**Proceedings Paper:**
Green, P.L. (2014) Bayesian system identification of MDOF nonlinear systems using highly informative training data. In: Allemang, R., (ed.) Topics in Modal Analysis II, Volume 8 : Proceedings of the 32nd IMAC, A Conference and Exposition on Structural Dynamics, 2014. 32nd IMAC, A Conference and Exposition on Structural Dynamics, 03-06 Feb 2014, Orlando, Florida USA. Conference Proceedings of the Society for Experimental Mechanics Series, 8. Springer, pp. 257-265. ISBN: 9783319047737. ISSN: 2191-5644. EISSN: 2191-5652.

https://doi.org/10.1007/978-3-319-04774-4_25

# Bayesian System Identification of MDOF Nonlinear Systems using Highly Informative Training Data

Dr. P.L.Green

University of Sheffield
Department of Mechanical Engineering, Sir Frederick Mappin Building, Mappin Street, S1 3JD
email: p.l.green@sheffield.ac.uk

October 25, 2013

## Abstract

The aim of this paper is to utilise the concept of 'highly informative training data' such that, using Markov chain Monte Carlo (MCMC) methods, one can apply Bayesian system identification to multi-degree-of-freedom nonlinear systems with relatively little computational cost. Specifically, the Shannon entropy is used as a measure of information content such that, by analysing the information content of the posterior parameter distribution, one is able to select and utilise a relatively small but highly informative set of training data (thus reducing the cost of running MCMC).

**Key words:** System Identification, Bayesian Inference, Markov chain Monte Carlo, Shannon Entropy, Nonlinear Dynamics

## 1 Introduction

This paper is concerned with the system identification of nonlinear dynamical systems using physics-based models. In this context the overall aim of system identification is to infer, using experimental data, a reliable and robust physical-law based model of a real system. This requires the selection of an appropriate model structure as well as estimation of the parameters within that model. This is a procedure which, as a result of measurement noise and modelling uncertainties, is best approached using probability logic. Adopting a Bayesian framework allows one to take a probabilistic approach to both parameter estimation and model selection.

Using Bayes' Theorem, one can express the parameter estimation and model selection levels of inference as

$$P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = \frac{P(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})P(\boldsymbol{\theta}|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})} \tag{1}$$

and

$$P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D})} \tag{2}$$

respectively. With regards to equation (1), $P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$ is termed the 'posterior distribution'. The posterior is a probability density function (PDF) which represents the probability that the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{N_\theta}$ is 'true' given some experimentally obtained training data $\mathcal{D}$ and a chosen model structure $\mathcal{M}$. It is proportional to the product of the 'likelihood' $P(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})$, and the 'prior' $P(\boldsymbol{\theta}|\mathcal{M})$. The prior is a PDF which represents one's prior knowledge of the parameters before the experimental data was realised. The likelihood is a PDF which, given a model structure $\mathcal{M}$ with parameters $\boldsymbol{\theta}$, represents the probability that

the data $\mathcal{D}$ was realised. Consequently then, defining the likelihood involves the selection of a noise model which represents the errors due to the measurement and modelling processes. The denominator of equation (1) is termed the 'evidence' - this is a constant given by

$$P(\mathcal{D}|\mathcal{M}) = \int ... \int P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) d\theta_1 ... d\theta_{N_\theta} \tag{3}$$

thus ensuring that the posterior will integrate to unity. The evidence term also appears in the numerator of equation (2) such that, having successfully evaluated equation (1) for model structure $\mathcal{M}$, the probability that that model is a good replication of the physics of the real system (relative to other competing model structures) can then be evaluated.

Unfortunately, it is often the case that the high-dimensionality and complex geometry of the posterior distribution makes evaluation of equation (3) difficult. With regards to parameter estimation, this problem can be overcome through the use of Markov chain Monte Carlo (MCMC) methods (such as the well-known Metropolis [1] and Hamiltonian Monte Carlo [2] algorithms) which allow one to generate samples from the posterior PDF in equation (1) without having to evaluate the evidence term. Indeed, there are also MCMC methods which allow one to tackle the issue of model selection - the Transitional MCMC algorithm proposed in [3] can be used to estimate the evidence term in equation (1) while the Reversible Jump MCMC algorithm [4] is capable of generating samples from a PDF of varying dimension (thus allowing one to simultaneously evaluate a set of competing model structures of varying levels of complexity).

While undoubtedly useful, the number of model runs required tend to make MCMC algorithms computationally expensive (thus restricting their use to relatively small models). The aim of this paper is to investigate whether this cost can be reduced through the use of small but highly informative sets of training data. To that end, the Shannon entropy is used to quantify the information content of a set of training data. It is shown that, in the Bayesian parameter estimation of a MDOF nonlinear dynamical system, this can reduce the cost of running MCMC algorithms.

## 2 Bayesian Framework

In this section the Bayesian framework for the parameter estimation of a $N_D$ DOF dynamical system is described. While this is not new, it will help to establish the notation used throughout this work.

This paper is concerned with systems whose state-space equations of motion are of the form:

$$\dot{\boldsymbol{x}} = \boldsymbol{y} \tag{4}$$

and

$$\dot{\boldsymbol{y}} = \boldsymbol{M}^{-1}(\boldsymbol{C}\boldsymbol{y} + \boldsymbol{K}\boldsymbol{x} + \boldsymbol{\eta} + \boldsymbol{f}) \tag{5}$$

where $\boldsymbol{x} \in \mathbb{R}^{N_D}$ and $\boldsymbol{y} \in \mathbb{R}^{N_D}$ represent the displacements and velocities of each DOF, $\boldsymbol{M}$ is the mass matrix, $\boldsymbol{C}$ is the linear damping matrix, $\boldsymbol{K}$ is the linear stiffness matrix, $\boldsymbol{\eta}$ is a vector which contains the nonlinear terms and $\boldsymbol{f}$ is a vector which describes the excitation force being delivered to each mass. The training data $\mathcal{D}$ consists of $N$ points of recorded time history from the force vector $\boldsymbol{f}$ as well as the resulting displacement measurements ($N$ from each DOF). For the $i$th degree of freedom, the measured and simulated displacement time histories will be written as

$$x^{(i)}_{1:N} = \{x^{(i)}_1, x^{(i)}_2, ..., x^{(i)}_N\} \tag{6}$$

and

$$\hat{x}^{(i)}_{1:N} = \{\hat{x}^{(i)}_1, \hat{x}^{(i)}_2, ..., \hat{x}^{(i)}_N\} \tag{7}$$

respectively. Drawing on the central limit theorem, it is assumed that each measured data point is corrupted by Gaussian noise of variance $\sigma^2$ such that, after some manipulation, the likelihood can be written as:

$$P(\mathcal{D}|\boldsymbol{\theta},\mathcal{M}) = (2\pi\sigma^2)^{-NN_D/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N_D} J_i(\boldsymbol{\theta})\right) \tag{8}$$

where

$$J_i(\boldsymbol{\theta}) = [x^{(i)}_{1:N} - \hat{x}^{(i)}_{1:N}][x^{(i)}_{1:N} - \hat{x}^{(i)}_{1:N}]^T. \tag{9}$$

Throughout the following analysis $\sigma$ is treated as an additional unknown parameter for which probabilistic estimates can also be realised.

From now on, for the sake of simplicity, the likelihood will be written as:

$$P(\mathcal{D}|\boldsymbol{\theta},\mathcal{M}) = \frac{1}{Z_L}\exp(\boldsymbol{J}(\boldsymbol{\theta})) \tag{10}$$

where

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N_D} J_i(\boldsymbol{\theta}). \tag{11}$$

and $Z_L$ is the likelihood normalisation constant.

# 3 Informative Training Data

With aim of being able to identify training data which is 'highly informative' with regards to one's parameter estimates, the Shannon entropy is used as measure of information content throughout this paper. A similar idea was also developed by MacKay in [5] (although this was within the context of machine learning).

The Shannon entropy of the posterior distribution is defined as

$$S = -\int P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})\ln(P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}))d\boldsymbol{\theta} \tag{12}$$

which cannot be evaluated as the geometry of the posterior is unknown. To proceed, a Taylor series expansion about the most probable parameter estimates ($\boldsymbol{\theta}_0$) is used to approximate the posterior as being Gaussian. Before this can be accomplished one must first define the prior distribution - throughout this paper Gaussian priors of the form

$$P(\boldsymbol{\theta}|\mathcal{M}) = \frac{1}{Z_P}\exp\left(-\frac{1}{2}[\boldsymbol{\theta}-\boldsymbol{\theta}_0^-]\boldsymbol{B}[\boldsymbol{\theta}-\boldsymbol{\theta}_0^-]^T\right) \tag{13}$$

are used. $Z_P$ is a normalising constant, $\boldsymbol{\theta}_0^-$ represent the mean of one's prior parameter estimates and $\boldsymbol{B}$ is a diagonal, user-defined covariance matrix. Recalling the definition of the likelihood (equation (10)), the posterior distribution can now be written as

$$P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = \frac{1}{Z_L Z_P P(\mathcal{D}|\mathcal{M})}\exp(G(\boldsymbol{\theta})) \tag{14}$$

where

$$G(\boldsymbol{\theta}) = \boldsymbol{J}(\boldsymbol{\theta}) - \frac{1}{2}[\boldsymbol{\theta}-\boldsymbol{\theta}_0^-]\boldsymbol{B}[\boldsymbol{\theta}-\boldsymbol{\theta}_0^-]^T \tag{15}$$

($\boldsymbol{J}(\boldsymbol{\theta})$) was defined in equation (11)). Expanding $G(\boldsymbol{\theta})$ about $\boldsymbol{\theta}_0$ using the Taylor series (as one does when using Laplace's method) then, after some manipulation, one can approximate the posterior as being Gaussian:

$$P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})^* = \frac{1}{Z^*}\exp\left(-\frac{1}{2}[\boldsymbol{\theta}-\boldsymbol{\theta}_0]\boldsymbol{A}[\boldsymbol{\theta}-\boldsymbol{\theta}_0]^T\right) \tag{16}$$

where the asterisk is used to represent the approximation and $Z^*$ is the posterior normalising constant. The elements of the matrix $\boldsymbol{A}$ are given by

$$\boldsymbol{A}_{i,j} = \left. \frac{\partial^2 G(\boldsymbol{\theta})}{\partial(\theta_i)^2} \right|_{\boldsymbol{\theta}_0}. \tag{17}$$

which can be approximated using finite difference methods. It is then relatively simple to shown that the Shannon entropy of $P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})^*$ is given by

$$S = \ln(Z^*) + \frac{N_\theta}{2}. \tag{18}$$

Assuming that the matrix $\boldsymbol{A}$ is diagonal (which is equivalent to ignoring parameter correlations) and ignoring terms which do not change as a function of the training data (as the aim is to analyse the information content as a function of the training data) one can write equation (18) in a simpler form:

$$S = \sum_{i=1}^{N_\theta} \ln\left(\frac{1}{A_{i,i}}\right). \tag{19}$$

It is important to note that the Shannon entropy also acts as a measure of uncertainty - a highly informative data point will cause a *decrease* in the entropy. It is also interesting to note that the inverse of $\boldsymbol{A}$ is the covariance matrix of $P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})^*$. Consequently then, by monitoring the Shannon entorpy, one is actually monitoring the diagonal elements of the covariance matrix of $P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})^*$ as a function of the training data. This is equivalent to monitoring the width of the Gaussian - an informative data point is one which causes are large reduction in the width of the posterior.

# 4 Potential Issues

## 4.1 Most Probable Parameter Estimates

The method presented in Section 3 relies on one having a reasonable estimate of $\boldsymbol{\theta}_0$. Consequently, the accuracy of the information estimates will depend on the accuracy of one's prior knowledge - this is something which seems intrinsically Bayesian. Throughout this work, with the aim of improving one's prior estimates of $\boldsymbol{\theta}_0$ the Data Annealing (DA) algorithm [6] has been utilised. Essentially, the DA algorithm is similar to the well-known Simulated Annealing algorithm except that, to save computational cost, the annealing procedure is achieved through the gradual introduction of training data into the likelihood. Additionally, in a similar fashion to 'Fast Simulated Annealing' [7], the DA algorithm also utilises a proposal distribution with relatively heavy tails to reduce the changes of becoming stuck in 'local traps' (regions of high probability mass which are not globally optimum). As a result, DA can be used to improve one's information estimates with relatively little computational cost.

The assumption that there is a single set of optimum parameters is one of the potential issues with the method outlined in this paper. In reality there may be several sets of, or in fact a continuous set of, optimum parameters (this are referred to as being 'locally identifiable' and 'unidentifiable' cases in [8]). The danger with using small sets of training data is that one may inadvertently provoke a situation where no one single optimum parameter vector exists. As a first step to addressing this issue, the DA algorithm can be run multiple times so that multiple estimates of the optimum parameter vector are established - it should be ensured that these estimates are reasonably repeatable. Secondly, once a highly informative subset of the training data has been selected, one should utilise an MCMC algorithm which, while being more expensive than 'traditional' MCMC, is capable of sampling from PDFs with complex geometries (in fact, one could argue that these algorithms should *always* be employed over more traditional methods). Such algorithms include Adaptive Metropolis Hastings (AMH) [9], Transitional Markov chain Monte Carlo (TMCMC) [3] and Asymptotically Independent Markov Sampling (AIMS) [10]. TMCMC is utilised throughout the following analysis as it is both well-established and works well when higher-dimensional problems are considered.

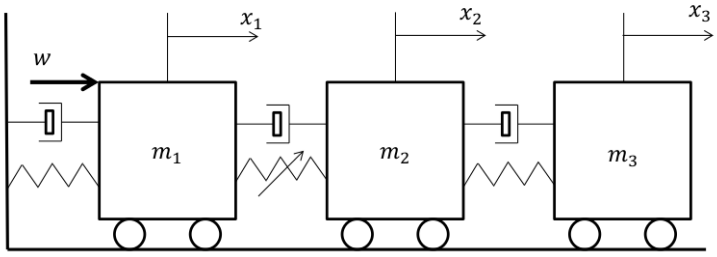


**Fig. 1: 3 DOF nonlinear system**

Table 1: Mean and standard deviation of Gaussian priors.

| Parameter | Prior Mean | Prior Standard Deviation |
|---|---|---|
| $k_1$ | 50 | 10 |
| $k_2$ | 50 | 10 |
| $k_3$ | 50 | 10 |
| $c_1$ | 0.05 | 0.01 |
| $c_2$ | 0.05 | 0.01 |
| $c_3$ | 0.05 | 0.01 |
| $k^*$ | 1000 | 100 |
| $\sigma$ | 0.05 | 0.02 |

# 5 Nonlinear System

To demonstrate the concept of highly informative training data, the parameter estimation of a nonlinear 3DOF system will be analysed (Figure 1). From left-to-right, the masses are connected by springs with linear stiffness coefficients $k_1$, $k_2$, $k_3$, and dampers with linear damping coefficients $c_1$, $c_2$, $c_3$. The second spring also has a cubic stiffening component with nonlinear stiffness coefficient $k^*$. The first mass is excited with a Gaussian white noise signal ($w$ in Figure 1). Consequently, the relevant system matrices are:

$$\boldsymbol{K} = \begin{bmatrix} -k_1 - k_2 & k_2 & 0 \\ k_2 & -k_2 - k_3 & k_3 \\ 0 & k_3 & -k_3 \end{bmatrix} \tag{20}$$

$$\boldsymbol{C} = \begin{bmatrix} -c_1 - c_2 & c_2 & 0 \\ c_2 & -c_2 - c_3 & c_3 \\ 0 & c_3 & -c_3 \end{bmatrix} \tag{21}$$

$$\boldsymbol{\eta} = \begin{Bmatrix} -k^*(x_1 - x_2)^3 \\ -k^*(x_2 - x_1)^3 \\ 0 \end{Bmatrix} \tag{22}$$

$$\boldsymbol{f} = \begin{Bmatrix} w \\ 0 \\ 0 \end{Bmatrix}. \tag{23}$$

The parameter values were generated from Gaussian distributions (the moments of which are shown in Table 1). These distributions were then used as priors throughout the following analysis. Care was taken to ensure that the author did not know the true parameter values - only the size of the prior was known. The 'full' set of output data consisted of 1000 displacement measurements (from each DOF). The next section details the selection of a highly informative subset of this data.

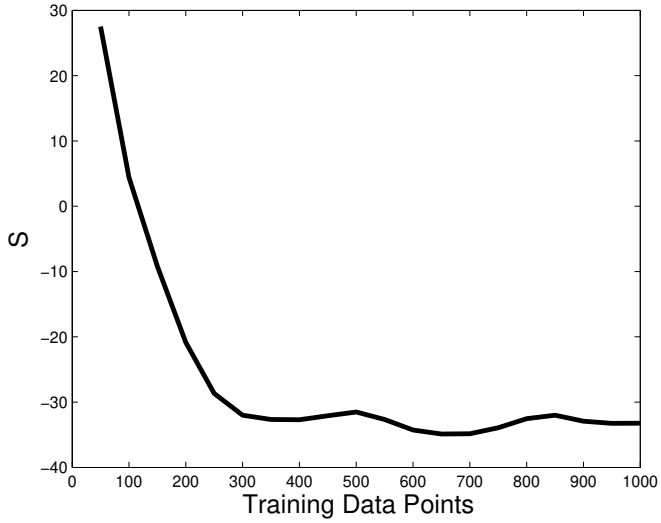


**Fig. 2: Shannon Entropy as a function of the number of points in the training data**

Table 2: True parameter estimates.

| Parameter | True Value |
|---|---|
| $k_1$ | 40.58 |
| $k_2$ | 54.30 |
| $k_3$ | 49.70 |
| $c_1$ | 0.054 |
| $c_2$ | 0.068 |
| $c_3$ | 0.040 |
| $k^*$ | $1.1 \times 10^3$ |
| $\sigma$ | 0.05 |

# 6 Results

Five Data Annealing runs were carried out - this took around 6 minutes to compute. The estimates of the most probable parameter estimates were found to be reasonably consistent. The Shannon entorpy of the training data was then calculated (every 50 points were analysed). This process also took roughly 6 minutes.

The Shannon entropy is plotted as a function of the number of points in the training data in Figure 2. According to the figure, the majority of the learning is achieved using the first 300 data points - thus leaving the remaining 700 points relatively uninformative.

To test this hypothesis, the TMCMC algorithm was used to generate samples from the posterior for varying amounts of training data. The resulting prior and posterior samples (TMCMC is initiated with samples from the prior) are shown respectively in Figures 3, 4, 5 and 6 for the cases where 100, 200, 300 and 1000 points of training data were used. It is clear that, through using 300 data points instead of 100, a significant amount of information has been gained. However, by comparing Figure 5 with Figure 6, it is clear that little benefit can be gained through using the additional points 700. This is as predicted in Figure 2. It is also interesting to note that the posterior has multiple modes for the 100 point case but that, through the use of more training data, it appears to be uni-modal for all the other cases.

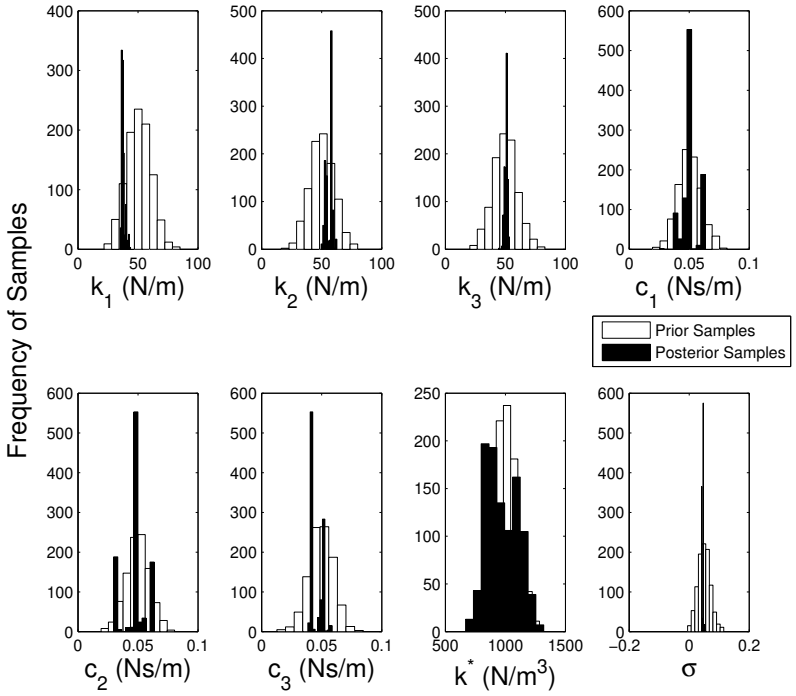


**Fig. 3: TMCMC results using 100 points of training data**

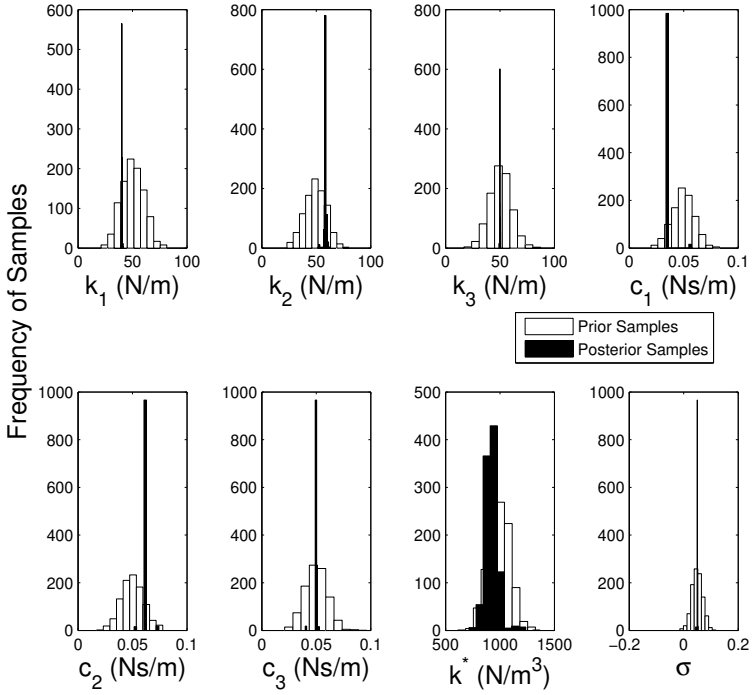


**Fig. 4: TMCMC results using 200 points of training data**

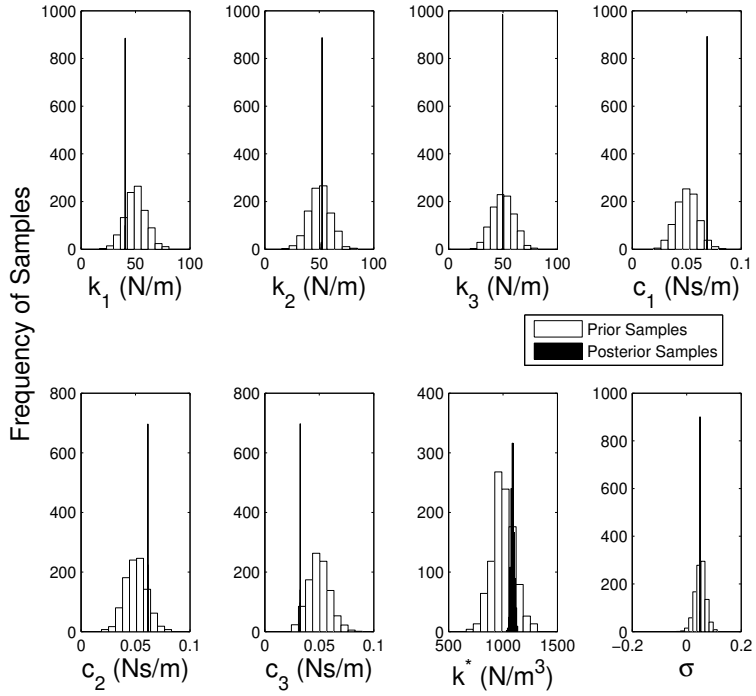


**Fig. 5: TMCMC results using 300 points of training data**

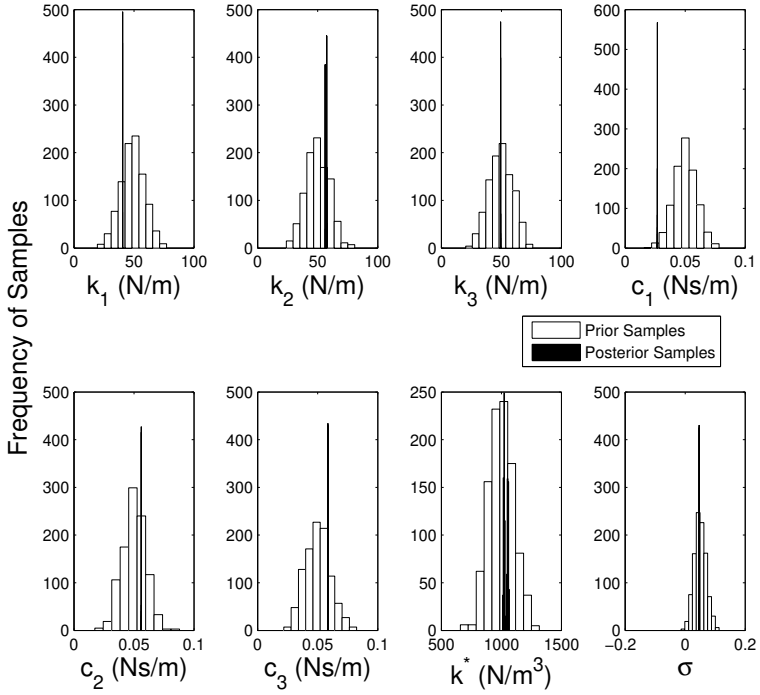


**Fig. 6: TMCMC results using 1000 points of training data**

# 7 Conclusions

This paper was concerned with the Bayesian parameter estimation of nonlinear dynamical systems through the use of Markov chain Monte Carlo (MCMC) methods. Using the Shannon entropy as an information measure it was shown that, by electing to use small amounts of 'highly informative' training data, the computational cost of running MCMC algorithms can be greatly reduced. This was then demonstrated with regards to the probabilistic parameter estimation of a MDOF nonlinear system. It was also shown that, through the use of small amounts of training data, one may induce multi-modal posterior distributions. This was addressed though the use of Transitional MCMC which is able to sample from posterior distributions with complex geometries.

# References


[1] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21:1087, 1953.

[2] S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

[3] J. Ching and Y.C. Chen. Transitional Markov Chain Monte Carlo Method for Bayesian Model Updating, Model Class Selection, and Model Averaging. *Journal of Engineering Mechanics*, 133(7):816–832, 2007.

[4] P.J. Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711–732, 1995.

[5] D.J.C. MacKay. Information-Based Objective Functions for Active Data Selection. *Neural computation*, 4(4):590–604, 1992.

[6] P.L. Green. Bayesian System Identification of a Nonlinear Dynamical System using a Novel Variant of Simulated Annealing (**under review**). *Mechanical Systems and Signal Processing*, 2013.

[7] H. Szu and R. Hartley. Fast Simulated Annealing. *Physics letters A*, 122(3):157–162, 1987.

[8] J.L. Beck and L.S. Katafygiotis. Updating Models and their Uncertainties. i: Bayesian Statistical Framework. *Journal of Engineering Mechanics*, 124(4):455–461, 1998.

[9] J.L. Beck and S.K. Au. Bayesian Updating of Structural Models and Reliability using Markov Chain Monte Carlo Simulation. *Journal of Engineering Mechanics*, 128(4):380–391, 2002.

[10] J.L. Beck and K.M. Zuev. Asymptotically Independent Markov Sampling: a New Markov Chain Monte Carlo Scheme for Bayesian Inference. *International Journal for Uncertainty Quantification*, 3(5), 2013.