A Bayesian Approach to System Identification using Markov Chain Methods

Brett Ninness* Thomas Brinsmead†

Abstract

This paper takes a Bayesian approach to the problem of dynamic system estimation, and illustrates how posterior densities for rather arbitrary system parameters or properties (such a frequency response, phase margin etc) may be numerically computed. In achieving this, the key idea of constructing an ergodic Markov chain with invariant distribution equal to the desired posterior is one borrowed from the mathematical statistics literature. An essential point of the work here is that, via the associated posterior computation from the Markov chain, error bounds on estimates are provided that do not rely on asymptotic in data length arguments, and hence they apply with arbitrary accuracy for arbitrarily short data records.

Technical Report EE02009, Department of Electrical and Computer Engineering, University of Newcastle, AUSTRALIA

1 Introduction

A dominant force in both the practise and underlying understanding of modern methods for system identification of dynamic systems has been work within the context of a Maximum Likelihood framework and associated approximations, such as prediction error approaches [11, 24, 2, 9, 18]. In particular, the software developed as part of this latter effort [12, 1] has become an industry standard.

A key aspect of this approach is that any quantification of the accuracy of the associated system estimates relies on employing asymptotic in data length expressions as if they applied for finite data lengths. For example, the Gaussian distribution commonly achieved by estimates in the infinite limit, is usually assumed to hold for whatever finite data length is available.

While these techniques enjoy widespread acceptance, there has recently arisen a body of work that has sought to derive methods and supporting theory which apply for arbitrarily short data records. To give some examples, the so-called 'bounded error', or 'set estimation' techniques [17, 16, 14, 26] were developed to handle cases where measurement corruptions were of constrained magnitude. More recent work has examined how finite data applicable bounds may be computed for prediction error methods [27, 28], and has also turned to concepts from machine learning theory [4, 3, 25], which have

*This work was supported by the Australian Research Council. This author is with the School of Electrical Engineering & Computer Science, University of Newcastle, Australia and can be contacted at email:brett@ee.newcastle.edu.au or FAX: +61 2 49 21 69 93.

†Senior Research Associate, CRC for Coal in Sustainable Development, Complex Systems Research Group, Philosophy, School of Liberal Arts, Faculty of Education and the Arts, University of Newcastle, Australia and can be contacted at email:tsb543@alinga.newcastle.edu.au or FAX: +61 2 49 21 69 28

originated largely within the Computer Science community, and are also closely linked to ideas of 'information based complexity' [15].

This paper is directed at the same issue of dynamic system identification in a finite data record setting, but takes a different approach to the problem. In particular, the perspective here is that, especially for very short data lengths, it is sensible to take a Bayesian approach to quantifying the manner in which prior knowledge and data-based information are combined to yield posterior information about system properties.

While there are strong scientific and philosophical arguments for this strategy [19], it has historically foundered on the difficulty of actually computing the posterior distribution. Indeed, the celebrated Kalman Filter is well known as one of the few instances where simple computation of the posterior is possible.

However, in relation to this, the introduction of so-called Markov Chain Monte–Carlo methods in the mathematical statistics literature has recently caused something of a minor revolution in that field by offering a means for numerically computing posterior distributions for very complex modelling scenarios [22, 21, 7]. The essential idea there is to invent a method for constructing a Markov Chain which converges to an invariant density equal to the desired posterior. Sampling from this chain then provides a means for computing posteriors with respect to this density via sample averages from the simulated chain, hence the 'Monte–Carlo' epithet.

The contribution of this paper is to provide a tutorial introduction to the key ideas in this area, and then illustrate how they may be successfully applied to the problem of Bayesian estimation of dynamic systems, and in doing so provide estimation error quantifications that apply for arbitrarily short data records.

The ideas behind this paper have begun to have a significant impact in the field of signal processing, and have already been considered in a context of analysing dynamic systems; see for example[23] and the references therein. In particular [10] considers the state estimation of Markov systems, and the more recent work [5, 6] examines extensions to state estimation of jump-Markov systems. However, there seems to be no extant literature examining the parameter estimation, subsequent controller design and estimation-error quantification for finite data set applications that are considered here.

2 Problem Formulation

A very wide class of commonly used linear and time invariant structures that are employed to model the relationship between an observed input data record $\{u_t\}$ and output data record $\{y_t\}$ may be described as [11, 24, 2, 9]

$$y_t = G(q,\theta)u_t + H(q,\theta)e_t = \frac{B(q,\theta)}{A(q,\theta)}u_t + \frac{C(q,\theta)}{D(q,\theta)}e_t \tag{1}$$

where the numerator and denominator polynomials are of the form

$$\begin{aligned} A(q,\theta) &= q^n + a_{n-1}q^{n-1} + \cdots + a_1 q + a_0, && (2)\\ B(q,\theta) &= b_{n-1}q^{n-1} + b_{n-2}q^{n-2} + \cdots + b_1 q + b_0, && (3)\\ D(q,\theta) &= q^n + d_{n-1}q^{n-2} + \cdots + d_1 q + d_0, && (4)\\ C(q,\theta) &= q^n + c_{n-1}q^{n-2} + \cdots + c_1 q + c_0 && (5) \end{aligned}$$

and the parameter vector θ is defined as

$$\theta = [a_0, b_0, d_0, c_0, a_1, b_1, d_1, c_1, \cdots, a_{n-1}, b_{n-1}, d_{n-1}, c_{n-1}].$$

If all of the parameter elements in (2)-(5) are free and to be estimated, then (1) is known as a Box–Jenkins structure, but otherwise more restricted cases such as FIR, ARX, ARMAX and Output–Error types may also be described by (1).

In (1), the signal $\{e_t\}$ is an independent stochastic process that models possible corruptions of the measurements $\{y_t\}$. By virtue of the fact that according to the parameterisation (4), (5) the model $H(q,\theta)$ for correlation in this corruption is monic, then by simple algebra the relationship (1) may be rewritten as

$$y_t = \widehat{y}_{t|t-1}(\theta) + e_t \tag{6}$$

where

$$\widehat{y}_{t|t-1}(\theta) = H^{-1}(q,\theta)G(q,\theta)u_t + \left[1 - H^{-1}(q,\theta)\right] y_t. \tag{7}$$

In fact, by the independence assumption on $\{e_t\}$, the quantity $\widehat{y}_{t|t-1}(\theta)$ is the mean square optimal one-step ahead predictor of y_t according to the model (1), and this suggests a strategy of finding an estimate $\widehat{\theta}_N$ of the parameter vector θ by minimising the error between this predictor, and the observed y_t over N observed data points. That is,

$$\widehat{\theta}_N \triangleq \arg\min_{\theta} V_N(\theta). \tag{8}$$

where

$$V_N(\theta) = \frac{1}{2N}\sum_{t=1}^{N} f(y_t - \widehat{y}_{t|t-1}(\theta)) \tag{9}$$

and f is some positive valued function, with $f(x) = x^2$ being the most common choice. This strategy is known as the 'prediction error method', and enjoys wide popularity not only because of sophisticated software implementations [12], but also due to a rather complete body of supporting theory. In particular, under rather mild assumptions [13, 2, 11] on the distribution of the stochastic component $\{e_t\}$ and on the function f

$$\sqrt{N}(\widehat{\theta}_N - \theta_\circ) \xrightarrow{\mathcal{D}} \mathcal{N}(0, P) \qquad \text{as } N \to \infty. \tag{10}$$

where

$$\lim_{N\to\infty} \widehat{\theta}_N = \theta_\circ \triangleq \arg\min_{\theta} \lim_{N\to\infty} \mathbf{E}\{V_N(\theta)\} \qquad \text{w.p.1} \tag{11}$$

and, provided the data was in fact generated by a system obeying (1) for $\theta = \theta_\circ$, then

$$P^{-1} = \frac{1}{\sigma^2} \lim_{N\to\infty} \frac{\mathrm{d}^2}{\mathrm{d}\theta \mathrm{d}\theta^T} \mathbf{E}\{V_N(\theta)\}\Big|_{\theta=\theta_\circ} \tag{12}$$

where $\sigma^2 = \mathbf{E}\left\{e_t^2\right\}$.

Even though these are strictly asymptotic in data length N results, in practise it is common to assume that the equalities in (10)–(12) hold approximately for the finite data available, so that multi-dimensional Gaussian confidence regions may derived as quantifications of the error between $\widehat{\theta}_N$ and any underlying true $\theta_\circ$.

An added complication in this approach arises when the estimated quantities of interest are not the parameters themselves, but one or more functions of the parameters, such as frequency response, phase margin, etc. In this case, the ubiquitous strategy is to form a first order Taylor expansion in this function of interest about the assumed true parameter, and then couple this with (10) in order to derive error bounds. The success of this approach depends on the Taylor expansion being accurate by virtue of $\|\widehat{\theta}_N - \theta_\circ\|$ being small, and again this depends on the data length N being large.

3 Non-Asymptotic Estimation Error Quantification

As mentioned in the introduction, especially over the last decade, the requirement of large N has been found objectionable by many researchers, who have sought to avoid its need in the error quantification process by a variety of means.

This paper lies within that realm of investigation, and but adopts a new Bayesian computational approach. For this purpose note that for a very general class of models, of which the one adopted here in (1) for the sake of concreteness is a subsumed case, the likelihood function associated with an observed data record can be decomposed as

$$p(Y_N \mid \theta) = p(y_0 \mid \theta) \prod_{t=1}^{N} p(y_t \mid Y_{t-1}, \theta) \tag{13}$$

where $Y_N \triangleq \{y_1, \cdots, y_N\}$. Therefore, by (6) and assuming that e_t is distributed as

$$e_t \sim p_e(\cdot) \tag{14}$$

then

$$p(Y_N \mid \theta) = p(y_0 \mid \theta) \prod_{t=1}^{N} p_e(y_t - \widehat{y}_{t|t-1}(\theta)). \tag{15}$$

Consequently, by Bayes' rule the posterior distribution $p(\theta \mid Y_N)$ is given as

$$p(\theta \mid Y_N) = \frac{p(Y_N \mid \theta)p(\theta)}{p(Y_N)} = \frac{p(y_0 \mid \theta)p(\theta)}{p(Y_N)} \prod_{t=1}^{N} p_e(y_t - \widehat{y}_{t|t-1}(\theta)) \tag{16}$$

where $p(\theta)$ is the a-priori distribution of θ and $p(Y_N)$ is a normalising constant given as

$$p(Y_N) = \int_{\mathbf{R}^{4n}} p(Y_N \mid \theta)p(\theta)\, \mathrm{d}\theta. \tag{17}$$

Now, for a given density p_e, and in the case of the model structure (1), since its associated one-step ahead predictor has the explicit formulation (7), then the left hand side of (16) can, in principle, be computed for any desired θ, and hence the posterior density $p(\theta \mid Y_N)$ may be directly computed.

If the density of only a particular i'th parameter element θ^i is required, then this too can be evaluated via numerical computation of the integral

$$p(\theta^i \mid Y_N) = \int_{\mathbf{R}^{4n-1}} p(\theta^1, \cdots, \theta^i, \cdots, \theta^{4n} \mid Y_N) \mathrm{d}\theta^1 \cdots \mathrm{d}\theta^{i-1} \mathrm{d}\theta^{i+1} \cdots \mathrm{d}\theta^{4n}. \tag{18}$$

For a large dimensional model, this could involve a rather heavy computational load. For example, since typically around thirty points on a histogram are needed to represent it accurately, then (18) implies the numerical evaluation of around thirty multidimensional integrals, and this latter dimension could be relatively large: a fifth order model would imply a nine dimensional integral.

Furthermore, since (16) involves a product term over N terms, then this product can be very large when p_e is greater than one, and very small when it is less than one. This large dynamic range implies great numerical difficulties in its evaluation. One strategy to circumvent this is to instead work with $\log p(\theta \mid Y_N)$, but then the marginal density (18) cannot be computed since the logarithm and integral operator don't commute.

Nevertheless, if possible, the evaluation of (16) and (18) provide a means for providing parameter estimation error quantification applicable to arbitrarily short data records. Indeed, from a Bayesian perspective, the calculation of the posterior (16) is an optimal strategy in that it completely characterises the information available from the data and prior knowledge about the parameter θ.

However, suppose interest is centred not on the parameters themselves, but on a function of them such as (for example) system phase margin $\phi_m(G, K)$ for a given closed loop controller $K(q)$. Then it is not at all clear how one might tractably compute the posterior density

$$p(\phi_m(G, K) \mid Y_N). \tag{19}$$

The contribution of this paper is to illustrate how all the posteriors (16), (18) and (19) together with others that are rather arbitrary functions of θ may be computed in a straightforward manner. As already mentioned, the methods employed here are borrowed from the mathematical statistics literature where they are known collectively as Markov Chain Monte–Carlo approaches, and in the particular case employed here, rely on the employment of the Metropolis–Hastings algorithm. The following section is devoted to a tutorial introduction to these underlying ideas.

4 Markov Chain Methods

In general, except for the strictly Gaussian case, it is impossible to analytically formulate the posterior density specified by (15). The contribution of this paper is to illustrate how it may still be computed numerically. This is achieved by drawing on a large body of work[22, 21, 8] that has generated significant interest in the mathematic statistics community, but has not yet been widely appreciated outside this domain, and in particular appears not to have previously been investigated for the dynamic system estimation applications considered in this paper.

The key idea is to examine a realisation $\{\theta_t\}$ of a Markov-Chain with transition probability

$$p(\theta_k = \theta \mid \theta_{k-1}) \tag{20}$$

such that

$$\lim_{t\to\infty} p(\theta_t = \theta \mid \theta_0) = p(\theta \mid Y_N) \qquad \forall\theta_0 \tag{21}$$

and to then use this simulated realisation $\{\theta_t\}$ as if it were a random sample from $p(\theta \mid Y_N)$. Provided (as will be shown presently) that the required distributional convergence holds, this will lead to consistent estimates of various quantities. For example, it allows the numerical computation and consistent estimation of the conditional expectation $\mathbf{E}\{g(\theta) \mid Y_N\}$ as

$$\mathbf{E}\{g(\theta) \mid Y_N\} = \int_{\mathbf{R}^{4n}} g(\theta)p(\theta \mid Y_N)\mathrm{d}\theta \approx \frac{1}{M}\sum_{t=k+1}^{k+M} g(\theta_t) \tag{22}$$

where g is an arbitrary measurable function, or

$$p(g(\theta) \in A \mid Y_N) \approx \frac{1}{M}\sum_{t=k+1}^{k+M} \chi_{g^{-1}(A)}(\theta_t) \tag{23}$$

where χ is the indicator function and A is a g-measurable set. That is, the left hand side of (23) is the sample histogram of a realisation from the Markov chain used as an estimate of the posterior density.

4.1 Convergence of Markov Chains

The above Markov Chain is time homogeneous, in that $p(\theta_t \mid \theta_{t-1})$ does not depend on t. Suppose that it has the desired ergodic property that for some density $\pi(\theta)$

$$\pi(\theta) = \lim_{t\to\infty} p(\theta_t = \theta \mid \theta_0). \tag{24}$$

Then by straightforward computation

$$\begin{aligned}
\pi(\theta) &= \lim_{t\to\infty} p(\theta_t = \theta \mid \theta_0) \\
&= \lim_{t\to\infty} \int_{\mathbf{R^n}} p(\theta_t = \theta \mid \theta_{t-1}) \cdot p(\theta_{t-1} \mid \theta_0) \mathrm{d}\theta_{t-1} \\
&= \int_{\mathbf{R^n}} \lim_{t\to\infty} p(\theta_t = \theta \mid \theta_{t-1}) \cdot \lim_{t\to\infty} p(\theta_{t-1} \mid \theta_0) \mathrm{d}\theta_{t-1} \\
&= \lim_{t\to\infty} \int_{\mathbf{R^n}} p(\theta_t = \theta \mid \xi) \cdot \pi(\xi)\, \mathrm{d}\xi.
\end{aligned}$$

Therefore, if the Markov chain converges, then it does so to an *invariant* density π, if it exists for such a chain, that satisfies the defining equation for an invariant density of

$$\pi(\theta) = \int p(\theta \mid \xi)\pi(\xi)\, \mathrm{d}\xi. \tag{25}$$

Note that here, and in what follows, to be absolutely general (and correct), we should not assume the existence of absolutely continuous measures and hence integrable densities. Therefore (for example) the above should be a Lebesgue integral with respect to a measure given by a probability distribution. Such formalism would unnecessarily distract from the presentation of the essential ideas, and hence will be avoided, although it can be injected whenever appropriate to rigorise, but perhaps obfuscate the core intuition.

To continue then, suppose that for some density π, the Markov transition probability satisfies the reversibility condition (also called the 'detailed balance' condition)

$$\pi(\xi)p(\theta \mid \xi) = \pi(\theta)p(\xi \mid \theta). \tag{26}$$

Then clearly

$$\int p(\theta \mid \xi)\pi(\xi)\, \mathrm{d}\xi = \int p(\xi \mid \theta)\pi(\theta)\, \mathrm{d}\xi = \pi(\theta) \int p(\xi \mid \theta)\, \mathrm{d}\xi = \pi(\theta) \tag{27}$$

and hence the reversibility condition (26) on π is sufficient for it to be an invariant density with respect to $p(\cdot \mid \cdot)$, in which case it is a candidate for the ergodic limit (24). However, for the convergence (24) to actually occur, the Markov chain defined by $p(\cdot \mid \cdot)$ needs to also satisfy the conditions of *irreducibility* and *aperiodicity*, which are defined as follows [22].

The kernel $p(\cdot \mid \cdot)$ with invariant density π is *irreducible* if, for any initial state θ_0, it has positive probability of entering any set to which π assigns positive probability. Intuitively then, irreducibility is analogous to a concept of reachability, and it is a sufficient condition for realisations of the Markov chain to satisfy

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} g(\theta_k) = \int g(\theta)\pi(\theta)\, \mathrm{d}\theta \tag{28}$$

with probability one, for π almost all θ_0, and for any measurable function g such that

$$\int |g(\theta)|\pi(\theta)\,\mathrm{d}\theta < \infty. \tag{29}$$

The kernel is $p(\cdot \mid \cdot)$ is periodic if there are regions of the state space that it can visit only at certain regularly spaced times; otherwise it is aperiodic. In this case of aperiodicity, distributional convergence occurs in that

$$\lim_{t\to\infty} \sup_{\theta} |p(\theta_t = \theta \mid \theta_0) - \pi(\theta)| = 0 \tag{30}$$

for π almost all θ_0.

4.2 The Metropolis–Hastings Algorithm

Given the afore-going discussion, the challenge now is to construct a Markov chain with irreducible transition kernel $p(\cdot \mid \cdot)$ for which $\pi = p(\theta \mid Y_N)$ is an invariant distribution so that (22) is a valid estimator, and for which (23) is also appropriate if $p(\cdot \mid \cdot)$ is also irreducible and aperiodic.

Fortunately, constructing such Markov chains can achieved far more straightforwardly than might prima facie be thought possible. The key is to use the Metropolis–Hastings algorithm.

To explain this method, the notation θ^i will denote the i'th element of $\theta \in \mathbf{R}^n$, while θ^{-i} will denote all of θ *except* for θ^i. With this in mind, the Metropolis–Hastings algorithm involves the following steps.

1. Initialise θ_0 at some value;
2. At iteration k, consider a candidate value ξ_k^i for the i'th element θ_k^i of θ_k which is drawn from an arbitrary *proposal* density $q_i(\xi_k^i \mid \theta_{k-1})$. That is, find a possible realisation for θ_k^i as
$$\xi_k^i \sim q_i(\cdot \mid \theta_{k-1}); \tag{31}$$
3. Set $\xi_k^{-i} = \theta_{k-1}^{-i}$ and compute the acceptance probability
$$\alpha(\xi_k^i \mid \theta_{k-1}) = \min\left\{1,\ \frac{p(\xi_k^i \mid \theta_{k-1}^{-i}, Y_N)}{p(\theta_{k-1}^i \mid \theta_{k-1}^{-i}, Y_N)} \cdot \frac{q_i(\theta_{k-1}^i \mid \xi_k)}{q_i(\xi_k^i \mid \theta_{k-1})}\right\}; \tag{32}$$
4. Accept the proposed ξ_k^i and set $\theta_k^i = \xi_k^i$ with probability $\alpha(\xi_k^i \mid \theta_{k-1})$;
5. Move to another element i of θ_k and return to step 2. If steps 2-4 have already been performed for all elements θ^i of θ for the current value of k, then increment k and return to step 2.

Note that this is an intuitively obvious method for trying to generate samples from a particular density $p(\theta^i \mid Y_N)$ - make a random choice, check to see if it is more likely to have come from that density than the previous random choice, if so keep it, if not discard it with a certain level of randomness dependent on how unlikely it is to have come from the target density.

More formally though, we now show that the Metropolis-Hastings algorithm as described above achieves the detailed balance condition (26) for $\pi = p(\theta \mid Y_N)$. In particular, note that under the Metropolis–Hastings scheme, θ_k^i can either transit to ξ_k^i or remain unchanged at θ_{k-1}^i. Clearly the probabilities for these events are respectively (assuming in the first instance that $\xi_k^i \neq \theta_{k-1}^i$)

$$p(\theta_k^i = \xi_k^i \mid \theta_{k-1}) = \alpha(\xi_k^i \mid \theta_{k-1}) q(\xi_k^i \mid \theta_{k-1}) \tag{33}$$

$$p(\theta_k^i = \theta_{k-1}^i \mid \theta_{k-1}) = 1 - \int_{\mathbf{R}} \alpha(\xi_k^i \mid \theta_{k-1}) \cdot q(\xi_k^i \mid \theta_{k-1})\,\mathrm{d}\xi_k^i. \tag{34}$$

Therefore, for the case of $\theta_k^i \neq \theta_{k-1}^i$,

$$\begin{aligned} p(\theta_{k-1} \mid Y_N)p(\xi_k^i \mid \theta_{k-1}) &= p(\theta_{k-1}^{-i} \mid Y_N)p(\theta_{k-1}^i \mid \theta_{k-1}^{-i}, Y_N)q_i(\xi_k^i \mid \theta_{k-1}) \times \\ &\quad \min\left\{1, \frac{p(\xi_k^i \mid \theta_{k-1}^{-i}, Y_N)}{p(\theta_{k-1}^i \mid \theta_{k-1}^{-i}, Y_N)} \cdot \frac{q_i(\theta_{k-1}^i \mid \xi_k)}{q_i(\xi_k^i \mid \theta_{k-1})}\right\} \\ &= \min\left\{p(\theta_{k-1}^{-i} \mid Y_N)p(\theta_{k-1}^i \mid \theta_{k-1}^{-i}, Y_N)q_i(\xi_k^i \mid \theta_{k-1}),\right. \\ &\quad \left. p(\theta_{k-1}^{-i} \mid Y_N)p(\xi_k^i \mid \theta_{k-1}^{-i}, Y_N)q_i(\theta_{k-1}^i \mid \xi_k)\right\} \end{aligned} \tag{35}$$

and similarly

$$\begin{aligned} p(\xi_k \mid Y_N)p(\theta_{k-1}^i \mid \xi_k) &= p(\xi_k^{-i} \mid Y_N)p(\xi_k^i \mid \xi_k^{-i}, Y_N)q_i(\theta_{k-1}^i \mid \xi_k) \times \\ &\quad \min\left\{1, \frac{p(\theta_{k-1}^i \mid \xi_k^{-i}, Y_N)}{p(\xi_k^i \mid \xi_k^{-i}, Y_N)} \cdot \frac{q_i(\xi_k^i \mid \theta_{k-1})}{q_i(\theta_{k-1}^i \mid \xi_k)}\right\} \\ &= \min\left\{p(\xi_k^{-i} \mid Y_N)p(\xi_k^i \mid \xi_k^{-i}, Y_N)q_i(\theta_{k-1}^i \mid \xi_k),\right. \\ &\quad \left. p(\xi_k^{-i} \mid Y_N)p(\theta_{k-1}^i \mid \xi_k^{-i}, Y_N)q_i(\xi_k^i \mid \theta_{k-1})\right\}. \end{aligned} \tag{36}$$

Therefore, comparing (35) and (36), noting that the $\min\{\cdot,\cdot\}$ operation is symmetric, and recalling that $\xi_k^{-i} = \theta_{k-1}^{-i}$, then implies that the Metropolis–Hastings algorithm yields a Markov chain for which the reversibility condition

$$p(\theta_{k-1} \mid Y_N)p(\xi_k^i \mid \theta_{k-1}) = p(\xi_k \mid Y_N)p(\theta_{k-1}^i \mid \xi_k) \tag{37}$$

holds.

Consequently, if the chain is also ergodic, then by the developments in §4.1 it converges to the invariant distribution $p(\theta \mid Y_N)$.

Whether or not the convergence itself occurs depends further on the chain being irreducible and aperiodic, and this cannot be checked in abstraction for the Metropolis–Hastings construction. Rather it must be verified on a case by case basis. However, this task can be greatly simplified to one of checking the aperiodicity and irreducibility of the proposal density $q_i(\xi_k^i \mid \theta_k)$ which, as established in [20], is a sufficient condition for the aperiodicity and irreducibility of the Markov chain achieved by the Metropolis–Hastings algorithm.

5 Application to Dynamic System Estimation

With this background in place for methods to construct Markov processes with given invariant densities, we now turn to the application of these ideas for the system identification purposes discussed in §2.

There are two items of importance in this context, the specification of the proposal density $q_i(\cdot \mid \cdot)$, and the evaluation of the acceptance probability $\alpha(\xi_k^i \mid \theta_{k-1})$. In relation to the former, note that if

$$q_i(\xi_k^i \mid \theta_{k-1}) = p(\xi_k^i - \theta_{k-1}^i) \tag{38}$$

holds, then

$$q_i(\theta_{k-1}^i \mid \xi_k) = p(-(\xi_k^i - \theta_k^i)). \tag{39}$$

so that if the proposal density is symmetric in that $p(x) = p(-x)$, then the acceptance probability (32) simplifies to

$$\alpha(\xi_k^i \mid \theta_{k-1}) = \min\left\{1, \quad \frac{p(\xi_k^i \mid \theta_{k-1}^{-i}, Y_N)}{p(\theta_{k-1}^i \mid \theta_{k-1}^{-i}, Y_N)}\right\}. \tag{40}$$

In this special case, the Metropolis–Hastings algorithm is simply known as the 'Metropolis' method. This paper suggests the intuitively reasonable proposal density choice implied by the random walk strategy

$$\xi_k^i = \theta_{k-1}^i + \nu_k \tag{41}$$

where ν_k is a random variable with symmetric density $p_\nu(x) = p_\nu(-x)$; for example a Gaussian density. This implies the simplified acceptance probability computation (40).

This leaves only the question of computing the posterior ratio

$$\frac{p(\xi_k^i \mid \theta_{k-1}^{-i}, Y_N)}{p(\theta_{k-1}^i \mid \theta_{k-1}^{-i}, Y_N)}. \tag{42}$$

However, via the developments in §2 leading to (16)

$$p(\theta^i \mid \theta^{-i}, Y_N) = \frac{p(y_0 \mid \theta)p(\theta)}{p(\theta^{-i} \mid Y_N)p(Y_N)} \prod_{t=1}^{N} p_e(y_t - \widehat{y}_{t|t-1}(\theta)). \tag{43}$$

Therefore, denoting by ϕ_k the parameter vector formed from the union of θ_k^{-i} and ξ_k^i

$$\frac{p(\xi_k^i \mid \theta_{k-1}^{-i}, Y_N)}{p(\theta_{k-1}^i \mid \theta_{k-1}^{-i}, Y_N)} = \frac{p(y_0 \mid \phi_k)p(\phi_k)}{p(y_0 \mid \theta_k)p(\theta_k)} \cdot \frac{\prod_{t=1}^{N} p_e(y_t - \widehat{y}_{t|t-1}(\phi_k))}{\prod_{t=1}^{N} p_e(y_t - \widehat{y}_{t|t-1}(\theta_k))}. \tag{44}$$

Computing this is then straightforward once the density p_e is specified, and once a means for evaluating the one-step ahead predictor $\widehat{y}_{t|t-1}(\theta)$ is available; in the linear system case profiled in §2, the latter is simply given by (7).

Note in particular the simplifying aspect of the acceptance probability $\alpha(\xi_k^i \mid \theta_{k-1})$ depending only on the *ratio* of probabilities, in that normalising constants $p(Y_N)$ and certain densities $p(\theta^{-i} \mid Y_N)$ need not be computed. Also, numerical problems associated with large dynamic ranges of the ratio components are avoided once they are computed in a ratio-combined fashion.

Furthermore, note that the proposal density (41) where ν_k has Gaussian distribution clearly assigns non-zero probability to any value of ξ_k^i for any given θ_{k-1} and hence is irreducible with respect to any other density, including the posterior $p(\theta^i \mid Y_N)$. Additionally, it imposes no periodic constraints, and hence is aperiodic.

Therefore, according to the results of [20], the Markov chain realised by the Metropolis algorithm using the above acceptance probability computation and proposal density is also $p(\theta^i \mid Y_N)$ irreducible and aperiodic, and hence is ergodic with invariant distribution $p(\theta^i \mid Y_N)$.

That is, after a suitable 'burn in' period in which the Markov chain is allowed to converge to its invariant distribution, a realisation $\{\theta_k, \cdots, \theta_M\}$ will be a sample path of a stochastic process distributed as $p(\theta \mid Y_N)$, and hence can be used for estimating various posterior distributions and quantities. For example, the conditional expectation of θ given Y_N can be computed as

$$\mathbf{E}\{\theta \mid Y_N\} \approx \frac{1}{M-k} \sum_{t=k}^{M-1} \theta_t \tag{45}$$

while the posterior density of the system phase margin $\phi_m(G,K)$ for a given controller $K(q)$ can be simply computed as the sample histogram of the values $\phi_m(G(\theta_t),K)$ for $t \in [k,M]$.

With regard to the latter, it should be made clear that the phase margin is only being used as a rather arbitrary example to emphasise that computations with respect to almost any function of the parameters are easily performed. That is, once the main computational load of finding a realisation $\{\theta_k\}$ has been performed, further operations such as (45) or histogram computation are rather trivial.

6 Simulation Example

To illustrate the application of these ideas, we begin by addressing the very simple illustrational example of

$$y_t = \left(\frac{b}{q-a}\right) u_t + e_t \tag{46}$$

where $b = 0.2$, $a = 0.8$ and $\{e_t\}$ is a zero mean i.i.d. process with $\sigma^2 = \mathbf{E}\{e_t^2\} = 0.01$. Suppose further that the available data consists of $N = 20$ samples of $\{y_t\}$ and $\{u_t\}$ when u_t is a piecewise constant signal, transiting $1 \mapsto 0$ at $t = 10$. This is illustrated in figure 1, where the solid line is the noise free response, and the samples around this line are the noise corrupted data assumed to be available.

In the case where the density $p_e(\cdot)$ governing e_t is uniform, and with prior distribution on $\theta = [b,a]$ being one that assigns zero weight to $b < 0$ and $|a| > 1$, then the posterior distributions for these parameters given the data realisation shown in figure 1 are illustrated in figure 2.

There, the solid line shows the marginal posterior density for b and a computed via numerical computation of the integral (18) via the use (16) to obtain the posterior joint density. The bar graph, is the sample histogram of 10^5 realisations from the Markov chain, constructed via the material presented in §4.2 and §5 to have invariant density $p(\theta \mid Y_N)$. Note the close agreement between these two numerically computed estimates of the posterior.

If an Output–Error model structure is fitted to this data via the methods outlined (1)–(9) with $f(x) = x^2$, and then the asymptotic results (10)–(12) are, as is commonly done, used in a finite data setting in order to provide error quantification, then the results of this strategy are as shown by the dash-dot Gaussian-curve lines in figure 2. While these quantifications are not strictly comparable to the posterior distributions, since they evaluate different quantities, it would still seem interesting to compare the two in terms of their utility for informing a user of what system information can be extracted from the available data.

Finally, suppose that a closed loop PI controller $K(q)$ is to be designed for this plant, so as to achieve at least $\phi_m(G,K) = 105^\circ$ of phase margin, and suppose that

$$K(q) = 1 + \frac{0.1}{q-1} \tag{47}$$

is a candidate for this task. Then to assess the suitability of this proposal, one could seek to know the posterior density

$$p(\phi_m(G,K) \mid Y_N) \tag{48}$$

in order to assess what information the data contains in relation to the question of whether (47) will attain the design objective. While this would seem a daunting task from an analytical point of view, the same Markov chain realisation $\{\theta_k\}$ used to form the histogram estimates of posterior densities in figure 2 can be simply employed to estimate (48) via its sample histogram, which is shown in the

top diagram of figure 3. Clearly, there seems to be good evidence from the data that the controller (47) will achieve the required 105° phase margin. Finally, for the sake of completeness, the computed posterior density for the gain margin $g_m(G, K)$ for this same scenario is shown in the bottom diagram of figure 3, and indicates that the data strongly supports a conclusion that the controller (47) achieves a gain margin of greater than 7.5.

7 Conclusion

This paper presents a preliminary investigation of the employment, in a systems and control setting of new ideas and techniques developed in the mathematical statistics literature. While the early results presented here appear to have promise, there are many questions and issues to be investigated, of which a few that spring to mind are:

- How might Markov Chain convergence be accelerated, and how can one assess that convergence has occurred?
- How should the proposal density be chosen, and how does it affect the preceding issues?
- What if the parameters affecting $p_e(\cdot)$, such as variance σ^2 are unknown? Clearly, they can be included in the posterior parameters to be estimated, but how does this affect the remaining parameters?
- How can these ideas be sensibly married with control design methods in order to seek a data-to-controller synthesis solution?
- How might this be extended to modelling scenarios more sophisticated than the linear time invariant one considered here for the purposes of illustration? Achieving this is likely to be quite straightforward, since the key step of evaluating the posterior via Bayes' rule is invariant to the complexity of the underlying model, although the computation of the associated one step ahead predictor might not be.

The resolution of some of these questions, especially the initial ones, can be aided by drawing on existing literature. Addressing all of the above issues will be the subject of further work by the authors. It should be admitted that a drawback of the methods proposed is their computational demands. However we would point out that they are eminently parallelisable, and that current widely available computing powers far exceed those of a decade ago when the issues underlying this paper gained impetus. Furthermore, to date these computing resources continue to grow at Moore's law rate.

References

[1] *Frequency Domain System Identification Toolbox for MATLAB*, `http://elecwww.vub.ac.be/fdident/index.html`.

[2] P. CAINES, *Linear Stochastic Systems*, John Wiley and Sons, New York, 1988.

[3] M. C. CAMPI AND P. R. KUMAR, *Learning dynamical systems in a stationary environment*, Systems Control Lett., 34 (1998), pp. 125–132. Learning theory.

[4] M. C. CAMPI AND M. VIDYASAGAR, *Learning with prior information*, IEEE Trans. Automat. Control, 46 (2001), pp. 1682–1695.

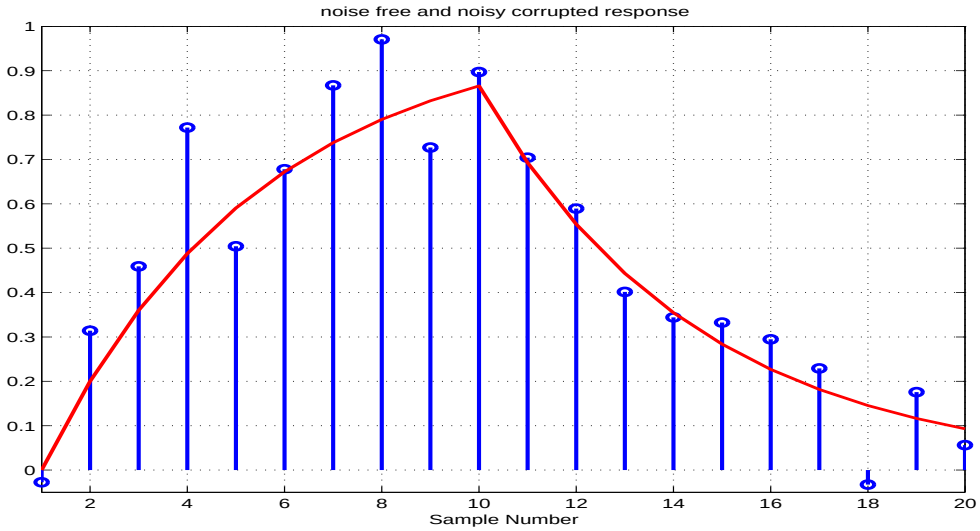

Figure 1: *System response: Solid line is noise free, sampled dots are the noise corrupted measurements actually available.*

[5] A. DOUCET, N. GORDON, AND V. KRISHNAMURTHY, *Particle filters for state estimation of jump markov linear systems*, IEEE Transactions Signal Processing, 49 (2001), pp. 613–524.

[6] A. DOUCET, A. LOGOTHETIS, AND V. KRISHNAMURTHY, *Stochastic sampling algorithms for state estimation of jump Markov linear systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 188–202.

[7] W. GILKS, S. RICHARDSON, AND D. SPIEGELHALTER, *Markov Chain Monte Carlo in practice*, Chapman and Hall, 1996.

[8] P. GREEN, *Complex Stochastic Systems*, Chapman and Hall/CRC, 2001, ch. A Primer on Markov Chain Monte Carlo.

[9] E. HANNAN AND M. DEISTLER, *The Statistical Theory of Linear Systems*, John Wiley and Sons, New York, 1988.

[10] J. S. LIU AND R. CHEN, *Sequential Monte Carlo methods for dynamic systems*, J. Amer. Statist. Assoc., 93 (1998), pp. 1032–1044.

[11] L. LJUNG, *System Identification: Theory for the User, (2nd edition)*, Prentice-Hall, Inc., New Jersey, 1999.

[12] ——, *MATLAB System Identification Toolbox Users Guide, Version 5*, The Mathworks, 2000.

[13] L.LJUNG AND P.E.CAINES, *Asymptotic Normality of prediction error estimators for approximate system models*, Stochastics, 3 (1979), pp. 29–46.

[14] M. MILANESE AND A. VICINO, *Optimal inner bounds of feasible parameter set in linear estimation with bounded noise*, IEEE Transactions on Automatic Control, 36 (1991), p. 759.

[15] ——, *Information based complexity and nonparametric worst-case system identification*, Journal of Complexity, 9 (1993), pp. 427–446.

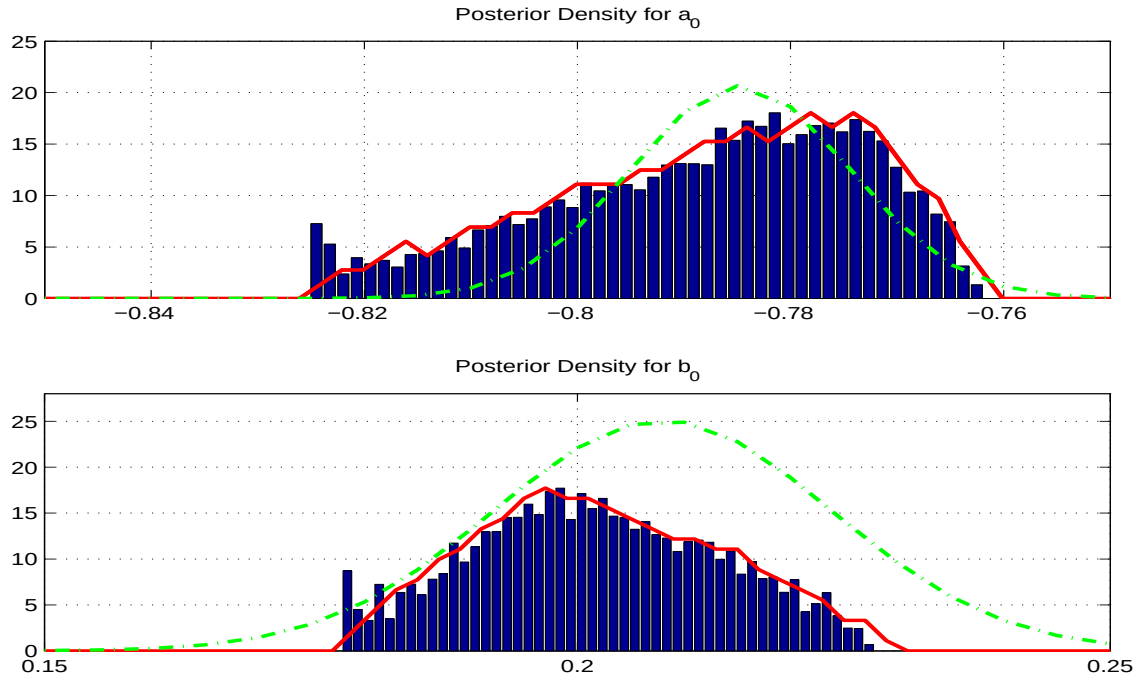

Figure 2: *Posterior marginal densities of parameters computed as sample histograms, together with (solid) line another evaluation of the posterior marginals computed via numerical integration of the joint posterior, and (dash-dot line), the parameter information that would be inferred from the data via a prediction error approach with asymptotic in N quantifications used for finite N.*

[16] J. NORTON, *Identification and application of bounded parameter models,* Automatica, 23 (1987), pp. 497–507.

[17] ——, *Identification of parameter bounds of ARMAX models from records with bounded noises,* International Journal of Control, 42 (1987), pp. 375–390.

[18] R. PINTELON AND J. SCHOUKENS, *System Identification: A Frequency Domain Approach,* IEEE Press, 2000.

[19] C. P. ROBERT, *The Bayesian Choice,* Springer Verlag, 2 ed., 2001.

[20] G. O. ROBERTS AND A. F. M. SMITH, *Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms,* Stochastic Process. Appl., 49 (1994), pp. 207–216.

[21] A. SMITH AND G. ROBERTS, *Bayesian compuations via the gibbs sampler and related markov chain monte carlo methods,* Journal of the Royal Statistical Society-Series B, 55 (1993), pp. 3–23.

[22] L. TIERNEY, *Markov chains for exploring posterior distributions,* Ann. Statist., 22 (1994), pp. 1701–1762. With discussion and a rejoinder by the author.

[23] J.-Y. TOURNERET AND O. C. (GUEST EDS), *Special section on markov chain monte carlo (mcmc) methods for signal processing,* Signal Processing, (2001).

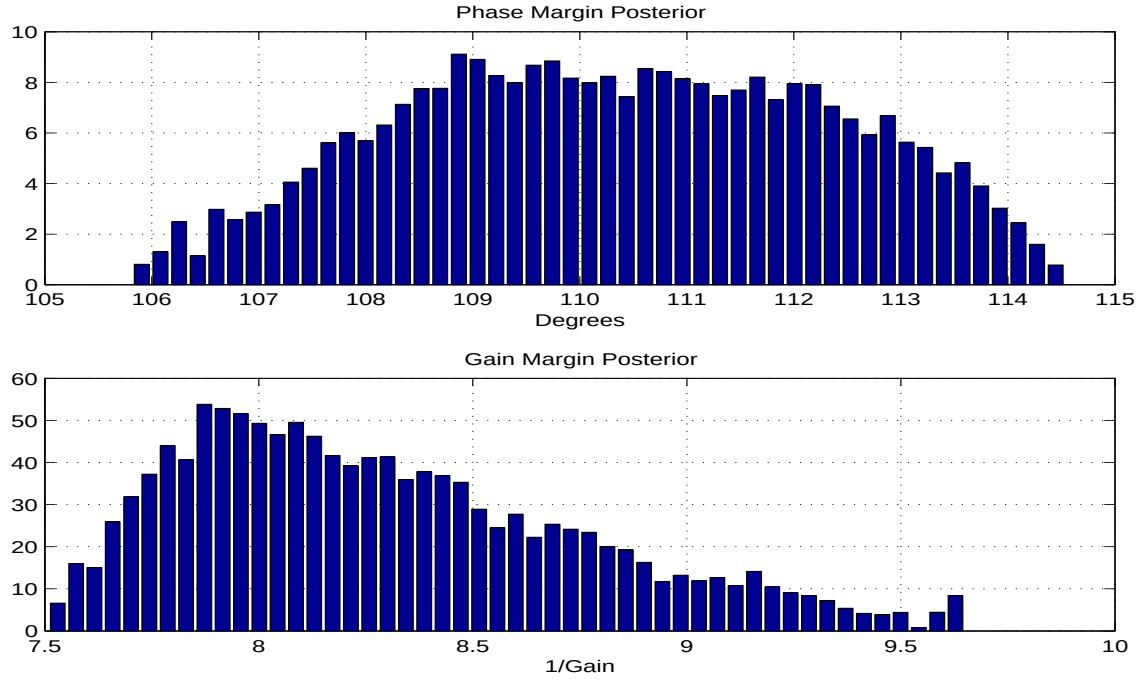

Figure 3: *Posterior distributions for phase margin $\phi_m(G, K)$ and gain margin $g_m(G, K)$ for a given PI controller.*

[24] T.SÖDERSTRÖM AND P.STOICA, *System Identification,* Prentice Hall, New York, 1989.

[25] M. VIDYASAGAR, *An introduction to some statistical aspects of PAC learning theory,* Systems Control Lett., 34 (1998), pp. 115–124. Learning theory.

[26] E. WALTER AND H.PIET-LAHANIER, *Exact recursive polyhedral description of the feasible parameter set for bounded-error models,* IEEE Transactions on Automatic Control, AC-34 (1989), pp. 911–914.

[27] E. WEYER, *Finite sample properties of system identification of ARX models under mixing conditions,* Automatica J. IFAC, 36 (2000), pp. 1291–1299.

[28] E. WEYER, R. C. WILLIAMSON, AND I. M. Y. MAREELS, *Finite sample properties of linear model identification,* IEEE Trans. Automat. Control, 44 (1999), pp. 1370–1383.