wang-Landau algorithms are stated in a generalised form.

This parameter estimation process begins with creating the surrogate model. In this implementation we make use of an in-house developed software package, *MoDS*, to generate quadratic response surfaces using evaluations of the original model. The surrogate model

Table 3: *Model parameters to be estimated with physical limits and limits chosen for construction of response surfaces.*

			Physical limits		Chosen limits		
Parameter	Model value	Unit	low	up	low	up	Scaling
θ_1	τ	$10^{-3}\,\mathrm{kg}^{-1}$	0	∞	1	2	linear
θ_2	λ	–	0	∞	1	3	linear
θ_3	β	–	0	∞	1	3	linear
θ_4	$x_{50\mathrm{d}}$	μm	0	∞	5	20	linear
θ_5	$\sigma_{50\mathrm{d}}$	–	1	∞	1	3	linear
θ_6	$k_{x_{50\mathrm{d}}}$	$100\,\mathrm{m}^6/\mathrm{kg}^2$	0	∞	0.5	3	linear
θ_7	$k_{\sigma_{50\mathrm{d}}}$	$100\,\mathrm{m}^6/\mathrm{kg}^2$	0	∞	0.01	0.1	linear
θ_8	k_{disc}	kg^{-1}	0	∞	0.5	5.5	linear



MoDS

Evaluate Original model: Values from Table 2

Generate surrogate model (CCD)

Uses: Experimental process conditions; Initial parameter ranges

Quadratic response surface values: (β₀, β_i, β_{ij})

MCMC Sampling

Evaluate Surrogate model:

Metropolis-Hastings

OR

Wang-Landau

Uses: Experimental data; Initial parameter ranges

Parameter estimates

Figure 2: *Overview of parameter estimation process used in this work.*

is then used in lieu of the original model with the MCMC algorithms. Both parts of this process employ initial estimates of likely ranges of the parameters which are based on physical limits or past experience which appear in **Table 3**.

The entire process with component parts is summarised in **Figure 2**.

4.1 Definitions

We begin by introducing some terminology and notation. The model simulates an experimental process on which experiments have been performed. Any experiment is characterised by a number of settings prescribed by the experimenter, *i.e.* a vector of

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

with M components. The resulting measurements can be expressed as

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

with L components.

The corresponding model predictions produce a vector of

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

which also has L components.

The model responses depend on the process conditions and on a vector of

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

with P components. 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 as a sequence of process conditions we denote this by superscript indices in parentheses: The n^{th} experiment, performed at the process conditions $\xi^{(n)} = \big(\xi_1^{(n)}, \dots, \xi_M^{(n)}\big)^\top$, yields responses $\eta^{\text{exp},(n)} = \big(\eta_1^{\text{exp},(n)}, \dots, \eta_L^{\text{exp},(n)}\big)^\top$, and the n^{th} model evaluation, also performed at the process conditions $\xi^{(n)}$, yields responses $\eta^{(n)} := \eta(\xi^{(n)}, \theta) = \big(\eta_1^{(n)}, \dots, \eta_L^{(n)}\big)^\top$.

4.2 Response surfaces

The methods by which the parameter estimates can be generated are determined in part by the operational behaviour of the model. In this case, direct evaluation of the model is prohibitively slow which suggests the use of a surrogate model; in our case we will make use of quadratic response surfaces.

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:

$$\eta_l^{(n)}(\theta) = \beta_{l,0}^{(n)} + \sum_{i=1}^{P} \beta_{l,i}^{(n)}\, \theta_i + \sum_{i=1}^{P} \sum_{j \ge i}^{P} \beta_{l,ij}^{(n)}\, \theta_i\, \theta_j, \tag{11}$$

where $\beta_{l,0}^{(n)}$, $\beta_{l,i}^{(n)}$, and $\beta_{l,ij}^{(n)}$ are the coefficients of the constant, linear, and quadratic terms, respectively [11, 16, 29].

We determine the coefficients of our surrogate models (11) by performing least-squares fitting 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 [30]. Construction of response surfaces has previously been used to good effect under similar circumstances for a wide variety of applications [11, 15, 16, 24, 28, 29, 34].

4.3 Bayesian preliminaries

We estimate the model parameters with a Bayesian methodology based on the construction given in [5] for multi-response experimental data. The construction used is derived in detail in [29].

A Bayesian approach is based on the idea that uncertainty in the model parameters θ can be represented by a probability density $p(\theta)$, called a *prior* distribution. When provided with new information, such as experimental data, which have a corresponding probability density $p(\eta^{\text{exp}}|\theta)$, the model parameters can be updated, resulting in a *posterior* distribution $p(\theta|\eta^{\text{exp}})$ of the unknown parameters. This is the concept that is expressed by Bayes' Theorem:

$$p(\theta|\eta^{\text{exp}}) \propto p(\eta^{\text{exp}}|\theta)p(\theta). \tag{12}$$

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.

4.3.1 Likelihood

The likelihood embodies the distribution of the experimental data, where we will make the common assumption that the experimental response is equal to the model response plus a Gaussian error:

$$\eta^{\text{exp},(n)} = \eta\big(\xi^{(n)}, \theta\big) + \varepsilon^{(n)} \qquad \text{with} \qquad \varepsilon^{(n)} \sim \mathcal{N}_L(0, \Sigma), \tag{13}$$

where $\varepsilon^{(n)}$ is the L-dimensional vector of the measurement errors which are normally distributed with zero mean and covariance matrix Σ. Further, Σ is assumed to be independent of the process condition ξ, but the L components of the error vector for any one experiment may be correlated. Then the likelihood can be shown to be: [5, 29]

$$p\big(\eta^{\text{exp},(1)}, \ldots, \eta^{\text{exp},(N)}\big|\theta, \Sigma\big) = (2\pi)^{-NL/2}(\det \Sigma)^{-N/2} \exp\big\{-\tfrac{1}{2}\text{tr}\big[\Sigma^{-1}S(\theta)\big]\big\}, \tag{14}$$

where:

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

4.3.2 Prior distributions

The choice of prior distributions is a non-trivial, much-debated subject in the literature. From (14) we have two objects, θ and Σ, for which we need to state prior beliefs. We begin by assuming their independence:

$$p(\theta, \Sigma) = p(\theta)\, p(\Sigma). \tag{15}$$

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

$$p(\theta) = \frac{1}{|\mathcal{C}|} \mathbb{1}_{\{\theta \in \mathcal{C}\}}, \tag{16}$$

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

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

$$p(\Sigma) \propto (\det \Sigma)^{-\alpha-(L+1)/2} \exp\left[-\tfrac{1}{2}\mathrm{tr}\big(\Sigma^{-1}\Psi\big)\right], \tag{17}$$

where $\alpha > 0$ and $\Psi \in \mathbb{R}^{L\times L}$ are positive definite arbitrary parameters.

4.3.3 Posterior distributions

The application of Bayes' Theorem gives the posterior densities (up to constant factors) as:

$$\begin{aligned} &p\big(\theta, \Sigma \big| \eta^{\exp,(1)}, \ldots, \eta^{\exp,(N)}\big) \\ &\qquad \propto (\det \Sigma)^{-\alpha-(N+L+1)/2} \exp\left\{-\tfrac{1}{2}\mathrm{tr}\big[\Sigma^{-1}\big(S(\theta) + \Psi\big)\big]\right\} \cdot \mathbb{1}_{\{\theta\in\mathcal{C}\}}. \end{aligned} \tag{18}$$

As we are interested in the marginal posterior density for θ, we integrate over all positive definite matrices Σ which gives:

$$\begin{aligned} p\big(\theta \big| \eta^{\exp,(1)}, \ldots, \eta^{\exp,(N)}\big) &= \int_{\Sigma \text{ pos. def.}} p\big(\theta, \Sigma \big| \eta^{\exp,(1)}, \ldots, \eta^{\exp,(N)}\big) \mathrm{d}\Sigma \\ &\propto \big[\det\big(S(\theta) + \Psi\big)\big]^{-\alpha-N/2} \cdot \mathbb{1}_{\{\theta\in\mathcal{C}\}}. \end{aligned} \tag{19}$$

Using these constructs we have expressions for the posteriors (18) and (19) only up to constant positive factors. While these normalisation factors can in principle be found, it is not necessary when an appropriate sampling method is used.

Further we define a constant ε and assign values to α and Ψ in (19), such that:

$$\alpha = \varepsilon, \tag{20}$$

$$\Psi = \begin{pmatrix} 2\,\varepsilon & \ldots & 0 \\ 0 & \ddots & 0 \\ 0 & \ldots & 2\,\varepsilon \end{pmatrix}. \tag{21}$$

4.4 Sampling algorithms

4.4.1 Metropolis-Hastings

The Metropolis-Hastings algorithm creates a continuous space discrete time Markov Chain with a stationary distribution identical to the distribution of interest, called $\pi(x)$. In simple terms, this means sequentially choosing values of x, or states, to move into such that a state x is visited with frequency density $\pi(x)$. A generalised statement of the algorithm is given in **Algorithm 1**. Notice that (23) has the factor $\pi(x')/\pi(x)$ implying that we need only the density $\pi(x)$ up to a constant normalisation factor. Hence, if we wish to sample from the posteriors given in (18) or (19), we need only substitute $\pi(x)$ with those expressions.

To apply Algorithm 1 to our application, we set the states x to be θ and the density of interest to be the posterior density $p(\theta \,|\, \eta^{\text{exp}})$. We also use the proposal density $q(x \to x')$ (as explained in Step 1 of Algorithm 1) to be $q(x \to x') = q(\theta \to \theta') = \delta \frac{1}{|\mathcal{X}|} \mathbb{1}_{\{\theta' \in \mathcal{X}\}} + (1 - \delta)\, \frac{1}{|\mathcal{X}(\theta)|} \mathbb{1}_{\{\theta' \in \mathcal{X}(\theta)\}}$, where $\mathcal{X}(\theta)$ is the intersection of the hypercube $\mathcal{X}$ and the hypercube with edgelength Δ centred at θ. The idea behind this choice is to jump completely uniformly in $\mathcal{X}$ with probability δ to ensure coverage of the entire space but, to reduce rejections, we jump with probability $(1 - \delta)$ to a subspace of the hypercube centered on the current position in $\mathcal{X}$.

An important requirement for implementation is that the created Markov Chain must fulfil the condition of ergodicity and the condition of balance, which is usually extended to a condition of detailed balance. The condition of ergodicity requires that from any state x one can move, with some number of intermediate steps, to any other state x' and if the chain runs long enough it will return to x at some future point [21]. The condition of detailed balance requires that the transition probability $q(x \to x')$ between each pair of states x and x' is such that:

$$\pi(x)q(x \to x') = \pi(x')q(x' \to x). \tag{22}$$

This means that the probability of moving from $q(x \to x')$ is the same as moving from $q(x' \to x)$. It can be shown that the Metropolis-Hastings algorithm has this property, given weak conditions on q and π [35]. One says that the chain has *converged* when these conditions have been met. In practice, it is not possible to establish with full certainty that a chain has converged. However, there are diagnostic methods, which will be employed in section 6 that may indicate non-convergence.

4.4.2 Wang-Landau

The Wang-Landau algorithm is an extension of the Metropolis-Hastings algorithm which attempts to improve mixing by adaptively sampling from alternative distributions. A full description of a generalised form of the Wang-Landau algorithm can be found in [4]. The basic idea is that we begin with a state space $\mathcal{X}$ and a probability measure π. One can then create a partition such that $\mathcal{X} = \cup \mathcal{X}_i$ where $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ and π is reweighted in each $\mathcal{X}_i$. The primary difficulty in this method is in calculating the weights so that sampling is appropriately distributed across the partitions. The Wang-Landau algorithm addresses this issue by simultaneously computing the weights and sampling from the new distribution. A generalised form of the Wang-Landau algorithm is set out in **Algorithm 2**.

This method naturally lends itself to simulated tempering where distributions are generated that are close to π but easier to sample from [18]. The implementation in this paper is such that one selects a so-called 'temperature ladder' consisting of T temperature

Algorithm 1: Metropolis-Hastings algorithm

```
Choose a **proposal density** $q(x \to x')$ which is a probability density function for
choosing a new state $x'$ given that we are currently positioned in the state $x$.

Set $i = 0$. Start with any initial state $x^{(0)}$ for $x$.

**while** $i < N$ **do**
    **Propose** a new state $x'$ sampled from (any) *proposal density* $q(x \to x^{(i)})$.
    Compute the quantity
        $$\alpha_{\text{accept}} := \frac{\pi(x')\, q(x' \to x^{(i)})}{\pi(x^{(i)})\, q(x^{(i)} \to x')}. \qquad (23)$$
    Perform a **rejection step**, i. e. with probability $r := \min\{1, \alpha_{\text{accept}}\}$, *accept* the
    proposed state $x'$, i. e., set $x^{(i+1)} = x'$, otherwise, set $x^{(i+1)} = x^{(i)}$.
    Set $i \leftarrow i + 1$.

STOP.
```

classes. For each step of the ladder, the posterior density function is taken to a power $1/T$, which has the effect of smoothing the distribution. Prior to the execution of each Metropolis-Hastings step, a temperature class is selected and the acceptance/rejection is decided based on the smoothed distribution. The benefits of using this method is improved mixing of the sampling due to the curve smoothing effect as one ascends the ladder, however its usage incurs additional complications. A method is required which will select a temperature class for each iteration that will ensure good mixing between the temperature classes. This is accomplished by checking after each iteration if an approximately 'equal' distribution has been obtained and adjusting the weights attached to each temperature class before selecting the temperature for the next iteration. Additionally, only the samples generated in the first temperature class are pertinent to the distribution of interest. This being the case, we discard all samples generated in the other temperature classes which can amount to a substantial portion of the samples generated.

5 Details of implementation

To simplify calculations throughout, the model parameters are coded using a linear transform to restrict the range of each parameter θ_i by transforming it to a corresponding θ_i' such that $-1 < \theta_i' < 1$. The linear transformation used for each parameter with given bounds is:

$$b_{i,1} = \frac{2.0}{\theta_{i,\text{upperbound}} - \theta_{i,\text{lowerbound}}}, \tag{26}$$

$$b_{i,0} = 1.0 - b_{i,1} \times \theta_{i,\text{upperbound}}, \tag{27}$$

$$\theta_i' = \frac{\theta_i - b_{i,0}}{b_{i,1}}. \tag{28}$$

The values used with this model for $\theta_{i,\text{lowerbound}}$ and $\theta_{i,\text{upperbound}}$ are listed as the chosen limits in **Table 3**.

As this is a linear mapping, the properties of vector addition and scalar multiplication are preserved. Samples are generated by the sampling algorithm while restricted to this interval and then decoded when the results are presented.

5.1 Surrogate model

The construction of the second order response surfaces is based on a Central Composite Design (CCD). In addition to the cornerpoints of a full factorial design, the centre point and two points in each direction are evaluated. With the assumption that the full factorial design is based on [-1, 1], the CCD becomes rotatable if the axis points are a certain distance away from the centre point. This distance parameter D needs to be equated by the following equation if rotatability of the design is required,

$$D = 2^{k/4}, \tag{29}$$

Algorithm 2: Wang-Landau algorithm

```
Set $i = 0$. Set constants $a_0 = 0$, $\kappa = 0$ and select $c \in (0, 1)$.
Set $T$ = number of partitions/temperature classes.
Choose a **proposal density** $q_\rho((x, T^{(i)}) \to (x', T'))$ which is a probability density function for choosing a new state $(x', T')$ given that we are currently positioned in the state $(x, T^{(i)})$ and $T'$ is a function of $\rho_i(.)$.
Define function $f_{a,b}(j) = \frac{1}{b-a} \sum_{k=a+1}^{b} \mathbb{1}_{\{T^{(k)}=j\}}$ when $a \leq b$ and 0 otherwise.
Define $\{\gamma_n\}$, as some positive decreasing sequence.
Define $T$-dimensional vector $\phi_0$ where $\phi_0(j) \in \mathbb{R}$ and $\phi_0(j) > 0$ for $j = 1, \ldots, T$.
Define function $\rho_i(n)$ such that:
    $$\rho_i(n) = \frac{\phi_i(n)}{\sum_{j=1}^{T} \phi_i(j)}. \quad (24)$$
Start with any initial state $(x^{(0)}, T^{(0)})$.
while $i < N$ do
    **Propose** a new state $(x', T')$ sampled from *proposal density* $q_\rho((x^{(i)}, T^{(i)}) \to (x', T'))$.
    Compute the quantity
        $$\alpha_{\text{accept}} := \frac{\pi(x', T')}{\pi(x^{(i)}, T^{(i)})} \times \left( \frac{q_\rho((x', T') \to (x^{(i)}, T^{(i)}))}{q_\rho((x^{(i)}, T^{(i)}) \to (x', T'))} \right). \quad (25)$$
    Perform a **rejection step**, i. e. with probability $r := \min\{1, \alpha_{\text{accept}}\}$, *accept* the proposed state $(x', T')$, i. e., set $(x^{(i+1)}, T^{(i+1)}) = (x', T')$, otherwise, set $(x^{(i+1)}, T^{(i+1)}) = (x^{(i)}, T^{(i)})$.
    For $j = 1, \ldots, T$; Set $\phi_{i+1} = \phi_i(j) \left(1 + \gamma_{a_i} \mathbb{1}_{\{T^{(i+1)}=j\}}\right)$.
    Calculate $\rho_{i+1}(j)$.
    if $Max_{1 \leq j \leq T} |f_{\kappa,(i+1)}(j) - \frac{1}{T}| \leq \frac{c}{T}$
    then
        Set $\kappa = i + 1$, and $a_{i+1} = a_i + 1$.
    else
        $a_{i+1} = a_i$.
    Set $i \leftarrow i + 1$.
Discard samples from all temperature classes, except $T^{(i)} = 1 \; \forall \; i$.
STOP.
```

where k is the number of dimensions of the design; in this application k is set equal to the number of model parameters being estimated. However, it is possible for the values of the uncoded parameters outside [-1, 1] to take on physically impossible values, e. g., become negative. Hence, the rotatability of the CCD is ignored for the forthcoming parameter estimation and points selected such that

$$D = 0.5\,. \tag{30}$$

5.2 MCMC settings

The number of samples generated and the number samples at the beginning of the chain that are discarded to eliminate the influence of the starting position, called 'burnin', are typically of interest in MCMC implementations. We have chosen to forego investigation of these elements and instead use arbitrarily large values for both factors and then validate after the fact that these sizes are sufficient. For all of the sample runs, 6 million samples were generated for Metropolis-Hastings and 5 million were generated for Wang-Landau over all of the temperature classes. It is important to note that the practice of discarding the samples generated in any class except $T = 1$ when employing Wang-Landau implies that the number of samples we have for analysis is significantly less than the 5 million generated. The amount of burnin will be chosen after the remaining algorithm specific settings have been established. Additionally, preliminary testing has established a value of $\varepsilon = 0.001$ to be used in (21) as a value which works well in this system. The scripts for both algorithms have been implemented in R.

The criterion we will use to establish the remaining settings is by examining the acceptance rates. It is a known result that an optimal rate of convergence, under certain conditions, is at an acceptance rate of 0.234 [33]. We will use this as a guideline to determine appropriate algorithm specific settings.

5.2.1 Metropolis-Hastings acceptance rate analysis

The values for δ and Δ, which are the probability of a large jump and the edgelength, respectively, will be established by performance.

Samples were generated using all combinations of:

1. δ values: 0.3, 0.4, 0.5; and
2. Δ values: 0.004, 0.005, 0.006, 0.007.

The resulting acceptance rates can be seen in **Figure 3**. We can observe a downward shift in the acceptance rates when δ increases. In addition, we can clearly see that the acceptance rates decrease as the edgelength Δ increases. From this plot we select the combination of settings which are closest to our theoretical best acceptance rate for further scrutiny. The selected combinations of settings are when $\delta = 0.3$ and $\Delta = 0.005$. Further, with 6 million samples available for analysis, we will set the burnin at 1 million samples.

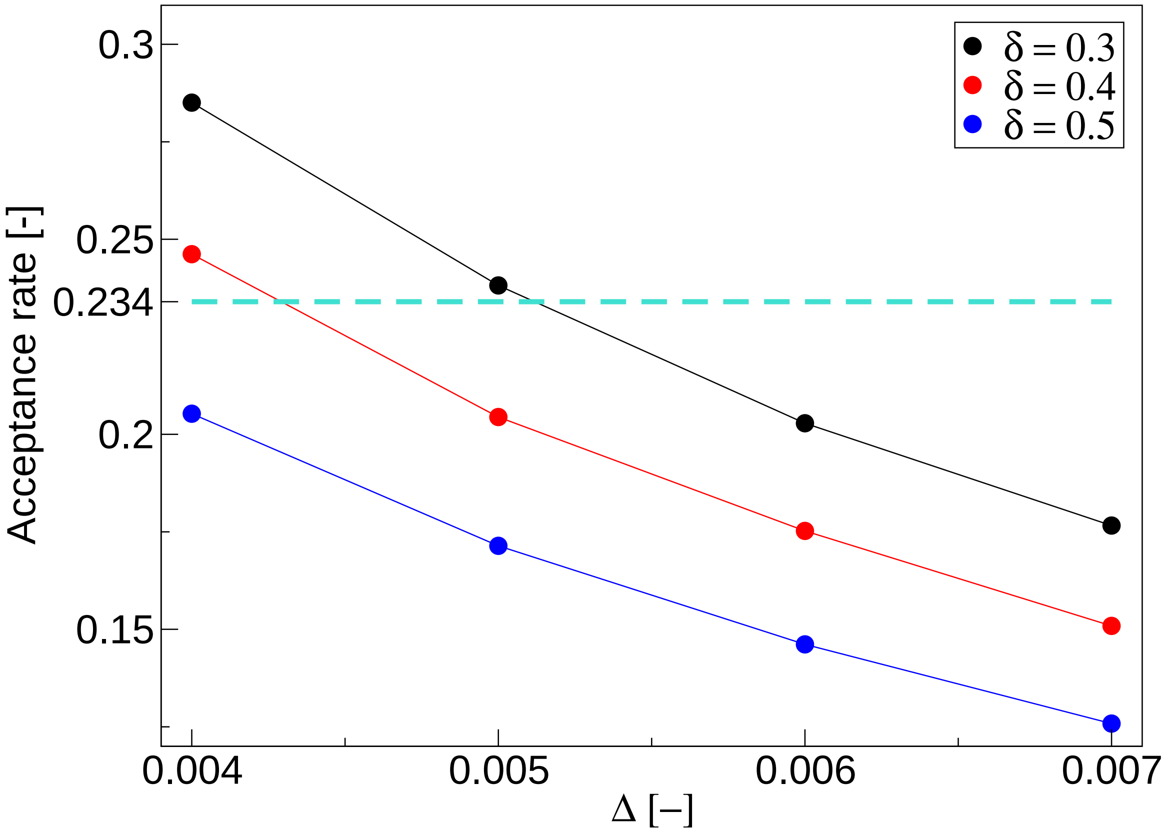


Figure 3: *Acceptance rates for Metropolis-Hastings;* $\varepsilon = 0.001$.

5.2.2 Wang-Landau acceptance rate analysis

The Wang-Landau algorithm also has settings and functions that must be established beforehand.

Our first concern is the spacing of the temperatures. In this case, we have defined a constant ratio between adjacent temperatures called I_{inv} to create geometric growth in the temperatures. In this specific case we have set $I_{\text{inv}} = 0.2$.

The quantity for δ retains its Metropolis-Hastings definition as the probability of a large jump however, with Wang-Landau, each temperature class $j = 1 \ldots T$ requires an individual edgelength Δ_j. We set these as:

$$\Delta_j = \sqrt{\frac{1}{I_{\text{inv}}^{j-1}}} \times \Delta_1, \tag{31}$$

where Δ_1 is the edgelength for the first temperature class, or the distribution of interest.

The mechanism for selecting the next temperature class using the weights ρ_i is the Gibbs sampler. For details about Gibbs sampling see [12].

We also need to determine initial values to update the weights, ρ_i, for the temperature classes. Specifically, from [4] we have:

$$c = 0.4, \tag{32}$$

$$a = 2.0, \tag{33}$$

$$tol = 1.0 \times 10^{-5}, \tag{34}$$

where the sequence γ_{a_n} is defined as:

$$\gamma_{a_n} = a^{a_n} \qquad \gamma_{a_n} < tol, \tag{35}$$

$$\gamma_{a_n} = n^{-1} \qquad \text{otherwise.} \tag{36}$$

As done previously, with the inclusion of a third factor T, samples were generated using all combinations of:

1. δ values: 0.3, 0.4, 0.5;
2. Δ_1 values: 0.004, 0.005, 0.006, 0.007; and
3. T: 15, 25, 35.

The plot of the acceptance rates appears in **Figure 4** where we can observe the same trends with respect to Δ and δ as with Metropolis-Hastings. The effect of the number of temperatures, T, causes the choice to be less clear using this criteria, however an additional consideration is that the larger the number of temperature classes, the more samples we will discard. In this case, from the combinations that are near the optimal rate, we selected the one that discards the fewest samples. The settings selected are $\delta = 0.5$, $\Delta = 0.004$, $T = 15$. The 5 million samples generated yielded 342,820 samples in the lowest temperature class for analysis, of which an additional 100,000 are discarded for burnin.

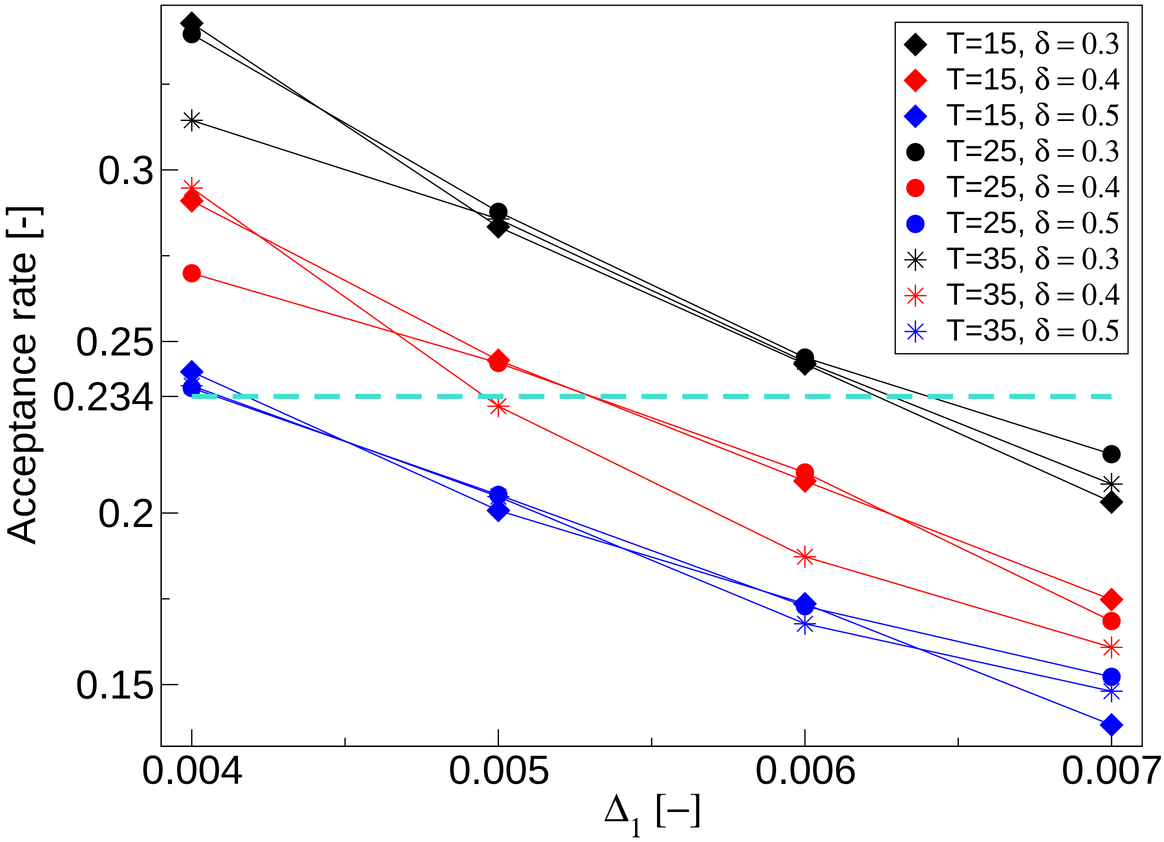


Figure 4: *Acceptance rates for Wang-Landau;* $\varepsilon = 0.001$, $I_{\text{inv}} = 0.2$.

6 Results

6.1 Diagnostic plots and convergence

The choices that have been made in the implementation need to be validated by examining the generated chains to acertain if they are exhibiting undesirable behaviour. Two common methods for quick assessment of a chain are the examination of the autocorrelation function (ACF) and trace plots. It should be emphasised that these plots are not guarantees of convergence or that the chain is well behaved. They are useful to see if the required properties are *not* being met and may suggest ways to fix any problems.

6.1.1 ACF Plots

The first set of diagnostic plots that we will examine are the ACF plots. These figures show the autocorrelation function for each parameter plotted against the number of iterations between the values, where the distance between steps is referred to as the 'lag.' Ideally we would like the autocorrelation to quickly decline to zero.

By comparing the ACF plots for Metropolis-Hastings in **Figure** 5 and Wang-Landau in **Figure** 6, it is readily seen that there is a significant difference in the behaviour of the autocorrelation of the eight model parameters. Although the Wang-Landau results are far from the ideal case, combined with the fact that a lag of 3000 could easily be considered excessive, the rate of decay is much faster than for Metropolis-Hastings. The Metropolis-Hastings results show the autocorrelation function is decreasing for all of the parameters, although quite slowly and not at a rate as satisfying as with Wang-Landau. In either case, there is a strong suggestion that a large number of samples will be necessary.

6.1.2 Trace plots

Next we turn our attention to the adequacy of the number of samples generated, the burnin and convergence of the chain. We may make a quick assessment of these factors by the use of trace plots, which are the sample values plotted against the step index. Ideally, there should be no relationship between the value of the sample and its location in the chain. Trends in the early part of the chain would indicate insufficient burnin and overall patterns may suggest a lack of convergence. Loosely speaking, we want trace plots that look like white noise, with no detectable pattern.

The plots shown in **Figure** 7, for Metropolis-Hastings, overall demonstrate encouraging behaviour. There is no sudden shift in the pattern to indicate insufficient burnin and the chains appear to be meeting the conditions of detailed balance and ergodicity. The plots for θ'_2, θ'_3 and θ'_5 demonstrate a near textbook definition of what we want to see. The other plots, while not as ideal, could be considered adequate, although the patterns suggest that we have a slowly mixing chain which will require a large number of samples. One feature that should be noted is the tendency of the plot for θ'_4 to shift toward the upper limit. A similar, if not as marked, downward trend can be seen in θ'_7. It is likely that this can be ascribed to the pre-chosen boundaries defined for these parameters being off-centre.

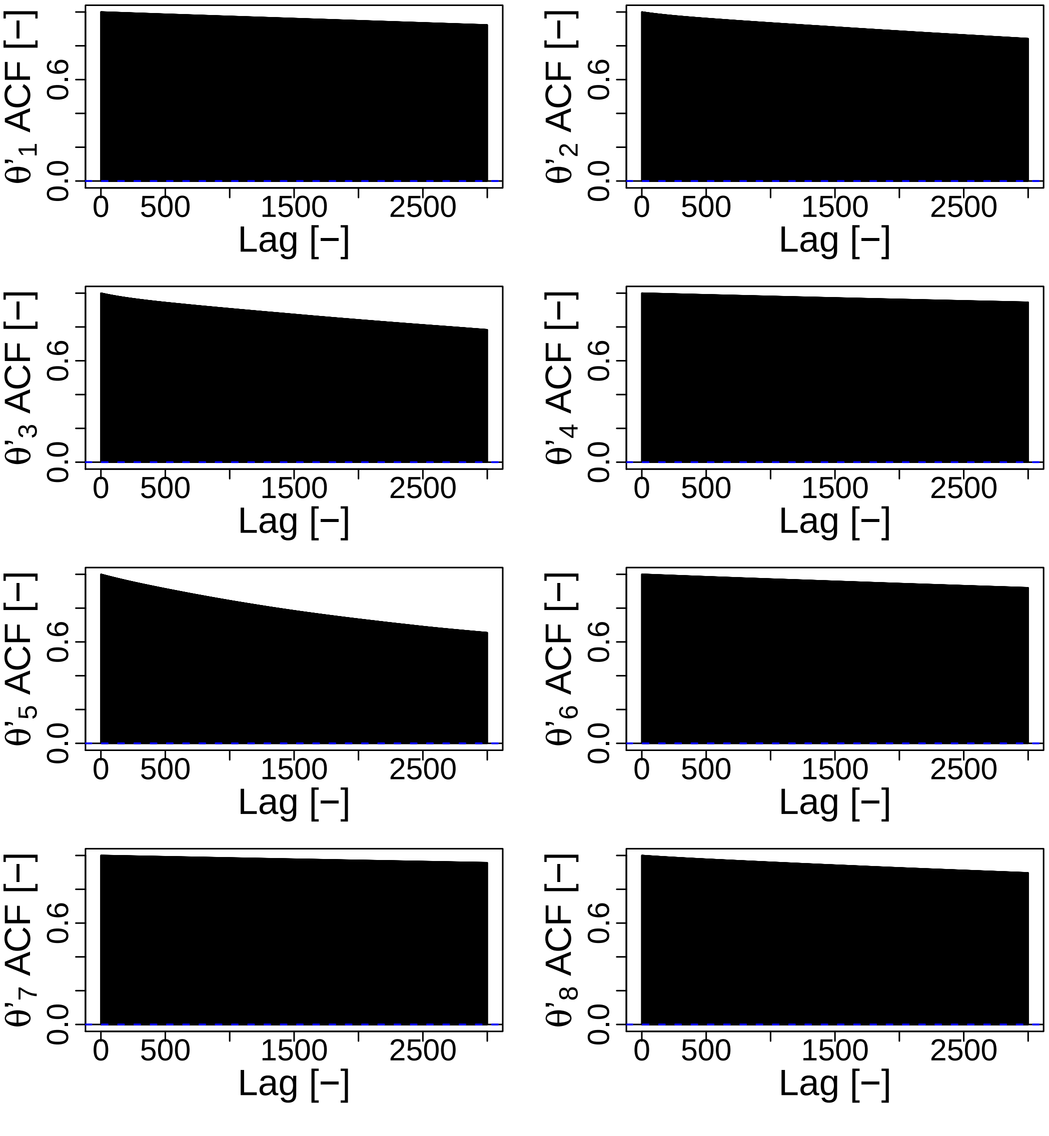


Figure 5: *Autocorrelation function for Metropolis-Hastings:* $\delta = 0.3$, $\Delta = 0.005$.

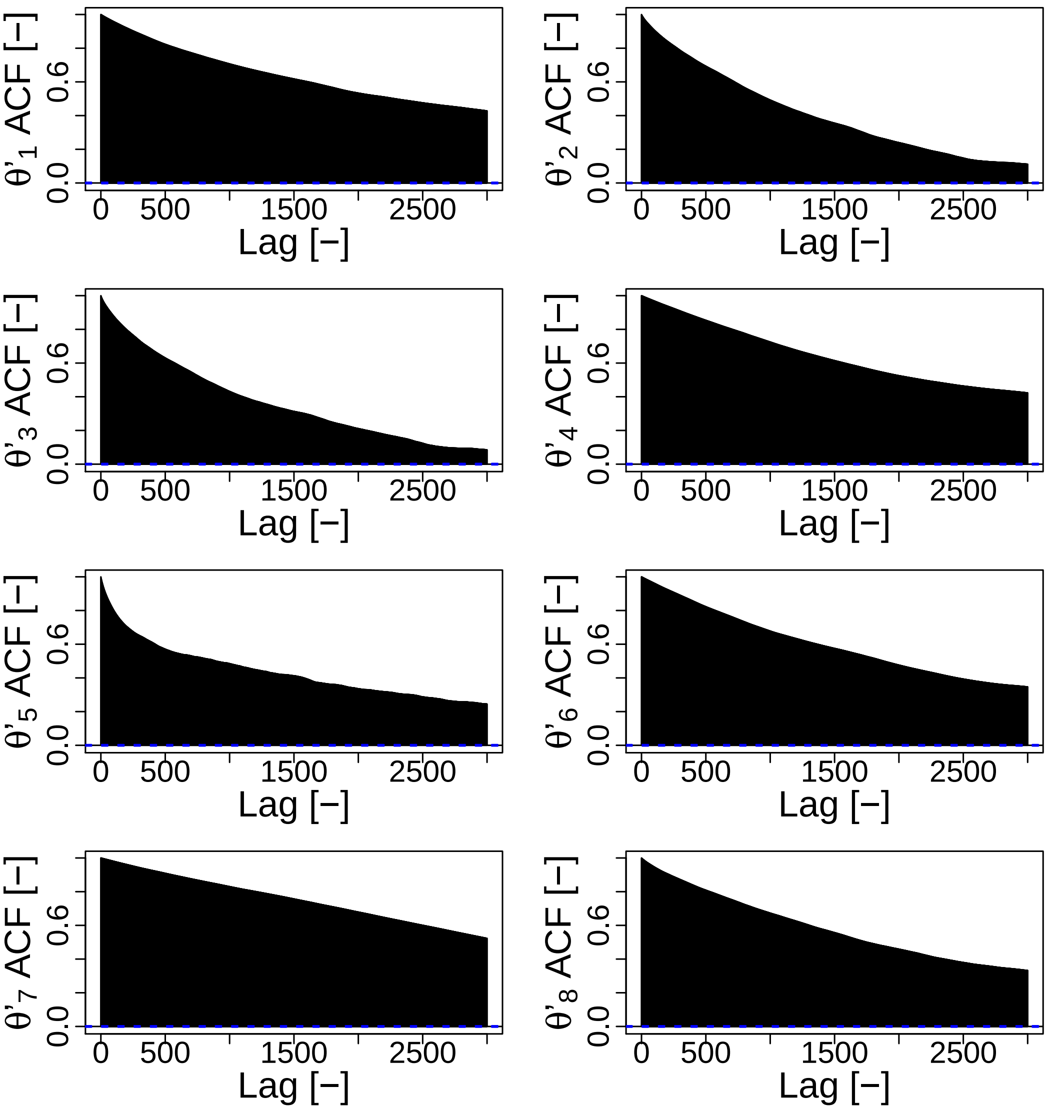


Figure 6: *Autocorrelation function for Wang-Landau:* $\delta = 0.5$, $\Delta = 0.004$, $T = 15$.

The corresponding Wang-Landau plots in **Figure** 8 are not as classically perfect, however there is a significant difference in the available number of samples. Where the Metropolis-Hastings case has 5 million samples, after burnin of 1 million has been removed, the Wang-Landau case has 242,820, after a burnin of 100,000 samples. However, the same patterns can be observed in that θ'_2, θ'_3 and θ'_5 appear to be mixing well while θ'_4 and θ'_7 are vertically shifted. Overall, both sets of plots are encouraging, if suggestive that a change in bounds for θ'_4 and θ'_7 should be attempted in future.

6.2 Analysis of the parameter distributions

6.2.1 Density plots

One method of examining the distribution generated by a Markov Chain is by viewing the density plots for each parameter individually. It is highly desirable for the density plots to appear to be unimodal, as then we will have a single optimal parameter value. Samples in **Figure** 9 and **Figure** 10 appear to be creating a bell-shaped curve, except for θ'_4 and θ'_7. In these cases, our observations from the previous plots are reiterated. In particular, the distributions for those two parameters appear to be left and right truncated, respectively. While this reinforces the suggestion that the imposed boundaries are truncating the distribution, the truncation appears to be happening in the tails. This being the case, it is conceivable that our parameter estimates will be usable.

6.2.2 Marginal posterior 2-D plots

Another consideration is the two-dimensional marginal posterior plots. The results for Wang-Landau are shown in **Figure** 11 and **Figure** 12. The Metropolis-Hastings case presented some difficulties in generating plots due to the large number of samples. The results for Metropolis-Hastings shown in **Figure** 13 and **Figure** 14 represent a random sample of 500,000 from the total of 5 million samples available. Both of these sets of plots illustrate that a central ‘good’ region has been found by the algorithms. It should be noted that it is not appropriate to make a direct comparison between the two algorithms based on these plots, as the kernel density estimation for Metropolis-Hastings is based on more points, which consist of a subset of the available points. Nevertheless, we see again that realisation of θ'_4 and θ'_7 indicate that the bounds may benefit from further investigation, however there are no other glaring errors or evidence of multi-modal solutions. Further, the similarities between the corresponding plots indicate that both methods are creating a realisation of the same distribution. One feature to note is the linear appearance of the plots produced by both algorithms for θ'_2 vs. θ'_3. This shape suggests correlation between these two parameters, which if verified indicates a method to simplify the model.

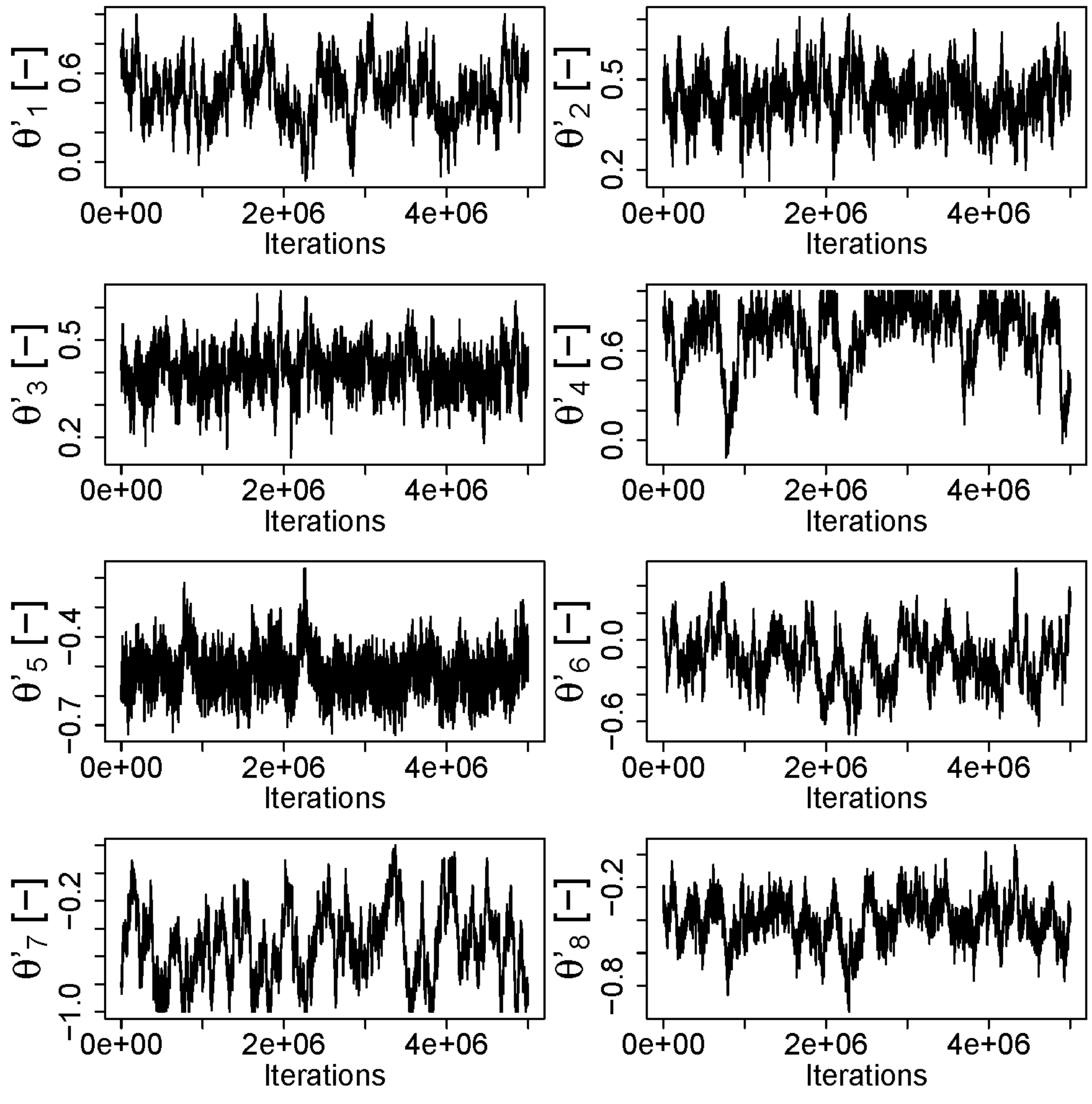


Figure 7: *Trace plots; Metropolis-Hastings:* $\delta = 0.3$, $\Delta = 0.005$.

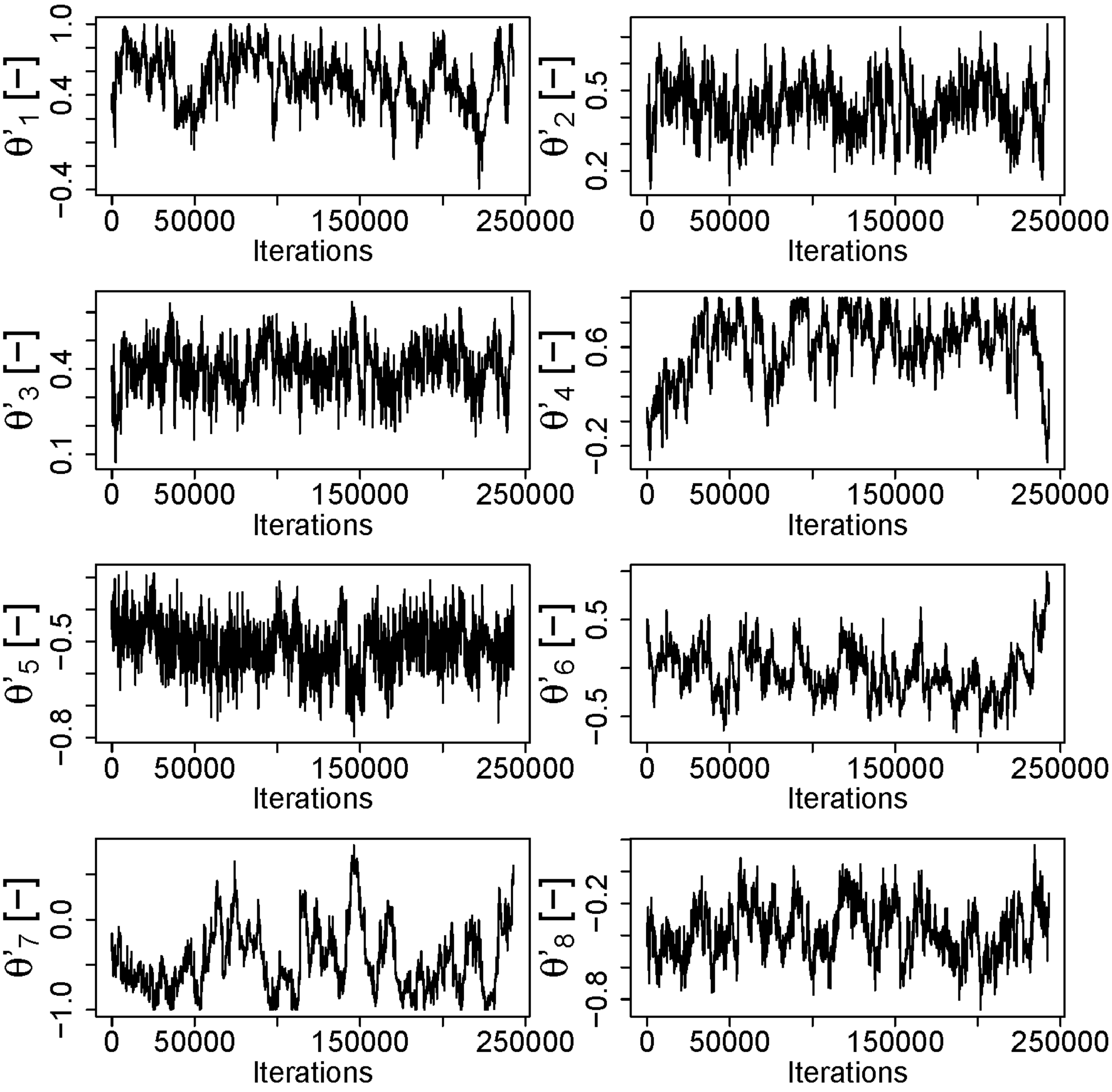


Figure 8: *Trace plots; Wang-Landau:* $\delta = 0.5$, $\Delta = 0.004$, $T = 15$.

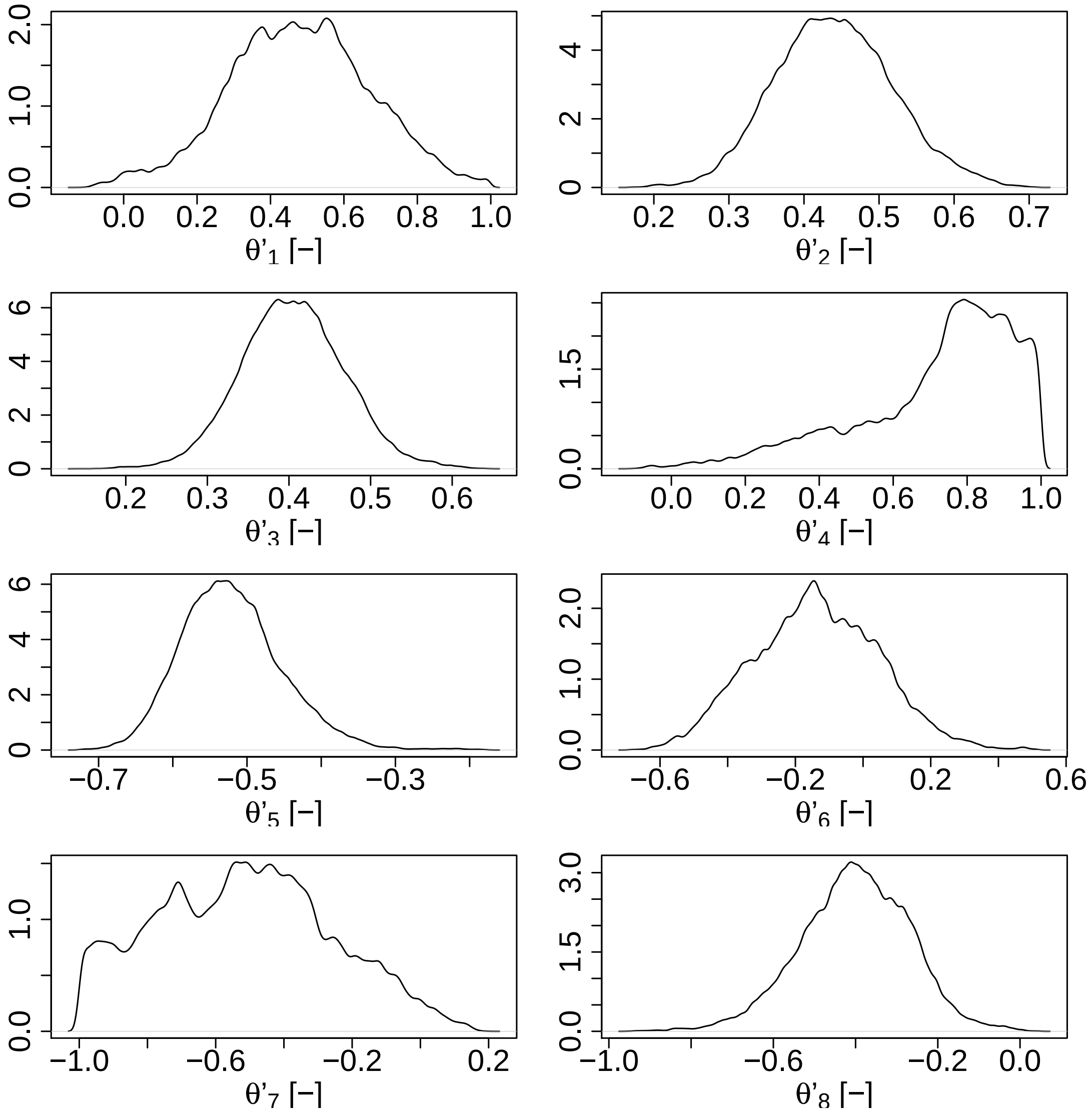


Figure 9: *Marginal posterior probability densities; Metropolis-Hastings:* $\delta = 0.3$, $\Delta = 0.005$.

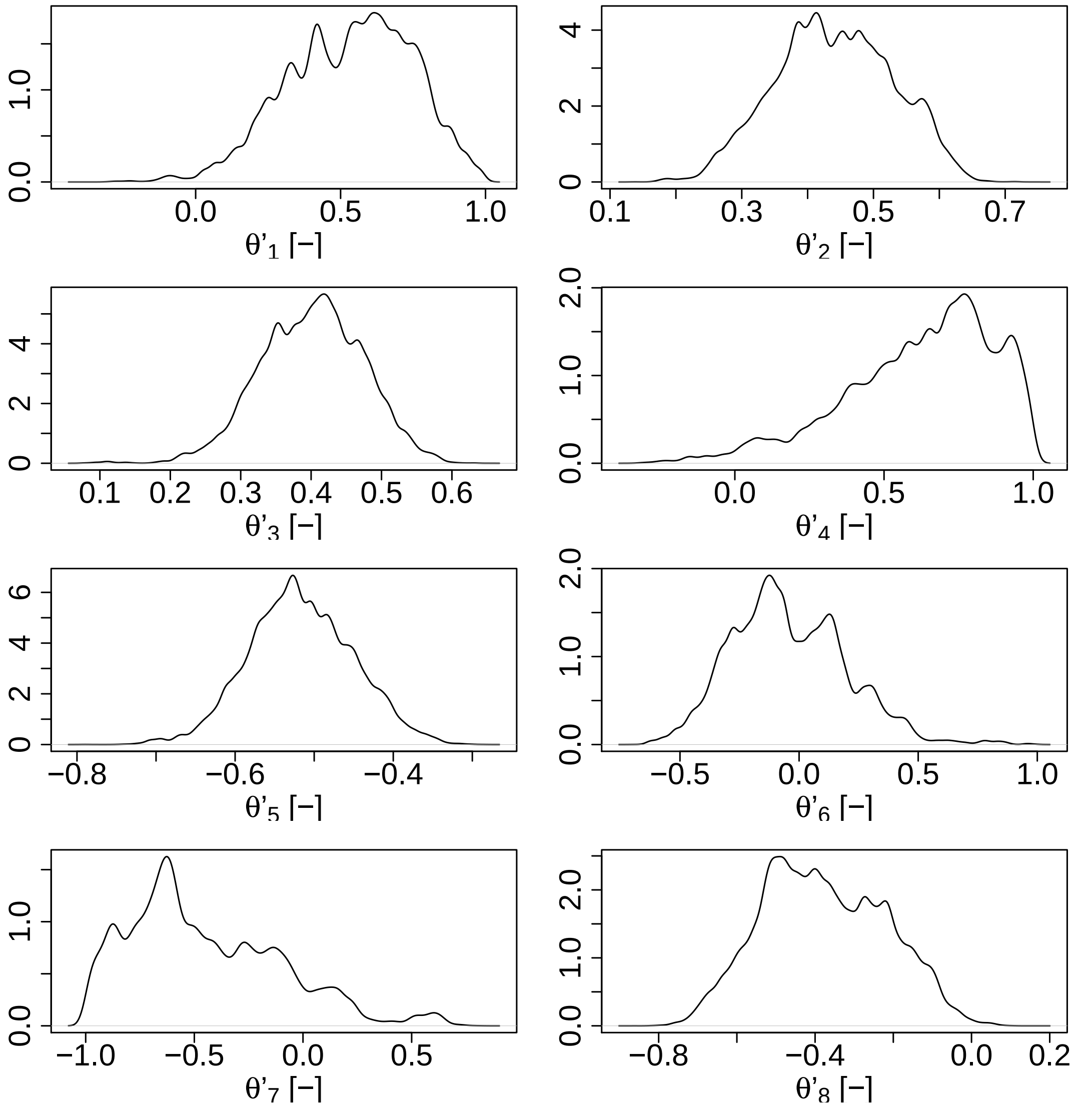


Figure 10: *Marginal posterior probability densities; Wang-Landau:* $\delta = 0.5$*,* $\Delta = 0.004$*,* $T = 15$.

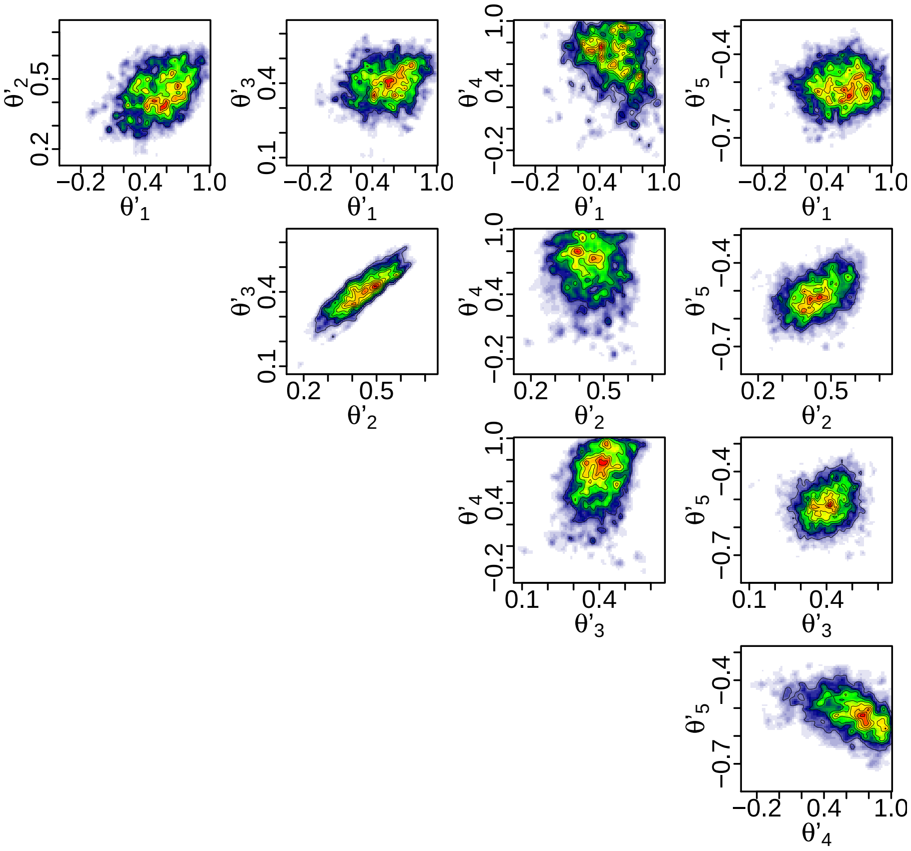


Figure 11: *Wang-Landau paired marginal posterior; 1 of 2.*

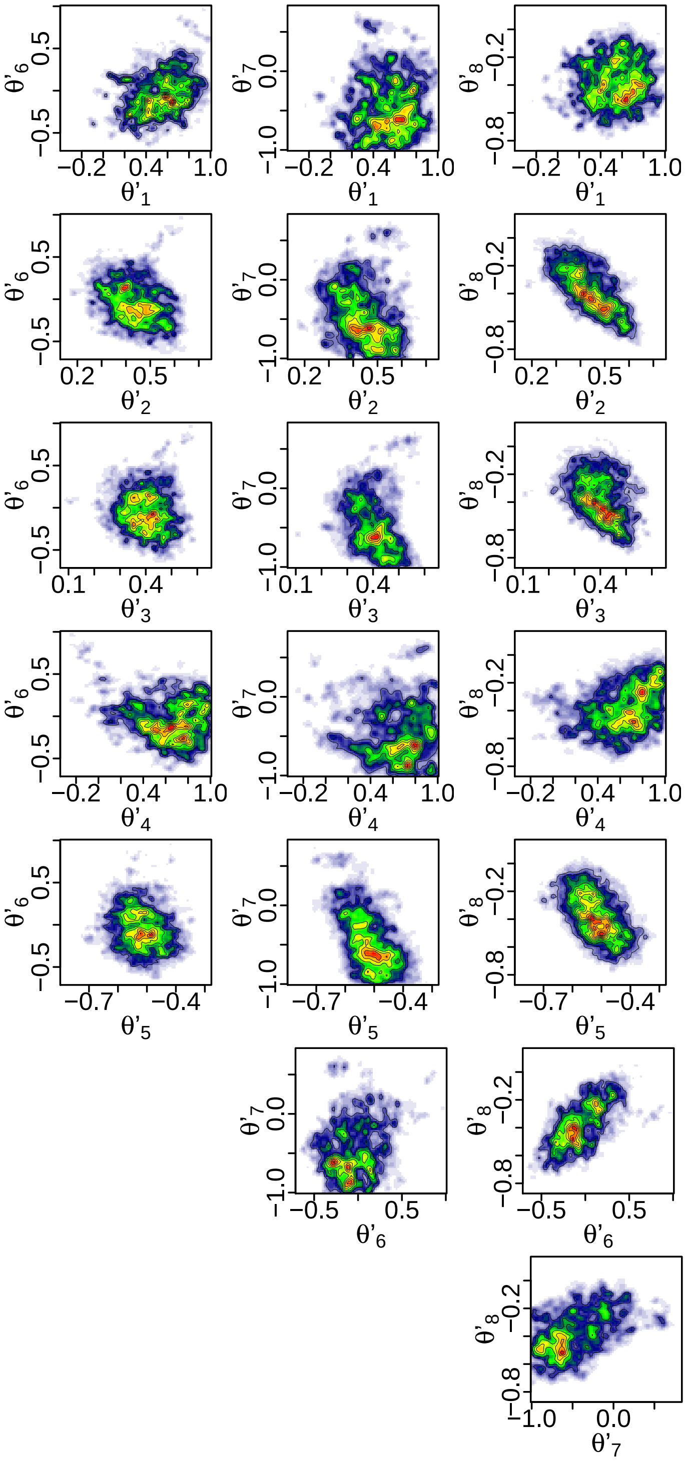


Figure 12: *Wang-Landau paired marginal posterior; 2 of 2.*

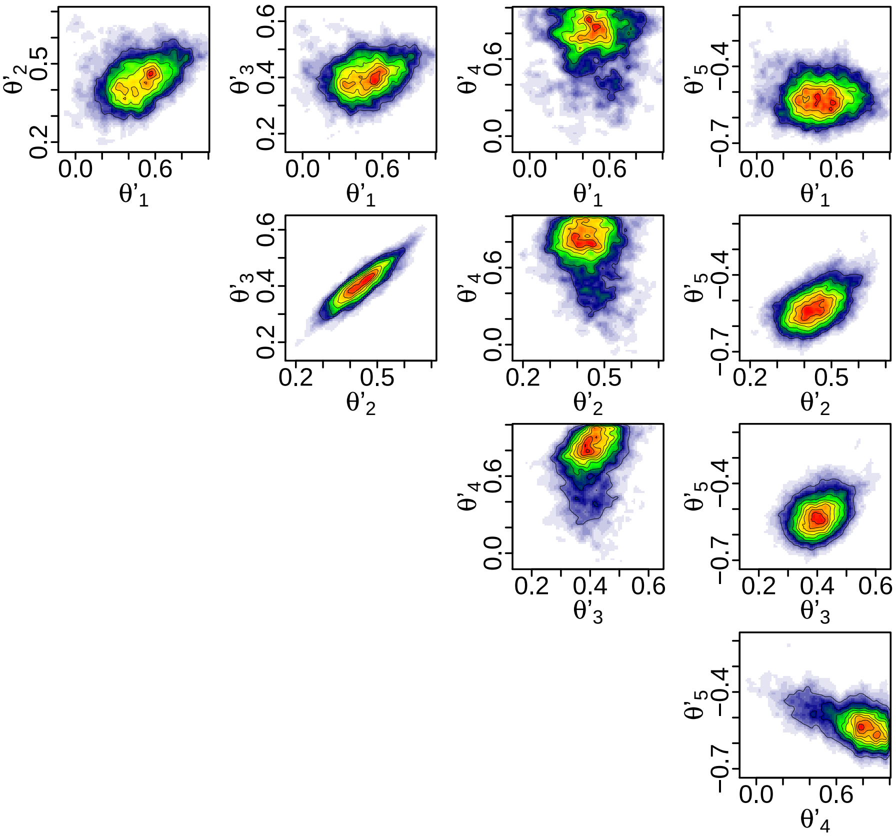


Figure 13: *Metropolis-Hastings paired marginal posterior; subset of 500,000 samples; 1 of 2.*

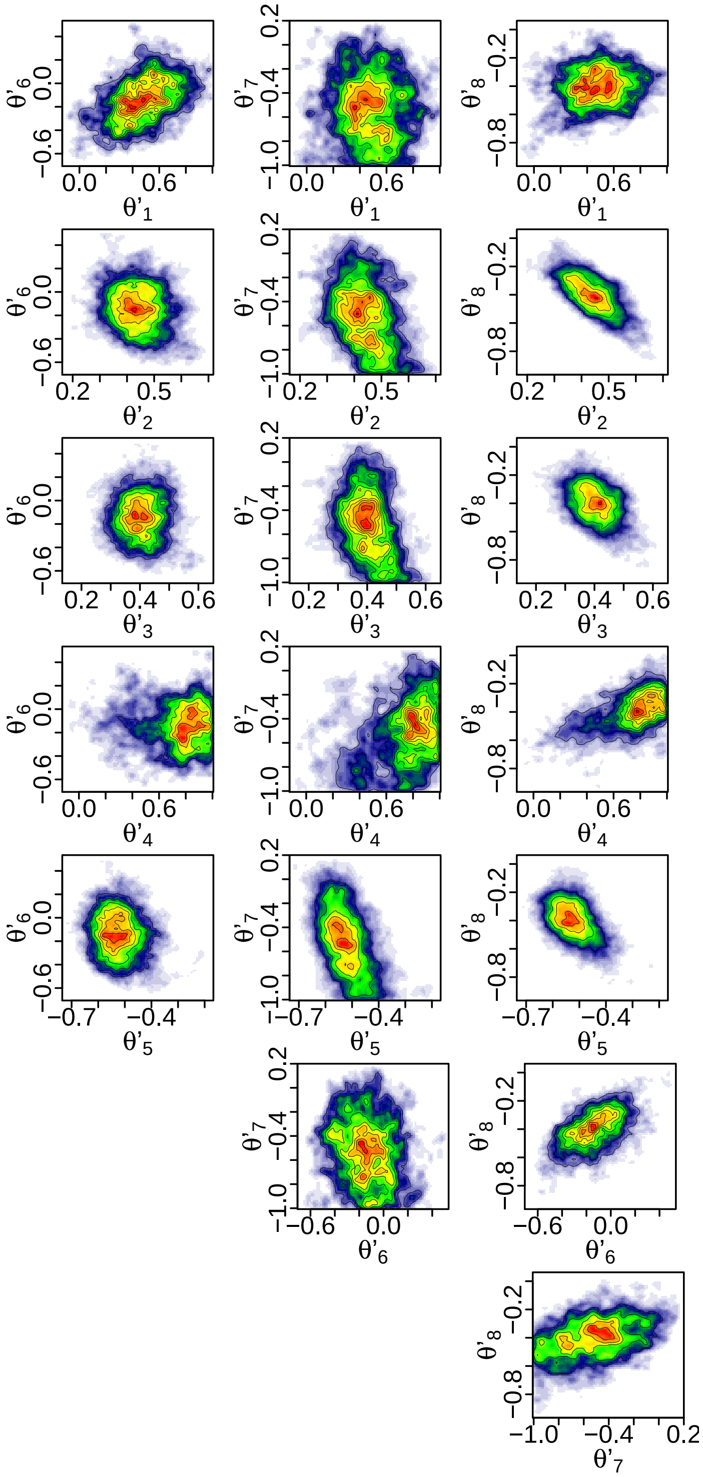


Figure 14: *Metropolis-Hastings paired marginal posterior; subset of 500,000 samples; 2 of 2.*

6.3 Parameter estimates and high probability density regions

We now turn to extracting estimates for the model parameters from the realised distributions. The usual summary statistics, e. g., mean and standard deviation, in general do not yield very useful values in these circumstances. Instead, the value of highest probability density (HPD) together with the bounds of the high probability density region (at some given confidence level) should be used [7]. 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 multi-modal, it may not make sense to pick a best estimate or construct error bars. In this application, the evidence of the one- and two-dimensional density plots suggest that it is reasonable to do so.

In this case, we will use $100\,(1\text{-}\alpha)$ posterior credibility intervals, with $\alpha = 0.0455$, which corresponds to 2 standard deviations. The values for the estimated parameters are the points of maximum density from their respective posterior marginal densities with error bars that represent a credibility region.

The constructed estimates and credibility regions for both sampling algorithms are displayed in **Figure 15**. First, it should be noted that the two sampling algorithms yield largely the same estimates and ranges which supports the claim that the Markov Chain has converged and the samples are drawn from the distribution of interest. The vertical shift of θ'_4 and θ'_7 are apparent in that, while the estimated model parameter values appear to be within the credibility region, the error bars collide with the predefined boundaries. Additionally, the ranges of the error bars are far wider than for the other estimates, with the bars being skewed. This suggests that the probability weight is not being equally distributed, again likely due to sub-optimal bounding for the model parameters.

6.4 Varied starting position

A last aspect that we should examine is the choice of initial values. By beginning the chain at different points, we can establish if the sample size and burnin are sufficient, or if they are having an impact on the end results. **Table 4** shows the uncoded values for the Wang-Landau HPD point of maximum density with errors for the chain we have been analysing for five randomly generated starting vectors. The previous observations are borne out in that the results for θ_4 and θ_7, the HPD regions are bounded on one edge by the predefined boundaries, although the credibility regions themselves are similar. The chains that have been generated largely exhibit similar bounds and points of maximum density. Of the five alternate chains, Run 3 exhibits the most significant deviation from the other chains, which may be attributed to a particularly poor set of initial conditions.

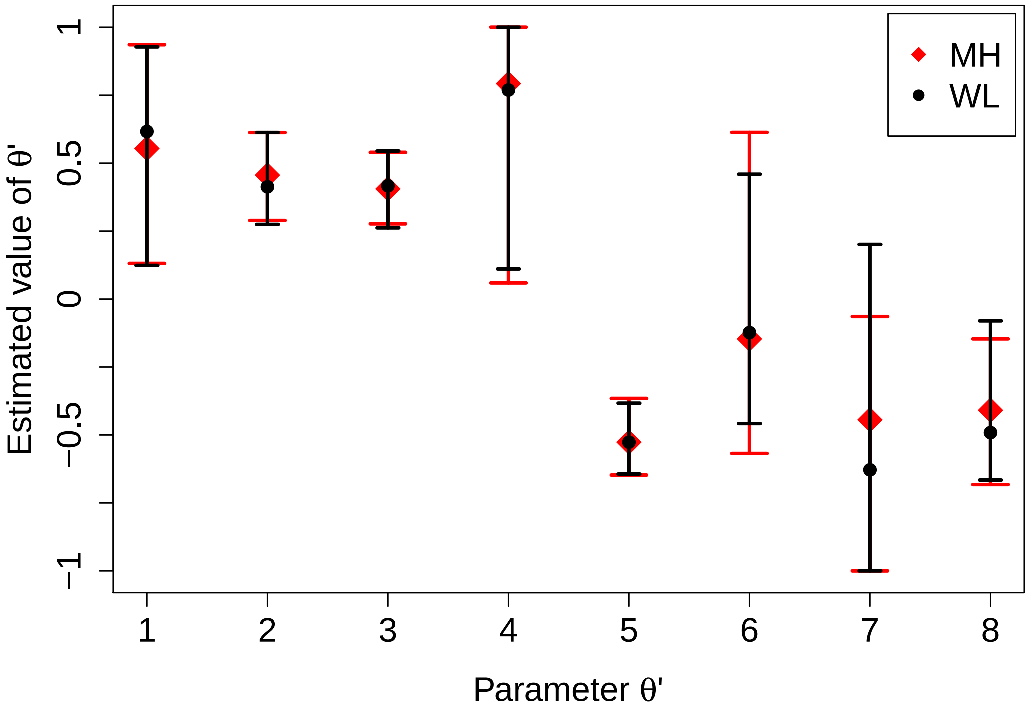


Figure 15: *Comparison of Metropolis-Hastings (MH) and Wang-Landau (WL) high probability density (HPD) region*

Table 4: *HPD decoded estimates of parameters with 6 different randomly chosen initial values; Wang-Landau implementation.*

	θ_1/τ	θ_2/λ	θ_3/β	$\theta_4/x_{50\mathrm{d}}$	$\theta_5/\sigma_{50\mathrm{d}}$	$\theta_6/k_{x_{50\mathrm{d}}}$	$\theta_7/k_{\sigma_{50\mathrm{d}}}$	$\theta_8/k_{\mathrm{disc}}$
	$[10^{-3}\,\mathrm{kg}^{-1}]$	[-]	[-]	$[\mu\mathrm{m}]$	[-]	$[100\,\mathrm{m}^6/\mathrm{kg}^2]$	$[100\,\mathrm{m}^6/\mathrm{kg}^2]$	$[\mathrm{kg}^{-1}]$
HPD lower credibility bound								
Original	1.5619	2.2746	2.2618	13.3302	1.3565	1.1776	0.0100	1.3356
Run 1	1.5797	2.2841	2.2528	11.9020	1.3863	1.0999	0.0100	1.2761
Run 2	1.5035	2.2905	2.2547	12.6522	1.3563	1.0699	0.0103	1.3136
Run 3	1.4983	2.2765	2.2489	9.5094	1.3865	0.9809	0.0108	1.2024
Run 4	1.4779	2.2278	2.2292	10.5739	1.3526	1.1219	0.0101	1.2653
Run 5	1.6042	2.2655	2.2607	11.2073	1.3568	1.2445	0.0100	1.3620
HPD maximum								
Original	1.8082	2.4129	2.4168	18.2681	1.4736	1.5959	0.0267	1.7720
Run 1	1.7495	2.4560	2.4403	15.4501	1.4853	1.5029	0.0223	1.8307
Run 2	1.7926	2.4599	2.4129	18.6497	1.5010	1.6057	0.0239	1.7622
Run 3	1.8669	2.4560	2.4325	14.5695	1.5519	1.8601	0.0297	1.8796
Run 4	1.7593	2.4795	2.4090	17.9159	1.5284	1.6204	0.0216	2.0851
Run 5	1.8063	2.4481	2.3933	15.4795	1.4775	1.8454	0.0273	2.1341
HPD upper credibility bound								
Original	1.9637	2.6127	2.5447	19.9998	1.6171	2.3241	0.0640	2.7995
Run 1	1.9653	2.6300	2.5480	19.8982	1.6484	2.6828	0.0640	2.5886
Run 2	1.9565	2.6226	2.5335	19.9999	1.6176	2.4573	0.0672	2.7989
Run 3	1.9732	2.6717	2.5523	19.2560	1.7072	2.6714	0.0610	2.8026
Run 4	1.9808	2.6564	2.5621	19.9150	1.6690	2.4793	0.0628	2.7629
Run 5	1.9776	2.6351	2.5756	19.9559	1.6556	2.7134	0.0769	2.8000

7 Conclusions

We have demonstrated a Bayesian approach to parameter estimation for a computationally expensive jet-milling model. The method uses both the experimental data and the uncertainties associated with the measurements to create a posterior distribution for the model parameters. The Metropolis-Hastings and Wang-Landau sampling algorithms were used to independently create a realisation of the distribution. Both algorithms produce consistent model parameter estimates. Their behaviour is compared and assessed.

The parameter estimation process starts by constructing a posterior distribution for our model parameters. A realisation of the posterior distribution is generated using the Metropolis-Hastings and Wang-Landau sampling algorithms. These algorithms require a large number of model evaluations and would be intractable using the original computationally expensive model. We make use of quadratic response surfaces as surrogate models in order to obtain parameter estimates in a reasonable timeframe. The model parameter estimates are extracted from the realised distribution in the form of credibility regions, described by a point of maximum density with associated error bars.

The two sampling algorithms were demonstrated to produce similar model parameter estimates. The behaviour of the algorithms was assessed by examination of autocorrelation function plots, trace plots and parameter marginal posterior density plots. The autocorrelation functions indicated that the Wang-Landau algorithm showed a markedly superior rate of decay. The trace plots suggested that good mixing was occurring for both algorithms; which supports the belief that the Markov Chain has converged and the samples are being generated from the distribution of interest for each algorithms. One- and two-dimensional marginal density plots suggested a unimodal bell-shaped distribution for each parameter. The fact that both algorithms produce the same results, strongly suggests that the Markov chains have converged and gives us further confidence in the model parameter estimates. In addition, the parameter estimates were shown to be largely independent of the initial values for the chains by examination of the credibility regions.

The process of obtaining the parameters estimates and testing their validity has suggested characteristics of the model and surrogate model that should be investigated or altered. The 2-D marginal posterior density plots suggest the existence of correlation between at least two of the parameters. A high degree of correlation between parameters indicates that the model could be simplified without degrading its functionality. The tendency of two of the parameter samples to be vertically shifted in the trace plots as well as the truncation in the density plots indicate that the bounds placed on the model parameter space are impacting the Markov Chains. As the truncation appears to be confined to the tails of the distributions, estimates obtained using point of maximum density should be usable; however the error bars are likely to be affected. This behaviour suggests that the initial bounds for the parameters used to generate the surrogate model need to be modified for the affected parameters. It also demonstrates a drawback of this methodology in that you need to define an initial range for the parameters. While bounds can be defined by preliminary testing, the ability to begin with less specific information is desirable.

In the course of this investigation, we have taken arbitrarily large values for the number of samples generated and burnin for the Markov Chains. Further study of minimum values for both elements, particularly in reference to the implementation of Wang-Landau would be both useful and of interest. The Wang-Landau algorithm contains a number of numerical parameters which were arbitrarily established, such as I_{inv} and the method of generating the sequence Δ_j. Similarly, for both algorithms ε was set equal to 0.001, as a value that worked well for the system considered in this paper. Further investigation of these values would be of interest. Additionally, many more sophisticated diagnostic techniques have been described in the literature and could be implemented to further assess the behaviour of the Markov Chains and the validity of the results.

Acknowledgements

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

Nomenclature

Roman symbols

a	Wang-Landau numerical parameter	-
a_n	Wang-Landau numerical parameter	-
A_0	Nozzle area	m^2
$b_{i,0}$	Variable for defining linear transform to $[-1,1]$	-
$b_{i,1}$	Variable for defining linear transform to $[-1,1]$	-
$\overline{B}$	Breakage transition matrix	-
c	Wang-Landau numerical parameter	-
$\mathcal{C}$	n dimensional Hypercube where n is the number of parameters (theoretical)	-
D	Distance parameter for Central Composite Design	-
E_k	Kinetic energy	kJ/kg
E_sp	Specific energy	kJ/kg
$\overline{I}$	Identity matrix	-
I_inv	Inverse temperature ratio	-
k	Number of dimensions of Central Composite Design	-
k_disc	Discharge constant	kg^{-1}
k_r	Initial PSD parameter	-
k_x50d	Hold-up constant (on cut size)	$100\text{m}^6/\text{kg}^2$
$k_{\sigma\text{50d}}$	Hold-up constant (on spread)	$100\text{m}^6/\text{kg}^2$
m	Initial PSD parameter	-
$\dot{m}_\text{feed}$	Feed mass flow rate	kg/h
$\dot{m}_\text{gas}$	gas mass flow	kg/s
M_w	Molecular weight	kg/kmol
M_holdup	Concentration of solid	kg/m^3
$\mathcal{N}_L(x,y)$	Gaussian distribution with mean x and covariance matrix y	-
p_grind	Grinding pressure	barg
$\overline{P}_\text{disc}$	Discharge probability	-
$q(x \rightarrow y)$	Transition probability from state x to state y	-
R	Universal gas constant	J/(kmol K)
$\overline{S}$	Breakage frequency vector	1/s
S_i	ith breakage frequency	1/s
T	Number of temperature classes	-
T_gas	Gas temperature	K
tol	Lower limit for resolution	-
v_gas	velocity of gas	m/s
x_{10}	Particle class (10th pencentile)	μm
x_{50}	Particle class (50th pencentile)	μm
x_50d	Cut size	μm
x_{90}	Particle class (90th pencentile)	μm

x_i	Discrete particle size classification	-
X_{solid}	Fraction of solid material	-
$\overline{X}_{w,\text{feed}}$	Mass fraction characterisation	μm
$\mathcal{X}$	n dimensional Hypercube with edgelength Δ, where n is the number of parameters estimated	-

Greek symbols

α	Parameter for Inverse-Wishart distribution	-
β	Empirical parameter	-
$\beta_{l,0}^{(n)}$	Constant coefficient of response surface for lth response of model at n process condition	-
$\beta_{l,i}^{(n)}$	i Linear coefficients of response surface for lth response of model at n process condition	-
$\beta_{l,ij}^{(n)}$	i Quadratic coefficients of response surface for lth response of model at n process condition	-
γ_n	Positive increasing sequence	-
δ	Probability of making a large jump	-
Δ	Edgelength for MCMC jumps	-
Δ_i	Wang-Landau edglength for nth temperature	-
$\varepsilon^{(n)}$	Vector of measurement errors	-
η	Vector of model responses	-
η^{exp}	Vector of experimental responses	-
η_i^{exp}	ith experimental response	-
$\eta_i^{\text{exp},(n)}$	nth experimental response for ith process conditions	-
$\eta^{\text{exp},(n)}$	Experimental responses for ith process conditions	-
η_i	ith model response	-
$\eta^{(n)}$	Model responses for ith process conditions	-
$\eta_i^{(n)}$	nth model response for ith process conditions	-
θ	Vector of model parameters	-
θ_i	ith model parameter	-
θ_i'	ith linearly transformed (coded) value for θ	-
κ	Wang-Landau numerical parameter	-
$\varkappa$	Heat capacity ratio	-
λ	Empirical parameter	-
π	Markov Chain with stationary distribution identical to distribution of interest	-
ρ_i	Wang-Landau function to select next temperature class	-
σ_{50d}	Spread	-
Σ	Covariance matrix	-
τ	Empirical parameter	$10^{-3}\,\text{kg}^{-1}$
ϕ_n	Wang-Landau frequency parameter	-

ξ	Vector of experimental process conditions	-
ξ_i	ith experimental process condition	-
$\xi_i^{(n)}$	nth experiment performed at ith process conditions	-
$\xi^{(n)}$	Experiments performed at ith process conditions	-
Ψ	Positive definite matrix for Inverse-Wishart distribution	-

References

[1] H. Adi, I. Larson, and P. Stewart. Use of milling and wet sieving to produce narrow particle size distributions of lactose monohydrate in the sub-sieve range. *Powder Technology*, 179(1–2):95–99, 2007. doi:10.1016/j.powtec.2007.01.020.

[2] T. Allen. *Particle Size Measurement*, volume 1. Chapman and Hall, London, England, 5th edition, 1997.

[3] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, 2008. doi:10.1007/s11222-008-9110-y.

[4] Y. F. Atchadé and J. S. Liu. The Wang-Landau algorithm in general state spaces: Applications and convergence analysis. *Statistica Sinica*, 20(1):209–233, 2010.

[5] K. J. Beers. *Numerical Methods for Chemical Engineering*. Cambridge University Press, Cambridge, 2007.

[6] H. Berthiaux, C. Varinot, and J. Dodds. Approximate calculation of breakage parameters from batch grinding tests. *Chemical Engineering Science*, 51(19):4509–4516, 1996. doi:10.1016/0009-2509(96)00275-8.

[7] G. Blau, M. Lasinski, S. Orcun, S. 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.

[8] 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.

[9] 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.

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

[11] A. Braumann, P. L. W. Man, and M. Kraft. The inverse problem in granulation modelling – two different statistical approaches. *AIChE Journal*, 57(11):3105–3121, 2011. doi:10.1002/aic.12526.

[12] G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992. doi:10.2307/2685208.

[13] V. Ĉerný. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45 (1):41–51, 1985. doi:10.1007/BF00940812.

[14] M. K. Cowles and B. P. Carlin. Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434): 883–904, 1996.

[15] S. G. Davis, A. V. Joshi, H. Wang, and F. Egolfopoulos. An optimized kinetic model of H_2/CO combustion. *Proceedings of the Combustion Institute*, 30(1):1283–1292, 2005.

[16] 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(1):47–73, 1992. doi:10.1016/0360-1285(92)90032-V.

[17] D. Gajda, C. Guihenneuc-Jouyaux, J. Rousseau, K. Mengersen, and D. Nur. Use in practice of importance sampling for repeated MCMC for poisson models. *Electronic Journal of Statistics*, 4:361–383, 2010. doi:10.1214/09-EJS527.

[18] C. Geyer and E. Thompson. Annealing Markov Chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920, 1995.

[19] W. R. Gilks and G. O. Roberts. Strategies for improving MCMC. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 89–114. Chapman and Hall, Boca Raton, Florida, 1996.

[20] W. R. Gilks, G. O. Roberts, and E. I. George. Adaptive direction sampling. *The Statistician*, 43(1):179–189, 1994.

[21] W. R. Gilks, S. Richardson, and G. O. Roberts. Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 89–114. Chapman and Hall, Boca Raton, Florida, 1996.

[22] H. J. C. Gommeren, D. A. Heitzmann, J. A. C. Moolenaar, and B. Scarlett. Modelling and control of a jet mill plant. *Powder Technology*, 108(2–3):147–154, 2000. doi:10.1016/S0032-5910(99)00213-2.

[23] 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.

[24] T. Jansson, A. Andersson, and L. Nilsson. Optimization of draw-in for an automotive sheet metal part: an evaluation using surrogate models and response surfaces. *Journal of Materials Processing Technology*, 159(3):426–434, 2005.

[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] P. L. W. Man, A. Braumann, and M. Kraft. Resolving conflicting parameter estimates in multivariate population balance models. *Chemical Engineering Science*, 65(13):1038–4045, 2010. doi:10.1016/j.ces.2010.03.042.

[27] N. Midoux, P. Hošek, L. Pailleres, and J. R. Authelin. Micronization of pharmaceutical substances in a spiral jet mill. *Powder Technology*, 104(2):113–120, 1999. doi:10.1016/S0032-5910(99)00052-2.

[28] 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.

[29] S. Mosbach, A. Braumann, P. L. W. Man, C. A. Kastner, G. P. E. Brownbridge, and M. Kraft. Iterative improvement of Bayesian parameter estimates for an engine model by means of experimental design. *Combustion and Flame*, In Press, 2011. doi:10.1016/j.combustflame.2011.10.019.

[30] 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.

[31] R. H. Perry, D. W. Green, and J. O. Maloney. *Perry's chemical engineers' handbook*. McGraw-Hill, New York, 7th edition, 1997.

[32] G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18:349–367, 2009.

[33] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7(1):110–120, 1997.

[34] 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.

[35] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8(1):1–9, 1998.

[36] A. Vikhansky, M. Kraft, M. Simon, S. Schmidt, and H.-J. Bart. Droplets population balance in a rotating disc contactor: An inverse problem approach. *AIChE Journal*, 52(4):1441–1450, 2006. doi:10.1002/aic.10735.

[37] F. Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters*, 86(10):2050–2053, 2001. doi:10.1103/PhysRevLett.86.2050.