This is a repository copy of *Bayesian System Identification of Nonlinear Systems: Informative Training Data through Experimental Design*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/81830/

Bayesian System Identification of Nonlinear Systems: Informative Training Data through Experimental Design

P.L.Green[1], K.Worden[1]

[1]Department of Mechanical Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD

email: p.l.green@sheffield.ac.uk, k.worden@sheffield.ac.uk

ABSTRACT: This paper addresses the situation where one is performing Bayesian system identification on a nonlinear dynamical system using a set of experimentally-obtained training data. To be more specific, an investigation is performed to find the optimum form of excitation that should be used during generation of the training data. To that end, the Shannon entropy is used as an information measure such that, through analysing the information content of the posterior parameter distribution, the `informativeness' of different sets of training data can be assessed. In the current work the form of excitation is parameterised thus allowing the choosing of an appropriate excitation to be phrased as an optimisation problem (where one is aiming to maximise the information content of the training data).

KEY WORDS: Nonlinear Dynamics, System Identification, Bayesian Inference, Shannon Entropy

1   INTRODUCTION

Defining y and x respectively as the input and output of a system, the process of system identification involves using a set of observed data ($\mathcal{D}$) to infer a mapping

$$\mathcal{M}(\boldsymbol{\theta})\colon y \to x \qquad (1)$$

which can be used to approximate the behavior of the system of interest. In the context of the current work, $\mathcal{M}$ is a mathematical model of a (potentially nonlinear) dynamical system, $\boldsymbol{\theta} \in \mathbb{R}^{N_\theta}$ is a vector of parameters within that model and $\mathcal{D}$ is input and output data which has been obtained experimentally. In the case where one has already selected a model structure, probabilistic estimates of the model parameters can be realised using Bayes' theorem:

$$P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = \frac{P(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})P(\boldsymbol{\theta}|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})} \qquad (2)$$

where the prior $P(\boldsymbol{\theta}|\mathcal{M})$ is a probability density function (PDF) which represents one's knowledge of the parameters before the data $\mathcal{D}$ was witnessed. $P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$ is known as the posterior and represents one's knowledge of the parameters after the data $\mathcal{D}$ was witnessed. The transformation from prior to posterior is controlled by the likelihood, $P(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})$, which represents the probability of witnessing the data $\mathcal{D}$ given one's choice of parameters and model (a more thorough description of the likelihood is given in the following section). The denominator of equation (2) is known as the evidence - it is a normalising constant which is obtained through marginalising the posterior PDF over $\boldsymbol{\theta}$.

In addition to the above, a Bayesian framework allows one to adopt a probabilistic approach to model selection; in the case where there are a set of competing model structures the probability of each model structure can also be assessed using Bayes' theorem:

$$P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D})}. \qquad (3)$$

As a result of difficulties in evaluating the evidence (a consequence of the curse of dimensionality), as well as the often complex geometry of the posterior parameter distribution (particularly when the system is nonlinear), it is now common practice to utilise Markov chain Monte Carlo (MCMC) methods when addressing Bayesian inference problems. MCMC methods involve the creation of a Markov chain whose stationary distribution is equal to the posterior parameter distribution (thus allowing one to generate dependent samples from $P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$). There exists a varied assortment of MCMC methods – some of which can be used to generate samples from the posterior parameter distribution [1-3] while others are also capable of addressing model selection [4-6].

While undoubtedly useful, MCMC methods often require many model runs and, as such, tend to be computationally expensive. In recent work [7,8] it was suggested that the computational cost of MCMC methods could be reduced through the use of relatively small amounts of 'highly informative' training data. This involved using estimates of the Shannon entropy to measure the information content of a set of data such that, in subsequent MCMC simulations, one could use a relatively small subset of the data which was still highly informative with regard to one's parameter estimates / choice of model.

The current paper aims to extend the concept of 'highly informative training data' towards experimental design. Specifically, in a situation where one is dynamically exciting a structural system with the aim of performing system

identification, it aims to answer the question: which excitation will provide the most informative training data? In this preliminary study the term 'informative' is used with regard to one's parameter estimates - the authors intend to extend the concepts within this paper towards model selection in future work.

To summarise, the method proposed herein involves running simulations of a real system such that, through measuring the information content of the resulting training data, the optimum excitation type for subsequent laboratory experiments can be determined.

2   BAYESIAN FRAMEWORK

Consider the situation where one is analysing a dynamical system which possesses N_D degrees of freedom (DOF). During an experimental investigation, N data points of system response are measured at each DOF. When forming the likelihood it is common practice to assume that, as a result of measurement and modelling error, each data point is independently[1] corrupted by Gaussian noise of variance σ^2 (drawing on the central limit theorem). Thereupon, by defining $x_i^{(j)}$ and $\hat{x}_i^{(j)}(\boldsymbol{\theta})$ as the ith data point at the jth DOF according to the measured data and the model response respectively, the likelihood is given by:

$$P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) \propto \exp(-J_L(\boldsymbol{\theta})) \tag{4}$$

where

$$J_L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{N_D} \left(x_i^{(j)} - \hat{x}_i^{(j)}(\boldsymbol{\theta}) \right)^2. \tag{5}$$

The prior is a user-defined PDF which will be written as:

$$P(\boldsymbol{\theta}|\mathcal{M}) \propto \exp(-J_P(\boldsymbol{\theta})). \tag{6}$$

Using these definitions for the likelihood and prior, the posterior is:

$$P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) \propto \exp(-J(\boldsymbol{\theta})) \tag{7}$$

where $J(\boldsymbol{\theta}) = J_L(\boldsymbol{\theta}) + J_P(\boldsymbol{\theta})$.

3   PROPOSED METHODOLOGY

3.1   Shannon Entropy

For a PDF $\pi(\boldsymbol{\theta})/Z$ (where Z is a normalising constant), the Shannon entropy (which will be referred to as 'the entropy' from now on) is defined as

---

[1] Although it is a strategy that has been adopted here, it is interesting to note that one does not necessarily have to assume that each data point is uncorrelated – see [9] for a detailed discussion.

$$S = -\int \frac{\pi(\boldsymbol{\theta})}{Z} \ln \left(\frac{\pi(\boldsymbol{\theta})}{Z} \right) d\boldsymbol{\theta}. \tag{8}$$

As with statistical physics, equation (8) can be viewed as a measure of uncertainty – the lower the entropy the more confidence one has in $\boldsymbol{\theta}$. In the current paper the PDF $\pi(\boldsymbol{\theta})/Z$ is chosen to be the posterior parameter distribution such that a reduction in entropy represents an increase in confidence with regard to one's parameter estimates.

As one frequently finds that the geometry of the posterior is complex and that the evidence term is difficult to evaluate, the exact entropy of the posterior is often unattainable. In such cases, it is advantageous to approximate the posterior as being Gaussian. Using a second-order Taylor series expansion of $J(\boldsymbol{\theta})$ about the most probable parameter estimates (denoted $\boldsymbol{\theta}_0$) it is possible to shown that:

$$P^*(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) \propto \exp\left(-\frac{1}{2} \Delta\boldsymbol{\theta}[\boldsymbol{H}]\Delta\boldsymbol{\theta}^T \right) \tag{9}$$

where the asterisk indicates that the posterior has been approximated, $\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta}_0$ and $[\boldsymbol{H}]$ is the Hessian matrix:

$$[\boldsymbol{H}]_{k.l} = \frac{\partial^2 J(\boldsymbol{\theta})}{\partial\theta_k \partial\theta_l} = \frac{\partial^2 J_L(\boldsymbol{\theta})}{\partial\theta_k \partial\theta_l} + \frac{\partial^2 J_P(\boldsymbol{\theta})}{\partial\theta_k \partial\theta_l}. \tag{10}$$

Throughout this paper, the elements of $[\boldsymbol{H}]$ are estimated using finite difference methods. From equation (9) it can be seen that the inverse of $[\boldsymbol{H}]$ is equal to the covariance matrix of the approximated posterior. It should be noted that uniform priors are used throughout this work and, as such, the second derivative in equation (10) is always equal to zero.

Substituting the Gaussian approximation of the posterior into equation (8), one finds that the entropy of $P^*(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$ is given by

$$S = \frac{N_\theta}{2}(1 + \ln(2\pi)) + \frac{1}{2}\ln(\det([\boldsymbol{H}]^{-1})). \tag{11}$$

It is possible to show that, in the case where $[\boldsymbol{H}]$ is diagonal, minimising the Shannon entropy is equivalent to minimising the diagonal elements of the posterior covariance matrix (therefore maximising one's confidence in the parameter estimates) [8].

3.2   Algorithm

The algorithm used to analyse the effectiveness of different excitation parameters is now described using pseudo-code. Before the algorithm can be implemented, one must select an excitation type ($\mathcal{M}_e$) (a square wave for example) as well as a vector of parameters ($\boldsymbol{\theta}_e$) which control certain features of the excitation (amplitude and frequency for example). Additionally, one needs an estimate of the measurement noise variance (σ^2) which is likely to occur during experimental testing (this is based on the assumption that only the measured output has been corrupted by noise).

Finally, one must raster ($\boldsymbol{\theta}_e$) into a grid of values which are to be investigated: $(\boldsymbol{\theta}_e^{(1)}, \boldsymbol{\theta}_e^{(2)}, ..., \boldsymbol{\theta}_e^{(N_e)})$ before proceeding as follows:

For $n = 1 : N_e$
- Generate excitation $y(\boldsymbol{\theta}_e^{(n)})$
- Generate $\hat{x}(\boldsymbol{\theta}_0)$, the response of the model to input y.
- Corrupt the model response with measurement noise – this is equivalent to simulating the data that one would typically witness experimentally
- Estimate the entropy of the posterior parameter distribution

End

The optimum excitation parameters are those which minimise the entropy of the posterior.

3.3 Probabilistic Entropy Estimates

It is worth observing that, as a result of the measurement noise, there will always be uncertainty in one's entropy estimates - this will inevitably lead to uncertainty in the optimum choice of excitation parameters.

To address this, the authors considered using a Bayesian framework to realise probabilistic estimates of the optimum $\boldsymbol{\theta}_e$:

$$P(\boldsymbol{\theta}_e | S, \mathcal{M}_e) = \frac{P(S|\boldsymbol{\theta}_e, \mathcal{M}_e) P(\boldsymbol{\theta}_e|\mathcal{M}_e)}{P(S|\mathcal{M}_e)} \qquad (12)$$

where, in this case, the likelihood would be defined as

$$P(S|\boldsymbol{\theta}_e, \mathcal{M}_e) = (2\pi\sigma_e^2)^{-1/2} \exp\left(-\frac{1}{2\sigma_e^2}(S - E[S])^2\right). \qquad (13)$$

However, this would require estimates of the average entropy as well as the variance term σ_e^2. This is further complicated by the fact that σ_e^2 is likely to change depending on the excitation parameters (a low amplitude response will lead to a higher signal-to-noise ratio and, as such, more uncertainty in the entropy estimates).

Consequently, rather than adopt this methodology, the authors chose to generate probabilistic entropy estimates using a frequentist approach (by generating an ensemble estimate of the entropy). This makes use of the fact that, as numerical simulations are used to analyse the effect of the excitation type, it is possible to conduct many experiments under the same conditions.

3.4 Reducing Numerical Errors

As the elements of $[\boldsymbol{H}]$ are approximated using finite difference methods, the entropy estimation process will always be prone to numerical errors. It was found that these errors could be reduced significantly using a simple strategy. Recalling that $[\boldsymbol{H}]^{-1}$ is the covariance matrix of the approximated posterior then it follows that any estimate which

leads to the diagonal elements of $[\boldsymbol{H}]^{-1}$ being negative must be false. By instructing the algorithm to ignore such results, it was found that the uncertainty in the entropy estimates could be greatly reduced.

4 RESULTS

In this section, the proposed method is demonstrated on several nonlinear dynamical systems.

4.1 SDOF System: Coulomb Nonlinearity with Square Wave Excitation

Initially, a base-excited SDOF system with nonlinear damping was considered (Figure 1). The equation of motion of the system is

$$\ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x + \alpha\,\mathrm{sgn}(\dot{x}) = -\ddot{y} \qquad (14)$$

where $\ddot{y}$ is a user-defined base acceleration, x is the resulting displacement of the mass, ζ is the damping ratio, ω_n is the natural frequency and α modulates the level of Coulomb damping in the system. The magnitude of these parameters is shown in Table 1.

For the first part of this investigation, $\mathcal{M}_e$ was chosen such that a square-wave base acceleration of amplitude 4 m/s^2 was used to excite the system:

$$\ddot{y} = \mathrm{sgn}[4\sin(2\pi f_e t)]. \qquad (15)$$

The frequency of the square wave was left to be determined such that, in the context of this paper, $\theta_e = f_e$. Each simulation of a real test was corrupted with Gaussian measurement noise of standard deviation $\sigma = 1 \times 10^{-4}$ which, depending on the excitation frequency, resulted in a signal-to-noise ratio of between 15 and 60.
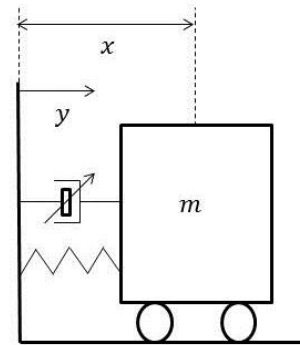

Figure 1: Schematic diagram of the system described by equation (14).

Table 1: Parameters of the dynamical system described by equation (14).

Parameter	Magnitude	Units
ζ	0.01	-
ω_n	$2\pi 10$	rad/s
α	0.1	N/kg

Treating all of the model parameters ($\boldsymbol{\theta}$) as unknown, the entropy of the posterior parameter distribution was estimated for a range of different excitation frequencies. The ensemble average entropy, as well as confidence bounds, is shown in Figure 2.
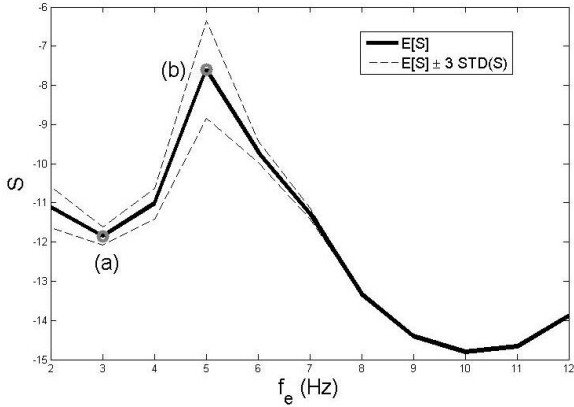

Figure 2: Variation of the Shannon entropy with the frequency of a square-wave excitation for the system described by equation (14).

It is immediately obvious that the entropy of the posterior will be minimised if one excites the system at its resonance frequency (10 Hz). This is simply because this maximises the signal-to-noise ratio – the relatively large response of the system is as far from the noise floor of the measurement noise as possible. It is important to note that Figure 2 shows the entropy of the entire posterior distribution – it does not show how one's confidence in each *individual* parameter estimate changes with f_e (this is addressed in section 4.3 of the current work).

Two other points are of particular interest ((a) and (b) on Figure 2). According to the entropy estimates, point (b) (a square-wave excitation of 5 Hz) should yield relatively uncertain parameter estimates relative to point (a) (a square-wave excitation of 3 Hz). This was confirmed using MCMC simulations (the Metropolis algorithm specifically), the results of which are shown in Figure 3. The MCMC results have been normalised such that the Gaussian approximations of the posterior, which form an essential part of the entropy estimation, can be shown alongside. One can see that excitation (a) has indeed led to less uncertain parameter estimates relative to excitation (b). It is interesting to note that the Gaussian approximations are fairly poor, although they have still been able to predict which excitations will lead to better parameter estimates – this is a trend which the authors encountered throughout all the subsequent investigations.
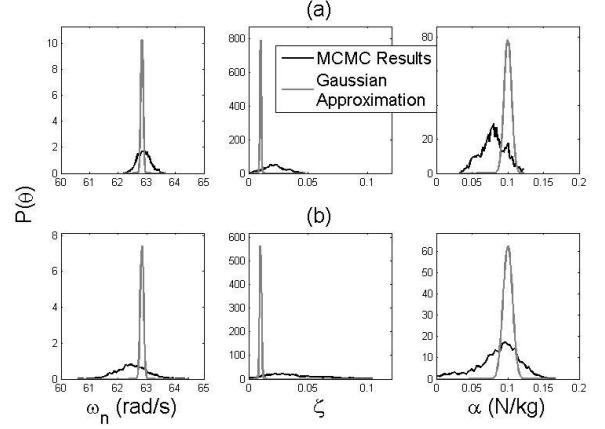

Figure 3: MCMC results (black lines) and Gaussian approximations of the posterior distribution (grey lines) for the parameters of the system shown in equation (14). Plots (a) and (b) refer to the excitations (a) and (b) on Figure 2.

4.2 SDOF System: Coulomb Nonlinearity with Sinusoidal Excitation

In this section the natural frequency and damping ratio of the system was treated as known such that the magnitude of Coulomb damping was the only parameter to be found. Additionally, a sinusoidal excitation at the resonance frequency of the system was employed:

$$\ddot{y} = Y_e \sin(20\pi t), \qquad (15)$$

where the amplitude (Y_e) was left as a parameter to optimise. The resulting entropy estimates, as a function of Y_e, are shown in Figure 4. It is clear that, as one would expect, low amplitude excitations (such as point (a) on Figure 4) are beneficial when attempting to identify Coulomb-type nonlinearities. However it is also important to observe that, in the low amplitude regions, the response of the system is heavily corrupted by measurement noise – this has greatly increased the variance of the entropy estimates. It is also interesting to note that, above the amplitude indicated by point (b), the entropy is relatively unaffected by the amplitude of the excitation. This may be because the system is dominated by the linear response.
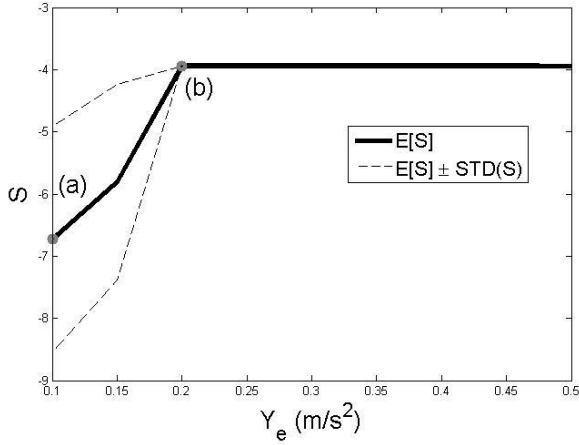
Figure 4: Variation of the Shannon entropy with the frequency of a sine-wave excitation for the system described in equation (14).

Figure 5: Variation of the Shannon entropy with the frequency of a sine-wave excitation for the system described in equation (15).

4.3 SDOF System: Coulomb and Duffing-type Nonlinearities and Sinusoidal Excitation

In this example a hardening Duffing-type spring was added to the system such that its equation of motion was now:

$$\ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x + \alpha\text{sgn}(\dot{x}) + \beta x^3 = -\ddot{y} \quad (15)$$

where β controls the magnitude of the nonlinear spring. Throughout the following analysis, β was set equal to 1×10^5. Once again, a sinusoidal excitation was employed at a frequency of 10 Hz (and whose amplitude was to be optimised). The natural frequency and damping ratio of the system were considered known such that the parameters to be identified were α and β.

In this case, rather than tracking the entropy of the entire posterior distribution, it was assumed that α and β were uncorrelated such that

$$P(\alpha, \beta | \mathcal{M}, \mathcal{D}) = P(\alpha | \mathcal{M}, \mathcal{D}) P(\beta | \mathcal{M}, \mathcal{D}). \quad (16)$$

This allowed the entropy of the individual elements of the posterior (denoted S_α and S_β) to be tracked separately (Figure 5). As with the previous example, the entropy of $P(\alpha | \mathcal{M}, \mathcal{D})$ is lowest for low excitations but, above a certain level, is relatively insensitive to the excitation amplitude. As one would expect, the entropy $P(\beta | \mathcal{M}, \mathcal{D})$ appears to be a strictly decreasing function of the excitation amplitude. The entropy of $P(\beta | \mathcal{M}, \mathcal{D})$ at lower levels of excitation is not shown on Figure 5 as, because of the large amounts of uncertainty involved, numerical overflow issues were encountered.
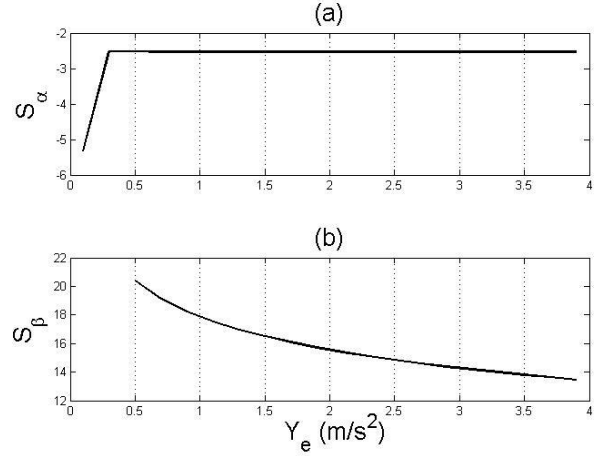
Figure 5 presents an interesting conclusion. It suggests that, if one is wishing to simultaneously identify a Coulomb and Duffing-type nonlinearity using this type of excitation and $S_\alpha \approx 2.5$ can be considered a reasonable level of uncertainty with regard to the Coulomb damping estimate, then a large amplitude excitation is preferable. This is because the confidence one has in the magnitude of α is insensitive to the excitation amplitude and so, therefore, one should simply choose the excitation which maximises one's confidence in β.

To validate this conclusion two sets of training data were created using excitation amplitudes of $Y_e = 1$ and $Y_e = 4$ m/s^2 respectively. MCMC simulations were then used to generate samples from the resulting posterior distributions. Figure 6 confirms that the use of a higher amplitude excitation has greatly increased one's confidence in the value of β, while it has had relatively little influence on one's confidence in the value of α.
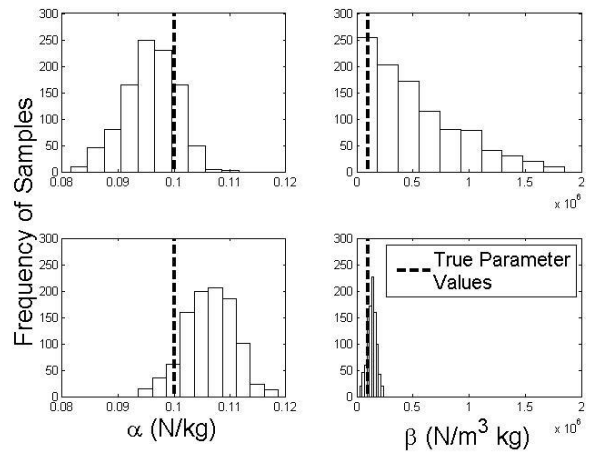

Figure 6: MCMC samples from the posterior shown in equation (16) where training data was generated using an amplitude of (a) 1 and (b) 4 m/s^2. Dashed black lines represent the true parameter values.

5   DISCUSSION AND FUTURE WORK

The authors intend to extend the preliminary study detailed in this paper in a variety of ways. For example, much of the analysis relies on the assumption that one already has a reasonable estimate of the optimum parameter vector $\boldsymbol{\theta}_0$. This may seem to be a somewhat circular argument as the overall aim of the proposed methodology is to design an experiment which allows one to obtain accurate estimates of $\boldsymbol{\theta}_0$. However, the ultimate goal of this work is to provide much more generic results. For example, in the case where one wishes to infer whether there is a combination of Coulomb and Duffing-type nonlinearities (of any magnitude) present in a system, it is hoped that the type of analysis detailed in this paper will be able to generate statements such as 'a large amplitude excitation at the resonance frequency of the structure is required to test for these nonlinearities'. Additionally, the authors wish to extend the method such that it can aid in selecting the bandwidth of random excitations or the frequency and phase of multi-sine excitations.

Studying Figure 6, it is clear that the most probable parameters predicted by the MCMC simulations are biased. This highlights that, through using the method outlined in this paper, one defines the optimum excitation as being that which minimises parameter uncertainty (rather than that which minimises the bias in the most-probable parameter estimates). It would be interesting to see if, in future work, the phrase 'optimum excitation' can be defined as that which produces parameter estimates whose bias, as well as uncertainty, has been minimised.

Ultimately, before this work can be pursued further, the aim is to investigate how the proposed methodology can aid the system identification of real, laboratory-based systems.

6   CONCLUSIONS

Broadly speaking, this paper is concerned with the Bayesian system identification of dynamical systems using experimentally-obtained training data. An investigation is performed to find the optimum form of excitation that should be used during the generation of training data. This is achieved by using the Shannon entropy as an information measure such that, by estimating the entropy of the posterior parameter distribution, the information content of different sets of training data can be analysed. Using a series of simulations it is shown that such an approach can allow one to design experiments which, ultimately, will result in parameter estimates with minimal uncertainty.

REFERENCES

[1]   Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*, 1087.

[2]   Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, *195*(2), 216-222.

[3]   Hukushima, K., & Nemoto, K. (1996). Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, *65*(6), 1604-1608.

[4]   Marinari, E., & Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *EPL (Europhysics Letters)*, *19*(6), 451.

[5]   Ching, J., & Chen, Y. C. (2007). Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging. *Journal of engineering mechanics*, *133*(7), 816-832.

[6]   Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711-732.

[7]   Green, P. L. (2014). Bayesian system identification of MDOF nonlinear systems using highly informative training data. *Proceedings of IMAC XXXII, conference and exposition on structural dynamics – Orlando, Florida, 3-6 February 2014.*

[8]   Green, P.L. (2014). Bayesian system identification of dynamical systems using highly informative training data. *Mechanical systems and signal processing* (under review).

[9]   Simoen, E., Papadimitriou, C., & Lombaert, G. (2013). On prediction error correlation in Bayesian model updating. *Journal of Sound and Vibration*.