Available online at www.eccomasproceedia.org
Eccomas Proceedia UNCECOMP (2017) 176-186

ECCOMAS
Proceedia

UNCECOMP 2017
2nd ECCOMAS Thematic Conference on
Uncertainty Quantification in Computational Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Rhodes Island, Greece, 15–17 June 2017

ZERO-VARIANCE SIMULATED ANNEALING FOR BAYESIAN SYSTEM IDENTIFICATION

Peter L. Green[1]

[1] Institute for Risk and Uncertainty, School of Engineering, University of Liverpool, Liverpool, L69 7ZF, United Kingdom
e-mail: p.l.green@liverpool.ac.uk

Keywords: Markov chain Monte Carlo, Zero Variance, Hamiltonian Monte Carlo, Simulated Annealing

Abstract. *Markov chain Monte Carlo (MCMC) algorithms are a set of methods which allow samples to be generated from generic probability distributions. They have been used to aid the simulation of rare events, the Bayesian system identification of systems which are nonlinear and/or are approached using a Bayesian hierarchical structure and the training of a variety of machine learning algorithms (for example). The current paper discusses the 'Zero-Variance method', which can be used to greatly reduce the sample variance of quantities that are estimated using Monte Carlo methods. The ability of this approach to increase the efficiency of gradient based MCMC methods is illustrated. Finally, a Zero-Variance version of the well-known simulated annealing algorithm is employed. The algorithm is demonstrated on the Bayesian system identification of a nonlinear dynamical system.*

Peer-review under responsibility of the organizing committee of UNCECOMP 2017.
doi: 10.7712/120217.5361.17287

1 INTRODUCTION

Over recent years Bayesian approaches to system identification have been adopted across a wide range of applications within structural dynamics. Markov chain Monte Carlo (MCMC) algorithms, which can be used to generate samples from generic probability distributions, often form a fundamental component of these methods.

Recent works, for example, have focused on the application of hierarchical Bayesian frameworks to aid the identification of structures subject to changes in ambient temperature and excitation amplitude [1], damage detection from noisy and/or incomplete modal data [2] and the multilevel identification of sets of nominally identical systems [3]. These works either make use of MCMC [2] [3] or cite it as an avenue for future work [1]. MCMC can also be applied in various other contexts, such as the efficient simulation of rare events (subset simulation) [4].

For the current paper, it is sufficient to consider the situation where one wishes to generate samples from the posterior distribution

$$p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{N_\theta}$ is a vector of parameters which are to be inferred from a set of observations, D. It is assumed here that closed-form solutions for the posterior are unavailable (such that it is necessary to generate samples from $p(\boldsymbol{\theta}|D)$).

2 ZERO-VARIANCE PRINCIPLE

Say $g(\boldsymbol{\theta})$ is a quantity whose expected value, with respect to a target distribution $\pi(\boldsymbol{\theta})$, is of interest. In the context of this paper $\boldsymbol{\theta}$ is a vector of a model's parameters and the expected value of $g(\boldsymbol{\theta})$ is typically estimated using

$$\mathrm{E}_\pi[g(\boldsymbol{\theta})] \approx \frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{\theta}^i) \tag{2}$$

where $\{\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^N\}$ are samples from the posterior which have been realised using MCMC. This estimate is, of course, subject to statistical error which will reduce if larger N is used (or if correlations between successive MCMC samples is reduced).

The zero-variance (ZV) principle [5] suggests that $g(\boldsymbol{\theta})$ should be transformed into a different function, $\tilde{g}(\boldsymbol{\theta})$, whose expected value is still equal to $\mathrm{E}_\pi[g(\boldsymbol{\theta})]$ but whose variance is reduced. $\tilde{g}(\boldsymbol{\theta})$ can then be used to estimate the quantity of interest with less statistical error, relative to if $g(\boldsymbol{\theta})$ had been used.

This method gets its name from the fact that, ideally, the transformation would lead to $\tilde{g}(\boldsymbol{\theta})$ having zero variance (thus eliminating the statistical error). In practice, however, this is usually impossible to achieve and approximations to this ideal transformation must be employed. It should be noted that a very general treatment of the zero-variance principle and the nature of the transformations that can be utilised are outlined in [5] but that this is somewhat beyond the current paper, which aims to investigate the applicability of the ZV method to Bayesian system identification problems within structural dynamics. The current work instead focuses on the transformations that are described in [6] [7] – the reader simply needs to be

aware that these are specific examples of the wide variety of transformations that are potentially available.

2.1 First order transformation

At this point it is convenient to denote the negative score of the target distribution as $\boldsymbol{z}$:

$$\boldsymbol{z}(\boldsymbol{\theta}) = -\frac{\partial}{\partial \boldsymbol{\theta}} \log \pi \tag{3}$$

Noting that the expected score is zero, the following transformation is defined:

$$\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \boldsymbol{\alpha}^T \boldsymbol{z}(\boldsymbol{\theta}) \tag{4}$$

where $\boldsymbol{\alpha}$ is a vector of parameters which require identification. The values of $\boldsymbol{\alpha}$ that minimise $\mathrm{Var}[\tilde{g}]$ are

$$\boldsymbol{\alpha} = -(\mathrm{Var}[\boldsymbol{z}])^{-1}\mathrm{Cov}(g, \boldsymbol{z}) \tag{5}$$

where

$$\mathrm{Cov}[g, \boldsymbol{z}] = \mathrm{E}[g\boldsymbol{z}] - \mathrm{E}[g]\mathrm{E}[\boldsymbol{z}] \tag{6}$$

(proved in the appendix). Consequently then, samples from the target must be used to realise a Monte Carlo estimate of $\boldsymbol{\alpha}$ *before* the transformation can be applied. This may appear to be a somewhat circular argument, as the resulting estimates of $\boldsymbol{\alpha}$ will still be subject to statistical error. For now, it should simply be observed that, for the examples investigated in the current paper, the results were very insensitive to the statistical variation of $\boldsymbol{\alpha}$. A more formal investigation of this property is a topic of future work.

With regard to the reduction in variance that can be achieved, it is straightforward to show that

$$\mathrm{Var}[\tilde{g}] = \mathrm{Var}[g] + 2\boldsymbol{\alpha}^T\mathrm{Cov}[g, \boldsymbol{z}] + \boldsymbol{\alpha}^T\mathrm{Var}[\boldsymbol{z}]\boldsymbol{\alpha} \tag{7}$$

such that, after substituting in the optimum values of $\boldsymbol{\alpha}$ and rearranging, the reduction in variance is found to be

$$\mathrm{Var}[\tilde{g}] - \mathrm{Var}[g] = -\mathrm{Cov}[g, \boldsymbol{z}]^T(\mathrm{Var}[\boldsymbol{z}])^{-1}\mathrm{Cov}[g, \boldsymbol{z}] \tag{8}$$

This confirms that $\mathrm{Var}[\tilde{g}] \leq \mathrm{Var}[g]$, as desired (as $(\mathrm{Var}[\boldsymbol{z}])^{-1}$ is positive definite).

2.2 Second order transformation

A potential second order transformation, defined in [6] [7], is given by

$$\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \boldsymbol{\alpha}^T \boldsymbol{z}(\boldsymbol{\theta}) + \boldsymbol{\theta}^T \boldsymbol{B}\boldsymbol{z}(\boldsymbol{\theta}) + \boldsymbol{C} \tag{9}$$

where $\boldsymbol{B}$ is symmetric and $\boldsymbol{C}$ must be chosen such that $\mathrm{E}_\pi[\tilde{g}(\boldsymbol{\theta})] = \mathrm{E}_\pi[g(\boldsymbol{\theta})]$. To define $\boldsymbol{C}$ it is first noted that

$$\boldsymbol{\theta}^T \boldsymbol{B}\boldsymbol{z}(\boldsymbol{\theta}) \equiv \sum_i B_{ii}\theta_i z_i + \sum_{i \neq j} B_{ij}\theta_i z_j \tag{10}$$

(where summations are taken up to N_θ and the notation $z_i \equiv z(\theta_i)$ has been adopted). As

$$\mathrm{E}_\pi[\theta_i z_i] = 1 \tag{11}$$

$$\mathrm{E}_\pi\left[\theta_i z_j\right] = 0, \quad i \neq j \tag{12}$$

then, for the estimator to be unbiased, it follows that $\boldsymbol{C} = -\mathrm{Tr}(\boldsymbol{B})$. The second order transformation therefore becomes

$$\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \boldsymbol{a}^T \boldsymbol{z}(\boldsymbol{\theta}) + \boldsymbol{\theta}^T \boldsymbol{B}\boldsymbol{z}(\boldsymbol{\theta}) - \mathrm{Tr}(\boldsymbol{B}) \tag{13}$$

At this point it is convenient to rearrange all of the transformation's parameters into a single vector, such that equation (5) can be used to estimate their optimum value. This is achieved as follows:

$$\boldsymbol{\theta}^T \boldsymbol{B}\boldsymbol{z}(\boldsymbol{\theta}) - \mathrm{Tr}(\boldsymbol{B}) = \sum_i B_{ii}\theta_i z_i + 2\sum_{j>i} B_{ij}\theta_i z_j - \boldsymbol{b}^T \boldsymbol{1} \tag{14}$$

$$= \boldsymbol{b}^T \boldsymbol{u}(\boldsymbol{\theta}) + \boldsymbol{c}^T \boldsymbol{v}(\boldsymbol{\theta}) \tag{15}$$

where $\boldsymbol{b} = \mathrm{diag}(\boldsymbol{B})$,

$$\boldsymbol{u}(\boldsymbol{\theta}) = \boldsymbol{\theta} \circ \boldsymbol{z}(\boldsymbol{\theta}) - \boldsymbol{1} \tag{16}$$

($\circ$ is the Hadamard product), the $\frac{1}{2}(j-1)(j-2) + i$ th element of $\boldsymbol{c}$ is equal to B_{ij} $(j > i)$ and the $\frac{1}{2}(j-1)(j-2) + i$ th element of $\boldsymbol{v}(\boldsymbol{\theta})$ is equal to $2\theta_i z_j$. This allows one to write the second order transformation neatly as

$$\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \boldsymbol{\alpha}^T \boldsymbol{w}(\boldsymbol{\theta}) \tag{17}$$

where

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \end{pmatrix}, \qquad \boldsymbol{w}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{z}(\boldsymbol{\theta}) \\ \boldsymbol{u}(\boldsymbol{\theta}) \\ \boldsymbol{v}(\boldsymbol{\theta}) \end{pmatrix} \tag{18}$$

In a similar manner to the first order case, the reduction in variance can be shown to be

$$\mathrm{Var}[\tilde{g}] - \mathrm{Var}[g] = -\mathrm{Cov}[g, \boldsymbol{w}]^T (\mathrm{Var}[\boldsymbol{w}])^{-1} \mathrm{Cov}[g, \boldsymbol{w}] \tag{19}$$

2.3 Illustrative examples

The method will first be demonstrated on some simple examples, where samples from the target can be generated directly (without having to resort to MCMC). The interested reader may also like to consult the technical report [7] where other relatively simple examples are discussed. Note that the following only explores the first and second order expansions that were outlined in the previous section.

First, Monte Carlo samples are used to estimate the mean of the distribution $\pi(\theta) \propto \exp\left(-\frac{1}{2}\theta^2\right)$ such that $g(\theta) = \theta$ and $z(\theta) = \theta$. For the first order case, it is straightforward to show that $\mathrm{Var}[z] = 1$ and $\mathrm{Cov}[g, z] = 1$. This implies that the optimum choice of transformation parameter is $\alpha = -1$. For the second order case $v(\theta) = 0$ (as $\theta \in \mathbb{R}^1$) and so

$$\boldsymbol{w}(\theta) = \begin{pmatrix} z(\theta) \\ \theta z(\theta) - 1 \end{pmatrix} = \begin{pmatrix} \theta \\ \theta^2 - 1 \end{pmatrix} \Rightarrow \mathrm{Var}[\boldsymbol{w}] = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \tag{20}$$

Also,

$$\mathrm{Cov}[g, \boldsymbol{w}] = \mathrm{E}\begin{pmatrix} \theta^2 \\ \theta(\theta^2 - 1) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{21}$$

and so, for the second order case, the optimum transformation parameters are

$$\boldsymbol{\alpha} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \tag{22}$$

This implies that, in the current example, the second order term has no influence and so the second order ZV expansion will not be able to outperform the first order ZV expansion. This is verified by simulation in Figure 1 where 100 estimates of the mean are realised, each using 100 samples from the target. Monte Carlo estimates of $\boldsymbol{\alpha}$ are used throughout. Relative to standard Monte Carlo estimates, the reduction in the statistical error of the estimates is quite remarkable.

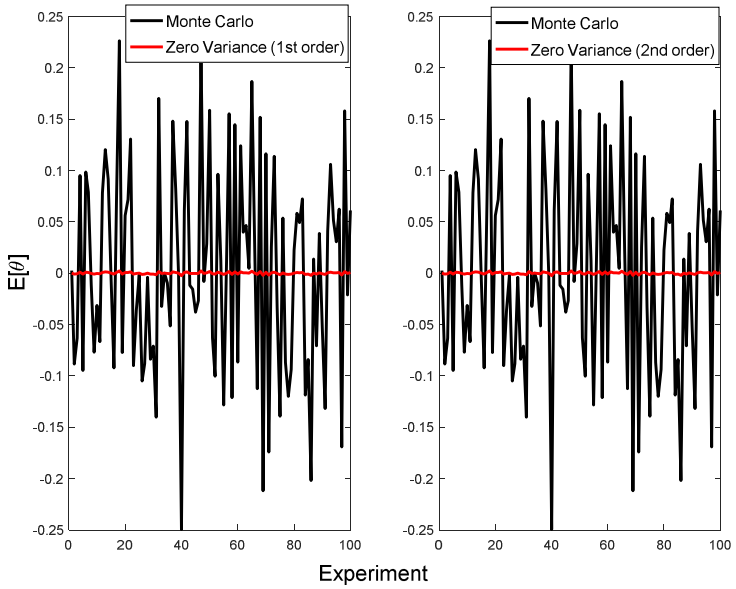

Figure 1. Estimating the mean of $p(\theta) = N(\theta; 0,1)$

The scenario where the aim is to estimate $\mathrm{E}[\theta^2]$ is now considered. For the first order case $\mathrm{Var}[z] = 1$ and $\mathrm{Cov}[g, z] = \mathrm{E}[\theta^3] = 0$ and so $\alpha = 0$ (implying that there is no advantage to using the first order expansion). For the second order expansion, $\mathrm{Var}[\boldsymbol{w}]$ is the same as in equation (20) and

$$\mathrm{Cov}[g, \boldsymbol{z}] = \mathrm{E}\left[\theta^2 \begin{pmatrix} \theta \\ \theta^2 - 1 \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \tag{23}$$

which gives optimum transformation parameters

$$\boldsymbol{\alpha} = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \tag{24}$$

These results are verified by simulation in Figure 2 where it is shown that, to achieve a reduction in the statistical error of the estimated quantity of interest, the second order transformation must be employed. Again, the reduction in statistical error is impressive.

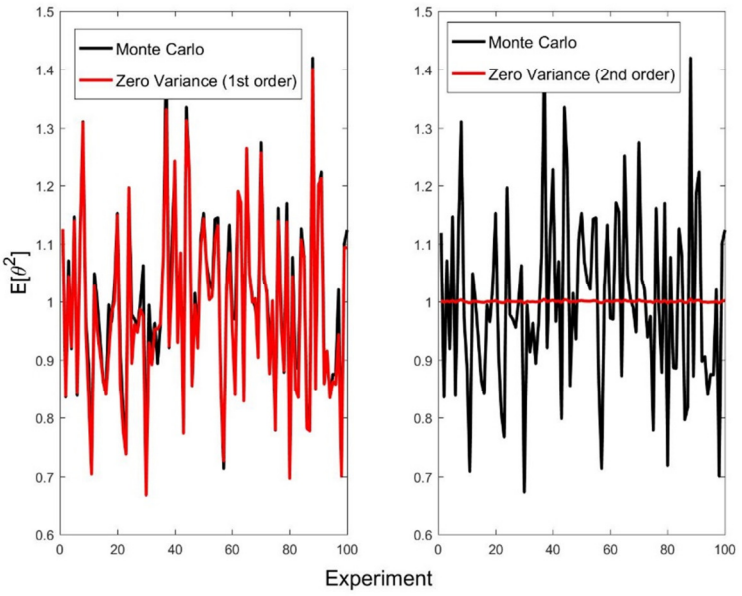

Figure 2. Estimating the variance of $p(\theta) = N(\theta; 0,1)$

3 ZERO-VARIANCE BAYEISAN SYSTEM IDENTIFICAITON USING MCMC

In the following the aim is to realise a Bayesian estimate of the nonlinear stiffness, k_3, of a Duffing oscillator:

$$\ddot{x} + c\dot{x} + kx + k_3x^3 = F(t) \tag{25}$$

where $F(t)$, in this case, was a random excitation generated from a zero-mean Gaussian with unit variance. The training data, D, consisted of the excitation time history as well as noisy observations of the system's displacement response (shown in Figure 3).

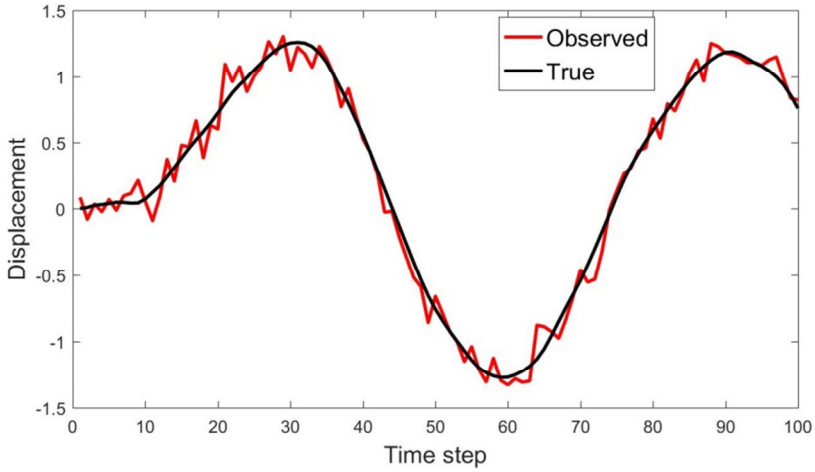

Figure 3. Training data for Duffing oscillator example.

The likelihood was created by assuming the following noise model:

$$y_i = x_i(\theta) + \varepsilon_i, \qquad \varepsilon_i = N(\varepsilon_i; 0, \sigma^2) \tag{26}$$

where $\theta = k_3$ in this case, y_i denotes the ith noisy observation of the displacement and the noise variance, σ^2, is assumed to be known. The true value of the nonlinear stiffness was $k_3 = 100$ N / m^3, while the prior was $U(k_3; 0{,}1000)$. The other model parameters were set equal to $k = 10$ N / m and $c = 0.1$ Nm / s. Samples from the posterior were generated using a simulated annealing MCMC algorithm whose annealing schedule is self-adaptive and is designed to introduce the information contained in the measurements at a constant rate (where information is measured using the Shannon entropy). This is essentially the algorithm outlined in [9] but applied to the situation where the set of measurement data does not grow with time. Throughout, the score of the target distribution was estimated using a finite difference approximation.

The resulting MCMC samples were used to estimate the mean and variance of the posterior (using both standard Monte Carlo and ZV estimation methods). ZV estimates of the variance were realised by defining

$$g_1(\theta) = \theta, \qquad g_2(\theta) = \theta^2 \tag{27}$$

before computing

$$\mathrm{E}[g_2] - \mathrm{E}^2[g_1] = \tilde{g}_2 - \tilde{g}_1^2 \tag{28}$$

For the sake of completeness, the following results are also compared with 'long run' Monte Carlo estimates (realised with sample size of 10000).

Estimates of the mean and variance are shown in Figure 4 and Figure 5 respectively. The variance of the ZV estimates are far reduced relative to the Monte Carlo estimates. The ZV estimates are also centered on the results realised using the long run Monte Carlo simulations.

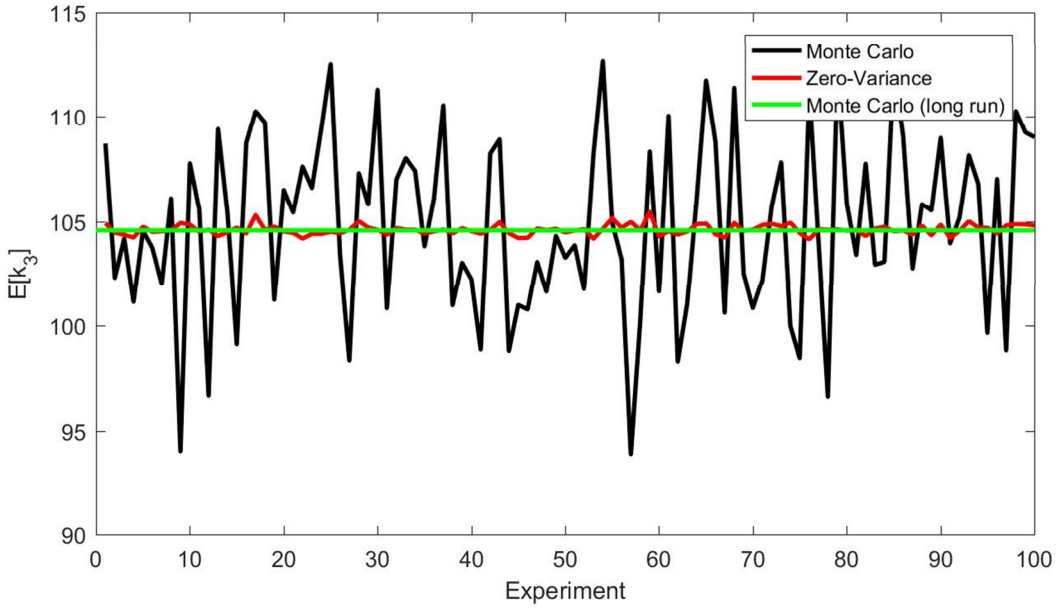

Figure 4. Estimating the posterior mean of the nonlinear stiffness coefficient, k_3.

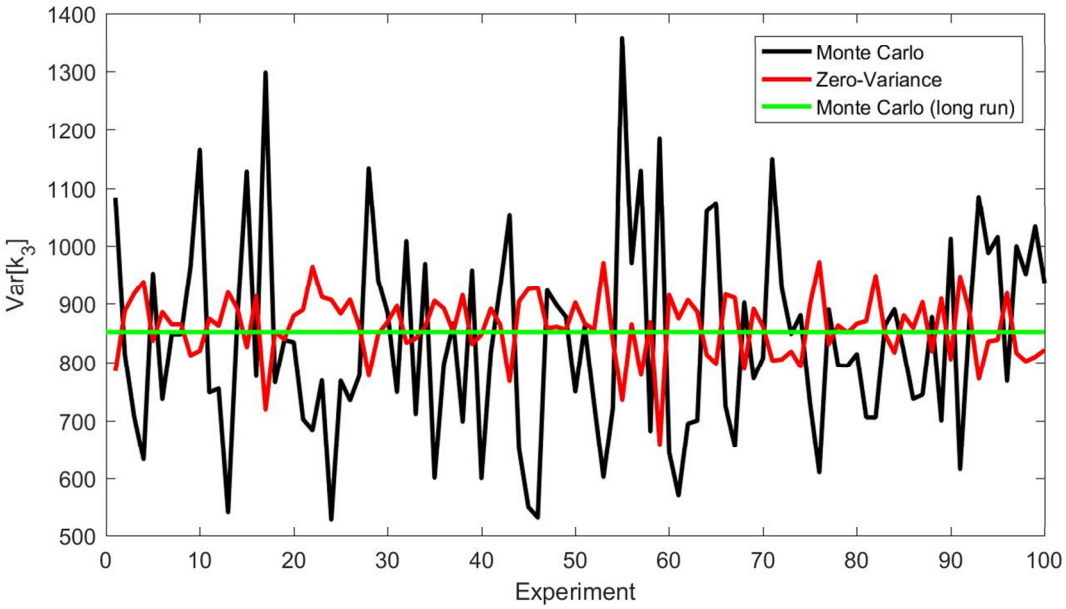

Figure 5. Estimating the posterior variance of the nonlinear stiffness coefficient, k_3.

4 DISCUSSION

The transformations used throughout this paper all make use of the score of the target distribution. The exploitation of the information contained in the score appears to be an important factor in the successful application of the method. With this in mind it seems sensible that, alongside the zero-variance method, a MCMC algorithm which uses the gradient to improve its performance should be employed (Hamiltonian Monte Carlo, for example, is probably the most well-known example of such an algorithm). This is explored in detail in the paper [7]. For future work the author aims to apply the zero-variance method to samples that are generated using an importance-sampling-based Sequential Monte Carlo sampler (recently applied to growing data sets, within the context of mechanical engineering, in [10]). The potential to combine the statistically efficient ZV method with a sampling algorithm that is well suited to parallelisation is particularly attractive.

The current paper demonstrates examples where, despite the statistical errors associated with estimating the parameters $\boldsymbol{\alpha}$, the zero-variance method greatly outperformed standard Monte Carlo. For future work, it would be useful to gain insight into how far the method can be pushed (in other words: how few samples is it possible to 'get away with' when estimating $\boldsymbol{\alpha}$). The current results certainly indicate that it is worth investigating the potential of the ZV method further.

With regard to the simulated annealing algorithm used in the current paper, the self-adaptive annealing schedule relies on estimating the variance of the negative log-likelihood [9]. Interestingly, it was found that the ZV methods investigated here were unable to reduce the statistical error of these estimates and so, as a result, were not able to improve the consistency of the simulated annealing algorithm. For the simple Gaussian examples that were considered here, estimating the variance of the negative log-likelihood would be equivalent to estimating

$\mathrm{E}_{\pi}[\theta^4]$, hinting that a higher order ZV expansion may be required to improve the repeatability of the simulated annealing algorithm (another interesting avenue of future work).

REFERENCES

[1] I. Behmanesh and B. Moaveni, "Accounting for environmental variability, modeling errors, and parameter estimation uncertainties in structural identification," *Journal of Sound and Vibration,* no. 374, pp. 92-110, 2016.

[2] Y. Huang, J. L. Beck and H. Li, "Bayesian System Identification based on Hierarchical Sparse Bayesian Learning," *Computer Methods in Applied Mechanics and Engineering,* no. 318, pp. 382-411, 2017.

[3] J. B. Nagel and B. Sudret, "Hamiltonian Monte Carlo and Borrowing Strength in Hierarchical Inverse Problems," *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering,* vol. 2, no. 3, 2015.

[4] S. K. Au and J. L. Beck, "Estimation of small failure probabilities in high dimensions by subset simulation," *Probabilistic Engineering Mechanics ,* vol. 16, no. 4, pp. 263-277, 2001.

[5] R. Assaraf and M. Caffarel, "Zero-variance principle for monte carlo algorithms," *Physical review letters,* p. 83(23) 4682., 1999.

[6] A. Mira, R. Solgi and D. Imparato, "Zero variance Markov Chain Monte Carlo for Bayesian estimators," *Statistics and Computing,* pp. 23.5: 653-662., 2013.

[7] T. Papamarkou, A. Mira and M. Girolami , "Zero variance differential geometric Markov chain Monte Carlo algorithms," *Bayesian Analysis,* pp. 9.1: 97-128., 2014.

[8] A. Mira, P. Tenconi and D. Bressanini, "Variance reduction for MCMC.," No. qf0310. Department of Economics, University of Insubria, 2003.

[9] P. L. Green, "Bayesian system identification of dynamical systems using large sets of training data: A MCMC solution," *Probabilistic Engineering Mechanics,* pp. 42: 54-63., 2015.

[10] P. L. Green and S. Maskell, "Estimating the parameters of dynamical systems from Big Data using Sequential Monte Carlo samplers," *Mechanical Systems and Signal Processing,* 2017.

APPENDIX

To derive the optimum values of $\boldsymbol{\alpha}$ it is first convenient to show that:

$$\mathrm{E}[\tilde{g}^2] = \mathrm{E}[g^2] + 2\boldsymbol{\alpha}^T\mathrm{E}[g\boldsymbol{z}] + \boldsymbol{\alpha}^T\mathrm{E}[\boldsymbol{z}\boldsymbol{z}^T]\boldsymbol{\alpha} \tag{29}$$

therefore

$$\frac{\partial}{\partial\boldsymbol{\alpha}}\mathrm{E}[\tilde{g}^2] = 2\mathrm{E}[g\boldsymbol{z}] + 2\mathrm{E}[\boldsymbol{z}\boldsymbol{z}^T]\boldsymbol{\alpha} \tag{30}$$

Also:

$$\mathrm{E}[\tilde{g}]^2 = \mathrm{E}[g]^2 + 2\boldsymbol{\alpha}^T\mathrm{E}[g]\mathrm{E}[\boldsymbol{z}]+\boldsymbol{\alpha}^T\mathrm{E}[\boldsymbol{z}]\mathrm{E}[\boldsymbol{z}^T]\boldsymbol{\alpha} \tag{31}$$

therefore

$$\frac{\partial}{\partial\boldsymbol{\alpha}}\mathrm{E}[\tilde{g}]^2 = 2\mathrm{E}[g]\mathrm{E}[\boldsymbol{z}] + 2\mathrm{E}[\boldsymbol{z}]\mathrm{E}[\boldsymbol{z}^T]\boldsymbol{\alpha} \tag{32}$$

Consequently, setting

$$\frac{\partial}{\partial\boldsymbol{\alpha}}\mathrm{Var}[\tilde{g}] = \mathbf{0} \tag{33}$$

it can be shown that

$$(\mathrm{E}[\boldsymbol{z}\boldsymbol{z}^T] - \mathrm{E}[\boldsymbol{z}]\mathrm{E}[\boldsymbol{z}^T])\boldsymbol{\alpha} + (\mathrm{E}[g\boldsymbol{z}] - \mathrm{E}[g]\mathrm{E}[\boldsymbol{z}]) = \mathbf{0} \tag{34}$$

$$\Rightarrow \boldsymbol{\alpha} = -(\mathrm{Var}[\boldsymbol{z}])^{-1}\mathrm{Cov}[g, \boldsymbol{z}] \tag{35}$$

where

$$\mathrm{Cov}[g, \boldsymbol{z}] = \mathrm{E}[g\boldsymbol{z}] - \mathrm{E}[g]\mathrm{E}[\boldsymbol{z}] \tag{36}$$