Perspectives and Advances in Parameter Estimation of Nonlinear Models

Milena Clarissa Cuéllar Sánchez

Supervisor: Professor Leonard A. Smith

Department of Statistics

The London School of Economics and Political Science

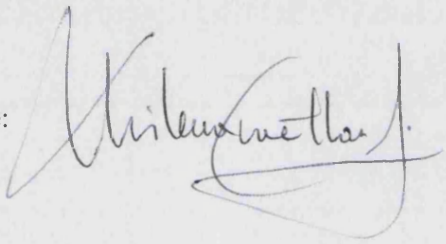
University of London

2007

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Statement of Authenticity

I confirm that this Thesis is all my own work and does not include any work completed by anyone other than myself. I have completed this document in accordance with the Department of Statistics instructions and with in the limits set by the School and the University of London.

Signature: 

Date: 05.02.2007

Abstract

Nonlinear methodologies to estimate parameters of deterministic nonlinear models are investigated in the case where experimental observations are available and uncertainty sources are present, e.g. model inadequacy, model error and noise. The problem of parameter estimation is interpreted from a nonlinear dynamical time series analysis perspective; however deterministic and probabilistic techniques originated outside the nonlinear deterministic framework are studied, implemented and discussed.

Conceptually, the Thesis is divided in two parts that explore two fundamentally different approaches: (a) Bayesian and (b) Geometrical estimation. Both approaches attempt to estimate parameters and model states in the case where the system and the model used to represent it are identical, i.e. Perfect Model Scenario (PMS), even though the implications of the results obtained are considered for Imperfect Model cases. The performance of the resulting model parameter estimates in control monitoring and forecasting of the corresponding system is assessed in an application-oriented fashion and contrasted where possible with system observations, in order to look for a consistent way to combine probabilistic and deterministic approaches. Given the presence of uncertainty in the model used to represent a system and in

the observations available, combined methodologies enable us to best interpret the resulting estimates in a probabilistic framework as well as in the context of a particular application.

The first conceptual part relates to the REMIND project, which is to find a way to meld advances in nonlinear dynamics with those in Bayesian estimation for both mathematical systems and real industrial settings, i.e. for control monitoring the UK's electricity grid system efficiently. Bayesian inference is used to estimate model parameters and model states using Markov Chain Monte Carlo (MCMC) techniques. For the observations of grid frequency and demand, the operational constraints of the data sets are maintained through the estimation process, for example in the situation where the data are provided at rates that restrict on-line storage and post processing. When MCMC is applied to the Logistic Map, curious behaviour of the convergence of the Markov Chain and in the resulting parameter and states estimates are observed and are suspected to be a consequence of high multimodality in the resulting posterior, which in turn generates estimates with a low dynamical informational content. In the case when the MCMC is applied to a UK's grid frequency dynamical model, the technique is implemented in such a way that gradually transform from the PMS case into a more realistic model representation of the system. Convergence of the MCMC algorithm for the grid frequency models is highly dependent on the quality of opera-

tional data, which fails to provide the information required by the tailor-made MCMC implementation. In addition, sanity checks are proposed to establish meaningful convergence of MCMC analyses of time series in general.

The second conceptual part explores a new approach to parameter estimation in nonlinear modelling, based on the geometric properties of short term model trajectories, whilst keeping track of the global behaviour of the model. Geometric properties are defined in the context of indistinguishable states theory. Parameter estimates are found for low dimensional chaotic systems by means of Gradient Descent methods (GD) in the PMS. Some of the advances are made possible by means of improving the balance between information extracted from the observations and from the dynamical equations.

As a result of this investigation, it is noted that, even with perfect knowledge of system and noise in both models, the uncertainty in the dynamics cannot be distinguished from the uncertainty in the observations. In addition, the Geometric approach and Bayesian approach of the problem of model parameter and state estimation for nonlinear models in the PMS are compared aiming to distinguish them based on dynamical features of the estimates. In the Bayesian formulation there are still fundamental challenges when a perfect model is not available.

Contents

A la memoria de mi abuelita Alcira

y dedicada a cada uno de mis abuelos:

a mi abuelito Carlos, el papá de mi mamá;

a Mariachi y el abuelo Rafael, los padres de mi papá.

Cada uno ha contribuido con la historia de su vida,

de diferentes maneras y proporciones

a lo que mis padres han sido y son ahora,

y en consequencia lo que yo soy.

Todos han tenido vidas extraordinarias y mágicas

que han cultivado en mí curiosidad

y capacidad de seguir siempre adelante

disfrutando y admirando cada momento,

solo mirando atrás para reir de lo que ha sido

y mirando hacia adelante con respeto,

sin tener miedo de lo que será.

Acknowledgements

There are many people that has been a part of the whole process of my *doing a Ph.D.*. Thus, I will like to thank them all, going one by one as far as I remember them all, if I do not remember you, be sure that as soon as I print the last version I will remember you and I will thank you from my heart.

I would like to greatly acknowledge Luis, my partner, friend, husband, colleague, and many other things at the same time; without his support and encouragement, at all levels, this story might have been very different. These long four years have been a great adventure for us, where laughter was and continues being the most important aspect of it all.

To my family, that has been always there for me. Being apart from them has been a great challenge but at the same time a confirmation of the strong motivation they had cultivated in me (and all my sisters) to go always forward looking for new challenges in both personal and professional life. I would like to thank my father, Rafael, for his encouragement and support towards my professional development; my mother, Olga, for her inspiring will to go beyond any trouble in life; my aunt Fulvia for her discipline and motivation given to all her nieces and nephews to pursue professional goals; my sisters: Mónica Concepción, Marcela Cristina, Maria Carolina y Mayra Constanza, for all their love and caring for me; and my niece Natalia for her happiness and joy for life.

Many people have taken the place of family members during these more than four years: my sister-in-law Maria for her support in many practical and logistic aspects of my life in UK; her children Violeta and Fidel for their caring and closeness to me; Nati, my mother-in-law, and Luis' sisters: Pilar, Teresa, Begoña and Alicia; that had taken me inside their family and I will appreciate that forever. I feel at home with them.

To Anna Andrianova for her great friendship, she has been there for me in good and bad times. She has made my life in London, in the personal and academic sense, warm enough to continue. To her boyfriend, David Mobbs, for his sharp sense of humour and brightness in many discussions we had.

To my friend Alexandra Olaya, for all the times we shared laughing. During this time, we have been reflecting each other path in "parallel" worlds. She is my foreing sister.

To Lenny whom has believed in me and has supported me through many moments and situations. I learned many things from him when he has been close or away. He has the spark to influence the life of people around him. I will be always grateful to him.

To my colleagues in LSE, everyone has touched both my personal and professional life and my main reason to thank them is that they have showed me the diversity of ways life can be approached: Kevin Judd, Devin Kilminster, Edward Tredger, Sarah Higgings, Antje Weisheimer and Frank Kwasniok. To Liam Clarke I own special thanks. He has the talent to find simple metaphors to explain things and has been a guiding light through the NGT project and the time I spent in CATS. To Jochen Brocker for his friendship and useful and sharp comments in many informal discussions, and to Du Hailiang for his natural brightness and hard work during joint work and also that I learned how to properly fry spring rolls.

To the NGT team that supported us through the REMIND project and the Smith Institute, specially to Melvin Brown, Hai Bin Wan, and Ahmad Chebbo. I will also like to thank: Imelda Noble, Esther Heyhoe, Lyn Grove, Inesh, Emma and Paul, Tina, Effie, Claudio Zappia, Mónica and Duncan McLeod.

This project was financially supported by a grant awarded to LSE by the National Grid Transco, Plc. My scholarship was enmarked in the EPSRC, REMIND project managed by the Smith Institute for Industrial Mathematics and System Engineering. Additional financial support was awarded to me annually by the Department of Statistics at LSE, throught Graduate LSE Scholarships. Without this financial support all this adventure could not have been possible.

Milena Cuéllar
London, 29th July 2006

Chapter 1

The Problem

1.1 Introduction

The behaviour of natural and artificial systems has captured the attention of humans since the beginning of mankind. Every moment of every day, rationalisation of surrounding phenomena is performed.

Observations are constantly obtained and used in order to characterise and attempt to control the surrounding reality. Sets of rules are assigned to represent this reality with the aim of reproducing the observed phenomena. These sets of rules are the models that represent reality and differ from subject to subject depending on many different aspects such as current circumstances, awareness and interests, among others. Once phenomena are characterised according to the chosen model, the environment becomes more

secure and malleable—or it feels so at any rate. Although no model can change reality directly, such models help understand the world.

The information obtained from observations is incomplete and is frequently corrupted. Furthermore, it is normally impossible to acquire all the relevant information needed to assign a consistent set of rules. As a consequence, the model is just an approximation of reality. Models used to represent reality are sometimes mistakenly thought of as reality itself, given the success of the model. Even when the chosen model does not provide a very accurate representation of reality, it is taken as reality itself.

Modelling reality plays a key role in the scientific method and in the applications of its laws and developments towards particular applications. In this very restricted context, reality is the system where the phenomena of interest happens, and observations are quantities that can be measured, *i.e.* quantities with units. Observations contain information on the system and are used to formulate a model or to refine a previously formulated one. Models are mathematical structures that formally described the system.

In the process of gathering all relevant information about a system, many obstacles are encountered. For example, storage capacity is finite, measurements cannot be performed in continuous time, measurement devices have finite resolution, not all relevant variables can be observed and there are often systematic errors during the experimental process. Therefore the in-

formation available is uncertain. In addition, given that only observations are available, discriminating the nature of a system, *i.e.* stochastic or deterministic, is impossible [64]. Therefore an a priori choice of a stochastic or a deterministic model is arbitrary.

Although uncertainty makes the view of reality fuzzy, it also can be used to extract useful information. In more optimistic terms, uncertainty in the observations and in the model chosen to represent the system may highlight the relative ability of a model to represent the system in terms of forecasting or control monitoring.

This Thesis focuses on the situation where model parameter values are to be found once a model has been formulated to represent a system of interest. Parameter estimation is a fundamental problem in both stochastic and deterministic frameworks although it is approached in different ways.

This work follows the idea that there is no "proper methodology" for parameter estimation when the only source of information is a time series. Furthermore, it considers two model choices to be used to represent the dynamical system: (1) separable dynamical models that are deterministic but contain a separable stochastic component referred to throughout this Thesis as *measurement noise*, and (2) stochastic dynamical models that contain a non-separable noise component called *dynamical noise* in addition to the traditional deterministic and separable noise components. The use of either

of the two model choices is justifiable when formulated towards coping with uncertainty sources in the system. Given a particular system, validation of model choice and definition of "good" estimates are made from *out of sample* performance.

The use of dynamical models that can be either stochastic or deterministic leads to approaches to solve the problem of parameter estimation that meld methodologies from statistical and nonlinear times series analysis frameworks.

In this Thesis, the problem of parameter estimation is interpreted from a non-linear time series analysis perspective; however, techniques originated outside the nonlinear deterministic framework are studied, implemented and discussed for dynamical systems.

This Thesis is structured as follows. Chapter 1 formulates the problem in simple terms. Chapter 2 introduces the techniques used at different stages in the research for particular applications and provides an overview of what is new in this Thesis.

Chapter 3 uses Bayesian methodologies of parameter estimation to estimate parameters for the Logistic map [24]. The Chapter presents a correct formulation of the problem of parameter estimation in Bayesian terms and implements a tailored MCMC routine for this case [23].

Chapter 4 presents and describes a new methodology [88] that uses in-

distinguishable state theory [54, 55] and ensemble construction to search for parameter estimates in nonlinear models.

Chapter 5 contrasts system state estimates obtained by two different approaches, Bayesian and dynamical. It provides interesting results which lead to an extensive plan of further work.

Chapter 6 describes the attempts and results to estimate parameters for a simple model of electricity grid frequency dynamics [20] using Bayesian methodologies and real experimental data [22].

Finally, Chapter 7 lists the new results and general outcomes of this work, highlighting further research in this area.

1.2 Statement of the Problem

Let S be a time series of observations of a system's temporal evolution. The temporal evolution of the system is given by the map, $\tilde{f}(\tilde{x}_t; \tilde{\boldsymbol{\theta}}) : \mathbb{R}^m \longrightarrow \mathbb{R}^m$, where at time t, the system state is given by

$$\tilde{x}_t = \tilde{f}(\tilde{x}_{t-1}; \tilde{\boldsymbol{\theta}}), \tag{1.1}$$

$\tilde{x}_t \in \mathbb{R}^m$ are the system states at time $t \in \mathbb{Z}$ and $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^\ell$ is the fixed value for the true parameter vector of the system.

Estimation of the true parameter value $\tilde{\boldsymbol{\theta}}$ can be performed in several possible scenarios. Each scenario represents a relationship between the sys-

tem's temporal evolution and the model chosen to represent it. The possible scenarios are:

1. *The Perfect Model Scenario* (PMS):

   The system and the model share the same mathematical structure. Therefore, $x_t \equiv \tilde{x}_t$ for all t; thus $f \equiv \tilde{f}$. Given that the system's temporal evolution is given by $\tilde{f}$ for a fixed value of the parameter vector $\tilde{\boldsymbol{\theta}}$, the model chosen to represent the system is chosen from the model class $\tilde{f}(\,\cdot\,;\boldsymbol{\theta})$.

   In the special case where the temporal evolution of the system states, $\tilde{x}_t$, is modelled by the deterministic map, $f = f(x_t;\boldsymbol{\theta}) : \mathbb{R}^n \longrightarrow \mathbb{R}^n$,

$$x_{t+1} = f(x_t;\boldsymbol{\theta}), \tag{1.2}$$

   where x_t for all t are the model states and $\boldsymbol{\theta} \in \mathbb{R}^\ell$ is the model parameter vector. The map in equation (1.2) can be written in terms of an unknown initial condition x_0 by the $t-$fold composition of f as

$$x_{t+1} = f^t(x_0;\boldsymbol{\theta}). \tag{1.3}$$

   At time t, it is assumed that all components of $\tilde{x}_t$ are observed, *i.e.* $n = m$, and the data point $s_t \in \mathbb{R}^m$ is recorded. In other words, the sampling rate is constant. The length of the data set S is $N \in \mathbb{N}$, and is equal to the number of times that a system trajectory is observed.

Given that each observation is subject to noise, the measurement noise component is

$$s_t = \tilde{x}_t + \eta_t, \tag{1.4}$$

where $\eta_t \in \mathbb{R}^m$ and $\eta_t \sim IID(0, \sigma_\eta^2)$, an independent and identically distributed random variable with known variance σ_η^2.

Notice that in some cases, the system under observation is physically under the influence of internal random fluctuations. Therefore the system states, $\tilde{x}_t$, are randomly perturbed by dynamical noise. If the unknown initial true state of the system is $\tilde{x}_0$, the additive dynamical noise is mathematically represented as

$$\begin{aligned}
\tilde{x}_{t=1} &= f(\tilde{x}_0 + \tilde{\delta}_0; \boldsymbol{\theta}), \\
\tilde{x}_2 &= f(\tilde{x}_1 + \tilde{\delta}_1; \boldsymbol{\theta}), \\
&\vdots \\
\tilde{x}_t &= f(\tilde{x}_{t-1} + \tilde{\delta}_{t-1}; \boldsymbol{\theta}), \\
\tilde{x}_{t+1} &= f(\tilde{x}_t + \tilde{\delta}_t; \boldsymbol{\theta}).
\end{aligned} \tag{1.5}$$

where $\tilde{\delta}_t \in \mathbb{R}^m$ is a random variable with unknown mean and variance, $\tilde{\delta}_t \sim IID(0, \sigma_\delta^2)$.

The perfect model scenario case can be dressed in two ways:

(1) The observations of the system (1.1) only contain measurement

noise, the PMS is a separable dynamical model given by equations (1.2) and (1.4), explicitly,

$$s_t = f^t(\tilde{x}_0; \boldsymbol{\theta}) + \eta_t, \tag{1.6}$$

for $t > 0$. Note that in this case, the sequence of states, $\{\tilde{x}_t\}_{t>0}$, is indeed a trajectory of the system.

(2) The observations of the system (1.1) contain both additive noise components, the PMS is a stochastic dynamical model given by equations (1.2) to (1.5). Explicitly,

$$s_t = \tilde{x}_t + \eta_t, \tag{1.7}$$

where $\tilde{x}_t = f(\tilde{x}_{t-1} + \tilde{\delta}_{t-1})$ thus

$$s_t = \underbrace{f(f(\ldots f(f(\tilde{x}_0 + \tilde{\delta}_0) + \tilde{\delta}_1) \cdots + \tilde{\delta}_{t-2}) + \tilde{\delta}_{t-1})}_{t \text{ times}} + \eta_t, \tag{1.8}$$

for $\tilde{\delta}_t \sim IID(0, \sigma_\delta^2)$. Therefore, the sequence of states, $\{\tilde{x}_t\}_{t>0}$, is no longer a trajectory of the system

The admission of the presence of dynamical noise in the observations may sometimes be seen as a "shadow" in low dimensions of higher dimensional dynamics with small amplitude [56]. In some sense, the random perturbation called dynamical noise affects the system states before the random perturbation called measurement noise affects the

system states during the experimental process of gathering system observations. The effects on the dynamics from the presence of both noise types are studied in detail in [11] for chaotic systems.

2. *The Imperfect Model Scenario* (IMS):

   The system is approximately represented by the model. Therefore, $f \simeq \tilde{f}$ and $x_t \neq \tilde{x}_t$. Perfect models are not available in cases where data comes from physical systems ([55] and references therein). The "Laws of Physics" are only a useful approximation of the system under well defined conditions [15]. Model inadequacy arises when the model chosen to represent the system is structurally incorrect, is phenomenological not derived from any physical principles, does not include unknown and not observed dynamical components of the system, involves coarse measured variables or variables representing averages, among other factors.

   The only information available about the system states is provided by the observations recorded at time t. Given that the system $\tilde{f}$ is represented by an imperfect model, of the same dimension, $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$, it is assumed the observation are $s_t \in \mathbb{R}^m$, *i.e.* live in the same space $\mathbb{R}^m$, and are recorded at a constant rate, with no loss of generality. The length of the data set S is $N \in \mathbb{N}$ and is equal to the number of

times the system trajectory is observed.

The measurement noise is mathematically represented by

$$s_t = x_t + \eta_t, \tag{1.9}$$

for $\eta_t \sim IID(0, \sigma_\eta^2)$ and the x_t's, $t > 0$, are the imperfect model states. Comparing equation (1.4) for the PMS and the equation above shows that the difference is that in the IMS, systems states are not available, only imperfect model states.

In this Thesis, model *inadequacy* or impersection is represented by an artificial dynamical noise component in the model, written as

$$x_{t+1} = f(x_t; \boldsymbol{\theta}) + \delta_{t+1}, \tag{1.10}$$

for $\delta_t \in \mathbb{R}^m$, $\delta_t \sim IID(0, \sigma_\delta^2)$ where σ_δ^2 is uknown. δ_t is called artificial dynamical noise (*c.t.* equation (1.5)), and by no means, it does represent any random perturbation physically happening in the system.

In this context, dynamical noise is then interpreted as *model error* rather than as a property of the observations.

3. *The Real Model Scenario* (ReMS):

In this case, the system is complex; $\tilde{f}$, does not exist and there is no model that includes all relevant degrees of freedom for the description

of the system dynamics; Observations are noisy and finite, and all available models are a simplification of the current state of the system.

The ReMS is a special case of the IMS, and it is known that the model used to represent the system is an *ignored-subspace* model [55] since it does not include an unknown and unobserved dynamical component of the system, and involves coarse measured variables or variables representing averages. This scenario is exemplified in Chapter 6 for the grid frequency dynamics.

Two different approaches are used throughout this Thesis in the attempt to estimate model parameter estimation from observations:

A. *The Naive Realistic Approach* (NRA):

Given a system in any of the scenarios described in parts 1 to 3, the model is assumed to be a perfect model describing the system. Differences between system and model are neglected and model error is not taken into account.

B. *The Naive Statistical Approach* (NSA):

Given a system in the PMS, the model is assumed to be stochastic even though the perfect model is known to be a nonlinear deterministic model. Although the stochastic model is inadequate to describ the system, it is used to ease numerical calculation and analysis and to

cope better with some uncertainty sources in a statistical framework.

The dangers of assuming a model class as perfect ignoring the natural difference between system and model are clearly posed by Chatfield [19] in Chapter 3 and Chapter 13, respectively:

> *"... there is a real danger that the analyst will try many different models, pick the one that appears to fit best ... but then make predictions as if certain that the best-fit model is the true model."*

> *"When a model is selected using the data, rather than being specified a priori, the analyst needs to remember that (1) the true model may not have been selected, (2) the model may be changing through time or (3) there may not be a 'true' model anyway. It is indeed strange that we often implicitly admit that there is uncertainty about the underlying model by searching for a 'best-fit' model, but then ignore this uncertainty when making predictions. In fact it can readily be shown that, when the same data are used to formulate and fit a model, as is typically the case in time-series analysis, then least squares theory does not apply. Parameter estimates will typically be biased, often quite substantially. In other words, the properties of an estimator may depend, not only on the selected model but also on the sesection process."*

Chapter 3 of this Thesis explores issues related to parameter estimation in the PMS as defined in part 1 and also it uses intentionally the NSA to obtain estimates of parameters and non-observed variables in order to compare with earlier results shown in [70]. Chapter 4 estimates model parameters for chaotic maps in the PMS from observations that contain only measurement noise as formulated in [68]. Attempts to implement methodologies of parameter estimation in the IMS for the special case of the real model scenario ReMS in part 3, are presented in Chapter 6 and intentionally the NRA is used

to formulate a dynamical model. Once the model is formulated, the NSA is used to obtained parameter estimates from real data sets, assuming as PMS the imperfect model class formulated for the grid frequency dynamics.

The final solution to all parts of the problem, in particular part 2, *i.e.* the imperfect model scenario, where even the existence of "optimal" parameter values is doubtful, is beyond of the scope of this Thesis.

The results presented in this docuement highlight the importance of melding methodologies. The future paths to follow if "good" is defined in dynamical rather than statistical terms, are described. Finally, advances towards obtaining "better" parameter estimates in nonlinear systems are made.

Chapter 2

Background

Assume that a time series of observations of the system of interest is available. The features of the system's temporal evolution are to be used to characterise the system for purposes of forecasting and control monitoring tasks. Looking at the time series of interest, the system under study is represented as a mathematical structure or model. Once this model-system relation is set, the problem of model parameter estimation from time series is understood as a model fitting problem. Uncertainty is present in the observations and in the model chosen to be the representation of the system. Methodologies used to solve this problem should include considerations of uncertainty sources to increase the reliability of the resulting estimates.

Uncertainty plays a key role in the unfolding of the dynamics and in the resulting reliability of the estimates. The aim of this Chapter is to describe

the methods used at every stage of this investigation on how to find parameter estimates for nonlinear models. The Chapter is a list of recipes of the relevant methods for parameter estimation, and discussion of the issues related to the problem of interest is found in the main Chapters of this work.

A statistical approach to this problem is called *inference* or model evaluation. In that context, inference is the process of updating probabilities of outcomes based upon the relationships in the model and the evidence known about the situation at hand [9]. This Chapter presents the statistical methodologies from the Bayesian and the Frequentist perspectives in section 2.1 and 2.2 respectively.

Traditionally, in the nonlinear dynamical perspective, uncertainty in the observations given by noise presence is accounted for by noise reduction methods [58, 16, 36, 26, 92] whilst methods to account for uncertainty in the model are still in development and subject to continuous progress [19, 87, 75, 54, 55, 51]. section 2.3 presents the way indistinguishable states are found for the chaotic Logistic map [67] by means of the gradient descent (GD) algorithm following the work of Judd and Smith [54, 55].

In general terms, there is no method which could be labelled as a "proper" or "correct" approach without assessing the performance of the resulting estimates for a given application. In addition, independently of the methodology used to find parameter estimates, even in simple scenarios, is impossible to

identify the nature of all uncertainty sources. In most cases, there is a trade-offs between the relevant information obtained while reducing uncertainty on the "uncertainty" of the nature of the uncertainty sources themselves. A successful methodology is one that balances such trade-off for a given scenario in the context of a particular application.

2.1 Bayesian Parameter Estimation

Bayesian inference is a statistical approach to estimate and predict a behaviour of interest [95, 70, 12]. In this framework, probabilities are interpreted neither as frequencies, proportions nor likely events. Instead, this approach can be seen as a way to formally model a system in terms of probability distributions. These probability distributions combine "common-sense" knowledge and observational evidence [29].

From the Bayesian point of view, there is no fundamental distinction between variables and parameters in the model used to describe the situation of interest. In the first instance, parameter and model variables are both naively assumed to be random variables, if the model is a dynamical nonlinear system.

A distinction is made, however between *observable* and *non-observable* random variables in the model. An observable random variable is one which

can be measured in the experimental process of observation, *i.e.* it can be replaced by data values. Often the non-observable random variables are called parameters, regardless of being model parameters or model variables.

In order to translate the statement of the parameter estimation problem for a dynamical system to the Bayesian framework, model variables, parameters and observations are classified either as observables or non-observables. The notation defined in Chapter 1 is going to be stretched in the Bayesian framework in a way that all observables are collected in S whilst all non-observables in the Bayesian parameter vector $\boldsymbol{\theta}$.

The statement of the problem in Chapter 1, section 1.2, provides a definition of the system dynamics of interest in equation (1.1), a data set S of noisy observations (see equation (1.4)) and a model to represent the system dynamics given by equation (1.2).

From this statement of the problem and the principles of Bayesian inference, classification as observables and non-observables is made as follows, and it is independent of the scenario the model is placed.

- *Observables*, S:

  The observations $\{s_t\}_{t=1}^{N}$ are assumed to be a realisation of a random variable regardless of the dynamical information contained in each s_t. In addition, the mean and variance of the noise process are observables

constrained to the known values of zero and σ_η^2, respectively. Therefore, the observables in S are all observations $\{s_t\}_{t=1}^N$ and the two known parameters of the noise process, producing a set of observables with $N+2$ elements.

- *Non-observables*, $\boldsymbol{\theta}$:

  After finding the observables contained in S, the rest of model parameters and variables are all included in the Bayesian parameter vector $\boldsymbol{\theta}$. The parameter vector includes the model states $\{x_t\}_{t=1}^N$, the initial condition x_0 and the ℓ model parameters in $f(x_t;\cdot)$, making $\boldsymbol{\theta}$ of dimension $N+\ell+1$.

Note that for particular examples, the parameter vector $\boldsymbol{\theta}$ will also include *hyper-parameters* [95], parameters of the random process associated to a component of $\boldsymbol{\theta}$, which in turn will increase the dimension of the Bayesian parameter vector. Choice of hyper-parameters related to components of $\boldsymbol{\theta}$ is drawn from relevant *background information* on the system to be modelled, and it is denoted as I, following notation in [83].

Bayesian statistical inference requires setting up a joint probability distribution, $p(S,\boldsymbol{\theta}|I)$, of all random variables [95]. The joint probability density function (PDF) can be decomposed into the product

$$p(S,\boldsymbol{\theta}|I) = p(S|\boldsymbol{\theta},I)p(\boldsymbol{\theta}|I), \tag{2.1}$$

where $p(\boldsymbol{\theta}|I)$ is known as the *prior* distribution of all non-observables and $p(S|\boldsymbol{\theta}, I)$ is called the *Likelihood*: a conditional PDF of all observables given the non-observables. As noted before, the prior contains all the information about parameters which is obtained by having knowledge about the situation before observing a data set S. The information coming from the experiment is contained in the Likelihood.

The prior and the Likelihood are updated via Bayes' theorem [4] to a probability distribution of the parameters, given a realisation of the data set S, as follows:

$$p(\boldsymbol{\theta}|S, I) = \frac{p(S|\boldsymbol{\theta}, I)\ p(\boldsymbol{\theta}|I)}{\int p(S|\boldsymbol{\theta}, I)\ p(\boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\theta}} \tag{2.2}$$

where $p(\boldsymbol{\theta}|S, I)$ is called the *posterior* probability distribution. The posterior is the distribution that contains all the samples from the prior that best resemble the data given the data set S [95], and the relevant background information I. The background information I is also referred to as *prior information.*

The denominator in equation (2.2) is a normalisation constant with respect to $\boldsymbol{\theta}$. This distribution is known as the *marginal* distribution, $m(S|I)$, given by

$$m(S|I) = \int p(S|\boldsymbol{\theta}, I)\ p(\boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\theta} \tag{2.3}$$

The explicit inclusion of the background information I as a variable in

each of the PDF involved in equation (2.2) is made to stress the fact that once the background information is different, the functional form of the posterior is changed.

In practice, a major technical difficulty in the implementation of Bayesian methods is the high dimensional integration involved in $m(S|I)$ and in the calculation of any expected value of the posterior distribution. The numerical implementation of Bayesian methods involves sampling algorithms that draw realisations from the posterior distribution. The majority of these algorithms are formulated in terms of non-normalised distributions.

In these terms, it is more convenient to write equation (2.2) as

$$p(\boldsymbol{\theta}|S,I) \propto p(S|\boldsymbol{\theta},I)\, p(\boldsymbol{\theta}|I), \tag{2.4}$$

$$\propto p(S,\boldsymbol{\theta}|I). \tag{2.5}$$

Once a realisation of $S = S'$ is obtained, equation (2.5) is evaluated on S so that the posterior distribution $p(\boldsymbol{\theta}|S,I)$ is a function of $\boldsymbol{\theta}$ only. Thus the posterior distribution $p(\boldsymbol{\theta}|S,I)$ is denoted by $\pi_S(\boldsymbol{\theta}|I)$. This new notation for the posterior emphasises the fact that the posterior distribution (2.5) is different for a different realisation of S and what is considered in a particular case as relevant background information I. The PDF, $\pi_S(\boldsymbol{\theta}|I)$ is known as the *full joint posterior* distribution.

Formally, let $S = \{s_t\}_{t=1}^N$ be the set of observations of the system of in-

terest, where $s_t \in \mathbb{R}^m$ and $N \in \mathbb{Z}$ is the length of data available. From (2.5) it follows that

$$\pi_S(\boldsymbol{\theta}|I) \equiv p(\boldsymbol{\theta}|S = \{s_t\}_{t=1}^N, I), \tag{2.6}$$

is only a function of the parameter vector $\boldsymbol{\theta}$ and the prior information I.

The joint posterior distribution $\pi_S(\boldsymbol{\theta}, I)$ is the object of interest in the Bayesian framework since, in principle, any inference of any parameter in the model could be made from the knowledge of $\pi_S(\boldsymbol{\theta}|I)$. In general, inference on $\boldsymbol{\theta}$ translates into the calculation of expected values of an arbitrary function of $\boldsymbol{\theta}$, $g(\boldsymbol{\theta})$. The expectation of $g(\boldsymbol{\theta})$ is defined by

$$E_{\pi,I}\left[g(\boldsymbol{\theta})\right] = \int g(\boldsymbol{\theta})\ \pi_S(\boldsymbol{\theta}|I)\ \mathrm{d}\boldsymbol{\theta}. \tag{2.7}$$

For future reference, lets introduce some detailed notation for the components of the parameter vector $\boldsymbol{\theta}$. Let $\theta_{.i} \in \mathbb{R}^k$ be the ith component of $\boldsymbol{\theta}$ for $1 \leq k \leq \ell$ and $i = 1, \ldots, \ell$, and let $\theta_{.-i}$ be all components of $\boldsymbol{\theta}$ excluding the ith component $\theta_{.i}$. In general, $\theta_{.i}$ is not scalar, but for simplicity it is assumed to be scalar thus $\theta_{.i} \in \mathbb{R}\ \forall i$, with no loss of generality. For example, in this notation and for $g(\boldsymbol{\theta}) = \theta_{.i}$, equation (2.7) is written as

$$E_{\pi,I}[\theta_{.i}] = \int \theta_{.i}\ \pi_S(\boldsymbol{\theta}|I)\ \mathrm{d}\boldsymbol{\theta} \tag{2.8}$$

which clearly corresponds to the posterior mean of the ith component of $\boldsymbol{\theta}$.

The calculation of such expected values as equation (2.7) involves at least two high dimensional integrals, one to obtain the marginal distribution $m(S|I)$ and one to project $g(\boldsymbol{\theta})$ onto the measure induced by the full posterior $\pi_S(\boldsymbol{\theta}|I)$. Such calculation could not be performed analytically and it becomes one of the main practical difficulties when making inference from a posterior.

In order to address this analytical intractability of the Bayesian formulation, numerical integration of (2.7) is carried out by a Monte Carlo approximation. This approximation involves getting random samples of $\pi_S(\boldsymbol{\theta}|I)$ by suitable sampling algorithms. In particular, samples are taken to be the states of a suitably constructed Markov chain such that $\pi_S(\boldsymbol{\theta}|I)$ is its stationary distribution. The numerical implementation of Bayesian methods using realisations of Markov chains as samples drawn from the full joint posterior distribution (2.6) and Monte Carlo integration to calculate expected values (2.7) are known as Markov Chain Monte Carlo methods, or MCMC in short.

2.1.1 MCMC Techniques

Monte Carlo integration calculates the expectation $E_{\pi,I}[\cdot]$ in (2.7) by drawing samples $\{\boldsymbol{\theta}^{(j)}, j = 1, \ldots, T\}$ from the posterior $\pi_S(\boldsymbol{\theta}|I)$. In turn, posterior

samples are used to evaluate the expectation defined in (2.7) as follows

$$E_{\pi,I}[g(\boldsymbol{\theta})] \approx \frac{1}{T}\sum_{j=1}^{T} g(\boldsymbol{\theta}^{(j)}). \tag{2.9}$$

The set of posterior samples for the parameter vector $\boldsymbol{\theta}$ is denoted by $\{\boldsymbol{\theta}^{(j)}, j = 1, ..., T\}$ and can be generated by any process which draws "sufficiently" independent samples throughout the support of $\pi_S(\cdot|I)$ in the correct proportions. Sufficient independence can be understood to mean that samples of each of the components $\theta_{.i}$ of $\boldsymbol{\theta}$ are independent from each other *i.e.* $p(\theta_{.i}, \theta_{.i'}) \approx p(\theta_{.i}) \times p(\theta_{.i'}), \forall i \neq i'$. Now, sufficient independency is achieved by constructing a Markov chain such that $\pi_S(\boldsymbol{\theta}|I)$ is its stationary distribution.

Under precise regularity conditions (see [29] and references therein) a Markov chain is constructed such that when $T \to \infty$, the following asymptotic results are reached with probability one.

$$\lim_{T\to\infty} \boldsymbol{\theta}^{(T)} \longrightarrow \boldsymbol{\theta} \sim \pi_S(\boldsymbol{\theta}|I) \tag{2.10}$$

and

$$\lim_{T\to\infty} \left[\frac{1}{T}\sum_{j=1}^{T} g(\boldsymbol{\theta}^{(j)})\right] \longrightarrow E_{\pi,I}[g(\boldsymbol{\theta})], \tag{2.11}$$

As such, the averages of chain values are equivalent to estimates of parameters in the limiting distribution π. For detailed discussion of this point see [91, 95, 29].

MCMC is an iterative process in which samples of the components of $\boldsymbol{\theta}$ are obtained from the states of a suitable Markov chain. Each state of the chain is a complete sample of the parameter vector $\boldsymbol{\theta}$ at a given iteration j. Most of the effort is put in the generation of suitable chain states since inference is reduced to calculate geometric averages.

A suitable Markov chain is one whose *transition* probability, $p(\boldsymbol{\theta}^{(j+1)}|\boldsymbol{\theta}^{(j)}, I)$, converges to the joint posterior $\pi_S(\boldsymbol{\theta}|I)$ in the limit $T \rightarrow \infty$[66]. The transition probability is the conditional probability that the current state of the chain $\boldsymbol{\theta}^{(j)}$ becomes the state $\boldsymbol{\theta}^{(j+1)}$. This probability is also known as the *transition kernel* of the chain, and it is denoted by $K(\boldsymbol{\theta}^{(j+1)}|\boldsymbol{\theta}^{(j)})$.

Typically, the Markov chain takes values in $\mathbb{R}^\ell$, since $\boldsymbol{\theta} \in \mathbb{R}^\ell$, and a Markov chain is constructed using the algorithm developed by Hastings [40], which is a generalisation of the method developed by Metropolis *et al.* in 1953 [69], and is known as the *Metropolis-Hastings* (MH) algorithm [40, 69, 30].

The MH algorithm generates a sequence $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{T}$ as follows:

Set initial conditions: $\boldsymbol{\theta}^{(j=0)}$
Loop ($j = 1, \ldots, T$) {
  Sample a candidate state $j+1$: $Y \sim q(\cdot|\boldsymbol{\theta}^{(j)})$
  Sample a uniform random variable: $U \sim \mathcal{U}(0,1)$
    If: $U \leq \alpha(\boldsymbol{\theta}^{(j)}, Y)$ then: $\boldsymbol{\theta}^{(j+1)} = Y$
      otherwise: $\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)}$
  Increment j
}

The algorithm proceeds at each time j by choosing the next state of the chain $\boldsymbol{\theta}^{(j+1)}$ by first sampling a candidate state Y from a *proposal* distribution $q(\cdot|\boldsymbol{\theta}^{(j)})$ which depends on the current state $\boldsymbol{\theta}^{(t+1)}$. The candidate Y is accepted with a probability of acceptance given by

$$\alpha(\boldsymbol{\theta}^{(j)}, Y) = \min\left\{1, \frac{\pi(Y|I)\ q(\boldsymbol{\theta}^{(j)}|Y)}{\pi(\boldsymbol{\theta}^{(j)}|I)\ q(Y|\boldsymbol{\theta}^{(j)})}\right\}. \tag{2.12}$$

The proposal distribution $q(\cdot|\cdot)$ could be in any functional form and the stationary distribution of the chain is $\pi(\cdot|I)$ provided that the transition kernel for the Metropolis-Hastings algorithm satisfies the detailed balance equation

$$\pi(x)\ K(x,y) = \pi(y)\ K(y,x) \tag{2.13}$$

when the chain moves from $x = \boldsymbol{\theta}^{(j)}$ to $y = \boldsymbol{\theta}^{(j+1)}$. The detailed balance equation in (2.13) constrains the rates of moves through states in detail for every possible pair of states. Therefore, once $\boldsymbol{\theta}^{(t)} \sim \pi_S(\theta|I)$ is obtained, it is assured that $\boldsymbol{\theta}^{(t+1)}$ is also sampled from $\pi_S(\theta|I)$. For technical details, see [9, 29, 95].

It is often more convenient and efficient, from the computational point of view, to divide $\boldsymbol{\theta}$ in components $\{\theta_{.1}, \theta_{.2}, \ldots, \theta_{.i}, \ldots, \theta_{.\ell}\}$ and then update these components one at a time. It is not necessary for each component to be scalar. When the parameter vector is divided into components, the updating process used for constructing the appropriate Markov chain is known

as the *single-component Metropolis-Hastings* algorithm and was the original structure proposed by Metropolis [69].

At each time step, the algorithm updates each of the ℓ components of $\boldsymbol{\theta}$ using the Metropolis-Hastings algorithm. In the jth iteration, the state of the chain is updated one by one for each of the ℓ components of $\boldsymbol{\theta}$. Iteration j updates $\theta_{.i}^{(j)}$ for $i = 1, \ldots, \ell$ choosing $Y_{.i}$ from $q_i(Y_{.i}|\theta_{.i}^{(j)}, \theta_{.-i}^{(j)})$ as a candidate for the updated value $\theta_{.i}^{(j+1)}$ *i.e.* the ith component of the next state in the chain. Explicitly, the term $\theta_{.-i}^{(j)}$ is

$$\theta_{.-i}^{(j)} = (\theta_{.1}^{(j+1)}, \ldots, \theta_{.i-1}^{(j+1)}, \theta_{.i+1}^{(j)}, \ldots, \theta_{.\ell}^{(i)}). \tag{2.14}$$

Following (2.12), the candidate is accepted with probability

$$\alpha(\theta_{.i}^{(j)}, \theta_{.-i}^{(j)}, Y_{.i}) = \min\left\{1, \frac{\pi(Y_{.i}|\theta_{.-i}^{(j)}, I)\ q_i(\theta_{.i}^{(j)}|Y_{.i}, \theta_{.-i}^{(j)})}{\pi(\theta_{.i}^{(j)}|\theta_{.-i}^{(j)}, I)\ q_i(Y_{.j}|\theta_{.i}^{(j)}, \theta_{.-i}^{(j)})}\right\}, \tag{2.15}$$

where $\pi(\theta_{.i}^{(j)}|\theta_{.-i}^{(j)}, I)$ are the *full conditional* distributions for $i = 1, \ldots, \ell$.

If $Y_{.i}$ is accepted then $\theta_{.i}^{(j+1)} = Y_{.i}$, otherwise $\theta_{.i}^{(j+1)} = \theta_{.i}^{(j)}$ and the chain does not move. Note that no other component of $\theta^{(j)}$ is changed in step j. For clarity, the definition of the full conditional distributions is presented later in this section.

By analogy with the description of the MH algorithm, the general structure of the single-component algorithm is presented as follows:

Set initial conditions: $\boldsymbol{\theta}^{(j=0)}$
Loop $(j = 1, \ldots, T)$ {
  Set coordinate iterator $i = 1$
  Loop $(i = 1, \ldots, \ell)$ {
    Sample a candidate state $j+1$: $Y_{.i} \sim q_i(Y_{.i}|\theta_{.i}^{(j)}, \theta_{.-i}^{(j)})$
    Sample a uniform random variable: $U \sim \mathcal{U}(0,1)$
    If: $U \leq \alpha(\theta_{.i}^{(j)}, \theta_{.-i}^{(j)}, Y_{.i})$ then: $\theta_{.i}^{(j+1)} = Y_{.i}$
      otherwise: $\theta_{.i}^{(j+1)} = \theta_{.i}^{(j)}$
    Increment i }

Increment j }.

Note that $\theta_{.-i}^{(j)}$ mixes the values of the components previously already updated in the current iteration j (see equation (2.14)) and components updated in the previous iteration $j-1$ when $1 < i \leq \ell - 1$.

Synoptically, given an arbitrary set of starting values $\theta_{.1}^{(0)}, \ldots, \theta_{.\ell}^{(0)}$ the first iteration of the updating process looks as follows:

$$
\begin{aligned}
\text{simulate } \theta_{.1}^{(1)} &\sim \pi(\theta_{.1}|\ \theta_{.2}^{(0)}, \ldots, \theta_{.\ell}^{(0)}, I) \\
\text{simulate } \theta_{.2}^{(1)} &\sim \pi(\theta_{.2}|\ \theta_{.1}^{(1)}, \theta_{.3}^{(0)}, \ldots, \theta_{.\ell}^{(0)}, I) \\
&\vdots \\
\text{simulate } \theta_{.i}^{(1)} &\sim \pi(\theta_{.i}|\ \theta_{.1}^{(1)}, \ldots, \theta_{.i-1}^{(1)}, \theta_{.i+1}^{(0)}, \ldots, \theta_{.\ell}^{(0)}, I) \\
&\vdots \\
\text{simulate } \theta_{.\ell}^{(1)} &\sim \pi(\theta_{.\ell}|\ \theta_{.1}^{(1)}, \ldots, \theta_{.\ell-1}^{(1)}, I)
\end{aligned}
$$

and yields a chain with states $\boldsymbol{\theta}^{(j)} = (\theta_{.1}^{(j)}, \ldots, \theta_{.\ell}^{(j)})$ after j cycles. Consequently, if all the full conditional distributions are available, all that is required is to sample iteratively from them.

The full conditional distribution of $\theta_{.i}^{(j)}$ under $\pi(\cdot, I)$ is defined as

$$\pi(\theta_{.i}|\theta_{.-i}, I) = \frac{\pi_S(\boldsymbol{\theta}|I)}{\int \pi_S(\boldsymbol{\theta}|I)\ \mathrm{d}\theta_{.i}}. \tag{2.16}$$

Note that the normalisation constant in (2.16) is independent of $\theta_{.i}$, since from equation (2.6) it is clear that the posterior distribution is proportional to the joint PDF of all observables and non-observables in the model.

$$\pi(\theta_{.i}|\theta_{.-i}, I) \propto \pi_S(\boldsymbol{\theta}|I) \propto p(S = \{s_t\},\ \boldsymbol{\theta}|I). \tag{2.17}$$

In other words, the full conditional distribution of $\theta_{.i}$, given the values of the other components $\theta_{.-i}$, corresponds to those terms of $\pi_S(\boldsymbol{\theta}|I)$ in which $\theta_{.i}$ appears explicitly. This feature makes full conditional distributions straightforward to calculate. The process becomes even more straightforward when a *conjugate* functional form can be easily identified *i.e.* the full conditional distribution is in a *closed form.*

Let $\mathcal{F}$ be a family of probability distribution functions $f(x|\boldsymbol{\theta})$ (indexed by $\boldsymbol{\theta}$). A class $\mathcal{F}^\star$ of prior distributions $f^\star$ is a *conjugate family* for $\mathcal{F}$ if the posterior distribution is in the class $\mathcal{F}^\star$ for all $f \in \mathcal{F}$, all priors $f^\star \in \mathcal{F}^\star$, and all $x \in \mathcal{X}$ [17, 29]. When this happens, the functional form of the distribution is said to be in a closed form. This class is closed under product, therefore if $f \in \mathcal{F}$, the product $f \cdot f^\star \in \mathcal{F}^\star$.

For example, if the Likelihood distribution belongs to the same conjugate family as the prior, then the resulting posterior will also belong to the same

conjugate family. The role of conjugate families of distributions in the practical implementation of the MCMC methodology is very important and will be clearly visible in the example presented later in section 2.1.2 and in the applications of Bayesian perspectives for the Logistic map in Chapter 3 and for National Grid dynamics in Chapter 6.

One of the most important issues surrounding the implementation of MCMC techniques is the choice of the proposal distribution $q(\cdot)$[95, 91, 61, 80]. For computational efficiency, $q(\cdot)$ should be chosen so as to be easily evaluated and sampled from, and with associated high probability of acceptance given by equation (2.12).

A common way to choose the proposal distribution q is the process called *Gibbs sampler* [32, 30]. The Gibbs sampler is a special case of the single-component MH algorithm, taking as the proposal distribution for the i^{th} component of $\boldsymbol{\theta}$ its corresponding full conditional as defined in (2.16). The candidate for the next step of the chain is drawn from the full conditional. Therefore the MCMC with the Gibbs sampler is

$$q_i(Y_{.i}|\theta_{.i}, \theta_{.-i}) = \pi(Y_{.i}|\theta_{.-i}, I). \tag{2.18}$$

Substituting equation (2.18) into equation (2.12) gives

$$\alpha(\theta_{.i}^{(j)}, \theta_{.-i}^{(j)}, Y_{.i}) = 1, \tag{2.19}$$

i.e. Gibbs sampler candidates are always accepted as the next step in the

chain.

Having chosen the proposal distribution, the next step is to find out how the generated Markov chain is converging towards a stationary distribution, $\pi_S(\boldsymbol{\theta}|I)$. This problem is known as chain *mixing* and is directly related to the mix values of past and present updates of components $\boldsymbol{\theta}$. Depending on the relation between the proposal distribution $q = \pi(\cdot|\cdot)$ and the full posterior $\pi_S(\cdot|I)$, the mixing can happen at different rates. The slower the mixing of the chain, the larger the number of iterations needed to achieve convergency.

For the majority of applications, the complexity of the models makes it impossible to calculate analytically the rate of convergence of a particular transition kernel towards the stationary distribution, and it is therefore necessary to develop numerical tools to check chain convergence (for a review of these tests see [13]).

To monitor convergence where only one realisation of the chain is available, the output of the Monte Carlo calculation using the posterior samples from the MH algorithm iteration are plotted and then the number of iterations of the algorithm when the mixing appears to be achieved is chosen by eye [95]. When parallel runs are available leading to several realisations of the same chain, the *Gelman and Rubin* (GR) *statistic* [31] can be used to assess convergency.

Let τ be the *burn-in* time when the mixing is finished. The chain $\{\boldsymbol{\theta}^{(j)}, j =$

$\tau + 1, ..., T\}$ contains all $T - \tau - 1$ samples from the full posterior $\pi_S(\boldsymbol{\theta}|I)$. Thus the estimator in equation (2.9) is given by

$$E_{\pi,I}[g(\boldsymbol{\theta})] \approx \frac{1}{T-\tau} \sum_{j=\tau+1}^{T} g(\boldsymbol{\theta}^{(j)}). \tag{2.20}$$

In practice, τ should be large enough to obtain consistent estimates. The size of τ is limited by computational resources. Evaluations of the full conditional distributions make the numerical implementation costly in terms of running time and computational resources, which grow as the complexity of the model increases.

Note that the conditioning of the full conditional distributions changes at each iteration (since $\theta_{\cdot -i}^{(j)}$ changes), each full conditional distribution $\pi_i(\cdot|\theta_{\cdot -i}^{(j)}, I)$ being used only once per iteration j and for each component t. Generating random samples for such distributions is, in general, time consuming, even more so when analytical reduction to a closed form for the full conditional distribution is not possible.

When the full conditional distribution is in a closed form, standard algorithms should be used to generate random samples. When this is not the case, the *Accept/Reject Algorithm* is used for sampling from a general density $f_Y(y)$. This algorithm is presented in [17] as the theorem 3.2.1 in section 3.2.1.4.

The following section will exemplify the Bayesian perspectives in a sim-

ple example presented in Chapter 5 of [95]. The simple structure of this simple example gives insight on the implementation of the single-component Metropolis-Hastings algorithm, as well as on the definition, calculation and properties of the full conditional distributions, including the convergence of the transition kernel of the chain to the joint posterior distribution $\pi_S(\boldsymbol{\theta}|I)$.

2.1.2 Simple Example

Assume there is access to a data set of observations corresponding to realisations of a random variable Y. In order to infer any moment of Y, requires setting up a probability model to represent it. Let $y = \{y_1, .., y_N\}$ be N realisations of a random variable Y. Based on I, the background expert knowledge of the process associated with Y, and before any realisation of the process is observed, the y_t's are chosen to be normally distributed with unknown mean μ and variance σ^2

$$y_t \sim \mathcal{N}(\mu, \sigma^2), \quad t = 1, \ldots, N. \tag{2.21}$$

where $\{y_t\}_{t=1}^N$ are conditionally independent given μ and σ^2.

For unknown parameters, no information is available, so the following *non-informative* priors are set reflecting I,

$$\mu \sim \mathcal{N}(0, 1), \tag{2.22}$$

$$\frac{1}{\sigma^2} \sim \mathcal{G}a(2.01, 1), \tag{2.23}$$

where independency between μ and σ^2 is assumed and $\mathcal{G}a(\alpha, \beta)$ is the generic notation for a *Gamma* distribution with mean α/β and variance α/β^2. From (2.23) is clear that σ^2 follows an *Inverted Gamma* distribution.

Equations (2.21) to (2.23) form a two-parameter Bayesian model. This model consists of one observable y and two non-observables μ and τ. Following the notation introduced in last section, the parameter vector is $\boldsymbol{\theta} = (\theta_{.1}, \theta_{.2}) = (\mu, \tau)$.

From equations (2.4) and (2.5), the joint distribution of y, μ, and τ is given by

$$p(y, \mu, \sigma^2|I) = \prod_{t=1}^{N} p(y_t|\mu, \sigma^2, I)\, p(\mu|I)\, p(\sigma^2|I). \tag{2.24}$$

Substituting equations (2.21) to (2.23) into (2.24),

$$\begin{aligned} p(y, \mu, \sigma^2) &= \left[\prod_{t=1}^{N} \mathcal{N}(\mu, \sigma^2)\right] \times \mathcal{N}(0, 1) \times \mathcal{G}a(2.01, 1), \\ &= \prod_{t=1}^{N} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_t - \mu)^2}{2\sigma^2}\right\}\right] \times \\ &\quad \frac{1}{\sqrt{2\pi}} e^{-\mu^2/2} \times \frac{1}{\Gamma(2.01)\sigma^2} e^{-1/\sigma^2} \end{aligned} \tag{2.25}$$

is obtained.

If more relevant information about $\boldsymbol{\theta}$ were available, equation (2.25) had looked different since the different prior choices. The background information is translated into the priors and Likelihood terms by "relevant" probability distributions reflecting such information [83]. At this point, were the func-

tional form of the posterior is written, the explicit notation of the dependency of equation (2.24) on I is dropped for clarity.

Once the data set $S = \{y_t\}_{t=1}^{N}$ is observed, (2.25) is evaluated on the realisation of Y and from equation (2.6) it follows that the functional form of the joint posterior distribution $\pi_S(\boldsymbol{\theta})$ is proportional to

$$\pi_S(\boldsymbol{\theta}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{N}{2}+1} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y_t-\mu)^2 - \frac{\mu^2}{2} - \frac{1}{\sigma^2}\right\}. \tag{2.26}$$

As pointed out earlier, to construct the full conditional distribution associated with the parameter $\theta_{.i}$ one only needs to pick out the terms in (2.26), where $\theta_{.i}$ appears.

Choosing the correspondingly appropriate terms in (2.26), for this two-parameter Bayesian model the full conditional distribution for the mean parameter is

$$\pi(\mu|\sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y_t-\mu)^2 - \frac{\mu^2}{2}\right\}, \tag{2.27}$$

whilst for the variance parameter is

$$\pi(\sigma^2|\mu) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{N}{2}+1} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y_t-\mu)^2 - \frac{1}{\sigma^2}\right\}. \tag{2.28}$$

Rewriting equations (2.27) and (2.28), it is easily found that the full conditional distributions for μ and σ^2 are proportional to Normal and Inverted

Gamma distributions respectively. These full conditionals are then given by

$$\begin{aligned} \pi(\mu|\sigma^2) &\propto \exp\left\{-\frac{\sigma^2+N}{2\sigma^2}\left(\mu-\frac{\sum_{t=1}^N y_t}{\sigma^2+N}\right)\right\}, \\ \mu &\sim \mathcal{N}\left(\frac{\sum_{t=1}^N y_t}{\sigma^2+N}, \frac{\sigma^2}{\sigma^2+N}\right) \end{aligned} \tag{2.29}$$

and

$$\begin{aligned} \pi(\sigma^2|\mu) &\propto \left(\frac{1}{\sigma^2}\right)^{1+\frac{N}{2}} \exp\left\{-\frac{1}{\sigma^2}\left[1+\frac{1}{2}\sum_{t=1}^N (y_t-\mu)^2\right]\right\}, \\ \frac{1}{\sigma^2} &\sim \mathcal{G}a\left(2+\frac{N}{2}, 1+\frac{1}{2}\sum_{t=1}^N (y_t-\mu)^2\right). \end{aligned} \tag{2.30}$$

Note that the prior distributions are in the conjugate family of the Likelihood $p(y|\mu,\sigma^2)$ in (2.21), and therefore the full conditionals are reduced to a closed form. According to (2.18) the proposal distributions, q, are chosen as the full conditionals (2.29) and (2.30).

To generate the first state of the Markov chain given an arbitrary initial condition $\boldsymbol{\theta}^{(0)} = (\mu^{(0)}, \tau^{(0)})$

$$\begin{aligned} \text{simulate } \mu^{(1)} &\sim \mathcal{N}\left(\frac{\sum_{t=1}^N y_t}{\sigma^{2(0)}+N}, \frac{\sigma^{2(0)}}{\sigma^{2(0)}+N}\right) \\ \text{simulate } \tau^{(1)} &\sim \mathcal{G}a\left(2+\frac{N}{2}, 1+\frac{1}{2}\sum_{t=1}^N (y_t-\mu^{(1)})^2\right), \end{aligned} \tag{2.31}$$

using standard sampling algorithms for Normal and Gamma densities. After T iterations of both simulations in (2.31), a Markov chain $\{\boldsymbol{\theta}\}_{j=1}^T = (\mu^{(T)}, \tau^{(T)})$ values from which any inference can be made. From the equa-

tions in (2.31) the single-component MH algorithm described earlier is easily implemented.

Three parallel runs of the single-component MH algorithm using the Gibbs sampler were performed for $T = 1.1 \times 10^5$ iterations. A set of N=100 observations, randomly distributed as

$$y_t \overset{iid}{\sim} \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2) \tag{2.32}$$

for $t = 1, \ldots, 100$ and true parameter values $\tilde{\boldsymbol{\theta}} = (\tilde{\mu}, \tilde{\sigma^2}) = (1.2029, 27.4045)$ were observed. Each run generates a Markov chain for the parameter vector $\boldsymbol{\theta} = (\theta_{.1}, \theta_{.2})$.

Once the chain is obtained, the convergency and mixing have to be assessed by choosing the burn-in time τ where the samples drawn from the full posterior distribution are to be considered as samples of the posterior. Typically, convergency is physically assessed by plotting the output of a Monte Carlo approximation of any summary statistic, taking the Markov chain states obtained as samples, when only one chain is available. Plotting the Monte Carlo approximation of a summary statistic is the simplest way to check for convergence and mixing, *i.e.* the parameter space is explored with the support of the full posterior.

Given that the MCMC algorithm was run three times, three chains were obtained and the "eye" tests for convergency were applied for all chains.

Figure 2.1 plots the estimated mean (solid line) for $\theta_{.1} = \mu$ and $\theta_{.2} = \sigma^2$ in the left and right panels, respectively, as function of the iteration time. The estimated mean $\pm$ the standard deviation of the distribution are plotted as the envelopes of the Monte Carlo mean estimate. For clarity, only the first 100 iterations of the MCMC algorithm are shown. In this Figure it is clear how the initial posterior samples obtained by sampling the full conditional distributions in (2.29) and (2.30) start converging to a stable state. As more iterations pass, the non-informative priors, assigned for the components of $\boldsymbol{\theta}$, are sharpening toward a stationary distribution.

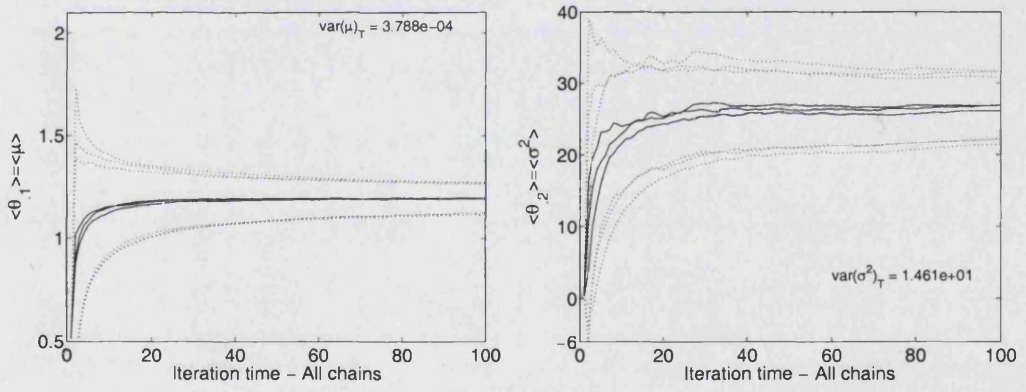


Figure 2.1: The Monte Carlo mean is plotted for each component of the parameter vector $\boldsymbol{\theta}$ as function of the iteration time, for $T = 1, \ldots, 100$. Left and right panels show the Monte Carlo mean for $\theta_{.1} = \mu$ and $\theta_{.1} = \sigma^2$, respectively for all three chains. The envelopes of the mean correspond to the Monte Carlo mean $\pm$ the estimated standard deviation of the resulting distributions.

From the initial prior setting in equations (2.22) and (2.23), the prior

variance is taken to be $\text{var}(\mu)^{(0)} = 1.00$ and $\text{var}(\sigma^2)^{(0)} = 9.80 \times 10^1$, whereas the posterior variance tends to be reduced to $\text{var}(\mu) = 3.80 \times 10^{-4}$ and $\text{var}(\sigma^2) = 1.46 \times 10^1$ as $T \gg 1$. Although it is clear that the uncertainty in the parameters μ and σ^2 is shrinking, the point at which the mixing is complete is not clearly identified if only one chain is available.

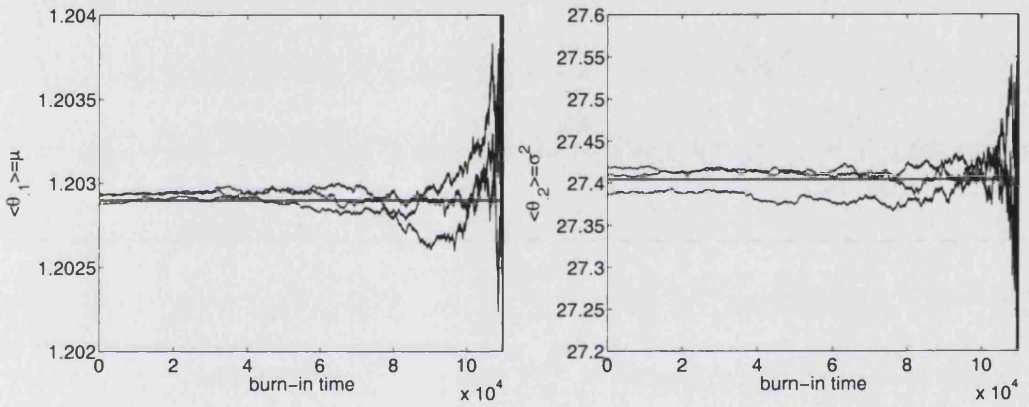


Figure 2.2: Monte Carlo mean for the components of $\boldsymbol{\theta}$ as function of the burn-in time. Estimated values for $\theta_{.1} = \mu$ and $\theta_{.1} = \sigma^2$ are plotted in the left and right panels, respectively. Each trace in a panel corresponds to estimated values from one chain. The horizontal grey line locates the true parameter value.

In order to detect the stability of the chain(s), Figure 2.2 shows the Monte Carlo approximated mean for μ and σ^2 from left to right as a function of the number of samples at the end of the chain used to calculated the estimations, *i.e.* burn-in time. The $x-$axis is the burn-in time and take values ranging 0 to T and is the number of samples at the beginning of the chain that have been neglected when calculating estimates. Thus for example, for a

burn-in time equal to 10 iterations, the mean for the last $T - 10$ samples states of the chain is plotted. The smaller the sample size, the bigger the burn-in time and therefore the variation in the estimated mean. Each trace for this running mean corresponds in each panel to the estimates of a chain for a given parameter, a horizontal grey solid line locates the true value of the parameter $\theta_{.i}$. Estimation for all three chains seem to stabilise after few iterations for both parameter components. In addition, for $\theta_{.2} = \sigma^2$ one of the chains tends to be slightly down shifted from the true values but only with an error of 1×10^{-2}. In all cases, τ could be set for a value of less than 2000 iterations. To refine further the value of the burn-in time, zooms of the samples and the Monte Carlo estimates should be made. In the typical case where only one chain is available, burn-in time estimation is made from plots like Figure 2.2 but looking only at one chain in isolation. Therefore, convergency assessment by eye can provide unreliable estimates of burn-in times, which may in turn affect the posterior estimates.

The theory described in the last section, assured that the chain generated by the Metropolis-Hastings algorithm in conjuction with the Gibbs sampler will reach the required target distribution, *i.e.* the posterior distribution but it does not give any details on when this will happen. Theoretical results are obtained asymptoticaly and it is a fact that the simulation of the chain cannot be run for infinite times. Henceforth, *convergency diagnostics* is a key part

of the MCMC techniques. Any test that is used to diagnose convergency provides a final conclusion when the chain has not reach convergence to the target distribution but conclusions are always ambigouos for complete convergence. It is common that the chain may appear that has reached convergence however there is always the possibility that the chain is actually trapped for a finite time in a region or mode of the posterior rather than properly exploring the parameter space [31, 13].

In the MCMC literature, there are available several quantitative tests to diagnose convergency and mixing for a given chain or sets of chains. The more prominent tests used in several packages and software available for implementation of MCMC techniques include: Gelman and Rubin (GR) statistic [31], Geweke time series test [33], Heildelberger and Welch test [44], and the Raftery and Lewis test [78] among others.

In order to quantify the convergence of the algorithm, the GR statistic [31, 95] is used throughout this Thesis to assess convergency in a more quantitative way. If only one chain is available, and is long enough, the GR statistic can be calculated by fragmenting the chain into segments and consider each of the segments as a chain itself [31, 95, 13].

The GR statistic is based on the idea that the best way to identify non-convergency [31, 13] is the simulation of multiple sequences for distinct and overdispersed starting points that can be generated in several ways [31, 2, 49].

These dispersed points are used as initial conditions for the several chains to be generated. Given that the Markov chain is built in such a way that asymptotically the chain states are a sample from the target distribution, chains starting from different initial conditions intuitively should show the same behaviour. The variance within the chains should be the same as the variance across the chains [95] for all scalar summaries of interest from the resulting empirical distributions.

The GR statistic is defined for a single summary statistic of interest. Let ξ be the summary statistic, *e.g.* the sample mean, median, etc. It is assumed there are available c parallel simulations of the same chain, each of length T. For a single summary statistic ξ, it is denoted ξ_{uv} as the summary statistic up to the v^{th} iteration time in the u^{th} chain, for $u = 1, \ldots, c$ and $v = 1, \ldots, T$.

For the c parallel sequences of length T of the same summary statistic, both the *between-sequence* variance B and the *within-sequence* variance W are calculated.

The between-sequence variance B is defined as

$$B = \frac{T}{c-1} \sum_{u=1}^{c} (\bar{\xi}_{u.} - \bar{\xi}_{..})^2, \tag{2.33}$$

where

$$\bar{\xi}_{u.} = \frac{1}{T} \sum_{v=1}^{T} \xi_{uv} \tag{2.34}$$

is the mean of the summary statistic within the u^{th} chain, and

$$\bar{\xi}_{..} = \frac{1}{c}\sum_{u=1}^{c}\bar{\xi}_{u.} \tag{2.35}$$

is the average of the summary statistic accross all parallel chains. Note that the between-sequence variance in equation (2.33) contains a factor of T since it is defined in terms of the within-sequence means, $\bar{\xi}_{u.}$, of equation (2.34), the averages of T values ξ_{uv}.

In the other hand, the within-sequence variance W is defined as:

$$W = \frac{1}{c}\sum \varsigma_u^2, \tag{2.36}$$

where

$$\varsigma_u^2 = \frac{1}{T-1}\sum_{v=1}^{T}(\xi_{uv} - \bar{\xi}_{u.})^2, \tag{2.37}$$

which represents the estimate of the average dispersion of the summary statistic ξ within a sequence u.

Once B and W are calculated from equations (2.33) and (2.36) respectively, two estimates of the variance of summary statistic ξ in the target distribution are constructed following [31]. These two estimates correspond to the upper and lower bounds of the variance of the summary statistic, $\text{var}(\xi)$, and are denoted by $\widehat{\text{var}}(\xi)$ and the already defined W.

Using equations (2.33) and (2.36), the upper bound of the variance is defined by:

$$\widehat{\text{var}}(\xi) = \frac{T-1}{T}W + \frac{1}{T}B, \tag{2.38}$$

and is an *unbiased* under stationarity conditions, *i.e.* the initial conditions of the each parallel chains were actually drawn from the target distribution. Equation (2.38) is an *overestimate* of the variance of ξ since the initial conditions of the chains are overdispersed relative to the target distribution.

In the other hand, the across-sequence variance W in equation (2.36) is the lower bound of $\text{var}(\xi)$ since it is an *underestimate.* For finite iteration times T the individual chains have not had time to explore the parameter space under the support of the target distribution, having less variability.

Convergency is reached for $T \rightarrow \infty$ when $W \rightarrow \text{var}(\xi) \leftarrow \widehat{\text{var}}(\xi)$. For finite iteration time, T, equation (2.38) is unbiased, and if $B = W$ then convergence is assumed to be reached.

Gelman and Rubin [31] proposed to monitor the convergence of the Markov chain by monitoring the shinking factor of the upper bound for the variance of ξ, $\widehat{\text{var}}(\xi)$. This ratio is the well known and used Gelman-Rubin statistic, GR, the ratio between the upper and lower bounds of the standard deviation of ξ. Then GR statistic is defined as:

$$GR = \sqrt{\frac{\widehat{\text{var}}(\xi)}{W}}, \tag{2.39}$$

and calculated as a function of the iteration time. Replacing the estimated upper and lower bounds on the variance of the summary statistic, $\widehat{\text{var}}(\xi)$, given by equations (2.38) and (2.36) into equation (2.39), the GR statistic is

written as:

$$GR = \sqrt{1 - \frac{1}{T}\left(1 - \frac{B}{W}\right)}, \tag{2.40}$$

and asymtotic results can be considered.

As $T \rightarrow \infty$ the GR value tends to 1, therefore the scale reduction between both bounds reduces to 1, *i.e.* the Markov chains are overlapping, the within and across-sequence variance are equal ($B = W$) and convergence is improving.

For finite T there are three possible behaviours for the values of the GR statistic:

If $B = W$ then GR= 1, convergence is reached, *i.e.* parallel chain sequences are overlapping.

If $B/W < 1$ then GR$\approx$ 1 for finite $T \gg 1$ and convergence can be assumed to be reached.

If $B/W > 1$ then GR$>$ 1 since the within-sequence variance is larger than the across-sequence variance. This case indicates poor convergence for some of the chains where the initial conditions are trapped in a local region rather than exploring the full posterior. More iterations of the algorithm are required to let the samples explore the parameter space.

In practice, values of GR close to 1 do not secure convergence of the samples to the target distribution, and it can be that by chance B/W is

about 1 [13] eventually in highly dimensional and complex models and GR statistic should be monitor graphically along with the within and accross-sequence variance B and W. Eventhough convergency diagnosis is developing constantly, complete convergence is impossible to assess, leaving mist in the air, states of the Markov chain are considered to be samples of the posterior when convergence is not rejected any of the tests used.

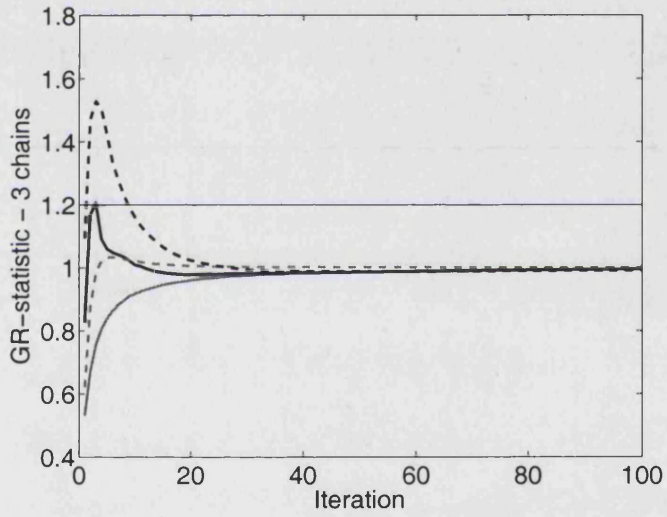


Figure 2.3: GR-statistic for the estimated mean and variance of each component of $\boldsymbol{\theta}$ as a function of the iteration time. The horizontal line locates the pass mark of the test. Solid lines are used for median estimates whilst dashed lines for variance estimates. Grey colour corresponds to $\theta_{.1} = \mu$ and black colour to $\theta_{.1} = \sigma^2$.

In this example, three chains are available and GR values for two summary statistics: the median and the variance of the resulting MCMC samples for $\boldsymbol{\theta}$ as a function of the iteration time. Figure 2.3 shows the trace of the GR

statistic for $\theta_{.1} = \mu$ in grey colour and in black for $\theta_{.2} = \sigma^2$. Median and variance estimations are plotted using solid and dashed lines, respectively. Convergency is reached, in the GR sense, when $GR < 1.2$ [95] and this pass mark corresponds to the horizontal line in the Figure. For all summary statistics and parameter vector components, the GR statistic tends to 1 after approximately 50 iterations of the MCMC routine, and for $T \gg 1$ all GR values tend to one. From this result, together with the analysis of Figures 2.1 and 2.2, the burn-in time is chosen to be $\tau = 50$. The number of chain states taken to be samples from the full posterior (2.25) is 2000, taking the iterations for $T = \tau + 1, \ldots, 2000 + \tau$ for the Monte Carlo approximation.

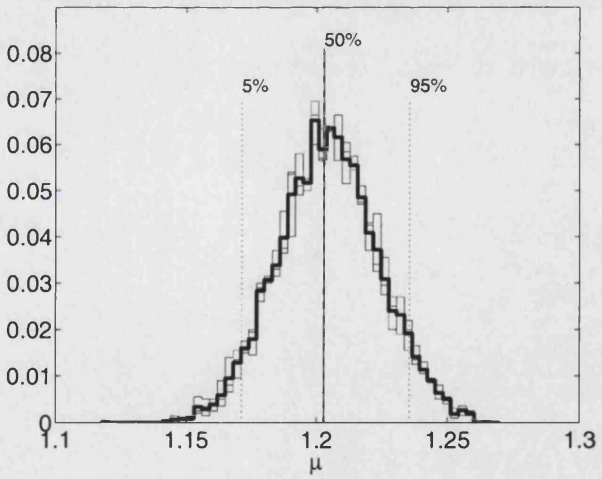


Figure 2.4: Histogram for the samples of $\theta_{.1} = \mu$ collected form all tree MCMC chains in solid black line. Lighter histograms in the background correspond to each of the chains in isolation. The vertical lines locate the 5%, 50% and 95% isopleths for the resulting distribution.

From these 2000 samples, any inference of $\theta_{.1} = \mu$ and $\theta_{.2} = \sigma^2$ can be made. Figures 2.4 and 2.5 shows the histograms for the resulting samples from the posterior of (2.25), obtained by MCMC for $\theta_{.1} = \mu$ and $\theta_{.2} = \sigma^2$, respectively. The solid black line is the histogram of samples collected from all three runs for $\tau+1 \leq T \leq 2000+\tau$, and it is composed of 2000×3 samples. The background histograms correspond to each of the chains separately. The true parameter value is marked with a vertical solid line. In addition, the 5%, 50%, and 95% are shown as vertical dotted and dashed lines.

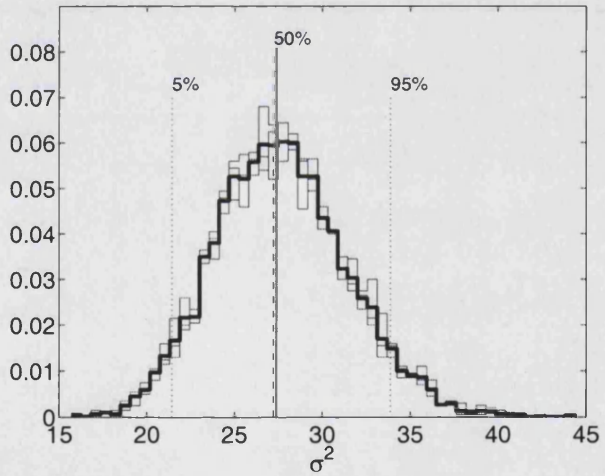


Figure 2.5: Histogram for the samples of $\theta_{.2} = \sigma^2$ collected form all tree MCMC chains in solid black line. Lighter histograms in the background correspond to each of the chains in isolation. The vertical lines locate the 5%, 50% and 95% isopleths for the resulting distribution.

The inferences of the parameter vector from the MCMC samples are summarised in Table 2.1 where the last row shows that error between the

$\boldsymbol{\theta}$	μ	σ^2
$\tilde{\boldsymbol{\theta}}$	1.2027	27.4045
$\langle\boldsymbol{\theta}\rangle$	1.2029	27.3754
median($\boldsymbol{\theta}$)	1.2028	27.2265
5% Isopleth	1.1711	21.4185
95% Isopleth	1.2354	33.8735
var($\boldsymbol{\theta}$)	3.74×10^{-4}	1.43×10^{1}
$\left\lvert\langle\boldsymbol{\theta}\rangle - \tilde{\boldsymbol{\theta}}\right\rvert$	2.70×10^{-4}	2.90×10^{-2}

Table 2.1: Inferences for the components of $\boldsymbol{\theta}$ from 6000 MCMC samples.

estimated mean and the true value of $\theta_{,i}$ is small enough to consider the estimates reliable.

The procedure presented here for this simple example is also applied in the applications described in Chapters 3 and 6 of the Logistic map and in the simplified model for the electrical grid frequency dynamics.

2.2 Maximum Likelihood Parameter Estimation

Maximum Likelihood parameter estimation technique is widely used for model fitting from a Frequentist approach. It can be regarded as a way of quantifying the "common-sense" idea that some sets of parameters will result in model traces that resemble the dynamics contained in the data more than others.

The Maximum Likelihood technique is employed in the PMS as follows. Given a set of observations S of the system of interest, a model that corresponds exactly to that system and the model's parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{\ell}$. For a large subset of $\mathbb{R}^{\ell}$, traces of the model can be obtained by using each point in the subset as a model parameter value. In this instance, it makes sense to think that some parameter values are more likely to generate model traces that closely match the data than others. This notion is quantified by using probabilities and asking the question:

Given a particular value of the model parameters $\boldsymbol{\theta}$,

what is the probability that S *may occur?*

The probability in question is then the conditional probability of the parameter $\boldsymbol{\theta}$ given a particular set of observations S, denoted by $p(S|\boldsymbol{\theta})$. This probability is identified with the Likelihood $L(\boldsymbol{\theta}|S)$ of a particular data

set S given a set of parameters $\boldsymbol{\theta}$. With this identification, the best estimate for the parameter $\boldsymbol{\theta}$ given the observations is the one that maximises the Likelihood $L(\boldsymbol{\theta}|S)$.

Assuming that there is a data set, $S = \{s_t\}_{t=1}^N$, of IID observations known to be normally distributed with unknown mean μ and variance σ^2. The probability of the data set S given the parameters μ and σ^2, can be written as:

$$\begin{aligned} P(\{s_t\}_{t=1}^N|\mu,\sigma^2) &= \prod_{t=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(s_t-\mu)^2}{2\sigma^2}\right\}, \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^N (s_t-\mu)^2\right\}, \quad (2.41) \end{aligned}$$

where the parameter vector is $\boldsymbol{\theta} = (\mu, \sigma^2)$. In turn, the PDF of equation (2.41) is identified with the Likelihood function $L(\boldsymbol{\theta}|\{s_t\}_{t=1}^N)$.

Maximisation of the Likelihood function identified with the PDF in equation (2.41) implies the minimisation of the argument of the exponential function. In other words, finding the maximum of the Likelihood is equivalent to finding the minimum of the negative log-Likelihood of equation (2.41), also known as the *cost function*. The cost function or log-Likelihood is then written as:

$$C_S(\mu,\sigma) = N\log(\sigma^2) + \frac{1}{\sigma^2}\sum_{t=1}^N (s_t-\mu)^2 . \quad (2.42)$$

The uncertainty in the estimates of $\boldsymbol{\theta}$ when the minimum of (2.42) is found is presented as error bars in the cases where the length of S is large

enough or several realisations of S are available [17].

In the case where the variance σ^2 is known or somehow given, equation (2.42) is only a function of μ and the Normal Likelihood is maximise for the value of μ that minimises $C_S(\mu)$.

In general, likely parameter values for a fixed set of observations S can be found using maximum Likelihood techniques by designing a cost function that measures the agreement between the data and the model, given a particular set of model parameter values. Depending on the context of the model parameter estimation problem, the definition of "agreement" can change drastically.

Widely used cost functions to estimate parameters in deterministic nonlinear models are the Least Squares and Total Least Squares cost functions. In this document these are regarded as special cases of the maximum Likelihood method and are presented in detail in Chapter 4, section 4.1.

2.3 Indistinguishable States

This section is devoted to describing briefly the methodology used to obtain indistinguishable states [54, 55] of a dynamical system by a gradient descent (GD) algorithm. Indistinguishable states are found for a set of noisy observations of the system of interest. These states are considered to be

indistinguishable from the noise reduction properties of the GD algorithm. It is guaranteed that a trajectory that shadows the true trajectory of the system is obtained.

Despite the fact that finding indistinguishable states is not by itself a methodology regarded as related to parameter estimation, Chapter 4 uses this idea to make advances into the solution of the problem of parameter estimation in nonlinear systems. Here, the problem is placed in the perfect model scenario (PMS) where the reality of the system under study matches exactly the model chosen to represent it. See Chapter 1 in section 1.2 for the definition of the PMS.

The ideas related to the indistinguishable state theory are closely related to noise reduction, state estimation and ι-shadowing but they differ from the approach presented in Chapter 4 in both technical details and motivation.

Reproducing equation (1.2), let $x_t \in \mathbb{R}^m$ be the system's state variable which evolves by the map

$$x_{t+1} = f(x_t; \boldsymbol{\theta}), \tag{2.43}$$

where the map is $f : \mathbb{R}^m \to \mathbb{R}^m$ and the model parameters are comprised in a vector $\boldsymbol{\theta} \in \mathbb{R}^\ell$. Here, $\boldsymbol{\theta}$ is considered to be known and fixed and attention is focused on the true trajectory of the system generated by it.

The noisy observations are considered only to contain a measurement

noise (*i.e.* additive) component. Each of the N observations available are given by

$$s_t = x_t + \eta_t, \tag{2.44}$$

for $t = 1, \ldots, N$, $s_t \in \mathbb{R}^m$, and each random perturbation η_t is assumed to be IID Normal random variable, $\eta_t \sim \mathcal{N}(O, \sigma_\eta^2)$, with no loss of generality.

Following the work of Judd and Smith in [54] regarding the calculation of indistinguishable states in the PMS, given a known or unknown noise model and a set of observations, an ensemble of indistinguishable states may be found. Such an ensemble is formed by states belonging to the maximum Likelihood trajectory. Following (2.43), a Maximum Likelihood trajectory is one that belongs to the set of all possible states y_N indistinguishable from x_N given the entire history of observations S, where x_N and y_N are system states

The set of all possible indistinguishable states is defined as

$$H(x_N) = \left\{ y_N \in \mathbb{R}^m : \sum_{t=1}^{N} h(y_t - x_t) < \infty \right\}. \tag{2.45}$$

The state y_N is indistinguishable from x_N when it belongs to $H(x_N)$. The definition of the set of indistinguishable states in equation (2.45) is constructed from the Likelihood of y_t and x_t being indistinguishable given all observations in S and is written as

$$h(y_t - x_t) \;=\; -\log q(y_t - x_t). \tag{2.46}$$

The Likelihood of y_t and x_t being indistinguishable can be regarded as the information gained when an observation is made at time t. In turn, $q(y_t - x_t)$ is the distribution of the distances between the states x_t and y_t and it is proportional to the joint probability density of the two being indistinguishable from each other, explicitly given by

$$q(y_t - x_t) \propto \int \rho(s_t - x_t)\rho(s_t - y_t)\mathrm{d}s_t, \tag{2.47}$$

where $\rho(\cdot)$ is a generic PDF, and the normalisation constant in (2.47) is calculated when $y_t = x_t$ and it is given by $\int[\rho(s_t)]^2\mathrm{d}s_t$.

From Figure 2.6 (reproduced from Figure 1 in [54]), the set of indistinguishable states for a set of observations with bounded noise are all points in the overlap (shaded area). For unbounded noise models and typical nonlinear systems, $H(x_N)$ is non-trivial and is a subset of the unstable set of x_N as showed in [54]. In general, a set of indistinguishable states can be found for any $t = 1, \ldots, N$.

An ensemble of such states is found by the minimisation of the mismatch between the state estimate u_t and the one-step forecast $f(u_t; \boldsymbol{\theta})$. Formally, for the perfect model defined in (2.43) and a set of observations S, there exists a *pseudo-state* $u_t \in \mathbb{R}^m$ such that

$$u_{t+1} - f(u_t; \boldsymbol{\theta}) = 0. \tag{2.48}$$

For a time series of length N, (2.48) conforms a set of equations for the

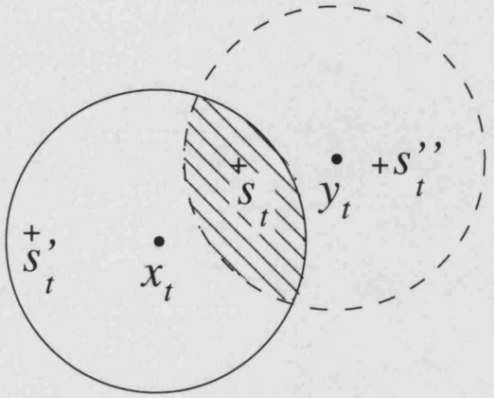


Figure 2.6: Two states of a system trajectory are indistinguishable from each other at time t given an bounded noise model density when an observation s_t falls into the shaded area. If an observation falls outside of the overlap at time t, such observation will distinguish x_t from y_t.

unknown *pseudo-orbit* $u = (u_1, \ldots, u_N)$ in a sequence space of dimension $\mathbb{R}^{N \times m}$. For a finite set of observations, $S = \{s_t\}_{t=1}^{N}$, define the mismatch cost function as

$$C_{MM}(u) = \frac{1}{N-1} \sum_{t=1}^{N-1} (u_{t+1} - f(u_t; \boldsymbol{\theta}))^2 . \tag{2.49}$$

One solution of the system of equations in (2.48) can be given by attempting to find the minimum of (2.49). As shown in theorem 2 of [54] and in [79], $C_{MM}(u)$ has no local minimum other than where $C_{MM}(u) = 0$, *i.e.* when u is a deterministic trajectory. This implies that finding the solution of $\min_u(C_{MM}(u))$ is equivalent to finding the minimum of the mismatch by gradient descent as follows.

2.3.1 Gradient Descent Algorithm

Finding the minimum of the mismatch is equivalent to solving

$$\dot{u} = -\nabla C(u), \tag{2.50}$$

where $C = C_{MM}(u)$, with the initial condition for $u^{(0)} = S$. Note that the initial condition $u^{(0)}$ refers to the place where the iteration of the gradient descent algorithm starts, so it is iteration time rather that the system's evolution time.

Differentiating the mismatch cost function at each time gives the explicit recurrence relation which will generate a pseudo-orbit $z^{(j)}$ at the jth iteration of the algorithm. Note that the $\{z_t\}$'s are the $\{u_t\}$'s that minimise equation (2.50) by Gradient Descent methods. Explicitly,

$$\frac{\partial C}{\partial z} = \frac{2}{N+1} \times \begin{cases} -(z_{t+1} - f(z_t))\ \mathrm{d}_1 f(z_t), & t = 1 \\ -(z_t - f(z_{t-1})) + (z_{t+1} - f(z_t))\ \mathrm{d}_t f(z_t), & 1 \le t \le N-1 \\ -(z_t - f(z_{t-1})), & t = N \end{cases},$$

where $\mathrm{d}_t f(z_t)$ is the transpose of the Jacobian derivative of the map f evaluated at z_t. Solving (2.50) by the Euler approximation, each iteration of GD produces a point in the sequence space $\mathbb{R}^{N\times m}$ from iteratively calculating

$$z_t^{(j+1)} \longmapsto z_t^{(j)} - \frac{2\Delta}{N-1} \times \begin{cases} (z_{t+1}^{(j)} - f(z_t^{(j)}))\ \mathrm{d}_t f(z_t^{(j)}), & t = 1 \\ (z_t^{(j)} - f(z_{t-1}^{(j)})) - (z_{t+1}^{(j)} - f(z_t^{(j)}))\ \mathrm{d}_t f(z_t^{(j)}), & 1 \le t \le N-1 \\ (z_t^{(j)} - f(z_{t-1}^{(j)})), & t = N \end{cases}$$

for a suitable step size Δ.

For more details on the algorithm and its use for nonlinear noise reduction purposes see [26], for its use with purposes of state estimation see [54, 79, 55, 51] and for details on how to use the methodology in the context of parameter estimation in nonlinear systems see Chapter 4.

Chapter 3

Bayesian Inference and Chaotic Dynamics

The application of Bayesian methods to the estimation of parameters in chaotic maps was first addressed by Berliner in 1991 [5]. In 1992 a second paper was published by Berliner [7] and was extensively commented upon and criticised by other statisticians [6, 25, 35, 37, 93]. This discussion saw the beginning of a new philosophical debate on the uses and boundaries of Bayesian methods on dynamical systems that is still active.

In Berliner's paper [5], noise free chaotic systems (in particular, the Logistic map) were presented in a didactic way. The aim was to point out how statisticians could understand chaotic behaviour using time series and the apparent risks involved in the inference process from such *chaotic Likelihoods.*

In particular, this work explores the dependency of the chaotic Likelihood's behaviour on the length of the data set and the unknown initial conditions as commented in [89]. In recent years, among the dynamical community, new interest on the application of Bayesian methods has arisen for the case of noisy chaotic time series as presented in several works, for example [12, 39, 70].

At present, time series analysis is carried out using both probabilistic and deterministic methods as the distinction between deterministic and random behaviour is difficult to ascertain. Sources of uncertainty in the observations from the systems under study and of the model's correspondence with reality make for a challenging task. At this point, any methodology that generates parameter estimates or forecasts that are consistent with reality is valid and useful.

One reason that Bayesian methodology is used in the analysis of nonlinear time series is that it incorporates in a "natural way" considerations of unknown parameters that can be interpreted as experimental information. For example, different types of noise present in the signal of interest [5, 42] can be incorporated into the modelling process. When the Bayesian perspective is applied numerically using Markov Chain Monte Carlo (MCMC) techniques, it seems that the use of stochastic models on chaotic systems provides impressively correct and unbiased parameter estimations [70, 12].

Implementation of the methodology suggests that it could help to deal

with noise components (measurement and dynamical noise) and system characterisation (*e.g.* parameter estimation and state estimation) simultaneously. Examples of the use of Bayesian methods in the analysis of complex dynamical systems such as discrete and continuous chaotic systems [60, 63, 70, 43, 74, 12, 94, 84], population models [90], sea clutter [41], ecosystem inverse problems [28], cardiorespiratory models [3, 62], and electricity grid dynamics [22], among others, are increasingly present in the literature.

Originally, Berliner [5], and later in [8, 52], addressed and presented a discussion about the possible flaws, shortcomings and inconsistencies of Bayesian estimations for nonlinear models. This discussion is re-addressed in this Chapter from a new perspective. Consistency tests are suggested in an effort to help recognise the validity and limitations of this approach in the case of parameter estimation for a grid frequency model, in conjunction with non-linear parameter estimation based on cost function methods [68, 76] and indistinguishable states theory.

Section 3.1 introduces chaotic maps from a Bayesian point of view. From the paradigm of Bayesian *state-space* modelling, a probabilistic model is presented to estimate parameters for a chaotic dynamical model. In particular, Section 3.2 discusses the case of the Logistic map's probability model. Samples for the posterior given the Logistic map's model and its corresponding noisy observations are generated using MCMC methods.

The numerical implementation of MCMC is carried out in two stages. Section 3.2.1.3 presents how to use the publicly available software, *Bayesian inference Using Gibbs Sampler* (WinBUGS). WinBUGS software is available free of charge from the BUGS project including manuals, tutorial and many examples. See ulr: `http://www.mrc-bsu.cam.ac.uk/bugs`.

The use of this software is shown to be inappropriate when dealing with chaotic Likelihoods, given that convergence of the posterior's samples is not robust enough, among other reasons. Moreover, high resolution estimation of the unknown initial condition is required as commented in [52].

A "Tailored" implementation of the algorithm is developed where each of the unknown parameters of the probabilistic model is obtained by sampling its corresponding full conditional distribution, as presented in Section 3.2.1.4.

The performance of MCMC techniques for the Logistic map is studied through a series of experiments. First, a perfect model experiment is set up in Section 3.2 in order to explore the sources of uncertainty included in the formulation and to detect inconsistencies in the probability model. Secondly, Section 3.3, presents several experiments performed by using the formulation developed in 3.2. These are used to distinguish deterministic from random behaviour. The Bayesian parameter estimation approach is used, in conjunction with surrogate data methods, to pin-point which features present in the data, the inferences are based on. In particular, the MCMC model

based method is applied both to time series data, and to surrogate data sets (*e.g.* random draws from the observed time series) for which the dynamical relationship described by the model does not usefully apply. Surprisingly, the MCMC estimations for both data sets are often indistinguishable. The results of this work are to be published elsewhere [24]. In section 3.4, several possible origins of the shortcomings of the Bayesian approach are discussed for this ambiguity, and the use, in general, of Bayesian perspectives in chaotic systems is summarised.

3.1 State-Space Modelling: Bayesian Framework

In order to illustrate the Bayesian approach, assume that there exists a dynamical system that is observed for a period of time. Moreover, the model that represents the system corresponds to the system itself, *i.e.* the model and reality are the same. Observations of the system are obtained by measuring the model variables at regular time intervals in order to produce a time series. It is assumed that the data set is normally perturbed by measurement and dynamical noise. Both noise components are taken to be independent and identically distributed random variables (IID).

Following the definition of the Perfect Model Scenario in Chapter 1, the dynamics follow a deterministic model of the form

$$x_t = f^t(x_0; \boldsymbol{\theta}), \qquad t = 0, 1, \ldots, \tag{3.1}$$

where $x_t \in \mathbb{R}^m$, $f : \mathbb{R}^m \to \mathbb{R}^m$ is a discrete map on t, $\boldsymbol{\theta} \in \mathbb{R}^\ell$ is the vector of unknown parameters, and $x_0 \in \mathbb{R}^m$ the unknown initial state (see equation (1.2) and (1.3) in Chapter 1, section 1.2). In a statistical framework, the deterministic model could be naively seen as a Nonlinear Auto-Regressive process (NAR) if the states x were to be considered random variables.

Let S be a data set *not yet* observed from a dynamical system f subject to observational and dynamical noise. Dynamical noise is seen as a stochastic perturbation of the deterministic states x_t given in general by:

$$x_t = f(x_{t-1}; \boldsymbol{\theta}) + \delta_t, \ \delta_t \sim IID(0, \sigma_\delta^2), \tag{3.2}$$

where $x_{t-1} = f(x_{t-2}; \boldsymbol{\theta}) + \delta_{t-1}$ and σ_δ is the amplitude or standard deviation of the perturbation (see equation (1.10) in Chapter 1, section 1.2).

In addition, the measurement noise component is

$$s_t = x_t + \eta_t, \ \eta_t \sim IID(0, \sigma_\eta^2), \tag{3.3}$$

with σ_η as its corresponding amplitude. The deterministic states x_t are often called *latent variables* since they are not observed directly (see equation (1.4) in Chapter 1, section 1.2).

The joint probability distribution for the dynamical model f [83], $p^{(f)}(S|\boldsymbol{\theta}, I)$, is updated via Bayes' theorem [4] to the distribution of the data given the model parameters, so that

$$p^{(f)}(\boldsymbol{\theta}|S, I) \quad \propto \quad p^{(f)}(S|\boldsymbol{\theta}, I) \cdot p^{(f)}(\boldsymbol{\theta}|I), \tag{3.4}$$

where the parameter vector $\boldsymbol{\theta}$ includes all the dynamical model parameters, the noise model parameters and any unknown parameters associated with the probabilistic model itself. Note that the parameter vector $\boldsymbol{\theta}$ contains all the non-observables whilst S contains only the observed parameters or variables. The superscript (f) in the probability distribution notation emphasises the fact that the Bayesian perspective is model dependent, and I is the background information about the dynamical system modelled by f.

The left hand side of equation (3.4) is known as the posterior distribution, which is decomposed into two terms: the Likelihood and the prior probability densities, respectively. The Likelihood contains the information of the data set given the model, whilst the prior contains information of the system itself and the model chosen to represent it (see section 2.1) and the background information I.

Assume now that a realisation $S = \{s_t\}_{t=1}^{N}$ is observed. Evaluating (3.4) on S gives the joint posterior distribution

$$\pi_S(\boldsymbol{\theta}|I) \quad \propto \quad p^{(f)}(S = \{s_t\}_{t=1}^{N} \, , \, \boldsymbol{\theta}|I), \tag{3.5}$$

where $\pi_S(\boldsymbol{\theta}|I)$ contains all the samples from the prior that best resemble the data, given the parameter vector $\boldsymbol{\theta}$ and are consistent with the background information I. The prior contains all the physical knowledge on the parameters involved. In principle, any inference of $\boldsymbol{\theta}$ could be made from (3.5). Note that, from now on, the superscript (f) in any probability density related to the probabilistic model is dropped for the sake of clarity and not because the dependency is forgotten.

In summary, the Bayesian framework does not make a fundamental distinction between model variables and model parameters, since all are naively considered as random variables in the context of dynamical systems. In addition, random variables are classified and collected in two categories: observables in S and non-observables in $\boldsymbol{\theta}$. Any random variable which is not observed is included in $\boldsymbol{\theta}$ (this includes, among other quantities, the system states x_t, the dynamical model parameters and the noise model amplitudes) and all are inferred in the process from the full posterior distribution of equation (3.5).

The Bayesian inference process can be seen as composed by:

1. Classification of all random variables in the model into observables and non-observables, collected in S and $\boldsymbol{\theta}$, respectively.

2. Forming the posterior distribution $p(S|\boldsymbol{\theta}, I)$.

a. Writing Likelihood terms, $p(\boldsymbol{\theta}|S, I)$, for all observables in S and background information I.

b. Writing prior terms, $p(\boldsymbol{\theta}|I)$, for all non-observables in $\boldsymbol{\theta}$ and prior information I.

3. Obtaining a realisation of S to evaluate the posterior with.

4. Generating samples of the full posterior distributions $\pi(\boldsymbol{\theta}|I)$.

In practice, random samples of $\pi_S(\boldsymbol{\theta}|I)$ are drawn using Markov Chain Monte Carlo techniques which are briefly described in section 2.1.1 and for the Logistic map in 3.2.1. section 2.1.2 exemplifies all these points for a simple two parameter probability model.

Up to this point, equation (3.5) is conformed by two terms: the Likelihood and the prior, with no apparent hurdle to overcome. Prior terms are not subject to any restriction, other than the one given by the model f and the expert knowledge, I, of the system of interest. Unfortunately, the Likelihood associated with noisy chaotic observations is highly complex, multimodal and strongly dependent on the initial condition x_0 [5]. This fact makes it difficult to interpret the results obtained by the methodology in dynamical terms if the observations are known to contain *only* measurement noise as it is the case in 3.2 and 3.2.1.

3.2 Example: Bayesian Inference for the Logistic Map

Before applying the probabilistic methodology to dynamical systems analysis, the PMS is defined for noisy observations that are suspected of being generated by a system governed by the Logistic map [67] following the general definition of the dynamical system of interest in Chapter 1.

Let S be a set of noisy observations of the system state x_t. The underlying states follow the one dimensional map $f = f(x_{t-1}, a)$ such that

$$x_t = 1 - ax_{t-1}^2 = f(x_{t-1}; a) = f^t(x_0; a). \tag{3.6}$$

Equation (3.6) is known as the Logistic map, where $a \in [0, 2]$ is the logistic parameter, $x_0 \in [-1, 1]$ its initial condition and f^t is the $t-$fold composition of the map.

The observations are, in general, subject to measurement and dynamical noise such that

$$x_t = f(x_{t-1}; a) + \delta_t, \qquad \delta \overset{iid}{\sim} \mathcal{N}(0, \sigma_\delta^2), \tag{3.7}$$

$$s_t = x_t + \eta_t, \qquad \eta \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2), \tag{3.8}$$

where the measurement noise variance σ_η^2 is known. Note that equations (3.7) and (3.8) are included for the sake of generality, *i.e.* they do not correspond to the conditions of the PMS used in the examples of this Chapter.

It is important to clarify that such a general setting is used to formulate the Bayesian methodology in the correct terms but the MCMC techniques will be implemented for the PMS conditions mentioned above, *i.e.* reality is (3.6) and the observations are given by (3.8) only.

In the Bayesian framework and for the general setting where both observational and dynamical noise components are present, random variables are classified into observables and non-observables. The non-observable random variables correspond to the components of $\boldsymbol{\theta}$, given by

$$\boldsymbol{\theta} = \left(x_0, x_1, x_1, \ldots, x_N, a, \sigma_\eta^2, \sigma_\delta^2\right). \tag{3.9}$$

From further refinements of the probability model, the components of $\boldsymbol{\theta}$ will eventually increase/decrease in dimension as prior information is included through the modelling process.

Once a realisation of length N, $S = \{s_t\}_{t=1}^N$, is observed, the Likelihood in (3.4) is evaluated on S so (3.5) becomes

$$\pi_S(\theta|I) \propto \left[\prod_{t=1}^{N} p(s_t|x_t, a, \sigma_\eta^2, I)\; p(x_t|x_{t-1}, a, \sigma_\delta^2, I)\right] \times \tag{3.10}$$

$$p(x_0, I)\; p(a, I)\; p(\sigma_\eta^2, I)\; p(\sigma_\delta^2, I), \tag{3.11}$$

under normality and independency assumptions for η_t and δ_t and each of the components of $\boldsymbol{\theta}$.

The Likelihood terms (3.10) show explicitly the contribution of both noise components in equations (3.7) and (3.8). The Likelihood accounts for un-

certainty sources in the observations. Explicit contributions for dynamical noise are given by

$$\begin{aligned} \delta_t &\sim \mathcal{N}(0, \sigma_\delta^2) \\ p(x_t|x_{t-1}, a, \sigma_\delta^2, I) &= \frac{1}{\sqrt{2\pi\sigma_\delta^2}} \exp\left\{-\frac{1}{2\sigma_\delta^2}\left(f^t(x_{t-1}; a) - x_t\right)^2\right\}, \end{aligned} \tag{3.12}$$

and for the measurement noise component by

$$\eta_t \sim \mathcal{N}(0, \sigma_\eta^2), \tag{3.13}$$

$$p(s_t|x_t, a, \sigma_\eta^2, I) = \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \exp\left\{-\frac{1}{2\sigma_\eta^2}(s_t - x_t)^2\right\}. \tag{3.14}$$

The probabilistic nature of this approach allows for the inclusion of a stochastic transition over time of the dynamical states in (3.12) and it could be interpreted in several ways. This multiplicity of meanings or interpretations has led to some misunderstandings and inconsistencies as in [70, 12], since the PMS is not well defined. The implications of such terms are discussed in sections 3.2.1.2 and 3.2.1.3. In the case where the noise model consists only of measurement noise, the stochastic evolution in time of the system's states in equation (3.12) is an artificial construct that needs to be introduced in the probabilistic model, and the use of Bayesian techniques is made in the NSA. As will be shown in section 3.2.1.2, that is required in order to make the MCMC's numerical implementation feasible. As discussed in section 3.2.1.2, the *ad hoc* "prior" knowledge of stochastic transition of the

system's states is included for the same reason. Therefore, it is not a natural feature of Bayesian methodology. Only where both noise components are present, *i.e.* dynamical and measurement noise, should the prior of the system's states include both dynamical and measurement noise.

At this stage, the discussion of the application of the Bayesian perspectives for the problem of the Logistic map is developed for the case where both measurement and dynamical noise are present in the observations following the general framework. Later in this Chapter, it will be reduced to a simpler case, where only measurement noise is present in the signal. This reduction is performed in order to make a consistent comparison with previous implementations of Bayesian methodologies to chaotic systems, in particular the work presented in [70].

The prior terms in (3.11) represent given uncertainty sources related to the model used to represent the data. Using prior knowledge of the model used, the priors are set as follows. A wide informative prior is chosen for the initial condition such that

$$\begin{aligned} x_0 &\sim \mathcal{U}(-1,1), \\ p(x_0, I) &= \mathbb{I}(x_0)_{[-1,1]}, \end{aligned} \tag{3.15}$$

given that it is known that the initial condition must lie in the real interval

$[-1, 1]$, where $\mathbb{I}(x)_{[u,v]}$ is the indicator probability function given by

$$\mathbb{I}(x)_{[u,v]} = \begin{cases} 0, & u < x \\ \frac{1}{v-u}, & u \leq x \leq v \\ 0, & x < v \end{cases} \tag{3.16}$$

For the logistic parameter a, a non-informative prior is assigned to be

$$\begin{aligned} a &\sim \mathcal{N}(0, 1), \\ p(a, I) &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-a^2/2}, \end{aligned} \tag{3.17}$$

where $\mathcal{N}(\mu, \sigma^2)$ is the Normal distribution with mean μ and variance σ^2.

The variance of the dynamical noise process in (3.7) is set to a prior that reflects the expectation for dynamical noise to be small and close to zero. The variance of the dynamical noise is written as σ_δ^2 and defined to be the prior for the dynamical noise variance given by

$$\begin{aligned} \frac{1}{\sigma_\delta^2} &\sim \mathcal{G}a(\alpha, \beta), \\ \sigma_\delta^2 &\sim \mathcal{I}\mathcal{G}a(\alpha, \beta), \\ p(\sigma_\delta^2, I) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma_\delta^2}\right)^{\alpha+1} \mathrm{e}^{-\beta/\sigma_\delta^2}, \end{aligned} \tag{3.18}$$

where $\mathcal{G}a(\alpha, \beta)$ is the Gamma distribution with scale and shape parameters α and β. The random variable, $X \sim \mathcal{G}a(\alpha, \beta)$, is always positive and close to zero. If $Y = 1/X$ then Y follows an inverted Gamma distribution $\mathcal{I}\mathcal{G}a(\alpha, \beta)$[17]. Here, the Gamma parameters are $\alpha = 2.01$ and $\beta = 0.0005$,

which corresponds to a Inverted Gamma distribution, denoted by $\mathcal{I}Ga(\alpha, \beta)$, with mean and variance of 4×10^{-4} and 2×10^{-3}.

The variance of the measurement noise process, σ_η^2, is set to be constant and equal to the square of the known noise amplitude. When the amplitude of the measurement noise process is unknown, the prior for σ_η^2 can be chosen to be an Inverted Gamma distribution.

Replacing equations (3.12) to (3.18) into the full posterior distribution $\pi_S(\boldsymbol{\theta}, I)$, the full probability model for the PMS of the Logistic map is

$$\begin{aligned}\pi_S(\boldsymbol{\theta}, I) \propto & \left(\frac{1}{\sqrt{2\pi\sigma_\eta^2}}\right)^N \exp\left[-\frac{1}{2\sigma_\eta^2}\sum_{t=1}^{N}(s_t - x_t)^2\right] \times \\ & \left(\frac{1}{\sqrt{2\pi\sigma_\delta^2}}\right)^N \exp\left[-\frac{1}{2\sigma_\delta^2}\sum_{t=1}^{N}(f(x_{t-1}; a) - x_t)^2\right] \times \\ & \mathbb{I}(x_0)_{[-1,1]} \times \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-a^2/2} \times \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma_\delta^2}\right)^{\alpha+1} \mathrm{e}^{-\beta/\sigma_\delta^2}, \quad (3.19)\end{aligned}$$

for a parameter vector, $\boldsymbol{\theta} \in \mathbb{R}^{N+3}$.

Generating samples from (3.19) is difficult because calculation of the marginal distribution implies a high dimensional integration ($\sim \mathbb{R}^{N+3}$), *i.e.* the normalisation constant of the posterior, as discussed in section 2.1.1.

MCMC techniques are independent of the normalisation constants. Hence what is important is the functional dependency of the posterior, and how components of $\boldsymbol{\theta}$ depend on other components of $\boldsymbol{\theta}$.

Figure 3.1 shows the graphical representation of this model by a *Directed*

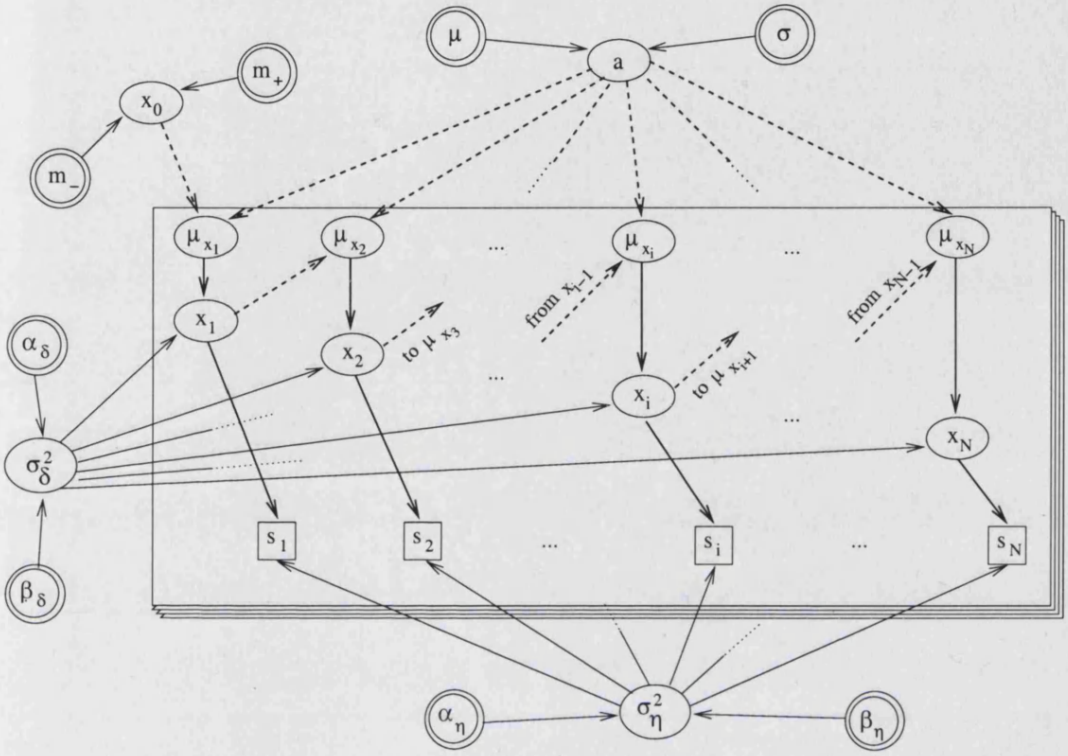


Figure 3.1: DAG for the Logistic map probability model structure. Squares represent model observables, circles non-observables, *i.e.* the model parameters, and double circles hyper-parameters included to model the probability model parameters. Dashed lines represent deterministic relations while solid lines stand for probabilistic dependency. The layered panels represent algorithmic iterations.

Acyclic Graph (DAG), where Directed means that each link between nodes is an arrow. It is Acyclic because it is impossible to return to a node after leaving it [95]. A graphical representation of model structure may be produced before any structural assumptions are made and helps to clarify the dependency or independency relations between parameters and observations

Likelihood terms	
Measurement noise: η_t	$\eta_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2)$
Dynamical noise: δ_t	$\delta_t \overset{iid}{\sim} \mathcal{N}(f(x_t;\boldsymbol{\theta}), \sigma_\delta^2)$
Prior terms	
Initial condition: x_0	$x_0 \sim \mathbb{I}(x_0)_{[-1,1]}$
Logistic parameter: a	$a \sim \mathcal{N}(0,1)$
Dynamical noise variance: σ_δ^2	$\sigma_\delta^2 \sim \mathcal{IG}a(\alpha, \beta)$

Table 3.1: Likelihood and prior terms for the PMS Logistic map probability model.

in the model. The graph is included here as a mnemonic resource for the probability model of the Logistic map. Relations among nodes in the graph resemble kinship relations, for details see [95].

In Table 3.1 all Likelihood and prior terms defined for the PMS are collected.

3.2.1 MCMC for the Logistic Map: In Practice

In practice, given the level of complexity of the probability models developed for dynamical systems (*i.e.* Logistic map), there are two possibilities for obtaining samples of the full posterior distribution given in general by (3.5). Perhaps the best option for the first approach to MCMC techniques is to use

a "black box" package. The advantage is that samples are obtained by specifying the Likelihood and prior terms using a correct syntax. For example, once the syntax of the package is familiar, the MCMC implementation of the probability model given by equation (3.19) involves only writing the Likelihood and prior terms explicitly. Changes in prior information or even noise model assumptions will only neglect some of the terms already described here and/or add some additional ones. Packages like WinBUGS [95] are very useful for learning and training purposes since they prevent the user from the generating random samples from non-standard probability functions. For deeper studies, it is often unavoidable for the researcher to use "Tailored" routines in order to perform more refined calculations.

Another option is a personalised MCMC implementation for the model of interest. In this work, both ways of generating samples from the posterior were used, given that some inconsistencies where found when using WinBUGS for the model described in section 3.2.

In any case, when using a "black-box" package or "Tailored" set of routines, the single component Metropolis-Hasting algorithm using a Gibbs sampler (see section 2.1.1) is used. In that case the steps to follow to obtain samples from the posterior in equation (3.19) are:

1. Calculate the full posterior distributions as in section 3.2.1.1 for each

of the components of $\boldsymbol{\theta}$ in equation (3.9).

2. Select a sampling method for each component of $\boldsymbol{\theta}$.

3. Set up the iterative process by which a Markov chain state is generated in each iteration t, by updating the full conditional distribution of each component of $\boldsymbol{\theta}$.

4. Approximate any inference of $\boldsymbol{\theta}$ by a Monte Carlo approximation.

As explained in section 2.1.1 and in Chapter 2, full conditional distributions play a key role in many applications of MCMC techniques. They are taken as the proposal distribution when using the Gibbs sampler, and are responsible for the update of each component of $\boldsymbol{\theta}$. section 3.9 calculates the full conditional distribution for the probabilistic model of the Logistic map.

The study of the Bayesian methodology for the inference of nonlinear dynamics involved the reproduction of the results obtained in [70, 12], particularly in the use of WinBUGS of Meyer and Christensen in [70].

Given that the parameter estimation problem stated in [70] only contains measurement noise, it is inconsistent with the probability model used, since the perfect model described at the begining of these chapter includes dynamical noise as well. In consequence, the PMS described in section 3.2 is taken as one where only measurement noise is present in the Logistic map's observations. section 3.2.1.2 presents a discussion of the reasons for the potentially

misleading concepts used by Meyer and Christensen in order to justify their model assumptions and corresponding parameter estimates. Using one PMS formulation (with additive dynamical noise) to simulate observation known to be generated in a different PMS (without dynamical noise), misuses the technique and attempts to solve the problem in a NSA. In addition, a correct but not numerically tractable Bayesian formulation is developed which is consistent with the problem of estimating model parameters for the Logistic map.

For consistency, the WinBUGS package was used to generate samples of the posterior in the NSA in order to make a comparison with results in [70]. No consistency with the estimates presented in [70] was found. The failure of WinBUGS when applied to solve numerically probability models from chaotic time series is presented in section 3.2.1.3 a brief discussion of its shortcomings is as well presented. From now on, the explicit dependency of the full joint posterior and full conditional distribution on the background information I is dropped off since it is already reflected in the way the posterior is constructed in this section.

3.2.1.1 Full Conditional Distributions: PMS Logistic Map

section 2.1.1 defines the full conditional distribution $\pi_S(\theta_{.i}|\theta_{.-i})$ as the distribution of the ith component of $\boldsymbol{\theta}$ conditioned on all the remaining components. From equation (3.5), $\boldsymbol{\theta}$ has distribution $\pi_S(\cdot|\cdot)$, the full conditional is then defined by

$$\pi_S(\theta_{.i}|\theta_{.-i}) = \frac{\pi_S(\boldsymbol{\theta})}{\int \pi_S(\boldsymbol{\theta})\, \mathrm{d}\theta_{.i}}, \tag{3.20}$$

where, for the Logistic map, the probability model has $i = 1, \ldots, N+3$ components, with N as the length of the observations available. The Gibbs sampler takes (3.20) as the proposal distribution from which the next step of the chain is updated. The proposal or full condition distribution is constructed from picking the terms in (3.5) which explicitly depend on $\theta_{.i}$.

Table 3.2 summarises the full conditional distributions for each of the components of $\boldsymbol{\theta}$. These full conditionals are calculated from (3.19) by just picking the terms where the corresponding component appears. Given that the MCMC implementation is independent of the normalisation constant, all that is needed is the functional form of the posterior distribution and the corresponding full conditionals.

The components of the $\boldsymbol{\theta}$ after the modelling process are:

$$\boldsymbol{\theta} = \left(x_0, x_1, \ldots, x_t, \ldots, x_N, a, \sigma_\delta^2\right). \tag{3.21}$$

In order of appearance of the components of $\boldsymbol{\theta}$ (3.21), the full conditionals are listed in Table 3.2.

$\boldsymbol{\theta}$ component	Full Conditionals functional form	PDF	PDF parameters
$\theta_{.1} = x_0$	$\mathrm{e}^{-A_1 x_0^4 - B_1 x_0^2}$	Quartic exponential	$A_1 = \frac{a^2}{2\sigma_\delta^2}$ $B_1 = \frac{a(x_1-1)}{\sigma_\delta^2}$
$\{\theta_{.i}\}_{i=2}^{N} = \{x_t\}_{t=1}^{N-1}$	$e^{-A_1 x_t^4 - B_i x_t^2 + C_i x_t}$	Quartic exponential	$B_i = \frac{\sigma_\delta^2+\sigma_\eta^2+2a\sigma_\eta^2(x_{t+1}-1)}{2\sigma_\delta^2\sigma_\eta^2}$ $C_i = \frac{s_t\sigma_\delta^2+\sigma_\eta^2(1-ax_{t-1}^2)}{\sigma_\delta^2\sigma_\eta^2}$
$\theta_{.N+1} = x_N$	$\mathrm{e}^{-(x_N-\mu_N)^2/2\sigma_N^2}$	$\mathcal{N}(\mu_N, \sigma_N^2)$	$\mu_N = \frac{\sigma_\delta^2 s_N+\sigma_\eta^2(1-ax_{N-1}^2)}{\sigma_\delta^2+\sigma_\eta^2}$ $\sigma_N^2 = \frac{\sigma_\delta^2\sigma_\eta^2}{\sigma_\delta^2+\sigma_\eta^2}$
$\theta_{.N+2} = a$	$\mathrm{e}^{-(a-\mu_a)^2/2\sigma_a^2}$	$\mathcal{N}(\mu_a, \sigma_a^2)$	$\mu_a = \frac{\sum_{t=1}^{N} x_{t-1}^2(1-x_t)}{\sigma_\delta^2+\sum_{t=1}^{N} x_t^4}$ $\sigma_a^2 = \frac{\sigma_\delta^2}{\sigma_\delta^2+\sum_{t=1}^{N} x_t^4}$
$\theta_{.N+3} = \sigma_\delta^2$	$(1/\sigma_\delta^2)^{\alpha_\delta+1}\mathrm{e}^{-\beta_\delta/\sigma_\delta^2}$	$\mathcal{IG}a(\alpha_\delta, \beta_\delta)$	$\alpha_\delta = \frac{N}{2} + \alpha$ $\beta_\delta = \beta + \frac{1}{2}\sum_{t=1}^{N}(x_t - 1 + ax_{t-1}^2)$

Table 3.2: Table of full conditional distributions for the probability model in the PMS for the Logistic map.

For example, the construction of the full conditional distribution of the

first component of $\boldsymbol{\theta}$, $\theta_{.1} = x_0$, is as follows. Checking equation (3.19) for the terms where x_0 appears, it is found that the full conditional distribution of the initial condition is proportional to:

$$\pi_S(x_0|x_1,\ldots,x_t,\ldots,x_N,a,\sigma_\delta^2) \;\propto\; \mathrm{e}^{-\frac{1}{2}(f(x_0;a)-x_1)^2}\mathbb{I}(x_0)_{[-1,1]}. \quad (3.22)$$

After some algebraic manipulation, the functional form is proportional to a quartic exponential distribution shown in Table 3.2. It is known that the indicator distribution is not conjugate to the Normal or Gamma prior conjugate family, and therefore equation 3.22 is not in a closed form. As described in section 2.1, the indicator distribution is not conjugate to the Normal distribution therefore the product of both does not belong to the conjugate family.

When a full conditional is not in closed form, a sampling random algorithm has to be implemented [9, 17]. Only the full conditional distributions corresponding to the parameter vector components to the initial condition x_0 and the latent variables $\{x_t\}_{t=1}^{N-1}$ are not in closed form. Details on how samples are obtained when closed form of the full conditionals are not availabel for the components of $\boldsymbol{\theta}$ are given in [23] and in section 3.2.1.4. In the case that the full conditional distribution is in a closed form, any standard routine for generating random samples can be used, as is the case for the rest of the remaining components of $\boldsymbol{\theta}$.

3.2.1.2 Naive Statistical Approach for the Logistic Map

It is important to note that the motivation for this study comes from the works cited in [70, 12]. In both papers, the outstanding performance of the MCMC techniques when estimating parameters, unobserved components of the state vector and reducing noise in the reconstructed dynamics of chaotic systems, caught the interest in the use Bayesian perspectives for parameter estimation and even for inference of the nonlinear dynamics of chaotic time series. Both papers used WinBUGS as the numerical tool to implement the Bayesian inference process of any parameter in the corresponding model. Their performance of the methodology was apparently higher that any other methodology or technique that has been used in this framework for parameter estimation [54, 26, 27]. It looked suspiciously successful.

In terms of an analytical formulation, and even in prior formulation, the probability model presented by Meyer and Christensen and the one developed in section 3.2 are quite similar. The difference is subtle and it was only identified when problems arose in the numerical implementation and an interpretation of the estimates obtained was made.

The fundamental difference has roots in the PMS conditions presented in [70] and the general model setting scenario described earlier in Chapter 1 and in this Chapter in section 3.1. The model used in [70] aims to solve a dif-

ferent problem than the one they described. In short, Meyer and Christensen want to solve the problem of parameter estimation of a chaotic system from noisy observations when there is *only* measurement noise in the signal. Note that if literally compared, *i.e.* equation by equation, both probability models appear equivalent. Intuitively, the two probability models should look different since the observations are conditioned differently and it is assumed that dynamical noise is not present in the observations.

Perfect model conditions are going to be described in detail, in order to clearly point out how Meyer and Christensen's formulation is incorrect and how it is inappropriate to solve the problem of model parameter estimation of nonlinear models from observations subject only to measurement noise. In addition, based on these perfect model conditions a Bayesian probability model suitable to solve this problem (as stated in McSharry and Smith [68]) is going to be constructed. Please follow this description in parallel with the one that starts in equation (3.6).

Assume a set of noisy observations S of the system state x_t is obtained. The system states follow the one dimensional map $f = f(x_{t-1}, a)$ such that

$$x_t = 1 - ax_{t-1}^2 = f(x_{t-1;a}) = f^t(x_0; a), \tag{3.23}$$

is the Logistic map.

The observations S are known to be subject to measurement noise only

and are therefore given by

$$s_t = x_t + \eta_t, \qquad \eta \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2), \tag{3.24}$$

with known variance σ_η^2. There is no dynamical noise.

Recalling that in the Bayesian framework where there is no distinction between parameters and variables, all model variables and parameters are considered random variables and in conjunction with the conditions of equations (3.23) and (3.24), the modelling process proceeds by classifying the random variables as observables and non-observables. The non-observed random variables correspond to the components of $\boldsymbol{\theta}$ and are given by:

$$\boldsymbol{\theta} = \left(x_0, x_1, x_1, \ldots, x_N, a, \sigma_\eta^2\right). \tag{3.25}$$

Once a realisation of length N, $S = \{s_t\}_{t=1}^N$, is observed, equation (3.4) is evaluated on S so (3.5) becomes

$$\pi_S(\theta, I) \propto \left[\prod_{t=1}^{N} p(s_t|x_t, a, \sigma_\eta^2, I)\right] \times \tag{3.26}$$

$$p(x_0, x_1, \ldots, x_N, I)\; p(a, I)\; p(\sigma_\eta^2, I), \tag{3.27}$$

under normality and independency assumptions for η_t each of the components of $\boldsymbol{\theta}$.

The Likelihood terms in (3.26) contain the noise contributions. In this case the noise consists only of measurement noise. Explicitly this contribution

is

$$\eta_t \sim \mathcal{N}(0, \sigma_\eta^2), \tag{3.28}$$

$$p(s_t|x_t, a, \sigma_\eta^2, I) = \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \exp\left[-\frac{1}{2\sigma_\eta^2}(s_t - x_t)^2\right]. \tag{3.29}$$

Now, the priors for each of the components in $\boldsymbol{\theta}$ are set as in equations (3.15) for the initial condition, (3.17) for the logistic's model parameter and the amplitude of the measurement noise is constrained to be a known constant.

The priors for all latent states, $x_1, \ldots, x_t, \ldots, x_N$, of the Logistic map are still to be chosen. Consistent with the background information I that there is no dynamical noise, let the prior of the latent state x_t be equal to a Dirac delta function of the form:

$$\begin{aligned} x_t &\sim \delta(x - x_t), \\ p(x_t, I) &= \begin{cases} x_t, & x = x_t \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{3.30}$$

In addition, there is no reason why any other prior should be assumed since it is known with certainty that the model is perfect for the dynamics and the noise.

It is at this point that the interpretation of the probability model for the Logistic map starts to be confusing in [70]. Up to the choice of the prior for the system states, the use of Bayesian perspectives for parameter estimation

of chaotic systems do not force nor prevent the choice of any prior or noise model in the formulation. Formally, there are no misleading assumptions in the probability model from equation (3.23) to (3.30). In turn, replacement of the Likelihood and prior terms into (3.5) gives the full joint probability distribution $\pi_S(\boldsymbol{\theta}, I)$

$$\pi_S(\boldsymbol{\theta}) \propto \left(\frac{1}{\sqrt{2\pi\sigma_\eta^2}}\right)^N \exp\left[-\frac{1}{2\sigma_\eta^2}\sum_{t=1}^{N}(s_t - x_t)^2\right] \times \mathbb{I}(x_0)_{[-1,1]} \times p(x_1)\ldots p(x_t)\ldots p(x_N) \times \frac{1}{\sqrt{2\pi}}e^{a^2/2}, \quad (3.31)$$

dropping the dependency in the prior information since priors are conviniently chosen.

Equation (3.31) is the posterior distribution that contains the information from which any inference of the parameter vector $\boldsymbol{\theta}$ in (3.25) can be drawn. This posterior distribution represents the values of $\boldsymbol{\theta}$ that best resemble the noisy observations S given that the dynamics follow the logistic map and the noise model is composed of only one component: additive perturbations, *i.e.* measurement noise, and the prior information I.

This formulation is as realistic as it could be consistently with the formulation of the problem scenario. The problem aims to find parameter estimates for an exactly known dynamical model, (3.23), and additive measurement noise model (3.24). Any other liberties given by the setting of priors are independent of the problem and consequently places the Bayesian attempt

to solve the problem as naive (NSA) even though valid if estimates obtained are consistently interpreted.

In the idyllic case where (3.31) is analytical, the solution of the problem is just as far as the evaluation of at least two integrals; one for the calculation of the normalisation constant of the posterior, *i.e.* the marginal distribution, and another to calculate any expectation of a function g of $\boldsymbol{\theta}$, *i.e.* to infer the parameter vector or any of its components.

Since it is known that the states x_t follow the Logistic map and that the prior chosen do not dependent on the dynamics, the Likelihood in (3.31) is explicitly as follows

$$\begin{aligned} p(\boldsymbol{\theta}|S) &\propto \exp\left[-\frac{1}{2\sigma_\eta^2}\sum_{t=1}^{N}(s_t - f(x_{t-1};a))^2\right], \\ &\propto \exp\left[-\frac{1}{2\sigma_\eta^2}\sum_{t=1}^{N}(s_t - f^t(x_0;a))^2\right]. \end{aligned} \tag{3.32}$$

Replacing (3.32) into the full posterior distribution (3.31), follows that

$$\begin{aligned} \pi_S(\boldsymbol{\theta}) \propto \left(\frac{1}{\sqrt{2\pi\sigma_\eta^2}}\right)^N \prod_{t=1}^{N}\exp\left[-\frac{1}{2\sigma_\eta^2}(s_t - f^t(x_0;a))^2\right] \times \\ \mathbb{I}(x_0)_{[-1,1]} \times p(x_1)\ldots p(x_t)\ldots p(x_N) \times \frac{1}{\sqrt{2\pi}}\mathrm{e}^{a^2/2}. \end{aligned} \tag{3.33}$$

In order to generate samples of the posterior such that inferences of $\boldsymbol{\theta}$ are calculated, the next step is to performed the calculation of the full conditional distributions. Going in order of the components in equation (3.25), the full conditional distribution for the first component of $\boldsymbol{\theta}$, $\pi_S(\theta_{.1}|\theta_{.-1})$,

corresponding to the initial condition x_0 is proportional to

$$\pi_S(x_0|x_1,\ldots,x_N,a) \;\propto\; \exp\left[-\frac{1}{2\sigma_\eta^2}\sum_{t=1}^{N}(s_t - f^t(x_0;a))^2\right]. \qquad (3.34)$$

In order to see clearly the functional form of the initial condition full posterior, the negative logarithm of (3.34) is taken to obtain

$$\log\left(\pi_S(x_0|x_1,\ldots,x_N,a)\right) \;\propto\; \frac{1}{2\sigma_\eta^2}\sum_{t=1}^{N}(s_t - f^t(x_0;a))^2, \qquad (3.35)$$

explicitly the polynomial

$$\begin{aligned}\propto\;& (s_1 - \underbrace{(1-ax_0^2)}_{\text{once}})^2 + (s_2 - \underbrace{(1-a(1-ax_0^2)^2)}_{\text{twice}})^2 + \cdots + \\ & (s_N - \underbrace{(1-a(1-a(1-a(1-\ldots a(1-ax_0^2)^2\ldots)^2)^2)^2)^2)}_{N\text{ times}})^2. \qquad (3.36)\end{aligned}$$

Henceforth the full conditional for x_0 is an exponential distribution

$$\pi_S(x_0|x_1,\ldots,x_N,a) \propto \exp\left[b_0 + b_1x_0 + b_2x_0^2 + \cdots + b_{2^{2N}}(x_0)^{2^{2N}}\right]. \qquad (3.37)$$

Similarly, for the logistic parameter a, the full conditional is

$$\pi_S(a|x_0,x_1,\ldots,x_N) \;\propto\; \exp\left[c_0 + c_1a_0 + c_2a_0^2 + \cdots + c_{2^N}a^{2^N}\right]. \qquad (3.38)$$

It is clear that both full conditional distributions are numerically intractable even for $N \sim 3$. This fact is directly related to the remark of Berliner when referring to the wild behaviour of chaotic Likelihoods and it is exemplified there with numerical calculations [5]. Such high order polynomials result in "wild" behaviour and large numbers of modes in the chaotic

Likelihood. Note that it is easy to see that any other Likelihood associated with a chaotic map which is quadratic bears similar intractability issues.

If the priors are not chosen to be Dirac delta function nor any probability density function in which the latent states x_t are IID, the more sensible choice will be to include dynamical noise in the probability model despite the fact that the observations do not contain any. Once the dynamical noise is included the direct dependency of the Likelihood on the dynamics is broken as explicitly calculated in section 3.2.1.1.

As it is clearly shown by equations (3.37) and (3.38), the justification of the introduction of dynamical noise in the probabilistic model for the Logistic map is ill posed and has to be included carefully. Meyer and Christensen literally argued in [70] that in order... *"To develop this idea within a proper statistical paradigm requires treating the system states as stochastic instead of deterministic. We therefore consider the more realistic case that the system dynamics are subject to random disturbances."* This argument is only true if the given PMS problem that is aimed to be solved is the one of parameter estimation of chaotic systems from observations with both components of noise.

From the formulation of the problem presented in [70], such an argument only follows in order to achieve numerical tractability and it does not imply any correctness or proper implementation of Bayesian perspectives [24]. On

the contrary, given the assumptions it is an incorrect formulation of the Bayesian approach to find parameter estimates for the Logistic map from observations that only contain measurement noise.

Although, the perfect model is known, to solve the problem, an imperfect model of the system is used in order to facilitate numerical calculation. This situation is referred to in Chapter 1 as the NSA.

Their inclusion of stochastic transition over time for the latent variables of the Logistic map is more justifiable and consistent with the approach if it is seen as a choice for the prior of the system states. Otherwise, it is just an *ad hoc* condition to force Bayesian perspectives into the dynamical settings.

Despite the inconsistencies in the problem formulation posed by Meyer and Christensen, the probability model which includes dynamical noise is correct when dynamical noise is also known to be present in S. When that is not the case, as in [70] the probability model is incorrect but it is useful in the sense that it provides a feasible numerical implementation. In order to have a tractable numerical probability model in the Bayesian framework, all experiments and simulations are going to be performed using this NSA.

Consistently the artificial dynamical noise included in the model could be seen as a term which accounts for model error and the resulting inferences should be interpreted accordingly. It is important to note that once the standard deviation of the dynamical noise σ_δ component tends to zero the

regime whith only measurement noise is reached. Even though that limit is justifiable in dynamical terms, in the Bayesian framework it implies a fundamentally different probability model.

3.2.1.3 Using WinBUGS: Chaotic Bugs

The BUGS (Bayesian inference Using Gibbs Sampling) project is concerned with flexible software for the Bayesian analysis of complex statistical models using MCMC methods. The project began in 1989 in the MRC Biostatistics Unit and led initially to the 'Classic' BUGS program, and then onto the WinBUGS software developed jointly with the Imperial College School of Medicine at St Mary's, London. The project first developed a DOS based program and later Windows based software, now widely used. At present, the software can also be run on from Unix-based platforms. For more information on the project please refer to `http://www.mrc-bsu.cam.ac.uk/bugs/`.

As a learning tool, WinBUGS is a very useful resource in conjunction with [95] for familiarisation with Bayesian modelling issues and numerical techniques involved in the process of posterior sampling. Most of the comments included in this section are a result of various personal communications with Renate Meyer and Nelson Christensen from the Department of Statistics, University of New Zealand, and Andrew Thomas from Imperial College,

one of the main developers of the BUGS software. Contacts with Meyer and Christensen were made when many trials of different versions of their model were performed using BUGS/WinBUGS did not produce any output which resembled the results presented in [70]. Kindly, Christensen shared one of the original versions used for the calculations presented in their paper.

The study of the lack of convergence of the MCMC output for the probability model of the Logistic map and technical sampling errors with the algorithms when multimodal priors are used (*e.g.* slice sampling [71]) were done in collaboration with Andrew Thomas in a couple of meetings and several personal communications [21]. As a result of these discussions, it was decided to build a "tailored" implementation of the MCMC techniques in order to gain control on the process of monitoring sampling process and identify any possible faults. This implementation is described in 3.2.1.4.

In particular, BUGS or WinBUGS performance is put under strain when faced to high complex Likelihoods and uncertain convexity in the full conditionals, the case when nonlinearities are present [21]. A. Thomas performed several corsections to the software in order to improve the sampling algorithm used in the such cases and also in the code interpreter.

A data set from the Logistic map of 100 points is generated for $a = 1.85$ and initial condition $x_0 = 0.3$. The Bayesian model implemented in WinBUGS is applied to several data files with noise levels varying from 0

to 2, in variance. The data files contain only measurement noise for two noise processes Gaussian and Uniform. The noise level is defined as $l \equiv \sigma_{noise}/\sigma_{signal}$, where σ is the standard deviation. Denote the Gaussian and Uniform noise amplitude as a function of the noise level as $\epsilon_g(l)$ and $\epsilon_u(l)$, respectively.

To obtain a sample from the posterior distribution for the Logistic noisy data, 1.1×10^5 iterations of the Gibbs sampler are performed. This number of iterations include a burn-in time of 1.0×10^4 iterations following [70]. To avoid highly correlated values and to reduce the size of the output, the resulting chain is thinned by taking every 20th observation which yields a final sample size of 5000 points and took an average of 3 minutes on a Pentium IV, 2GHz processor PC for each data file. Only traces of the parameters corresponding to a, x_0 and τ^2 were recorded, the Logistic parameter, the initial condition and the precision parameter respectively. Note that τ^2 corresponds to the variance of the dynamical noise parameter $\sigma^2 = 1/\tau^2$.

To assess convergence, the GR statistic is calculated for the three parameters by fragmenting the chain into two parts of 2500 points each. Following Meyer and Christensen [70], after 1.0×10^4 iterations, the mixing of the chain is assumed to have finished. Figure 3.2 shows the GR statistic calculated for the median, variance and 97.5% isopleth, for a and σ^2. The horizontal line is the pass mark for the test. Given that the samples are approximately

converging to samples of the full posterior, there is no clear evidence of convergence as shown in the Figure 3.2.

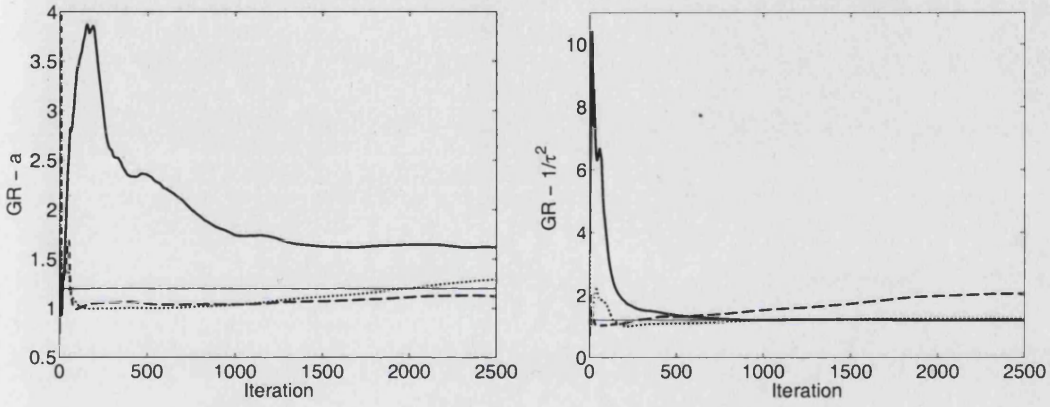


Figure 3.2: GR statistic for the Logistic parameter a and dynamical noise variance σ^2 from BUGS output. The GR statistic is calculated for the median (solid line), variance (dashed line) and 97.5% (dotted line). The horizontal line indicates the pass mark of the test.

For both parameters, the GR for the median presents very slow convergence, whilst for the GR statistic for the variance and 97.5% reached convergency after approximately 100 iterations but start loosing this property as is shown by increasing GR statistic values. Figure 3.2 can be interpreted as evidence that the chain has not finish the mixing process. Note that the chain has been thinned down, therefore values in the x-axis of the Figure have to be multiplied by 20 and offset by 1×10^4.

Figure 3.3 provides a clear evidence that the convergency properties of the samples is intermittent. For median (solid line), the GR test oscillates from

values that are higher and lower than the pass mark (black line, GR< 1.2). The chain has not mixed properly, therefore any inference from this sample is not reliable. There is no evidence that the samples are drawn from the stationary distribution of the chain, *i.e.* the full posterior (3.19).

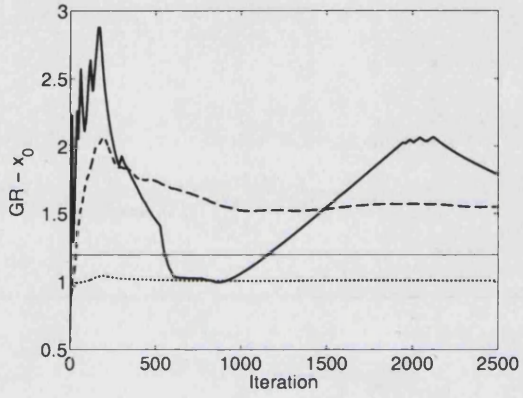


Figure 3.3: GR statistic for the initial condition x_0 of the Logistic map from BUGS output. The GR statistic is calculated for the median (solid line), variance (dashed line) and 97.5% (dotted line). The horizontal line indicates the pass mark of the test.

Reasons for this are related to the multimodality of the resulting full conditionals [21]. Random generation from multimodal distributions carries the risk of getting trapped in one of the modes, and as a consequence are not visited often or even ever visited, generating correlated samples. Once this is realised, the slice sampling algorithm was introduced in WinBUGS to generate more accurate samples from multimodal distributions. These changes in WinBUGS were performed by A.Thomas. Unfortunately, it did

not improve the quality of samples obtained for the Logistic map's Bayesian model at the time this study was made.

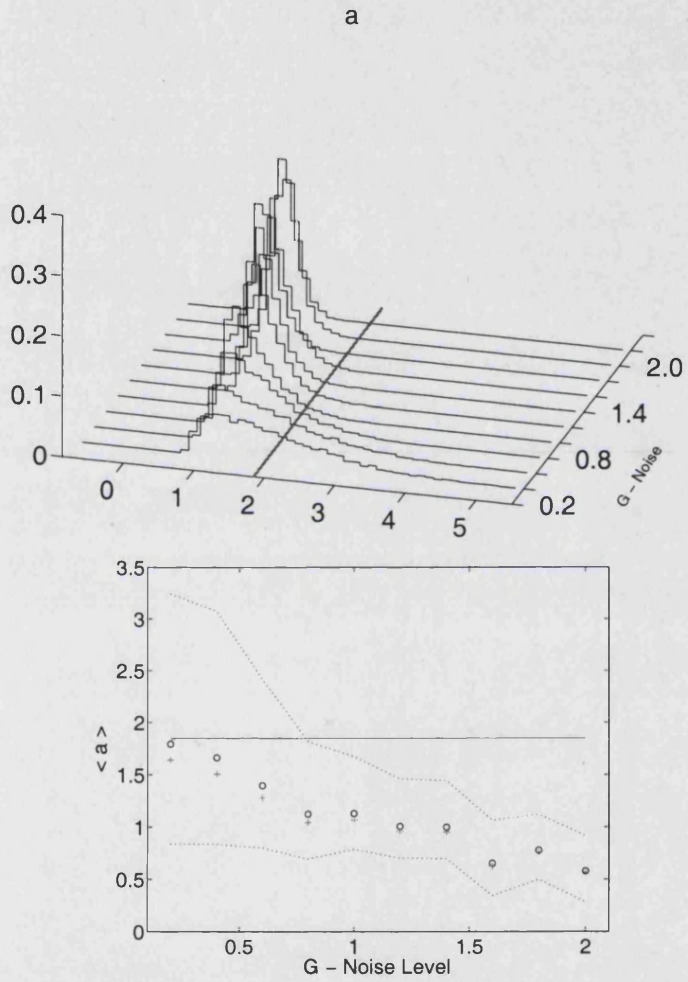


Figure 3.4: Logistic parameter, a. Top: Histograms of BUGS samples for each noise level considered, the line in the floor of the histograms indicates the true parameter value $\tilde{a} = 1.85$. Bottom: Mean ($\circ$), median ($+$), and envelopes of 5% and 95% isopleths (dotted lines) as a function of the noise level.

Despite the mixing not being complete, inference of some parameter com-

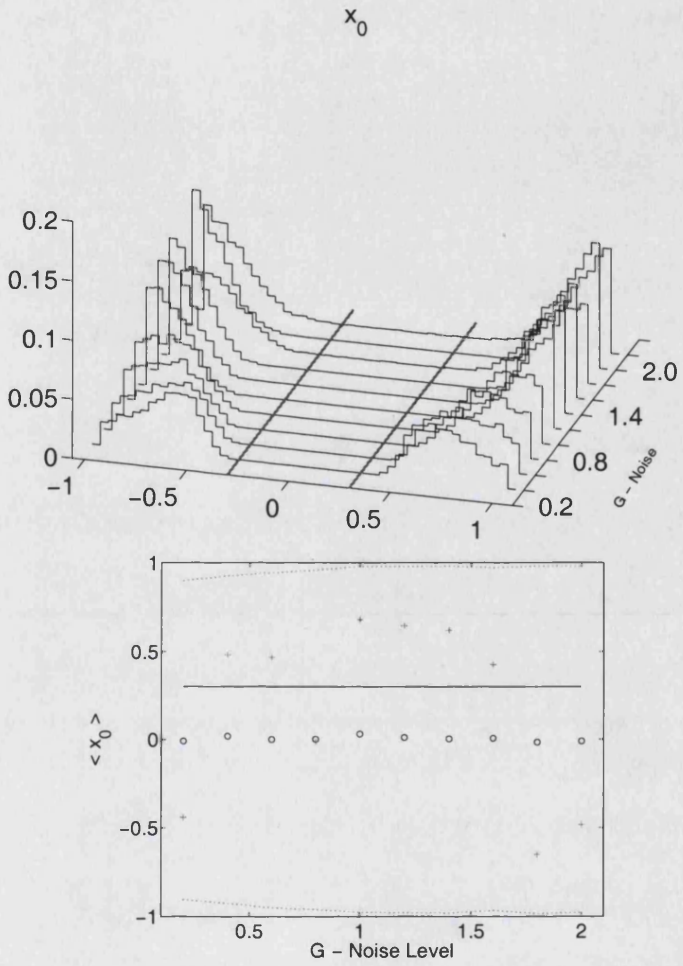


Figure 3.5: Initial condition, x_0. Top: Histograms of BUGS samples for each noise level considered, the line in the floor of the histograms indicates the true parameter value $\tilde{x_0} = 0.3$. Bottom: Mean (∘), median (+), and envelopes of 5% and 95% isopleths (dotted lines) as a function of the noise level.

ponents in $\boldsymbol{\theta}$ are made in order to compare with results obtained in Meyer and Christensen [70]. Figure 3.4 and Figure 3.5 shows the histograms (top) and traces (bottom) for the Monte Carlo estimates of the Logistic parameter

a and the initial condition x_0 as function of the noise level.

As expected, the estimates of a decrease as the noise level is higher; surprisingly the 95% envelopes are wider for lower noise levels even though the uncertainty is lower. This fact is clear as well in the histogram (top) since the empirical distribution obtained is sharper as the noise increases. Comparing with the corresponding figures of Meyer and Christensen in [70], the results are clearly not consistent despite the fact that they share some qualities.

After some more trials on different models, initial conditions, constant nodes, the convergence results obtained and the discussions with Andrew Thomas, it was necessary to implement a "tailored" MCMC routine. In particular, the problems concerning the implementation of the model in WinBUGS were related to intermittent convergence due to errors in the *Greedy sampling* algorithms used in the software itself [21].

Furthermore, well after the "tailored" implementation was on track serious errors were found by developers of the software. Principally, mistakes were found in the code interpreter of WinBUGS. Details can be found in the discussion list of the WinBUGS project at `http://www.mrc-bsu.cam.ac.uk/bugs/`.

3.2.1.4 MCMC "Tailored" Implementation

This section presents the actual implementation of Bayesian parameter estimation techniques for the PMS described in section 3.2. The state-space Bayesian model is resumed in Figure 3.1 and explicitly in Table 3.1.

The implementation of the single component Metropolis-Hastings algorithm [69] generates a sequence of Markov chain states for the parameter vector $\boldsymbol{\theta}$. Compared with the general Metropolis-Hastings algorithm described in section 2.1.1, when the Gibbs sampler is used, the candidates for the next chain state are always accepted when drawn from the full conditional distributions. Each iteration of the algorithm generates a state of $\boldsymbol{\theta} \in \mathbb{R}^{N+3}$ as in equation (3.21). The j^{th} state of the chain is denoted by $\boldsymbol{\theta}^{(j)}$.

Each of the full posteriors listed in Table 3.2 is used only once since as soon as one component is updated the full conditional distribution used to update the next component is evaluated at all other components values updated in the current iteration j and at the last one $j-1$.

Note that the dynamical system information is included in the probability model only through the observations S and the dynamical model structure is explicitly in the probability (or state-space) even when there are no system observations available. Whilst the iteration time is evolving, the chain is eventually becoming a sample of the posterior. In the first iteration, the

initial chain state $\boldsymbol{\theta}^{(0)}$ starts moving along each dimension at a time since only one component is changed. In contrast, the dynamics are refined for that window of length N for each iteration j. A new pseudo-trajectory of length N is generated for each iteration j. There is no temporal evolution of the states but a new realisation of the deterministic system states which best resemble the data S given the dynamical model chose. There is no dynamical evolution involved.

The algorithm that generates a chain of J states for each of the components of $\boldsymbol{\theta}$ is presented in Figure 3.2.1.4.

```
# Set chain initial conditions
```

$$\boldsymbol{\theta}^{(0)} = \left(x_0{}^{(0)}, x_1{}^{(0)}, \ldots, x_t{}^{(0)}, \ldots, x_N{}^{(0)}, a^{(0)}, \sigma_\delta^{2(0)}\right)$$

```
# Generate the Markov chain states
Loop (j = 1 to J)
{
```

$$\texttt{Sample } \boldsymbol{x_0}^{(j)} \sim \pi_S(x_0|x_1{}^{(j-1)}, \ldots, x_N{}^{(j-1)}, a^{(j-1)}, \sigma_\delta^{2(j-1)})$$

$$\texttt{Sample } \boldsymbol{x_1}^{(j)} \sim \pi_S(x_1|\boldsymbol{x_0}^{(j)}, x_2{}^{(j-1)}, \ldots, x_N{}^{(j-1)}, a^{(j-1)}, \sigma_\delta^{2(j-1)})$$

$$\vdots$$

$$\texttt{Sample } \boldsymbol{x_t}^{(j)} \sim \pi_S(x_t|\boldsymbol{x_0}^{(j)}, \boldsymbol{x_1}^{(j)}, \ldots, \boldsymbol{x_{t-1}}^{(j)}, x_{t+1}{}^{(j-1)}, \ldots, x_N{}^{(j-1)}, a^{(j-1)}, \sigma_\delta^{2(j-1)})$$

$$\vdots$$

$$\texttt{Sample } \boldsymbol{x_N}^{(j)} \sim \pi_S(x_N|\boldsymbol{x_0}^{(j)}, \boldsymbol{x_1}^{(j)}, \ldots, \boldsymbol{x_t}^{(j)}, \ldots, \boldsymbol{x_{N-1}}^{(j)}, a^{(j-1)}, \sigma_\delta^{2(j-1)})$$

$$\texttt{Sample } \boldsymbol{a}^{(j)} \sim \pi_S(a|\boldsymbol{x_0}^{(j)}, \boldsymbol{x_1}^{(j)}, \ldots, \boldsymbol{x_t}^{(j)}, \ldots, \boldsymbol{x_N}^{(j)}, \sigma_\delta^{2(j-1)})$$

$$\texttt{Sample } \boldsymbol{\sigma_\delta^{2(j)}} \sim \pi_S(\sigma_\delta^2|\boldsymbol{x_0}^{(j)}, \boldsymbol{x_1}^{(j)}, \ldots, \boldsymbol{x_t}^{(j)}, \ldots, \boldsymbol{x_N}^{(j)}, \boldsymbol{a}^{(j)})$$

```
}
```

MCMC techniques can be seen as an alternative method for generating pseudo-orbits for the map f. This point is discussed in more detail in section 3.3 and Chapter 5 where pseudo trajectories generated using MCMC methodology and gradient descent (in Chapter 4) are compared.

From the description of the MH algorithm, the generation of the chain seems to be a straight forward procedure, all full conditionals are known explicitly as summarised in 3.2. All components of $\boldsymbol{\theta}$ are easily sampled from their corresponding full conditionals, except for the components corresponding to the initial condition and latent states where full conditional are not in a closed form.

For the components corresponding to $\{x_t\}_{t=0}^{N-1}$ a sample generating routine was implemented. The routine might be useful in other implementations of Bayesian state-space models for quadratic maps, since the quartic exponential comes from the quadratic term in the map [23].

In those cases, the full conditional distribution is not in a closed form, so standard algorithms for random sampling generation cannot be used. Therefore, in order to generate samples for $\{x_t\}_{t=0}^{N-1}$, the *Accept/Reject Algorithm* for sampling from a general density $f_Y(y)$ is used.

This algorithm is presented in [17] as the following theorem:

THEOREM 3.2.1 *Let $Y \sim f_Y(y)$ and $V \sim f_V(v)$, where $f_Y(y)$ and $f_V(v)$ have common support with*

$$M = \sup_y \frac{f_Y(y)}{f_V(v)} < \infty. \tag{3.39}$$

From Theorem 3.2.1, the steps to be followed in order to generate a random variable $Y \sim f_Y(y)$ are shown in Figure 3.2.1.4.

`Step 1. Sample a uniform random variable:` $U \sim \mathcal{U}(0,1)$

`Step 2. Sample a random variable:` $V \sim f_V$

`Step 3. If:` $U \leq \frac{1}{M}\frac{f_Y(v)}{f_V(v)}$ `then:` $Y = V$

`otherwise: Go to step 1.`

Note that when the Accept/Reject algorithm is used to generate random samples of Y, it is necessary to ensure that the condition (3.39) is fulfilled in all points belonging to the range where $f_Y(y)$ is defined, otherwise the algorithm will generate samples very slowly. Also note that this algorithm does not depend on any of the normalisation constants of $f_Y(y)$ and $f_V(v)$.

The functional form of $f_Y(y)$ is

$$f_Y(y) \propto \mathrm{e}^{-Ay^4+By^2+Cy}, \tag{3.40}$$

where the exponent is a quartic polynomial. For all $\{x_t\}_{t=0}^{N-1}$, $A > 0$ and for the initial condition x_0, $C = 0$.

The quartic exponential in (3.40) requires that some conditions on A, B and C to hold in order to obtain a convex probability density function and for many values of A, B and C equation (3.40) is symmetric and bimodal. Details on how to constrain the choice of $f_V(v)$ in order to obtain samples of $f_Y(y)$ with a low frequency of resection (in step 3 of of the accept/reject algorithm) are discussed in [23].

The fact that (3.40) is bimodal makes the convergence of the chain slow and could be the reason why WinBUGS did not produce reliable Markov chain states. As described in section 3.2.1.3, sampling algorithms, such as the ones used in WinBUGS, could fail to generate reliable samples from such probability distributions.

Before doing any inference of $\boldsymbol{\theta}$, the MCMC output has to be tested for convergence. Such tests will determine the burn-in time τ, *i.e.* the time when the mixing of sampled components of $\boldsymbol{\theta}$ is reached or is stable. After τ iterations, each state of $\boldsymbol{\theta}$ is a state that is drawn from the posterior. In this work, the convergence tests used are "by-eye" methods where the trace is plotted against the iteration time, and the GR-statistic [31] calculated for parallel runs of the algorithm. Once reliable samples of the chain are generated following the procedure described for the Metropolis-Hastings algorithm, any inference of $\boldsymbol{\theta}$ can be obtained.

3.3 Distinguishing Dynamics

Several experiments are performed on data sets with different qualities and characteristics. The inferences obtained from each experiment test the performance of the Bayesian methodology for the Logistic map given that the answer is known. Even though the conditions of the experiments are set in the PMS, *i.e.* data is generated in the PMS, and following the discussion in 3.2.1.2, the Bayesian probability model formulated in the NSA Logistic's noisy data is used. Recall, the experiments are performed in the NSA, since the data contains only measurement noise.

Table 3.3 lists the types of data sets used to feed the Bayesian state-space model for the Logistic map. Type 1 observations are noisy segments of length N of a Logistic trajectory for fixed a and known initial condition x_0. Two values of the Logistic parameter are used, $a = 1.35, 1.85$, corresponding to periodic and chaotic trajectories. Initial condition is taken to be $x_0 = 0.3$.

Type 2 data sets are *surrogates* generated to be normally distributed with mean and variance corresponding to $\mu_x = 0.227$ and $\sigma_x^2 = 0.372$. These values are the mean and variance of the true dynamical states. On top of that a noise component is added to the data set in order to simulate noisy data of Type 1 and lengths $N = 100, 512$ per noise level.

Type 3 data is a second group of surrogates generated as a random draw

of the type 1 observations of length $N = 100$ of type 1. Randomisation of the Logistic observations is obtained by shuffling the time index $t = 1, \ldots, N$, and different realisations are generated by shuffling the time indices for several random seeds.

All three data types are statistically indistinguishable up to second order, *i.e.* shared mean and variance of the noise free Logistic trajectory.

Type	Description
1	$s_t = x_t + \eta_t$, where $x_t = 1 - ax_{t-1}^2$, $t = 1, \ldots, N$. $\eta_t \overset{iid}{\sim} \mathcal{N}(0, {\sigma_\eta}^2)$ for $0 \leq {\sigma_\eta}^2 \leq 2$
2	$r_t = q_t + \eta_t$, where $q_t \overset{iid}{\sim} \mathcal{N}(\mu_x, \sigma_x^2)$, $t = 1, \ldots, N$. $\mu_x = \text{mean}(\{x_t\}_{t=1}^N)$ and $\sigma_x^2 = \text{var}(\{x_t\}_{t=1}^N)$. $\eta_t \overset{iid}{\sim} \mathcal{N}(0, {\sigma_\eta}^2)$, with $0 \leq {\sigma_\eta}^2 \leq 2$.
3	m_t is a random draw of type 1 data sets.

Table 3.3: List of data types on which the MCMC implementation is applied.

Despite the wide spectrum of data sets used to obtained MCMC estimates of the Logistic parameter, this section focuses on the comparison of the results for data sets of type 1 and 2. A detailed and extensive description of the results obtained for the other data sets are to be presented in [24] and elsewhere. A more complete analysis is planned for future research.

Figures 3.6 and 3.7 show in the left panel the reconstruction of noisy observations (+) on top of the reconstruction of noise free states (grey dots) for both data types, 1 and 2, respectively. In the right panel of Figures 3.6 and 3.7, two histograms are overlapped, one for the noisy observations (black) and another for the noise free states (grey).

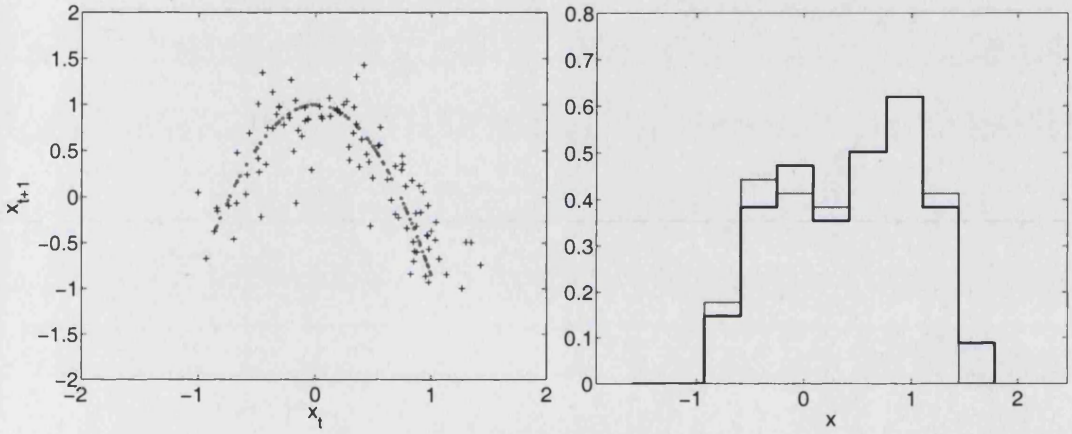


Figure 3.6: Data set type 1: Noisy observations. Right: Reconstruction of the noise free states (grey dots) and the noisy observations (+) for noise level of 0.2. Left: Histograms of the noise free states in grey and the noisy observations in black.

The left panel of Figure 3.7 clearly shows that there is no evidence of any structure resembling the dynamics of the Logistic map; neither is the *invariant measure* evident from the histogram in the right panel due its nonexistence in the data and to the short length of the data sets.

Application of the Bayesian techniques to both data types is made in

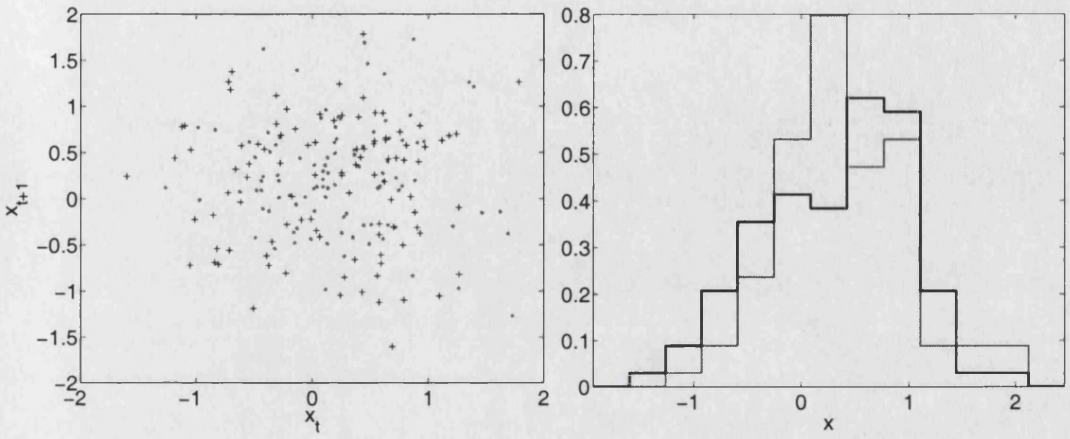


Figure 3.7: Data set type 2: Surrogate data. Right: Reconstruction of the noise free states (grey dots) and the noisy observations (+) for noise level of 0.2. Left: Histograms of the noise free states in grey and the noisy observations in black.

order to test MCMC parameter estimation performance for the Logistic map. MCMC is applied to type 1 observations of length $N = 100$ in order to check for sensitivity of the burn-in time in at least five realisations of the noisy observations. For each of the data types described in Table 3.3, a chain for the parameter vector $\boldsymbol{\theta}$ is generated from $T = 1.1 \times 10^5$ iterations of the algorithm.

Inference of any component of $\boldsymbol{\theta}$ is reliable once convergence of the chain is assessed, *i.e.* burn-in time is estimated. In order to asses chain convergence and since there is only one realisation of the chain for each data type and each noise level, the Gelman-Rubin (GR) statistic [31] is calculated by segmenting the chain into two segments of equal length (*i.e.* 5.5×10^4). See definition

in Chapter 2, section 2.1. The closer the GR statistic value is to 1, the more overlapping of the parallel Markov chains is evident [95]. Generally, GR values less than 1.2 are considered to be acceptable, passing the mark for convergence assessment [95, 13]. In the case that GR $>$ 1.2 for "all" summary statistics of interest, it is safe to continue with further iterations until GR $\in [1.0, 1.2]$.

In section 3.2.1.3 where the WinBUGS software is used, the output is thinned each 20th chain state to reduce the correlation between samples if the convergence is not fully reached. In this section, all chain states for $t > \tau + 1$ are taken for inference purposes.

Figures 3.8 to 3.11 show the GR values for two summary statistics of the MCMC output as a function of the iteration time. Each Figure is composed of two panels, in the left the GR values of the MCMC output from type 1 (Logistic) data are plotted whilst in the right the corresponding estimates are plotted for type 2 (surrogates) data, both with noise level $\sigma_\eta^2 = 0.2$. The GR statistic is calculated for the median (dashed line) and variance (dotted line) of the resulting chain, and the pass mark of the test is represented by a horizontal line. Figure 3.8 plots GR values for estimations of the Logistic parameter a, Figure 3.9 for the dynamical noise variance σ_δ^2, Figure 3.10 for the initial condition x_0, and Figure 3.11 for the state estimate for x_{37}.

In both panels of Figure 3.8, convergence is fully reached for the median

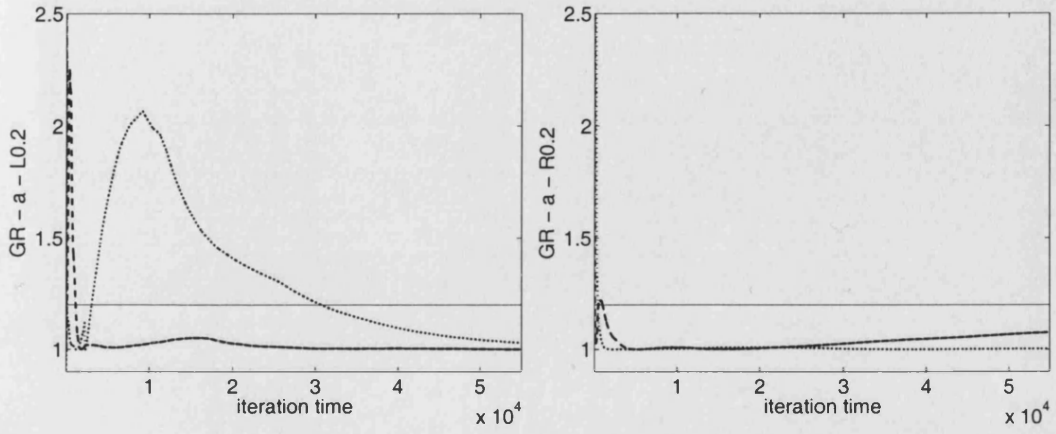


Figure 3.8: GR statistic for the median (dashed line) and variance (dotted line) of the Logistic parameter a. MCMC output is generated for type 1 (left) and type 2 (right) data with a noise level of $\sigma_\eta^2 = 0.2$. The horizontal line corresponds to the pass mark for the test.

estimates, in the GR sense, after 1×10^4. The chain is satisfactorily converging in median towards to the posterior distribution since the GR-values tend to one for both data types.

In contrast, the chain is converging in variance towards the posterior only for the data type 2 (surrogates) after few iterations. As seen in the left panel of Figure 3.8, the GR values for the variance of the distribution of a-estimates for the data type 1, are less than one only after approximately 2000 iterations. Suddenly, the GR statistic looses its skill at around 10^4 iterations, *i.e.* GR > 1, and shows a slow decrease towards 1, *i.e.* slow convergence of the chain. Although this is the case, $GR \leq 1.2$ soon after 3×10^4 iterations.

Convergence in variance to the posterior is slow and non-uniform when type 1 data is used to feed the MCMC algorithm.

Convergence in variance is reached faster when the observations are of type 2, *i.e.* when there is no dynamical information in the observations. The difference between variance convergence rates in both chains shows that the MCMC technique shrinks the uncertainty on the Logistic parameter more efficiently for the surrogate data set than the Logistic noisy data set.

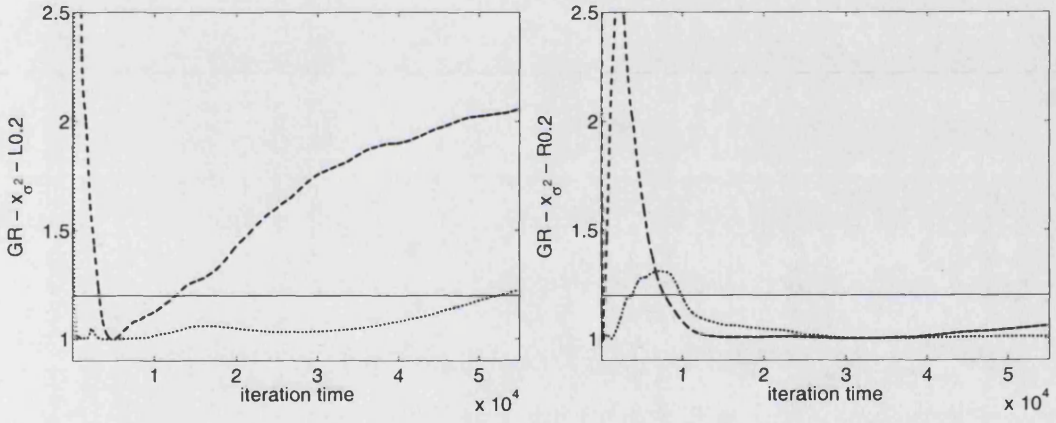


Figure 3.9: GR statistic for the median (dashed line) and variance (dotted line) of the dynamical noise variance σ_δ^2. MCMC output is generated for type 1 (left) and type 2 (right) data with a noise level of $\sigma_\eta^2 = 0.2$. The horizontal line corresponds to the pass mark of the test.

Similar results are obtained for the GR statistic of the dynamical noise variance σ_δ^2 chain obtained using type 1 data sets. This plot is shown in the left panel of Figure 3.9. Poor and slow convergence is also shown for the

median and variance estimates of this chain. In contrast, convergence is very fast and uniform when the data set is of type 1. As featured in Figure 3.8, convergence is reached faster when the observations used to evaluate the posterior do not contain dynamical information. This behaviour of the GR statistic is consistent for all noise levels and data types considered.

Given that convergence has to be assessed in all components of $\boldsymbol{\theta}$ in order to validate the samples obtained by MCMC as samples of the posterior [31, 95, 13]; Figures 3.10 and 3.11 plot GR values for other components of $\boldsymbol{\theta}$ as a function of the iteration time. As described in last sections, the other components of the parameter vector are directly related with the dynamical nature of the Logistic model. This components correspond to the initial condition, x_0, and the latent states, x_t.

The GR values calculated for the mean and variance of the distribution of estimates of the initial condition, x_0, and the latent state, x_{37}, display similar features of convergence as the ones described for the Logistic parameter a, and the dynamical noise amplitude σ_δ^2.

It is clear from Figure 3.10 that whilst convergence is fully reached in mean and variance for the chain generated using type 1 observations after 10^4 iterations; the chain generated using type 2 data, *i.e.* noisy Logistic observations, is converging weakly, GR ≈ 1.2.

In the inset of the left panel of Figure 3.10, a zoom of the GR values

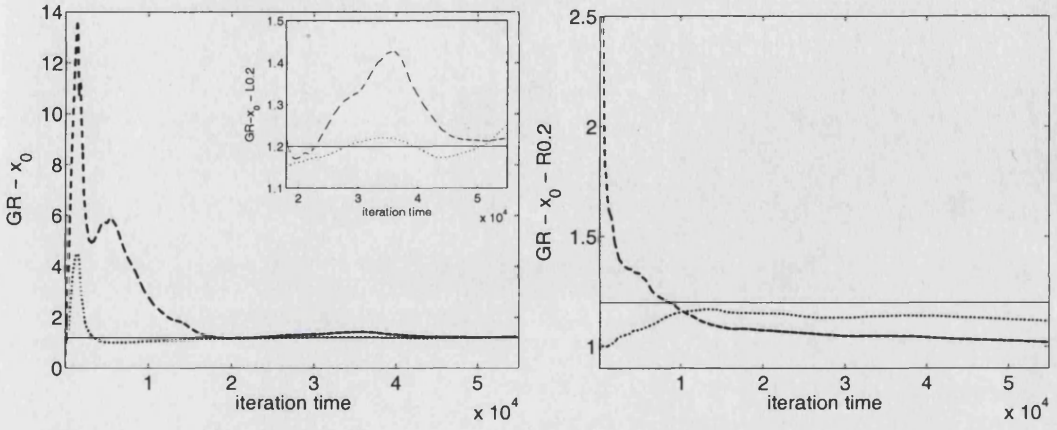


Figure 3.10: GR statistic for the median (dashed line) and variance (dotted line) of the initial condition x_0. MCMC output is generated for type 1 (left) and type 2 (right) data with a noise level of $\sigma_\eta^2 = 0.2$. The horizontal line corresponds to the pass mark for the test.

for iterations between 1.8×10^4 and 5.5×10^4 is shown. In that period of iteration time, convergence is weak and oscillating around the pass mark of the test.

Figure 3.11 shows similar features of convergence of the chain obtained for the latent state x_{37}. Poor convergence is reached for both the median and variance of the resulting distribution of samples of x_{37} and for any $t = 1, \ldots, 100$.

In addition, the GR values calculated for the WinBUGS output obtained in section 3.2.1.3 and the convergence plots display there, clearly show that convergence reached by the "Tailored" implementation of the MCMC tech-

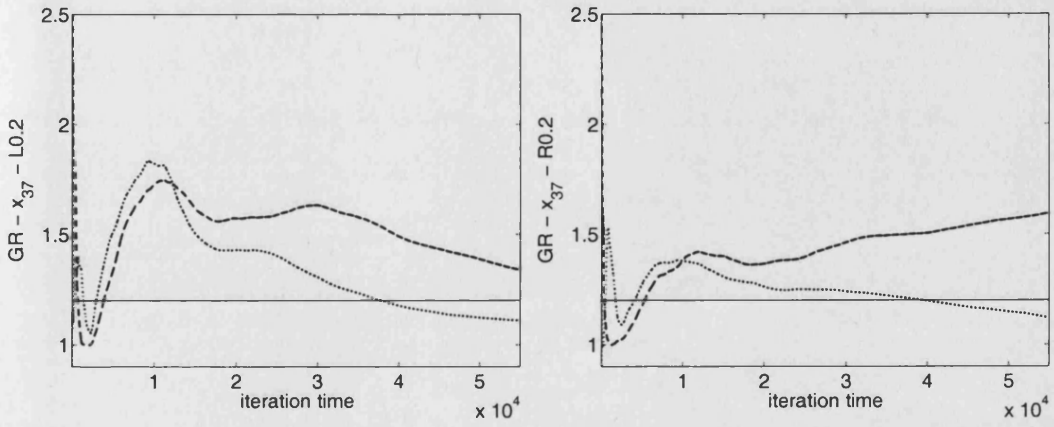


Figure 3.11: GR statistic for the median (dashed line) and variance (dotted line) of the latent state, x_{37}. MCMC output is generated for type 1 (left) and type 2 (right) data with a noise level of $\sigma_\eta^2 = 0.2$. The horizontal line corresponds to the pass mark for the test.

niques is more robust than for the convergence display in the chain obtained by WinBUGS software. Even though, oscillatory behaviour of the stable state of the chain (as shown by results obtained from the WinBUGS implementation of MCMC), in the case of the "tailored" implementation, oscillations are smaller in amplitude. Reasons for itermittency or oscillation of chain convergence are directly related with the multimodality of the chaotic Likelihood associated with the Logistic map as discussed in section 3.2.1.3 and in [5]. Mitigation of this effect is obtained by the "Tailored" implementation since the sampling algorithms using in WinBUGS [21] are no longer used.

GR values for the chain obtained using all data types, noise levels and

parameter vector components consistently show that higher quality convergence is reached when type 2 observations rather than type 1 observations are used to generate a chain by the "Tailored" implementation of MCMC. There is no evidence that this distinction could bring any light onto the distinction of deterministic and random behaviour. On the contrary, it highlights the type 1 data as time series generated from a deterministic system in this particular example and future research is planned to tackle such features of the MCMC estimates.

Convergence is assessed by the GR test and the burn-in time is set to be $\tau = 1 \times 10^4$ iterations. All samples obtained after τ iterations of the algorithm are considered as samples from the posterior, and inferences can be made by Monte Carlo approximations on the resulting chain of length 1×10^5.

Remark that this section is focused on the deterministic structure of the state estimates rather than the Logistic parameter estimates themselves. As a consequence, evidence of determinism is searched for the resulting MCMC distributions of estimates for the latent states of the Logistic map.

The reconstruction of the estimated x_t's for several noise levels is shown in Figure 3.12 after discharging τ burn-in samples. The Figure is composed by 4 panels, in each panel the reconstruction of the true states and the median state estimates both the type 1 and 2 data sets are shown. From left

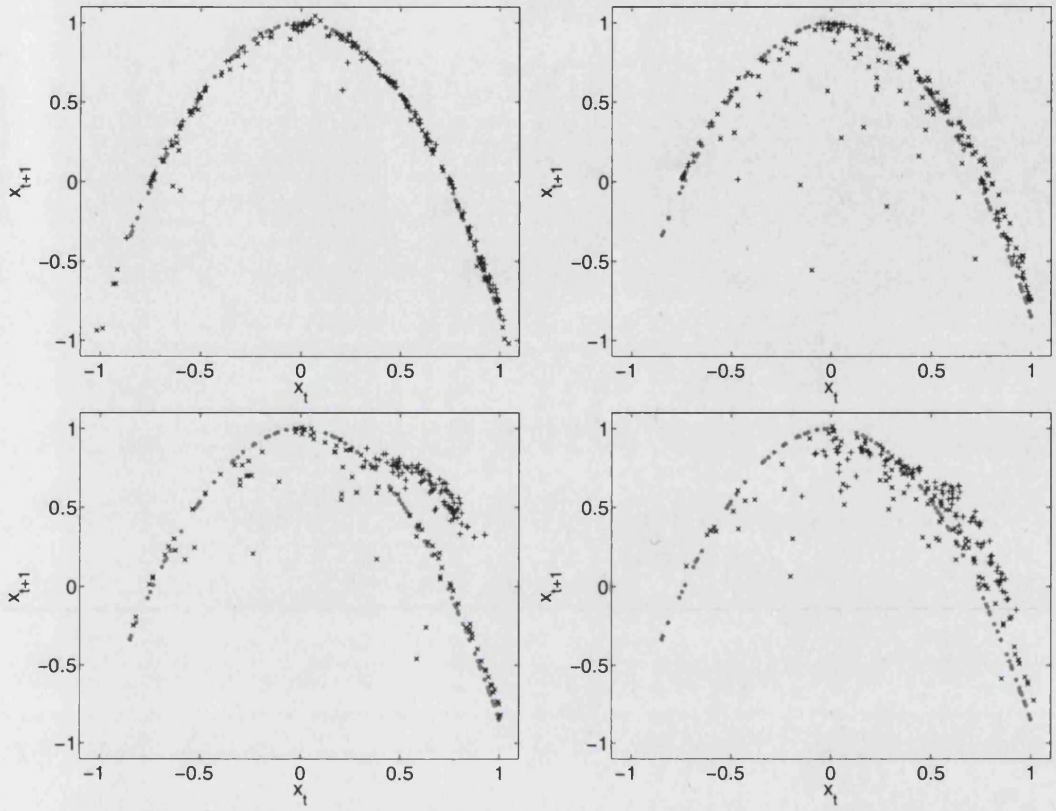


Figure 3.12: Delay plots of the Logistic and estimated MCMC states for noise levels of $\sigma_\eta^2 = 0.4, 0.8, 1.2, 2.0$, left to right. Grey dots are used for true Logistic states, (+) and (×) for the median Logistic states estimates from type 1 and type 2 data sets, respectively.

to right, the panels correspond to noise levels of $\sigma_\eta^2 = 0.4, 0.8, 1.2$ and 2.0, equivalent to 20%, 40%, 60% and 100% of the noise free signal approximately. The symbols in each panel correspond as follows: grey dots to true Logistic states, (+) and (×) to median state estimates from MCMC output using noisy Logistic observations and surrogates, respectively.

Despite the presence of high noise amplitudes, the short length of the data set and even the non-existence of the dynamics in the surrogates, the reconstructions clearly resemble the Logistic map structure. For some noise levels, the Logistic structure is even more evident in the reconstruction of the state estimates obtained from surrogates than in the reconstruction of the noisy one obtained from Logistic data. The median estimates of the Logistic states tend to cluster around the stable point of the Logistic map. The order of the standard errors for the calculated posterior medians corresponding to each of the x_t's for $t = 1, \ldots, N$ are consistently of the same order for both data sets, as can be seen for the state x_{37} in Figure 3.13.

In addition, the Bayesian methodology provides a robust de-noising feature (see Figure 3.12), based not on embedded measurement variables [57, 27, 48] but on the probabilistic structure of the state-space.

Posterior estimates for several components of the parameter vector $\boldsymbol{\theta}$ are calculated and shown in Figures 3.14 to 3.17. These Figures shown the posterior estimations and their corresponding uncertainty measures for the same components of $\boldsymbol{\theta}$, namely: Logistic parameter a in Figure 3.14, initial condition x_0 in Figure 3.15, absolute value of the initial condition in Figure 3.16 and dynamical noise amplitude σ_δ^2 in Figure 3.17.

Each of these Figures displays two panels. The left panel plots posterior estimates for data type 1 and the right panel for data type 2. The summary

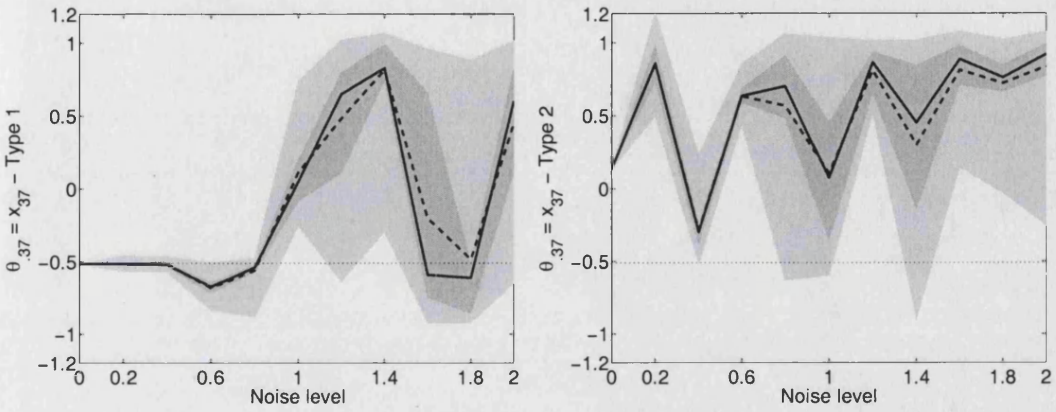


Figure 3.13: Posterior mean estimates for $0 \leq \sigma_\eta^2 \leq 2$ the Logistic state x_{37} for data type 1 (left) and 2 (right) as a function of the noise level $0 \leq \sigma\eta^2 \leq 2$. Mean and median are displayed as a solid and dashed black line. The light grey area covers the values between the isopleths of 2.5% and 97.5% and the darker grey area covers values between the 25% and 75% isopleths. The true state value is marked with a horizontal line.

statistics calculated are the median and the mean, plotted in a solid and dashed black line, respectively. True parameter values are marked by a horizontal dotted line. The light grey areas correspond to the values between the 2.5% and 97.5% isopleths of the resulting posterior distribution and the darker grey areas to the values between the 25% and 75% isopleths. All posterior estimations are calculated for only one realisation of measurement noise.

Posterior estimates for the Logistic parameter a in Figure 3.14, show that for both data types the Bayesian methodology produces a satisfactory value

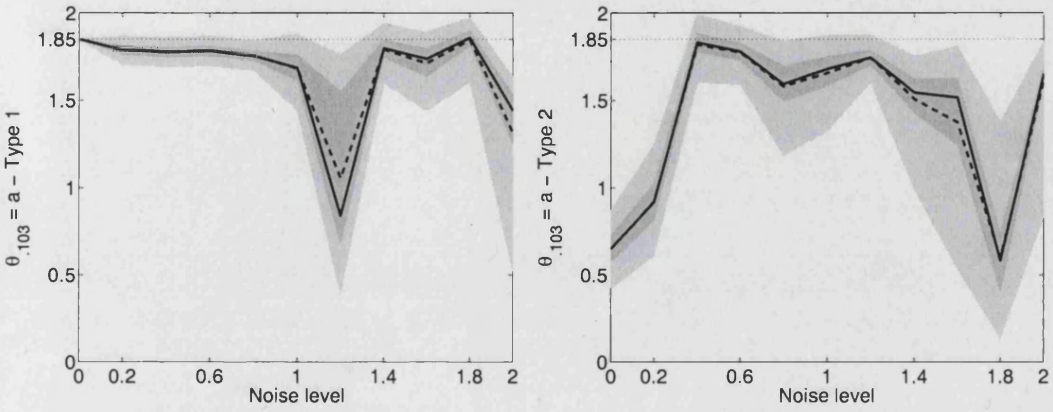


Figure 3.14: Logistic parameter, a, posterior mean estimates as a function of the noise level, $0 \leq \sigma_\eta^2 \leq 2$ for both data types, 1 (left) and 2 (right). Mean and median are displayed as a solid and dashed black line, respectively. The light grey area covers the values between the isopleths of 2.5% and 97.5% and the darker grey area covers values between the 25% and 75% isopleths. The true parameter value is marked with a horizontal line.

of a for several noise levels. For both types of input data, the performance of the technique on parameter estimation is markedly superior in comparison with traditional methods (*i.e.* such as Least Squares, *e.g.* Figure 1 in [68]) and for high noise levels.

Surprisingly, the left panel in Figure 3.14 shows that the posterior estimates obtained for the surrogates, where there is no dynamical information, are consistent in behaviour with the estimates in the right panel. In both cases, estimates stay close to the true value even for large noise levels. If one was given only Figure 3.14 in isolation then it would be very difficult task to

discern which of the two of the original data sets corresponds to the Logistic observations (*i.e.* contains a chaotic structure in the state space), and the possible answer to that question might be certainly ambiguous.

Figure 3.15 clearly shows the multimodality of some of the posterior pro-sections onto components of the parameter vector $\boldsymbol{\theta}$. Given a noise free Logistic map state at time t, calculation of the state at time $t-1$ for a known value of a implies finding $f^{-1}(x_{t-1};a)$, *i.e.* $x_{t-1} = \pm\sqrt{1-x_t}$. For a noisy Logistic observation, the same calculation implies accounting for the experimental uncertainty of the observations by calculating the state at time $t-1$ as for the case of the noise free Logistic state over an ensemble of values around the true state. In this case, the estimated value for the state at time $t-1$ would result in a bimodal distribution around $\pm x_{t-1}$. From this argument it is easy to see how the multimodality is natural in the posterior estimates of the latent Logistic states x_t.

The resulting posterior values of the initial condition x_0 display multimodality and symmetry with respect to zero as expected for the Logistic map from the argument in the paragraph above and from the discussion of Berliner in [5] and in section 3.2.1.2. Going backwards in time to find the initial condition that generated the true trajectory involves binary decisions each time the inverse of the map is solved. For both data types, results are very similar even though the posterior distribution tends to be wider when

surrogates are used to feed MCMC.

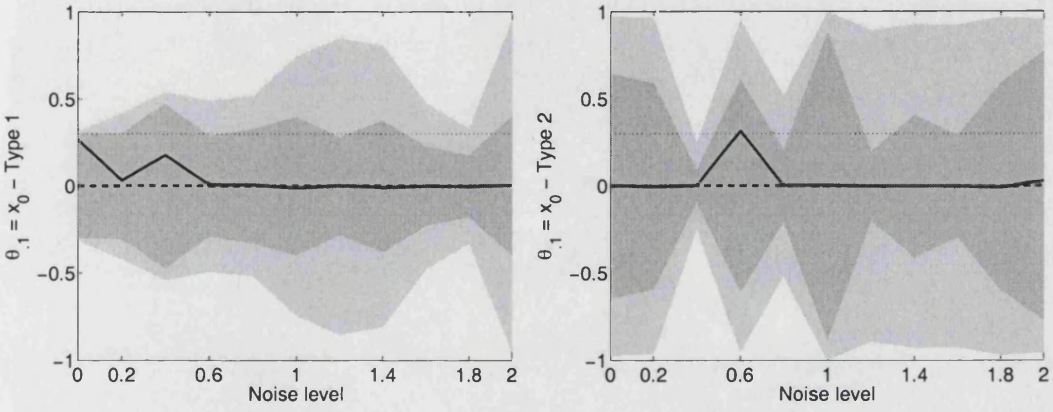


Figure 3.15: Posterior mean estimates for $0 \leq \sigma_\eta^2 \leq 2$ the initial condition x_0 for data type 1 (left) and 2 (right) as a function of the noise level $0 \leq \sigma\eta^2 \leq 2$. Mean and median are displayed as a solid and dashed black line. The light grey area covers the values between the isopleths of 2.5% and 97.5% and the darker grey area covers values between the 25% and 75% isopleths. The true value of the initial condition is marked with a horizontal line.

For this form of the Logistic map, $f(x_{t-1}; a) = 1 - ax_{t-1}^2$, and only in the PMS, it is valid to calculate estimates of the initial condition using the absolute values of the posterior obtained by MCMC. All results presented here generalise to all one-dimensional quadratic maps since all are dynamically equivalent in behaviour when parameter values are in the chaotic regime [73].

Figure 3.16 shows posterior estimates for the absolute value of the Logistic map's initial condition, $|x_0|$, for both data types used. In this case, location of the posterior modes is closer to $|x_0| = 0.3$ for estimates obtained from noisy

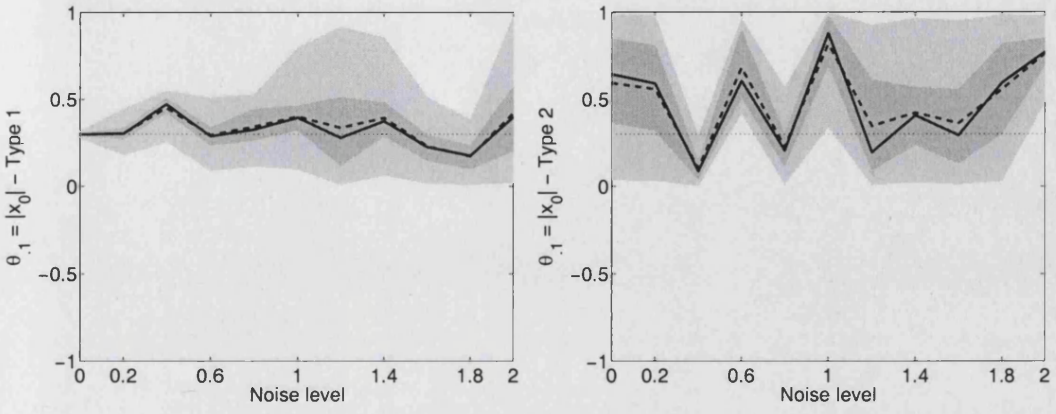


Figure 3.16: Posterior mean estimates for $0 \leq \sigma_\eta^2 \leq 2$ the initial condition x_0 for data type 1 (left) and 2 (right) as a function of the noise level $0 \leq \sigma\eta^2 \leq 2$. Mean and median are displayed as a solid and dashed black line. The light grey area covers the values between the isopleths of 2.5% and 97.5% and the darker grey area covers values between the 25% and 75% isopleths. The absolute value of the initial condition is marked with a horizontal line.

Logistic observations than for surrogates for all noise levels. It is surprising that estimates from surrogates oscillate around the true value and are always less than 1, *i.e.* even though there is no dynamical information in the data used, the methodology drifts initial chain states to the relevant domain for the values of x_0.

Figure 3.17 shows the posterior estimates for the amplitude of the dynamical noise parameter, σ_δ^2. In both cases, as expected from prior information, the dynamical noise variance satifies $0 < \sigma_\delta^2 \ll 1$. The dynamical noise component keeps close to zero for all noise levels when estimates are obtained

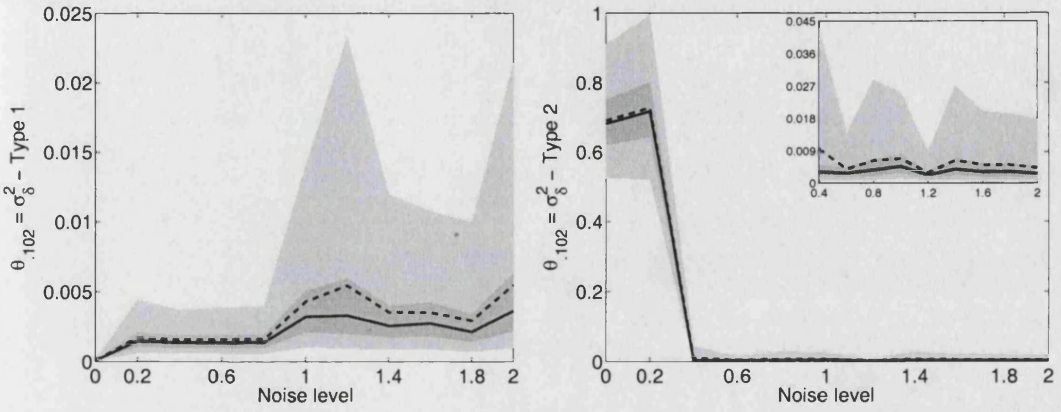


Figure 3.17: Posterior mean estimates for $0 \leq \sigma_\eta^2 \leq 2$ the initial condition σ_δ^2 for data type 1 (left) and 2 (right) as a function of the noise level $0 \leq \sigma\eta^2 \leq 2$. Mean and median are displayed as a solid and dashed black line. The light grey area covers the values between the isopleths of 2.5% and 97.5% and the darker grey area covers values between the 25% and 75% isopleth.

from noisy Logistic observations. Although, this is the case as well for estimates obtained from the surrogates for noise levels $0.4 \leq \sigma_\eta^2 \leq 2.0$, the dynamical noise variance is $\mathcal{O}(\sigma_\delta^2) = 1$ for $\sigma_\eta^2 = 0, 0.2$. Coincidentally, this behaviour corresponds to low estimates obtained for data type 2 of the Logistic parameter a, see Figure 3.14, for the same noise levels. This behaviour can be seen as a result of a particular realisation of the surrogates and from the non-uniform chain convergence (see Figure 3.9) obtained for surrogates for small noise levels.

In addition, this fact stresses the role of the dynamical noise inclusion in the Bayesian model of the Logistic map as a control parameter that ac-

counts for model error. Therefore, posterior distribution width is interpreted as model uncertainty. It is evidence of the strain experienced by the MCMC simulation to obtain parameter estimates from observations that are represented by a model that is far away from the real system where the observations are obtained from.

Figure 3.18 shows the histograms of the posterior distributions obtained for the Logistic parameter a and the initial condition x_0. The Figure is composed by four panels, each displaying two histograms, one corresponding to the posterior estimates using data type 1 in black and another using data type 2 in grey. The left column shows histograms for the Logistic parameter a and the right column for the initial condition x_0. The rows of the Figure display the posterior histograms for noise levels corresponding to $\sigma_\eta^2 = 0.4, 2.0$, from top to bottom. Thus, for example, the panel in the top right corner corresponds to the histograms of the initial condition from data types 1 (grey) and 2 (black) with noise level of $\sigma_\eta^2 = 0.4$. True parameter values are marked with a horizontal line.

The histograms are consistent with the convergence features discussed earlier and the ambiguity presented by the methodology when discerning between deterministic and random behaviour.

Histograms for the Logistic parameter a, show that the method is unable to distinguish between the deterministic data set and the surrogates.

Histograms obtained for both data types are very similar in shape and location of the mode or modes. Even more, wider posterior distributions are produced for noisy Logistic observations than when surrogates are used to generate estimates.

In the left column, the initial condition histograms clearly show the high complexity of the posterior obtained; posterior distributions with many modes and high tails given the quadratic and chaotic nature of the map. As for the case of the Logistic parameter a, estimates from surrogates display a better and smoother behaviour than those observed for data type 1, even for high noise levels.

For both data sets and all noise levels, the positions of the modes are located in the close neighbourhood of the true value for a and x_0.

Multiple peaks and coarseness in the resulting densities are evidence of the slow and non-uniform convergence of the chain to the stationary distribution [85] and of the inadequacy of the model used to represent the observations. In contradiction, it seems that the observations of type 1, the ones that contained dynamical observations, are more inaccurately represented by the Bayesian model than the surrogates.

If one was given only the convergence analysis, posterior estimates and histograms of the resulting chain obtained for both data types, and was asked

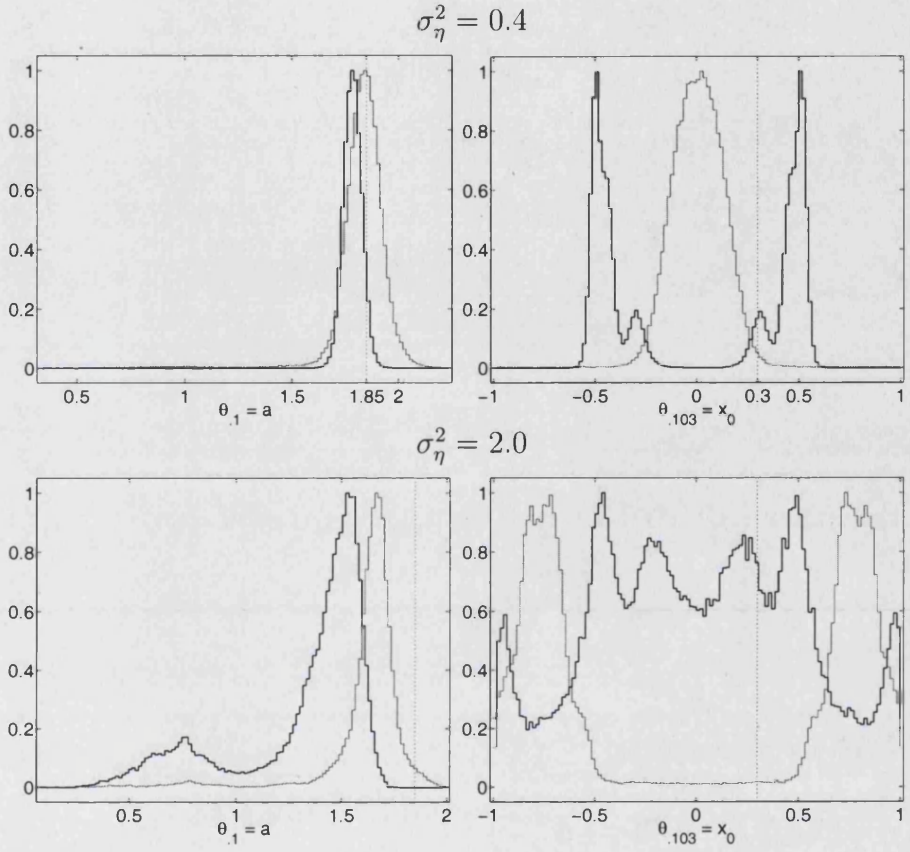


Figure 3.18: Histograms of the posterior samples obtained by MCMC for the Logistic parameter a (left column) and the initial condition x_0 (right column). Histograms in the top row correspond to posterior components obtained for a noise level $\sigma_\eta^2 = 0.4$ and for $\sigma_\eta^2 = 2.0$ in the bottom row. Each panel displays two histograms, one for the posterior samples obtained from noisy Logistic observations in black and surrogates in grey. True parameter values are marked with a vertical line.

to choose which of the two data types contains deterministic features from the MCMC chain it produces; the scale tends to point to a type 2 data

set as a candidate of containing deterministic features, even in the PMS. Current work [24] and future research is planned to tackle this ambiguity by formulating tests based in Ockham's razor ideas [72, 65]. This ambiguity is directly related with the open problem of distinction from deterministic and random behaviour and also to the discussion presented in Chapter 5.

The Bayesian approach seems to provide too good estimates and features a correctness in the parameter estimates that seems to be too good as it is presented in the relevant literature [70, 12]. From the study performed here, it seems too good to be true.

Complexity of the analytical posterior [5, 89] as discussed in section 3.2.1.2, numerical issues related to the resolution of the initial condition [52], inclusion of dynamical noise to solve numerically the posterior as discussed in section 3.2.1, and convergence issues [21, 24] of the resulting chains, make estimations obtained by MCMC techniques suspicious numbers that should be taken with caution.

Even more, it provides "good" and "close" estimates for parameter values in cases where it is known that they do not even exist. This study emphasises the care that should be taken when applying Bayesian methodologies to deterministic systems, as also discussed in Chapter 6 and in [24, 22]. Specially, in real cases where the existence of the model is questionable and model parameters may not have a correspond pair to the reality of the system [87].

This work does not confirm that the Bayesian method, through MCMC techniques, provides "unbiased parameter estimations" for deterministic chaotic systems [70, 12]. Instead, it warns that the estimations are provided by a "biased numerical implementation" when applied to noisy chaotic observations. Remark that this is true only in the case that the observations used contain only measurement noise. Here, bias is understood in the sense that, for the probability model studied, the correct model is impossible to be implemented numerically, therefore, a Naive Statistic Approach is used instead.

3.4 Summary

The results described in this chapter clearly reflect a series of ambiguities when the Bayesian methodology is applied to a signal from nonlinear, potentially, chaotic systems. In particular, Bayesian methodologies are correctly imported to the nonlinear framework to solve the problem of model parameter estimation from experimental observations within the PMS. Remark that the parameter estimation problem is defined in the PMS defined precisely by two conditions: the model representing the data corresponds exactly to the system from which the observations are taken and experimental observations contain both measurement and dynamical noise, as described in section 3.1.

If observations of the system contain only measurement noise, the Bayesian

formulation is forced to change fundamentally in interpretation since the observations are not obtained in the PMS. If the formulation is not changed and it is applied to observations with only measurement noise then is only done to make the problem numerically tractable. This fundamental change is made for the sake of numerical tractability, not with the interest of finding a way to model the observations in a "more realistic" way as argued in [70].

In section 3.2.1.2, the assumption of stochastic transition over time is seen as an unavoidable step in the Bayesian modelling process for deterministic chaotic time series. Without this assumption the *chaotic* Likelihood [5] is not well behaved, preventing the explicit calculations of the full conditionals.

Henceforth, several difficulties in the numerical implementation of the MCMC technique for chaotic models are implied since the resulting posterior and full conditional distributions display multimodality. Multimodal densities are difficult to sample from and in addition deeply affect convergence of samples as samples of the posterior [13, 95, 71, 21, 23], one of the major flaws of the version of WinBUGS used in 2002. As detailed in section 3.2.1.3, multimodality makes the process slow and less reliable since the convergence is not uniformly reached in some components of the parameter vector $\boldsymbol{\theta}$, and traces of the WinBUGS and MCMC output seem to jump in the state-space in an intermittent way.

Given these convergence issues, the MCMC algorithm is implemented in-

dependently of the WinBUGS package, developing a sampling routine for the initial condition and the latent states [23] specially for these components. Such a numerical routine is designed from the Accept/Reject algorithm to samples from quartic exponential distributions, as described in section 3.2.1.4.

In summary, the estimations obtained for the Logistic parameter and the latent states resemble the Logistic dynamics. Even when the technique is applied to Logistic surrogate data (see Table 3.3 in section 3.3). Reasons for this correctness of the parameter estimations for the latent states had been argued in terms the length of the data and sensitive dependency on initial conditions [5, 70, 53].

From the results presented in section 3.3, the inability to distinguish random from deterministic behaviour is a natural feature of the Bayesian methodology, since no matter how wide the posterior results in each case data type used to feed MCMC, it always generates samples that are consistent with the true trajectory of the Logistic map. This coincidence in dynamical behaviour obtained from estimates calculated from experimental observation of different nature is strongly related by the presence of the deterministic equations in the probabilistic model.

A deeper research to clarify the reasons of this ambiguity of identification is planned as a future research. The research is going to be focused to formu-

late tests which can identify dynamical information of any sort by comparing the resulting posterior distributions.

As a first approach, it is going to be use ideas related to the Ockham's razor approach [72, 65]. Instead of using it for choosing a model, it will cut the wider posterior given the same model class. Experiments and test performed in the PMS are promising at highlighting this difference since uncertainty sources are minimised.

Unfortunately, when faced with real scenarios (see chapter 6), it is almost impossible to decide which is the data set containing random or deterministic behaviour since the existence of a perfect model to represent the system is questionable (not to say such model does not exist) [87], and [86, 18] in [18].

It can also be argued that when a methodology is imported from one paradigm to another, misinterpretations are easily obtained, and the process of doing that, this should be made with care. Explicitly, Bayesian methodology is imported from a statistical paradigm to the deterministic paradigm in the framework of nonlinear time series analysis. For example, the term noise is understood differently and plays a very different role in the nonlinear and statistical analysis of time series.

Once the methodology is imported to a different paradigm interpretation of results are often mislead. For example, it is strongly and incorrectly pointed out by Meyer and Christensen [70] that in order... *"To develop this*

idea within a proper statistical paradigm requires treating the system states as stochastic instead of deterministic. It is therefore consider the more realistic case that the system dynamics are subject to random disturbances.". Severe implicit assumptions on the nature of deterministic chaotic latent states are made when dynamical noise is introduced in the probability model.

One of the points of the study presented in this chapter is not to defend or attack the Bayesian paradigm, much more, it is to recognise the advantages of assuming the states to be stochastic instead of deterministic. Randomness introduces useful redundancy in the restricted observational information available and much more when the problem is not formulated in the PMS.

At the same time, probability perspectives included carefully in deterministic realms could help to account for sources of uncertainty in observations and model-reality pairing.

This discussion does not disqualify the Bayesian approach for parameter estimation. On the contrary, it brings new insights on its effectivness when combined with conventional linear and non-linear methods of parameter estimation and noise reduction as explored in chapter 6.

This work brings new insights on the effectiveness of the Bayesian approach when combined with conventional linear and nonlinear methods of parameter estimation and noise reduction.

Many thanks to Thomas Andrew of Imperial College for his useful dis-

cussions, insights and comments related to the WinBUGS package and the MCMC technique and to Luis Fernandez of Universidad de los Andes at Bogotá, in the discussion and mathematical formalities in the development of the sampling routine to generate samples of Quartic Exponential densities.

Chapter 4

Distilling Information in the Parameter Space

Parameter estimation is typically formulated from a statistical point of view as the tunning of model parameters to mimic observations. In this framework, the problem of parameter estimation is seen as the problem of model fitting. Traditionally, estimates are obtained through calculation of the "best" model parameter values that fits the data and uncertainty measures are given by confidence intervals. In general, such estimations are related only to conditional information contained in the experimental observations given a particular model structure. Therefore estimates are often biased when observations are perturbed with noise. In practice, noise is always present.

This chapter focuses on the idea that, in a nonlinear framework, sta-

tistical methods should be introduced and combined with already existing methodologies in order to enhance the information content of the dynamical model trajectories and their distributions. The dynamical information content is enhanced by generating distribution of model states and parameter estimates instead of single "best" guesses.

"Better" parameter estimation for nonlinear models is pursued by balancing the contribution of information from the dynamical equations and the observations available following the ideas presented by McSharry and Smith in [68].

The information from the model and the observations is combined by the trajectories the model admits in the parameter space. When the balance is obtained, estimates cope simultaneously with both observational and model uncertainty for a given parameter value. In addition, the combination of information from both model and obsevartions in the parameter space generates model states and noise model estimates. If a parameter space region features biased parameter values it is because estimates for the system states and the noise model are also biased. On the contrary, "better" estimates might reflect a fine balance of dynamical and experimental information as discussed in section 4.1. Given the "quality" of a particular region, forecasts and/or condition monitoring of the system are deeply affected.

Even in the PMS when system and reality are matched with each other,

i.e. the model's functional form is known to match the system exactly, a value for the model parameters, must be estimated from corrupted information contained in the observations of the state variables. Observational uncertainty has to be accounted for one way or another.

The simplest case of parameter estimation is to attempt to solve the problem in the PMS where the parameter values that generated the data are known exactly. To solve the problem means finding the "true" parameter values and system trajectories to arbitrary precision. Unfortunately, even under PMS conditions and observations of sufficient length it is not possible to solve the problem with full certainty [68, 88]. It has been found that the inverse problem can only be solved when the dynamical system is known entirely, there are observations available for infinite duration into the past and the noise process is known exactly, as shown in [54] and references therein.

The challenge is then to find a consistent way of extracting useful information when the noise model and the true functional form of the system are known exactly given that there is no available information on the parameter values that generated the observations. This work contributes to its solution by introducing a simple method for extracting significant information in the parameter space via the trajectories admited. The method extracts information of short to moderate time scale dynamics between a model and observations of the system. The information extracted is then combined with

information of the long-term dynamics [68, 88]. It is interesting to note that fundamental questions of parameter estimation in nonlinear systems remain open [16, 36, 92, 12, 76], even in the PMS.

For this study, the problem of parameter estimation in the context of nonlinear dynamical models is stated in the sequence space. The sequence space is a higher dimensional space than the state-space. In contrast to the state-space where a point corresponds to a system's state, a point in the sequence space is a sequence of states or a trajectory segment. A sequence of states or pseudo-orbits are generated by gradient descent [54, 55], which by definition, resemble the observations and the dynamics for a given parameter value. The quality of any parameter value is then viewed through the evaluation of particular properties in the parameter space or statistics that may contain dynamical information. The distribution of shadowing pseudo-orbits, the implied noise level distribution and the pseudo-orbit mismatch distribution (see section 4.2.1 and sections therein) are calculated to explore the quality of the pseudo-orbits in the parameter space for a given value of $\boldsymbol{\theta}$.

This study is presented in two parts, in section 4.1 the statistical approach is described by a discussion of Maximum Likelihood techniques and how information of dynamical trajectories could be introduced. Following McSharry and Smith (*Phys. Rev. Lett.* **83**, 1999) in [68], the statistical

methodology for parameter estimation in nonlinear models focusing on the geometric properties of trajectories in the short term while capturing the global behaviour of the model is described in section 4.1.1.

In section 4.2, the geometric approach is presented. In that framework, for a given parameter value of $\boldsymbol{\theta}$, both pseudo-orbits and true trajectories are identified via gradient descent in the extended state-space or sequence space. Both pseudo-orbits and the shadowing trajectories are contrasted with observations in order to obtain insight on both dynamical information and noise model parameters. A comparison with the observations could be seen as a prosection of the sequence state into the parameter space which in turn induces a structure in the parameter space. Such structure highlights areas in the parameter space where the estimates are considered as candidates for "good" estimations (see section 4.2.1 for definition of "good") and mapping between sequence space and parameter space is made through relevant distributions and its summary statistics.

Balance in the dynamical and experimental information of the estimation process is made though the study of relevant statistics that may contain dynamical information and noise process information as presented in 4.2.1.2 and 4.2.1.1, for the time steps a trajectory that started in the state estimate shadows the dynamics and also is consistent with implied noise levels present in the signal. Advances in parameter estimation presented in this chapter

are possible by improving the balance in extracting information from the dynamic equations and the observations.

All experiments fall within the PMS. Candidate parameter values are evaluated upon how well corresponding trajectories and pseudo-orbits mimic the observations. Trajectories are found by gradient descent [77, 26, 79, 54, 55]. For each parameter value the trajectories obtained are quantified by

i. the ability of model trajectories to ι-shadow [34, 87] uncertain observations,

ii. how well model pseudo-orbits approximate relevant trajectories,

iii. the consistency of an implied observational noise distribution with the noise model (when one is known [68, 76]),

and are described in section 4.2.

Even with perfect knowledge of both the dynamical and the noise model, it is not possible to disentangle uncertainty in the dynamics from uncertainty in a given set of observations. Perspectives including a relevant Bayesian formulation of the problem in the PMS are commented as here and in chapter 3 but still there remain fundamental challenges of parameter estimation when a perfect model class is unavailable. In particular, both methodologies, Bayesian and nonlinear time series analysis, generate points in the sequence

space during the process of parameter estimation with similar noise reduction characteristics. Both methodologies produce similar estimates of model parameters and system states, one should be able to discern which method is more relevant given the scope of any particular study. In other words, it is important to find which method is more efficient on providing reliable solutions to the problem of parameter estimation given the context of each study. To chose one method or the other, there are two possibilities, one in terms of *parsimony* or in terms of the dynamical informational content of the estimates whithin in nonlinear time series analysis framework. A more extensive discussion of improper application of Bayesian techniques is presented in chapter 5, where results from the presentation of how proper Bayesian methodologies can be seen in the PMS in chapter 3 are contrasted with ones presented here.

Figure 4 shows a caricature of how the problem of parameter estimation is viewed in this chapter. Once de-noised segments of trajectories are generated given a particular noise model and dynamical equations of the system, they are represented in the sequence space by points, each point with as many components as the length of the trajectory.

For a given value of $\boldsymbol{\theta}$, in the parameter space a distribution of points in the sequence space is defined. For each distribution of points a relevant statistic g can be defined such that it maps distributions of points in the

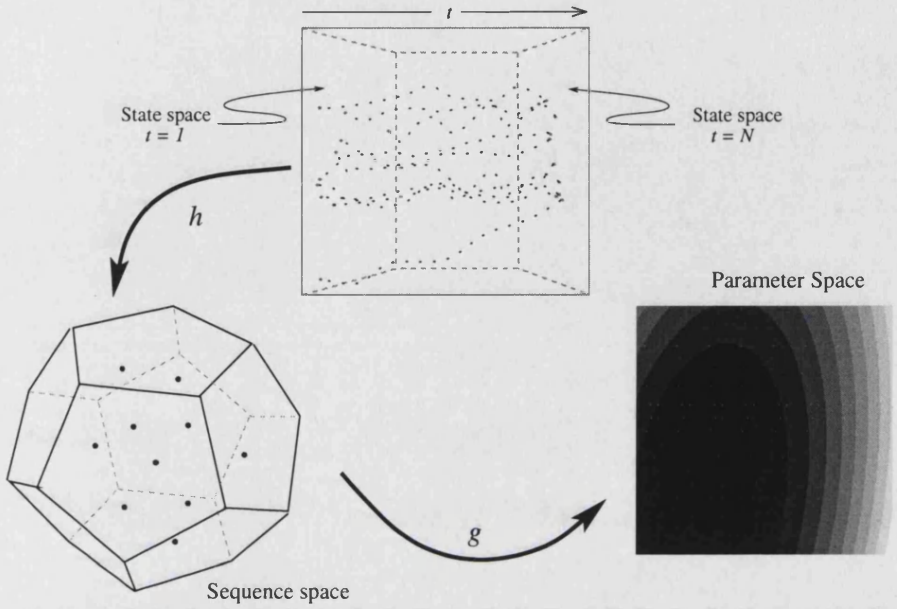


Figure 4.1: The diagram represents how the information form the dynamics and the observations is carried for particular values of the parameter vector $\boldsymbol{\theta}$ to the sequence state and finally to the parameter space via some mapping of relevant statistics represented by g.

sequence state to a point in the parameter space. From the value that g takes, structures in the parameter space are induced.

The procedure for parameter estimation can be easily generalised as an iterative process. The first iteration of that process are the results presented in this chapter where most of the effort is put on the quantification of the trajectories obtained given a parameter value and from there on distinguishing "good" areas in the parameter space. Later iterations will shrink the chosen parameter space areas to narrow the set of candidates for parameter

estimates.

This study is made in cooperation with Hailiang Du, Leonard A. Smith and Kevin Judd and some of the principal results are presented in [88] and summarised in the last section of this chapter.

4.1 Statistical Approach

In this section the statistical approach of point estimation to estimate model parameters using cost functions is discussed from Maximum Likelihood perspectives. In this framework, a probability density function is constructed for the observations given a known model of the system of interest and specified noise model. Parameter estimates are obtained by considering only the information content of the observations. Unfortunately, dynamical information contained in the observations is corrupted by the presence of noise in the form of measurement and/or quantisation errors.

As presented in section 2.2, a PDF is constructed such that the probability of a set of observations given a dynamical model is maximised for a set of parameter values. Given a set of parameters the goal is to find the probability that the set of observations can occur as a function of the parameter $\boldsymbol{\theta}$. Such a probability function, $P(\{s_t\}|\boldsymbol{\theta})$ is the conditional probability of the set of parameters $\boldsymbol{\theta}$ given a particular data set $\{s_t\}$. In turn, this probability is

identified with the Likelihood $L(\boldsymbol{\theta}|\{s_t\})$ of a particular data set given a set of parameters. This identification translates to the problem of parameter estimation to find those values of $\boldsymbol{\theta}$ that maximise the Likelihood.

In this Chapter, the problem of parameter estimation is formulated in the PMS therefore the model and the system share the same mathematical structure. The temporal evolution of the model states are given by equation (1.2) and is reproduced in equation (4.1) as follows

$$x_{t+1} = f(x_t; \boldsymbol{\theta}), \tag{4.1}$$

x_t, for all t, are the model states and $\boldsymbol{\theta} \in \mathbb{R}^\ell$ is the model parameter vector.

In a noise free setting, observations are a function of the model states $s_t = H(x_t)$, $H(x_t) \in \mathbb{C}^2$, where $H(x_t)$ is the *measurement function.* With no loss of generality, if the measurement function is the identity, *i.e.* $H(x) = 1$, then $s_t = x_t$ and it is easy to see that only a segment of $\ell + 1$ observations is sufficient to estimate the true parameter value $\tilde{\boldsymbol{\theta}}$ [68].

In the presence of noise each of the N observations of the model states is given by

$$s_t = x_t + \eta_t, \quad t = 1, \ldots, N \tag{4.2}$$

where each perturbation η_t is an IID Normal random variable, $\eta_t \sim \mathcal{N}(O, \sigma_\eta^2)$. It is known that the observations are generated by the perfect model for the true parameter vector $\tilde{\boldsymbol{\theta}}$. In order to find a Maximum Likelihood estimate

of the parameter vector in the PMS, the *Least Squares* (LS) and *Total Least Squares* (TLS) are defined as functions of the model parameter vector $\boldsymbol{\theta}$.

The LS cost function can be seen to be related to the Likelihood probability function as follows. Assume the observations s_t are IID normally distributed with mean $f(s_{t-1};\theta)$ and variance σ^2, for $t = 2,\ldots,N$. Therefore, at time $t > 0$, the probability of the data set given the parameters is written as:

$$P(\{s_t\}|\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{N-1}(s_{t+1} - f(s_t;\boldsymbol{\theta}))^2\right\}. \tag{4.3}$$

The PDF in (4.3) is identified with the Likelihood function $L(\boldsymbol{\theta}|\{s_t\})$. Maximising (4.3) with respect to $\boldsymbol{\theta}$ is equivalent to minimising the negative logarithm of (4.3) with respect to the parameter vector $\boldsymbol{\theta}$, the parameters of the model f. The Likelihood in equation (4.3) is written as

$$P(\{s_t\}|\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{N-1} d_t^2(\boldsymbol{\theta})\right\}. \tag{4.4}$$

when $d_t(\boldsymbol{\theta}) = |s_{t+1} - f(s_t;\boldsymbol{\theta})|$ are Euclidean distances, and the cost function to be minimised is the so-called least squares cost function, $C_{LS}(\boldsymbol{\theta})$ given by

$$C_{LS}(\boldsymbol{\theta}) = \sum_{t=1}^{N-1} d_t^2(\boldsymbol{\theta}). \tag{4.5}$$

The LS cost function in equation (4.5) is the one-step prediction error of a observations segment of length N and it is also known as the *Root Mean Square* (RMS) error.

Least squares is represented graphically in Figure 4.2. In the Figure, an observation segment of length N is marked by black dots, the image under the model $f(\cdot)$ of the observation s_t is marked by white filled dots, and the distance between the observation s_t and the iteration of the map $f(s_{t-1};\boldsymbol{\theta})$ is represented by a vertical line. The LS cost function is the average of the one-step prediction error for a segment of observations and the solution of the minimisation of equation (4.5) is the parameter estimate of the true parameter vector. Due to sensitivity of the dynamics on initial conditions and the presence of noise, estimates are biased, for higher noise levels lead to larger errors once the map is iterated.

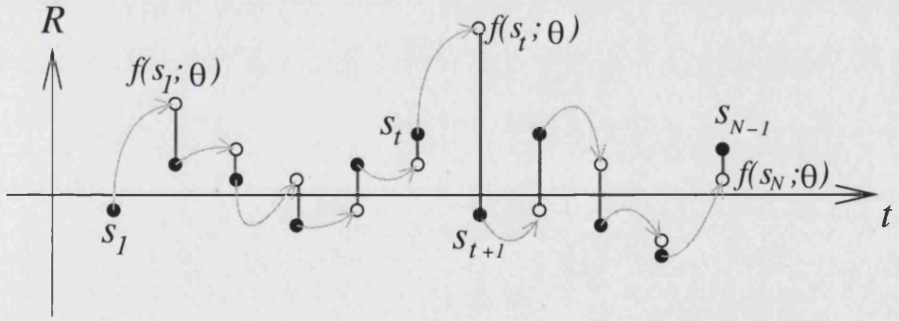


Figure 4.2: The LS cost function is the one-step prediction error. The black dots represent the observations, the white filled circles the iteration of the map $f(\cdot)$ from the observation s_t and the vertical lines are the distances $d_t^2(\boldsymbol{\theta})$ that defined $C_{LS}(\boldsymbol{\theta})$ for a given trajectory segment.

In the free noise PMS , the estimates converge to $\tilde{\boldsymbol{\theta}}$ in the limit of $N \to \infty$. In general, any cost function, $C(\boldsymbol{\theta})$, could be minimised provided that a

relevant distance, $d_t^2(\boldsymbol{\theta})$, is defined in order to account for the uncertainty of the observations and the likely model state assumptions.

Throughout this Chapter, two 2-dimensional maps are considered as toy systems to generate noisy observations, the *Hénon* map [45] and the *Ikeda* map [47] in equations (4.6) and (4.7), respectively. Noisy data is generated by these maps with noise variances, $0.01 \leq \sigma_\eta^2 \leq 0.05$ and $N = 512$ observations.

The Hénon map is a 2-dimensional map, $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, given by

$$f(x_{t+1}, y_{t+1}) = \begin{pmatrix} c - ax_t^2 + y_t \\ \\ bx_t \end{pmatrix}, \tag{4.6}$$

where the map is seen as a generalisation in two dimensions of the Logistic map (see equation (3.23) in Chapter 3, section 3.2.1.2). The parameter vector is $\boldsymbol{\theta} \in \mathbb{R}^3$ with components (a, b, c). The true parameter vector is $\tilde{\boldsymbol{\theta}} = (1, 1.4, 0.3)$.

The Ikeda map is a complex map. The Real representation of the map is 2-dimensional map, $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, given by

$$f(x_{t+1}, y_{t+1}) = \begin{pmatrix} a + \mu(x_t \cos\alpha - y_t \sin\alpha) \\ \\ \mu(x_t \sin\alpha - y_t \cos\alpha) \end{pmatrix}, \tag{4.7}$$

where α is given by

$$\alpha = \kappa - \frac{\eta}{1 + x_t^2 + y_t^2}. \tag{4.8}$$

The parameter vector is $\boldsymbol{\theta} \in \mathbb{R}^4$ with components (a, μ, κ, η). The parameter vector is $\boldsymbol{\theta} = (1, 0.83, 0.4, 0.6)$.

Each panel in Figure 4.3 plots the LS value as a function of one component of the parameter vector, a for Henon map (top) and μ for the Ikeda map (bottom), fixing the other components of $\boldsymbol{\theta}$ to their corresponding true values. The RMS error is plotted for two noise levels, 1% and 5%, in solid and dashed lines respectively. The true parameter value is marked with a vertical line.

In each panel of Figure 4.3, it is clear how the minimum of the LS cost function and resulting parameter estimate, *i.e.* LS minimum, approaches to zero as the noise level increases for both maps and noise levels used.

In both cases, the higher the noise level the flatter the LS curve making more difficult the identification of parameter estimate by finding the minimum in the LS cost function. As the noise level increases, the estimated parameter value is more distant from the true value, as seen in Table 4.1.

Map	Noise (%)	$\lvert\tilde{\boldsymbol{\theta}} - \theta\rvert$
Hénon	1	5.5×10^{-3}
	5	9.3×10^{-2}
Ikeda	1	1.3×10^{-3}
	5	6.6×10^{-1}

Table 4.1: LS estimates for Hénon and Ikeda map from noisy observations with noise levels of 1% and 5% for $\theta = a, \mu$ respectively to each map. The last column displays, $|\tilde{\theta} - \theta|$, the absolute value of the true parameter and the LS estimate.

Figure 4.3 and Table 4.1, shows also that the Ikeda map to be more sensitive than the Henon map, to the noise level when estimating μ using the LS cost function since for the Ikeda map the LS minimum is located in values one order of magnitude less than the true value, $\mu = 0.83$. In contrast, for the Hénon map, the location of the LS minimum for both noise levels is differing only in the second decimal place, for each noise level. In any case, model parameter estimates are biased and show a dependency on the noise level present in the observations.

Results presented in Figure 4.3 confirm the features of the LS cost function commented upon by McSharry and Smith [68]. In addition, in [68] it is shown for the Logistic map that even for an infinite data set the LS a-estimator depends explicitly on the noise amplitude, as seen here for Hénon and Ikeda estimates, numerically.

Another special case of the maximum Likelihood method, is the total Least Squares (TLS) cost function. The TLS cost function is given by

$$C_{TLS}(\boldsymbol{\theta}) = \sum_{t=1}^{N-1} \min_{x \in \mathbb{R}^m} d_t^2(x; \boldsymbol{\theta}), \tag{4.9}$$

where $d_t^2(x; \boldsymbol{\theta})$ is the Euclidean distance between the points (s_t, s_{t+1}) and $(x, f(x; \boldsymbol{\theta}))$

$$\min_{x \in \mathbb{R}^m} d_t^2(x; \boldsymbol{\theta}) = (s_t - x)^2 + (s_{t+1} - f(x; \boldsymbol{\theta}))^2, \tag{4.10}$$

where x is a real value that minimises d_t^2 for a fixed $\boldsymbol{\theta}$, under the assumption

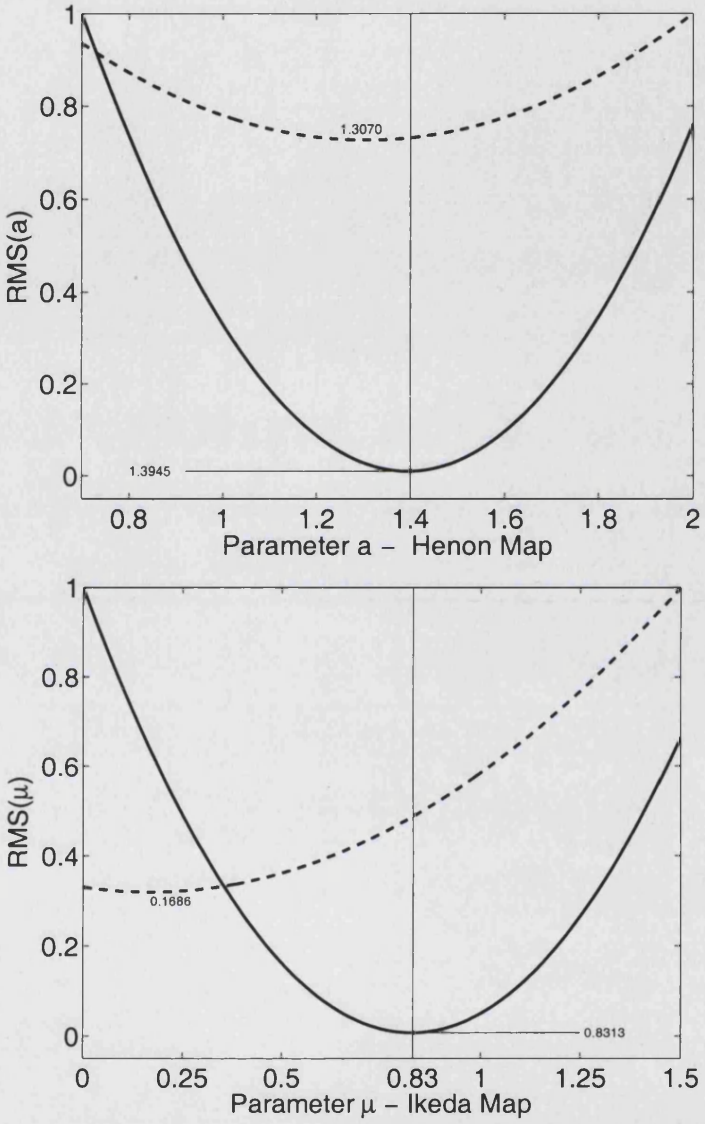


Figure 4.3: LS cost function as a function of one parameter vector component, a for Hénon map (top) and μ for the Ikeda (bottom). LS cost function is plotted for data sets of length $N = 512$ and noise levels of 1% (solid line) and 5% (dashed line). The true parameter value is marked with a vertical line. The LS estimator is written in place.

that s_t and s_{t+1} are independent. Figure 4.4 represents schematically this distance in (4.10).

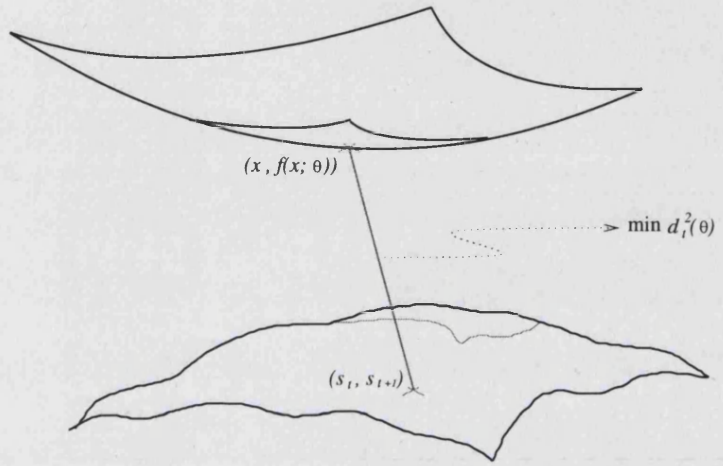


Figure 4.4: The TLS cost function is the minimum average distance between the surface defined by the observation pairs, (s_t, s_{t+1}), and a point in the model hypersurface, $(x, f(;\boldsymbol{\theta}))$. For each pair and fixed value of $\boldsymbol{\theta}$, a value of x is found such that $d_t^2(x;\boldsymbol{\theta})$ is minimum.

The minimum of $C_{TLS}(\boldsymbol{\theta})$, $\boldsymbol{\theta}_{\min}$, is identified with the maximum probability that the sequence of $N-1$ points, $\{(s_t, s_{t+1})\}_{t=1}^{N-1}$ corresponds to the set of model states $(x, f(x;\boldsymbol{\theta}_{\min}))$.

The TLS cost function ignores any dependency of the Likelihood of the system states x_t since the only dependency is through the map $f(x;\boldsymbol{\theta})$. The corresponding Likelihood, when the TLS cost function is used, is

$$P(\{s_t, s_{t+1}\}_{t=1}^{N-1}|\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N-1/2}} \exp\left\{-\frac{1}{2\sigma^2} C_{TLS}(\boldsymbol{\theta})\right\}. \qquad (4.11)$$

The dependency of the conditional probability (4.11) in the so-called latent variables, x, (since the system states are not measured directly) is accounted for in the TLS cost function (4.10) by making the dependency of the probability of observing (s_t, s_{t+1}) explicit on the model parameters $\boldsymbol{\theta}$, the model states x_t and its image under the map $f(x_t; \boldsymbol{\theta})$.

For several realisations of the noise process, the distributions of the estimates are consistently biased for both the LS and TLS cost functions (see Figure 2. in [68]).

As discussed and presented in [68], in summary, the estimates are biased for the following reasons:

- In the case of LS estimates, not even complete knowledge of the dynamical equation of the system and the noise model are enough to obtained unbiased estimators. Non-linearities in $f(\cdot)$ combined with observational noise, make the estimator dependent on the noise amplitude explicitly even for infinite data sets.

- The TLS cost function includes information regarding the latent variables or true system states, $\tilde{x}_t$, but fails to provide unbiased estimates of $\boldsymbol{\theta}$ since no restrictions are imposed on the distributions of x in (4.9) nor is the knowledge of the probability function of the system states (*i.e.* invariant measure) included in the Likelihood.

- Both cost functions, LS and TLS ignore any explicit dependency of the Likelihood on the system states.

In [68], McSharry and Smith define a new cost function that improves LS and TLS cost functions by including dynamical information of the system contained in the model. This cost function includes the information about the system's states $\tilde{x}$ by introducing the invariant measure, $\mu(x, \boldsymbol{\theta})$, induced by the map $f(\cdot)$, explicitly in the Likelihood. The next section, section 4.1.1 discusses the so-called *Maximum Likelihood* (ML) cost function.

Before introducing the ML cost function, it is interesting to point out an additional feature of the cost function approach. In addition to the parameter estimates obtained using the TLS cost function, model state estimates are obtained as well in the process. The estimates for model states are the result of the minimisation condition of equations (4.9) and (4.10) that defined the TLS cost function.

This feature is also present in other methodologies of parameter estimation, model states estimates are produce as well as model parameter estimates. This is the case of the Bayesian approach described in Chapter 3 and the new geometrical approach of parameter estimation described later in this Chapter in section 4.2.

Figure 4.5 shows a reconstruction of the model states estimates, x_t (squares),

that minimised $C_{TLS}(\boldsymbol{\theta})$ for $1.4 \leq \boldsymbol{\theta} \leq 2.2$, from several realisations of Logistic noisy observations for Gaussian noise with $\sigma_\eta^2 = 0.2$ and true parameter value $\tilde{a} = 1.85$.

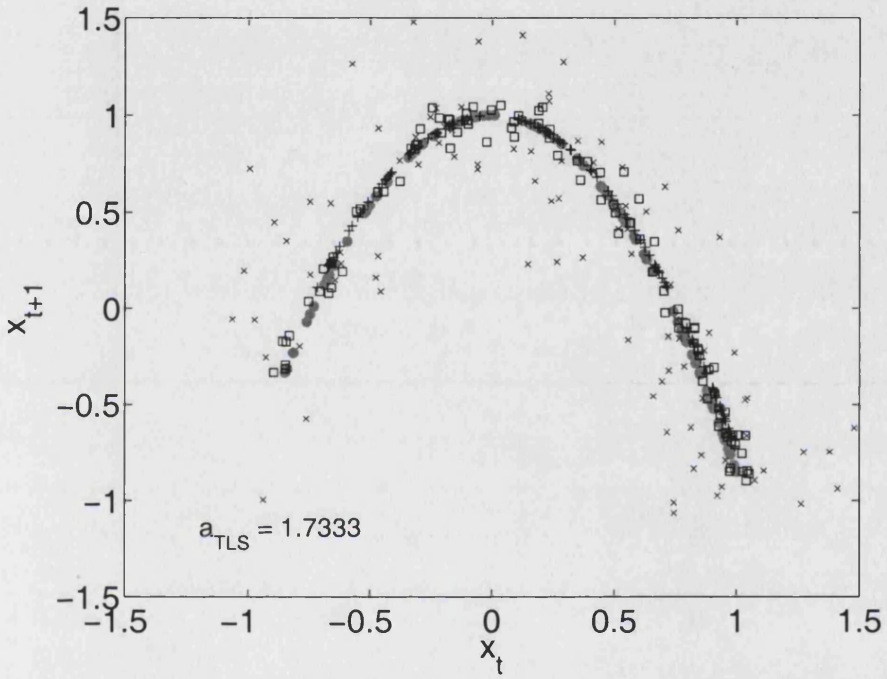


Figure 4.5: Reconstruction of TLS Logistic states estimates for $1.4 \leq \boldsymbol{\theta} \leq 2.2$ from several realisations of observations with Gaussian noise with noise of variance of 0.2 and $N = 100$ points. Grey dots: true Logistic states, grey crosses: noisy observations, black pluses: Logistic trajectory using TLS estimate for a, and squares: TLS Logistic states estimates resulting from the minimisation of equation (4.10). The TLS estimate is $a_{TLS} = 1.7333$.

Figure 4.5 shows several Logistic series with $N = 99$ points reconstructed in the same axes. The true Logistic trajectory for $\tilde{a} = 1.85$ is plotted with

grey dots. Grey crosses are used to reconstruct the noisy observations with noise variance $\sigma_\eta^2 = 0.2$. For several realisations of the noise process, TLS estimates for a are calculated thus the mean value of the a-estimates for all noise realisations is taken as the TLS estimation for the Logistic parameter. The mean value is equal to $a_{TLS} = 1.7333$ and is included as text in the plot. Using this value for $a = a_{TLS}$ a trajectory is generated from $x_0 = 0.3$ and is plotted using black pluses. Finally, squares are used to plot the TLS model state estimates, *i.e.* the solution for the minimisation of equation (4.10), which makes interesting the Figure.

As it is clear in Figure 4.5, the Logistic structure appears in the plot when the TLS model state estimates are reconstructed (squares), it appears as in the TLS estimates some dynamical information is present. In fact, the minimisation process in (4.10) is performing a type of noise reduction, which is a false statement. The presence of the dynamics in the resulting estimates does not imply in any sense that the cost function is considering any temporal dynamical correlation of the system states, as commented in [68] and seen straightforwardly from the definition of the TLS cost function. As in section 3.3, the resulting estimates can be misleadingly taken as de-noised states instead as indicators of the presence of deterministic equations in the cost function.

Further more, several questions could be formulated to discover possible

dynamical features of such model state estimates:

- Is this series of estimates, a pseudo-orbit that shadows the observations or a random sample that only looks like the attractor?
- Can the pseudo-orbit obtained by TLS estimates be considered as a shadowing pseudo-orbit of the model?
- In the case the state estimates are taken as a cleaned trajectory (*i.e.* with less implied noise than the originally present in the signal) of the model, is it useful for infering the future evolution of the system?
- In the case the state estimates are taken as a random sample of the model, do they sample of invariant measure of the model?
- Is it possible to extract any dynamical information of these system state estimates in either of the two cases above?

These and other questions related to the relevance of the content of dynamical information in a time series are going to be left for future research. A discussion of these issues is presented in Chapter 5.

Given that it is known that the LS and TLS cost functions do not include explicitly the temporal correlation of the model states, the next section presents a way to include information about the invariant measure of the model in the Likelihood for the parameter vector $\boldsymbol{\theta}$ following [68].

4.1.1 Inclusion of Global Behaviour

Solving the problem in the PMS provides the advantage of the knowledge of two facts: the model states, x_t are correlated in time, and the invariant measure induce a measure on the state space. Ignoring these facts results in biased estimates of parameter values, as in the case of TLS estimates. McSharry and Smith [68] include this knowledge to obtain a consistent formulation of the Maximum Likelihood approach for nonlinear models. Explicitly, the invariant measure weights the values of the model states in the Likelihood PDF.

Since the model is known exactly, the invariant measure $\mu(x, \boldsymbol{\theta})$ of the latent variables x_t of the model is always obtainable. This information is incorporated in (4.9) by integrating the dependency on the latent state x_t out of d_t^2 in (4.10). Formally, in terms of the conditional probability of $\boldsymbol{\theta}$ given the pairs (s_t, s_{t+1}) yields for all $t = 1, \ldots, N-1$ the following expression

$$L(\boldsymbol{\theta}|\{s_t, s_{t+1}\}_{t=1}^{N-1}) = \prod_{t=1}^{N-1} \int_x P(\{s_t, s_t + 1\}_{t=1}^{N-1}|\boldsymbol{\theta}) \mathrm{d}\mu(x, \boldsymbol{\theta}), \tag{4.12}$$

for a fixed $\boldsymbol{\theta}$ from which the maximum Likelihood cost function is given by

$$C_{ML}(\boldsymbol{\theta}) = -\sum_{t=1}^{N-1} \log \int_x \exp\left\{-\frac{d_t^2}{2\sigma^2}\right\} \mathrm{d}\mu(x, \boldsymbol{\theta}), \tag{4.13}$$

where d_t^2 is

$$d_t^2(\boldsymbol{\theta}) = (s_t - x)^2 + (s_{t+1} - f(x; \boldsymbol{\theta}))^2. \tag{4.14}$$

In Practice, the integral in (4.13) is replaced by a sum over a model trajectory of length $\tau \gg N$, where τ is as long as the computational constrains allows. Explicitly, the cost function in equation (4.13) is approximated by the following expression:

$$C_{ML}(\boldsymbol{\theta}) \approx -\sum_{t=1}^{N-1} \log \sum_{k=1}^{\tau} \exp\left\{-\frac{1}{2\sigma_\eta^2}\left[(s_t - x_k)^2 + (s_{t+1} - f(x_k;\boldsymbol{\theta})^2\right]\right\}, \tag{4.15}$$

an average over a long post-transient trajectory of the map $f(\cdot)$ for the distances between the point (s_t, s_{t+1}) and $(x_k, f(x_{k+1};\boldsymbol{\theta}))$ for $k = 1, \ldots, \tau$ where $\tau >> N$.

Despite $C_{ML}(\cdot)$'s name being misleading, by requiring consistency between the data and the PDF, *i.e.* invariant measure, of the latent variables, the ML cost function outperforms estimates from the minimisation of the LS and TLS cost functions. The ML cost function consistently yields better estimates for all noise levels considered and for the Logistic [67], *Moran-Ricker* [73] and Hénon [45] in comparison with LS and TLS cost functions (see Figures 2 and 3 in [68]).

Noted that the PMS conditions also include the fact that if the model, $f(x;\boldsymbol{\theta})$, which describes the system's state evolution is an application $f : \mathbb{R}^m \to \mathbb{R}^m$ then all components of the state vector are measured making the observation $s_t \in \mathbb{R}^m$, otherwise the problem is formulated in the IMS.

Observations and latent states must live in the same space $\mathbb{R}^m$ in order to consistently define the distance in (4.14).

For observations where not all components of $x_t \in \mathbb{R}^m$ are measured, *i.e.* $s_t \in \mathbb{R}^n$ for $n < m$, a representation of the map in delay coordinates is required or a corresponding *Nonlinear Auto-Regressive* (NAR) process representation of $f(\cdot)$. For $f(\cdot)$ following a NAR process, $NAR(\tau) = f(x_{t-\tau}, x_{t-\tau+1}, \ldots, x_{t-1}; \boldsymbol{\theta})$, where τ is a time lag.

When a delayed version of the perfect model is available additional terms in (4.13) should be included where the definition of the distance (4.15) requires it. If a delayed version of the perfect model is not available, $C_{ML}(\boldsymbol{\theta})$ is not suitable to be used unless further delay reconstruction of the observations is constructed.

This subtle point is exemplified for the case of the Hénon map. Consider the Hénon map's [45] formulation in delay coordinates

$$f(x_t, x_{t-1}; \boldsymbol{\theta}) = 1 - ax_t^2 + bx_{t-1}. \tag{4.16}$$

The delay version of Hénon in (4.16) makes the map $f : \mathbb{R}^2 \to \mathbb{R}$. Only observations s_t of x_t are available for $\tilde{\boldsymbol{\theta}} = (\tilde{a}, \tilde{b})$. In the case that the formulation in equation (4.16) is considered as the perfect model instead of equation (4.6), the distance in the ML cost function (4.13) has to be defined as

$$d_t^2(\boldsymbol{\theta}) = (s_t - x_t)^2 + \left(s_{t+1} - f\left((x_t, x_{t-1})\,; \boldsymbol{\theta}\right)\right)^2 + (s_{t-1} - x_{t-1})^2, \tag{4.17}$$

for consistency. The distance calculated in (4.17) is not the same distance as in (4.15) but the distance in a delay reconstructed sequence space, in other words, it is the distance between the point $(\{s_{t-1}, s_t\}, s_{t+1})$ of delay reconstructed observations and the corresponding point in the hyper-surface defined by the delay coordinate version of the map given by $(\{x_{t-1}, x_t\}, x_{t+1})$ where $x_{t+1} = f(\{x_{t-1}, x_t\}; \boldsymbol{\theta})$.

The correct calculation of of the $C_{ML}(\boldsymbol{\theta})$ cost function, for the PMS conditions defined in Chapter 1 using the 2-dimensional version of the Hénon map [45] in equation (4.6), is given by equation (4.14) and using approximation (4.15) for observations and system states living in $\mathbb{R}^2$ is written as

$$d_t^2(\boldsymbol{\theta}) = \overbrace{\left(s_t^x - x_k\right)^2 + \left(s_{t+1}^x - f_x(x_k, y_k; \boldsymbol{\theta})\right)^2}^{x-\text{component}} + \underbrace{\left(s_t^y - y_k\right)^2 + \left(s_{t+1}^y - f_y(x_k, y_k; \boldsymbol{\theta})\right)^2}_{y-\text{component}}. \tag{4.18}$$

Equation (4.17) compared with equation (4.18) contains an additional term. Note that the third term in (4.17) requires information on the past two states of the system in order to obtain estimates.

Although the additional delay term is in (4.17), there are no major qualitative differences of the values the ML cost function takes in the parameter space $\mathcal{Q}_H$; interpretations of those estimates have to be made carefully. Equation (4.17) can be seen as a formulation of the problem in the IMS where all system variables are not observed and the model is an approximation of the

system, *i.e.* system and model do not belong to the same model class. Figure 4.6 shows the results for the calculation of (4.13) using (4.18) in the left panel, and using (4.17) in the right panel, for a two dimensional parameter space and a uniform grid of 201 nodes in each coordinate. The darker the colour in the plot, the lower the value of the ML cost function. The white plus marks the true parameter value $\tilde{\boldsymbol{\theta}}$.

In both plots, the information of the invariant measure of the map as a function of the parameter vector $\boldsymbol{\theta}$ is reflected as complex structures or "tongues" appearing for values of $\boldsymbol{\theta}$. When the parameter value is located in a non chaotic regime or the trajectories escape of the attractor, the value of the cost function is systematically higher than in other areas of the parameter space.

As pointed out earlier, both panels of Figure 4.6 look qualitatively similar but they are fundamentally different. In the case that equation (4.18) is used in the right panel, the PMS scenario and each model state has a corresponding observation, *i.e.* model states and observations at time t are dimension 2, producing a smaller area of minimum values containing the true parameter value than the one in the left panel. Uncertainty in the parameter estimates can be associated with the area in the level curves of the plots. Thus when equation (4.17) is used, uncertainty is higher since the problem is solved in the IMS.

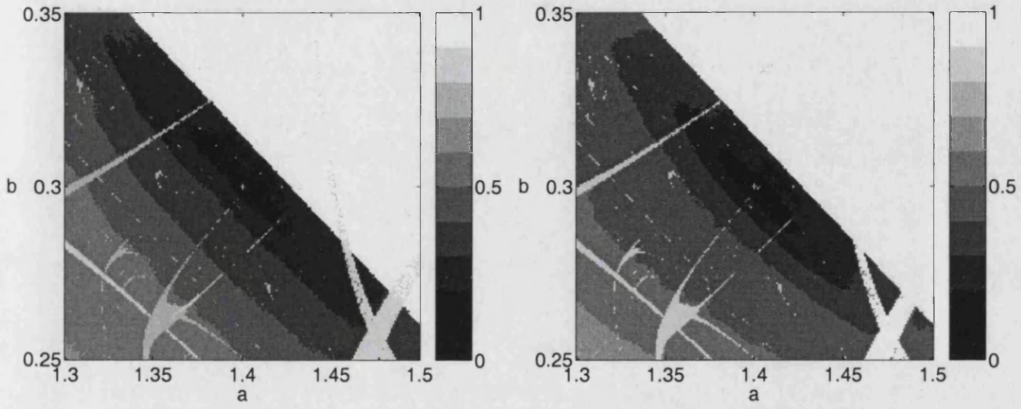


Figure 4.6: Value of the C_{ML} cost function in Hénon map's 2-dimensional parameter space, using equation (4.17) in the left and equation (4.18) in the right. Values of the cost function are calculated using $N = 512$ observations and a noise level of 0.05. The white plus marks the true parameter value.

Although, relevant dynamical information is being included successfully by the ML cost function for parameter estimation, there still remains some weak assumptions in the derivation and use of the ML cost function, in particular the fact that the pairs (s_t, s_{t+1}) are taken as not correlated in time.

Work in this disection, not taking into account Bayesian perspectives but where statistical methods are melded with dynamical views are rare but appearing more frequently in the literature. For example [76] and references therein.

The maximum Likelihood approach fails to produce accurate single point estimations for nonlinear model parameters in the PMS even though dy-

namical information is included as for the case of the ML cost function but still there is uncertainty entangle with the dynamics of interest. Disentangling uncertainty from observations and model error is considered in the next section of this Chapter.

4.2 Geometric Approach

In the PMS, dynamical information is entangled with the observational uncertainty. This section presents a new method to distil dynamical information for parameter estimation via characterisation of the parameter space by the trajectories the model admits.

Although the problem of parameter estimation is formulated and solved in the PMS, this new method is constructed in a way that future extensions outside the PMS are easily followed, even though are not detailed in this work.

Parameter estimation in the PMS by this new geometrical methodology is obtained by the following steps:

1. Measure a set of observations of the system for $t = 1, \ldots, N$.

2. Generate a grid in the parameter space of the model.

3. Generate indistinguishable states for each observation available and

parameter values $\boldsymbol{\theta}$ in the parameter space grid.

4. Use the indistinguishable states found for each time t and parameter value $\boldsymbol{\theta}$, as initial conditions of model trajectories.

5. Characterise the parameter space by defining dynamical quality measures of the resulting model trajectories.

6. Choose parameter space areas with "desired" qualities as parameter estimates.

7. Start again in step 2 to increase resolution of interesting/special parameter space areas.

Steps 1 to 3 are explained below whilst the rest of the steps are explain in detail in section 4.2.1.

This new method is posed as an iterative process in which dynamical information is included in the estimation process and observational uncertainty is simultaneously accounted for. The methodology and results presented here correspond to the first iteration of that process which find parameter estimates along with state estimates through a search in the parameter space and can be found in the work presented by L.A. Smith, M.C. Cuéllar, H. Du and Kevin Judd in [88]. Results for further iterations of the method are planned for future work.

The *parameter space* $\mathcal{Q}$ is understood as the space where the parameter vector $\boldsymbol{\theta}$ lives, *i.e.* $\mathcal{Q} \subset \mathbb{R}^{\ell}$. In the PMS, each point $\boldsymbol{\theta} \in \mathcal{Q}$ admits certain trajectories of the map $f(\cdot;\boldsymbol{\theta})$ given uncertainty in the observations $s_1, \ldots, s_t, \ldots, s_N$. For a given $\boldsymbol{\theta} \in \mathcal{Q}$ and initial condition, the map $f(\cdot;\boldsymbol{\theta})$ generates a trajectory $\{x_t \in \mathbb{R}^m\}_{t=1}^N$ by t-folding of the map.

The method focuses in trajectories generated by special candidates for initial conditions. An "optimal" candidate to generate a model trajectory is chosen in a way that it balances the trade-off of dynamical information between uncertainty sources present in the problem.

Following notation of Chapter 2, given a noise model and a segment of N observations, indistinguishable states theory [54] provides that the point y_t in the state space is indistinguishable from the true system state x_t, if $y_t \in H(x_t)$, *i.e.* belongs to the set of of all possible indistinguishable states (see equation (2.45)). Indistinguishable for an observation s_t at time t, it is either a measurement of x_t or y_t.

For $N \to \infty$, states y_N indistinguishable from the model state x_N are found with probability one. For $N < \infty$, a sequence of indistinguishable states, $\{y_t \in \mathbb{R}^m\}_{t=1}^N$, can be found by variational methods [54]. In particular, using a *gradient descent* (GD) method [79].

In this section, a GD algorithm is used to generate such a sequence of indistinguishable states in the PMS. The GD algorithm produces states from

the set of indistinguishable states, $H(x_t)$ which are at the same time Maximum Likelihood states, consistent with equation (4.4) [54].

The gradient descent method is an iterative algorithm where one iteration produces a sequence of states as long as a set of observations is available. Such a sequence of states is a point in the sequence space. Let $\{u_t \in \mathbb{R}^m\}_{t=1}^N$ denote the sequence of indistinguishable states generated by GD or GD states. Thus from now on, u_t denotes an GD indistinguishable state of the true model state x_t.

The $\{u_t\}_{t=1}^N$'s in the sequence space are obtained by GD such that the *one-step prediction error* is minimum for a given segment of N observations, and it shadows the true trajectory of the map $f(\cdot;\tilde{\boldsymbol{\theta}})$.

The one step prediction error for points in the shadowing trajectory is given by

$$d_t^2 = |u_{t+1} - f(u_t;\boldsymbol{\theta})|, \tag{4.19}$$

where $|\cdot|$ is the Euclidean distance in the state-space. The one-step prediction error, d_t^2, is known also as the *mismatch error.*

If $d_t^2 = 0$ then the sequence of points $\{u_t\}_{t=1}^N$ in the state space is a trajectory of the map $f(\cdot)$. Otherwise, $\{u_t\}_{t=1}^N$ is no longer a trajectory and it is said to be a *pseudo-orbit.* A pseudo-orbit generated by GD is denoted by $\{z_t\}_{t=1}^N$. An example of pseudo-orbit is a sequence of noisy observations,

$s_1, \ldots, s_t, \ldots, s_N$.

The GD algorithm (see [54, 55, 26, 79, 51] for details) is a minimisation algorithm and it is used in this context to minimise the the *Mismatch cost function*, $C_{MM}(\boldsymbol{\theta})$, given by

$$C_{\mathrm{MM}}(\boldsymbol{\theta}) = \frac{1}{N-1} \sum_{t=1}^{N-1} |u_{t+1} - f(u_t, \boldsymbol{\theta})|, \tag{4.20}$$

a temporal average of the mismatch error over the time interval of interest.

Although there are no local minima for C_{MM} in the sequence space, the GD algorithm is run for a pre-determined time T and thus a trajectory is not obtained, only a pseudo-orbit.

For a suitable number of iterations, T, the resulting sequence of states, $\{z_t^{(T)}\}_{t=1}^N$, is taken to be the pseudo-orbit that minimises (4.20), *i.e.* the pseudo-orbit with the lowest value of $C_{MM}(\boldsymbol{\theta})$ or $d_t^2 \ll 1$ for all t.

The GD algorithm can be applied on a temporal sliding window varying in sizes from 2 to N. Preliminary studies performed on the sensitivity of the GD and resulting estimates to the sliding window lengths, suggests that for medium range windows, *i.e.* between 2 and N, the combination of future and past information by GD is more efficient for points in the middle of a trajectory segment. Extensive studies on the sensitivity to the sliding window size of the GD algorithm and resulting quality of parameter estimates are planned for future research. Most of the results presented in this section

are from GD runs for a window with length equal to the duration of the observations.

For consistency and comparison with other results in [83, 26, 42, 68, 79, 54, 76, 52, 55, 43], the Hénon [45] and Ikeda [47] maps are considered as systems to generate noisy observations in the PMS. These maps are given by the equations (4.6) and (4.7) for Hénon and Ikeda respectively.

The Hénon map has a corresponding parameter space $\mathcal{Q}_H \subset \mathbb{R}^2$ for the parameter vector $\boldsymbol{\theta} = (a, b)$.

In the case of the Ikeda map, the corresponding Real representation of the map has an associate parameter space, $\mathcal{Q}_I \subset \mathbb{R}^4$, *i.e.* the parameter vector is $\boldsymbol{\theta} = (a, \mu, \kappa, \eta)$. From a study of the sensitivity of the Ikeda attractor to values of $\boldsymbol{\theta}$, the parameter space is reduced to be $\mathcal{Q}_I \subset \mathbb{R}^2$ by setting $a = 1$ and $\eta = 0.6$ leaving $\boldsymbol{\theta} = (\mu, \kappa)$. For further reduction in the dimension of $\mathcal{Q}_I$, $\kappa = 0.4$, when estimations are performed in a parameter space $\mathcal{Q}_I \subset \mathbb{R}$.

Observations of length N are generated for several noise levels for both maps, with the true parameter vector as $\tilde{\boldsymbol{\theta}} = (0.83, 0.4)$ and $\tilde{\boldsymbol{\theta}} = (1.4, 0.3)$ for Ikeda and Hénon respectively.

Pseudo-orbits z_t are generated using gradient descent for a window size equal to the length of the observations available, $N = 512$. In order to set the number of iterations, T, that the GD algorithm is to run, the mismatch cost function in (4.20) is used as an indicator of convergence. Convergence

is considered to be achieved when $C_{MM}(\boldsymbol{\theta}) < 10^{-3}$ and when the difference between the pseudo-orbit state z_t for an iteration T is approximately equal to the same state in the last iteration $T-1$, *i.e.* $|z_t^{(T)} - z_t^{(T-1)} \leq 10^{-12}|$.

Figure 4.7 shows the value of the Ikeda's parameter μ identified by the location of the minimum of $C_{MM}(\boldsymbol{\theta})$ (4.20) as function of the number of times the GD algorithm has been iterated, in a 1-dimensional grid of the parameter space. It shows μ estimates for two noisy Ikeda observations, 1% (solid line) and 5% (dashed line) noise levels.

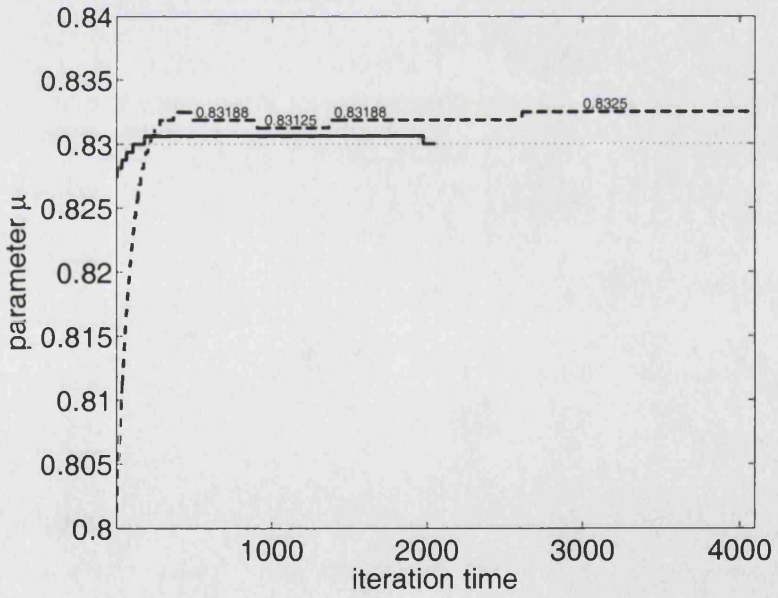


Figure 4.7: Location of the minimum of $C_{MM}(\mu)$ cost function as a function of iteration time for the GD algorithm. Location is shown for two noisy observations of the Ikeda map, 1% noise level in solid line and 5% noise level in dashed lines. The true value of μ is marked with a horizontal dotted line.

It is clear from the figure, how the location of the minimum moves toward the true value of the Ikeda map ($\tilde{\mu} = 0.83$) as the iteration time of the GD algorithm increases. For both noise levels shown, the minimum identifies the true value of μ up to 2 decimal places. From this plot and the fact that there is no minimum for a finite set of observations, satisfactory convergence is considered to be achieved after $T = 2048$ iterations.

Figure 4.8 shows an example of how the ensemble of pseudo-orbits looks like compared with the observations of Ikeda with noise of 1% and for $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ after $T = 2048$ iterations of the map. The figure is composed by 4 panels, right panel show results for 1% noise level and left for 5%. In the top row, a segment of the true Ikeda trajectory is plotted with pluses and the mean of the distribution of pseudo-states is plotted with circles. The bottom row of the figure shows the reconstruction of the mean of the resulting pseudo-orbits plotted with grey dots and overlapped with the reconstruction of the true trajectory of Ikeda with pluses.

Although, for both noise levels the distribution of pseudo-states is very sharp, for a 1% noise level, the GD algorithm tends to a trajectory closer to the true trajectory than for the case of 5% noise level, in trace and in the reconstructed state space.

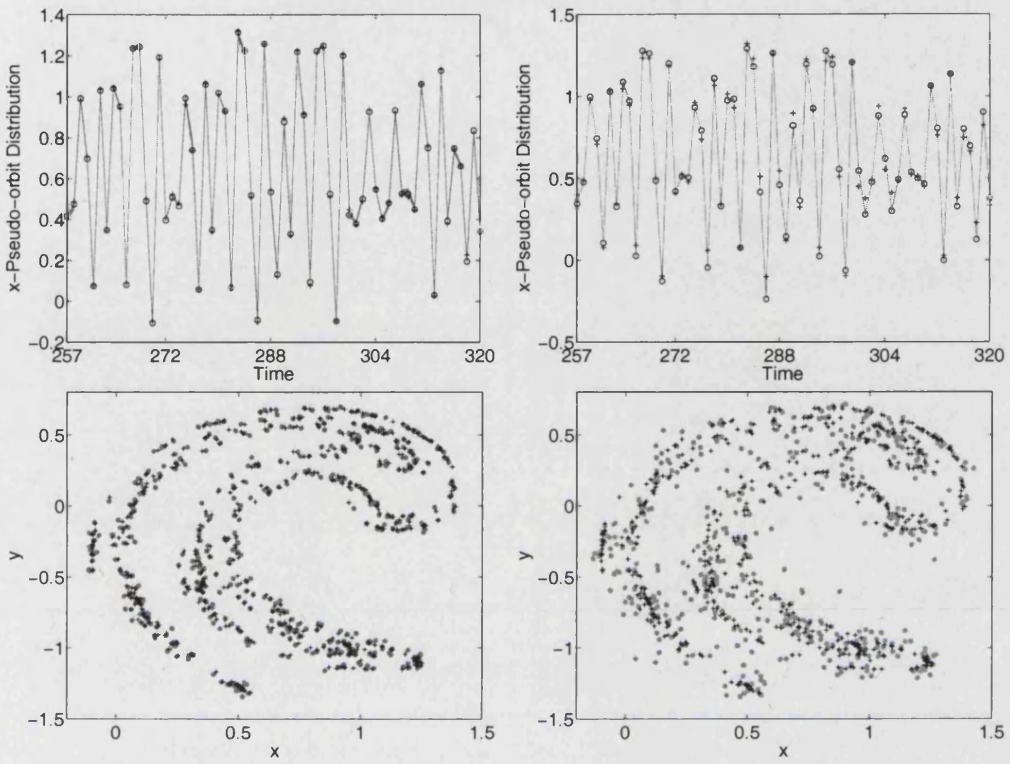


Figure 4.8: The ensemble of the x component of the pseudo-orbits is plotted for the true parameter value $\mu = 0.83$ in the top row for 512 Ikeda observations with noise levels of 1% (left) and 5% (right) after 2048 iterations of the GD algorithm. A true trajectory segment of 64 points is plotted with pluses and the pseudo-states with circles. The 99.5% and 0.5% isopleths are plotted grey. The bottom row plots the reconstruction of the median of the pseudo-orbits distribution (grey dots) and the true states (+).

4.2.1 Search in the Parameter Space

Once satisfactory pseudo-orbits have been generated for each point in the parameter space, a point in the sequence space is obtained. The point is a

pseudo-orbit in the state space that the model admits for a particular parameter value. Then the parameter space is characterised in a way that for each parameter value study, a measure or summary statistic of the corresponding distribution is assigned to it to quantify the quality of the pseudo-trajectories compared with the observations or the model trajectory. Henceforth, each point in the parameter space is paired with an empirical distribution.

Clearly, pseudo-orbits obtained for a given parameter value and a given noise level are in fact shadowing the observations and a trajectory of the model since the GD algorithm is formulated to minimise the Mismatch cost function [54]. For points in the parameter space close to the true parameter value, the minimum value of $C_{MM}(\boldsymbol{\theta})$ is smaller than the value it takes in other areas of the parameter space. Therefore, the quality of the trajectories the model admits in the parameter space will vary depending on how well the pseudo-orbits shadow the true trajectory and the observations simultaneously.

A *quality measure* is a function of the parameter vector $\boldsymbol{\theta}$. It is a summary statistic of the distribution of the calculation of a quality measure or a summary statistic of the transformation of this distribution.

The method uses quality measures of the pseudo-orbits defined by, for example, summary statistics of the *shadowing time* distribution of the model trajectories starting in each point of the pseudo-orbit (see section 4.2.1.2)

and the *implied noise level* distribution given by the difference between the pseudo-orbits and the observations (see section 4.2.1.1).

The values of the quality measures, induce a structure in the parameter space. Such structure is then used to choose the estimator for the value of the model parameter.

Areas in $\mathcal{Q}$ are distinguished by the value a quality measure takes. The new method benefits the process of parameter estimation since dynamical information available in the PMS is made redundant in order to obtain a distribution instead of a single-guess estimate.

4.2.1.1 Implied Noise Level

In the PMS, the noise process is known and it is assumed to be additive (*i.e.* measurement noise) and IID normal as in equation (4.2). The noise model is Gaussian with mean zero and known standard deviation σ_η, and variance of the noise model, σ_η^2, is referred as the noise level.

Figure 4.9 is a diagram of the noise model for a noise level of 1%. Given that the noise model is unbounded, any realisation of the observations is possible and in several cases it differs very much from the system state $x_t = \tilde{x}_t$. This means that the distance between the state x_t and the observation s_t is bigger than $\sigma_\eta/2$ with a non negligible probability. The unfilled circles

in the Figure represent five realisations of the observation of x_t.

Note that in nonlinear systems it may be possible to extract a smaller implied noise level assuming an incorrect noise model through noise reduction techniques. For example, assuming that the noise process is uniform when it is known to be normally distributed. If apparent improvement of the observations is obtained in the case of low noise levels then it is completely justifiable to believe in an incorrect model noise, otherwise the use of an incorrect noise model set the problem outside the PMS.

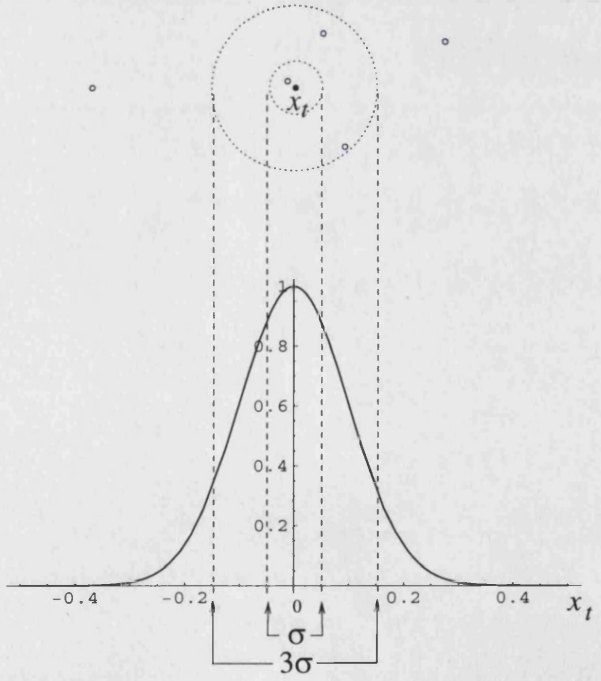


Figure 4.9: The noise model is assume to be Normal with standard deviation σ. The Figure shows 5 realisations (white dots) of the observation of x_t (black dot) with a noise level of 1%.

By gradient descent, pseudo-orbits $\{z_t\}_{t=1}^N$ are generated such that they are indistinguishable from the system trajectory segment $\{x_t\}_{t=1}^n$ given a noise model as equation (4.2) (see Figure 4.9) for values of $\boldsymbol{\theta}$. This implies that the $\{z_t\}_{t=1}^N$ contain a net lower noise level than the original observations $\{s_t\}_{t=1}^N$ since they are constructed to be on average closer to the true trajectory of the system.

Figure 4.10 shows schematically the relative distance of a pseudo-orbit state, z_t, an indistinguishable state y_t, the true state x_t, and the observation s_t. After several iterations of the GD algorithm a pseudo-orbit state z_t is obtained. Since y_t is indistinguishable from the system state x_t given the

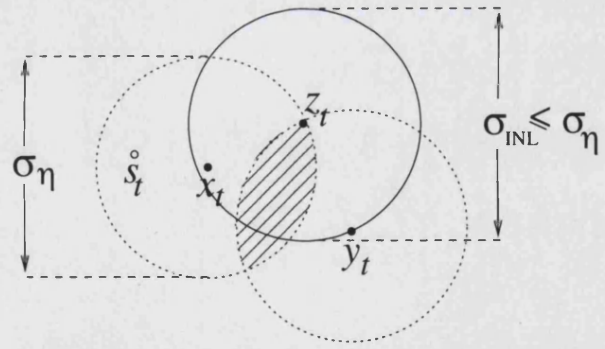


Figure 4.10: Diagram the relative location of a pseudo-orbit state z_t with respect to the true state x_t, an indistinguishable state y_t and the observation s_t. The z_t is located at least at half of the distance between y_t and x_t.

noise model, z_t should be located in the shaded area in the diagram.

Note that the GD algorithm could be seen as performing noise reduc-

tion [26, 27], therefore the distance between the observation s_t and z_t should be less than $\sigma_\eta/2$ for realisations of the observations inside the circle centred in the true state x_t. For realisations of the observations outside the circle around x_t, $z_t \leq x_t + \sigma_\eta/2$ at least.

Figure 4.11 shows the value of the Mismatch and Least Squares (or RMS) cost functions for the Ikeda map in a 1-dimensional parameter space defined by $0.79 \leq \mu \leq 0.87$, taking the $\{z_t\}_{t=1}^N$ at the iteration 2048 of the GD. The vertical line marks the location of the true value of the Ikeda parameter. Broken lines correspond to RMS values and solid lines for the mismatch. The plots show results for two noisy observations of Ikeda map for noise level of 1% in black and 5% in grey.

As expected, the location of the minimum for the MM cost function for both noise levels is closer to the true value than for the RMS cost function since the mismatch is minimised by gradient descent. This implies that the net noise level of the obtained pseudo-orbits z_t is less than the one that the observations contain. Note that these curves for the Ikeda's RMS values are the same as the ones in the bottom panel of Figure 4.3. The MM cost function is evaluated in the pseudo-orbits states z_t.

The sensitivity of the estimate of μ, $\hat{\mu} = \min_\mu C(\mu)$, to the noise levels is lower when the C_{MM} is used instead of C_{LS}. This fact confirms that the pseudo-orbits obtained from GD iterations contained less net noise level than

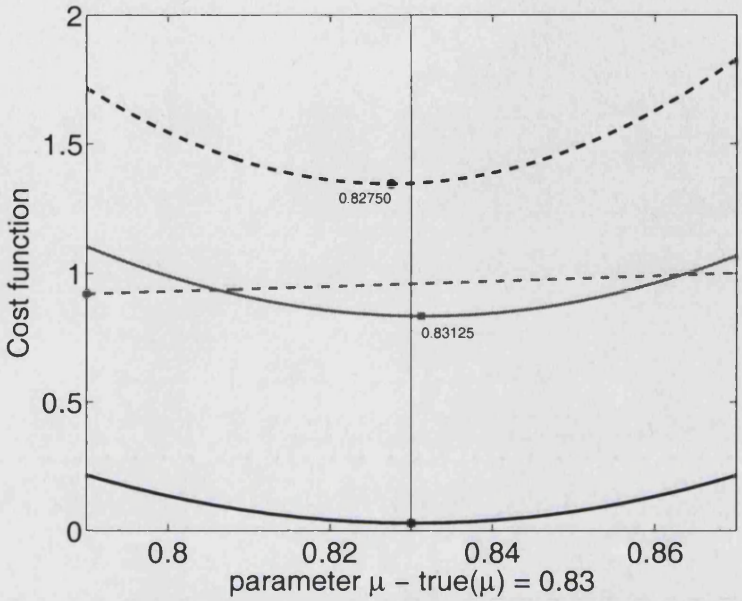


Figure 4.11: Mismatch (solid lines) and RMS (dashed lines) cost function values for Ikeda's parameter μ for observations with 1% (black lines) and 5% (grey lines) noise levels. The true parameter value is located with a vertical line.

the original observations even for high noise levels as 5% (≈ 2.0 in variance). Table 4.2 shows the summary of estimates and errors for both cost functions plotted in Figure 4.11.

In order to quantify the quality of the pseudo-orbits as being less noisy than the observations, lets define the Implied Noise Level (INL) as one of the relevant quality measure of estimated points in the parameter space. The INL quality measure is going to be used, both for monitoring convergence of the GD algorithm and for highlighting dynamical information contained in the pseudo-orbits.

	1%	5%
$\hat{\mu} = \min_{\mu} C_{MM}(\mu)$	0.83000	0.83125
$\lvert\tilde{\mu} - \hat{\mu}_{MM}\rvert$	0	1.25×10^{-3}
$\hat{\mu} = \min_{\mu} C_{LS}(\mu)$	0.82750	0.79000
$\lvert\tilde{\mu} - \hat{\mu}_{LS}\rvert$	2.5×10^{-3}	4.00×10^{-2}

Table 4.2: Estimates and error for the Ikeda's parameter μ for the MM and LS cost functions.

The implied noise level is defined as

$$INL(t;\boldsymbol{\theta}) = \sqrt{|s_t - z_t|^2}, \tag{4.21}$$

the distance between the observations and the pseudo-orbit state at time t for a given fixed parameter value $\boldsymbol{\theta}$, where $|\cdot|^2$ is the Euclidean distance in the state space. At $T = 2048$ iteration of the GD algorithm, the calculation of (4.21) generates an empirical distribution of $N = 512$ values of the implied noise level for a segment of observations and a pseudo-orbit of the same length.

Figure 4.12 shows the summary statistics for the quality measure distribution, $INL(\boldsymbol{\theta})$ resulting distribution for 1% (upper panel) and 5% (lower panel) noisy observations of the Ikeda map after approx. 2048 iterations of the GD algorithm. The names of the summary statistics calculated are in-

cluded in the right hand axes at the same level of the corresponding curve. The standard deviation of the noise model is marked as "True" with a horizontal line, the mean and median of the distribution are plotted in solid lines whilst the 95 and 99 isopleths in broken lines.

In both cases the mean and median of the distribution of the INL distribution are lower than the true noise level, confirming the lowest net noise level in the pseudo-orbit after $T = 2048$ GD iterations. The minimum is located close to the true parameter value $\tilde{\mu} = 0.83$. The location of the minimum of each summary statistic is marked with a dot and from the Frequentist perspective it could be used as an estimate of the parameter value and uncertainty in the parameter estimates.

The improvement of the estimation at this moment is made by the introduction of dynamical information by the "pre-processing" of the observations by means of the GD algorithm. Estimates can be performed at this point but uncertainty in the model is still not accounted for despite the fact that GD algorithm exploits the dynamics of the model $f(\cdot)$ to improve the observations in the generated pseudo-orbits by minimising the Mismatch. Relevant dynamical information can still be extracted from the pseudo-orbits by selecting the pseudo-orbits states that best shadow the model trajectory as presented in next section.

The calculation of the distributions of the INL and the shadowing times

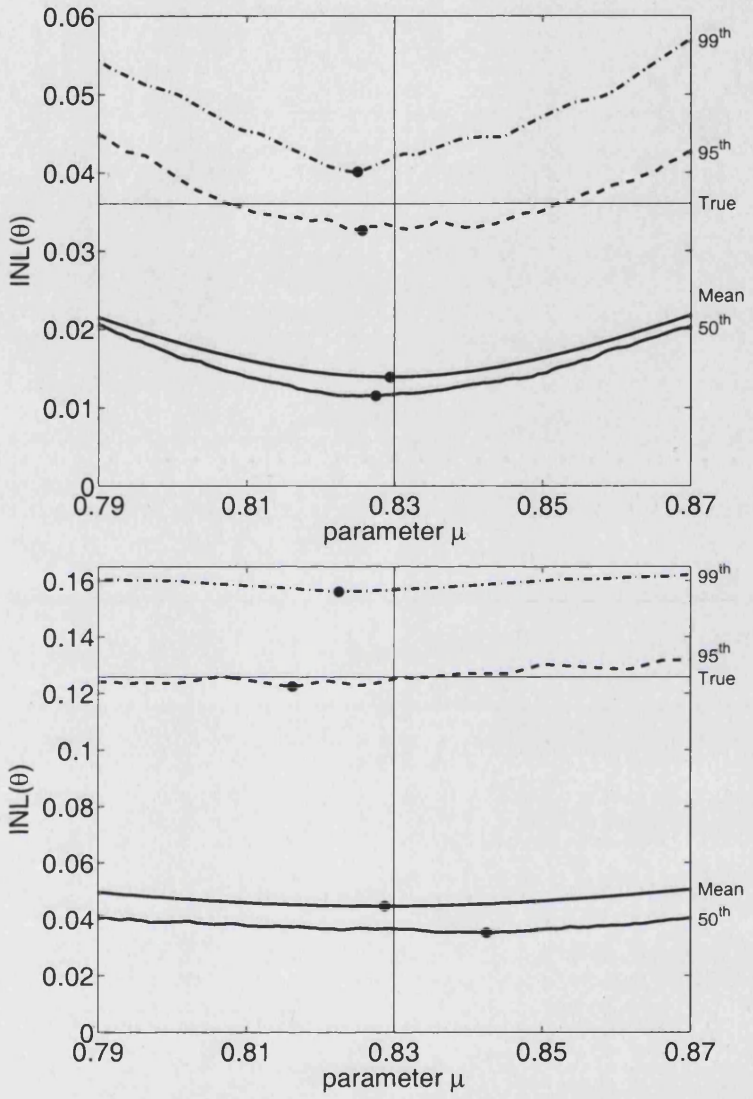


Figure 4.12: Implied noise level $INL(\boldsymbol{\theta})$ summary statistics for observations of 1% (top) and 5% (bottom) noise levels. The mean and the median of the INL distribution are plotted with solid lines, σ_η is marked with a horizontal line and the 95 and 99 isopleths with broken lines. The minimum of each summary statistic is marked with a dot.

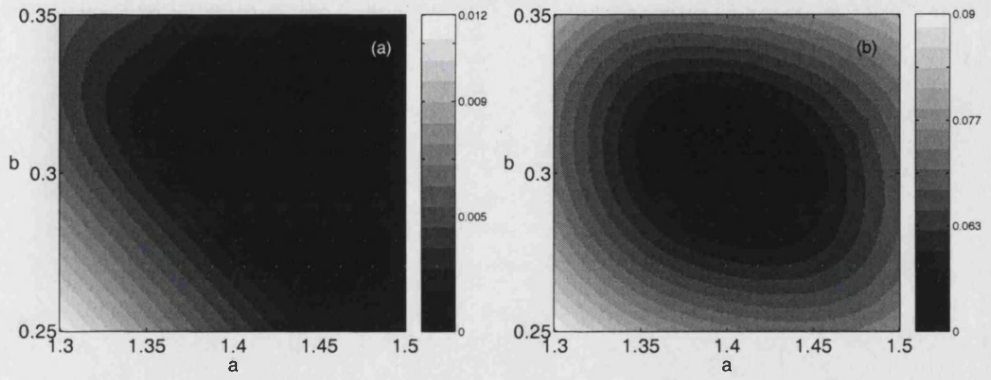


Figure 4.13: Pseudo-orbit information in the parameter space for the Hénon map. GD on 1024 observations with 5% noise level. Mismatch (left) and implied noise level (right) standard deviations [88].

for dimensionality of the parameter space $\mathcal{Q}$ larger than one is presented in [88] for the Hénon map and reproduced here in Figure 4.13. Results for Ikeda are also available.

Figure 4.13 shows the implied noise level for the resulting pseudo-trajectory's mismatch (left) and noise level standard deviations (right) for the Hénon map in a 2-dimensional parameter space. The true parameter value is located in the centre of the parameter space plotted, *i.e.* $\tilde{\boldsymbol{\theta}} = (1.4, 0.3)$.

The structure of the parameter space points to a candidate area, that in fact contains the true parameter value which had generated the data. The minimum area for the implied noise level offers a well defined area from which to select the parameter estimates, information on the short term dynamics is reflected in the pseudo-orbits in the area where the mismatch is

minimised. Even though, the correct neighbours of the true parameter value are highlighted by the INL and MM cost function, further distilation of long to medium term dynamics has to be performed by calculating the shadowing time distributions.

4.2.1.2 Shadowing Distributions

The dynamical information contained in the model structure chosen to be the system under study (PMS) is extracted via the generation of shadowing trajectories.

Whether or not a model shadows the observations $\{s_t\}_{t=1}^{N}$ [58, 82, 34, 87] as a goal of the study is similar to the one posed at this stage of the study, *i.e.* from suitable candidates, shadowing trajectories are generated. The difference is rooted in the fact that the shadowing trajectories start in pseudo-orbit states, z_t, thus they will be admissible and consistent with the dynamical and noise model rather than to find any other shadowing trajectory. Details of this difference between ι-shadowing and shadowing time distributions as function of parameter values are not addressed in depth here.

At this point, the uncertainty in the model has to be accounted for in order to obtain more reliable parameter estimates and in-sample forecasting. The

task of finding shadowing trajectories is posed in a restricted sense. Given a segment of observations, the goal is to find the trajectories the system admits. A trajectory is admitted if the the residuals defined by the trajectory starting in a candidate initial state and the observations are consistent with the noise model. The largest admissible shadowing trajectory, starting at time t, is of length τ for some model states $\{x_i\}_{i=1}^{\tau}$, the distribution of its residuals r_i for $i = 1, \ldots, \tau$ is consistent with the dynamical noise model.

Formally, let $\{c_i\}_{i=1}^{\tau}$ be a shadowing trajectory starting at time t from a a candidate state. For simplicity, take the candidate for the initial state of the admissible shadowing trajectory as the pseudo-orbit state $c_1 = z_t$. The time series of the *residuals* r_i is defined by

$$r_i = |s_i - f^i(c_1, \boldsymbol{\theta})|, \tag{4.22}$$

for $i = 1, \ldots, \tau$.

In order to find the length τ of the admissible shadowing trajectory, the noise model plays a key role. For bounded (*i.e.* uniform) noise, a candidate c_1 generates a shadowing trajectory with shadowing time τ if the residuals r_i are less than the bound. In other words, there exists a cylinder around the observations $\{s_t\}_{t=t+1}^{\tau}$ such that it contains the shadowing trajectory for τ time steps.

For IID Normally distributed noise, as equation (4.2), there are a variety

of approaches to calculate the shadowing time. The simple method used here is based on *thresholds* defined by the distribution of residuals for a candidate c_1.

Given that the noise model is unbounded, any observation is conceivable, so relevant shadows are searched for within a certain probability bound [10]. The probability bound is defined by testing the null hypoThesis

$$H_0 : r_{i+j-1} \sim \mathcal{N}(0, \sigma_\eta^2), \tag{4.23}$$

for $i + j = 1, \ldots, n_s$, where $n_s \leq N$. The maximum shadowing time n_s of the trajectory is calculated such that the following conditions hold:

1. The 90% isopleth of the residual distribution falls below the 99^{th} percentile of the distributions of 0.90 isopleths given n_s draws from a Gaussian distribution,

2. The median of the residual distribution falls below the corresponding 90^{th} percentile for the median of the noise model.

Holding these two conditions simultaneously implies a resection rate of 0.001 provided that $n_s < 100$. This way of calculating the shadowing time distribution for unbounded noise is still open to development and refinement.

Figure 4.14 shows the results for the conditions described above. It shows the distribution of several isopleths for the Ikeda map demonstrating the resolution of the test.

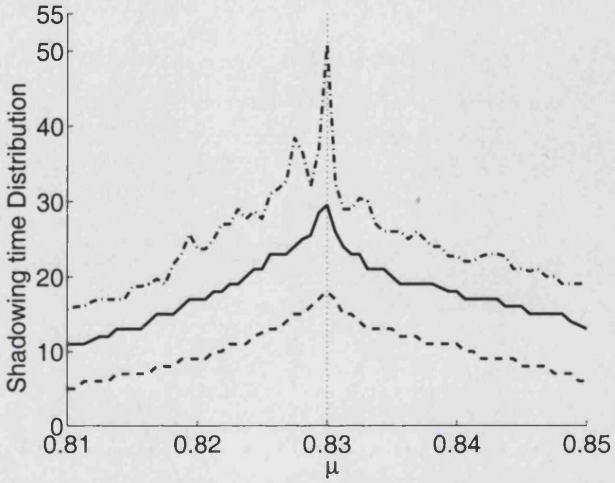


Figure 4.14: Shadowing Time Distribution for Ikeda observations with 5% noise level. The dotted line is the median, 90% and 99% shadowing isopleths in solid and dotted lines respectively [88].

For a given segment of observations, most interest is put in the longest shadowing trajectory, thus the shadowing time is defined as

$$\tau_s = \max_x \left[\tau(c_i)\right], \tag{4.24}$$

where the maximum is taken over all c_i values tested. The results presented in Figure 4.14 only test three candidates: the pseudo-orbit z_t, the image of the previous point on the pseudo-orbit $f(z_{t-1}; \boldsymbol{\theta})$, and the average of these two. Thus a distribution of shadowing time for a given parameter vector $\boldsymbol{\theta}$ is obtained.

When the method is applied to a high dimensional parameter space, for example $\mathcal{Q} \subset \mathbb{R}^2$, the results are compared with results obtained in [68]

and section 4.1.1. The right panel of Figure 4.15 shows the 50^{th} isopleth of the distribution of shadowing time obtained for 1024 Hénon observations after using the GD algorithm. Similar results are obtained for Ikeda for the distribution of shadowing time. Further exploration of the shadowing time distributions as a function of the parameter space of two-dimensional chaotic maps is planned for future research.

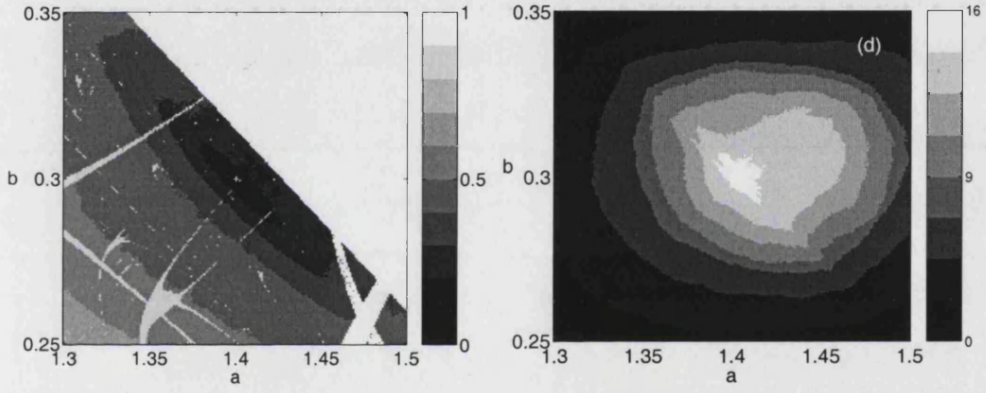


Figure 4.15: Information from a pseudo-orbit determined via gradient descent applied to a 1024 observations of the Hénon map with a noise level of 5%. (c) at the left is a cost function based on the model's invariant measure (after Fig.4(b) of ref [68]). (d) at the right shows the median of shadowing time distribution.

This result is one of the main results of this work and it is presented in [88] in collaboration with L.A. Smith, H. Du and K. Judd. The statistics of the shadowing time distribution provide a sharp indication of the location of the true parameter value. From Figure 4.14 it is clear that the information about the short term dynamics shows independently that the choice of a

particular isopleth is not important. Despite the fact that refinements are to be performed in the calculation of the thresholds for the shadowing time, the location of the true parameter value can be robustly estimated.

In addition, it clarifies and complements the information obtained by estimating the invariant measure of the map by the calculation of the $C_{ML}(\boldsymbol{\theta})$ cost function [68] discussed in section 4.1.1. As discussed in that section, the "tongues" and complex structure in the parameter space are due to sensitivity of the dynamics to the parameter values, although the minimum is located in the relevant regions of $\mathcal{Q}$.

In addition, the methodology can be extended and improved on technical details such as the calculation of thresholds, the use of sliding windows with lengths shorter than that of the observations, and the implementation of subsequent iterations of the method. Additional iterations of the method imply re-applying the GD algorithm to candidates of shadowing trajectories in the relevant areas of the parameter space and/or calculating the shadowing time for higher resolution grids in the interesting areas in order to discover finer structure and to reduce the candidates of true parameters [88].

In the light of the IMS, this new methodology can give insight in the areas of the parameter space where a model performs better than others. With only one imperfect or inadequate model to represent the system of interest, the information provided by the corresponding invariant measure is

not informative by itself. In the IMS, the quantification of the goodness of a parameter should be re-formulated in its usefulness rather than its error estimates. The existence of true parameter values is an utopian idea.

4.3 Summary

This Chapter focuses on how dynamical information can be introduced into traditional statistical methodologies of parameter estimation in the PMS, in particular in a cost function based approach. "Pre-processing" of observations exploiting the knowledge of the perfect model enhances the identification of parameter estimates via the definition of new cost functions. In addition, once pseudo-orbits are found by gradient descent [79] using indistinguishable states theory [54, 55], the calculation of shadowing time distributions for model trajectories starting in the pseudo-orbit states, distils dynamical information of the invariant measure of the model in the parameter space [88].

Variations in the shadowing time distribution with parameter values yield more insight than maps of root-mean-square forecast error, or other linear statistics which have well-known shortcomings in nonlinear models (see [68] and references therein) as discussed in section 4.1.

The problem of parameter estimation in nonlinear systems differs fun-

damentally from its well-understood equivalent in linear systems. In linear stochastic systems, nearby parameter values result in similar short term dynamics and similar asymptotic distributions. In the nonlinear dynamical framework, the uncertainty on any parameter estimate is not necessarily completely described by the difference between true parameter and the estimates. Uncertainty measures determine how good a parameter estimate generates numerical data most resembling the observed data.

Within the PMS, the method presented in this Chapter might profitably be recast in a Bayesian framework as partly discussed in Chapter 5, where state estimates obtained by this geometrical approach to parameter estimation and Bayesian methodologies are compared. Difficulties in traditional Likelihood based approaches in this context are clarified by Berliner [5] and listed in section 4.1. Coherent Bayesian formulations that condition the probabilities extracted, upon all the information available [83, 42, 43] are still to be developed. This formulation explores a wide new research area of time series analysis where statistical methodologies enter in contact with nonlinear time series analysis as seen in [16, 36, 92, 12, 76, 88].

For the systems considered here, the information available includes the fact that the data is generated by a deterministic system whose mathematical form is known. To assume some stochastic model for the sake of applying a "Bayesian technique" is to fail to grasp the fundamentals of Bayesian anal-

ysis. The dangers are discussed and illustrated in Chapter 3 and 5, [52], in addition results to be presented in [24].

Figure 4.15 shows clearly how the maps of shadowing time provide complimentary information quantifying the time scales on which the model dynamics reflect the observed behaviour [88]. Better parameter estimates can be obtained by quantifying the realism of both the short term dynamics by the shadowing trajectories and the long term dynamics through the invariant measure mapped in the parameter space.

It is important to emphasise that only in the perfect model case is it certain that a balance between the information in the dynamic equations and the information in the observations exists. When in the IMS, the invariant measure is not expected to be informative.

By considering shadowing times, the identification of parameter values which can reproduce the dynamics, quantification of the time scales on which they can shadow, and extracting information for improving the model class itself can be performed. This is a significant step forward. It remains to explore in more detail the performance of this approach in a variety of systems, model classes and noise levels.

Many thanks to Hailiang Du for his collaboration in parallel calculations and plots showed in this Chapter, as well to Lenny Smith and Jochen Broecker for many discussions and insightful comments.

Chapter 5

Gradient Descent vs Markov Chain Monte Carlo

Given a set of observations and any model scenario, no matter the model type and methodology used, parameter estimates are always obtained. Estimates are numbers that only have meaning in the context of the application of interest. At one point, what it is interesting and challenging is the quantification of the quality of the resulting estimates more than the process of generating estimates.

Throughout this Thesis, it has been studied and exemplified that any methodology to find parameter estimates is valid, if and only if it is consistently formulated, interpreted and contrasted out-of-sample with the behaviour of the system not used for the estimation. Although the use of a

particular methodology to estimate model parameters is always justifiable, it requires a detailed formulation of the problem scenario in which it is used (see Chapter 1 for the definitions of scenarios and approaches used in this Thesis). Both methodology and problem scenario formulation constitutes an *approach* for parameter estimation from time series. Failure to provide a consistent problem scenario results in misuses of a methodology and misinterpretation of the resulting estimates. In-sample performance of the model trajectories generated using estimated parameter values is only a safety check for the methodology used to generate estimates in the PMS. Whilst out-of-sample performance is a test for reliability and efficiency of the performance of the resulting estimates in forecasting and control monitoring tasks.

When the problem of parameter estimation is approached in the NSA. The PMS formulation clearly states that some model parameters are not paired with any system parameter. Thus, the estimates resulting for those "non-paired" parameters often are ignored or misinterpreted since they cannot be translated in the context of the system, *i.e.* dynamical noise variance parameter introduced in the model for the Logistic map when the observations only have measurement noise. As discussed in Chapter 1, section 3.2.1.2.

In the case that the system is complex and a Naive Realistic Approach (NRA) is necessarily used (see Chapter 6), consistent estimates may be found

across several models in order to account for *model impersection error.* The interpretation of the resulting estimates in context of the system is impossible for the majority of model parameters. The perfect model of the system is unreachable, therefore system parameters are unknown [15, 86, 18]. Furthermore, system observations do not even correspond to any model parameter or variable. If other models are not used and cross-compared, parameter estimates lack meaning outside the context of the model.

The interest of this Thesis is focused on model parameter estimation for nonlinear models. Through the process of parameter estimation, parameter estimates are generated along with estimates of model states (see Chapter 3, 4 and 6). Obtaining simultaneously, model state estimates and model parameter values, opens a new application of parameter estimation techniques to generate an *ensemble* of states and in-sample checks. In one way or another, most parameter estimation techniques produce additional traces of model trajectories for a given temporal window, *i.e.* exploring the model state space consistently with the observations and the model.

Depending on the methodology, some uncertainty measures are produced in the process of parameter estimation for both model state and parameter estimates and in turn they can be used for ensemble characterisation. Without uncertainty measures, estimates for model states parameters are meaningless.

In this chapter, the interest is put in the state estimates obtained through the process of parameter estimation instead of the parameter estimates themselves. Several questions can be asked about these state estimates produced naturally by a particular methodology of parameter estimation if they want to be seen from a dynamical perspective. Among other questions, it is relevant to ask:

1. What are the set of state estimates?

    i. A pseudo-orbit.

    ii. A model trajectory segment.

    iii. A random realisation of a certain stochastic process.

2. What is the dynamical information content of the state estimates?

3. Are the state estimates samples of the invariant measure of the model?

4. Is there any deterministic structure in the resulting state estimates for a given temporal window?

5. How uncertain are the state estimates?

To tackle these questions a comparative study is presented in this Chapter. The state estimates produced by different methodologies are compared to each other and to the true trajectory.

State estimates are produced for the Logistic map from observations that are known to contain only measurement noise. The state estimates are produced using the following methods:

1. MCMC techniques are used to estimate parameters in the PMS. As noted in Chapter 3, the PMS is defined for observations containing both additive noise components, dynamical and observational. Therefore, parameter estimation is made using the NSA.

2. Parameter estimates are obtained in the PMS by the new geometric methodology described in Chapter 4 based on the indistinguishable states theory [54, 55].

This study aims to compare system state estimates obtained by GD and MCMC algorithms. The study of dynamical features of the TLS state estimates are left to future work. From the results presented in Chapter 4, in the case of TLS state estimates there is evidence that the presence of any Logistic structure in the reconstructed state estimates is the result of the explicit inclusion of Logistic map equations in the Likelihood.

The comparison between GD and MCMC state estimates (items 1 and 2 respectively) is set as a "blind" test. The test is blind since it is assumed that there is no information about the approach used to obtain the series of distributions of system states. Information about the methodology will imply

a direct comparison of the methodologies. Comparing the methodologies is out of the scope of this work, GD and MCMC belong to *non commensurable* paradigms [59]. Here, only numbers are compared.

The empirical distributions of state estimates for a given temporal window can be seen as an ensemble of model states for all and each t. The goal of the study is to find out which of the ensembles of state estimates contains "more" dynamical information, *i.e.* how much they reflect the invariant measure of the model. Quality points are assigning from values that a quality measure takes. The rulers to measure "dynamical information" content in the resulting state estimates are defined based on geometric ideas described in Chapter 4, *i.e.* implied noise, mismatch values and shadowing time distributions, and it is open to further refinement.

The study proposed here is directly related to the problem of distinguishing between random and deterministic behaviour even though it is directed formulated as follows:

> *Given two empirical distributions of state estimates, which of the distributions of state estimates contains "more" dynamical information and reflects "better" the invariant measure of the model?*

Despite the fact that distinction between random and deterministic behaviour is open and likely unsolvable, in the context of this Thesis, is it possible to distinguish some dynamics in any of the sets by comparison of several pre-defined quality measures?

Noisy Logistic observations are generated for $t = 1, \ldots, N$ for $N = 100$, with noise levels ranging from 0 to 2, and for the true parameter value $a = 1.85$ and the initial condition $x_0 = 0.3$.

The noisy observations are used to generate system state estimates $\{z_t\}_{t=1}^{N}$ using both GD algorithm and the Bayesian implementation for the Logistic map in the PMS and solved using a NSA. Both algorithms, GD and MCMC, are run for approximately 1.1×10^5 iterations. In order to assure convergence, the first 1×10^4 iterations are discarded in both cases. Analysis of the estimates is made to empirical distributions composed by $T = 1 \times 10^5$ values for each system state at time $t = 1, \ldots, N$. Each state estimate is a distribution of T values.

Denote the state estimates generated using the geometric approach (*i.e.* GD algorithm) by $\{z_t^{(g)}\}_{t=1}^{N}$ and estimates using the Bayesian perspective (*i.e.* MCMC algorithm) by $\{z_t^{(m)}\}_{t=1}^{N}$.

Figure 5.1 shows the estimated Logistic states for observations with a noise level of 0.2 using GD in the left column and MCMC in the right column. The three states shown are $x_{50} = 9.945 \times 10^{-3}$ in the first row, $x_{18} = -0.313$ in the second row, and $x_{68} = 0.548$ in the last row. The mean of the resulting distribution of states is marked with a black solid vertical line and with a grey dashed line for the median. Each plot includes the error between the

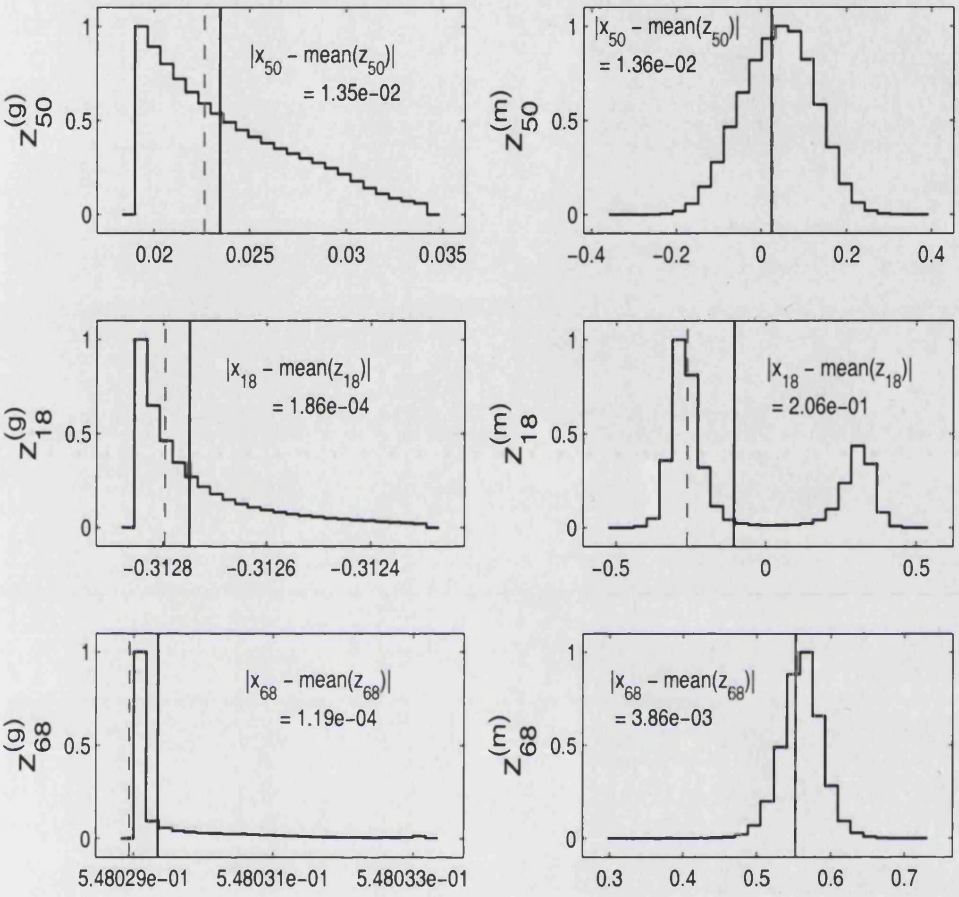


Figure 5.1: Histograms from three Logistic states estimated using GD (left) and MCMC (right). The three states are $x_{50} = 9.945 \times 10^{-3}$ (first row), $x_{18} = -0.313$ (second row) and $x_{68} = 0.548$ (third row). The mean and median of the resulting estimates are located with a black solid and a grey dashed vertical lines respectively. Note that the horizontal scales in each column are one order of magnitude different

true state and the mean of the estimates. This error at time t is defined as

$$E_t = |x_t - z_t|\,, \tag{5.1}$$

where $|\cdot|$ denotes the Euclidean distance between the true state x_t, and the state estimate z_t at time t, for a Logistic trajectory starting in $x_0 = 0.3$ and for the true Logistic parameter $a = 1.85$.

In Figure 5.1, the width of the distributions obtained by GD algorithm is several orders smaller than the width observed in the states estimates generated by MCMC algorithm, in some cases more than 3 orders of magnitude smaller. Table 5.1 presents several summary statistics for the three states, one close to zero, one negative and one positive. Figure 5.1 and Table 5.1 show typical results obtained using GD and MCMC techniques. In mean, both distribution of states estimates seem to perform similarly well, *i.e.* errors are small, however MCMC distributions tend to be wider and bimodal with symmetry around zero. These results are consistent for all noise levels studied, $0 \leq \sigma_\eta^2 \leq 2$.

Figure 5.2 shows the summary statistics for the trace of several summary statistics of the estimates distribution for $32 \leq t \leq 96$. The estimations are obtained from Logistic observations with a noise level of $\sigma_\eta^2 = 0.2$. The upper panel plots estimated states obtained using the GD algorithm whilst the lower panel plots the estimates obtained using MCMC. The plots show the 95^{th} and 75^{th} isopleths as light and dark grey areas respectively, the median is plotted with black crosses and the true states with black pluses.

x_t	Type	mean(z_t)	50%	99%	var(z_t)	std(z_t)	E_t
0.00995	GD	0.02346	0.02262	0.03287	1.37×10^{-5}	0.00370	0.01352
	MCMC	0.02352	0.02628	0.20227	0.00691	0.08313	0.01358
-0.31293	GD	-0.31275	-0.31280	-0.31235	1.58×10^{-8}	0.00013	0.00019
	MCMC	-0.10694	-0.26327	0.36179	0.07140	0.26721	0.20600
0.54791	GD	0.54803	0.54803	0.54803	1.03×10^{-12}	1.02×10^{-6}	0.00012
	MCMC	0.55177	0.55227	0.61948	0.00077	0.02769	0.00386

Table 5.1: Summary statistics for the resulting distributions of state estimates obtained from observations of noise level 0.2 using GD and MCMC algorithms for three true states: x_{50}, x_{18} and x_{68}.

The uncertainty in the MCMC estimates is bigger than the GD estimates. Wider distributions are obtained for the Logistic states estimates for all t when the MCMC is used.

Figure 5.3 plots the percentage of error calculated form equation 5.1 with respect to the true trajectory. In the Figure, the 95^{th} isopleth of the resulting distribution of errors is plotted as a grey area, where the upper panel shows the errors for the GD algorithm whilst the lower panel the errors obtained using MCMC. The black dots represent the median of the error at time t whilst the coloured area the 95^{th} isopleth of the error distribution.

The error is close to zero for most values of t when the GD is used to generate state estimates however there are six events when the error suddenly

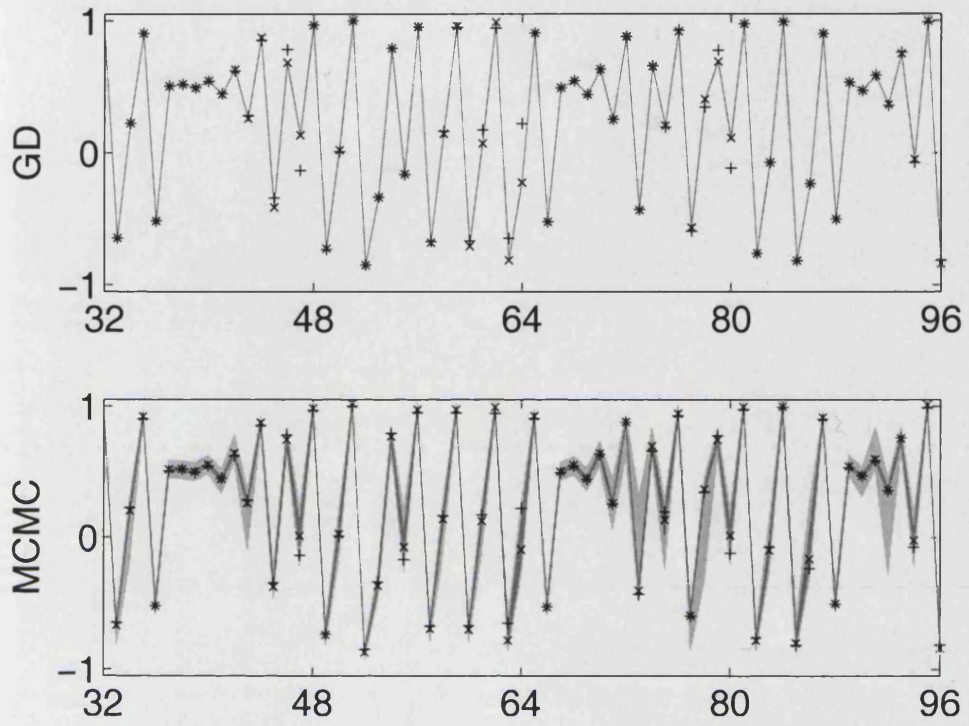


Figure 5.2: Logistic state estimates for observations with noise level of 0.2. The 99^{th} and 75^{th} percentiles correspond to the light and dark coloured areas. The median of the state estimates is plotted with black crosses and true Logistic states with the black pluses.

increases to values between 10% and 25% of the size of the attractor. Despite these jumps, the distribution of errors is sharp enough to reflect the complete convergence of the GD algorithm after a number of iterations, *i.e.* the values obtained for state estimates can be considered as indistinguishable states from the true trajectory.

When the MCMC algorithm is used, it is surprising that the median of

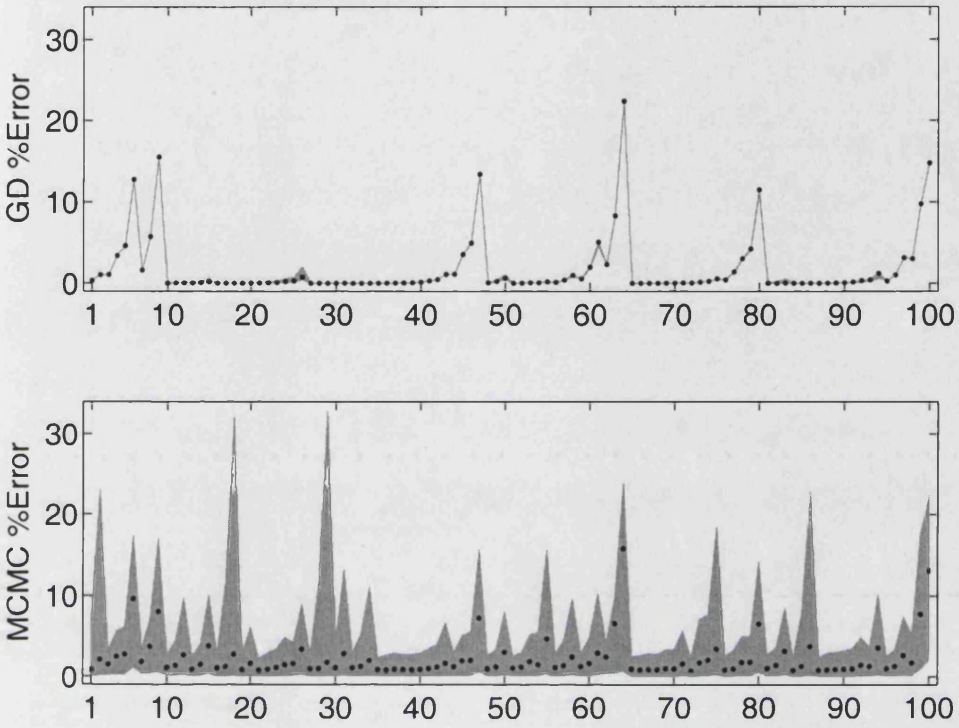


Figure 5.3: Percentage of error as a function of t, for $0 \leq t \leq 100$, for the Logistic state estimates using GD (top) and MCMC (bottom). Estimates are obtained from observations with noise level of 0.2. The 95^{th} isopleth is the grey area and the black dots are the median of the resulting distribution of errors.

the estimates is never bigger than 15% of the size of the attractor, *i.e.* for $1 \leq t \leq 100$. In fact, the GD state estimates are closer to the true trajectory in median than the MCMC estimates. At the same time error distributions are up to 2 orders of magnitude wider for the MCMC estimates than the ones obtained using GD.

Figures 5.1 and 5.3, clearly show how fast the GD algorithm is converging

towards pseudo-orbits close to the true trajectory whilst the MCMC algorithm is showing a weak and slow convergence to the true trajectory. As discussed in Chapter 3, convergence of the MCMC technique is not uniformly reached given the nonlinearities of the map and multimodality in the corresponding Likelihood. This behaviour is consistent with all noise levels studied, $0 \leq \sigma_\eta^2 \leq 2$.

After inspecting summary statistics for the traces and errors with respect to the true trajectory to assess qualities of the resulting empirical distributions of states, the *Average Implied Noise Level*, AINL , is calculated for both state estimates, $\{z_t^{(g)}\}_{t=1}^N$ and $\{z_t^{(m)}\}_{t=1}^N$, in order to assess the performance of each algorithm as a noise reduction method.

The AINL is defined by

$$AINL = \sqrt{\frac{1}{N}\sum_{t=1}^{N} |s_t - z_t|^2}, \tag{5.2}$$

where s_t is a noisy Logistic observation and z_t the estimated state. The square root is taken in (5.2) in order to obtain the same dimensional units as the model states x_t and the noise deviation σ_η. INL and AINL can be interpreted as an estimation of the original noise level present in the original observations in the case that the noise level is unknown.

For state estimates $\{z_t\}_{t=1}^N$ closer to the true trajectory, the AINL value is closer to the square root of the original noise level of the observations, *i.e.*

$AINL \equiv \sigma_\eta$ for perfect noise reduction.

Figure 5.4 shows the results for both, GD (left panel) and MCMC (right panel) algorithms, for each noise level a distribution of AINL values is obtained. Each panel plots the 95^{th} isopleth as a dark grey area whilst the 75^{th} isopleth as a light grey area. The median of the resulting distribution is plotted in a solid black line. The x-axis takes values for the deviation of the noise level, *i.e.* $0 \leq \sigma_\eta \leq \sqrt{2}$. The diagonal dotted line represents the original noise standard deviation. The diagonal is the line connecting the points $(\sigma_\eta, AINL) = (0, 0)$ and $(\sigma_\eta, AINL) = (\sqrt{2}, \sqrt{2})$.

In the case of the GD algorithm, the left panel of Figure 5.4 shows that the distribution of average distances between the estimated Logistic trajectories and the observations is consistent with the original noise level. Noise reduction of the algorithm is consistent and effective for all noise levels, as it has been noted also in [26].

In contrast, the left panel of Figure 5.4 shows that the MCMC algorithm is systematically overestimating the noise present in the observations. Overestimation of the noise level in the signal can be interpreted in two ways:

1. The algorithm is outperforming as a noise reduction method and the resulting state estimates are closer to the true trajectory in average and in median as seen in Figure 5.3.

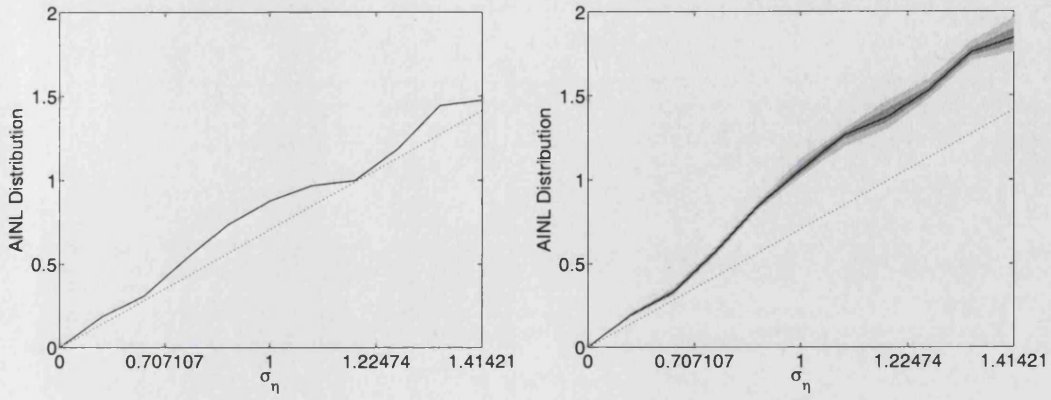


Figure 5.4: Average Implied Noise Level distribution summary for the GD (left) and MCMC (right) Logistic state estimates for noise levels form 0 to 2. The x-axis take values for the standard deviation of the noise level σ_η. The median is plotted in a solid black line, the 95^{th} and 75^{th} isopleths are plotted as light and dark grey areas respectively. The dotted diagonal marks perfect noise reduction.

2. The algorithm is in fact "over-reducing" noise but the resulting state estimates cannot be compared to the true trajectory since dynamical information might have been destroyed in the process.

Also note that the AINL values obtained for all noise levels when the GD algorithm is used, estimation of the original standard deviation of the noise process present is made from distributions of values less dispersed than the ones obtained when the MCMC is used. Table 5.2 lists mean and variance of the resulting distribution of AINL values as a function of the noise level for both GD and MCMC cases.

In order to start distinguishing dynamical information in each of the

Original Noise	$AINL_{GD}$		$AINL_{MCMC}$	
σ_η	Mean	Var.	Mean	Var.
0	9.16×10^{-4}	4.35×10^{-12}	9.72×10^{-4}	4.75×10^{-9}
0.44721	0.18516	6.17×10^{-8}	0.19422	3.82×10^{-5}
0.63246	0.30996	7.76×10^{-8}	0.33260	1.73×10^{-4}
0.77460	0.53204	1.85×10^{-7}	0.57546	2.07×10^{-4}
0.89443	0.73771	3.32×10^{-7}	0.84599	2.35×10^{-4}
1.00000	0.87741	3.36×10^{-8}	1.06164	9.64×10^{-4}
1.09545	0.96877	7.03×10^{-9}	1.25728	9.92×10^{-4}
1.18322	0.99885	1.22×10^{-6}	1.37638	2.67×10^{-3}
1.26491	1.18509	1.12×10^{-7}	1.53418	1.39×10^{-3}
1.34164	1.44311	5.15×10^{-7}	1.75973	1.11×10^{-3}
1.41421	1.47546	1.85×10^{-6}	1.85194	3.84×10^{-3}

Table 5.2: Mean and variance of AINL distributions as a function of the noise standard deviation for both GD and MCMC estimates.

estimated Logistic trajectories, the *Squared Mismatch* of the state estimates for each noise level is calculated. The Squared Mismatch value is defined from a modified version of the Mismatch cost function defined in equation (4.20), section 4.2, Chapter 4. The squared Mismatch value over a trajectory

is given by

$$C_{MM}^2 = \frac{1}{N-1} \sum_{t=1}^{N-1} |z_{t+1} - f(z_t; \tilde{a})|^2 . \tag{5.3}$$

and it is calculated for all noise levels, $0 \leq \sigma_\eta^2 \leq 2$, for both the estimates $\{z_t^{(g)}\}_{t=1}^{N-1}$ and $\{z_t^{(m)}\}_{t=1}^{N-1}$. An empirical distribution of $C_{MM}^2(\sigma_\eta^2)$ values is calculated for each time t for each of the states estimated for the Logistic trajectory.

Figure 5.5 shows the summary statistics of the resulting $C_{MM}^2(\sigma_\eta^2)$ distributions for both sets of state estimates. Squared Mismatch values are plotted in the left panel for estimates obtained using GD and in the right panel for estimates obtained using MCMC. Each panel plots the median in a black solid line and, the 95^{th} and 75^{th} isopleths in dark and light grey areas, respectively.

In the left panel of Figure 5.5, it is seen, for all noise levels, the GD squared mismatch values are less than 1×10^{-4}, showing a robust convergence of the GD algorithm to a pseudo-orbit close to the true model trajectory. This behaviour is consistent with the results shown in Figures 5.1 to 5.4.

For the MCMC estimates distribution of the squared Mismatch error (right panel), the squared mismatch values are three orders of magnitude larger than those obtained for the GD estimates reflecting poor performance of the one-step prediction error.

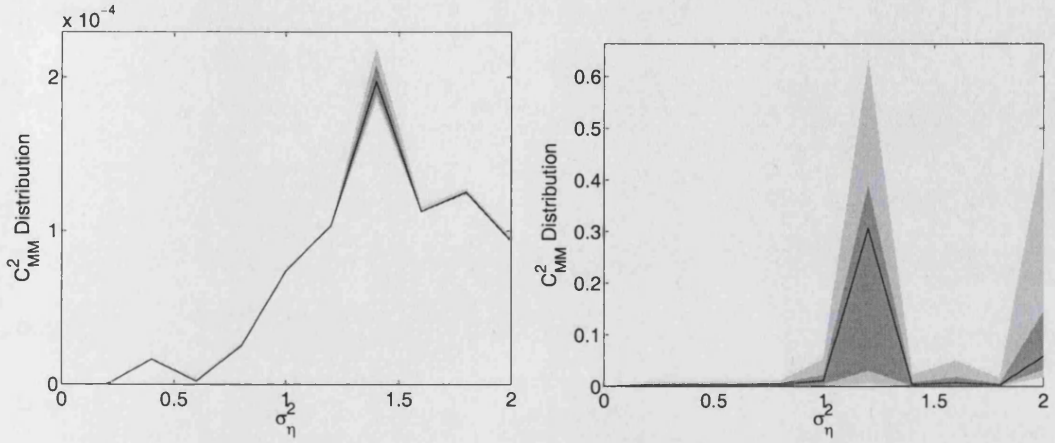


Figure 5.5: Mismatch distribution summary statistics for the GD (left panel) and MCMC (right panel) Logistic state estimates for noise levels form 0 to 2. The median is the black solid line, the 95^{th} and 75^{th} isopleths are plotted as light and dark areas respectively. Note that the vertical scales are four order of magnitude different.

It is clear from Figures 5.3, 5.2 and 5.5, that the states estimates obtained by the MCMC algorithm, $\{z_t^{(m)}\}_{t=1}^N$, and the ones obtained using GD, $\{z_t^{(g)}\}_{t=1}^N$, are both in the neighbourhood of the true trajectory, *i.e.* in both cases the obtained trajectory is a pseudo-orbit of the model.

The difference in amplitude between the values taken by the squared Mismatch by GD and MCMC estimates is clear evidence that the estimated Logistic trajectory obtained using the GD algorithm is closer to the stable set [54] of each point in the true model trajectory segment. Whilst in the case of MCMC, the estimated trajectory includes "less" information on the

invariant measure of the Logistic map for $a = 1.85$. Estimated states obtained by GD algorithm have a better prediction skill at one-step in time than those obtained using MCMC.

These results confirm that there is indeed a difference in the information content about the invariant measure between the two sets of estimates obtained using GD and MCMC. It highlights the fact that by construction of the state estimates one should expect to find more dynamical information in one distribution of state estimates than in the other, at least when looking at the squared Mismatch values calculated over a distribution of estimated trajectories.

Figure 5.6 shows the invariant measure histograms for a very long trajectory of the Logistic map (10^4 states) starting at $x_0 = 0.3$ for $a = 1.85$ in contrast to the histograms obtained from the state estimates produced by GD and MCMC algorithms. In the Figure, the true model trajectory's histogram is plotted in a black solid line, and the corresponding histograms for the GD and MCMC cases are plotted in a grey solid line and a light grey dashed line respectively.

Despite the differences described in this Chapter between both set of estimates when comparing the values taken by the AINL (see Figure 5.4) and Squared MM (see Figure 5.5), the invariant measure histograms resulting from GD and MCMC are qualitatively similar to each other and to the

true Logistic trajectory histogram. These similarities contradict conclusions drawn from Figures 5.1 to 5.5 about dynamical content of distributions of state estimates obtained by GD or MCMC. Reasons for this are still to be studied and should be tackled in future research.

Even though both MCMC and GD estimates can be seen as a (effective) samples of the invariant measure, MCMC estimates do not contain dynamical skill, *i.e.* shadowing time distribution is centered at longer times for GD estimates.

Further calculations planned in the disection of the proposed study include calculation of the shadowing time distributions in order to explore the forecasting skill of the state estimates obtained and their distance to the stable sets of the true trajectory. Forecasting skill of the state estimates provide insight from the dynamical point of view on the nature of both pseudo-orbits available from the iteration of both algorithms.

Although from the computational point of view, GD techniques are cheaper in computational and programming requirements than the MCMC algorithm, the interest of this study is the potential of the characterisation of the difference between the state estimates obtained. Such characterisation can then be exploited to efficiently meld statistical/probabilistic techniques and dynamical time series analysis techniques for parameter estimation, forecasting and control monitoring of dynamical systems.

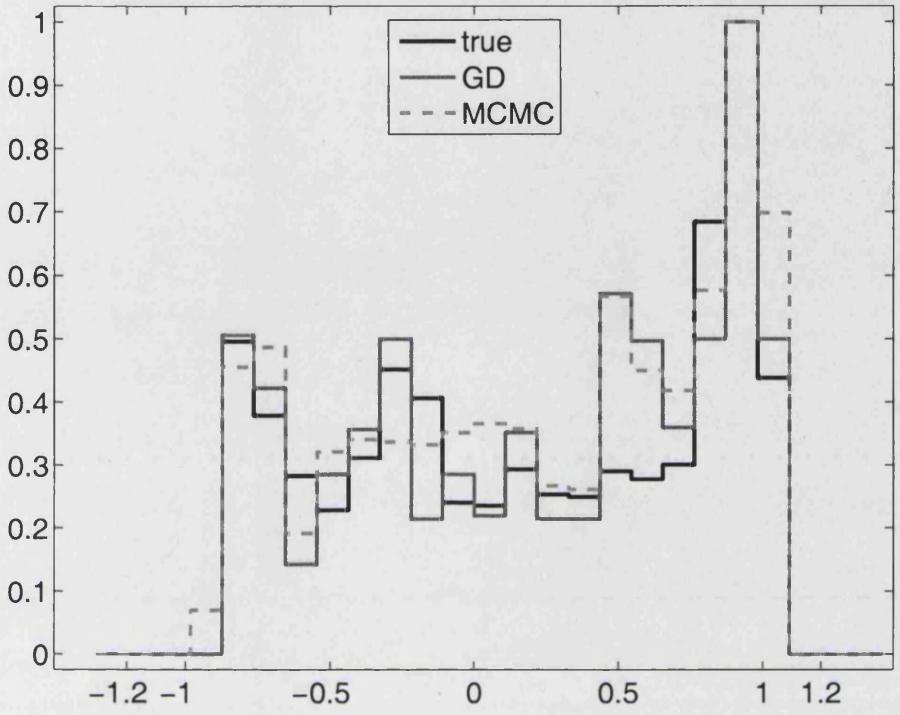


Figure 5.6: Invariant measure histograms for a long (10^4 points) true Logistic trajectory segment is plotted in a black solid line, and the histograms for the the GD and MCMC state estimates are plotted in a grey solid line and a light grey dashed line, respectively.

The brief discussion presented in this Chapter is only a milestone in the disection of extracting useful dynamical information from ensembles of dynamical states in the PMS.

In addition, it can also be extended into the general problem of distinguishing random from deterministic behaviour from time series given a model structure in the PMS or IMS. In particular, future results regarding this problem could be used to distinguish dynamical behaviour from pseudo-

orbits generated using the MCMC when the observations do not contain deterministic components as presented in section 3.3 [24].

This discussion provides insight on how deterministic dynamical components induced dynamical and deterministic structures in empirical probability distributions.

Chapter 6

Parameter Estimation from Real Time Series: The UK Electricity Grid Case

The work presented in this Chapter aims to produce a characterisation of the grid system that National Grid Transco (NGT) can incorporate into their grid management system. It is proposed that this characterisation will be based on a model for the grid system dynamics, specifically the grid frequency model, described in section 6.2. The estimated parameters of this model will then form the basis for a characterisation. The main results are presented in the work, *The Role of Operational Constraints on MCMC*

Parameter Estimation: The case of the UK electricity Grid, M.C. Cuéllar, L. Clarke, L.A. Smith and M. Brown, submitted to the IJEPES [22], and technical details are discussed in the REMIND Report [20].

NGT is the company that ensures the delivery of electricity to consumers across England and Wales, and is required [1] among other things, to maintain the electrical grid frequency around 50Hz. NGT is interested in monitoring, the condition of the grid frequency, *i.e.* determining the state of the grid system from observations. Estimates of the current state of the system can then be used to inform operational action to regulate grid frequency, a procedure known as *frequency response* [46].

In the context of the Real-timE Modelling of Non-linear Data-Stream (REMIND) project [20], the problem of condition monitoring of the grid frequency dynamics is translated into the problem of estimating parameters from grid frequency observations for a physical model of the system. Those estimates will then be used by grid frequency managers to take operational decisions on grid manipulation. This industrial problem is described in detail in 6.1 and the observational data available is presented in 6.1.1.

Note that in this study of the grid frequency system, a Real Model Scenario (ReMS) is used since the complexity of the system and all relevant degrees of freedom of the electrical grid cannot be completely captured in any physical model [15]. In addition, toy experiments are designed to esti-

mate parameters using a NRA where the simple physical model developed to represent the system is taken as a Perfect Model. In other words, the model and observations produced from forward simulation of the model conformed a PMS.

Mathematical models are the basis of condition monitoring, and a deterministic structural model is developed in section 6.2 where the physical understanding of the grid behaviour is presented as a set of differential equations. The stability of the grid frequency is understood in terms of the energy balance between electrical generation and demand in the grid system.

The structural model is assumed to be a *perfect model* that describes the grid system and condition monitoring (*i.e.* parameter estimation) is attempted by looking for parameter values that best describe the observations. Two different methodologies are used for this purpose, namely, least squares (LS) estimations (section 6.4) and Bayesian parameter estimation (section 6.5).

In practice, any model is only an approximation of the system it aims to describe (model $\neq$ system) then uncertainty in the state estimate is understood to be due to two distinct sources: observational uncertainty and model uncertainty, *i.e.* ReMS.

Before any condition monitoring methodology is used, a Perfect Model Scenario (PMS) is defined from the physical understanding of the system.

section 6.3 defines the PMS for the UK's electricity grid as a numerical integration of the grid frequency model described in section 6.2. PMS high resolution observations are generated by means of a forward simulation of the grid frequency model for the true parameter vales.

Due to operational constraints, one of the most significant features of these data sets is its coarse temporal resolution. Grid frequency is sampled at a rate of 1Hz (every second) and demand state data is obtained *in-situ* at an average sampling time on the order of minutes.

The investigation of the condition monitoring of the grid system from the PMS to real operational conditions is planned in stages. In each stage, the perfect model and the high resolution observations are transformed such that they gradually loose quality and temporal resolution and start to mimic practical conditions. Such operational conditions mean that the parameters are estimated in the PMS using a NRA for the grid from observations with the same sampling characteristics as the ones available for the grid system.

Depending on the methodology used for parameter estimation, Frequentist or Bayesian, transformations for both model and observations applied to several experiments that produced estimations reflecting different characteristic of the system. Sections 6.4 and 6.5 present those toy experiments and theircorresponding performance. In both cases, the first attempt at condition monitoring the grid frequency, is made in the PMS by a re-formulation of the

PMS using a NSA inside this NRA (see 6.5.1). Once consistent parameter estimates are obtained in the PMS, uncertainty is introduced to the parameter estimation process in two ways, to the observations (in section 6.4.1 and 6.4.2) and to changes in the perfect model (in section 6.4.2 and 6.5.1) by introducing terms that account for model error.

section 6.4 presents each of the stages designed to reproduce the real operational conditions in the parameter estimates for traditional *best guess* estimates. Least squares estimates are able to recover the true parameter values for the grid frequency system/model in the PMS for high resolution data.

Uncertainty in the observations is introduced by sub-sampling the demand data from which the perfect model is driven to obtain grid frequencies at 10Hz and sub-sampled at 1Hz. The missing demand values are interpolated using different interpolating schemes whilst keeping the grid frequency sampling rate constant at 1Hz.

In contrast, model uncertainty is introduced by checking the consistency of parameter estimates when some components of the model are changed such as the frequency response function and the integration scheme. After the model changes are performed, high resolution data is generated and then use to estimate the true model parameters.

sections 6.4.1 and 6.4.2 describes each of these experiments separately.

In all cases, "best guess" estimates fail to provide correct values for the true parameters and uncertainty measures of the estimates. There is no measure of the reliability of the methodology when the model is wrong.

All experiments presented in section 6.4 show that the Frequentist approach is unable to provide reliable estimates of the model parameters except when operating in the PMS with high temporal resolution data. In this approach, synthetic demand data is used to generate grid frequency from which the parameter values are tuned. Real demand data is available at operational rates and is insufficient for linear estimations of the parameters.

Bayesian parameter estimation offers an alternative methodology and seeks explicitly to account for uncertainty in the parameter values, observations and model error by taking a probabilistic approach. As presented in section 6.5, instead of producing a point estimate of the parameter value, the method produces a distribution of values that best resemble the data given a particular model.

section 6.5.1 describes one of the major thrusts of this Thesis, the development of a Bayesian parameter estimation implementation using *Markov chain Monte Carlo* techniques, for both the deterministic grid frequency model and its probabilistic counter-part. Given the stochastic (or probabilistic) nature of the approach, the PMS is re-defined in section 6.5.2 as an Euler approximation of the probabilistic version described in 6.5.1, *i.e.* NSA inside

a NRA.

Applying MCMC techniques to the condition monitoring of the grid frequency system made apparent a handful of fundamental difficulties related to the nature of the model and the quality of the data available [22, 20].

Most of the effort in this Chapter was placed on the investigation of methods to apply the MCMC algorithm to real data and to determine whether the lack of convergence of the parameter values were founded in:

1. The MCMC method itself.

2. The fact that the mathematical structure of the model is formulated in an NSA inside an NRA.

3. The quality of the operational data available, fails to provide the information required.

Discussion related to similar issues for chaotic system is also found in Chapter 3.

The MCMC approach works well in the PMS using a NRA when high resolution data is available. section 6.6 presents the attempt to use the MCMC methodology when restricted to realistic sampling rates. In this case, it is shown that MCMC significantly degrades performance even when the model structure is perfect. Despite the fact that estimates obtained from the MCMC algorithm cannot be interpreted in terms of physically plausible

parameter values, it is believed that the impact of these data quality issues may be generally under-appreciated within the MCMC parameter estimation community [22].

Implementing both methodologies (Frequentist and Bayesian) for parameter estimation, for model data sets with matching realistic sampling times, provides new insight into the role of the operational constraints on the parameter estimation. Reference [22] discusses the role of real operational conditions reflected in the quality of the experimental observations available. This discussion highlights additional constraints and uncertainty sources to the process of model parameter estimation of grid frequency dynamics.

In section 6.7, a summary is presented of the main results and issues that arose as a result of this investigation and further work and research related to this study are listed. Principally, to account for uncertainty in both model inadequacy and estimation of model parameters, the failure of MCMC methodologies in real operational circumstances may condition, among others, the quality of future data and the quantisation of model inadequacy by "stochastisation" of the deterministic model. It is important to note that the implementation of Gradient Descent (GD) techniques for parameter estimation as discussed in Chapters 4 and 5 has been left for future study.

To avoid confusion of scenarios and approaches, once the model of the grid frequency is formulated, it is assumed to be the system, the model is the

perfect model. Parameter estimation is attempted then in the ReMS but it is going to be referred as PMS all over this Chapter. When Maximum Likelihood or Bayesian techniques are used to estimate parameters, as explained earlier, the methodology and the ReMS conformed a NRA for the latter and a NSA inside NRA for the latter, and it is not going to be referred to by these terms but by the methodology used.

The project was funded by the NGT through a EPSRC Faraday project administered by the Smith Institute in Oxford.

6.1 The Problem

NGT is tasked with maintaining the integrity of supplied electrical energy to industrial and domestic customers across England and Wales. An important aspect of its operation is ensuring that the grid frequency of the electrical system does not exceed pre-defined bounds, as set out in NGT's transmission license [1]. Specifically this entails that the grid frequency be maintained at 50 ± 0.5 Hz. Grid frequency is itself dynamic, rising and falling with excess generation and excess demand respectively. To maintain a steady grid frequency, grid system operators schedule a proportion of generation to respond to changes in grid frequency. This feature of the grid system is called *frequency response* [46].

To effectively manage the provision of frequency response, operators would benefit greatly from additional knowledge on the state of the grid system. In particular, grid system operators are interested in estimates of the *grid system inertia* and the output of frequency responsive generating sets. The grid inertia characterises how stable the grid system is and therefore how sensitive the grid frequency is to sudden changes in demand.

Deviations in grid frequency arise when generation does not match demand exactly and since demand is continuously changing the grid frequency is rarely constant. Frequency response maintains the grid frequency within its prescribed limits by constantly adjusting generational output.

The provision of frequency response is produced by scheduling generators to run sub-optimally in order that they are able to increase their output if needed. This procedure results in large costs to NGT. Given an accurate picture of the state of the grid, response can be scheduled more efficiently.

The modelling process implies an idealisation of the grid frequency system and the experimental observations obtainable from the system. In this project, the observations available are grid frequency and demand. The characteristics of the data add constraints to the methodologies used to estimate parameters, and are discussed in 6.5.

The data available are extensive sets of grid frequency observations and demand data, and are described in the following section.

6.1.1 The Real Data

The data available is of grid frequency observations with a sampling rate of 1Hz for six months between July and December 2001, and two weeks in September 2004. Grid frequency data is sampled at a rate of 1Hz, *i.e.* one observation per second, with less than 0.5% of missing values and the same percentage of repeated values for a given time. Given the size of the electrical grid in the UK and the stability of the system, the grid frequency can be considered to be the same in all points of the grid [46].

Figure 6.1 shows the trace of grid frequency (upper panel) and demand (lower panel) for a window of 30 minutes on 1^{st} September 2004, between 15:17 to 15:47. The data points plotted in this Figure are adimensional values which represent fractional deviations from the *operational point* of the grid system ω_0, whilst in the case of the demand, values are normalised with respect to the *total load* of the grid system $\mathcal{G}$.

The upper plot shows a trace of 1801 points corresponding to the frequency values measured at each second during that half hour. The values clearly show finite resolution (2×10^{-4}) and magnitude of the variations from the operational point of the order of 10^{-2} approximately.

The lower plot shows a trace of 1290 points measured during half an hour at variable sampling times and where approximately 30% of values are

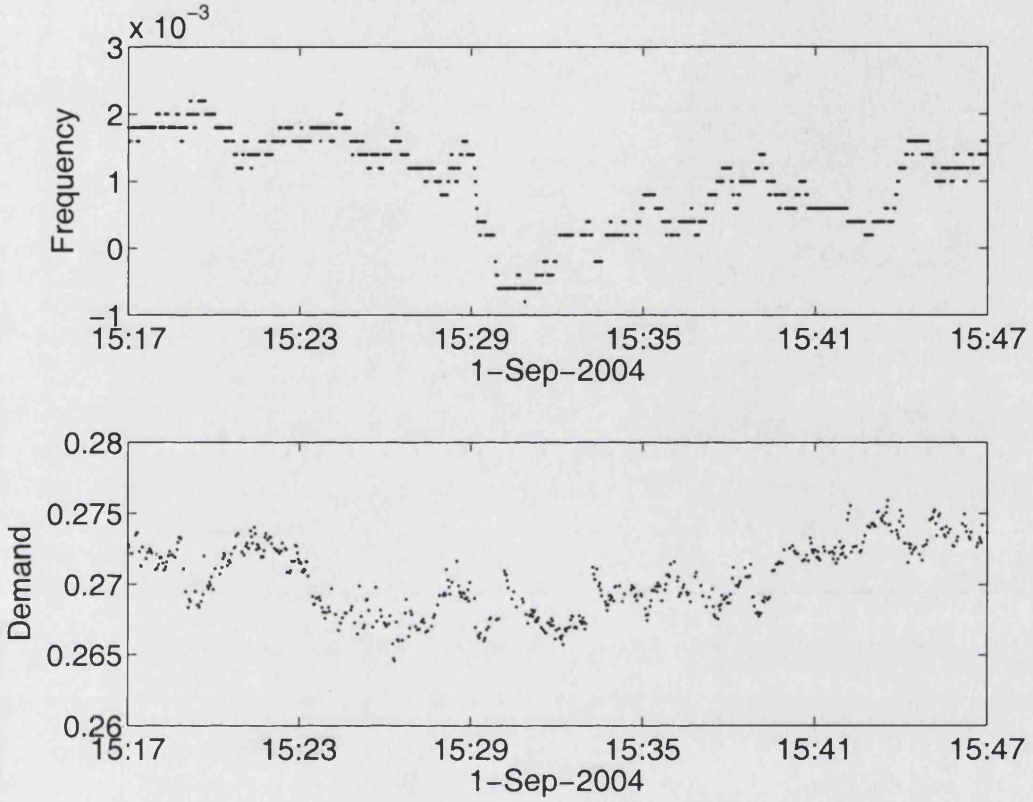


Figure 6.1: Trace of a 30 min. on 1^{st} September 2004 between 15:17 to 15:47. The grid frequency plotted in the upper panel is the fractional deviation from the operational point of the grid system ω_0 and the demand plotted in the lower panel is normalised with respect to the total load of the grid $\mathcal{G}$.

missing. For example, before and after 15:29 evident gaps where observations are missing can be seen. In contrast with the frequency observations, observations for demand available are in principle sampled at 1 minute, corresponding to a sampling rate of 0.0167 Hz approximately, a frequency rate two orders of magnitude smaller that the grid frequency sampling rate. For

the half an hour showed in Figure 6.1, the average sampling time is approximately 2 seconds.

Note that demand observations are not an instantaneous measurement of electrical demand at a given time but instead are an aggregated colsection from a number of measurements of electrical generation loaded into the grid system at each electrical substation. Demand is estimated from these measurements and "demand observations" are obtained by adding up collected values from all electrical substations to estimate demand for a particular moment in time. Sampling times in the real demand data set are a refsection of the time spent in the process of collecting measurements all over the electrical substations, and the interval of time between successive experimental observations are recorded.

The fact that grid frequency and demand observations are sampled at different rates increases the uncertainty of the grid system state at times when measurements of variables is not available. In addition, this difference in experimental information will unavoidably affect eventual parameter estimates.

Figure 6.2 shows typical histograms of the sampling time, Δt in seconds, of one week of demand observations during September 2004. The top panel shows the Δt histogram at which demand observations were collected during

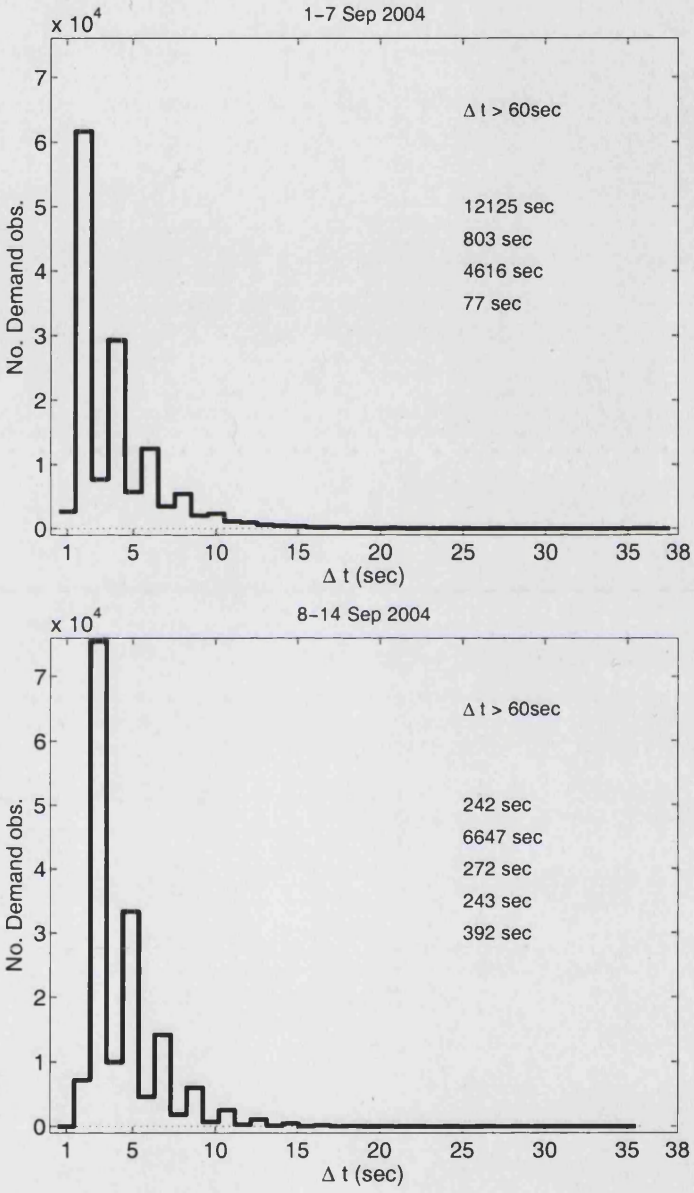


Figure 6.2: Distribution of sampling times for a window of 30 min. of demand observations. Top panel: 1 Sep. to 7 Sep. Bottom panel: 8 Sep. to 14 Sep.

the week between 1^{st} to 7^{th} September 2004, and the bottom panel shows the corresponding histogram for the week between 8^{th} to 14^{th} September 2004.

The sampling times vary in both cases from 1 to 15 seconds for approximately the 90% of the observations taken, corresponding to sampling rates varying from 1Hz to 0.067Hz. In both weeks, events were found where the time between two successive observations is larger than 1 minute. These events are listed in each panel where one of the events is a period of more than 3 hours with no demand observations (see top panel). Even though these events are rare, their cost is very high in terms of uncertainty of the grid system state at those times.

6.2 Grid Frequency Dynamics: Structural Model

The model of grid frequency is based on the generation and demand balance and the changes in kinetic energy that result from deviations of that balance. The constraints on grid frequency prescribed by NGT on the generation are then formulated within this framework.

A structural model is first developed, and expressed as a set of differential equations, to describe the generating mechanisms for various components of the grid system dynamics. Detailed formulation of the model can be found in [20].

The grid frequency model is composed by two parts. The first part describes the grid frequency dynamics - how grid frequency changes as the dif-

ference between generation and demand changes. The second part describes *frequency responsive generation.*

The dynamics of grid frequency is understood in terms of energy changes. Changes in grid frequency are driven by the changes in the kinetic energy of the rotors of the grid system's generators. All generators of the same type are seen as one generator. Therefore, the kinetic energy of the rotor is written as follows

$$E = \frac{1}{2} I\tilde{\omega}^2, \tag{6.1}$$

where I is the moment of inertia of the grid system and $\tilde{\omega}$ is the angular momentum of the generator's rotor. Excess generation increases the kinetic energy and correspondingly excess in demand reduces the kinetic energy. Hence changes in kinetic energy reflect the mismatch between generation and demand.

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{2} I\tilde{\omega}^2\right] = \frac{\mathrm{d}E(t)}{\mathrm{d}t} = \tilde{G}(t) - \tilde{D}(t), \tag{6.2}$$

where $\tilde{G}(t)$ and $\tilde{D}(t)$ are instantaneous generation and demand, respectively. Replacing equation (6.1) in (6.2) and assuming the moment of inertia I is a constant over time then

$$\frac{\mathrm{d}\tilde{\omega}}{\mathrm{d}t} = \frac{\tilde{G} - \tilde{D}}{I\tilde{\omega}}. \tag{6.3}$$

The moment of inertia, I, of the system's rotors is not observed and

must therefore be expressed through other physical quantities. Let H be the relaxation time taken for the grid frequency to be zero given no demand and a sudden lost of generation, written as

$$H = \frac{1}{2} I \tilde{\omega}^2 \mu^{-1}, \tag{6.4}$$

where μ is the mega-volt Ampere (MVA) of the grid frequency, with physical dimensions of the inverse of electrical power which is a well known constant for the UK's grid. Normally, the MVA is expressed as a constant per each type of generator.

Solving (6.4) for I and replacing it into (6.3), it follows that:

$$\frac{\mathrm{d}\tilde{\omega}}{\mathrm{dt}} = \frac{\tilde{\omega}(t)}{2H\mu} \left(\tilde{G}(t) - \tilde{D}(t) \right), \tag{6.5}$$

For practical reasons in the rest of this chapter H absorbs the constants in (6.5) containing all the knowledge of the inertia of the grid system.

In practice grid frequency, demand and generation deviate around the corresponding operational point in the following way

$$\tilde{\omega}(t) = \omega_0(1 + \omega(t)), \tag{6.6}$$

$$\tilde{G}(t) = \mathcal{G}(1 + G(t)), \tag{6.7}$$

$$\tilde{D}(t) = \mathcal{G}(1 + D(t)). \tag{6.8}$$

where $\omega_0 = 2\pi(50\text{Hz})$ is the operational point for grid frequency, and $\mathcal{G}$ is the operational point for demand and generation, called *nominal load.* For

the UK's grid system. the nominal load is around 40 Giga Watts at peak demand.

Substituting these terms into equation (6.5), gives

$$\frac{\mathrm{d}\omega}{\mathrm{dt}} = \frac{\omega_0 \mathcal{G}(1+\omega(t))}{H}(G-D), \tag{6.9}$$

where variables ω, G and D describe deviations from nominal levels. The changes in grid frequency are relatively small therefore second order terms on ω are neglected, *e.g.* $\omega_0\omega(t) \sim 0.2 \Rightarrow \omega(t) \sim 4 \times 10^{-4}$. Thus equation (6.9) is reduced to

$$\frac{\mathrm{d}\omega}{\mathrm{dt}} = \frac{\omega_0 \mathcal{G}}{H}(G-D). \tag{6.10}$$

Equation(6.10) constitutes the simplest form of the grid frequency model.

To maintain the grid frequency stability, generation must be continually matched to demand. In order to respond to instantaneous changes in frequency, some generators are prepared to respond to those changes, producing more or less energy. When grid frequency drops below the target frequency, frequency responsive generation increases. Conversely, generation is paired back when the grid frequency exceeds its target. For simplicity, the target frequency is taken to be constant and equal to ω_0. In terms of the normalised variables of equations (6.6) to (6.8), the target frequency is zero. When grid frequency is balanced, normalised generation and demand are zero.

The demand is modelled as composed by two components, one *resistive demand* and one *inductive demand.* Resistive demand relates to the electrical power delivered and used by consumers whilst inductive demand relates to industrial and generating motors that "feedback" and consume electrical power [46]. Details on demand modelling can be found in the REMIND project report [20].

Generation is modelled by type of generator, *i.e.* coal, gas, nuclear etc. All individual generators of the same type are grouped in one set and the set is seen as one generator.

Given that the model of equation (6.10) is written in normalised variables, only frequency response happens; model variables are different from zero. Henceforth, frequency response has to be included in the model. Frequency response is mostly scheduled on thermal generating plants, *e.g.* coal and gas fuelled generation [46]. Consequently, the model for frequency responsive generation is based on an understanding of the operation of *thermal generating sets.* Please refer to the technical report in [20] for details.

The power output from any thermal generating set is understood to be due to three components: the *internal energy*, $p(t)$, the *power source*, $Q(t)$ and *stored energy*, $r(t)$. These three components are referred in this document as the *internal variables* of the grid frequency dynamics model.

Assuming there are I_g generator types, for the i^{th} generating set, the

deviation in power output into the grid system is

$$P_i(t) = F_i(\omega(t); \alpha_i) + \beta_i p_i(t), \tag{6.11}$$

where $F_i(\omega(t); \alpha_i)$ describes the frequency scheduled power as a function of the deviations from the operational point, *i.e.* how generation responds to changes in grid frequency. β_i is a generation type specific parameter that measures the response of the delivered power to changes in the internal energy.

Note that $F_i(\omega(t); \alpha_i)$ is also called the *frequency response function* since it models the operation of frequency responsive generators. The frequency response function $F(\omega; \alpha)$ is the nexus between the grid frequency dynamics and the dynamics of the internal variables of the model. It links changes in grid frequency to changes in generation. The parameter α determines how fast the generating set responds to small changes in grid frequency.

Changes in the internal energy, $p_i(t)$, are due to a combination of factors: energy input to the generating set from the energy source *i.e.* fuel, energy drawn off the generating set by the grid system, and the transfer of stored energy. The change of internal energy is written as

$$\frac{\mathrm{d}p_i(t)}{\mathrm{dt}} = \kappa_i \left\{ Q_i(t) - P_i(t) + \sigma_i \left(r_i(t), p_i(t) \right) \right\}, \tag{6.12}$$

where κ_i is the internal energy parameter, $Q_i(t)$ is the power source, $\sigma_i(r_i(t), p_i(t))$ is the variable heat transfer function for the stored energy $r_i(t)$ and $P_i(t)$ is

the deviation in power delivered to (or drawn off by) the grid system.

The release of stored energy is modelled as a switch function

$$\sigma_i(X,Y) = s_i(X-Y)\left\{\frac{1-\rho}{2\tanh s_i(X-Y)}+\frac{1+\rho}{2}\right\}, \qquad (6.13)$$

where s_i is the release rate of stored energy and ρ is the energy storage transfer as a fraction of release time. When the internal energy drops, energy is quickly released from the stored energy. Conversely, energy is slowly transferred from the internal energy to the stored energy.

The rate of change of the power source aims to capture the plant's response to short term changes in the internal energy of the system and is given by

$$\frac{\mathrm{d}Q_i(t)}{\mathrm{dt}} = -\frac{1}{\tau_i}(\eta_i\, p_i(t) + Q_i(t)), \qquad (6.14)$$

where τ_i is a time constant that describes how quickly fuel can be added to the generating set and η_i describes the sensitivity of the controlling mechanism to changes in the internal pressure.

The rate of change of stored energy is modelled as

$$c_i\,\frac{\mathrm{d}r_i(t)}{\mathrm{dt}} = -\sigma_i\,(r_i(t), p_i(t)), \qquad (6.15)$$

where c_i is the capacity time constant of the energy store.

The power output to the grid by I_g, types of generating sets is expressed

as

$$G(t) = \sum_{i=1}^{I_g} \lambda_i P_i(t), \tag{6.16}$$

where λ_i is the proportion of generation attributable to the i^{th} generating set. In addition, it is assumed that the conversion process is 100% efficient.

Note that only generating sets which are frequency responsive will contribute to equation (6.16). Non-frequency responsive sets do not contribute to changes in deviations of the generation from its operational point.

These equations represent a basic understanding of the physical processes that constitute grid frequency dynamics. At the same time, the adoption of this model for the characterisation of the grid frequency system does not imply that this description is complete; there are processes and aspects of the system not described by these equations.

Figure 6.3 shows a diagram of the relations between the grid frequency and the internal variables of the model, *i.e.* equations (6.10 to (6.16)), for a given time t, for a grid system composed by two generating sets, one frequency responsive and one non-frequency responsive.

In the Figure 6.3, arrows pointing to a node, *i.e.* circles, squares and rounded squares, come from variables on which the node variable depends explicitly. Auto-dependencies are not included in the sketch. If auto-dependencies where included it would be represented by a loop leaving and arriving in the

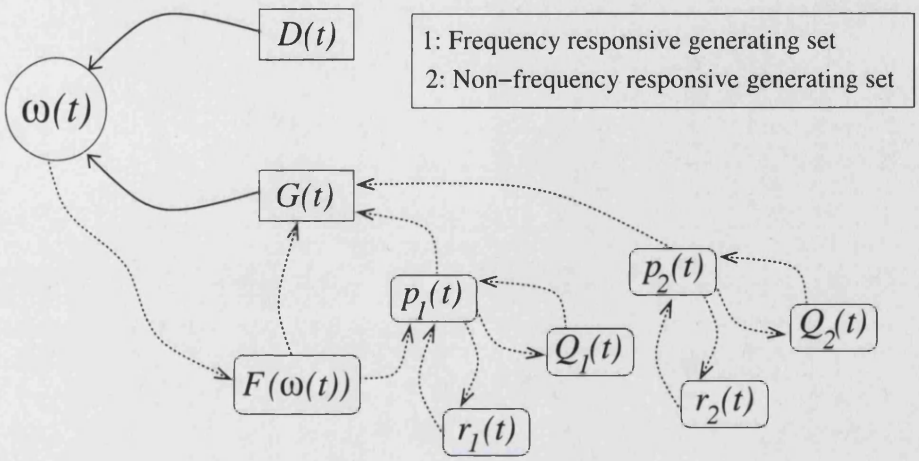


Figure 6.3: The diagram shows the dependency relations between the grid frequency and the internal variables of the model. Arrows pointing to a node, *i.e.* circles, squares or rounded squares, come from variables on which the node depends explicitly.

same node. Despite the fact that the frequency response function, $F(\omega(t);\alpha)$, is not a variable of the model, it is included to emphasise that it is the dynamical link between internal variables and grid frequency dynamics.

To be consistent with the notation used throughout this Thesis, let $\boldsymbol{\theta}$ be the parameter vector which is composed by all parameters of the grid system model. Thus for the model given by equations (6.10) to (6.16), the parameter vector is

$$\boldsymbol{\theta} = \left(\{H_i, \lambda_i, \alpha_i, \kappa_i, \beta_i, \tau_i, \eta_i, s_i, c_i\}_{i=1}^{I_g}, \rho \right), \tag{6.17}$$

henceforth $\boldsymbol{\theta} \in \mathbb{R}^{9I_g+1}$ for I_g types generating sets.

6.3 ReMS: Forward Simulation

A study of the identifiability of the model parameters $\boldsymbol{\theta}$ and the effects of data resolution is performed before carrying out an extensive program of Bayesian parameter estimation for the model using real data.

The study is done via a simulation approach, *i.e.* *forward simulation* of the structural model. In the forward simulation, the physical model of the grid frequency dynamics described in section 6.2 is taken to be the grid frequency system, *i.e.* theReMS is adopted (see Chapter 1).

The modelling process is made from considerations of continuous processes representing the understanding of the dynamics of the grid. Instantaneous changes in demand produce instantaneous changes in generation and in turn are reflected in instantaneous changes in grid frequency. Unfortunately, it is impossible to solve the grid frequency model analytically, therefore it has to be discretised when a numerical solution is sought. Analytical constraints makes it compulsory to use a version in discrete time from which the grid system is described. This discretised model is the one to be considered as the system itself.

Figure 6.4 shows a sketch for the ReMS forward simulation. Given that the model and the forward simulation are identified with the system and the state space of the system, respectively, the ReMS is taken to be a PMS only

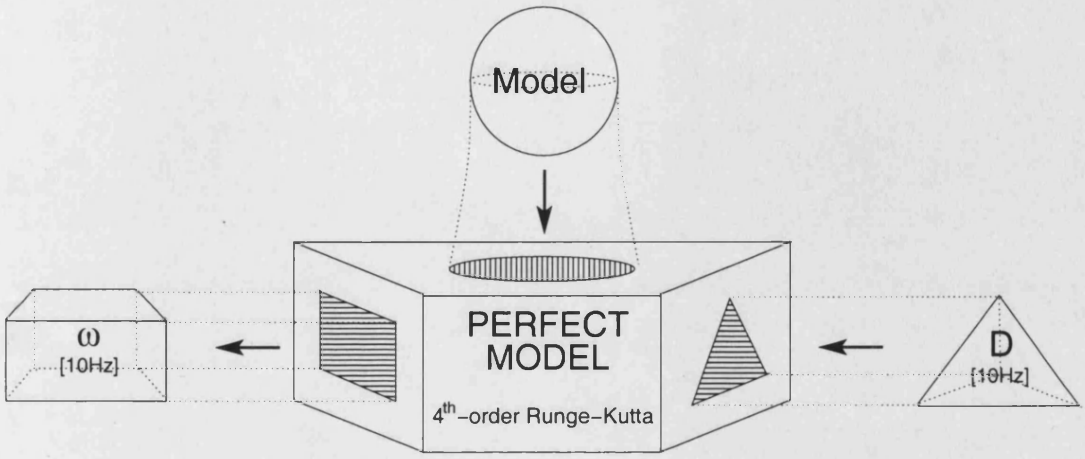


Figure 6.4: Sketch of the forward simulation for the grid frequency model. The synthetic demand data (pyramid in the right) drives the numerical integration of the model to obtained grid frequency (box in the left).

in this context thus a NRA is adopted. Whilst there is not any reference or comparison of any type with real observations, it can be said that model parameter estimation is attempted in the PMS.

Observations of the system are then generated by numerical integration of the model for a set of known parameter values, $\tilde{\boldsymbol{\theta}}$, and synthetic demand data. This process of generating a "true" trace of the grid system using the demand data and the model, is the forward simulation of the perfect model for the grid frequency dynamics. From the forward simulation, the task will then be to estimate the value of $\boldsymbol{\theta}$, the true parameter vector.

In detail, the system is composed by two generating sets, one frequency responsive and one not frequency responsive, representing the simplest con-

figuration of the grid system. Figure 6.3 shows the dependency of the internal variables, generation, demand and grid frequency for this configuration of the grid system.

The perfect model for the grid frequency is defined to be the 4^{th}-order Runge-Kutta approximation of the set of differential equations given by (6.10) to (6.16). The model is solved using an integration step of $h = 0.1$sec and is driven with synthetic demand at the same sampling rate of 10Hz (for details on how demand data is generated see REMIND report [20]) producing frequency data with a sampling rate of 10Hz (see Figure 6.4 for a schematic representation of the forward simulation).

Figure 6.5 shows typical traces of high resolution data, *i.e.* 10Hz sampling rate, of grid frequency in the upper panel and demand in the lower panel for a time interval of 30 minutes. High resolution in time translates to 1.8×10^4 data points of grid frequency and demand.

The true parameter values are shown in Table 6.1 for the simplest configuration of the grid model. Even though the non-frequency responsive generator set do not contribute to changes of grid frequency, corresponding parameter values are still listed.

From Table 6.1 there are only two generating sets, one frequency responsive and one not. Note that from equation (6.16) the parameter α describes how responsive is a set to changes in grid frequency, if there are non-frequency

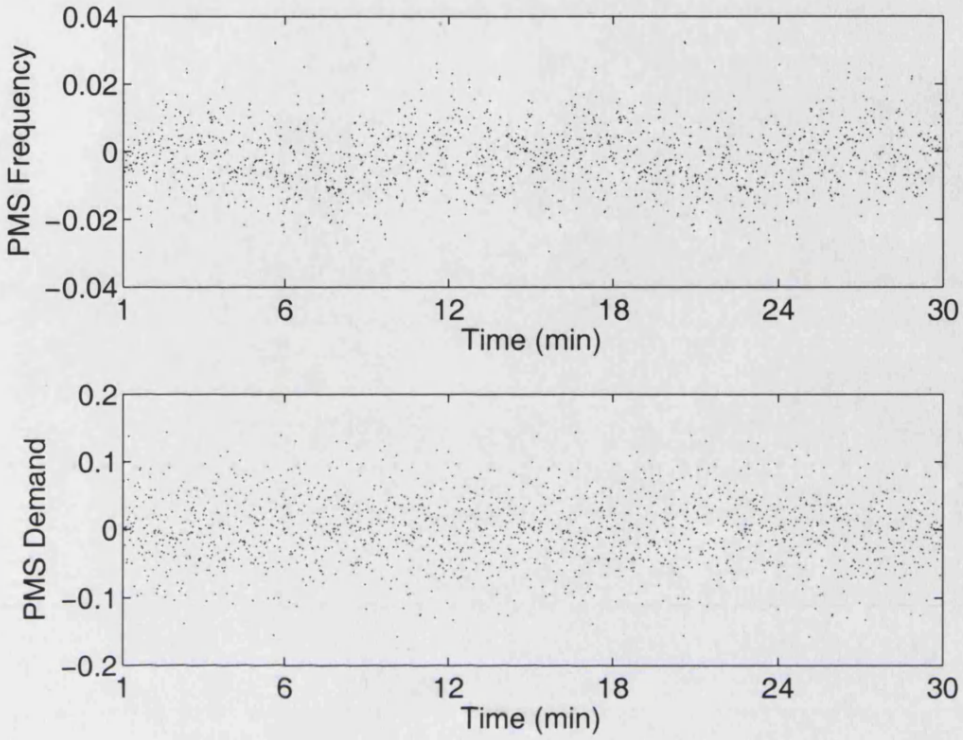


Figure 6.5: PMS high resolution grid frequency and demand observations sampled at 10Hz

responsive generating sets then grid frequency cannot be balanced *i.e.* it is always zero. The frequency responsive generating set produces 10% of the generation and is responsible for the changes in frequency when the mismatch between generation and demand is different from zero.

6.4 PMS Experiments

A study of the identifiability of the model parameters $\boldsymbol{\theta}$and the effects of data resolution in Least Squares parameter estimates in the PMS is performed

Generator type	Parameters									
i	τ	η	κ	β	c	s	ρ	λ	$\boldsymbol{\alpha}$	H
1	1	0	1.00	0.00	100	0	0.00	0.90	**0**	0.0373
2	60	50	1.30	0.80	100	100	0.01	0.10	**25**	0.0373

Table 6.1: Grid frequency model true parameter values that generate high resolution data in the PMS. Column in bold face highlights the parameter that determines which of the set is frequency responsive, *i.e.* $\boldsymbol{\alpha} \neq 0$.

before carrying out an extensive program of Bayesian parameter estimation. Each experiment explores different issues concerning model structure and temporal resolution of the available PMS data.

The issues explored are:

1. *Experiment 1: Data*

   section 6.4.1 presents this experiment where high resolution data is sub-sampled in a way that gradually starts mimicking real operational conditions. In one hand, section 6.4.1.1 studies the effect of lower sampling rates in grid frequency and demand. Whilst in the other, section 6.4.1.2 studies the effect when the PMS data has the same sampling resolution as the real data described in section 6.1.1.

2. *Experiment 2: Model*

section 6.4.2 presents how LS parameter estimates are affected by slight changes in the model class that represents the grid system in equations (6.10) to (6.16). In particular, sensitivity of estimates to choice of frequency response function is studied in section 6.4.2.1 and to the integration scheme used to solve the set of differential equations that comprise the physical understanding of the grid system in section 6.4.2.2.

Given that this part of the Thesis was developed inside the REMIND project, LS estimates in the PMS and the experiments listed below, results are qualitatively commented upon to justify the final Bayesian implementation. Details can be found in the REMIND report [20].

All model parameters (except parameters c and ρ) of the frequency responsive generation type are chosen to be estimated by the LS methodology (see Table 6.1 for a list of parameters).

Table 6.2 shows that LS parameter estimation in the PMS leads to the correct results as well and there is no uncertainty in the point estimates. True parameter values are recovered with no uncertainty.

Unfortunately, a perfect model does not exist in most practical cases and it is important to remember that these perfect conditions do not reflect real conditions of the grid frequency system. This first estimation of parameters is a safety check that in the PMS conditions things to work. To adopt a

Parameter	τ	η	κ	β	ρ	λ	α	H
True Value	60	50	1.30	0.80	100	0.10	25	0.0373
Estimated Value	60	50	1.30	0.80	100	0.10	25	0.0373
RMS/σ_ω^2	0.00							

Table 6.2: Table of fitted parameter values for experiment in the PMS showing: parameter, true value, estimated value and the fitting error

consistent NSA in which to use Bayesian techniques in this model of the grid system that later on can again be translated to an even more realistic NRA to estimate useful parameters values from real data. The following experiments explore sensible issues related to data and model.

6.4.1 Experiment 1: Data

This section enumerates the experiments and results for the LS methodology and its shortcomings as a motivation to implement alternative techniques of condition monitoring in real operational circumstances. Further details and plots can be found in [20].

6.4.1.1 Sub-sampled Frequency and Demand

Grid frequency observations are generated in the PMS at a rate of 10Hz and sub-sampled at 1Hz (as is the real data). The perfect model is then driven with the synthetic demand data sub-sampled at several rates and the missing values were completed by means of a number of interpolation schemes. These three experiments produced three different grid frequency traces which are then contrasted with the PMS observations in order to tune model parameter values. The experiments are:

1. Demand data sub-sampled at 1Hz (every second). The sampling rate is 10Hz filling the gaps with a piecewise constant function.

2. Demand data sub-sampled every 15 seconds (0.067Hz, 6 times the average of the one in real demand data). The sampling rate of 10Hz is achieved by using a piecewise constant interpolation.

3. Demand data sub-sampled at 1Hz (every second) and the missing values are linearly interpolated to obtain 10Hz sampling rate.

In all cases, the correct values of the parameter vectore are not identified, even though the smallest errors are obtained for the case where the missing values of demand are linearly interpolated.

6.4.1.2 Real operational Conditions

High resolution synthetic demand data is degraded to be sampled are in a typical 30 minute window. In other words, the real demand data sampling times distributions, as in Figure 6.2, is matched to the high resolution demand data. In order to obtain a demand data with 10Hz resolution to drive the estimating model, the missing values are set to perturbed values linearly interpolated between consecutive missing demand values. The size of the perturbations is defined to be half the magnitude of the difference between the existing demand values.

The failure in LS to provide correct estimates is due to the fact that the observations are not sampled at a sufficiently high rate even though the estimating model itself is perfect. Independently of the length of the observations, the sampling rate is not high enough to get the relevant dynamical information from the system.

6.4.2 Experiment 2: Perfect Model

This experiment is motivated by the intention to implement MCMC techniques to the grid frequency model. MCMC implementation imposes some constraints on the analytical forms used to model random variables, as detailed in section 6.5.1. Two of the strongest changes in the explicit form of

the model are related to the modelling of the frequency response function and the integration scheme.

6.4.2.1 Frequency Response Function

The observed grid frequency is generated using a 4^{th}-order Runge-Kutta integration scheme and the frequency response function is taken to be

$$F(\omega;\alpha) \;=\; -\tanh(\omega\alpha). \tag{6.18}$$

The Taylor expansion around zero of (6.18) is

$$F(\omega;\alpha) \cong -\alpha\omega + \frac{\alpha^3}{3}\omega^3 - \frac{2\alpha^5}{15}\omega^5 + \ldots. \tag{6.19}$$

Thus, given that the estimating model takes the frequency response function as a first order approximation; the estimating frequency response function is $F(\omega;\alpha) = -\alpha\omega$.

It is found that the correct parameter values of $\boldsymbol{\theta}$ are obtained when the high resolution data is used, and there is a strong indication that for $\alpha = 25$ the frequency response function is indistinguishable from equation (6.18), given that locally both the estimating model and the perfect model behave like $-\alpha\omega$.

The values that the model variable takes in both cases are not the same since there is some fitting error as shown in Table 6.4.2.1. The method does

however pick out the "correct" parameter values.

Parameter	τ	η	κ	β	Rel	λ	α	H
True Value	60	50	1.30	0.80	100	0.10	25	0.0373
Estimated Value	60	50	1.30	0.80	100	0.10	25	0.0373
RMS/σ_ω^2	9.40×10^{-4}							

Table 6.3: Table of fitted parameter values for Experiment 2 using a first order approximation of the frequency response function $F(\omega; \alpha)$

The relevance of this scenario is clear when the Bayesian methodology is applied. In particular, this approximation is convenient for MCMC implementation as it simplifies some analytical calculations.

6.4.2.2 Integration Scheme

The effect of the integration scheme used on the LS estimatesis explored, instead of using a 4^{th}-order Runge-Kutta integration scheme, an Euler approximation is used with same integration step used in the PMS, *i.e.* $h = 0.1$ sec. The estimating model is now a first order approximation of the differential equations that generated the data.

The observations used to estimate parameters of the Euler approximation

of the model are:

1. Observations of the grid frequency system are traces of ω from the 4^{th}-order Runge-Kutta integration of the model for grid frequency driven with demand data at 10Hz.

2. Demand data at 10Hz is used to drive the estimating model to generate grid frequency at 10 Hz. Grid frequency is then sub-sampled at 1Hz (real sampling rate) and used to tune parameter values for the PMS observations.

The main finding is that the change of the integration scheme impacts the results of the parameter fitting. The parameters τ, η and κ are the most affected as by the change of integration scheme the strength of the relation with internal variables of the model is directly reduced.

Some information is lost in the temporal evolution of the internal variables when the approximation of the system is made linear. In practice, the true parameter values are unknown and the resulting Least Squares estimates correspond to optimal parameter values in the Maximum Likelihood sense, *i.e.* those where the minimum of the LS cost function is located. Therefore, grid frequency system parameter estimates are biased.

As a result the LS method is unable to provide reliable estimates of the model parameters with corresponding uncertainty measures when the inte-

gration scheme is coarse.

From the results of the two experiments performed, the most dramatic effects were observed when the perfect model chosen to represent the grid frequency dynamics is approximated at first order and when the temporal resolution of the data is like the real one.

Neglecting the fact that there exists a significant mismatch between the estimated values for the parameters and the model parameter true values, the following Bayesian formulation of parameter estimations for such model is made using a first order discretisation of equation (6.10) and PMS data is used many times such that it gradually resembles real operational conditions. The problem of parameter estimation is clearly an NRA. Relevance of the approach in this context, is justified in terms of the insight such implementation can provide in a real application, as for the UK electricity grid. The main results of this attempt are presented in the next section and in [22].

6.5 Bayesian parameter Estimation for the UK's Grid System

The Bayesian methodology seeks to account for uncertainty in the parameter values and observations by taking, essentially, a probabilistic approach. In-

stead of identifying the "correct" parameter values, probability distributions of parameters are sought that are consistent with the data and the model structure.

The model is understood to encode the understanding of the grid dynamics; the equations describe the relationships between the variables. The most 'probable' parameter values of the model given the observations are then sought. The parameters are estimated in an iterative fashion. One starts from some initial understanding of the parameter values, encoded as a prior distribution, and the methodology iterates this distribution to the resulting posterior distribution.

The prior distribution can be uninformative, with no particular value favoured. In addition the method can be used to estimate the distribution on uncertain observables or non observed variables (*i.e.* latent variables). Consequently, the method has the potential to produce a distribution on the demand variable.

There are a number of differences between LS and Bayesian methodologies that it is important to make clear. Unlike the simulation approach (LS estimation), the Bayesian methodology uses information in the grid frequency observations to estimate the parameters due to the explicit dependency on the model structure and the observation from the system.

Figure 6.6 shows a sketch of the Bayesian approach as it is implemented

in this work. The model parameter distributions are conditioned on the observed demand data and the observed grid frequency.

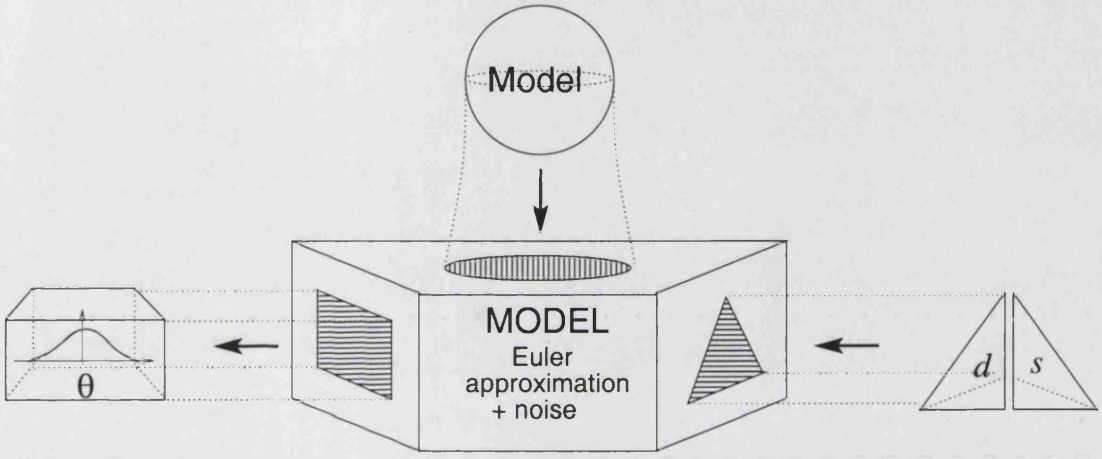


Figure 6.6: Sketch of the Bayesian implementation for the grid frequency model. Observations of demand and grid frequency are denoted by d and s respectively.

Comparing Figure 6.4 with Figure 6.6, Bayesian parameter estimation methodology does not use the perfect model but a realistic and "affordable" version of the model in order to make numerical implementation feasible and also include model error.

In addition, the Bayesian model for the grid frequency system needs both demand and grid frequency observations, denoted by d and s, respectively.

For parameter estimates it provides an estimate of each component in the parameter vector $\boldsymbol{\theta}$, *i.e.* the "best" o most probable parameter value given the model and the observations. Each component estimate is the form of an empirical distribution formed by MCMC samples.

As explained in detail in section 6.5.1, in the Bayesian framework the parameter vector does not only contain the model parameters of the physical encoding for the grid frequency dynamics in equation (6.17) but also all the non observables variables. Moreover, it also contains hyper-parameters associated with uncertainty of the components of $\boldsymbol{\theta}$.

The Bayesian methodology assumes that all observations, as well as any parameter or model variable not observed have associated uncertainties. This assumption enables the methodology to model the demand as a distribution. Given observations of the demand, the Bayesian methodology is then able to estimate unobserved values of demand, while at the same time estimating the uncertainty on those observations.

Bayesian parameter estimation proposes a coherent methodology within which model parameters, and the uncertainty in those estimates, can be determined despite the naive assumptions made in order to implement it numerically. In this work, the method is applied to a model for a real world physical system, namely the UK's electrical grid system, and investigates the robustness of the method to operational data constraints.

Presenting parameter estimates as a distribution allows us to quantify, to some extent, how good the model is; where tight posterior distributions on the estimated model parameters provide reassurance that the model is describing the observed behaviour.

Given high quality data and a model that is mathematically correct, estimations using the Bayesian technique in the PMS are found consistent with the known values as described in section 6.5.1.

section 6.6 shows that when the method is applied to operational data - using observations from the UK's electrical grid, the resulting posterior distributions are uninformative. Moreover, it is found that similar ambiguities are observed in experiments where a perfect model is available but data is sampled at operational sampling rates. In short, there is insufficient information in the available data to be able to identify the parameters in the context of the real system.

One of the most interesting aspects of this work is that it points out the need to meld physical and probabilistic modelling techniques to be applied in real applications where uncertainty is ironically a source of information.

In addition, it motivates to infuse more attention to the areas of research where the deterministic and stochastic approaches are in contact. Chapter 3 also discussed this aspect when a nonlinear models are involved.

The state of the grid, like many complex physical systems, is not directly observable and must be inferred from understanding of the system and the information available from observations. Given a model structure with undetermined parameters and observations of that system, the task is to seek those parameter values that "best" describe the observed behaviour. Here

"best" is understood to be application specific.

For NGT the "best" parameter values are those that result in a more effective scheduling of frequency response. Implicitly, parameter estimation is often carried out in the perfect model scenario, whereby it proceeds as if the system is drawn from the model class being used [19]. The result is that a correct, or optimal, set of parameters is assumed to exist and is actively sought. When the model is known to be an approximation of the system, the true parameter values are known not to exist. In an effort to address this fact it is sought to quantify the uncertainty in the parameter estimates.

The method is illustrated by walking through the application of a Bayesian techniques to the simplified grid frequency model described in section 6.2. The version adopted does not include the internal variables, as seen in Figure 6.3, the left lower part of the diagram shows an island of variables interrelated but only connected to the grid frequency ω by the frequency response function through the generation.

As a first attempt to implement the Bayesian methodology, the probabilistic model does not see the contribution of the internal variables as such but only through the high resolution data generated in the PMS.

This continuous model is then made discrete, dynamical noise is included in order that Bayesian parameter estimation produces a way to monitor model error. The melding of these two modelling approaches results in a

stochastic dynamical model.

The advantages of such an approach are two fold. Structural models, based on a physical understanding of the system, allow for the explicit inclusion of prior knowledge, *i.e.* background information. Statistical models, on the other hand, offer a more direct treatment of noise and the uncertainties inherent in the modelling process.

It is difficult, however, to incorporate domain knowledge when the model parameters have no intrinsic physical meaning. This is often the case when describing observed phenomena using standard stochastic processes such as auto-regressive models.

Here, the physical knowledge encapsulated in a structural model is fused with uncertain information from observations, whilst at the same time maintaining the uncertainty due to model impersection and inadequacy. The result is a probabilistic model with the advantage of probabilistic reasoning and the inclusion of domain specific information.

The model of equation (6.10) is a deterministic expression of the grid frequency dynamics. In the Bayesian framework the dynamics is modelled as a stochastic process and as such equation (6.10) must be formulated probabilistically. To do this, it is necessary to specify a stochastic term Γ to

give

$$\frac{\mathrm{d}\omega}{\mathrm{dt}} = \frac{\omega_0 \mathcal{G}}{H}(G - D) + \Gamma, \tag{6.20}$$

where Γ is a Weiner process [38]. The inclusion of this term makes the application of the MCMC techniques more tractable and hopefully makes it more useful in terms of uncertainty accounting, *i.e.* model inadequacy.

The model for the frequency responsive generation is left deterministic but in future work could also include further stochastic terms. As pointed out earlier, from the MCMC algorithmic representation of the parameter estimation process, the internal variables are invisible.

The following sections describe how Bayesian parameter estimation is applied to the grid frequency dynamics model of equation (6.20). Observables S and, non-observable, $\boldsymbol{\theta}$ quantities are identified based in the background information, I, and available and appropriate prior distributions are specified.

The available observable quantities are the grid frequency and demand data, s and d, respectively. The set of observations corresponding to these model variables is therefore

$$S = \{\{s_t\}_{j=1}^{N}, \{d_t\}_{t=1}^{N}\}, \tag{6.21}$$

where s_t and d_t correspond to observations of the model grid variables: grid frequency ω_t and demand D_t respectively. The observations of the grid fre-

quency, s_t, are not assumed to be the grid frequency variable, ω_t, themselves and are modelled as a sample from a normal distribution with mean ω_t and variance σ_s^2.

Similarly, observations d_t of the demand variable D_t are modelled as samples from a normal distribution $\mathcal{N}(D_t, \sigma_d^2)$. As such σ_s and σ_d represent the amplitude of the observational noise present in each data set.

Lets see how the inclusion of stochastic terms in (6.20) is revised step by step. Initially, a simple grid frequency model is considered in equation (6.10), reproduced as follows

$$\frac{\mathrm{d}\omega}{\mathrm{dt}} = \frac{\omega_0 \mathcal{G}}{\sum_i H_i \lambda_i}(G - D). \tag{6.22}$$

The model described in section 6.3 only contains two generating sets therefore equation (6.22) can be reduced further. In general, the parameter related with the proportion of generation λ_i for $i = 1, 2$ is constrained to the condition $\lambda_1 + \lambda_2 = 1$, *i.e.* the proportion of generation types in the grid generate all the electrical power as in equation (6.16). Introducing that constraint to the equation (6.22), it is reduced to

$$\frac{\mathrm{d}\omega}{\mathrm{dt}} = \frac{1}{H}(G - D), \tag{6.23}$$

assuming the inertia H of each generating set is the same for both types of generators, and has absorbed the constants ω_0 and $\mathcal{G}$.

Given the discrete nature of the observations, (6.23) is integrated using the Euler approximation

$$\omega_{t+1} = \omega_t + \frac{h}{H}(G_t - D_t), \tag{6.24}$$

for $h = 0.1$ sec, and higher order approximations results in posterior distribution with a more complex analytical form. Higher order approximations of equation (6.23) has not been used with MCMC techniques yet.

The continuous parametric model for grid frequency, described in section 6.2, is made discrete in time, through the adoption of an Euler integration scheme, in order to apply the Bayesian parameter estimation algorithm giving

$$\omega_{t+1} = \omega_t + \frac{h}{H}(G_t - D_t) + \gamma_t, \tag{6.25}$$

where t is the time index, h is the integration step and γ_t is an independent and identically distributed random variable with mean zero and variance σ_γ^2. The dynamical noise term described in equation (6.20) and (6.25) can be seen as well as accounting for uncertainty in the order of the approximation, *i.e.* model error.

It remains to specify prior distributions for each of the unknown quantities. For constrained components of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_i$, an informative prior reflects the prior knowledge in the form of a probability density or a constant value. Whilst for unconstrained components of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_k$, a non-informative prior is set

with the hope that the uncertainty about $\boldsymbol{\theta}_k$ will shrink as new information is included in the updating process.

Based on prior knowledge, the following non-observable quantities are constrained to be constant values: the integration step h is set to the sampling time scale of the observations, $\mathcal{G}$ to the scheduled load level, the operational point of the grid frequency ω_0 to $2\pi 50$.

Note that the only informative prior that is assigned to any component of $\boldsymbol{\theta}$ is to the frequency states ω_t. The information reflected in equation (6.25) is the representation of the fundamental Bayesian assumption that ω_t is a random variable that evolves according to

$$p(\omega_{t+1}|\omega_t) = \mathcal{N}\left(\omega_t + \frac{h}{H}(G_t - D_t), \sigma_\gamma^2\right). \tag{6.26}$$

The variables corresponding to generation G_t and demand D_t are assigned normal non-informative priors reflecting the range of values for the total generation and demand with corresponding variances σ_G^2 and σ_D^2. The inertia H is positive and appears explicitly in the parametric model as an H^{-1} term and is chosen to be an informative inverted Gamma prior.

The variances of the observational noise processes σ_s^2, σ_d^2, and of the dynamical noise processes σ_γ^2, σ_G^2 and σ_D^2, are assigned Gamma non-informative priors reflecting the knowledge that variances are positive and close to zero. From the considerations above, the parameter vector $\boldsymbol{\theta}$ for the grid frequency

equation is of $3N + 6$ components given by

$$\boldsymbol{\theta} = \left(\{\omega_t\}_{t=1}^N, \{D_t\}_{t=1}^N, \{G_t\}_{t=1}^N, H, \sigma_\gamma^2, \sigma_s^2, \sigma_d^2, \sigma_G^2, \sigma_D^2\right). \tag{6.27}$$

Any inference of the parameter values and states of the grid will be made from the posterior distribution. Assuming independency between the components of $\boldsymbol{\theta}$ and once the observations are available, from Bayes' Theorem the full posterior distribution can be written as

$$\begin{aligned} p(\boldsymbol{\theta}|S, I) \;=\; & \prod_{t=1}^{N} p(s_t|\boldsymbol{\theta}, I)\; p(d_t|\boldsymbol{\theta}, I) \times \\ & \prod_{t=1}^{N-1} p(\omega_{t+1}|\omega_t, I) \times \prod_{t=1}^{N} p(D_t|I) p(G_t|I) \times p(H|I) \times \\ & p(\sigma_\gamma^2|I)\; p(\sigma_s^2|I)\; p(\sigma_d^2|I)\; p(\sigma_G^2|I)\; p(\sigma_D^2|I). \end{aligned} \tag{6.28}$$

The first line in equation (6.28) corresponds to the Likelihood term, the second line to the prior distribution for the deterministic dynamical states represented in the probability model and the third line to the pure prior terms. Priors reflecting the uncertainty in the dynamical states of the grid system are represented by $p(\sigma_\gamma^2|I)\; p(\sigma_G^2|I)\; p(\sigma_D^2|I)$, the ones representing the uncertainty in the observations are $p(\sigma_s^2|I)\; p(\sigma_d^2|I)$ whilst $p(H|I)$ is the prior for the system inertia.

Equation 6.28 is a high dimensional distribution with as many dimensions as components in $\boldsymbol{\theta}$, (6.27). Once S is observed, N frequency and demand observations are available. When evaluating the posterior of equation 6.28

using the observed set S, this inference problem consists of characterising a full posterior of dimension $\ell = 3N + 6$. Even for a small ℓ an analytical solution is not feasible and numerical solutions are required.

There are several possible numerical implementations to solve equation (6.28) such as the Laplace approximation (see for example [9] pages 340–345 and references therein), Importance Sampling (see for example [9] pages 348–350) and Sampling-Importance-re-sampling [81], all known as non-MCMC techniques. Here, MCMC techniques are implmented as presented in the following section.

6.5.1 MCMC Implementation

This section deals with the intermediary steps needed to implement the MCMC algorithm with the Gibbs sampler of the posterior distribution for the grid frequency dynamics model in an iterative fashion. section 2.1.1 presents how the MCMC algorithm generates samples from the posterior distribution and here these ideas are going to be presented briefly again. Samples of each of the components are generated using the single component Metropolis-Hastings algorithm [40, 69, 30] with the Gibbs sampler [32, 30] which has been proved useful in a wide range of applications of Bayesian inference [95].

To implement MCMC, the following steps are carried out:

1. Classification of model variables and parameters as observables or non-observables, as in Section 6.5.

2. Construction of a probability model, *i.e.* a posterior distribution, as in Section 6.5, equation (6.28).

3. Set up prior and Likelihood terms for parameter in $\boldsymbol{\theta}$ and observables in S, (6.21), based on background information, I. Likelihood and prior terms are included explicitly in section 6.5.1.1 and section 6.5.1.2, respectively.

4. Calculation of the full conditional distributions.

5. Implementation of the sampling algorithms for each of the components of the parameter vector $\boldsymbol{\theta}$, where it is necessary, *i.e.* when the full conditional is not in a closed form. This is the case only for the Inertia parameter, H.

6. Convergence tests (*e.g.* GR-statistic [31, 95]) to determine a suitable burn-in time, $\tau_{\boldsymbol{\theta}}$, as in Section 6.5.2.

7. Iteration of the MCMC algorithm for $j \geq \tau_{\boldsymbol{\theta}}$ to obtain samples from the posterior distribution, section 6.5.2.

Specifically, MCMC generates samples for each component of the parameter vector $\boldsymbol{\theta}$ in equation (6.27) individually by means of the single-component

Metropolis-Hastings algorithm [69]. MCMC algorithmic representation of the high dimensional posterior distribution in equation (6.28) translates to obtaining samples from a suitable Markov chain for each component of the parameter vector $\boldsymbol{\theta}$.

MCMC is implemented for 30 minute windows of observations, given that 30 minutes is the largest time scale at which the relative populations of generation types are held "constant" [46], *i.e.* λ_i is constant. In this time interval, the parameter vector $\boldsymbol{\theta}$ is of dimension ℓ given by

$$\begin{aligned} \ell &= (3 \text{ variables}) \times (30 \text{ min}) \times (60 \text{ sec}) + (6 \text{ parameters}), \\ &= 5406 \text{ components.} \end{aligned} \tag{6.29}$$

In short, MCMC generates a sample of $\boldsymbol{\theta} \in \mathbb{R}^{\ell}$ from the posterior distribution $p(\boldsymbol{\theta}|S, I)$ once S is observed. In order to do this, a Markov chain is generated such that its state space dimension is ℓ and whose equilibrium distribution is the posterior $p(\boldsymbol{\theta}|S, I)$ [66]. After sufficiently many iterations the resulting states of the Markov chain can be taken as samples from the posterior of interest [66, 9]. Estimations for the expected value of any function $g(\boldsymbol{\theta})$ is then made by means of a Monte Carlo approximation.

To construct a suitable Markov chain the algorithm due to Hastings [40] which is a generalisation of the method of Metropolis [69] is employed as follows.

Let $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(j)}, \ldots\}$ in $\mathbb{R}^{\ell}$ be the states for a suitable Markov chain. A candidate $\boldsymbol{\theta}'$ for chain state $j+1$, is drawn from a proposal distribution $q(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}')$ such that the new candidate is accepted with probability $\alpha(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}')$. When the candidate is accepted $\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}'$ otherwise the chain does not move, $\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)}$, and the state is the same as in the last iteration.

There are several ways to choose the proposal distribution $q(\cdot|\cdot)$ [95, 91, 61, 80]. Moreover the choice of the candidate distribution is an important issue for the implementation of MCMC. The proposal should be easily evaluated and sampled from as having a high probability of acceptance $\alpha(\cdot)$ to ensure computational efficiency [95].

In this implementation, the proposal distribution is defined by the Gibbs sampler [32, 30]. The Gibbs sampler defines the proposal distribution to be the full conditional distribution [9, 95] defined in such a way that the candidates $\boldsymbol{\theta}'$ are accepted with probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1$.

The full conditional distributions, $\pi_S(\theta_{.i}|\theta_{.-i})$, are easily calculated from equation (6.28) by just picking the terms which implicitly contain $\theta_{.i}$. Given an initial condition $\boldsymbol{\theta}^{(0)}$ each iteration $j > t_0$ samples each component of the parameter vector from the full conditional distribution, *i.e.* $\theta_{.i}^{(j)} \sim \pi_S\left(\theta_{.i}|\{\theta_{.i'}^{(j-1)}|\ i' < i\}, \{\theta_{.i'}^{(j)}|\ i' < i\}\right)$ for $i = 1, \cdots, \ell$, see section 2.1.1 for details.

Note that the MCMC technique takes as an initial condition for the chain

a realisation of each of the components of $\boldsymbol{\theta}$ drawn from the corresponding priors, and data for a 30 minute window. The iteration time is denoted by $j \in \mathbb{Z}$, the dynamical time by $t \in \mathbb{Z}$, and the components of the parameter vector $\boldsymbol{\theta}$ are denoted by $i \in \mathbb{Z}$.

It is important to keep in mind that the grid system dynamical variables, *i.e.* ω, G and D, do not evolve in a deterministic way within an iteration of the MCMC algorithm. Instead, for the j^{th} iteration of the algorithm, a realisation of the whole dynamics for the 30min-window is contained in $\boldsymbol{\theta}^{(j)}$. After many iterations an ensemble of possible grid system states is available (see discussion of this interpretation of the traces obtained by MCMC techniques in Chapter 5). Unlike an ensemble approach, the probabilistic model *i.e.* the posterior distribution, explores the space of possibles states for the dynamics of the model during those 30 minutes consistently with the background information I.

For the purpose of calculating full conditionals, some terms already discussed in section 6.5 will be revisited and some that will not be relevant until sections 6.5.2 are going to be introduced, *e.g.* terms related to the inclusion of sub-sampled demand values that mimic operational constraints.

The next two sections list explicitly, Likelihood and prior terms in equation (6.28).

6.5.1.1 Likelihood Terms

The Likelihood terms are defined for the observed variables. In this case, both the grid frequency observations, s_t, and the demand observations, d_t, are assumed to be independent for all t.

- **Grid Frequency Observations: $\{s_j\}_{j=1}^{N}$**

Observations of grid frequency are available at a sampling time of 1sec.

$$s_t \sim \mathcal{N}(\omega_t, \sigma_s^2), \tag{6.30}$$

$$p(s_t|\boldsymbol{\theta}, I) = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left[-\frac{1}{2\sigma_s^2}(s_t - \omega_t)^2\right]. \tag{6.31}$$

Assuming independence and normality for the s_t, the Likelihood term for the set of grid frequency observations in the 30min-window is

$$p(\{s_t\}_{t=1}^{N}|\boldsymbol{\theta}, I) = \prod_{t=1}^{N} p(s_t|\boldsymbol{\theta}, I) \tag{6.32}$$

$$= \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left[-\frac{1}{2\sigma_s^2}(s_t - \omega_t)^2\right], \tag{6.33}$$

$$= \left(\frac{1}{2\pi\sigma_s^2}\right)^{N/2} \exp\left[-\frac{1}{2\sigma_s^2}\sum_{t=1}^{N}(s_t - \omega_t)^2\right]. \tag{6.34}$$

- **Demand Observations: $\{d_t\}_{t=1}^{N}$**

Assuming demand observations are available at a sampling time of 1sec (1Hz rate).

$$d_t \sim \mathcal{N}(D_t, \sigma_d^2), \tag{6.35}$$

$$p(d_t|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma_d^2}} \exp\left[-\frac{1}{2\sigma_d^2}(d_t - D_t)^2\right]. \tag{6.36}$$

In practice, observations of the demand are available at a variable sampling time as described in section 6.1.1. The sampling time varies up to 40 seconds due to some outlier sampling times (see Figure 6.2).

For missing demand observations, a process is defined to complete the gaps between existent demand observations while MCMC updates the components of the parameter vector.

Let Q be the set of times where demand observations exist,

$$Q = \{q_{(k)} \in \mathbb{Z} |\ q_{(k)} = t \text{ when } d_t \text{ is not a missing value}\},$$

where Q is an ordered set, *i.e.* $q_{(k)} < q_{(k+1)}$, and $k = 1, \ldots, \dim Q$. The dimension of the set Q, $\dim Q$, is equal to the number of existing demand values in the 30min-window at 1Hz and it is assumed $\{q_{(1)} = 1, q_{(\dim Q)} = N\} \in Q$, *i.e.* the first and last point in the window are not missing.

Missing observations of the demand are contained in intervals of the form $[d_{q_{(k)}}, d_{q_{(k+1)}}]$, the straight line that joins the limits of the interval is given by

$$d(t) = \frac{d_{q_{(k+1)}} - d_{q_{(k)}}}{q_{(k+1)} - q_{(k)}}\, t + \frac{q_{(k+1)} d_{q_{(k)}} - q_{(k)} d_{q_{(k+1)}}}{q_{(k+1)} - q_{(k)}}. \tag{6.37}$$

Let $Q^\star = \{t \in \mathbb{Z}\}_{t=1}^{N} - Q$ be the complement of Q, the set of time indexes of missing demand values. This set is

$$Q^\star = \{m \in \mathbb{Z} |\ m = t \text{ when } d_t \text{ is a missing value}\},$$

The missing demand observations, d_m for $m \in Q^\star$, are completed by a per-

turbed linear interpolation in the interval $[d_{(q_k)}, d_{(q_{k+1})}]$, independent and normally distributed over the straight line of equation (6.37), Figure 6.7 sketches this idea.

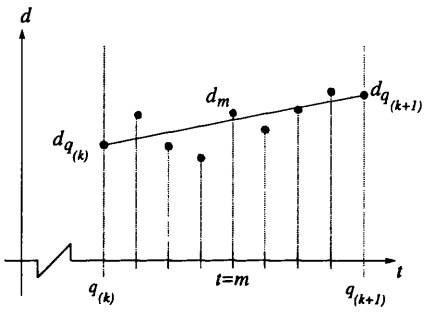


Figure 6.7: Graphical representation of the random linear interpolation process to generate missing demand observations.

The d_m's are distributed as

$$d_m \sim \mathcal{N}(d(m), \sigma_\delta^2), \tag{6.38}$$

$$p(d_m|I) = \frac{1}{\sqrt{2\pi\sigma_\delta^2}} \exp\left[-\frac{1}{2\sigma_\delta^2}(d_m - d(m))^2\right]. \tag{6.39}$$

where $d(m)$ is equation (6.37) evaluated at $t = m$, and σ_δ^2 is the amplitude of the random perturbation.

Assuming that d_m's are independent for all $m \in Q^\star$,

$$p(d^\star) = \prod_{m \in Q^\star} \frac{1}{\sqrt{2\pi\sigma_\delta^2}} \exp\left\{-\frac{1}{2\sigma_\delta^2}(d_m - d(m))^2\right\}, \tag{6.40}$$

where $d^\star$ is the set of missing values, *i.e.* $d^\star = \{d_m | \ \forall m \in Q^\star\}$.

Note that equation (6.40) is independent of any other component of $\boldsymbol{\theta}$ in

equation 6.27, it only depends on demand observations. In addition, in each iteration of the algorithm, new realisations d_m's are drawn from (6.38).

After considering all observables in S and the available background information I, the Likelihood term is:

$$\begin{aligned} p(S|\boldsymbol{\theta}, I) &= p(\{s_t\}_{t=1}^{N}|\boldsymbol{\theta}, I) \times p(\{d_t\}_{t=1}^{N}|\boldsymbol{\theta}, I), && (6.41) \\ &= p(\{s_t\}_{t=1}^{N}|\boldsymbol{\theta}, I) \times && (6.42) \\ &\quad p(\{d_{q_{(k)}}\}_{k=1}^{\dim Q}|\boldsymbol{\theta}, I) p(\{d_m\}_{m\in Q^\star}|I), && (6.43) \end{aligned}$$

in the case where demand observations are missing.

6.5.1.2 Prior Terms

At this stage of the probabilistic modelling the parameter vector $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta} = \left\{\{\omega_t\}_{t=1}^{N}, \sigma_\gamma^2, \{D_t\}_{t=1}^{N}, \sigma_D^2, \{G_t\}_{t=1}^{N}, \sigma_G^2, \sigma_d^2, \sigma_\delta^2, \sigma_s^2, H, \{d_m\}_{\forall m\in Q^\star}\right\}. \tag{6.44}$$

For each component in $\boldsymbol{\theta}$ a prior distribution is set according to the background information on the grid frequency system. In the process, additional hyper-parameters are introduced to refine the modelling of some priors and/or to account for uncertainty in some components of $\boldsymbol{\theta}$. In this section, each component of $\boldsymbol{\theta}$ in equation (6.44) is modelled in order of appearance.

- **Grid Frequency States:** $\{w_t\}_{t=1}^{N}$

The prior for the grid frequency states is presented in section 6.26 and is given by

$$p(\omega_{t+1}|\omega_t, I) = \mathcal{N}\left(\omega_t + \frac{h}{H}(G_t - D_t), \sigma_\gamma^2\right), \tag{6.45}$$

reflecting the expectation that the grid frequency states should evolve close to the value of the Euler approximation of the simple model (6.24), and also to control model error, see equation (6.45). This term, is in general, referred to as additive dynamical noise with amplitude σ_γ^2.

For ω_t, $t = 1, \ldots, N$, the prior is

$$p(\{\omega_{t+1}\}_{t=1}^{N-1}|\{\omega_t\}_{t=1}^{N-1}, I) = \prod_{t=1}^{N-1} p(\omega_{t+1}|\omega_t, I). \tag{6.46}$$

- **Dynamical Noise Variance for ω_t : σ_γ^2**

Regarding prior information about σ_γ^2, it is desirable to keep the model error close to zero and with small variation around the model state ω_t. Hence, σ_γ^2 follows an Inverse Gamma distribution:

$$\frac{1}{\sigma_\gamma^2} \sim \mathcal{G}a(\alpha_\gamma, \beta_\gamma), \tag{6.47}$$

$$p(\sigma_\gamma^2|I) = \frac{\beta_\gamma^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)}\left(\frac{1}{\sigma_\gamma^2}\right)^{\alpha_\gamma+1} \exp\left\{-\frac{1}{\beta_\gamma \sigma_\gamma^2}\right\}. \tag{6.48}$$

Mean and variance are set correspondingly to $\text{mean}(\sigma_\gamma^2) = 1 \times 10^{-4}$ and $\text{var}(\sigma_\omega^2) = 1 \times 10^{-3}$, which in turn, determine α_γ and β_γ[17].

- **Demand States: $\{D_t\}_{t=1}^N$**

The D_t's correspond to the demand model variables. The difference between d_t and D_t is that d_t is the observation of D_t at time t. For each value of the model demand at time j a non-informative prior distribution is set to be

$$D_j \sim \mathcal{N}(0, \sigma_D^2), \tag{6.49}$$

where $\sigma_D^2 = 1$. Therefore, the demand state follows a density of the form:

$$p(D_t|I) = \frac{1}{2}\exp\left\{-\frac{D_t^2}{2}\right\}. \tag{6.50}$$

The D_t's are independent and identically distributed, therefore for all $t = 1, \ldots, N$, the prior is:

$$p(\{D\}_{t=1}^N|I) = \left(\frac{1}{2}\right)^{\frac{N}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{N} D_t^2\right\}. \tag{6.51}$$

- **Generation States: $\{G_t\}_{t=1}^N$**

Demand and generation model variables given by equation (6.22) could be seen as "mirror" variables. The exact balance between G and D reflects no changes in the grid frequency ω. On the contrary, unbalance between G and D reflects into positive or negative deviations from the operational point of the grid.

Given this close relationship between generation and demand, the prior set for the generation state G_t is the same as the prior for the demand state

D_t given in equation (6.49) as the background information about G_t and D_t is the same since internal variables are not included explicitly in the implementation. Therefore, the prior for the generation states is

$$p(\{G_t\}_{t=1}^{N}|I) = \left(\frac{1}{2}\right)^{\frac{N}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{N} G_t^2\right\}, \tag{6.52}$$

under independency and normality assumptions of the G_t's for all t in the 30min-window.

- **Demand Observations Variance: σ_d^2**

The variance of the demand observations appears in the demand Likelihood term in equation (6.35). It describes how variable are the observations of demand. This hyper-parameter is set to be a known constant parameter. The prior associated with σ_d^2 is constrained to be the variance of the existing demand observations:

$$\sigma_d^2 = \text{var}\left(\{d_t\}_{\forall t \in Q}\right). \tag{6.53}$$

- **Missing Demand Observations Variance: σ_δ^2**

The variance σ_δ^2 appears in equation (6.40). It represents how spread the missing demand observations are around the line that joins existing demand observations. The prior chosen reflects that σ_δ^2 is always positive and should

be close to the variance of the existing demand observations. Hence

$$\frac{1}{\sigma_\delta^2} \sim \mathcal{G}a(\alpha_\delta, \beta_\delta), \tag{6.54}$$

$$p(\sigma_\delta^2|I) = \frac{\beta_\delta^{\alpha_\delta}}{\Gamma(\alpha_\delta)} \left(\frac{1}{\sigma_\delta^2}\right)^{\alpha_\delta+1} \exp\left\{-\frac{1}{\beta_\delta \sigma_\delta^2}\right\}. \tag{6.55}$$

The parameters α_δ and β_δ are found from setting the mean and variance equation (6.54) to mean$(\sigma_\delta^2) = \sigma_d^2$ and var$(\sigma_\delta^2) = 10\sigma_d^2$, respectively. Note that σ_δ^2 is the variance for the perturbation of the linearly interpolated values of the demand (see Figure 6.7).

- **Frequency Observations Variance: σ_s^2**

The variance of grid frequency observations is set to be a constraint parameter. The prior associated with σ_s^2 is set to be a constant value given by

$$\sigma_s^2 = \text{var}\left(\{s_t\}_{t=1}^N\right). \tag{6.56}$$

- **Inertia: H**

The Inertia parameter H appears explicitly as an H^{-1} term in the model (6.20) and from background information, $H > 0$ [46]. For convenience, the prior for H is chosen to be and Inverted Gamma distribution given by:

$$\frac{1}{H} \sim \mathcal{G}a(\alpha_H, \beta_H), \tag{6.57}$$

$$p(H|I) = \frac{\beta_H^{\alpha_H}}{\Gamma(\alpha_H)} \left(\frac{1}{H}\right)^{\alpha_H+1} \exp\left\{-\frac{1}{\beta_H H}\right\}. \tag{6.58}$$

From the calculation of its full conditional distribution it turns out, as seen in next section, that β_H is constrained to be equal to

$$\frac{1}{\beta_H} = \frac{h}{\sigma_\gamma^2} \sum_{t=1}^{N-1} (\omega_{t+1} - \omega_t)(G_t - D_t), \tag{6.59}$$

in order to obtained a closed form of the corresponding full conditional for H. The shape parameter is set to be a constant, *i.e.* $\alpha_H = 2.01$.

After setting Likelihood and prior terms, the full conditionals of each of the components of $\boldsymbol{\theta}$ can be easily calculated from the posterior. The full posterior distribution is then explicitly written by replacing all probability densities described in the last two sections into equation (6.28).

6.5.1.3 Full Conditional Distributions

The full posterior distribution is written explicitly as

$$
\begin{aligned}
p(\boldsymbol{\theta}|S) \propto & \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left\{-\frac{1}{2\sigma_s^2}(s_t-\omega_t)^2\right\} \times \\
& \prod_{t\in Q} \frac{1}{\sqrt{2\pi\sigma_d^2}} \exp\left\{-\frac{1}{2\sigma_d^2}(d_t-D_t)^2\right\} \times \\
& \prod_{t=1}^{N-1} \frac{1}{\sqrt{2\pi\sigma_\gamma^2}} \exp\left\{-\frac{1}{2\sigma_\gamma^2}\left[\omega_{t+1}-\omega_t-\frac{h}{H}(G_t-D_t)\right]^2\right\} \times \\
& \prod_{t\in Q^*} \frac{1}{\sqrt{2\pi\sigma_\delta^2}} \exp\left\{-\frac{1}{2\sigma_\delta^2}(d_t-d(t))^2\right\} \times \\
& \prod_{t=1}^{N} \frac{1}{2} \exp\left\{-\frac{D_t^2}{2}\right\} \times \prod_{t=1}^{N} \frac{1}{2} \exp\left\{-\frac{G_t^2}{2}\right\} \times \\
& \frac{{\beta_\gamma}^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)} \left(\frac{1}{\sigma_\gamma^2}\right)^{\alpha_\gamma+1} \exp\left\{-\frac{1}{\beta_\gamma\sigma_\gamma^2}\right\} \times \\
& \frac{{\beta_d}^{\alpha_d}}{\Gamma(\alpha_d)} \left(\frac{1}{\sigma_d^2}\right)^{\alpha_d+1} \exp\left\{-\frac{1}{\beta_d\sigma_d^2}\right\} \times \\
& \frac{{\beta_H}^{\alpha_H}}{\Gamma(\alpha_H)} \left(\frac{1}{H}\right)^{\alpha_H+1} \exp\left\{-\frac{1}{\beta_H H}\right\}. \qquad (6.60)
\end{aligned}
$$

Once the data S is available, the posterior in equation (6.60) is evaluated. Therefore, a probability density that is only function of $\boldsymbol{\theta}$ is obtained, *i.e.* $\pi_S(\boldsymbol{\theta})$, the full posterior distribution. Clearly, the posterior $\pi_S(\boldsymbol{\theta})$ is a high dimensional distribution. Any inference of $\boldsymbol{\theta}$ will involve calculations of its moments that in turn will involve high dimensional integration, in practice unachievable.

The definition of the full conditional distribution is given by equation

(2.16), Chapter 2 in section 2.1 and can be calculated easily from (6.60).

In all components of $\boldsymbol{\theta}$, the terms containing the relevant component were extracted from the posterior in equation (6.60) and manipulated such that a density probability function could be identified in a closed form when possible. Note that each full conditional distribution is only used once in any given iteration time t.

- **Grid frequency states**, $\{\omega_t\}_{t=1}^{N}$:

$\theta_{.i}$ i^{th} Component	PDF Functional Form	PDF Parameters
$\theta_{.1} = \omega_1$	$\mathcal{N}\left(\frac{B_{\omega_1}}{A_{\omega_1}}, \frac{1}{A_{\omega_1}}\right)$	$A_{\omega_1}=\frac{1}{{\sigma_s}^2}+\frac{1}{{\sigma_\gamma}^2}$ $B_{\omega_1}=\frac{s_1}{{\sigma_s}^2}+\frac{1}{{\sigma_\gamma}^2}\left[\omega_2-\frac{h}{H}(G_1-D_1)\right]$
$\theta_{.3t-2} = \omega_t,$ $t = 2,\ldots,N-1$	$\mathcal{N}\left(\frac{B_{\omega_t}}{A_{\omega_t}}, \frac{1}{A_{\omega_t}}\right)$	$A_{\omega_t}=\frac{1}{{\sigma_s}^2}+\frac{2}{{\sigma_\gamma}^2}$ $B_{\omega_t}=\frac{s_t}{{\sigma_s}^2}+\frac{1}{{\sigma_\gamma}^2}\left[\omega_{t-1}+\frac{h}{H}(G_{t-1}-D_{t-1})\right]+$ $\frac{1}{{\sigma_\gamma}^2}\left[\omega_{t+1}-\frac{h}{H}(G_t-D_t)\right]$
$\theta_{.N} = \omega_N$	$\mathcal{N}\left(\frac{B_{\omega_N}}{A_{\omega_N}}, \frac{1}{A_{\omega_N}}\right)$	$A_{\omega_N}=\frac{1}{{\sigma_s}^2}+\frac{1}{{\sigma_\gamma}^2}$ $B_{\omega_N}=\frac{s_N}{{\sigma_s}^2}+\frac{1}{{\sigma_\gamma}^2}\left[\omega_{N-1}+\frac{h}{H}(G_{N-1}-D_{N-1})\right]$

Table 6.4: Full conditionals for Grid Frequency states $\{\omega_t\}_{t=1}^{N}$.

- **Demand states, $\{D_t\}_{t=1}^N$:**

$\theta_{.i}$ ith Component	PDF Functional Form	PDF Parameters
$\theta_{.3t-1} = D_t$, for $t \in Q$ and $t \neq N$	$\mathcal{N}\left(\frac{B_D}{A_D}, \frac{1}{A_D}\right)$	$A_D = \frac{1}{{\sigma_d}^2} + \frac{h^2}{H^2{\sigma_\gamma}^2}$ $B_D = \frac{d_t}{{\sigma_d}^2} + \frac{h}{H{\sigma_\gamma}^2}\left[\omega_t - \omega_{t+1} + \frac{hG_t}{H}\right]$
$\theta_{.3N-1} = D_N$	$\mathcal{N}\left(\frac{B_D}{A_D}, \frac{1}{A_D}\right)$	$A_D = \frac{1}{{\sigma_d}^2} + 1$ $B_D = \frac{d_N}{{\sigma_d}^2}$
$\theta_{.3t-1} = D_t$, for $t \in Q^\star$	$\mathcal{N}\left(\frac{B_D}{A_D}, \frac{1}{A_D}\right)$	$A_D = \frac{h^2}{H^2{\sigma_\gamma}^2} + 1$ $B_D = \frac{h}{H{\sigma_\gamma}^2}\left[\omega_t - \omega_{t+1} + \frac{hG_t}{H}\right]$

Table 6.5: Full conditionals for Demand states $\{D_t\}_{t=1}^N$.

- **Generation states, $\{G_t\}_{t=1}^N$:**

$\theta_{.i}$ ith Component	PDF Functional Form	PDF Parameters
$\theta_{.3t} = G_t$, for $t = 1, \ldots, N-1$	$\mathcal{N}\left(\frac{B_G}{A_G}, \frac{1}{A_G}\right)$	$A_G = \frac{h^2}{H^2{\sigma_\gamma}^2} + 1$ $B_G = \frac{h}{H{\sigma_\gamma}^2}\left[\omega_t - \omega_{t+1} - \frac{hD_t}{H}\right]$
$\theta_{.3N} = G_N$	$\mathcal{N}(0,1)$	

Table 6.6: Full conditionals for Generation states $\{G_t\}_{t=1}^N$.

- **Grid Frequency Variance, $\sigma_\gamma{}^2$:**

$\theta_{.i}$	PDF	PDF
i^{th} Component	Functional Form	Parameters
$\theta_{.3N+1} = \sigma_\gamma^2$	$\mathcal{IG}a(A_\gamma, B_\gamma)$	$A_\gamma{=}(N-1+2\alpha_\gamma)/2$ $B_\gamma{=}\left[\frac{1}{2}\sum_{t=1}^{N-1}\left[\omega_{t+1}-\omega_t-\frac{h}{H}(G_t-D_t)\right]^2+\frac{1}{\beta_\gamma}\right]^{-1}$

Table 6.7: Full conditional for Grid Frequency variance σ_γ^2.

Note that if $X \sim \mathcal{G}a(\alpha, \beta)$ and $Y = 1/X$ then $Y \sim \mathcal{IG}a(\alpha, \beta)$ [17], where $\mathcal{IG}a(\cdot)$ is the Inverse Gamma Distribution [17].

- **Missing Demand Variance, σ_δ^2:**

$\theta_{.i}$	PDF	PDF
i^{th} Component	Functional Form	Parameters
$\theta_{.N+2} = \sigma_\delta^2$	$\mathcal{IG}a(A_\delta, B_\delta)$	$A_\delta{=}\frac{dim(Q^\star)}{2}+\alpha_\delta$ $B_\delta{=}2\beta_\delta\left\{2+\beta_\delta\sum_{t\in Q^\star}(d_t-d(t))^2\right\}^{-1}$

Table 6.8: Full conditional for Missing Demand Variance, σ_δ^2.

- **Grid System Inertia, H:**

The functional form of the full conditional PDF found for the Inertia parameter is not in a closed form, *i.e.* does not have the functional form of a

"known" probability density function. To generate samples of H from that PDF, there are two options:

1. Code a routine which generates random samples of such PDF found using, for example, the Acceptance/Resection algorithm [17].

2. Find a change of variable which will transform the PDF found into a closed form and then use an existing routine for random samples for a "known" probability density.

In order to avoid low convergence rates of the MCMC algorithm due to possible slow random generation [17, 9, 77, 14], option 2. is followed. In general, if such a change of variable is not found feasible then option 1. is the only choice.

$\theta_{.i}$	PDF	PDF
i^{th} Component	Functional Form	Parameters
$\theta_{.N+3} = H$	$\pi(\theta_{.H}\|\theta_{.-H}) \propto \left(\frac{1}{H^2}\right)^{\alpha} \exp\left\{-\frac{A}{H^2}\right\}$	$\alpha = \frac{\alpha_H + 1}{2}$ $A = \frac{h^2}{{\sigma_\gamma}^2} \sum_{t=1}^{N-1} (G_t - D_t)^2$

Table 6.9: Full conditional for the Grid System Inertia, H.

Let $f_X(x)$ be the density function corresponding to that associated with

the functional form of the full conditional found for the Grid Inertia:

$$f_X(x) \propto \left(\frac{1}{x^2}\right)^{\alpha} e^{-A/x^2},\ 0 < x < \infty, \tag{6.61}$$

where A and α are positive constants. Let $y = g(x)$ be the change of variable chosen to transform (6.61) in a close form, where $g(X) = X^2$. Therefore, $g^{-1}(y) = \sqrt{y}$ and $\frac{\mathrm{d}}{\mathrm{d}y}[g^{-1}(y)] = \frac{1}{2}y^{-1/2}$. From the theorem for distribution transformation [17, 9] and for $y \in (0, \infty)$, it is obtained that

$$\begin{aligned} f_Y(y) &\propto f_X(g^{-1}(y)) \left|\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)\right|, \\ &\propto \left(\frac{1}{y}\right)^{\alpha} e^{-A/y} \frac{1}{\sqrt{y}}, \\ &\propto \left(\frac{1}{y}\right)^{\alpha+1/2} e^{-A/y}. \end{aligned} \tag{6.62}$$

From equation (6.62) is clear that $1/y \sim \mathcal{G}a(\alpha_y, \beta_y)$. In other words, y follows an Inverse Gamma distribution with parameters $\alpha_y = \alpha - 1/2$ and $\beta_y = 1/A$. Replacing the PDF parameters in Table 6.9 into (6.62), it is obtained that

$$y \sim \mathcal{IG}a\left(\frac{\alpha_H}{2}, \frac{1}{A}\right). \tag{6.63}$$

Generation of random samples of H is made by generating random samples of y from (6.63), and a sample of the inertia parameter is given by $H = \sqrt{y}$.

Once full conditionals are available and samples can be drawn efficiently, the implementation of MCMC is now reduced to write two loops:

i. The outer loop runs over the iteration time j. Thus, after one iteration, the chain constructed for $\boldsymbol{\theta}$ has moved one step.

ii. The inner loop runs over the components of $\boldsymbol{\theta}$, i. Each component is sampled from the corresponding full conditional distributions listed in Tables 6.4 to 6.9 and equation (6.63). At each iteration j, and for each component $\theta_{.i}$, the full conditionals are evaluated in $\theta_{.-i}$, updated in the present and past iteration $j-1$.

Remark that full conditionals are one dimensional PDF's. Combination of chain components updated in the past and current iteration is key to understand chain mixing [13, 31]. The mixing of the chain takes place while the first several iterations of the algorithm combined posterior information thus convergence is approach [83, 66]. The mixing period is the burn-in time, $\tau_{\boldsymbol{\theta}}$. As in previous Chapters, direct inssection of the resulting Monte Carlo estimates and parallel runs of the chain are both used and considered to find the burn-in time of the chain.

The next sections present the results obtain for several scenarios of the grid system posed in such a way that gradually real operational conditions are reflected in S.

6.5.2 ReMS: MCMC Estimates

To investigate the performance of MCMC for the problem of estimating the parameters of a model for grid frequency dynamics a number of experiments is carried out. All experiments are designed in a similar way to the ones presented in section 6.4 when sensitivity of LS estimates in the ReMS was explored. Note that even though the problem of parameter estimation is formulated for the simple model of the grid system using NRA, the use of ReMS is indistinguishable from the PMS. The use of the PMS here is made in the NSA since PMS data is not generated by stochastic processes but from the numerical integration of the deterministic grid model.

All experiments are run for approximately a 34 minutes window, *i.e.* $N = 20480$ data points of demand and grid frequency data, the algorithm is initially run for an iteration time of $T = 1000$. Preliminary studies on the initial iteration time to run the MCMC algorithm were performed during the REMIND project [20]. Given the high dimensionality of the parameter vector, *i.e.* 5406 components, computational time vastly increases when T is increased. Real CPU time for 1000 iterations is approximately 2 hours for the simplest configuration of MCMC. The results of the experiments presented in this and following section are presented in [22].

1. *Experiment 1: PMS High Resolution Data*

It is set as a convergence safety check of the methodology in the PMS. Given that the model is approximated using the forward Euler method for a time step of $h = 0.1$, PMS data is the numerical integration of the deterministic model described in section 6.2 using the same h. The resulting demand and grid frequency data is used to evaluate the posterior distribution in equation (6.28) to obtain $\pi_S(\boldsymbol{\theta})$.

2. *Experiment 2: PMS Sub-sampled Demand Data*

   Investigates the effects of using data with limited sampling rates by sub-sampling the PMS data at several slower rates. The sampling rates studied correspond to sampling times of $\Delta t = 0.2, 0.4, 0.6, 0.8, 1$ seconds, respectively. For all cases, grid frequency data is kept at high resolution, *i.e.* 10Hz.

3. *Experiment 3: Real Operational Conditions*

   It studies the performance of the MCMC technique to calculate parameter estimates in the PMS when real operational conditions are reproduced for both demand and grid frequency. MCMC output is generated using $h = 1$sec using PMS sub-sampled data and real data. Performance of the technique is studied by comparing estimates obtained for both PMS sub-sampled data and real data. This experiment is presented in section 6.6.

The set up for Experiment 1 is designed to asses convergence of the Markov chain, and to establish that a satisfactory level of mixing has been achieved by the calculation of the Gelman-Rubin (GR) statistic [31] for several scalar summaries of the resulting posterior samples. The GR statistic is calculated for three parallel runs of the perfect model scenario implementation of MCMC. Here it is considered data generated by the deterministic model of equations (6.10) to (6.16) driven by synthetic demand. The parameter values for this model can be found in Table 6.1 of section 6.3.

The stochastic grid frequency dynamics model of equation (6.20) is then fitted to the data using the Bayesian techniques described in section 6.5.

Setting informative and uninformative priors on generation and demand for the perfect model case results in posterior distributions that not only converge but whose variances shrink on iterating the algorithm. Figure 6.8 shows the convergence for two scalar summaries from the posterior. The GR statistics are calculated for the Grid Inertia parameter I in the upper row (panels A and B) and for the Grid Frequency variance σ_γ^2 in the lower column (panels C and D). The GR statistic for the median is plotted in the left column (panels A and C) whilst the 97.5% percentile is plotted in the right column (panels B and D).

The variance of the dynamical noise σ_γ^2 and estimates for the model inertia H both converge satisfactorily after approximately 500 iterations, indicated

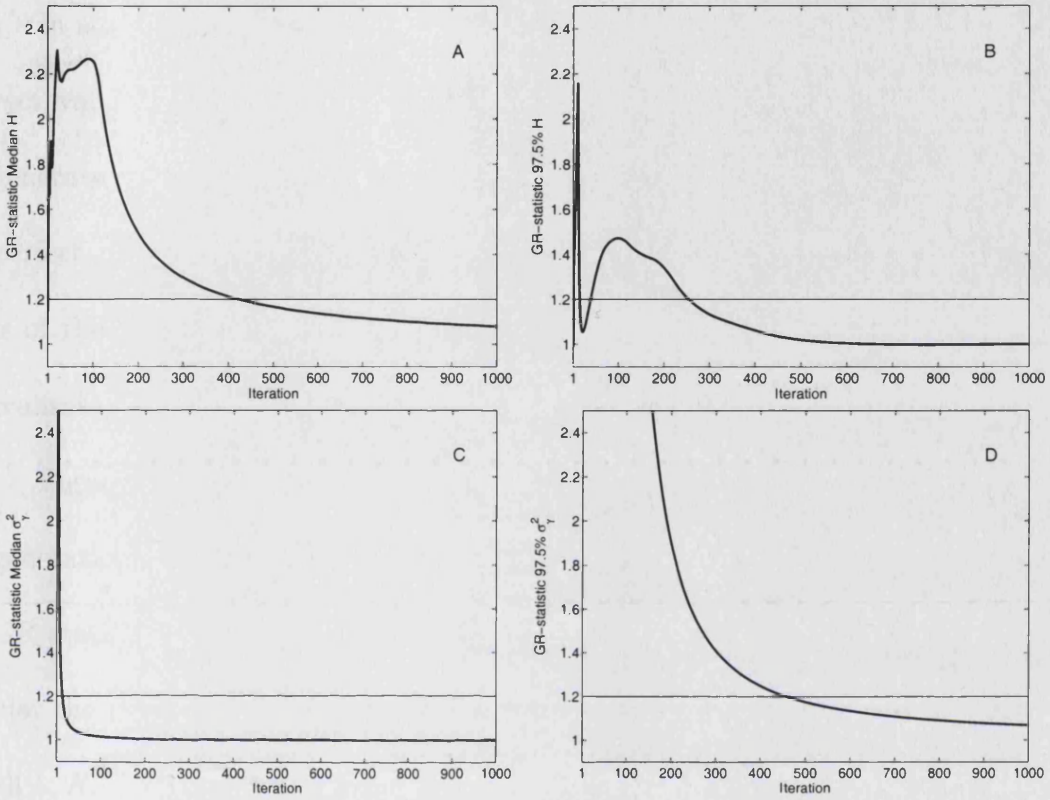


Figure 6.8: Convergence of *Experiment 1* for the inertia H (panels A and B), and dynamical noise variance σ_γ^2 (panels C and D). Plot A shows the GR statistic as a function of iteration for the median of the posterior for the inertia, H. Plot B shows the convergence of the 97.5^{th} percentile of the posterior for H. Figures C & D show the convergence in the median and 97.5^{th} percentile for the posterior distribution of the dynamical noise variance σ_γ^2, respectively.

by GR-statistic < 1.2. Similar convergence is achieved in all other components. Consistently, the output of the MCMC algorithm, tends to stabilise after a burn-in time of $\tau_{\boldsymbol{\theta}} = 500$ iteration in this first experiment.

In addition to displaying convergence, the estimates are close to the correct values. The estimated model inertia converges to the value used to generate the data; the mean of the posterior distribution is within 6% of the perfect model value. In addition, the estimated dynamical noise variance is of the order 10^{-7}. These results are consistent across the three different realisations.

Figure 6.9 plots the GR-statistic for the grid frequency (panel A) and generation (panel B) states. For each dynamical state and for each t, the GR statistic is calculated as a function of the iteration time, j. The panels plot the minimum and maximum of the GR statistic at iteration time j over all t. After 500 iterations of the algorithm the GR-statistic is well below the pass mark of the test, *i.e.* $GR < 1.2$, and is in the range of convergence for all t and for each dynamical state, both grid frequency and generation.

Note that the left hand plot, corresponding to the grid frequency states, shows a strong convergence of the MCMC samples for all values of t in the 30 minutes of interest. In the case of the generation states, convergence is achieved more slowly for all t given that an uninformative prior is set for the generation states, see (6.52). Note that after 100 iterations ω_t has already converged for all three runs whilst the corresponding generation states G_t have not.

It is clear, that given frequently sampled high resolution data the Bayesian

parameter estimation technique is able to produce consistent estimates for the inertia and generation. Moreover, uncertainties in the prior distributions shrink and the means of the marginal posterior distributions are close to the true values.

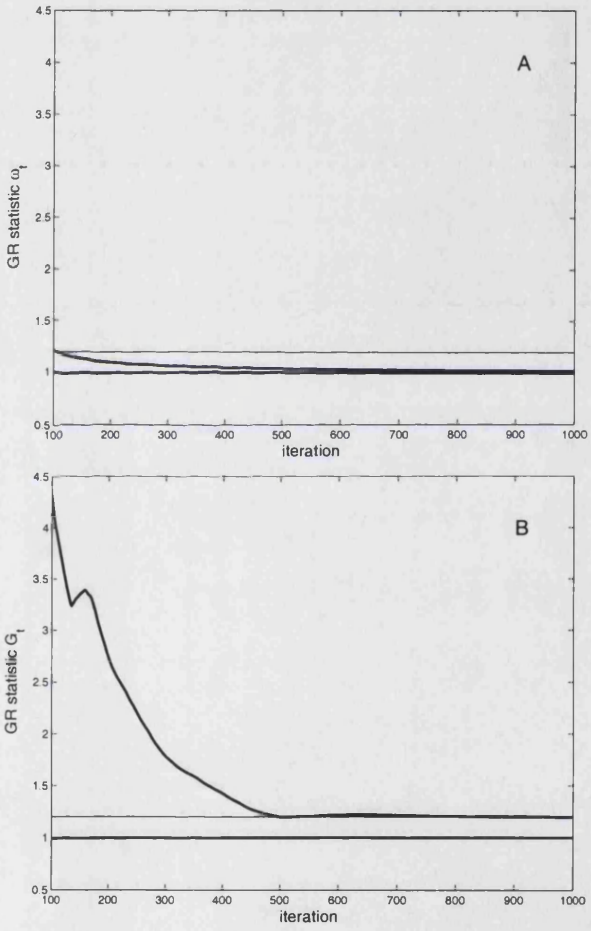


Figure 6.9: Convergence of temporal grid model states. Plot A shows convergence of the grid frequency ω_t, plot B shows convergence in the generation G_t as a function of iteration. The plots show the minimum and maximum GR statistics computed over all t.

Experiment 1 demonstrates that the method works in the situation where the model has the same structure as the observed system.

The motivation for Experiment 2 is to study the sensitivity and robustness of this MCMC implementation to changes in sampling rates. Sampling rates are decreased in a way that they resemble the operational conditions described in section 6.1.1. In practice, the rate at which data is sampled is constrained by the monitoring equipment in situ. Model demand data is sub-sampled at times ranging from 0.2 seconds to 1 second. This sub-sampled demand data is then linearly interpolated so that the sampling time is effectively 1Hz and is commensurate with the grid frequency observations.

Figure 6.10 shows the effect of sampling time on MCMC performance. The mean value of the marginal posterior distribution for H varies as the sampling time of demand is increased, increasing for sampling times greater than 0.6 seconds. In addition the variance of the posterior distribution increases as the sampling time decreases, indicating that the obtained estimates of H are less certain when sampling time is too coarse. The posterior distributions are still consistent with the target values; the mean of the posteriors is are all within 7% of the value that generated the data. Moreover, the variance of the posteriors, although increasing, is still relatively small. As such, the quality of the estimates is degraded but there is still sufficient information in the sub-sampled data stream that allows the identification of model

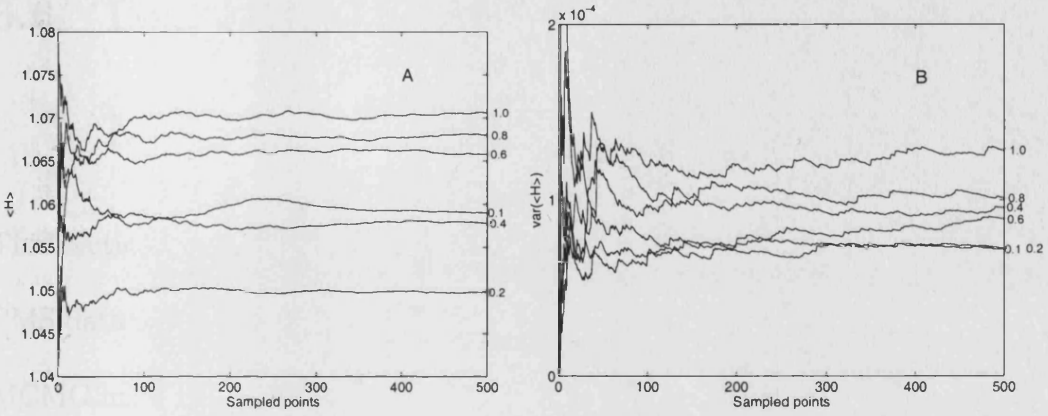


Figure 6.10: Effect of sampling time on the marginal posterior distribution for H. Plot A shows the mean of the posterior distribution for H. The values have been normalised so that 1 corresponds to the value that generated the data. Plot B shows the variance of the posterior empirical distribution for the inertia. It is clear, the mean varies as the sampling time changes and that the variance in the posterior increases as the sampling time decreases.

parameters demonstrating that the MCMC is of value in the PMS setting despite the fact that operational contrains were imposed.

Attempts to estimate parameters in this model scenario using a geometric approach as the one presented in Chapter 4 are planned for future work. Using this approach, uncertainty on both observations and model class is consistently generated while model trajectories are consistent with the observations and the impersection of the model. Gradient descent methods have been shown useful in characterisation of complex systems [51, 50] such as weather models.

6.6 Real Operational Conditions: MCMC Estimates

This section describes Experiment 3. This experiment uses real data and PMS data with the same operational conditions as the real data through the MCMC implementation.

In the third experiment, the practical constraints given the operationally available data are considered. In particular, it is examined the effect of operational sampling times on MCMC performance. To do this, parameter estimation is performed for two data sets. The first data set consists of observed grid frequency and demand data provided by NGT. The authors are grateful to Hai-Bin Wan and NGT for providing these data and assisting the interpretation of measured quantities.

The second data set consists of model data sampled at rates comparable with the observed data, that is, the sampling times are engineered to have the same distribution as the operational data.

A necessary condition for any meaningful analysis of operational data will be that adequate results are achievable using model data that is sampled operationally. That is, it is required that the posterior distributions given operationally sampled model data, where the parameter values that generated the data are known, be consistent with the perfect model experi-

ments of section 6.4. If this is not the case there is unlikely to be sufficient information in the operational data set to be able to identify the parameters.

The operational demand data sets have variable sampling times with a maximum sampling rate of 10Hz. A typical data set of demand observations contains several sampling times ranging from one second up to values of the order of minutes. For the same temporal window, corresponding grid frequency observations are sampled at 1Hz. These sampling rates are roughly a factor of 10 smaller than those used in the perfect model experiments of section 6.4.1 and in section 6.5.2. The third experiment requires a mixing period of $\tau_{\boldsymbol{\theta}} = 19500$ of the MCMC algorithm before the resulting distributions are considered to have converged [20].

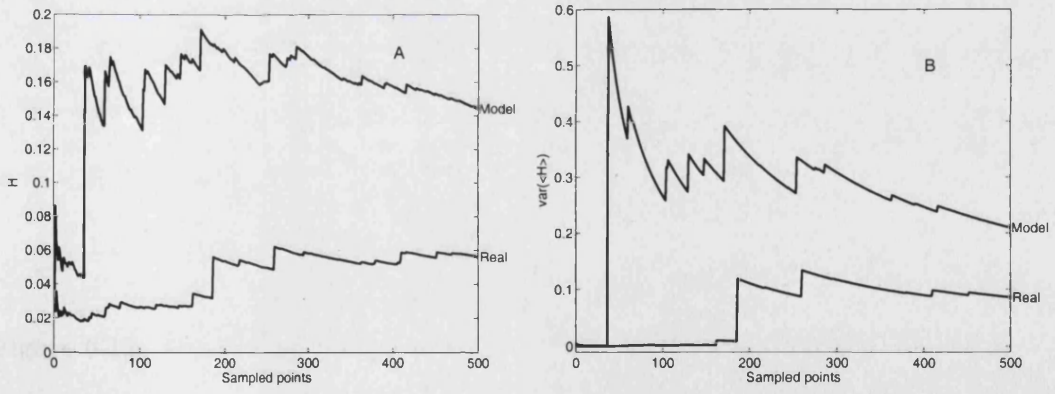


Figure 6.11: Summaries of the marginal posterior distributions for the inertia H using observed and operationally sampled model data after convergence is assesed. Plot A shows the mean of the posterior distributions. Plot B shows the variance of the posterior distributions.

Figure 6.11 summarises the estimates for the inertia H using observed data and operationally sampled model data. It is clear that the mean of the marginal posteriors are very far from 1 (in normalised values of the inertia), which corresponds to the value used to generate the model data. Moreover, the variance of the distributions is large compared to the perfect model experiment of section 6.4, Figure 6.10. Interpreting the estimated model parameters must be handled with care.

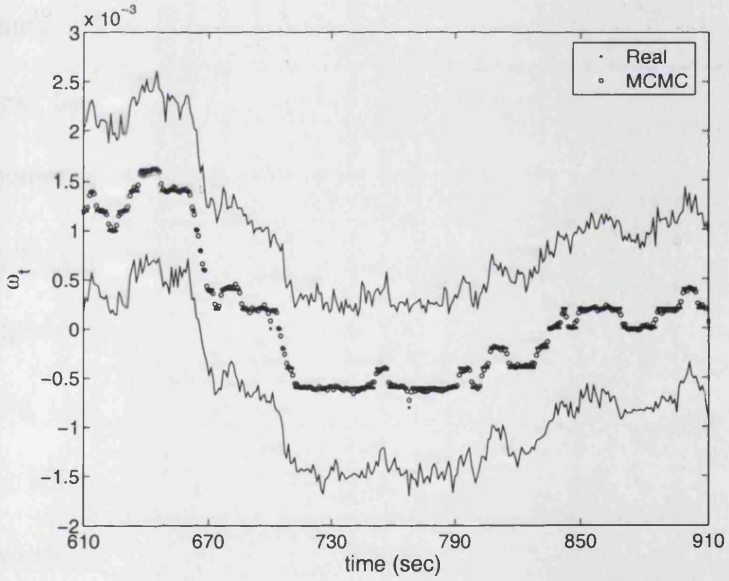


Figure 6.12: Grid frequency trace for real observations (dots) and mean of the MCMC output for the grid frequency states ω_t when using real observations. Solid lines correspond to the 99.5% and 0.5% percentiles of the distribution of states obtained for each time t.

Clearly the mean of the posterior for the model's grid inertia given sub-

sampled model data is far from the known value. The corresponding estimates given observed data behave in a similar fashion, the mean of the posterior is small compared to the perfect model experiments and the variance is large. The failure of the method to provide reasonable results for operationally sampled model data suggests that parameter estimates from the operational data streamsx are also compromised.

Figure 6.12 shows the output from MCMC using observed demand and grid frequency. Even though the uncertainty in the estimates of inertia are large the grid frequency traces obtained by sampling the posterior distribution are consistent with the observations used. The variance in the estimates of the grid frequency, although consistent, are larger than for the PMS. In practice, highly consistent behaviour in some components can mislead one into thinking that the method has converged to meaningful results for all components when, in fact, the marginal posterior distributions contain very little information.

6.7 Summary

In the context of the REMIND project, control monitoring of the grid frequency dynamics is translated into the problem of parameter estimation of a simple physical model presented in section 6.2 [20]. Model parameter esti-

mates are sought in order to based a characterisation of the system that can be incorporated into the National Grid Transco (NGT) management system.

Given the complexity of the grid system, model parameter estimation is performed in the ReMS, where the model adopted is only a representation of the system of interest it is assumed to be the Perfect Model. Therefore, the way parameter estimation is performed using any methodology is naively realistic (NRA) since some (if not all) parameters have no correspondent in the context of the system.

Observations of grid frequency and demand are provided by NGT. These observations reflect operational conditions of the grid system where data is gathered at 1Hz for grid frequency and demand is "observed" at variable sampling times. Uncertainty on the grid frequency dynamics is present in the available data therefore parameter estimation methodologies are used with the intent of obtaining meaningful parameter estimates in the limit that these operational conditions are reached.

Sensitivity of parameter estimates using traditional statistical methods, *i.e.* Least Squares, is studied for several data and model configurations via forward simulation. Section 6.4 shows that LS estimates are sensitive to the temporal resolution of the data and to coarse integration of the forward simulation; observational and model uncertainty is not accounted for. The "best guess" estimates fail to provide correct values when operational conditions

are reached.

In order to account for uncertainty on the model used to represent the system dynamics, a stochastic model is re-formulated based on the physical model developed in the context of Bayesian methodologies. Bayesian methodologies offer an alternative approach and seek to account for uncertainty in the parameter values, observations and model error by taking a naive probabilistic view of the system (NSA). As presented in section 6.5, the Bayesian methodology can, through the use of MCMC techniques instead of producing a point estimate of the parameter value, produce distributions of values that best resemble the data given a particular model.

section 6.5.1 describes one of the major thrust of this Thesis: A development of a Bayesian parameter estimation implementation using MCMC techniques for the deterministic model of the grid system.

Operational constraints on data quality, limit the application of MCMC in practice. Estimates of the parameters, and the uncertainty in those estimates, for the grid frequency dynamics can be achieved using Bayesian MCMC techniques given highly sampled data of demand and frequency. The quality of currently available operational data streams is critical: variations in the (multivariate) sampling rate lead to increases in the variance of the posterior estimates in the perfect model case.

Moreover, given operational sampling times, the posteriors are uninfor-

mative for model data. Similar results are observed when fitting the model to operational observations. The results presented in Figure 6.11 indicate how the real-world constraints limit this Bayesian parameter estimation technique. Although the results appear to be reasonable there is little information in the estimated parameter values in the perfect model scenario, making any interpretation of the operational data ambiguous. In short, there is insufficient information in the operational data stream to identify the parameters using this method.

Convergence in the posterior is not a sufficient test of relevance when applying Bayesian techniques to real world data. The perfect model experiment carried out above shows that the posterior distribution can converge even when the parameter values are effectively meaningless.

Further work will concentrate on the analysis of more complete operational data sets as they become available.

It is of fundamental interest to identify what exactly limits Bayesian techniques: the quality of the data, or structural errors in the underlying model. Resolving this issue on a case by case basis is of prime importance to the operational application of Bayesian techniques.

This work was supported by the EPSRC and National Grid Transco. Many thanks to Hai-Bin Wan and Ahmad Chebbo in NGT for providing the data and background information to develop the grid system model and

parameter estimation interpretation.

Liam Clarke contributed on the development of the project as part of the REMIND project and special acknowledgement goes to him for his continuous support, discussions and comments during the 3 years of the REMIND project.

Many thanks to Lenny Smith for his continuous commitment and support during the project and to Melvin Brown from the Smith Institute for acting as the technology translator of the corresponding Faraday project, and for acting as an "earth wire" between the REMIND project group at LSE and the contact group at NGT.

Chapter 7

Summary and Further Work

This Chapter summarises the results found and remarks are made in each of Chapters of this Thesis. A summary of new results and future work based in this research is provided.

Chapter 3. *Bayesian Inference and Chaotic Dynamics*

1. A correct Bayesian formulation of the problem of parameter estimation for the Logistic map is presented in the PMS when the system is under dynamical noise and the observations are noisy.
2. A posterior distribution for the Logistic map is written but shown to be numerically intractable due to the high order polynomial of the initial condition in the resulting exponential distribution.
3. Numerical intractability of the posterior agrees with early qualitative

properties of the "chaotic" Likelihoods described by Berliner in [5]. Multimodality and complex behaviour in the Logistic Likelihood are shown to depend explicitly on the initial condition and the length of the sequence of observations (see equation (3.36)).

4. Use of the Bayesian perspectives in the estimation of the parameter values of chaotic maps is correct when the system is under the influence of random perturbations, i.e. dynamical noise.

5. When the observations include only measurement noise, it is shown that the inclusion of a dynamical noise term is not a natural feature of the Bayesian approach but an artifact that makes the "chaotic" Likelihood numerically solvable by MCMC techniques.

6. The problem of parameter estimation presented by [68] is incorrectly formulated by Meyer and Christensen [70] in the Bayesian framework. Meyer and Christensen use a NSA to solve the problem in the PMS.

7. WinBUGS fail to provide convergent samples of the Logistic posterior.

8. Given the failure of WinBUGS to handle chaotic Likelihoods, numerical results presented in [70] and [12] are invalidated.

9. From this study, the WinBUGS development team took actions to correct deficiencies of the MCMC algorithm, specially, in order to cope with multimodal distributions.

10. A MCMC tailored implementation for the Logistic map was developed in order to produce posterior samples for the NSA of the Bayesian perspectives to solve the problem of parameter estimation for the Logistic map in the PMS.

11. A new sampling routine based on the Accept/Reject algorithm was developed to generate samples from the full conditional distributions of the initial condition and model states [23]. This routine is easily generalised to be used in any MCMC implementation of the Bayesian perspectives for quadratic maps.

12. Convergence of the Logistic posterior samples is improved by the MCMC tailored implementation.

13. Posterior estimates for the Logistic states generated for two types of noisy observations, statistically indistinguishable up to second order, resemble the Logistic structure in the delay reconstructed space even though one of the data types studied does not contain any deterministic component.

14. MCMC tailored implementation always generates pseudo-orbits of the Logistic map regardless the content of dynamical information of the observations used [24].

From the results of the study of the use of Bayesian methodologies in the nonlinear time series analysis framework, the derived further work includes:

1. Shadowing features of the MCMC technique are still to be studied and are partly tackled in Chapter 5. Is there any dependency of the width of the resulting posterior distributions if the surrogates correspond to Logistic observations in the chaotic or periodic regime? How can dynamical, (*i.e.* invariant measure), information be extracted from the estimates? Is it possible to find any trace of the deterministic/random behaviour in the estimates?

2. Reasons for good convergence of the MCMC parameter estimates for nonlinear systems are still to be found and they should be pursued further. Preliminary studies regarding this point include the work of Judd in [52].

Bayesian perspectives used in the nonlinear time series analysis framework provide insight in the importance of melding statistical and dynamical methodologies to find better parameter estimates. It explores the value of estimate distributions over "best guess" estimates. When estimate distributions are available, reliability and uncertainty measures are calculated from summary statistics. The NSA as the one use in Chapter 3 can also be used to find better forecasts and control monitoring techniques.

Chapter 4. *Distilling Information in the Parameter Space*

1. The use of pseudo-orbit states obtained by gradient descent [79] via indistinguishable states theory [54] instead of the original noisy observations

to estimate parameters using cost function based approaches produces better estimates that the traditional Maximum Likelihood estimates.

2. Summary statistics of the shadowing time distribution mapped into the parameter space yields more insight than maps of root-mean-square error (LS), which have well-known shortcomings in non-linear models (see [68] and references therein) as discussed in 4.1.

3. Figure 4.14 shows clearly how the maps of shadowing time provide complimentary information quantifying the time scales on which the model dynamics reflect the observed behaviour [88].

4. Better parameter estimates can be obtained by mapping summary statistics of the shadowing time distribution into the parameter space via effective quantification of the short and long term dynamics through the shadowing trajectories and the invariant measure respectively.

5. The balance between the information in the dynamic equations and the information in the observations exists only in the Perfect Model Scenario.

6. In the Imperfect Model Scenario, the invariant measure is not expected to be informative, only when other imperfect models are available.

7. This new method of parameter estimation, balances successfully the dynamical information of the model and the uncertainty in the observations. It is an example of the combination of nonlinear and statistical techniques

in order to obtain better parameter estimates for nonlinear models.

From the results of the study of how to distil dynamical information in the parameter space in the PMS, the derived further work includes:

1. Study of the information extraction from consideration of shadowing times, identification of parameter values which can mimic the dynamics, and quantification of the time scales on which the model can shadow the observations for improving the model parameter estimates when it is known to be imperfect.

2. In the light of the results obtained in this Thesis, the recast of the Bayesian perspectives in the nonlinear framework in the PMS, is calling for attention. Coherent Bayesian formulations that condition the probabilities extracted upon all information available [83, 4] are still to be developed.

3. Generalisations to more realistic and practical scenarios are still to be develop.

4. Refining the method to calculat shadowing time distributions when the noise model is unbounded.

5. Study the performance of the new geometric approach for a variety of nonlinear systems, model classes and noise levels to test robustness.

Methods which include explicitly dynamical information by the trajectories both model and observations admit, presented in Chapter 4, contribute to highlight the importance of melding different approaches whilst generating dynamical consistent estimates and clearly motivates further interest in the area.

Chapter 5. *Gradient Descent vs Markov Chain Monte Carlo*

1. The results presented in this Chapter are a milestone in the di of extracting useful dynamical information from ensembles of dynamical states in the PMS.
2. Through the process of parameter estimation, parameter estimates are generated along with estimates of model state estimates, as in the case of Total Least Squares, Gradient Descent and MCMC techniques.
3. A study of the value of state estimates, obtained through the process parameter estimation, as dynamical ensembles of states is proposed.
4. State estimates produced by MCMC, GD or TLS techniques explored the model state space consistently with the observations and the model chosen to represent the system of interest, here the Logistic map.
5. The quality "better" is defined based in the dynamical quality measures described in Chapter 4. Such quality measures include summary statis-

tics of empirical distributions of implied noise level, error between state estimates and true states, mismatch, and shadowing times.

6. From direct inspection of the resulting distributions, state estimates, for both MCMC and GD, the mean and median are close to the true states for all noise levels studied, see Figures 5.1 and 5.2 and Table 5.1.

7. Width of resulting distributions of state estimates and quality measures tend to be wider when MCMC is used than widths obtained using the GD algorithm.

8. The MCMC estimates are closer in median to the true trajectory than the GD estimates even though error distributions are up to 2 orders of magnitude wider than the width of GD errors.

9. Convergence of the state estimates to a pseudo-orbit close to the true trajectory is faster when the GD algorithm is used, MCMC algorithm shows a weak and slow convergence as commented in Chapter 4 for all noise levels studied.

10. Implied noise level and average implied noise level are interpreted as an estimation of the original noise level of the signal and can be used as an estimator of noise level in the time series in cases where it is unknown.

11. GD estimates reflect closely the original noise level in the signal proving the value of the GD algorithm as a noise reduction method [79], see

Figure 5.4.

12. Reasons of the over-performance of the MCMC techniques as a noise reduction method are still to be studied.

13. GD algorithm shows a robust convergence to a pseudo-orbit of the Logistic map for all noise levels studied when compared with MCMC results.

14. There is evidence that the pseudo-orbit obtained by MCMC techniques do not include information on the invariant measure of the Logistic map at least when the mismatch is calculated for both sets of estimates.

15. The results listed above confirm a difference in invariant measure informational content in the both sets of states estimates.

16. Surprisingly, histograms of the state estimates from both techniques are qualitatively similar to each other and the histogram of a very long true trajectory of the system.

17. Shadowing time distributions are planned to be calculated in future work for both, GD and MCMC, state estimates sets to measure the forecasting skill of each set and obtain insight on the nature of the pseudo-orbits obtained.

18. This study and future work has the potential to provide insight on how deterministic dynamical components induce dynamical and deterministic structures in empirical probability distributions. This study is directly

related to the work presented in 3.3 of Chapter 4.

Chapter 6. *Parameter Estimation from Real Time Series:*

The UK Electricity Grid Case

1. An implementation of the Bayesian methodology is obtained for the grid frequency model which via MCMC techniques produces consistent estimates in the PMS with high resolution model data.

2. Operational constraints on data quality limit the application of MCMC in practice.

3. Useful estimates of the parameters, and the uncertainty in those estimates, for the grid frequency dynamics can be achieved using Bayesian MCMC techniques given highly sampled data of demand and frequency.

4. The quality of currently available operational data sets prevents a complete characterisation/deployment of the present state of the grid system for forecasting and/or monitoring purposes. Variations in the (multivariate) sampling rate lead to increases in the variance of the posterior estimates in the perfect model case. Moreover, given operational sampling times the posteriors are uninformative for model data.

5. The perfect model experiments using the MCMC implementation and real data show that the posterior distribution converges to traces of the

system even when the model parameter values are not paired with system parameters.

6. Convergence in the posterior is not a sufficient test of relevance when applying Bayesian techniques to real world data.

Further work:

1. Analysis of some of the experiments posed in the PMS to test the performance of the MCMC technique that are not presented in this Thesis.
2. Analysis of more complete operational data sets as they become available.
3. Estimate parameters of the simple model developed for the grid system using geometrical approaches, *i.e.* the GD algorithm.

It is of fundamental interest to identify what exactly limits Bayesian techniques: the quality of the data, or structural errors in the underlying model. Resolving this issue on a case by case basis is of prime importance to the operational application of Bayesian techniques.

Finally, it can be said that any approach of parameter estimation has the potential to produce useful parameter estimates when properly formulated in a particular model scenario and methodology. If the model scenario is not stated before any attempt of parameter estimation, interpretation of resulting estimates are useless in the context of the system of interest.

The impossibility of knowing the perfect model of a system, finite and noisy observations, among other reasons, transforms the problem of parameter estimation in to challenging task where statistical and deterministic techniques are called upon to be melded. Uncertainty is the final source of information to obtain useful parameter estimates to then be used to produce informative forecasts and control monitoring strategies.

Efforts directed to develop methodologies that can produce reliable probabilistic parameter estimates, forecast, control monitoring strategies, among others are still scarce. This Thesis' principal contributions are aligned in this direction, to the goal of generating dynamical estimates from hybrid methodologies to effectively account for uncertainty in the modelling and characterisation of real systems.

References

[1] Electricity Act, *National Grid Transmission Licence*, 1989, Details in: `http://www.ofgem.gov.uk`.

[2] D. Applegate, R. Kannan, and N.G. Polson, *Random polynomial time algorithms for sampling from joint distributions.*, Technical Report 500, Carnegie-Mellon University., 1990.

[3] A. Bandrisvskyy, D.G. Luchinsky, P.V.E. McClintock, V.N. Smelyanskiy, and A. Stefanovska, *Inference of systems with delay and applications to cardiovascular dynamics*, Stochastics and Dynamics **5** (2005), no. 2, 321–331.

[4] T. Bayes, *An essay towards solving a problem in the doctrine of changes.*, Phil. Trans. Roy. Soc. London **53/54** (1958), 370–418/256–325, Published posthumously.

[5] L.M. Berliner, *Likelihood and Bayesian Prediction of Chaotic Systems*, J. Am. Statist. Ass. **86** (1991), no. 416, 938–952.

[6] ———, *Rejoinder (2): Statistics, Probability and Chaos*, Statist. Sci. **7** (1992), no. 1, 118–122.

[7] ———, *Statistics, Probability and Chaos*, Statist. Sci. **7** (1992), no. 1, 69–90.

[8] L.M. Berliner and S.N. Maceachern, *Examples of inconsistent Bayes procedures based on observations on dynamical-systems*, Statistics and Probability Letters **17** (1993), no. 5, 355–360.

[9] José M. Bernardo and Adrian F.M. Smith, *Bayesian Theory*, first ed., Wiley Series in Probability and Statistics, Wiley & Sons, Ltd., 2003.

[10] J.E. Borel, *Probability and certainty*, Physics and Mathematics, Walker, New York, 1963.

[11] R. Bowen and D. Ruelle, *The ergodic theory of axiom a flows.*, Inventiones Math. (1975), no. 29, 181–202.

[12] Christopher L. Bremer and Daniel T. Kaplan, *Markov Chain Monte Carlo estimation of non-linear dynamics from time series*, Physica D **160** (2001), 116–126.

[13] S.P. Brooks and A. Gelman, *General methods for monitoring convergence of iterative simulations*, Journal of Computational and Graphical Statistics **7** (1998), 434–455.

[14] B.P. Carlin and T.A. Louis, *Bayes and empirical Bayes methods for data analysis*, Chapman & Hall, London, 1997.

[15] Nancy Cartwright, *How the laws of physics lie*, Oxford University Press, 1983.

[16] M. Casdagli, S. Eubank, J.D. Farmer, and J. Gibson, *State space reconstruction in the presence of noise*, Physica D **51** (1991), 52–98.

[17] George Casella and Roger L. Berger, *Statistical Inference*, 2nd ed., Duxbury Advance Series, Thompson Learning, Duxbury, 2002.

[18] G. Cini Castagnoli and A. Provenzale (eds.), *International School of Physics "Enrico Fermi"*, vol. CXXXIII, Bologna, Societá Italiana di Fisica, IOS Press, 1997, See [86].

[19] Chris Chatfield, *The Analysis of Time Series: An Introduction*, Chapman & Hall, 2003.

[20] Liam Clarke and Milena C. Cuéllar, *REMIND Project. Grid Frequency System*, Tech. report, CATS, Department of Statistics, LSE, 2005.

[21] Personal communications by e mail and Imperial College. discussions with Andrew Thomas, 2002-2003, Personal Communication.

[22] Milena C. Cuéllar, Liam Clarke, Melvin Brown, and Leonard A. Smith, *The Role of the Operational Constraints on the MCMC Parameter Estimation: The case of the UK Electricity Grid*, International Journal of Energy and Power Sources (2005), 15p, Submitted to JASA.

[23] Milena C. Cuéllar and Luis Fernández, *An efficient algorithm for random sampling from quartic exponential distributions*, In preparation.

[24] Milena C. Cuéllar and Leonard A. Smith, *On the curious behaviour of MCMC in the Logistic map*, In preparation.

[25] C.D. Cutler, *Comment on Chaos and Statistics*, Statist. Sci. **7** (1992), no. 1, 91–94.

[26] M.E. Davies, *Nonlinear reconstruction by gradient descent*, Int. J. Bif. Chaos **3** (1992), no. 1, 113–118.

[27] ———, *Noise reduction schemes for chaotic time series*, Physica D **79** (1994), 174–193.

[28] M. Dowd and R. Meyer, *A Bayesian approach to the ecosystem inverse problem*, Ecological Modelling **168** (2003), no. 1-2 39 55, 17 pages.

[29] D. Gamerman, *Markov Chain Monte Carlo. Bayesian simulation for Bayesian inference*, 1st ed., Texts in Statistics, Eds. D. ChatfieD. Chatfield and J.V. Zidek. Chapman & Hall, London, UK, 1997.

[30] A.E. Gelfand and A.F.M. Smith, *Sampling based approaches to calculating marginal densities*, J. Am. Statist. Ass. **85** (1990), 398–409.

[31] A. Gelman and D.B. Rubin, *Inference from Iterative Simulation using Multiple Sequences (with discussion)*, Statist. Sci. **7** (1992), 457–511.

[32] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*, IEEE Trans. Patt. Anal. Mach. Intelligence **6** (1984), 721–740.

[33] J. Geweke, *Evaluating the accuracy of sampling-based approaches to calculating posterior moments.*, Bayesian Statistics (A.P. Dawid J.M. Bernardo, J.O. Berger and A.F.M. Smith., eds.), vol. 4, Oxford University Press., 1992.

[34] I. Gilmour, *Nonlinear model evaluation: ι-shadowing, probabilistic prediction and weather forecasting*, Ph.D. thesis, University of Oxford, 1998.

[35] C.W.J. Granger, *Comment on Chaos and Statistics*, Statist. Sci. **7** (1992), no. 1, 102–104.

[36] P. Grassberger, T. Schreiber, and C. Schaffrath, *Non-linear time sequence analysis*, Int J. Bifurcation and Chaos **1** (1991), 521–547.

[37] D. Greaffith, *Comment on Chaos and Statistics: Randomness in Complex Systems*, Statist. Sci. **7** (1992), no. 1, 104–108.

[38] G.R. Grimmett and D.R Stirzaker, *Probability and Random Processes*, Oxford University Press, 1992.

[39] D. Guegan, *Some Remarks on the Statistical Modelling of Chaotic Systems*, Non-linear Dynamics and Statistics, ch. 5, pp. 125–126, Birkhäuser, New York, 2001.

[40] W.K. Hastings, *Monte Carlo Sampling Methods using Markov Chains and their applications*, Biometrika **57** (1970), no. 1, 97–109.

[41] S. Haykin, R. Bakker, and B.W. Currie, *Uncovering nonlinear dynamics - the case study of a sea clutter*, Proceedings of the IEEE, vol. 90 (5), May 2002, pp. 860–881.

[42] J.P.M. Heald and J. Stark, *Nonlinear noise reduction - Bayesian perspective*, 1998 International Symposium on Nonlinear Theory and its Applications (Lausanne, Switzerland), vol. 3, 1998, pp. 1293–1296.

[43] ———, *Estimation of noise levels for models of chaotic dynamical systems*, Phys. Rev. Lett. **84** (2000), no. 11, 2366–2369.

[44] P. Heidelberger and P. Welch, *Simulation run length control in the presence of an initial transient.*, Operations Research **31** (1983), 1109–1144.

[45] M. Henon, *A two dimensional mapping with a strange attractor*, Communication in Mathematical Physics **50** (1976), 69–77.

[46] Paul Hurlock, *Briefing Note on Frequency Response Provision*, Tech. report, National Grid, UK, June 2003.

[47] K. Ikeda, *Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system*, Optical Communications **30** (1979), 257–261.

[48] L. Jaeger and H. Kantz, *Homoclinic tangencies and non-normal jacobians – effects of noise in non-hyperbolic chaotic systems*, Physica D **105** (1997), 79–96.

[49] C. Jennison, *Discussion of "Bayesian computation via the gibbs sampler and related markov chain monte carlo methods" in [85]*, Journal of the Royal Statistical Society, Series B **55** (1993), 54–56.

[50] K. Judd, C.A. Reynolds, and T.E. Rosmond, *Toward shadowing in operational weather prediction.*, Tech. Report NRL/MR/7530-04018, Naval Research Laboratory, 2004.

[51] K. Judd, L.A. Smith, and A. Weisheimer, *Gradient free descent: shadowing and state estimation with limited derivative information*, Physica D **190** (2004), 153–166.

[52] Kevin Judd, *Chaotic-time-series reconstruction by the Bayesian paradigm: Right results by wrong methods*, Phys. Rev. E **67** (2003), 026212.

[53] ———, *Nonlinear state estimation, indistinguishable states and the extended Kalman filter*, Physica D **183** (2003), 273–281.

[54] Kevin Judd and Leonard A. Smith, *Indistinguishable states I. Perfect Model Scenario*, Physica D **151** (2001), 125–141.

[55] ———, *Indistinguishable states II. The Imperfect Model Scenario*, Physica D **196** (2004), 224–242.

[56] Holger Kantz and Thomas Schreiber, *Nonlinear time series analysis*, 1^{st} ed., Cambridge Nonlinear Science Series 7, Eds. B. Chirikov, P. Cvitanović, F. Moss and H. Swinney. Cambridge University Press, Cambridge, UK, 1997.

[57] E. Kostelich and T. Schreiber, *Noise reduction in chaotic time-series data: A survey of common methods*, Phys. Rev. E **48** (1993), no. 3, 1752–1763.

[58] E.J. Kostelich and J.A. Yorke, *Noise reduction in dynamical systems*, Phys. Rev. A **38** (1988), 1649–1652.

[59] Thomas S. Kuhn, *The structure of scientific revolutions*, vol. 2, International encyclopedia of unified science, no. 2, Chicago : University of Chicago Press, 1962.

[60] S. Lele, *Estimating functions in chaotic systems*, J. Am. Statist. Ass. **89** (1994), no. 426, 512–516.

[61] C. Liu and J. Liu., *Discussion on the on the Gibbs sampler and other Markov chain Monte Carlo methods*, J.R. Statist. Soc. B **55** (1993), 82–83.

[62] D.G. Luchinsky, M.M. Millonas, V.N. Smelyankiy, A. Pershakova, A. Stefanovska, and P.V.E. McClintock, *Nonlinear statistical modeling and model discovery ofr cadiorespiratory data*, Phys. Rev. E **72** (2005), no. 2, 021905, Part 1.

[63] S.N. Maceachern and L.M. Berliner, *Asymptotic inference for dynamical systems observed with error*, Journal of Statisticsl Planning and Inference **46** (1995), no. 3, 277–292.

[64] Ernest Mach, *Knowledge and error: Scketches on the psychology of enquiry*, Vienna Circle Collection, ch. 15, p. 208, Ed. Brian F. McGuinness.

Springer, December 1975, "There is no way of proving the correctness of the position 'determinism' or 'indeterminism'. Only if science were complete of demonstrably impossible could we decide such questions. These are presuppositions we bring to the consideration of things...".

[65] David J.C. MacKay, *Information theory, inference and learning algorithms*, Cambridge University Press, 2003.

[66] Neal Noah Madras, *Lecture notes in monte carlo methods*, vol. viii, Fields Institute Monographs Series, no. 16, American Mathematical Society, 2002.

[67] R.M. May, *Simple mathematical models with very complicated dynamics*, Nature **261** (1976), 459.

[68] Patrick E. McSharry and L.A. Smith, *Better Nonlinear Models from Noisy Data: Attractors with Maximum Likelihood*, Phys. Rev. E **83** (1999), no. 21, 4285–4288.

[69] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, *Equation of State Calculations by Fast Computing Machines*, J.Chem.Phys. **21** (1953), 1087–1092.

[70] R. Meyer and N. Christensen, *Bayesian Reconstruction of Chaotic Dynamical Systems*, Phys. Rev. E **62** (2000), no. 3, 3535–3542.

[71] Radford M. Neal, *Slice sampling*, Ann. Statist. **31** (2003), no. 3, 705–767.

[72] William of Ockham, *Summa logicae*, Guillelmi de Ockham opera philosophica et theolog- ica, cura Instituti Franciscani Universitatis S. Bonaventurae. Edidit Stephanus Brown, adlaborante Gedeone Gl. S. Bonaventure, N.Y.: a impressa Ad Claras Aquas (Italia) 1967-1985. 1st publish in Paris 1448. The Summa logicae is not completely available in English yet: Part I is translated in Loux (1974) and Part II in Freddoso (1980). Part III has not yet been translated in English.

[73] E. Ott, *Chaos in dynamical systems*, Cambridge University Press, 1993.

[74] T. Ozaki, J.C. Jimenez, and V. Haggan-Ozaki, *The orle of the Likelihood function in the estimation of chaotic models*, Journal of Time Series Analysis **21** (2000), no. 4, 363–387.

[75] Tim N. Palmer, *Predicting Uncertainty in Forecasts of Weather and Climate*, Reports on progress in Physics **63** (2000), 71–116.

[76] V.F. Pisarenko and D. Sornette, *On Statistical Methods of Parameter Estimation for Deterministic Chaotic Series*, Phys. Rev. E **69** (2004), no. 036122, 12 pages.

[77] William H. Press, Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C: The art of scientific computation*, second ed., Cambridge University Press, Cambridge, 1999.

[78] A.L. Raftery and S. Lewis, *How many iterations in the Gibbs sampler?*, Bayesian Statistics (A.P. Dawid J.M. Bernardo, J.O. Berger and A.F.M. Smith., eds.), vol. 4, Oxford University Press., 1992.

[79] D. Ridout and K. Judd, *Convergence properties of gradient descent noise reduction*, Physica D **165** (2001), 27–48.

[80] C. Ritter and M.A. Tanner, *Facilitating the Gibbs sampler. the Gibs sampler and the Griddy-Gibbs sampler*, J. Am. Statist. Ass. **87** (1992), 861–868.

[81] D.B. Rubin, *Using the SIR Algorithm to Simulate Posterior Distributions*, Bayesian Statistics 3 (J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, eds.), Oxford University Press, 1988, with discussion, pp. 305–402.

[82] T. Sauer, J. A. Yorke, and M. Casdagli, *Embedology*, Journal of Statistical Physics **65** (1991), no. 3-4, 579–616.

[83] D.S. Sivia, *Data analysis: A bayesian tutorial*, Oxford University Press, October 1996.

[84] V.N. Smelyanksiy, D.G. Luchinsky, D.A. Tumicin, and A. Bandrivskyy, *Reconstruction of stochastic nonlinear dynamical models from trajectory measurements*, Phys. Rev. E **72** (2005), no. 2, 026202, Part 2.

[85] A.F.M. Smith and G.O. Roberts, *Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods*, J.R. Statist. Soc. B **55** (1993), no. 1, 3–23.

[86] Leonard A. Smith, *The maintenance of uncertainty*, in Castagnoli and Provenzale [18], See [86], pp. 177–246.

[87] ———, *Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems*, Nonlinear Dynamics and Statistics (Alister Mees, ed.), Birkhauser, 2000.

[88] Leonard A. Smith, Milena C. Cuéllar, Hailiang Du, and Kevin Judd, *Identifying dynamically coherent behaviours: A geometrical approach to parameter estimation in nonlinear models*, In preparation, 2005.

[89] Leonard A. Smith and Du Hailaing, *Berliner's paper solution*, In preparation.

[90] A.R. Solow, *On fitting a population model in the presence of observation error*, Ecology **79** (1998), no. 4, 1463–1466.

[91] L. Tierney, *Markov Chains for Exploring Posterior Distributions*, Ann. Statist. **22** (1994), 1701–1762.

[92] J. Timmer, *Modeling noisy time series: Physiological tremor*, Int. J. Bif. Chaos **8** (1998), 1505–1516.

[93] R.S. Tsay, *Comment on Chaos and Statistics: Simplicity and Nonlinearity*, Statist. Sci. **7** (1992), no. 1, 113–114.

[94] H.U. Voss, J. Timmer, and J. Kurths, *Nonlinear dynamical system identification from uncertain and inderect measurements*, International Journal Bifurcations and Chaos **16** (2004), no. 6, 1905–1933.

[95] S. Richardson W.R. Gilks and D.J. Spiegelhalter (eds.), *Markov chain Monte Carlo in practice*, first ed., Interdisciplinary Statistics, Chapman & Hall, London, 1997.

Glossary

$C_{ML}(\boldsymbol{\theta})$	Maximum Likelihood cost function.
$C_{MM}(\boldsymbol{\theta})$	Mismatch cost function.
D	Demand, model variable.
G	Generation, model variable.
S	Set of observations, time series.
$\mathcal{Q}$	Parameter space.
ω	Grid Frequency, model variable.
$\omega_0 = 2\pi 50$	Operational electricity grid frequency.
$\tilde{\omega}(t)$	Instantaneous grid frequency.
$\tilde{D}(t)$	Instantaneous demand.
$\tilde{G}(t)$	Instantaneous generation.
u_t	Indistinguishable state.
z_t	Pseudo-orbit state.
$\boldsymbol{\theta}$	Parameter vector

AINL Average Implied Noise Level.

BUGS Bayesian inference Using Gibbs Sampling.

GD Gradient Descent

IID Identically and Independently Distributed.

IMS Imperfect Model Scenario.

INL Implied Noise Level.

LS Least Squares.

MCMC Markov Chain Monte Carlo

MH Metropolis-Hastings

NAR Nonlinear Auto-Regressive process.

NGT National Grid Transco Plc.

NRA Naive Realist Approach.

NSA Naive Statistical Approach.

PDF Probability Density Function.

PMS Perfect Model Scenario.

REMIND Real-timE Modelling of NonlInear Data-streams.

ReMS Real Model Scenario.

RMS Root Mean Square.

TLS Total Least Squares cost function.

Subject Index

Index of Figures

Index of Tables