Mechanical Systems and Signal Processing 56-57 (2015) 109–122

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

Bayesian system identification of dynamical systems using highly informative training data

P.L. Green *, E.J. Cross, K. Worden

Department of Mechanical Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, United Kingdom

ARTICLE INFO

Article history:
Received 5 July 2013
Received in revised form
20 May 2014
Accepted 9 October 2014
Available online 4 November 2014

Keywords:
Nonlinear system identification
Bayesian inference
Markov chain Monte Carlo
Shannon entropy
Tamar bridge

ABSTRACT

This paper is concerned with the Bayesian system identification of structural dynamical systems using experimentally obtained training data. It is motivated by situations where, from a large quantity of training data, one must select a subset to infer probabilistic models. To that end, using concepts from information theory, expressions are derived which allow one to approximate the effect that a set of training data will have on parameter uncertainty as well as the plausibility of candidate model structures. The usefulness of this concept is then demonstrated through the system identification of several dynamical systems using both physics-based and emulator models. The result is a rigorous scientific framework which can be used to select 'highly informative' subsets from large quantities of training data.

1. Introduction

To be practically useful, any system identification method needs to be able to quantify and propagate the inevitable uncertainties which arise as a result of noise-contaminated measurements, as well as the fact that one's chosen model structure will never be able to perfectly replicate the physics of the system of interest. Consequently, system identification is best approached using probability logic such that, rather than searching for the 'perfect model', one is able to assess the relative plausibility of a set of models as well as the parameters within those models [1]. As a result of seminal papers in the machine learning [2] and structural dynamics [3] communities, it is now widely accepted that both levels of inference (parameter estimation and model selection) can be achieved using a Bayesian approach.

With regard to parameter estimation, the plausibility of a model parameter vector $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_{N_D}\}$ given a model structure $\mathcal{M}$ and training data $\mathcal{D}$ can be expressed using Bayes' Theorem:

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

One's belief in the plausibility of $\boldsymbol{\theta}$ before the training data were known is represented by the prior distribution $P(\boldsymbol{\theta}|\mathcal{M})$, while one's belief in the plausibility of $\boldsymbol{\theta}$ after the training data are known is represented in the posterior distribution $P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$. $P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})$ is termed the likelihood and represents the plausibility that the training data $\mathcal{D}$ was witnessed given

* Corresponding author.
E-mail addresses: p.l.green@sheffield.ac.uk (P.L. Green), e.j.cross@sheffield.ac.uk (E.J. Cross), k.worden@sheffield.ac.uk (K. Worden).

http://dx.doi.org/10.1016/j.ymssp.2014.10.003
0888-3270/

model structure $\mathcal{M}$ and parameter vector $\boldsymbol{\theta}$. The evidence $P(\mathcal{D}|\mathcal{M})$ is essentially a normalising constant given by

$$P(\mathcal{D}|\mathcal{M}) = \int \cdots \int P(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})P(\boldsymbol{\theta}|\mathcal{M})d\theta_1 \cdots d\theta_{N_D} \tag{2}$$

thus ensuring that the posterior probability distribution integrates to unity. When dealing with nonlinear systems, the evidence integral is often intractable and, as a result of the curse of dimensionality, cannot practically be evaluated numerically when the number of unknown parameters is greater than 3.

To surmount this issue one can choose to generate samples from the posterior distribution using Markov chain Monte Carlo (MCMC) methods, which can be implemented without having to evaluate Eq. (2). While many MCMC methods have been developed (see Refs. [4–6] for a comprehensive discussion), by far the most popular is the Metropolis–Hastings (MH) algorithm. This involves the evolution of an ergodic Markov chain through the parameter space such that it is able to converge to, and then generate samples from, the posterior distribution. Ensuring that the chain has converged to the globally optimum region of the parameter space (rather than a 'local trap') is a nontrivial problem which has led to the development of the well-known Simulated Annealing algorithm [7] (and its many variants [8–10]) and, more recently, the Adaptive Metropolis–Hastings [11], Transitional Markov chain Monte Carlo [12] and Asymptotically Independent Markov Sampling [13] algorithms.

Once converged, MCMC can then be used to generate samples from the posterior. These samples can be used to analyse parameter correlations, to propagate one's uncertainty in the parameter estimates and to conduct a sensitivity analysis of the model structure of interest (see [14,15] for example). While undoubtedly useful, MCMC tends to be expensive, as many model runs are usually required before one can build up a reasonable 'picture' of the posterior distribution. This is compounded by the fact that, by the nature of MCMC, the samples have not been generated independently and are in fact correlated with each another. Consequently, to avoid making biased estimates, it is often the case that many of the samples generated by MCMC need to be 'thrown away' such that the correlation between the remaining samples is reduced (this process is typically referred to as *thinning*). To alleviate this issue one may choose an alternative to the MH algorithm such as Hybrid Monte Carlo (HMC) [16] which tends to produce samples which are less strongly correlated than the MH algorithm (HMC is discussed in the context of structural dynamics in [17]). However, as HMC utilises estimates of the gradient of the posterior distribution – which incurs additional computational cost – the author's have found that the ability of HMC to outperform the MH algorithm is very dependent on the problem at hand.

To reduce the computational expense of Monte Carlo analysis one may choose to utilise emulators (also known as meta-models or surrogate models) which are inferred directly from the training data rather than from the underlying physics of the system (see [18] for example). The relatively simple structures of emulators often make them considerably easier to analyse, and computationally cheap when compared to physics-based models.

The work in this paper specifically addresses the situation where, to perform Bayesian system identification as part of some collaborative work, one is presented with a very large quantity of data from which to infer probabilistic models. In such a scenario – particularly if one is aiming to utilise physics-based models – it is usually desirable to select a small subset of the training data to reduce the computational cost of running MCMC.[1] In such a scenario one would ideally select a subset of data which is both short and *highly informative* with regard to one's parameter estimates. Consequently, the first aim of this paper is to provide a framework which allows one to view the information content – specifically with regard to one's parameter estimates – of large sets of training data *before* the application of MCMC. This allows one to select subsets of data which are both small, and from which one can learn a great deal about the parameters of a candidate model.

The second aim of this paper is with regard to the second level of inference: model selection. Whether using physics-based models or emulators, any system identification procedure will involve choosing a model $\mathcal{M}$ from a set of candidate model structures (as implied by Eq. (1)). This task is complicated by the fact that model performance cannot be judged simply by how well a model is able to replicate a set of training data as this will lead to overfitted models based on redundant parameter sets. This issue can be addressed by using model selection criteria such as the AIC [20] or the BIC [21], which reward model fidelity while also penalising model complexity. Alternatively, one can phrase the model selection problem using Bayes' Theorem:

$$P(\mathcal{M}_i|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M}_i)P(\mathcal{M}_i)}{P(\mathcal{D})} \tag{3}$$

where $\mathcal{M}_i$ is a model from a set of candidate model structures $\boldsymbol{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_{N_M}\}$. Assuming that there is no prior bias over any of the models in $\boldsymbol{M}$, one can then rate the relative plausibility of two competing model structures (models $\mathcal{M}_i$ and $\mathcal{M}_j$ for example) by computing a Bayes Factor:

$$\beta_{i,j} = \frac{P(\mathcal{M}_i|\mathcal{D})}{P(\mathcal{M}_j|\mathcal{D})}. \tag{4}$$

The Bayes Factor is a model selection criterion which, it can be shown, penalises overfitting without the introduction of *ad hoc* penalty terms (see [1,2,6,22] for more details). Recent work [23] has also shown that such an approach can also be used

[1] Clearly the computational savings achieved through this approach is dependent on the size of the chosen subset, relative to the full set of training data. in situations where the full set of training data is relatively small, the computational savings that could be made through the methods presented in this paper may be small relative to what can be achieved through the parallel implementation of MCMC algorithms (see [19] for example).